e-Infrastructures for data publishing in biodiversity science

Edited by

Vincent Smith & Lyubomir Penev

PENSOFT

Sofia–Moscow

2011

ZooKeys 150 (Special Issue)

e-Infrastructures for data publishing in biodiversity science

Edited by Vincent Smith & Lyubomir Penev

First published 2011
ISBN 978-954-642-619-2 (paperback)

Pensoft Publishers
Geo Milev Str. 13a, Sofia 1111, Bulgaria
Fax: +359-2-870-42-82
info@pensoft.net
www.pensoft.net

Printed in Bulgaria, November 2011

Contents

ZooKeys 150: 1–3 (2011)
doi: 10.3897/zookeys.150.2458
www.zookeys.org

EDITORIAL

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Collaborative electronic infrastructures to accelerate taxonomic research

Vincent S. Smith[1], Lyubomir Penev[2]

1 *Natural History Museum, Cromwell Road, London, SW7 5BD, U.K.* **2** *Pensoft Publishers, Sofia, Bulgaria*

Corresponding author: *Vincent S. Smith* (vince@vsmith.info)

Received 27 November 2011 | Accepted 28 November 2011 | Published 28 November 2011

Citation: Smith VS, Penev L (2011) Collaborative electronic infrastructures to accelerate taxonomic research. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 1–3. doi: 10.3897/zookeys.150.2458

The discipline of taxonomy lives in a state of perpetual beta, constantly evolving as species hypotheses change to reflect the latest character evidence. Likewise, the electronic infrastructures that underpin taxonomic research are evolving to reflect the latest technical advances. This collection of articles illustrates how advances to research infrastructures are reciprocally changing the practice of taxonomy. The infrastructures have, in part, been developed through the EU funded ViBRANT project. This is the latest in a series of EU funded initiatives that aim to move taxonomy towards a more collaborative, electronic framework in an effort to accelerate the pace of biodiversity research. Through a platform of web-based informatics tools and services, ViBRANT project partners are building a framework that allows distributed groups of scientists to create their own virtual research communities that support biodiversity science. To mark the first year of the ViBRANT project we have brought together a set of reviews, essays and research articles that reflect some of the project highlights and illustrate a number of associated activities. Fittingly, many of the contributions are from researchers who have no direct support from the ViBRANT project, but have used or reviewed some aspect of the infrastructure. This marks an important transition from many EU funded infrastructure projects have typically focused on technical developments, and less on the communities that use these systems.

A detailed review of data issues in the life sciences (Thessen and Patterson 2011) sets the tone for subsequent articles in this special issue, whose contributions broadly fall into three categories. The initial articles consider some of the major infrastructure

platforms that support the production and management of biodiversity data. These include the EDIT Platform for Cybertaxonomy, Wiki-based approaches including BioWikiFarm and the Scratchpads Virtual Research Environment. Later articles provide deeper coverage of specialist areas of interest to taxonomic and biodiversity researchers. The topics covered include the mark-up (Penev et al. 2011) and management (King et al. 2011) of taxonomic literature, geospatial assessment of species distributions (Bachman et al. 2011) and licensing issues specific to life science data (Hagedorn et al. 2011). Finally, the special issue closes with a series of research and review papers that provide detailed use cases illustrating how these research infrastructures are being put into practice. These articles make up the majority of this special issue and are subdivided into the sociological analysis of how people are using these infrastructures, as well as the practical experience of biodiversity researchers developing taxonomic data with these systems. Highlights from this section include citizen science approaches to collecting species information by the COMBER Marine observation network (Arvanitidis et al. 2011) and the Australian Bush Blitz programme (Lambkin and Bartlett 2011); use of new tools for data publishing like the Global Biodiversity Information Facility (GBIF) Integrated Publishing Toolkit (IPT) and the DRYAD Data Repository; new forms of publication via "data papers" that allow checklists and identification keys to be formally published as structured datasets (e.g., Narwade et al. 2011); and finally new taxonomic revisions and species descriptions constructed from within the collaborative systems like XPER^2 and Scratchpads.

This diverse collection of articles illustrates how the paradigm of scholarly communication in taxonomy is being changed by new electronic infrastructures. These support new ways to collaborate and disseminate taxonomic information, facilitating greater reuse of the underlying data. The infrastructures described here are in many cases experimental, but illustrate a number of possible trajectories for how taxonomic data might be assembled and disseminated in the future. These infrastructures will continue to change and evolve, but it now seems certain that the future of taxonomy is increasingly digital, to the point that non-digital work is becoming invisible and perhaps irreverent to the next generation of scholars.

Acknowledgements

We sincerely thank the authors and reviewers of these articles who have responded, often at very short notice, to our requests for assistance. We would also like to thank all members of the ViBRANT consortium, and in particular Dave Roberts, without whose help the compilation of this special issue would have been impossible. VS is indebted the Scratchpad team at the Natural History Museum (Simon Rycroft, Irina Brake, Ben Scott, Edward Baker, Laurence Livermore and Vladimir Blagoderov), and in particular his partner Helen Whitaker, who endured a "holiday" with VS while the final version of these manuscripts was being edited. This work was supported by the EU funded FP7 ViBRANT project (contract number RI-261532).

References

Arvanitidis C, Faulwetter S, Chatzigeorgiou G, Penev L, Bánki O, Dailianis T, Pafilis E, Kouratoras M, Chatzinikolaou E, Fanini L, Vasileiadou A, Pavloudi C, Vavilis P, Koulouri P, Dounas C (2011) Engaging the broader community in biodiversity research: the concept of the COMBER pilot project for divers in ViBRANT. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 211–229. doi: 10.3897/zookeys.150.2149

Bachman S, Moat J, Hill AW, de la Torre J, Scott B (2011) Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 117–126. doi: 10.3897/zookeys.150.2109

Hagedorn G, Mietchen D, Morris RA, Agosti D, Penev L, Berendsohn WG, Hobern D (2011) Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 127–149. doi: 10.3897/zookeys.150.2189

King D, Morse DR, Willis A, Dil A (2011) Towards the bibliography of life. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 151–166. doi: 10.3897/zookeys.150.2167

Lambkin CL, Bartlett JS (2011) Bush Blitz aids description of three new species and a new genus of Australian beeflies (Diptera, Bombyliidae, Exoprosopini). In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 231–280. doi: 10.3897/zookeys.150.1881

Narwade S, Kalra M, Jagdish R, Varier D, Satpute S, Khan N, Talukdar G, Mathur VB, Vasudevan K, Pundir DS, Chavan V, Sood R (2011) Literature based species occurrence data of birds of northeast India. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 407–417. doi: 10.3897/zookeys.150.2002

Penev L, Lyal CHC, Weitzman A, Morse DR, King D, Sautter G, Georgiev T, Morris RA, Catapano T, Agosti D (2011) XML schemas and mark-up practices of taxonomic literature. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 89–116. doi: 10.3897/zookeys.150.2213

Thessen AE, Patterson DJ (2011) Data issues in the life sciences. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 15–51. doi: 10.3897/zookeys.150.1766

ZooKeys 150: 5–14 (2011)
doi: 10.3897/zookeys.150.2431
www.zookeys.org

EDITORIAL

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

ZooKeys 150: Three and a half years of innovative publishing and growth

Terry Erwin[1], Pavel Stoev[2], Teodor Georgiev[3], Lyubomir Penev[2]

1 *Smithsonian Institution, Washington, DC, USA* **2** *Bulgarian Academy of Sciences & Pensoft Publishers, Sofia, Bulgaria* **3** *Pensoft Publishers, Sofia, Bulgaria*

Corresponding author: *Lyubomir Penev* (info@pensoft.net)

Received 21 November 2011 | Accepted 23 November 2011 | Published 28 November 2011

Citation: Erwin T, Stoev P, Georgiev T, Penev L (2011) ZooKeys 150: Three and a half years of innovative publishing and growth. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 5–14. doi: 10.3897/zookeys.150.2431

'ZooKeys publishes articles of the future'
Roderic Page, title of a blog post in iPhylo

On the 28th of November 2011, the open access journal ZooKeys published its 150th issue – an excellent occasion for the Editorial team to evaluate the journal's development and its position among systematic biology journals worldwide.

From the very beginning, ZooKeys was designed as an innovative journal aiming at developing new methods of publication and dissemination of taxonomy information, including publishing of atomized, semantically enhanced automated exports to global data aggregators, such as Encyclopedia of Life (EOL), the Global Biodiversity Information Facility (GBIF), Plazi, Species-ID and others. Since its launch on the 4th of July 2008, the journal provided registration of all new taxa and authors in ZooBank on a mandatory basis and continues to include their Life Science Identifiers (LSID) in the published articles (Penev et al. 2008). Also since its first issue, ZooKeys made it a routine practice of supplying all new taxa to the Encyclopedia of Life through XML mark up. In the subsequent years, the journal joined GBIF and the Taxonomic Databases Working Group (TDWG) in the development of common data publishing standards and workflows.

In 2009, ZooKeys initiated several pilot projects thereby setting foundations of semantic tagging of, and enhancements to, biodiversity articles using the TaxPub XML

schema, an extension of the DTD (Document Type Definitions) of the National Library of Medicine (USA) (Penev et al. 2009a; Catapano 2010). The first one was the milestone article 'The symphytognathoid spiders of the Gaoligongshan, Yunnan, China' (Miller et al. 2009) where, for the first time in systematic zoology, a unique combination of data publication and semantic enhancements was applied within the mainstream process of journal publishing. The article demonstrated how all primary biodiversity data underlying a taxonomic monograph could be published as a dataset under a separate DOI within the paper and the occurrence dataset could be integrated and accessed through GBIF data portal simultaneously with the publication. In the same year, data publication practices of online identification keys (Penev et al. 2009b) were exemplified by the pioneering articles of Sharkey et al. (2009) and that was shortly followed by others (van Noort and Johnson 2009; Stoev et al. 2010).

On the 30th of June 2010, ZooKeys published a special issue 'Taxonomy shifts up a gear: New publishing tools to accelerate biodiversity research' which marked the journal's brand new innovative publishing model, based on XML editorial workflow and on the TaxPub XML schema. From that time on, ZooKeys has been published in four formats – full-colour print version, PDF, HTML, and XML (Penev et al. 2010a). This happened simultaneously with the implementation in the editorial process of the Pensoft Mark Up Tool (PMT), a program specially designed for XML tagging and semantic enhancements (Penev et al. 2010b). Four papers using three different types of manuscript submission (Stoev et al. 2010; Blagoderov et al. 2010; Brake and Tschirnhaus 2010; Taekul et al. 2010) were used to exemplify the process.

Realizing the importance of Wiki environment for popularization and dissemination of the biodiversity data, in April 2011 ZooKeys undertook another major step towards its modernization. Three sample papers (Hendriks and Balke 2011; Stoev and Enghoff 2011; Bantaowong et al. 2011) demonstrated the automated integration of species descriptions at the day of publication to Species-ID – an open access Wiki-based resource for biodiversity information. This was achieved by programming a special tool, named Pensoft Wiki Convertor (PWC), which transforms the XML versions of the papers into MediaWiki-based pages (Penev et al. 2011a).

In October 2011, ZooKeys launched its *multiple-choice model* for publishing biodiversity data that provides a non-exclusive choice of mechanisms for the publication of data of different kinds and complexity, in cooperation with specialized data repositories and data aggregators, based on the previously published Pensoft Data Publishing Policies and Guidelines for Biodiversity Data (Penev et al. 2011b) One of the most important steps in this direction was the launch of an innovative route for publishing occurrence data and taxon checklists using an approved TDWG standard (Darwin Core), enriched metadata descriptions for the published datasets, and the possibility of downloading both data and metadata in a machine-readable form, the so-called Darwin Core Archive. This is supported by a specialized tool of GBIF, the Integrated Publishing Toolkit (IPT). Use of this tool allows the production of so-called "Data Paper" manuscripts that formally describe a dataset's metadata as a peer-reviewed and citable scholarly publication (Chavan and Penev in press).

A second important element of the *multiple-choice* data publishing model of ZooKeys was the integration of its data publishing workflow with the Dryad Digital Repository, thus providing an option to its authors to archive data files of different kinds and complexity (e.g., phylogenetic, morphometric, ecological, environmental, etc.).

The latest innovation of ZooKeys was announced just a few days before publication of this editorial. On the 22nd of November 2011, ZooKeys launched an automated export and indexing of identification keys metadata published in the journal in KeyCentral – a global database of keys and other identification resources for living organisms.

ZooKeys has shown a significant publication growth for the 41 months of its existence (Fig. 1). Starting with a mere 32 articles in 2008, the journal has rapidly increased its production to 180 in 2010 and 413 in 2011 (through the 28th of November). Likewise the number of published pages has grown from 657 in 2008, 3,738 in 2009, 4,831 in 2010 to 10,082 in 2011. The growth rate for 2011 in comparison to 2010 in the number of published pages is more than 100% and will most probably exceed 120% by the end of the year. For three and a half years, ZooKeys has published overall 19,308 pages (780 articles), a figure that is comparable to the number of pages published by Zootaxa during its first 41 months of activity (16,738 pages – see Zhang 2011 and http://www.mapress.com/zootaxa).

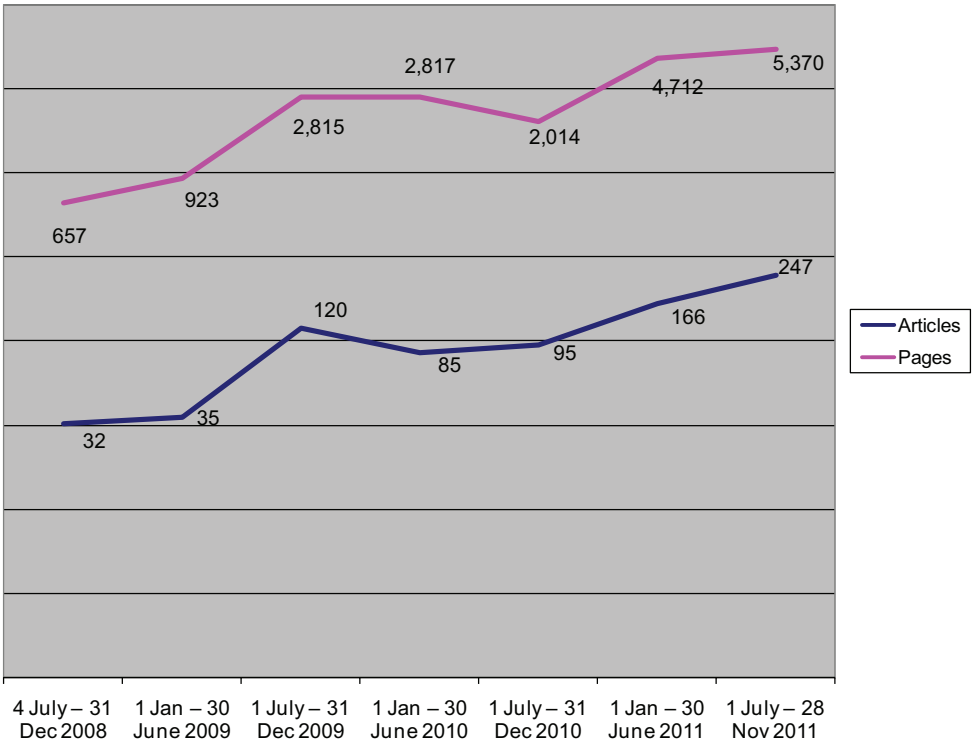


Figure 1. Total number of published articles and pages on six-month intervals.

Altogether, 1,558 new species-group, 192 new genus-group and 16 new family-group taxa have been published in the journal since its launch (Table 1). This makes overall 1,766 new taxa in total, or 43 new taxa per month on average. Comparing these figures with the Index of Organism Names of Zoological Record (accessed 18 November 2011) ZooKeys has published approximately 2.5% of all the 69,224 new animal taxa described from 2008 to 2011, and ranks second (immediately after Zootaxa) in the top 10 journals publishing new taxa. The data retrieved from ZooBank show that one third of all new names registered in ZooBank since June 2008 have been published in ZooKeys. The total number of ZooKeys authors registered in ZooBank up to issue 148 reached 754 (Richard Pyle, in litt.).

Table 1. New taxa published in ZooKeys that have been registered and assigned LSIDs in ZooBank (data for issues 1-148 provided by Richard Pyle, in litt.).

Categories	Number
Species-group names	1,558
Genus-group names	192
Family-group names	16
Total	**1,766**

Figure 2 summarizes the distribution of articles per large taxon. Unsurprisingly, the highest number of articles published in ZooKeys dealt with insects (584). The articles on Coleoptera (249) dominate and together with those dealing with Hymenoptera (122) make up approximately 48% of all ZooKeys articles. Those on Lepidoptera (77), Hemiptera (42) and Diptera (39) also form a significant share of the published volumes. Among the non-insect invertebrates the highest number of articles were pub-

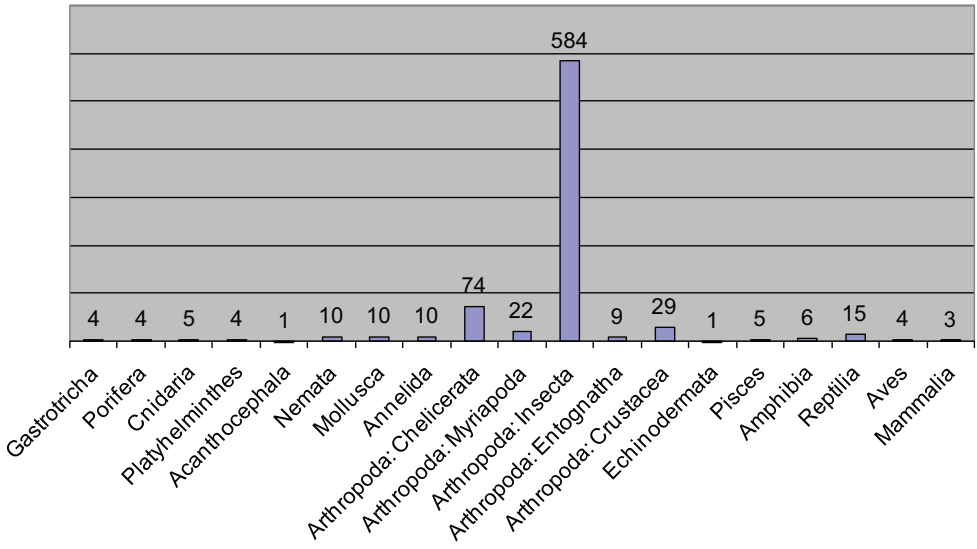


Figure 2. Distribution of the published articles per taxon.

lished on Chelicerata (74), followed by Crustacea (29) and Myriapoda (22). The total number of articles dealing with vertebrates is comparatively low (33), nearly half of them refer to reptiles (15).

The top 10 most accessed ZooKeys papers through the 20th of November 2011 are listed in Table 2. The 972 page monograph of Bouchard et al. (2011) 'Family-Group names in Coleoptera (Insecta)' is taking the first place reaching 8,623 page views on the 20th of November. In the top 3 most viewed articles are also the 'Data publication and dissemination of interactive keys' (Penev et al. 2009) and 'Cretaceous Crocodyliforms from the Sahara' (Sereno and Larsson 2009), with 7,716 and 6,275 page views, respectively.

Table 2. Top ten most viewed articles of ZooKeys (according to the ZooKeys website counter accessed on the 20th of November 2011).

Article	Page views
Bouchard et al. 2011 – Family-Group names in Coleoptera (Insecta)	8,623
Penev et al. 2009 – Data publication and dissemination of interactive keys under the open access model	7,716
Sereno and Larsson 2009 – Cretaceous Crocodyliforms from the Sahara	6,275
Baldwin et al. 2011 – Seven new species within western Atlantic *Starksia atlantica*, *S. lepicoelia*, and *S. sluiteri* (Teleostei, Labrisomidae), with comments on congruence of DNA barcodes and species	5,283
Achterberg and Long 2010 – Revision of the Agathidinae (Hymenoptera, Braconidae) of Vietnam, with the description of forty-two new species and three new genera	5,107
Hendrich and Balke 2011 – A simultaneous journal / wiki publication and dissemination of a new species description: *Neobidessodes darwiniensis* sp. n. from northern Australia (Coleoptera, Dytiscidae, Bidessini)	3,986
Hong et al. 2011 – A revision of the Chinese Stephanidae (Hymenoptera, Stephanoidea)	3,888
Wizen and Gasith 2011 – Predation of amphibians by carabid beetles of the genus *Epomis* found in the central coastal plain of Israel	3,818
Heads and Leuzinger 2011 – On the placement of the Cretaceous orthopteran *Brauckmannia groeningae* from Brazil, with notes on the relationships of Schizodactylidae (Orthoptera, Ensifera)	3,655
Murphy et al. 2011 – The dazed and confused identity of Agassiz's land tortoise, *Gopherus agassizii* (Testudines: Testudinidae) with the description of a new species and its consequences for conservation	3,653

In order to increase public awareness to the importance of taxonomy and biodiversity studies in general, in May 2011 Pensoft opened a press office and started active public relations (PR) activities. Authors are invited to draft press releases on their findings at the moment of acceptance of their publications. The Pensoft PR team offers support to the authors in "translating" the technical texts into a language that would be of interest for the public. Press releases are posted to a number of sites; the first place, EurekAlert!, is the world largest online distributor of science news supplying information to more than 7,500 mass media and independent science journalists. A list of the top 10 most accessed press releases of ZooKeys articles is

given in Table 3. The press release on the new Late Cretaceous family of wasps, Plumalexiidae, described in a Festschrift honouring the Russian paleontologist Alexandr Rasnitsyn has hitherto attracted the highest attention in the world media. Of similar high popularity in the world news outlets was the unique observation of oviposition behaviour of four ant parasitoids that was filmed for the first time and movies uploaded in YouTube (Durán and Achterberg 2011). Another ZooKeys article showing *Epomis* beetles preying on amphibians (Wizen and Gasith 2011) whose associated movies were posted on YouTube have been watched 344,325 times in 6 months. This is further evidence that taxonomic discoveries enjoy a lot of interest from the public, if they are properly and attractively distributed.

Table 3. Top 10 most accessed press releases of ZooKeys articles posted through EurekAlert! (from the EurekAlert! counter). The counter registers only the downloads from EurekAlert! mostly by science media and journalists. The actual number of readers may actually be much higher than this number.

Title	Author/s and year of publication of the original article	Date posted	Page views since posted
New family of wasps found in North American amber, closest relatives in southern hemisphere	Brothers 2011	26-Sep-2011	3,412
Death from above: Parasite wasps attacking ants from the air filmed for the first time	Durán and Achterberg 2011	29-Aug-2011	2,749
A living species of aquatic beetle found in 20-million-year-old sediments	Fikáček et al. 2011	6-Oct-2011	2,676
Chinese researchers identify insect host species of a famous Tibetan medicinal fungus	Wang and Yao 2011	8-Sep-2011	2,340
Small insects attacks and kill amphibians much bigger than themselves	Wizen and Gasith 2011	20-May-2011	2,309
A new species of fossil silky lacewing insects that lived more than 120 million years ago	Peng et al. 2011	5-Oct-2011	2,203
Jewel beetles, obtained from local people, turn out to be 4 species unknown to science	Bílý and Nakládal 2011	7-Jul-2011	1,921
A new species of a tiny freshwater snail collected from a mountainous spring in Greece	Radea 2011	1-Nov-2011	1,885
Unknown species and larval stages of extremely long-legged beetles discovered by DNA test	Freitag and Balke 2011	18-Oct-2011	1,437
Earliest psychomyiid caddisfly fossils, from 100-million-year-old Burmese amber	Wichard et al. 2011	5-Oct-2011	1,350

ZooKeys represents a new type of a journal whose mission is to create new horizons for taxonomists through modern technology and widespread promulgation of biodiversity data. Thanks to its continuously applied innovations, and especially owing to the commitment of its professional editorial team, the journal will continue

to facilitate and accelerate biodiversity research at the same pace, along with its sister journals PhytoKeys and MycoKeys. We sincerely thank all editors and reviewers for their selfless support and professional editorial work, as well as our hundreds of friends and colleagues that have been actively discussing with us and sharing their opinions on the 'ZooKeys' project throughout the years. Without your kind assistance the journal would never have become as popular as it is now and would never merit its consideration as one of the most technologically advanced journals in biological science.

Acknowledgements

Our sincere thanks are due in first place to all our authors, reviewers, editors and readers without whose support the success of ZooKeys would be simply impossible! We are also deeply indebted to all our colleagues and friends with whom we collaborated on development of several innovative workflows and projects. It is not possible to list all of them here, but we would like to especially mention Plazi (Donat Agosti, Terry Catapano, Guido Sautter, Robert A. Morris, Gregor Hagedorn), GBIF (Nickolas King, Vishwas Chavan, David Remsen, Eamonn O'Tuyama, Samy Gaiji, Tim Robertson and others), Encyclopedia of Life (Cynthia Parr, Katja Schulz), ZooBank (Richard Pyle), Dryad Data Repository (Todd Vision, Peggy Schaefer), ViBRANT FP7 project (Vincent Smith, David Roberts and all partners), Wikispecies (Stephen Thorpe), Wikimedia Commons (Andrew Leung), Species-ID and BioWikifarm (Gregor Hagedorn and Daniel Mietchen).

References

Achterberg (van) C, Long KD (2010) Revision of the Agathidinae (Hymenoptera, Braconidae) of Vietnam, with the description of forty-two new species and three new genera. ZooKeys 54: 1–184. doi: 10.3897/zookeys.54.475

Baldwin CC, Castillo CI, Weigt LA, Victor BC (2011) Seven new species within western Atlantic *Starksia atlantica*, *S. lepicoelia*, and *S. sluiteri* (Teleostei, Labrisomidae), with comments on congruence of DNA barcodes and species. ZooKeys 79: 21–72. doi: 10.3897/zookeys.79.1045

Bantaowong U, Chanabun R, Piyoros Tongkerd P, Sutcharit C, James SW, Panha S (2011) New earthworm species of the genus *Amynthas* Kinberg, 1867 from Thailand (Clitellata, Oligochaeta, Megascolecidae). ZooKeys 90: 35–62. doi: 10.3897/zookeys.90.1121

Bílý S, Nakládal O (2011) Four new species of the genus *Philanthaxia* Deyrolle, 1864 from Southeast Asia and comments on *P. iris* Obenberger, 1938 (Coleoptera, Buprestidae, Thomassetiini). ZooKeys 116: 37–47. doi: 10.3897/zookeys.116.1403

Blagoderov V, Brake I, Georgiev T, Penev L, Roberts D, Rycroft S, Scott B, Agosti D, Catapano T, Smith VS (2010) Streamlining taxonomic publication: a working example with Scratchpads and ZooKeys. ZooKeys 50: 17–28. doi: 10.3897/zookeys.50.539

Bouchard P, Bousquet Y, Davies AE, Alonso-Zarazaga MA, Lawrence JF, Lyal CHC, Newton AF, Reid CAM, Schmitt M, Ślipiński SA, Smith ABT (2011) Family-group names in Coleoptera (Insecta). ZooKeys 88: 1–972. doi: 10.3897/zookeys.88.807

Brake I, von Tschirnhaus M (2010) *Stomosis arachnophila* sp. n., a new kleptoparasitic species of freeloader flies (Diptera, Milichiidae). ZooKeys 50: 91–96. doi: 10.3897/zookeys.50.505

Brothers DJ (2011) A new Late Cretaceous family of Hymenoptera, and phylogeny of the Plumariidae and Chrysidoidea (Aculeata). In: Shcherbakov DE, Engel MS, Sharkey MJ (Eds) Advances in the Systematics of Fossil and Modern Insects: Honouring Alexandr Rasnitsyn. ZooKeys 130: 515–542. doi: 10.3897/zookeys.130.1591

Catapano T (2010) TaxPub: An extension of the NLM/NCBI Journal Publishing DTD for taxonomic descriptions. Proceedings of the Journal Article Tag Suite Conference 2010. http://www.ncbi.nlm.nih.gov/books/NBK47081/#ref2

Chavan V, Penev L (in press) Data Paper: Mechanism to incentivise discovery of biodiversity data resources. BMC Bioinformatics.

Durán JM, Achterberg (van) C (2011) Oviposition behaviour of four ant parasitoids (Hymenoptera, Braconidae, Euphorinae, Neoneurini and Ichneumonidae, Hybrizontinae), with the description of three new European species. ZooKeys 125: 59–106. doi: 10.3897/zookeys.125.1754

Fikáček M, Prokin A, Angus RB (2011) A long-living species of the hydrophiloid beetles: *Helophorus sibiricus* from the early Miocene deposits of Kartashevo (Siberia, Russia). In: Shcherbakov DE, Engel MS, Sharkey MJ (Eds) Advances in the Systematics of Fossil and Modern Insects: Honouring Alexandr Rasnitsyn. ZooKeys 130: 239–254. doi: 10.3897/zookeys.130.1378

Freitag H, Balke M (2011) Larvae and a new species of *Ancyronyx* Erichson, 1847 (Insecta, Coleoptera, Elmidae) from Palawan, Philippines, using DNA sequences for the assignment of the developmental stages. ZooKeys 136: 47–82. doi: 10.3897/zookeys.136.1914

Heads SW, Leuzinger L (2011) On the placement of the Cretaceous orthopteran *Brauckmannia groeningae* from Brazil, with notes on the relationships of Schizodactylidae (Orthoptera, Ensifera). ZooKeys 77 : 17–30. doi: 10.3897/zookeys.77.769

Hendrich L, Balke M (2011) A simultaneous journal / wiki publication and dissemination of a new species description: *Neobidessodes darwiniensis* sp. n. from northern Australia (Coleoptera, Dytiscidae, Bidessini). ZooKeys 79: 11–20, doi: 10.3897/zookeys.79.803

Hong C, van Achterberg C, Xu Z (2011) A revision of the Chinese Stephanidae (Hymenoptera, Stephanoidea). ZooKeys 110: 1–108. doi: 10.3897/zookeys.110.918

Miller JA, Griswold CE, Yin CM (2009) The symphytognathoid spiders of the Gaoligongshan, Yunnan, China (Araneae, Araneoidea): Systematics and diversity of micro-orbweavers. ZooKeys 11: 9–195. doi: 10.3897/zookeys.11.160

Murphy RW, Berry KH, Edwards T, Leviton AE, Lathrop A, Riedle JD (2011) The dazed and confused identity of Agassiz's land tortoise, *Gopherus agassizii* (Testudines, Testudinidae) with the description of a new species, and its consequences for conservation. ZooKeys 113: 39–71. doi: 10.3897/zookeys.113.1353

Noort (van) S, Johnson NF (2009) New species of the plesiomorphic genus *Nixonia* Masner (Hymenoptera, Platygastroidea, Platygastridae, Scelioninae) from South Africa. In:

Johnson N (Ed) Advances in the systematics of Hymenoptera. Festschrift in honour of Lubomir Masner. ZooKeys 20: 31–51. doi: 10.3897/zookeys.20.112

Penev L, Erwin T, Thompson FC, Sues H-D, Engel MS, Agosti D, Pyle R, Ivie M, Assmann T, Henry T, Miller J, Ananjeva NB, Casale A, Lourenzo W, Golovatch S, Fagerholm H-P, Taiti S, Alonso-Zarazaga M (2008) ZooKeys, unlocking Earth's incredible biodiversity and building a sustainable bridge into the public domain: From "print-based" to "web-based" taxonomy, systematics, and natural history. ZooKeys Editorial Opening Paper. ZooKeys 1: 1–7. doi: 10.3897/zookeys.1.11

Penev L, Erwin T, Miller J, Chavan V, Moritz T, Griswold C (2009a) Publication and dissemination of datasets in taxonomy: ZooKeys working example. ZooKeys 11: 1–8. doi: 10.3897/zookeys.11.210

Penev L, Sharkey M, Erwin T, van Noort S, Buffington M, Seltmann K, Johnson N, Taylor M, Thompson FC, Dallwitz MJ (2009b) Data publication and dissemination of interactive keys under the open access model: ZooKeys working example. ZooKeys 21: 1–17. doi: 10.3897/zookeys.21.274

Penev L, Roberts D, Smith VS, Erwin T (2010a) Taxonomy shifts up a gear: New publishing tools to accelerate biodiversity research. ZooKeys 50: i–iv. doi: 10.3897/zookeys.50.543

Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith VS, Brake I, Ryrcroft S, Scott B, Johnson NF, Morris RA, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress WJ, Thompson FC, Erwin T (2010b) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1–16. http://dx.doi.org/10.3897/zookeys.50.538

Penev L, Hagedorn G, Mietchen D, Georgiev T, Stoev P, Sautter G, Agosti D, Plank A, Balke M, Hendrich L, Erwin T (2011a) Interlinking journal and wiki publications through joint citation: Working examples from ZooKeys and Plazi on Species-ID. ZooKeys 90: 1–12. doi: 10.3897/zookeys.90.1369

Penev L, Mietchen D, Chavan V, Hagedorn G, Remsen D, Smith V, Shotton D (2011b). Pensoft Data Publishing Policies and Guidelines for Biodiversity Data. Pensoft Publishers, www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf

Peng Y, Makarkin VN, Wang X, Ren D (2011) A new fossil silky lacewing genus (Neuroptera, Psychopsidae) from the Early Cretaceous Yixian Formation of China. In: Shcherbakov DE, Engel MS, Sharkey MJ (Eds) Advances in the Systematics of Fossil and Modern Insects: Honouring Alexandr Rasnitsyn. ZooKeys 130: 217–228. doi: 10.3897/zookeys.130.1576

Radea C (2011) A new species of hydrobiid snails (Mollusca, Gastropoda, Hydrobiidae) from central Greece. ZooKeys 138: 53–64. doi: 10.3897/zookeys.138.1927

Sereno PC, Larsson HCE (2009) Cretaceous Crocodyliforms from the Sahara. ZooKeys 28: 1–143. doi: 10.3897/zookeys.28.325

Sharkey MJ, Yu DS, van Noort S, Seltmann K, Penev L (2009) Revision of the Oriental genera of Agathidinae (Hymenoptera, Braconidae) with an emphasis on Thailand including interactive keys to genera published in three different formats. ZooKeys 21: 19–54. doi: 10.3897/zookeys.21.271

Stoev P, Akkari N, Zapparoli M, Porco D, Enghoff H, Edgecombe GD, Georgiev T, Penev L (2010) The centipede genus *Eupolybothrus* Verhoeff, 1907 (Chilopoda: Lithobiomorpha:

Lithobiidae) in North Africa, a cybertaxonomic revision, with a key to all species in the genus and the first use of DNA barcoding for the group. ZooKeys 50: 29–77. doi: 10.3897/zookeys.50.504

Stoev P, Enghoff H (2011) A review of the millipede genus *Sinocallipus* Zhang, 1993 (Diplopoda, Callipodida, Sinocallipodidae), with notes on gonopods monotony vs. peripheral diversity in millipedes. ZooKeys 90: 13–34. doi: 10.3897/zookeys.90.1291

Taekul C, Johnson NF, Masner L, Polaszek A, Rajmohana K. (2010) World species of the genus *Platyscelio* Kieffer (Hymenoptera, Platygastridae). ZooKeys 50: 97–126. doi: 10.3897/zookeys.50.485

Wang X-L, Yao Y-J (2011) Host insect species of *Ophiocordyceps sinensis*: a review. ZooKeys 127: 43–59. doi: 10.3897/zookeys.127.802

Wichard W, Ross E & Ross AJ (2011) *Palerasnitsynus* gen. n. (Trichoptera: Psychomyiidae) from Burmese amber. In: Shcherbakov DE, Engel MS, Sharkey MJ (Eds) Advances in the Systematics of Fossil and Modern Insects: Honouring Alexandr Rasnitsyn. ZooKeys 130: 323–330. doi: 10.3897/zookeys.130.1449

Wizen G, Gasith A (2011) Predation of amphibians by carabid beetles of the genus *Epomis* found in the central coastal plain of Israel. In: Kotze DJ, Assmann T, Noordijk J, Turin H, Vermeulen R (Eds) Carabid Beetles as Bioindicators: Biogeographical, Ecological and Environmental Studies. ZooKeys 100: 181–191. doi: 10.3897/zookeys.100.1526

Zhang Zhi-Qiang (2011) Accelerating biodiversity descriptions and transforming taxonomic publishing: the first decade of Zootaxa. Zootaxa 2896: 1–7.

ZooKeys 150: 15–51 (2011)
doi: 10.3897/zookeys.150.1766
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Data issues in the life sciences

Anne E. Thessen, David J. Patterson

Center for Library and Informatics, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543 USA

Corresponding author: *Anne E. Thessen* (athessen@mbl.edu)

Academic editor: *Lyubomir Penev* | Received 7 July 2011 | Accepted 9 August 2011 | Published 28 November 2011

Citation: Thessen AE, Patterson DJ (2011) Data issues in the life sciences. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 15–51. doi: 10.3897/zookeys.150.1766

Abstract

We review technical and sociological issues facing the Life Sciences as they transform into more data-centric disciplines - the "Big New Biology". Three major challenges are: 1) lack of comprehensive standards; 2) lack of incentives for individual scientists to share data; 3) lack of appropriate infrastructure and support. Technological advances with standards, bandwidth, distributed computing, exemplar successes, and a strong presence in the emerging world of Linked Open Data are sufficient to conclude that technical issues will be overcome in the foreseeable future. While motivated to have a shared open infrastructure and data pool, and pressured by funding agencies in move in this direction, the sociological issues determine progress. Major sociological issues include our lack of understanding of the heterogeneous data cultures within Life Sciences, and the impediments to progress include a lack of incentives to build appropriate infrastructures into projects and institutions or to encourage scientists to make data openly available.

Keywords

life science, informatics, data issues, standards, incentives, escience

Introduction

The urgent need to understand complex, global phenomena, the data deluge arising from new technologies, and improved data management are driving an agenda to extend the Life Sciences with more data-driven discovery dimensions (National Academy of Sciences 2009). The agenda requires new attitudes, facilities and approaches to sharing and querying existing data (Hey et al. 2009; Kelling et al. 2009). This document

addresses some of the more proximate issues that some of the Life Sciences face as they progress towards this "Big New Biology".

Data-driven discovery refers to hypothesis-testing and the discovery of scientific insights through the novel management and analysis of pre-existing data. It relies on access to and reuse of data which will most likely have been generated to address other scientific problems. While still hypothesis-based, data-driven discovery contrasts with the more familiar process of scientific inquiry based on collecting new data - whether by experimentation or by making new observations. It introduces opportunities to address questions that demand a "scale" of data that cannot be acquired within a single project. It is cost-effective (Piwowar et al. 2011). Data-driven discovery is not new to biology, it is already part of exploring long term trends and is an integral part of the molecular field, but it is not the norm in most sub-disciplines. It requires a large open pool of data across the full breadth of the Life Sciences and into adjacent disciplines. The pool will probably be virtual, with tools accessing data from many repositories. Such a pool will allow biology to join the other "Big" (= data-centric) sciences such as astronomy and high-energy particle physics (Hey et al. 2009). Access to a pool will invite "New" logic, strategies and tools (a "macroscope") to discover those trends, associations, discontinuities, and exceptions that reveal aspects of the underlying biology which are unlikely to emerge from more reductionist approaches (De Rosnay 1975; Ausubel 2009; National Academy of Sciences 2009; Patterson et al. 2010; Sirovich et al. 2010). An additional benefit is that a pool, and the resources from which it is macerated, may reveal factors not intrinsic to biology which improve our acuity or introduce distortions into knowledge; that is, it can lead to a better understanding of scientific certainty (Evans and Foster 2011).

The emergence of a data-centric Big New Biology is not guaranteed. Current practices in much of the discipline are parochial, with data being generated by individuals or small teams, being called upon to develop insights that are communicated in a narrative style in scientific publications. These small sciences rarely have a formal data culture, data are rarely collected with reuse in mind, they may be discarded, although more recently some journals and some sub-disciplines retain publication-related subsets of data (White et al. 2008). Data sharing requires a stable and effective cyberinfrastructure and the enthusiastic participation of the scientific community (National Science Foundation 2003, 2006; Burton and Treloar 2009; European Science Foundation 2006; http://www.gloriad.org). Registries and repositories must grow to meet the challenges of making data discoverable and accessible. The emerging "Knowledge Organization Systems" (Morris 2010) need to effectively aggregate disparate data sets in part through evolving schemas that define categories of data across the Life Sciences and through ontologies that will intelligently model existing knowledge. Semantic web technologies are needed to achieve flexibility of reuse. Enhanced user interfaces with organizational, analytical and visualization tools will be needed to allow scientists to interact with the data and associated infrastructure. Most existing environments for data management are limited in scope, and need to be improved. The enthusiastic participation of professional biologists requires a readiness to make data available for

reuse, and to take advantage of new opportunities in their quest for understanding. The resulting new mesh of biological, computer and information sciences, as well as changes to current cultures, is envisioned as having the capacity achieve the data-centric architecture capable of building new bridges among the sub-disciplines of the Life Sciences and making biology big.

This document reviews technical and sociological issues for biologists in the light of this futuristic vision for the Life Sciences. Many elements, such as data trust and data types have technological and sociological components and in such cases we have combined them for clarity.

What is meant by data

The term "data" is not used consistently. For some it is limited to raw data, for others the term widens to include any kind of information or process that leads to insights. We prefer to limit the term to neutral, objective, raw data that are largely independent of context, analysis or observer. As data become constrained, filtered and selected, they acquire or are assigned a meaning in the context of what they apply to. This is part of the process that transforms data into information (Ackoff 1989). There is no clear point of transition.

Contextual categorization of data

The context in which biological data are acquired or generated is important to understanding how data can be appropriately reused. A context may be formed if observers select or interpret their records, because of the limitations of tools or instruments used, or because data are gathered in an unnatural setting such as an experiment or "in silico". Individuals and technologies are selective and capture a limited subset of all available data. Data are affected by choice of instrument and analytical processes. Some context can be represented through the addition of appropriate metadata to data. We categorize the following broad types of data reflecting the context of their origins.

A. Observational data relate to an object or event actually or potentially witnessed by an agent. An agent may be a person, team, project, initiative; and they may call upon tools and instruments. Scientists need to take responsibility to add metadata to the observational data, ideally identifying the agent, date, location, and contexts such as experimental conditions if relevant or the equipment used. Within the Life Sciences, metadata should include taxon names, the basis for identification and/or pointers to reference (voucher) material.

1. Descriptive data are non-experimental data collected through observations of nature. Ideally, descriptive data can be reduced to values about a specified aspect of a taxon, system, or process. Each value will be unique, having been made at one place, at one time, by one agent. Observations

may be confirmed but not replicated such that it is important to preserve these data. Preservation often does not occur as data of this type are discarded after completion of the research narrative - the publication. The OBOE project offers a formal framework for descriptive data (Madin et al. 2007a).

Descriptive data can be collected by instruments or by individuals. Data collected by individuals may not represent the world completely or accurately. Mistakes can be made, such as misidentification of taxa (MacLeod et al. 2010). Researchers may be selective about the data they seek to gather, either intentionally or unintentionally, such that data sets have limited applicability. Some individuals may discard data that are not in keeping with their expectations. Few or no raw data may be recorded, such that the information may only be available in an interpreted form. Descriptive data contribute to the "long tail" of small data sets, and often are not well suited to reuse.

2. ***Experimental data*** are obtained when a scientist changes or constrains the conditions under which the expression of a phenomenon occurs. Experiments can be conducted across a broad range of scales - from electrophysiological investigations of sub millisecond processes within cells (Bunin et al. 2005) to manipulations of oceanic ecosystems (Coale et al. 2004). The intent is to dissect the elements of the phenomenon by changing conditions to uncover causal relationships, or to identify variant and invariant elements of biological processes. The raw data that are produced are contextualized by the experimental framework, and may have limited or no value in other contexts. It is important for associated metadata to include information about source and storage of material before the experiment, experimental conditions, equipment, controls and treatments.

B. Processed data are obtained through a reworking, recombination, or analysis of raw data. There are two primary types.

1. Computed data result from a reworking of data to make them more meaningful or to normalize them. In ecology, productivity or the extent of the ecosystem are rarely measured directly. Rather they are computed using information or data from other sources to generate measurements of the amount of carbon or mass that is generated per unit area per unit time. While computed data may be held in the same regard as raw data, choices or errors in formulae or algorithms may diminish or invalidate the data created. The raw data that were used and information on how computed data were derived (provenance) are important for reproducibility. The metadata should provide this information. As computed data will grow as the virtual data pool expands, it will be helpful for sub-disciplines to develop appropriate protocols and advertize best practices.

2. Simulation data are generated by combining mathematical or computational models with raw data. Often models seek to make predictions of processes, such as the future distribution of cane toads in Australia under various

climatic projections. The proximity of predictions to subsequent observations is used to test the concepts on which the model is based and to improve the model and our associated understanding of biology. Metadata differ dramatically from other data types in that date of the run, initial conditions of the model, resolution of the model output, time step, etc. are important. Rerunning the model may require preservation of initial conditions, model software, and even the operating system (Shirky 2005). Simulation data become less useful as they age and can become a storage burden.

Sociological issues

As the study of human social behavior, sociology includes the study of the behavior and practices of scientists. If we are to promote a shift to a Big New Biology, we need to understand current data cultures to determine which elements favor a transformation, and which will hinder it.

1. Data cultures

The phrase "data culture" refers to the explicit and implicit data practices and expectations that determine the destiny of data. It relates to the social conventions of acquisition, curation, preservation, sharing, and reuse of data. If the goal is to make data digital, standardized and openly accessible in a reusable format, then current data cultures provide starting points to determine the changes that will be needed before that vision can be realized. While a comprehensive survey has yet to be undertaken, it is clear that there is no single data culture for the Life Sciences (Norris et al. 2008; Gargouri et al. 2010; Key Perspectives Ltd 2010; Feijen 2011). This is unsurprising given that Life Sciences range in scope and scale from the field biologist whose data are captured in short-lived notebooks as a prelude to a narrative explanation of observations to the molecular biologist whose data are born digital in near terabyte quantities and are widely shared through global data repositories.

2. Readying data for reuse

The preparation of data for reuse in a shared pool often involves a series of steps or stages that relate to the capture, digitization, structure, storage, curation, discoverability, access, and mobility of data. The situation with molecular data achieved by the International Nucleotide Sequence Database Collaboration comprising the DNA Data Bank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and the NCBI GenBank in the USA is exemplary (http://www.insdc.org/). Molecular data tend to be born digital, and are submitted in standard formats to centralized reposito-

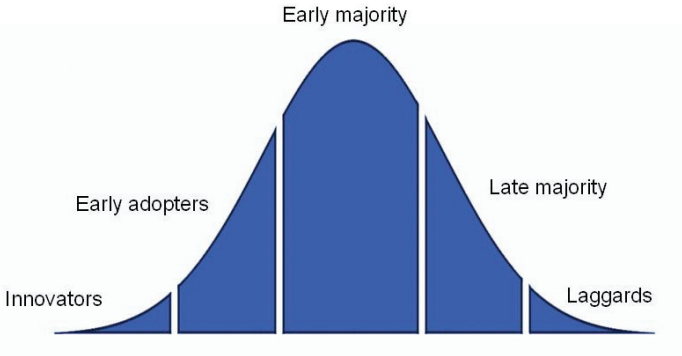


Figure 1. Rogers adoption curve describes the acceptance of a new technology. Life Sciences is still in the Early Adopters phase for accepting principles of data readiness.

ries in which they are freely available for reuse in a standard form. A rich diversity of tools, services and applications has evolved to analyze and visualize the data.

Yet, set in the context of Rogers adoption curve (Rogers 1983; Fig. 1), and as suggested by Harnad (2010), Life Sciences, generally, are closer to the early adopters stage of transition to data sharing than other sciences. It is still unusual for data created in most sub-disciplines to be made ready and openly available for sharing (Davis 2009). For these sub-disciplines to join Big New Biology, data practices must change to improve retention of data, their conversion to digital form and placement within schemes of widely agreed standards, and visibility and accessibility with few or no restrictions. The technical aspects of these practices are described in the technical issues section.

3. Agents

The term "agent" refers to individuals, groups or organizations - each influencing data cultures.

Scientists. As major producers and consumers of Life Sciences data, scientists are important participants in Big New Biology. Within the US there are almost 100,000 biologists (excluding agriculture and health sciences) working outside of academia (United States Department of Labor). The number within academia can be estimated from data on the approximately 2,500 colleges and universities (http://www.globalcomputing.com/american-universities.htm) that employ almost 300,000 academics in science and engineering, 40% of whom work in the Life Sciences (National Science Board 2010a). US research and development endeavors account for approximately one-third of the global effort (National Science Board 2010b). Consequently, changing data practices will directly or indirectly affect as many as 200,000 life scientists in the US and about half a million professionals worldwide (PARSE 2009).

As personal computers and Internet access have become integral components of biological research (Stein 2008), scientists' views and practices of data sharing have changed. Biologists are increasingly publishing data through repositories like GenBank (http://www.ncbi.nlm.nih.gov/genbank/), their own web sites, or are participating in collaborative environments such as those that allow data to be annotated (e.g. EcoliWiki, http://ecoliwiki.net/colipedia/index.php/Welcome_to_EcoliWiki or DNA Subway for genome annotation, http://dnasubway.iplantcollaborative.org/) or to capture field data using services such as provided by Artportalen (http://www.artportalen.se/default.asp) or eBird (www://ebird.org). An increasing number of databases are providing web services to mobilize data and new tools for visualizing data (e.g. GeoPhyloBuilder, https://www.nescent.org/sites/evoviz/GeoPhyloBuilder, Kidd and Liu 2008). Data processing and management pipelines such as Kepler (https://kepler-project.org/) and VisTrails (http://www.vistrails.org/index.php/Main_Page) are emerging. Yet, for these changes to dominate across the breadth of the discipline and influence the full life cycle of the data, researchers must feel comfortable with design and performance of software systems (Stein 2008). There must be good dialog between the biologists and computer programmers for new tools to be adopted (Lee et al. 2006). Increasingly, biologists will need to be trained in computer and information science (Stein 2008) and include archiving machine-readable data and appropriate metadata as part of their normal workflow (Whitlock 2011). Computer scientists, software engineers, and others who produce code need to develop sensitivity to biology and biological thinking if they are to provide tools that delight life scientists.

Scientists, especially those associated with small science, will need to be more engaged in mobilization of data than at present (Froese et al. 2003, Heidorn 2008, Costello 2009, Smith 2009). Many scientists do share specific data sets with close colleagues (Science staff editorial 2011), yet are insufficiently incentivized to share their data openly. In part, they perceive the risks of making data available as outweighing the rewards (Porter and Callahan 1994, Key Perspectives Ltd 2010). This is despite the fact that papers with openly available data gain more citations (Piwowar et al. 2007). While there are communal repositories for sub-disciplines other than molecular, such as Global Biodiversity Information Facility and Ocean Biogeographic Information System for occurrences data, the majority of sub-disciplines lack appropriate communal repositories.

Publishers. Publishers of scientific journals are increasingly involved in data management (Whitlock et al. 2010). Publishers may provide the same services for data that they provide for manuscripts (i.e. peer review, citability, etc. Vision 2010). Some journals require deposition of data as a condition of publication. An example is the joint data archiving policy (JDAP, http://datadryad.org/jdap). JDAP has grown from its original consortium of evolution and ecology journals to include more than a dozen journals (Vision 2010). Dryad (http://datadryad.org/; White et al. 2008), GenBank (http://www.ncbi.nlm.nih.gov/genbank/; Bilofsky and Christian 1988), Protein Data Bank (http://www.wwpdb.org; Berman et al. 2006) and TAIR (http://www.arabidopsis.org/; Rhee et al. 2003) are examples of repositories that benefit from deposition requirements from publishers. Publishers historically controlled the dissemination of the

narrative. Some limit access to articles while others, such as PLoS (http://www.plosbiology.org/static/help.action#xmlContent) and Pensoft (http://www.pensoft.net/journals.php) have moved to an open-access model. Although some publishers (http://www.articleofthefuture.com/, Ziegler et al. 2011) are experimenting with enhanced publication to allow researchers to share data sets, illustrations and audio files, we may presume that a publisher-driven model for data sharing is likely to incur charges for access to or submission of data. Many scientists feel this is inappropriate (Key Perspectives Ltd 2010). A model is offered by Thomson Reuters BIOSIS that indexes more than half a million Life Sciences abstracts yearly (http://thomsonreuters.com/content/science/pdf/BIOSIS_Factsheet.pdf). They are compiling metadata such as organism names and Enzyme Commission numbers that can be used to discover sources, and the publisher charges for its discovery services.

Funding agencies. Funding agencies worldwide have been called upon to finance informatics research and to promote tools and digital libraries that will underpin the shift towards a Big New Biology paradigm (Hey et al. 2009; National Academy of Sciences 2009). Funding agencies are accountable to the public and to the government (e.g. Coburn 2011). Data cost money and the reuse of data represents a better return for each research dollar invested (Piwowar et al. 2011). In recognition of the importance of data sharing to their investment, funding agencies are increasingly imposing data-sharing requirements on their researchers (Table 1). Yet, many funding agencies, especially outside the US and Europe, do not have data policies or plans to make data available. Of those that do, many require scientists to submit data management plans as a part of their proposals. The plans are designed to explain where data will be deposited, under what terms data may be accessed, and what standards will be used. Many agencies believe in open access to data at the end of a project and have specific timelines for data release. They often acknowledge that the data provider will have a period of exclusive "right of first use" of data.

Governments. The realization of a Big New Biology will require significant investment in and reorganization of technical and human infrastructure, the creation of new agencies, new policies and implementation frameworks, as well as national and transnational coordination. The scale of these developments will require governmental and intergovernmental participation. Issues that require high-level attention are illustrated by the OECD report that established GBIF (OECD 1999). GBIF has now about 60 national participants and influences national agendas. Especially relevant is the commitment to data sharing with its Suwon declaration (http://www2.gbif.org/SignedSUWONdeclaration_small.pdf). This underscores the importance of data sharing to science, conservation and sustainability. INSDC, which collates the sharing of molecular data via the US-based NCBI Genbank, the European EMBL, and the Japanese DDBJ, is another example of international informatics initiatives in the Life Sciences (http://www.insdc.org/policy.html).

Several countries have established governmental digital data environments inclusive of the data.gov environments (http://www.data.gov/, http://data.australia.gov.au/, data.gov.uk), or more specialist agencies such as Conabio in Mexico (http://www.

Table 1. List of funding agencies and characteristics of their data policies

Funding Agency	Country	Policy	Data Management Plan	Deposit	Standards Compliant	Attribution	Local Archive	Open Source	QA/QC	Confidentiality	IPR/Licensing	Metadata Deposit	Provides Data for Free	Free Access to Publications	Notes
Gordon and Betty Moore Foundation	US	http://moore.org/docs/GBMF_Data%20Sharing%20Philosophy%20and%20Plan.pdf	×			×					×				
Genome Canada	Canada	www.genomecanada.ca/medias/PDF/EN/DataReleaseandResourceSharingPolicy.pdf	×	×	×	×		×	×	×					Data must be made available no later than the publication date or the date the patent has been filed (which ever comes first) at the end of the project
National Institutes of Health	US	http://grants.nih.gov/grants/policy/data_sharing/	×							×					Applies to projects requesting > $500,000, data must be released no later than the acceptance of publication of the main findings from the final data set
Biotechnology and Biological Sciences Research Council	UK	www.bbsrc.ac.uk/publications/policy/data_sharing_policy.html	×	×	×						×	×			data release no later than publication or within 3 years of generation, Researchers are expected to ensure data availability for 10 years after completion of project
Natural Environment Research Council	UK	www.nerc.ac.uk/research/sites/data/policy.asp	×	×		×					×		×		Data must be made available within 2 years from the end of data collection
Wellcome Trust	UK	www.welcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm	×			×									
Department of Energy	US	http://genomicsgtl.energy.gov/datasharing	×	×	×	×	×	×		×		×			Requires deposit of 1) protocols 2) raw data 3) other relevant materials no later than 3 months after publication
Chinese Academy of Sciences	China	http://english.cas.cn/													Requires deposit or no further funding

Funding Agency	Country	Policy	Data Management Plan	Deposit	Standards Compliant	Attribution	Local Archive	Open Source	QA/QC	Confidentiality	IPR/Licensing	Metadata Deposit	Provides Data for Free	Free Access to Publications	Notes
Australian Research Council	Australia	http://www.arc.gov.au/default.htm													No policy
National Science Foundation	US		×												
Austrian Science Fund	Austria	http://www.fwf.ac.at/en/public_relations/oai/index.html		×										×	Data must be available no more than 2 years after end of project
NASA	US	http://science.nasa.gov/earth-science/earth-science-data/data-information-policy/						×		×					Data can be embargoed for 2 years
NOAA	US	http://www.ncdc.noaa.gov/oa/about/open-access-climate-data-policy.pdf		×									×		
Council for Scientific and Industrial Research	India	http://rdpp.csir.res.in/csir_acsir/Home.aspx													Plan being developed in 2010
North Pacific Research Board	US	http://www.nprb.org/projects/metadata.html		×									×		Data must be transferred to NPRB by the end of the project
Japan Science and Technology Agency	Japan	http://www.jst.go.jp/EN/index.html													None
National Research Foundation	South Africa	http://www.nrf.ac.za/													None

conabio.gob.mx/), ABRS, ERIN and ALA in Australia (http://www.environment.gov.au/biodiversity/abrs/, http://www.environment.gov.au/erin/, http://www.ala.org.au/), ITIS in US (http://www.itis.gov/) or the European Environment Agency (http://www.eea.europa.eu/data-and-maps).

In respect to the economics at this level, OECD, when establishing GBIF, compared the cost of the molecular informatics infrastructure (millions of dollars) against the benefits to pharmaceutical, health and agricultural businesses worth billions of dollars (OECD 1999). The costs of international cooperation on biodiversity informatics must be set against the estimated economic value of the world's natural capital of tens of trillions (millions of millions) of dollars (Costanza et al. 1997; TEEB 2010). The OECD estimates costs of sustaining infrastructure to be 25% of the costs of generating raw data. Yet, an allocation of as little as 5% of research funding could provide billions of dollars for data preservation (Schofield et al. 2010).

Universities. With in excess of 20,000 universities (and institutions modeled on Universities) worldwide (Webometrics Ranking of World Universities; http://www.webometrics.info/methodology.html), employing an estimated 5–10 million academics and associated researchers, universities form the largest research and development initiative. Collectively, Universities are a significant source of new data and given their international communal character, will be important as consumers of the data pool. The support, infrastructure and services that Universities provide will be a major determinant of the flow and fate of data. Some environments, such as the SURF foundation (http://www.surffoundation.nl/en/actueel/Pages/Researchersenhancetheirpublications.aspx) seek to unite research institutes through the application of new technologies. SURF serves the Dutch context and currently emphasizes 5 disciplines; Life Sciences are not included.

Universities may or may not regard themselves as owners (having IP rights) of data and so may regulate access to data generated in-house or as part of collaborative projects. Universities may or may not have policies that require the retention of research data for a limited period usually in the range of 3 to 7 years. The University of Melbourne policy is based on guidelines from the National Health and Medical Research Council/Australian Vice Chancellors' Committee and specifies that "Data must be recorded in a durable and appropriately referenced form" for a minimum of 5 years (http://www.unimelb.edu.au/records/research.html). The Chinese University of Hong Kong encourages researchers to deposit their data in the University Service Center upon completion of their research (http://www.usc.cuhk.edu.hk/Eng/SharingPolicy.aspx). US universities are bound to comply with the requirements of OMB Circular A-110 (Uniform Administrative Requirements for grants and agreements with Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations – http://www.whitehouse.gov/omb/circulars_a110). This specifies that financial records, supporting documents, statistics, and all other records produced in connection with a financial award, including laboratory data and primary data *are to be retained by the institution* for a specified period. OMB A-110 also states "The Federal awarding agency(ies) reserve a royalty-free, nonexclusive and irrevocable right to reproduce, publish, or otherwise use the

work for Federal purposes, and to authorize others to do so." Many universities have data policies that target administrative data and administrative agenda rather than on promoting the use of data for academic purposes (e.g. "(This) University must retain research data in sufficient detail and for an adequate period of time to enable appropriate responses to questions about accuracy, authenticity, primacy and compliance with laws and regulations governing the conduct of the research" – http://ora.ra.cwru.edu/University_Policy_On_Custody_Of_Research_Data.pdf). As their policies improve, Universities will need to play a significant role in educating staff and students as to the value of data. They will be the focus of reshaping the skill base on which the Big New Biology will rely (Doom et al. 2002). New trans-discipline curricula will ensure that biologists gain informatics skills and that computer scientists develop sensitivity to the challenges and needs in Biology.

Museums and herbaria. Museums and herbaria play special roles within the Life Sciences. Along with libraries, they have a mandate for the long-term preservation of materials. Those materials include several billion specimens of plants, animals and fossils collected by biologists over 3 centuries (Chapman 2005a; OECD 1999; Vollmar et al. 2010). Those collections provide invaluable information as to changing distributions of species, provide access to extinct species, and inform research into defining species. They have special value in some phenomena that motivate the agenda for Big New Biology, such as distribution of invasive species, consequences of deforestation, and so on. Chapman (2005a) provides an exhaustive treatment of potential and actual value of primary biodiversity records.

Citizen scientists. Citizen scientists are non-professionals who participate in scientific activities. The appealing richness of nature, its accessibility, and our reliance on natural resources ensures that biology attracts an especially high participation by the citizenry (Silvertown 2009). The academic skills of citizen scientists cover a massive spectrum, from those with casual interests in nature or science to individuals who publish in the scientific literature. The tens of millions of birders in the US (Kerlinger 1993) translates to more than 100 million worldwide. The number of recreational fishermen in marine waters approaches that of birdwatchers (Arlinghaus and Cooke 2009; Cisneros-Montemayor and Sumaila 2010), and an estimated 500 million people have livelihoods attached to fishing (ftp://ftp.fao.org/FI/brochure/climate_change/policy_brief.pdf). That suggests that the potential citizen scientist community exceeds 1 billion people. This remarkable pool can be called upon to add the "sightings" (occurrence of a given species at a particular location at a particular time) which can be used to monitor the changing distributions and abundances of endemic and invasive species. The Swedish ArtPortalen (http://www.artportalen.se/default.asp) has in 10 years compiled more than 26 million sightings at a rate of about 10,000 per day, illustrating the irreplaceable role of the citizen scientist. Several mobile phone apps exist that allow naturalists to record species occurrences in the field (BirdsEye from eBird, http://www.getbirdseye.com/ and Observer from WildObs, http://wildobs.com/about/observer).

Data on occurrences, or of the first occurrences of flowering or appearance of migratory species, can be called on to test scientific hypotheses as to the impact of climate change on the biosphere. Citizen scientists are significant monitors of endangered species – providing the first evidence that some presumed-extinct species, such as the coelocanth (http://www.extinctanimal.com/the_coelacanth.htm), Wollemi pine (http://www.wolganvalley.com/pdf/wolgan-valley/en/media-centre/fact-sheets/Wolgan%20Valley%20Wollemi%20Pine%20Fact%20Sheet.pdf?1=6), ivory-billed woodpecker (http://www.cryptomundo.com/cryptozoo-news/ibw-rainsong/), Lord Howe Island stick insect (http://www.kidcyber.com.au/topics/Lordhowestick.htm) and mountain pygmy possum (http://animaldiversity.ummz.umich.edu/site/accounts/information/Burramys_parvus.html) are still with us.

Repositories. A repository provides services for management and dissemination of data inclusive of, ideally, making data discoverable, providing access, protecting the integrity of the data, ensuring long term preservation and migrating to new technologies (Lynch 2003). Most repositories typically handle a specific data type at a particular granularity. Thousands of repositories already exist for managing Life Sciences data and hold tens of millions of items (Table 2; see Jones et al. 2006, repository66.org and http://datacite.org/repolist for more). However, it is estimated that less than 1% of ecology data is captured in this way (Reichman et al. 2011). Some sub-disciplines do not have repositories and the volume of data in some fields has led even exemplar repositories such as GenBank to question their capacity to host all data (http://www.ncbi.nlm.nih.gov/About/news/16feb2011; http://phylogenomics.blogspot.com/2011/06/sequenceshort-read-archive-sra-back.html).

Repositories range in functionality from basic data stores to collaborative databases that incorporate analysis functions (WRAM, Wireless Remote Animal Monitoring, www-wram.slu.se). Some repositories host heterogeneous data sets (such as oceanographic databases – http://woce.nodc.noaa.gov/wdiu/, http://www.nodc.noaa.gov/, http://www.ices.dk/ocean/), but those that provide normalization, standardization, atomization and quality control services (see below) will facilitate the reuse of data and will play a stronger role in data-intensive science. That many older repositories are difficult to access or are not maintained (Wren and Bateman 2008) reveals the need for appropriate funding and persistence strategies. Repositories can fail as a result of policy shifts, funding instability, management issues, or technical failures (Lynch 2003). Such failures can undermine acceptance of digital scholarly work by the community at large. As data repositories become more important over time, they must be trusted to provide high quality services reliably (Schofield et al. 2010). The trustworthiness of archives can be assessed using criteria catalogues (Klump 2011) available from organizations like the Digital Curation Center (Innocenti et al. 2007) and the International Standards Organization (ISO 2000). The Center for Research Libraries has assembled a list of ten principles for data repositories that addresses administrative and technical concerns (http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re).

Table 2. Examples of repositories for Life Sciences data.

Repository	Type of Life Sciences Data	location
AlgaeBase	algae names and references	http://www.algaebase.org/
ArrayExpress	microarray	http://www.ebi.ac.uk/arrayexpress/
Australia National Data Service	general research data	http://www.ands.org.au/
ConceptWiki	concepts	http://conceptwiki.org/index.php/Main%20Page
CSIRO	fisheries catch	http://www.marine.csiro.au/datacentre/
Data.gov	natural resources data	http://www.data.gov/
Diptera database	Dipteran information	http://www.sel.barc.usda.gov/diptera/biosys.htm
EMAGE	gene expression	http://www.emouseatlas.org/emage/
ENA	gene sequences	http://www.ebi.ac.uk/ena/
Ensembl	genomes	http://uswest.ensembl.org/index.html
Euregene	renal genome	http://www.euregene.org/
Eurexpress	transcriptome	http://www.eurexpress.org/ee/
EURODEER	movement of roe deer	http://sites.google.com/site/eurodeerproject/home
FishBase	fish information	http://www.fishbase.org/
GBIF	occurrences	http://www.gbif.org/
GenBank	gene sequences	http://www.ncbi.nlm.nih.gov/genbank/
GEO	microarray	http://www.ncbi.nlm.nih.gov/geo/
GNI	names	http://gni.globalnames.org/
INBIO	Costa Rican biodiversity	http://www.inbio.ac.cr/es/default.html
INSPIRE	spatial	http://inspire.jrc.ec.europa.eu/index.cfm
KEGG	genes	http://www.genome.jp/kegg/
Life Sciences Data Archive NASA	effects of space on humans	http://lsda.jsc.nasa.gov/
MassBank	mass spectra	http://www.massbank.jp/index.html?lang=en
MGI	mouse	http://www.informatics.jax.org/
MorphBank	images	http://www.morphbank.net/
OBIS	occurrences	http://www.iobis.org/
OMIM	human genes and phenotypes	http://www.ncbi.nlm.nih.gov/omim
PDB	molecule structure	http://www.pdb.org/pdb/home/home.do
PRIDE	proteomics	http://www.ebi.ac.uk/pride/
PubMed	citations	http://www.ncbi.nlm.nih.gov/pubmed/
Stanford Microarray Database	microarray	http://smd.stanford.edu/
tair	Arabidopsis molecular biology	http://www.arabidopsis.org/
TOPP	animal tagging	http://www.topp.org/topp_census
TreeBase	phylogenetic trees	http://www.treebase.org/
TROPICOS	plant specimens	http://www.tropicos.org/
UniProt	protein sequence and function	http://www.uniprot.org/
WILDSPACE	life history information	http://wildspace.ec.gc.ca/more-e.html
WRAM	wireless remote animal monitoring	http://www-wram.slu.se/

Technological issues

The second array of challenges that need to be addressed as we move towards Big New Biology are technical issues that affect the distribution, preservation, accessibility and reuse of data.

Making data accessible

The effective reuse of data requires that an array of conditions (Fig. 2) is optimized.

Data need to be retained. Relatively few data acquired historically have been retained in an accessible form by scientists, projects or institutions (Pullin and Salafsky 2010). The culture of disposing of data following publication, termination of a grant, relocation or retirement of a scientist is clearly incompatible with the vision of a data-centric biology. While work practices in some areas, such as those in which data are born digital, or institutions with a strong tradition of preserving records, include data retention or their submission to a repository, much of the small biology lacks such a culture (Key Perspectives Ltd 2010). There is as yet an unresolved debate as to whether all data should be retained, or if subsets of data should be selected for retention, or if retained data should be subject to periodic review for deaccessioning.

Data need to be digital. Digitization is a prerequisite for data mobility. Considerable amounts of relevant data are not yet in a digital format (Chavan and Krishnan

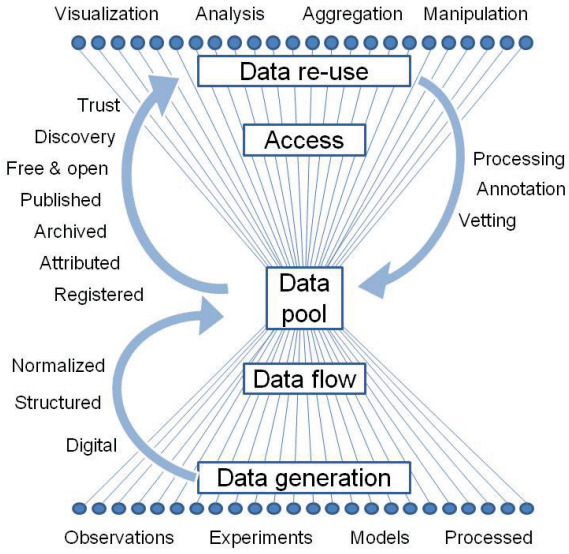


Figure 2. A Big New Biology can only emerge with a framework that optimizes reuse. Ideally, data should be in forms that can flow from source into a common pool and can flow back out to consumers, be subject to quality control, or be enhanced through analysis to rejoin the pool as processed data.

2001; Vollmar et al. 2010; Schofield et al. 2010; Heidorn 2008). Non-digital formats include notes, books, photographs and micrographs, papers, and specimens. The Biodiversity Heritage Library and similar projects are now in the process of digitizing some half billion pages of biology text (Gwinn and Rinaldo 2009). Digital metadata about non-digital materials have value as they make the data discoverable and increase incentives for digitization.

Data need to be structured. Digital data may be unstructured (e.g. in the form of free text or an image) or they may be structured into categories that are represented consecutively or periodically through the use of a template, spreadsheet or database. The simple structure of a spreadsheet allows records to be represented as rows. Data occur within the cells formed by the intersection of rows and columns defined by metadata (headers). A source may mix both structured and unstructured data such as when fields include free-form text, images, or atomic data. Unstructured data, such as the legacy data to be found in an estimated 500 million pages of text, can be improved through annotation with metadata provided by curators or through tools such as natural language processing tools.

Data should be normalized. Normalization brings information contained within different structures to the same format (or structure). Normalization may be as simple as consistently using one type of unit. Placing data within a template is a common first step to normalization. Normalization is a prerequisite for aggregating data. When data are structured and normalized, they can be mobilized in simple formats (tab delimited or comma delimited text files) or can be transformed into other structures to meet agreed upon standards. DiGIR is an early example of a data transformation tool (http://digir.sourceforge.net/). More contemporary tools, such as TAPIR or IPT from GBIF (http://ipt.gbif.org/) can output data in an array of normalized forms.

Data should be standardized. Standardization indicates compliance with a widely accepted mode of normalizing. Standards provide terms that define data and relationships among categories of data. Two basic types of standards that are indispensable for management of biological data are metadata and ontologies. Organizations such as TDWG develop new standards, and catalogs of standards and ontologies are available on the web (http://otter.oerc.ox.ac.uk/biosharing/?q=standards, http://wg.sti2.org/semtech-onto/index.php/The_Ontology_Yellow_Pages).

Metadata are terms that define data in ways that may serve different purposes, such as helping people to find data of relevance (that is they aid the discovery of data - Michener 2006), or allow data to be drawn together (federated). Metadata standards define how data should be named and structured, thus reducing the heterogeneity of terms. Standards may mandate the types of metadata that are appropriate for different types of data. Sets of metadata terms agreed upon by a community are referred to as controlled vocabularies, one of the most extensive bearing on the Life Sciences is the Ecological Metadata Language (EML; Fergraus et al. 2005). Scientific names are argued by some as having the potential to act as an extensive system of metadata (Patterson et al. 2010; See discussion below).

By articulating what metadata should be applied and how they should be formatted, standards introduce the consistency that is needed for interoperability and machine reasoning. For example, a marine bacterial RNA sequence collected from the environment ideally might be accompanied by metadata on location (latitude, longitude, depth), environmental parameters, collection metadata (collection event, date of collection, sampling device), and an identifier for the bacterium. Without such metadata, the scope of possible queries is much reduced. Examples of minimum reporting requirements have been established by the MIBBI project (Taylor et al. 2008). Numerous metadata guides are available within Life Sciences (Table 3). There are software programs available to assist in the collection and organization of metadata (such as Morpho, http://knb.ecoinformatics.org/morphoportal.jsp Higgins et al. 2002; Metacat, http://knb.ecoinformatics.org/software/metacat/, Jones et al. 2002; MERMAid, http://www.ncddc.noaa.gov/metadataresource/metadata-tools).

An ontology is a formal statement of relationships among concepts represented by metadata terms. Ontologies enable discovery of and reasoning on data through those relationships. Ontologies may use formal descriptive languages to define the relationships. Ontologies are regarded as having great promise (Madin et al. 2007b): "An ontology makes explicit knowledge that is usually diffusely embedded in notebooks, textbooks and journals or just held in academic memories, and therefore represents a formalization of the current state of a field. If ontologies are properly curated over the longer term, they will come to be seen as modern day (albeit terse) textbooks providing online and up-to-date biological expertise for their area. In another sense, they will provide the common standards needed for producing a strong biological framework for integrating data sets. Ontologies therefore provide the formal basis for an integrative approach to biology that complements the traditional deductive methodology" (Bard and Rhee 2004).

Ontologies are part of "Knowledge Organization Systems". Those relating to biodiversity have been discussed by Morris (Morris 2010). Ontologies contribute to the semantic annotation of data and the artificial intelligence it enables. As an example, a simple search for information about the bird - robin, seeks to match some or all of character string r-o-b-i-n or to character strings in text within a data object or annotating the data object. The system cannot discriminate among data on American robins, European robins, Robin Reliant cars, Robin Wright Penn, or Robin the boy-superhero. However, if the query for "robin" is placed in the context of an ontology, such as one that declares that a context is the Turdidae, an informed system is able to return only relevant results from appropriately annotated data. In addition to more precise searching, ontological structures allow the computer to perform inference, a form of artificial intelligence. For example, an ontology that establishes that turdidae is_a bird and wing is part_of a bird, allows the inference that an American robin has wings and that data on wings, flight, or migrations may be discoverable. Larger interconnected ontologies allow more complex inferences.

Many ontological structures are available for use in Life Sciences (Table 3). Some, such as the observational (http://marinemetadata.org/references/oboeontology, http://www.nceas.ucsb.edu/ecoinfo, https://sonet.ecoinformatics.org/) and

Table 3. Examples of standards and their location.

Standard	Location	Type
ABCD	http://www.bgbm.org/TDWG/CODATA/Schema/default.htm	Schema
Bioontology	http://www.bioontology.org/	Ontology Repository
BIRN	http://www.birncommunity.org/	
Cardiac Electrophysiology Ontology	http://bioportal.bioontology.org/ontologies/39038	Ontology
CMECS	Coastal and marine ecological classification standard http://www.csc.noaa.gov/benthic/cmecs/cmecs_doc.pdf	Vocabulary
Comparative Data Analysis ontology	http://sourceforge.net/apps/mediawiki/cdao/index.php?title=Main_Page	Ontology
Darwin Core	http://wiki.tdwg.org/twiki/bin/view/DarwinCore/	Metadata
Dublin Core	http://dublincore.org/	Metadata
Ecological Metdata Language	http://knb.ecoinformatics.org/software/eml/	Metadata
Environment Ontology	http://www.environmentontology.org/	Ontology
Evolution Ontology	http://code.google.com/p/evolution-ontology/	Ontology
Experimental Factor Ontology	http://www.ebi.ac.uk/efo/	Ontology
Federal Geospatial Data Committee	http://www.fgdc.gov/	Metadata
Fungal Anatomy	http://www.yeastgenome.org/fungi/fungal_anatomy_ontology/	Ontology
Gene Ontology	http://www.geneontology.org/	Ontology
Homology Ontology	http://bioportal.bioontology.org/ontologies/42117	Ontology
HUPO	http://www.psidev.info/index.php?q=node/159	Vocabulary
Infectious Disease ontology	http://www.infectiousdiseaseontology.org/Home.html	Ontology
International Standards Organization	http://www.iso.org	Metadata
Marine Metadata Interoperability	http://marinemetadata.org/	Metadata
Miriam	http://www.ebi.ac.uk/miriam/main/datatypes/	Vocabulary
National Biodiversity Information Infrastructure	http://www.nbii.gov/portal/community/Communities/NBII_Home/	Metadata
Ontology of Microbial Phenotypes	http://sourceforge.net/projects/microphenotypes/	Ontology
Open Biological and Biomedical Ontologies	http://www.obofoundry.org/	Ontology Repository
Phenotype Quality Ontology	http://obofoundry.org/wiki/index.php/PATO:Main_Page	Ontology
Plant Ontology	http://www.plantontology.org/	Ontology

Standard	Location	Type
SDD	http://wiki.tdwg.org/twiki/bin/view/SDD/Version1dot1	Schema
Species Profile Model	http://wiki.tdwg.org/SPM	Schema
Taxonomic Concept Schema	http://www.tdwg.org/activities/tnc/tcs-schema-repository/	Schema
TDWG	http://www.bgbm.org/TDWG/acc/Referenc.htm	Metadata
Teleost Anatomy Ontology	https://www.phenoscape.org/wiki/Teleost_Anatomy_Ontology	Ontology

taxonomic ontologies (below), have broad applicability - the first within the field of ecoinformatics and the second to biodiversity informatics. Users can adopt existing structures or create their own using an ontology editor such as Protégé (http://protege.stanford.edu/) or OBOEdit (http://oboedit.org/). The search engines, Swoogle (http://swoogle.umbc.edu/) and Sindice (http://sindice.com/), search over 10,000 ontologies and can return a list of those that contain a term of interest. Services such as these help users to determine if an existing ontology will meet his/her needs. Often, a user may need to use parts of existing ontologies or merge several ontologies into a single new one. Defining relationships between terms in different ontologies can be accomplished through the use of automated alignment tools such as SAMBO and KitAMO (Lambrix and Tan 2008). The development and integration of ontologies is best carried out using formal languages (such as OWL, http://www.w3.org/TR/owl-ref/) and by individuals versed in their logical foundations. The Biodiversity Information Standards (TDWG) organization (http://www.nhm.ac.uk/hosted_sites/tdwg/first_minutes.pdf) and GBIF have been prime movers in developing organizational frameworks for biodiversity information. Unfortunately, there are competing systems of standards and not all aspects of biology have established standards. Various efforts are under way to create broad scope ontologies (http://www.loa-cnr.it/index.html, http://www.tonesproject.org/, http://www.geneontology.org/). The promise of ontologies is as yet not fully realized as "The semantic web is littered with ontologies lacking ... data" (Joel Sachs, pers. comm.).

The system of latinized binomial names (such as *Homo sapiens*) introduced for species in the mid-18th century by Linnaeus is an extensive system of potential metadata for data management in the Life Sciences. They have been used to annotate virtually every statement about any of our current catalog of 2.2 million living and extinct forms of life (Raup 1991, Chapman 2009) until quite recently. Now they are being supplemented with molecular identifiers, but at this time they are well suited to form the basis of a names-based cyberinfrastructure for Biology (Patterson et al. 2008, 2010). This approach has been used for life-wide, data organization projects such as the Encyclopedia of Life (http://www.eol.org/). Placement of names within hierarchical classifications offers ontological frameworks that enable data aggregation, drilling down through data sets, and browsing through data. The conversion of names into a formal ontology has been explored through projects such as ETHAN (http://spire.umbc.edu/ont/ethan.php). Our current understanding of biodiversity

and the system of names is maintained by a specialist group of 5,000–10,000 professional taxonomists worldwide (Hopkins and Freckleton 2002), who generally are unaware of the informatics potential of names as a near universal indexing system for biological data. The Global Names Architecture is a new global initiative that links names databases and associated services to deliver names-based services to end users (Patterson et al. 2010).

Data will need to be atomized. Atomization refers to the reduction of data to minimal semantic units and stands in contrast to complex data such as images or bodies of text. In atomized forms, data may exist as numerical values of variables (e.g. "length of tail: 5.3 cm"), binary statements (e.g. "chloroplasts: absent"), or as the association with metadata terms from agreed upon vocabularies (e.g. "part of lodicules of lower floret of pedicellate spikelet of tassel"; *Zea mays* ontology ID ZEA:0015118, http://bioportal.bioontology.org/visualize/3294). Atomized data on the same subject can be brought together if the data are classified in a standard way. Atomization is necessary for machine-based analysis of data from one or more datasets. Many older data centers capture data as files (or packages of files) and the responsibility for extraction of data atoms falls to the user. This can be time consuming suggesting that, in the future, atomization needs to occur at or near the source of raw data, becoming part of the responsibilities of the author of the data, the software in which data are logged, or data centers that can provide services to transform data sets.

Data need to be published. Projects participating in a Big New Biology will increasingly make data visible and accessible (i.e. published). Scientists may publish data by displaying them in unstructured or structured formats on local, project, or institutional web sites; or they may seek to place data in central repositories. In science generally, over three-quarters of the published data are in local repositories (Science staff editorial 2011) which can provide few guarantees of persistence (see "Data are Archived" below). In such environments, the responsibilities for discovery of data, negotiations with copyright holders and acquisition of data lie with the consumer. This is time consuming and unlikely to be done on a large scale. Publication is better served through the use of central, domain-specific repositories because they are more likely to persist, provide better services, and offer the framework around which third-parties develop value-adding services. The molecular data environment consortium of ISNDC is a good example of this model. Only a small fraction of data are deposited in such environments (less than 10% of the science community generally - Science staff editorial 2011), with costs and absence of an organizational framework (metadata and archiving environments) being cited as reasons.

Publication of atomized data is essential for large scale data reuse. Data must be able to move from one computer to another in an intelligent way. As illustrated by the Global Biodiversity Information Facility (http://www.gbif.org/informatics/standards-and-tools/using-data/web-services/), scientific initiatives can add RSS feeds, web services, and APIs (Application Programming Interfaces) to their web sites to broadcast new data or to respond to requests for data. An API facilitates interaction between

computers in the same way that a user interface facilitates interactions between humans and computers. Without such services, data may need to be screen scraped from the web site, a process that is usually costly (because the solution for each site will differ) and, at worst, may require manual re-entry of data. A service-oriented approach is scalable but incurs overhead. They are probably best served through community repositories that can call on appropriate domain-specific knowledge.

Data must be archived. It is preferable that data, once published, are persistent (Feijen 2011). Projects, initiatives and host institutions have little incentive to preserve data for the long term as the process incurs a cost, and repositories that emerge within projects may have limited life spans (e.g. OBIS, http://www.iobis.org/). However, data archiving can be viewed as a good investment by funding agencies (Piwowar et al. 2011). Central repositories that are not dependent on short-term funding are better positioned to archive data making them persistent. The three global molecular databases that make up the International Nucleotide Sequence Database Collaboration provide an excellent example of how domain-specific repositories may operate. Because they are not funded through short-term projects, and because they mirror each other, such repositories guarantee the persistence of data, and empower scientists to develop projects that involve substantial analyses of shared data (Tittensor et al. 2010). Persistence can be assisted by institutions such as libraries and museums that specialize in the preservation of artifacts or by governmental intervention (the US-based National Institutes of Health support GenBank). An alternative solution to persistence is an effective business model that allows a data center to be sustained by income from services that it sells; or by providing essential services that ensure support from the community of users. Examples of commercial models include the Chemical Abstracts Service of the American Chemical Society (www.cas.org/) or Thomson Reuters' Zoological Record (http://thomsonreuters.com/products_services/science/science_products/a-z/zoological_record/).

Data will ideally be free and open. Open Access, the principle of providing unconstrained access to information on the web, improves the uptake, usage, application and impact of research output (Harnad 2008). Open Access has been applied widely to the process of publication, where it is seen as an alternative to the model in which publishers act as gatekeepers. Open Access has been applied less to data, and while this extension is natural, it is not straightforward (Vision 2010). Attitudes about sharing data freely within Life Sciences vary broadly. In sub-disciplines like genomics, data sharing is the norm with some researchers sharing their data immediately via blogs or wikis (http://www.carlboettiger.info/research/lab-notebook and http://pathogenomics.bham.ac.uk/blog/). Communities that value data sharing may have no formal recognition for such activities nor supportive technical infrastructure. Other communities have a strong sense of data ownership and are antagonistic to open data sharing. Researchers in these communities expect to be directly involved in any further analyses of their data. Databanks for these communities often require registration and/or a fee to gain access. Some data may be regarded as too sensitive to be made fully accessible (Key Perspectives Ltd 2010).

Data can be trusted. Once data are accessed, consumers may reveal errors and/or omissions. Biological data can be very dirty, especially if they were acquired without expectation that they would be shared later. Any data cleaning procedures should be documented to aid the consumer in assessing whether the source is "suitable for their purpose" (Chapman 2005b). The creation of "quality loops" allow comments to flow back to the source where data can be annotated or modified, and returned to users for renewed vetting. Webhooks (http://iphylo.blogspot.com/2011/02/web-hooks-and-openurl-making-databases.html) offer a mechanism to exploit APIs to have comments returned to source. Any editing of data can lead to the undesirable outcome that variant forms of the same data may coexist. To some extent, versioning of data sets can be used to discriminate between modified datasets, but users need to cite the version used in analyses (Zhang et al. 2007).

Data must be attributed. Scientists gain credit in part through attribution. The permanent association of identifiers with open data offers a means of linking attribution to the data and of tracking reuse (Cryer et al. 2009). The association of authors' names with data motivates contributions (or lack of credit demotivates them). Attribution favors the development of quality loops to correct errors or otherwise comment on the data. Special care is needed when attributing data resulting from the combination of one or more existing sets so that all intellectual investment is properly credited. Dryad, a JDAP partner, provides data citations through the use of DataCite DOIs with an unrestrictive Creative Commons Zero license, thus promoting clear citation and reuse of data (Vision 2010). Community norms can ensure proper attribution of CC0-licensed data (Fauchart and von Hippel 2008). The Panton Principles provide guidelines for licensing data (http://pantonprinciples.org/).

Data can be manipulated. A value of having large amounts of appropriately annotated data available on the web is that users can explore, in addition to search for, data. Data exploration may result from a desire to test a hypothesis. It is therefore desirable to have tools that draw data together, analyze or visualize them. Exploratory systems include: Humboldt (Kobilarov and Dickinson 2008) which operates like a faceted filter for Linked Data; Parallax which accesses data in Freebase and has the ability to interact with data on multiple web pages at once (Huynh and Karger 2009); and Microsoft Pivot (http://www.getpivot.com/) allows a user to interact with large amounts of data from multiple Internet sources.

Visualizations have the capacity to reveal patterns, discontinuities and exceptions that can inform us as to underlying biological processes, appropriateness of data sets, or consistency of experimental protocols. Visualizations can be used to display results with analyses of large data sets. Through visualizations we may help address the challenge stated by Fox and Hendler (2011) that "... many of the major scientific problems facing our world are becoming critically linked to the interdependence and interrelatedness of data from multiple instruments, fields and sources". The absence of effective visualization is creating a bottleneck within data-intensive sciences (Fox and Hendler 2011). Solutions need to be found in relatively simple low end visualizations (as wonderfully catalogued in http://www.visual-literacy.org/periodic_table/periodic_table.

html) to high end tools designed for the data deluge that themselves may call on graphics and visualization standards to be pipelined into rich, complex, and flexible aids. Many Life Sciences data sets can be drawn together and visualized using the geospatial element such as with LifeMapper (http://www.lifemapper.org/) or by OBIS and GBIF (inter alia; Webb et al. 2010). Geospatial metadata, along with temporal, publication, and names metadata are especially valuable as integrators of diverse data sets.

Data need to be registered and discoverable. Registries index data resources to alert potential users to their availability. Search engines, the normal indexers of web-accessible materials, are not good at revealing database contents - only about half of the open data in repositories are indexed by search engines (McCown et al. 2006). Discovery is made possible by the addition of coarse grained discovery metadata. Registry functions need to expose discovery metadata to make data sets more visible. As an example, GBIF provides registry level service for biodiversity data (http://www.gbif.org/informatics/standards-and-tools/integrating-data/resource-discovery/). Registries that cover software (http://en.bio-soft.net/geshi.html, http://www.equisetites.de/palbot/software/software.html) or web services (www.biocatalogue.org) are valuable in promoting awareness of tools for data capture, conversion and processing. Successful domain repositories, such as GenBank, have well-structured and detailed metadata that enable detailed search and enhanced discoverability. In the absence of such registries, researchers turn to peers, publications or the thousands of minor data sets available via the Internet. Under these circumstances, it is hard to know when, or if, all relevant data are found. There is a need for a broad-spectrum registry and indexing service (like a Google for data) where researchers can post pointers to their own data, search for desired data and have a means to quickly preview the results. Examples of this exist in Europe with OpenDOAR (http://www.opendoar.org/) and in India with Database of Biological Database (http://www.biodbs.info/), each with thousands of listings. Semantic annotation of data greatly increases discoverability, and is discussed below.

The semantic web and Big New Biology

The "semantic web" has many definitions, but here we think of it as a technical framework that promotes automated sharing and reuse of data across disciplines (Campbell and MacNeill 2010). The semantic approach has advantages of being flexible, evolvable, and additive. A semantic infrastructure will lead to machine-mediated answers to more complex queries than previously possible (Stein 2008). The foundations for automated reasoning lie in the annotation of data with agreed metadata, linked through a network of ontologies, and queried using conventions (languages) such as RDF, OWL, SKOS and SPARQL (Campbell and MacNeill 2010). The mass of appropriately annotated data that can be accessed through the Internet is referred to as LOD (Linked Open Data). Through common metadata, the data can be linked to form a Linked Open Data cloud. At this time, Life Sciences makes up 9% of the triples in LOD and 51% of the links (Bizer et al. 2011).

Berners-Lee has promoted four guidelines for linked data (Berners-Lee et al. 2006):

1. The use of a standard system of Uniform Resource Identifiers (URIs) as "names" for things
2. The use of HTTP URIs so that the names can be looked up on the internet and the data accessed
3. When a URI is looked up, it should return useful information using standards (RDF, SPARQL)
4. Links to other URIs so that users can discover more things.

A URI is a type of persistent identifier made up of a string of characters that unambiguously (at least in an ideal world, see Booth 2010 for discussion) represents data or metadata and can be used by machines to access the data. Different data sets can be linked when they refer to the same URIs. For example, several marine data sets could be linked because they identify the same investigator, sampling event, or location. The most useful classes of terms that are likely to serve the needs of the Life Sciences are georeferences (which can link data from the same location held in different repositories), names of taxa (the common denominator to the majority of statements about biodiversity), publications and identities of people that can be interconnected through devices such as FOAF (friend-of-a-friend) to find collaborators, relevant data, as well as to dig into the world of scientific literature, the latter being linkable through devices such as DOIs to show citation trends, influential publications, etc. (Patterson et al. 2010).

RDF is a language that defines relationships between things. Relationships in RDF are usually made in three parts (often called triples), Entity:Attribute:Value. A machine-readable form in RDF may be a statement that "American robin:has_color:red". Each term is ideally defined stringently by controlled vocabularies and ontologies, and each part represented within the triple as a URI. The "Value" can be a URI or a literal - the actual value. An advantage of RDF is that it allows datasets to be merged, for example TaxonConcept and Wikipedia (http://www.slideshare.net/pjdwi/biodiversity-informatics-on-the-semantic-web). A goal of the Linking Open Data project is to promote a data commons by registering sets in RDF. As of March 2011, the project had grown to 28 billion triples and 395 million RDF links (Bizer et al. 2011). The EU project, Linking Open Data 2, received €6.5 million to expand Linked Data by building tools and developing standards (http://lod2.eu/Welcome.html).

Transformation of data from printed narrative or spreadsheet to semantic-web formats is a significant challenge. Based on existing ontologies, there is enough information to create 10^{14} triples in biomedicine alone (Mons and Velterop 2009). At the time of writing, this quantity far exceeds the capacity of any system to process the information.

Life Sciences stand to benefit greatly from the advantages of linked data (Reichman et al. 2011), but need additional investment in mechanisms that ensure quality, provenance and attribution. Provenance identifies sources and, among other things, can ensure attribution and be part of quality control processes. Several software packages currently exist for tracking provenance (such as Kepler, https://kepler-project.org/; Taverna, http://www.taverna.org.uk/; VisTrails, http://www.vistrails.org/index.

php/Main_Page). Bechhofer et al. (2010) advocate the use of Research Objects (ROs) as a mechanism to capture additional value necessary to make the semantic web work for science. Provenance of ROs would satisfy recent calls for "open science" that argue that not only data should be open, but so should be associated methods and analyses (Reichman et al. 2011).

Semanticization enables nanopublication, a form of publication that extends traditional narrative publication (Groth et al. 2010) and allows attribution to be associated with the semantic web (Mons and Veltrop 2009). Nanopublications relate to publication of triples. A uniquely identifiable triple is a statement. A triple with a statement for a subject is called an annotation and a set of annotations that refer to the same statement is called a nanopublication. The annotations add attribution and context to the statement. The concept is not widely accepted.

Discussion

A Big New Biology holds much promise as a means to address some large proximate scientific challenges. Macroscopic tools will enable discovery of hidden features and better descriptions of relationships within the complexity of the biosphere. Yet, to date, progress towards the vision varies enormously from the successes with high-throughput biology to virtual stasis in some small science biology. Considerable effort is needed to catalog current practices, and to define the sociological transformations that will be required to improve the likelihood of success. If the transformation is to be purposeful, then it will need general oversight, discipline-specific reviews, and a description of the actual and desirable components of the Knowledge Organizational System for Biology and their relationships. Some obvious challenges relate to standards and associated ontologies, incentivizing participation, and assembling an appropriate infrastructure and skill base.

Standards and Ontologies. Data standards bring order to the virtual data pool on which a Big New Biology will rely. While complex and finely grained metadata are needed for analyses and for the world of Linked Open Data, the first challenge is to improve the discoverability of data. This process has traditionally been supported by word-of-mouth at conferences or in publications. With standards, registries can enable users to find data sets containing information about taxa, parameters, times, processes, or places of interest. If metadata are absent or incomplete, then the data sets cannot be discovered or reused and cannot contribute to Big New Biology.

Automated data discovery, aggregation and analysis require more comprehensive standards than those currently available for many of the Life Sciences. Instead of a comprehensive system of standards, there is a piecemeal system of metadata, vocabularies, thesauri, ontologies, and data transfer schemas that overlap, compete, and have gaps. Greatest progress is being made outside the Life Sciences (such as georeferencing), or in high-investment areas where data are born digital (such as in genomics, Taylor et al. 2008). Given the richness of biodiversity and interactions, a comprehensive

system of standards will necessarily be extremely complex, and be costly to implement. This creates a tension: whether to promote the comprehensive annotation of data with a significant overhead that deters participation versus pursuing a more minimalistic annotation that can set a grander process in motion. As the commitment to standards is not widespread, the minimalistic approach is more likely to gain traction. The perspective that "The semantic web is littered with ontologies lacking ... data" noted above warns us against starting with complex structures. Metadata and their interrelationships will need a framework that is designed to allow initial discipline-specific standards to become more finely grained and for the parts to merge into a dynamic grand schema. The world of Linked Open Data provides a good model for this, but given that few data are appropriately annotated, it has yet to realize its potential.

Two organizational frameworks for Life Sciences data are as yet under-exploited. The first is the system of georeferencing that is in use in rich applications in earth sciences, cartography, and so on. Information on occurrences of species is compiled in central databases such as GBIF and OBIS, has been and is being collected in vast quantities by a myriad of citizen scientists. Its potential is well illustrated by some large-scale applications such as the impressive charting of bird migrations (Marris 2010), meta-analyses of oceanic biota (Webb et al. 2010), or web sites that emphasize locally relevant biota (http://zipcodezoo.com/). Less well developed, but arguably with more potential for many sub-disciplines of the Life Sciences, is the transformation of taxonomic and phylogenetic knowledge into an information management system that uses Latin names and molecular identifiers as metadata and classifications and phylogenies as ontological frameworks for the metadata (Patterson et al. 2010).

Incentives. Despite widespread calls for scientists to make data more widely available, this has yet to happen for many sub-disciplines (Dittert et al. 2001, Harnad 2008, Mandavilli 2011, Piwowar 2011). Only about 10% of data make their way to open repositories (Savage and Vickers 2009, Science staff editorial 2011). A current impediment to data sharing is that the benefits derived are often greater for the consumer than the producer (Porter and Callahan 1994). Other reasons are the lack of resources, infrastructure, and incentives for sharing. Sociological, financial, legal and technical barriers must be surpassed for communities to become directly involved in populating and maintaining data pools, a requisite for success and scalability (Feijen 2011).

In surveys, (Froese et al. 2003, Kohnke et al. 2005, RIN 2008, Costello 2009), scientists give the following five reasons not to share data. The first relates to intellectual property: A scientist's funding and professional recognition relies on receipt of credit for work done. Until scientists receive credit for data publication, there will be little motivation to redirect efforts from more rewarding activities (such as exploring nature or writing papers) towards data mobilization. This problem can be solved with an infrastructure capable of creating citations for data and tracking data use (Froese et al. 2003). The second relates to legal and confidentiality issues as some data cannot be shared, such as data concerning people (Guttmacher et al. 2009) or location of endangered species (Froese et al. 2003), proprietary information, or because employers or funders claim that they have copyright over data. The infrastructure must have

mechanisms to protect necessary confidentiality. Some data can be anonymised, and in the case of endangered taxa, protection can be accomplished by fuzzing data, so that exact locations or identities are obscured (Froese et al. 2003). Thirdly, there is concern over misuse or misinterpretation of data, which, once in the literature, cannot be unpublished. This is not a new problem, but it will increase as data producers lose control and can no longer act as "gate-keepers". Part of the solution lies in developing stringent metadata and format standards such that data are released only when there are sufficient metadata to ensure that all users understand the context and limitations of the data. Until such time, disclaimers can alert consumers about inappropriate reuse (Froese et al. 2003, Smithsonian 2011). Fourthly, scientists are concerned that publication can expose errors in their data or weaknesses of analysis. Errors may include insufficient, inaccurate or inappropriate data encoding, metadata, or analysis. Third parties may reveal the selective or inappropriate use of data to emphasize particular arguments. Given the noisy and rich nature of biology, there can be no such thing as a perfect data set; all are incomplete. Errors or gaps uncovered by subsequent users can be dealt with openly and honestly, thereby enhancing the body of scientific data. Finally, there is the issue of sustainability. Project-based data repositories run a risk of being abandoned at the end of the funding cycle. This increases doubts that data curation activities are a good use of resources. It is cheaper to curate data properly than it is to gather it again (Heidorn 2008, Piwowar et al. 2011), and some data, such as data on past distributions of species, are irreplaceable and thus priceless. From an economic perspective, persistent discipline-specific repositories are attractive. There are considerable academic benefits from engaging with repositories. Scientists who share data often report increased book and/or photograph sales, increased web site hits and higher visibility for their projects (Froese et al. 2003). There is greater citation impact for open-access articles (Gargouri et al. 2010). In larger consortia, scientists (such as those studying phylogenetic relationships) who pool data are able to answer questions they could not answer if they were limited to the data that they themselves generated. Some publishers are incentivizing early data-sharing by granting an embargo to the data producers (Kaye et al. 2009) to alleviate fears of being "scooped" (Reichman et al. 2011). An emphasis on "carrots" such as these may be much more effective means of promoting data-sharing than the "sticks" (in the form of funding agency requirements, Kaye et al. 2009; Table 1).

Infrastructure. In addition to challenges to incentivize scientists in the direction of data-sharing, the infrastructure for a Big New Biology is incomplete. Funding agencies, like the National Science Foundation in the US, require projects to have plans for data management - a requirement that presumes data persistence. The infrastructure needed to guarantee persistence will require an investment well beyond the usual 3–5 year funding cycle into multi-decadal periods and coordination that has international dimensions. The infrastructure must include tools to capture data, policies, data standards, data identifiers, registration of discovery-level metadata, and APIs to share data (Fig. 3). There is as yet no index of data-sharing services (for some initial steps see datacatalogs.org and DataCite http://www.datacite.org/repolist) nor a framework in which such elements could be integrated. There is little assessment of which elements of data

plans will lead to persistence of data or their reuse. In the absence of these elements, principle investigators are left to make their own policies, use their own systems, and to finance the processes. As long as the response is piecemeal, there can be no assurances of interoperability, efficiency or persistence. At this time, research scientists need to be supported by data managers and data archivists. Institutional libraries and museums are well placed to shift their agendas to include data management and the preservation of digital artifacts and so may fill this gap, providing institutional, regional or discipline-based services. It is hoped that the ongoing NSF Data Net projects can contribute significantly to the infrastructure.

A new technical challenge is the lack of bandwidth to distribute data from modern data-intense technologies. The problem is illustrated by high throughput molecular biology with tera and petabyte scale data sets (Cochrane et al. 2009). Proposed solutions include Bio-Mirror (http://www.bio-mirror.net/) which consists of several servers holding the same data, or the Tranche Project (https://trancheproject.org/), which shares repository functions across servers. The latter has a high administrative overhead. Peer-to-peer sharing systems such as BitTorrent (Langille and Eisen 2010) overcome potential bandwidth problems by sharing data sets without a central repository. Users of BioTorrents benefit from lower bandwidth use, faster transfer times and data publication. Although terabit per second line rates are on the horizon (Hillerkuss et al. 2011), bandwidth problems are likely to persist as part of the interplay between the evolution of new data-generating instruments and the limitations of the infrastructure to make data freely available to all. We may expect to see a growth of specialist centers that will offer analysis, visualization, and data transformation services on behalf of the users.

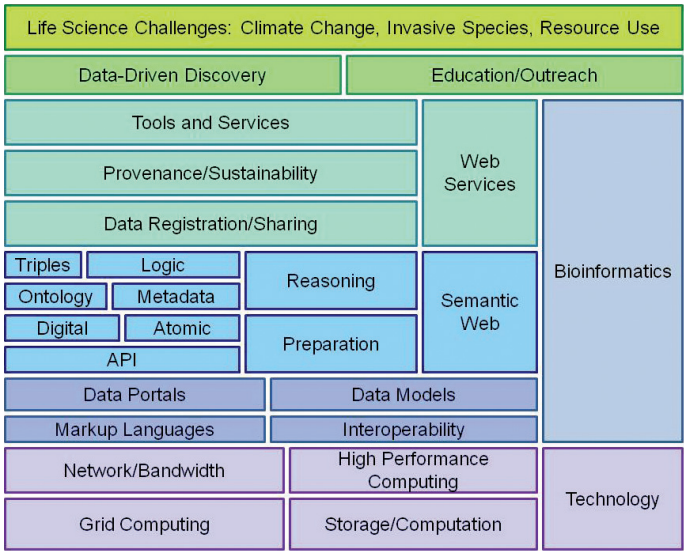


Figure 3. Technical infrastructure needed for Big New Biology to fully emerge (based on Sinha et al. 2010).

Conclusion

There is growing pressure from scientists, funding agencies and governments to use new information technologies to effectively manage the increasingly vast amounts of data emerging from new technologies, to integrate these with smaller data sets, and to enhance the communal nature of science. If successful, biology will be enriched with data-intensive dimensions better suited to address large scale and trans-discipline problems. The transition requires many technical advances and cultural changes. Progress on the technical front to date clearly demonstrates that technical issues can be resolved. The process of sociological adaptation is less convincing. Some sub-disciplines (molecular domains) have embraced data-intensive dimensions, some (environmental ecology) are in transition, and others (such as taxonomy) are just beginning. A much better understanding of the existing cultures is needed before we can promote solutions that will realign the traditions of each community with the common goal of shared data use. Training environments such as Universities need to create a new cadre of scientists trained in computer sciences and biology. Other pressing challenges to data integration relate to the development of comprehensive and agreed metadata and ontologies, and to the semanticization of data so that the discipline can take advantage of the Linked Open Data cloud. The long tail of small data sets presents a special challenge - that of bringing heterogeneous data sets together. At this time, the common denominators that are likely to be effective are georeferencing, citations, and names. All require further investment. None of the elements of the transition will come quickly or cheaply, but these transformations are needed if we are to make the Life Sciences less parochial and more capable of responding to major research challenges.

Acknowledgments

The authors would like to thank Dmitry Mozzherin, David Shorthouse, Nathan Wilson, Jane Maeinschein, Peter DeVries, Holly Miller, Vince Smith, Daniel Mietchen and members of the Data Conservancy Life Sciences Advisory Group (Mark Schildhauer, Bryan Heidorn, Steve Kelling, Dawn Field, Norman Morrison and Paula Mabee) for valuable comments. This work is supported by NSF award 0830976 The Data Conservancy (A digital research and curation virtual organization).

The topics raised here were explored during a workshop held in Woods Hole, Massachusetts attended by computer, information and biological scientists, and representatives of academia, the private sector and government. A longer "white paper" produced for the National Science Foundation Data Conservancy project is available (Thessen and Patterson 2011).

References

Ackoff R (1989) From data to wisdom. Journal of Applied Systems Analysis 16: 3–9. doi: 10.1002/9781444303179.ch3

Arlinghaus R, Cooke SJ (2009) Recreational fisheries: socioeconomic importance, conservation issues and management changes. In: Adams B (Ed) Recreational Hunting, Conservation, and Rural Livelihoods: Science and Practice. Blackwell, Oxford. doi: 10.1002/9781444303179.ch3

Ausubel JH (2009) A botanical macroscope. Proceedings of the National Academy of Science. 106: 12569. doi: 10.1073/pnas.0906757106

Bard JBL, Rhee SY (2004) Ontologies in biology: design, applications and future challenges. Nature Reviews Genetics 5: 213–222. doi: 10.1038/nrg1295

Bechhofer S, Ainsworth J, Bhagat J, Buchan I, Couch P, Cruickshank D, De Roure D, Delderfield M, Dunlop I, Gamble M, Goble C, Michaelides D, Missier P, Owen S, Newman D, Sufi S (2010) Why linked data is not enough for scientists. 6th IEEE e-Science conference.

Berman H, Henrick K, Nakamura H, Markley JL (2006) The worldwide protein data bank (wwPDB): ensuring a single uniform archive of PDB data. Nucleic Acids Research 35: D301–D303. doi: 10.1093/nar/gkl971

Berners-Lee T, Chen Y, Chilton L, Connolly D, Dhanaraj R, Hollenbach J, Lerer A, Sheets D (2006) Tabulator: exploring and analyzing linked data on the semantic web. Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI0), Athens, Georgia.

Bilofsky HS, Christian B (1988) The GenBank genetic sequence data bank. Nucleic Acids Research 16: 1861–1863. doi: 10.1093/nar/16.5.1861

Bizer C, Jentzsch A, Cyganiak R (2011) State of the LOD cloud. [http://www4.wiwiss.fu-berlin.de/lodcloud/state/]

Booth D (2010) Resource identity and semantic extensions: making sense of ambiguity. Semantic Technology Conference, San Francisco, USA http://dbooth.org/2010/ambiguity/.

Bunin VD, Ignatov OV, Gulii OI, Voloshin AG, Dykman LA, O'Neil D, Ivnitskii D (2005) Investigation of electrophysical properties of *Listeria monocytogenes* cells during the interaction with monoclonal antibodies. Biofizika 50: 316–321.

Burton A, Treloar A (2009) Designing for discovery and re-use: the ANDS data-sharing verbs approach to service decomposition. The International Journal of Digital Curation 4: 44–56.

Campbell LM, MacNeill S (2010) The semantic web, linked and open data: a briefing paper. JISC cetis. [http://wiki.cetis.ac.uk/images/1/1a/The_Semantic_Web.pdf]

Chapman AD (2005a) Uses of primary species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. [http://www.niobioinformatics.in/books/Uses%20of%20Primary%20Data.pdf]

Chapman AD (2005b) Principles of data quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. [http://niobioinformatics.in/pdf/workshop/Data%20Quality.pdf]

Chapman AD (2009) Numbers of Living Species in Australia and the World, 2nd edition. Australian Biological Resources Study, Australia.

Chavan V, Krishnan S (2001) Digitizing life: role of digital libraries in life conservation in developing world. Proceedings of the 4th International Conference on Asian Digital Libraries, December 10–12, 2001, Bangalore India, 330–340. [http://ncsi-net.ncsi.iisc.ernet.in/gsdl/collect/icco/index/assoc/HASHe590.dir/doc.doc]

Cisneros-Montemayor AM, Sumaila UR (2010) A global estimate of benefits from ecosystem based marine recreation: Potential impacts and implications for management. Journal of Bioeconomics 12: 245–268. doi: 10.1007/s10818-010-9092-7

Coale KH, Johnson KS, Chavez FP, Buesseler KO, Barber RT, Brzezinski MA, Cochlan WP, Millero FJ, Falkowski PG, Bauer JE, Wanninkhof RH, Kudela RM, Altabet MA, Hales BE, Takahashi T, Landry MR, Bidigare RR, Wang X, Chase Z, Strutton PG, Friederich GE, Gorbunov MY, Lance VP, Hilting AK, Hiscock MR, Demarest M, Hiscock WT, Sullivan KF, Tanner SJ, Gordon RM, Hunter CN, Elrod VA, Fitzwater SE, Jones JL, Tozzi S, Koblizek M, Roberts AE, Herndon J, Brewster J, Ladizinsky N, Smith G, Cooper D, Timothy D, Brown SL, Selph KE, Sheridan CC, Twining BS, Johnson ZI (2004) Southern Ocean iron enrichment experiment: carbon cycling in high- and low-Si waters. Science 304: 408–414. doi: 10.1126/science.1089778

Coburn TA (2011) The National Science Foundation: Under the microscope. A report by Tom A. Coburn, M.D. U.S. Senator, Oklahoma. [http://coburn.senate.gov/public/index.cfm?a=Files.Serve&File_id=f6cd2052-b088-44c3-b146-5baa5c01552a]

Cochrane G, Akhtar R, Bonfield J, Bower L, Demiralp F, Faruque N, Gibson R, Hoad G, Hubbard T, Hunter C, Jang M, Juhos S, Leinonen R, Leonard S, Lin Q, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Plaister S, Radhakrishan R, Robinson S, Sobhany S, Hoopen PT, Vaughan R, Zalunin V, Birney E (2009) Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Research 37: D19–D25. doi: 10.1093/nar/gkn765

Costanza R, D'arge R, de Groot R, Farber S, Grasso M, Hannon B, Limburg K, Naeem S, O'Neill RV, Paruelo J, Raskin RG, Sutton P, van den Belt M (1997) The value of the world's ecosystem services and natural capital. Nature 387: 253–260. doi: 10.1038/387253a0

Costello M (2009) Motivating online publication of data. BioScience 59: 418–426. doi: 10.3525/bio.2009.59.5.9

Cryer P, Hyam R, Miller C, Nicolson N, Ó Tuama É, Page R, Rees J, Riccardi G, Richards K, White R (2009) Adoption of persistent identifiers for biodiversity informatics: Recommendations of the GBIF LSID GUID task group, 6 November 2009. [http://www2.gbif.org/Persistent-Identifiers.pdf]

Davis PM (2009) Author-choice open access publishing in the biological and medical literature: a citation analysis. Journal of the American Society for Information Science and Technology 60: 3–8. doi: 10.1002/asi.20965

De Rosnay J (1975) Le macroscope: vers une vision globale. Seuil, Paris.

Dittert N, Diepenbroek M, Grobe H (2001) Scientific data must be made available to all. Nature 14: 393. doi: 10.1038/35106716

Doom T, Raymer M, Krane D, Garcia O (2002) A proposed undergraduate bioinformatics curriculum for computer scientists. ACM SIGCSE Technical Symposium on Computer Science Education (33rd, Covington, KY), 78–81. doi: 10.1145/563340.563368

ESF (European Science Foundation) (2006) Press Release: A cyberinfrastructure network for Europe. [http://www.esf.org/media-centre/press-releases/ext-single-news.html?tx_ttnews[tt_news]=129&cHash=98c6548070c4afa002061d23560e8f96]

Evans JA, Foster JG (2011) Metaknowledge. Science 331: 721–725. doi: 10.1126/science.1201765

Fauchart E, von Hippel E (2008) Norms-based intellectual property systems: The case of French chefs. Organization Science 19: 187–201. doi: 10.1287/orsc.1070.0314

Feijen M (2011) What researchers want. SURFfoundation. [http://www.surffoundation.nl/en/publications]

Fergraus EH, Andelman S, Jones MB, Schildhauer M (2005) Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. Bulletin of the Ecological Society of America 86: 158–168. doi: 10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2

Fox P, Hendler J (2011) Changing the equation on scientific data visualization. Science 331: 705–708. doi: 10.1126/science.1197654

Froese R, Lloris D, Opitz S (2003) Scientific data in the public domain. ACP-EU Fisheries Research Report 14: 267–271.

Gargouri Y, Hajjen C, Larivière V, Gingras Y, Carr L, Brody T, Harnad S (2010) Self-selected or mandated, open access increases citation impact for higher quality research. PLoS ONE 5: e13636. doi: 10.1371/journal.pone.0013636

Groth P, Gibson A, Velterop J (2010) The anatomy of a nanopublication. Information Services & Use 30:51–56. doi: 10.3233/ISU-2010-0613

Guttmacher AE, Nabel EG, Collins FS (2009) Why data-sharing policies matter. Proceedings of the National Academy of Science 106: 16894. doi 10.1073/pnas.0910378106

Gwinn NE, Rinaldo C (2009) The Biodiversity Heritage Library: sharing biodiversity literature with the world. IFLA Journal 35: 25–34. doi: 10.1177/0340035208102032

Harnad S (2008) Waking OA's "Slumbering Giant": The University's mandate to mandate open access. New Review of Information Networking 14: 51–68. doi: 10.1080/13614570903001322

Harnad S (2010) Open Access – Open Data: similarities and differences. [http://www.slideshare.net/oaod2010/oa-oa-self-archiving-oa-publishing-and-data-archiving]

Heidorn PB (2008) Shedding light on the dark data in the long tail of science. Library Trends 57: 280–299. doi: 10.1353/lib.0.0036

Hey T, Tansley S, Tolle K (2009) The Fourth Paradigm. Microsoft Research. Redmond, WA, 252 pp.

Higgins D, Berkley C, Jones MB (2002) Managing heterogeneous ecological data using Morpho. 14th International Conference on scientific and statistical database management (SSDBM'02), 69.

Hillerkuss D, Schmogrow R, Schellinger T, Jordan M, Winter M, Huber G, Vallaitis T, Bonk R, Kleinow P, Frey F, Roeger M, Koenig S, Ludwig A, Marculescu A, Li J, Hoh M,

Dreschmann M, Meyer J, Ben Ezra S, Narkiss N, Nebendahl B, Parmigiani F, Petropoulos P, Resan B, Oehler A, Weingarten K, Ellermeyer T, Lutz J, Moeller M, Huebner M, Becker J, Koos C, Freude W, Leuthold J (2011) 26 Tbit s^{-1} line-rate super-channel transmission utilizing all-optical fast Fourier transform processing. Nature Photonics 5: 364–371. doi: 10.1038/nphoton.2011.74

Hopkins GW, Freckleton RP (2002) Declines in the numbers of amateur and professional taxonomists: implications for conservation. Animal Conservation 5: 245–249. doi: 10.1017/S1367943002002299

Huynh DF, Karger DR (2009) Parallax and companion: set-based browsing for the data web. In: Proceedings of WWW '09.

Innocenti P, McHugh A, Ross S, Ruusalepp R (2007) Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE) audit toolkit: DRAMBORA. International Conference on Digital Preservation (iPRES), Beijing.

ISO (2000) ISO 9000:2000: quality management systems – fundamentals and vocabulary. Standard, International Organization for Standardization (ISO), Geneva, Switzerland.

Jones MB, Berkley C, Bojilova J, Schilhauer M (2002) Managing scientific metadata. Internet Computing IEEE 5: 59–68. doi: 10.1109/4236.957896

Jones MB, Schildhauer MP, Reichman OJ, Bowers S (2006) The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution and Systematics 37: 519–544. doi: 10.1146/annurev.ecolsys.37.091305.110031

Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P (2009) Data sharing in genomics – reshaping scientific practice. Nature Reviews Genetics 10: 331–335. doi: 10.1038/nrg2573

Kelling S, Hochachka WM, Fink D, Riedewald M, Caruana R, Ballard G, Hooker G (2009) Data-intensive science: a new paradigm for biodiversity studies. BioScience 59: 613–619. doi: 10.1525/bio.2009.59.7.12

Kerlinger P (1993) Birding Economics and Birder Demographics Studies Conservation Tools. In: Finch D, Stangel P (Eds) Proceedings of the Status and Management of Neotropical Migratory Birds. Rocky Mountains Forest and Range Experimental Station, Fort Collins, CO. USDA Forestry Service General Technical Report RM-229, 32–38.

Key Perspectives Ltd (2010) Data Dimensions: disciplinary differences in research data-sharing, reuse and long term viability. DCC Scarp Synthesis Report. ISSN 1759–586X. [http://hdl.handle.net/1842/3364]

Kidd DM, Liu X (2008) GEOPHYLOBUILDER 1.0: an ARCGIS extension for creating "geophylogenies". Molecular Ecology Resources 8: 88–91. doi: 10.1111/j.1471-8286.2007.01925.x

Klump J (2011) Criteria for the trustworthiness of data-centres. D-Lib Magazine vol. 17. doi: 10.1045/january2011-klump

Kobilarov G, Dickinson I (2008) Humboldt: exploring linked data. In: Proceedings of the WWW '08 Workshop on Linked Data on the Web.

Kohnke D, Costello MJ, Crease J, Folack J, Martinez Guingla R, Michida Y (2005) Review of the International Oceanographic Data and Information Exchange (IODE). Intergovernmental Oceanographic Commission (IOC) IOC/IODE-XVIII/18.

Lambrix P, Tan H (2008) Ontology alignment and merging. In: Burger A, Davidson D, Baldock R (Eds) Anatomy Ontologies for Bioinformatics: Principles and Practice. Springer, 134–149.

Langille MGI, Eisen JA (2010) BioTorrents: A file sharing service for scientific data. PLoS ONE 5(4): e10071. doi: 10.1371/journal.pone.0010071

Lee CP, Dourish P, Mark G (2006) The human infrastructure of cyberinfrastructure. Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work. doi: 10.1145/1180875.1180950

Lynch CA (2003) Institutional repositories: essential infrastructure for scholarship in the digital age. Libraries and the Academy 3: 327–336. doi: 10.1353/pla.2003.0039

MacLeod N, Benfield M, Culverhouse P (2010) Time to automate identification. Nature 467: 154–155. doi: 10.1038/467154a

Madin J, Bowers S, Schildhauer M, Krivov S, Pennington D, Villa F (2007a) An ontology for describing and synthesizing obervation data. Ecological Informatics 2: 279–296. doi: 10.1016/j.ecoinf.2007.05.004

Madin JS, Bowers S, Schildhauer SM, Jones MB (2007b) Advancing ecological research with ontologies. Trends in Ecology and Evolution 23: 159–168. doi: 10.1016/j.tree.2007.11.007

Mandavilli A (2011) Trial by twitter. Nature 469: 286–287. doi: 10.1038/469286a

Marris E (2010) Supercomputing for the birds. Nature 466: 807. doi: 10.1038/466807a

McCown F, Liu X, Nelson ML, Zubair M (2006) Search engine coverage of the OAI-PMH Corpus. IEEE Internet Computing, 10: 66–73 doi: 10.1109/MIC.2006.41

Michener WK (2006) Meta-information concepts for ecological data management. Ecological Informatics 1: 3–7. doi: 10.1016/j.ecoinf.2005.08.004

Mons B, Velterop J (2009) Nano-publication in the e-science era. In: Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009), Washington DC, USA. [http://www.surffoundation.nl/SiteCollectionDocuments/Nano-Publication%20-%20Mons%20-%20Velterop.pdf]

Morris R (2010) GBIFKOS Draft White Paper v 2010_11-25-0400. [http://community.gbif.org/pg/file/BMorris/read/10694/gbifkos-draft-white-paper-v-2010_11250400]

NAS (National Academy of Sciences) (2009) A New Biology for the 21st Century, 112 pp.

National Science Board (2010a) Science and Engineering Indicators 2010, Chapter 5, Academic Research and Development. [http://www.nsf.gov/statistics/seind10/c5/c5h.htm]

National Science Board (2010b) Globalization of Science and Engineering Research. [http://www.nsf.gov/statistics/nsb1003/]

Norris M, Oppenheim C, Rowland F (2008) The citation advantage of open access articles. Journal of the American Society of Information Science and Technology 59: 1963–1972. doi: 10.1371/journal.pbio.0040157

NSF (National Science Foundation) (2003) Revolutionizing science and engineering through cyberinfrastructure: report of the national science foundation blue-ribbon advisory panel on cyberinfrastructure. 84 pp. [http://www.nsf.gov/od/oci/reports/atkins.pdf]

NSF (National Science Foundation) (2006) NSF's Cyberinfrastructure Vision for 21st Century Discovery ver 5.0. NSF Cyberinfrastructure Council, 32pp. [http://www.nsf.gov/od/oci/ci_v5.pdf]

OECD (1999) Final Report of the megascience forum working group on biological informatics. OECD, Paris.

PARSE (2009) PARSE.Insight: INSIGHT into issues of permanent access to the records of science in Europe. [http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf]

Patterson DJ, Faulwetter S, Shipunov A (2008) Principles for a names-based cyberinfrastructure to serve all of biology. In: Minelli A, Bonato L, Fusco G (Eds) Updating the Linnaean Heritage: Names as Tools for Thinking About Plants and Animals, 153–163.

Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP (2010) Names are key to the Big New Biology. Trends in Ecology and Evolution 25: 686–691. doi: 10.1016/j.tree.2010.09.004

Piwowar HA (2011) Who shares? Who doesn't? Factors associated with openly archiving raw research data. PLoS ONE 6: e18657. doi: 10.1371/journal.pone.0018657

Piwowar HA, Day RS, Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. PLoS ONE 3: e308. doi: 10.1371/journal.pone.0000308

Piwowar HA, Vision TJ, Whitlock MC (2011) Data archiving is a good investment. Nature 473: 285. doi: 10.1038/473285a

Porter JH, Callahan JT (1994) Circumventing a dilemma: historical approaches to data-sharing in ecological research. In: Michener WK, Brunt JW, Stafford SG (Eds) Environmental Information Management and Analysis: Ecosystem to Global Scales. Taylor & Francis Ltd, London, 193–202.

Pullin AS, Salafsky N (2010) Save the whales? Save the rainforest? Save the data! Conservation Biology 24: 915–917. doi: 10.1111/j.1523-1739.2010.01537.x

Raup D (1991) Extinction: Bad Genes or Bad Luck? Norton and Co., New York.

Reichman OJ, Jones MB, Schildauer MP (2011) Challenges and opportunities to open data in ecology. Science 331: 703–705. doi: 10.1126/science.1197962

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Carcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, YoonJ, Zhang P (2003) The *Arabidopsis* information resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. Nucleic Acids Research 31: 224–228. doi: 10.1093/nar/gkg076

RIN (Research Information Network) (2008) To share or not to share: publication and quality assurance of research data outputs. A report commissioned by the Research Information Network. [http://www.rin.ac.uk/data-publication]

Rogers EM (1983) Diffusion of innovations. 3rd Edition. Free Press, New York.

Savage CJ, Vickers AJ (2009) Empirical study of data-sharing by authors publishing in PLoS journals. PLoS ONE 4: e7078. doi: 10.1371/journal.pone.0007078

Schofield PN, Eppig J, Huala E, Hrabe de Angelis M, Harvey M, Davidson D, Weaver T, Brown S, Smedley D, Rosenthal N, Schughart K, Aidinis V, Tocchini-Valentini G, Hancock JM (2010) Sustaining the data and bioresource commons. Science 330: 592–593. doi: 10.1126/science.1191506

Science staff editorial (2011) Challenges and opportunities. Science 331: 692–693.

Shirky C (2005) Making digital durable. [http://video.google.com/videoplay?docid=4000153761832846346&hl=en]

Silvertown J (2009) A new dawn for citizen science. Trends in Ecology and Evolution, 24: 467–471. doi: 10.1016/j.tree.2009.03.017

Sinha AK, Malik Z, Rezgui A, Barnes CG, Lin K, Heiken G, Thomas WA, Gundersen LC, Raskin R, Jackson I, Fox P, McGuinness D, Seber D, Zimmerman H (2010) Geoinformatics: transforming data to knowledge for geosciences. GSA Today 20: 4–10. doi: 10.1130/GSATG85A.1

Sirovich L, Stoeckle MY, Zhang Y (2010) Structural analysis of biodiversity. PLoS ONE 5:e9266. doi: 10.1371/journal.pone.0009266

Smith VS (2009) Data publication: towards a database of everything. BMC Research Notes 2: 113. doi: 10.1186/1756-0500-2-113

Smithsonian Institution (2011) Sharing Smithsonian digital scientific research data from biology. Smithsonian Institution Office of Policy and Analysis, Washington DC. [http://www.si.edu/openda/docs/Rpts2011/DataSharingFinal110328.pdf]

Stein LD (2008) Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. Nature Reviews Genetics 9: 678–688. doi: 10.1038/nrg2414

Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Clark AM, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Frimes F, Hancock JM, Hardy NW, Hermjakob H, Julian Jr. RK, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novère N, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert Jr. CJ, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nature Biotechnology 26: 889–896. doi: 10.1038/nbt.1411

TEEB (2010) The economics of ecosystems and biodiversity: Mainstreaming the economics of nature: A synthesis of the approach, conclusions and recommendations of TEEB. United Nations Environment Program.

Thessen AE, Patterson DJ (2011) Data Issues in the Life Sciences. [http://dataconservancy.org/sites/default/files/Data%20Issues%20in%20the%20Life%20Sciences%20White%20Paper.pdf]

Tittensor DP, Mora C, Jetz W, Lotze HK, Ricard D, van den Berghe E, Worm B (2010) Global patterns and predictors of marine biodiversity across taxa. Nature 466: 1098–1101. doi: 10.1038/nature09329

United States Department of Labor (0000) Occupational Outlook Handbook, 2010–11 Edition. [http://www.bls.gov/oco/ocos047.htm]

Vision TJ (2010) Open data and the social contract of scientific publishing. BioScience 60: 330–330. doi: 10.1525/bio.2010.60.5.2

Vollmar A, Macklin J, Ford LS (2010) Natural history specimen digitization: challenges and concerns. Biodiversity Informatics 7: 93–112.

Webb TJ, Vanden Berghe E, O'Dor R (2010) Biodiversity's big wet secret: The global distribution of marine biological records reveals chronic under-exploration of the deep pelagic ocean. PLoS ONE 5: e10223. doi: 10.1371/journal.pone.0010223

White HC, Carrier S, Thompson A, Greenberg J, Scherle R (2008) The dryad data repository: a Singapore framework metadata architecture in a DSpace environment. Proceedings of the International Conference on Dublic core and Metadata Applications 157–162.

Whitlock MC (2011) Data archiving in ecology and evolution: best practices. Trends in Ecology and Evolution 26: 61–65. doi: 10.1016/j.tree.2010.11.006

Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ (2010) Data archiving. The American Naturalist 175: 145–146. doi: 10.1086/650340

Wren J, Bateman A (2008) Databases, data tombs and dust in the wind. Bioinformatics 24: 2127–2128. doi: 10.1093/bioinformatics/btn464

Zhang M, Kihara D, Prabhakar S (2007) Tracing lineage in multi-version scientific databases. Technical Report CSD TR 06–013, Purdue University. doi: 10.1109/BIBE.2007.4375599

Ziegler A, Mietchen D, Faber C, von Hausen W, Schöbel C, Sellerer M, Ziegler A (2011) Effectively incorporating selected multimedia content into medical publications. BMC Medicine 9: 17. doi: 10.1186/1741-7015-9-17

ZooKeys 150: 53–70 (2011)
doi: 10.3897/zookeys.150.2193
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science

Vincent S. Smith, Simon D. Rycroft, Irina Brake, Ben Scott, Edward Baker, Laurence Livermore, Vladimir Blagoderov, David Roberts

Natural History Museum, Cromwell Road, London, SW7 5BD, U.K.

Corresponding author: *Vincent S. Smith* (vince@vsmith.infos)

Academic editor: *L. Penev* | Received 3 October 2011 | Accepted 24 November 2011 | Published 28 November 2011

Citation: Smith VS, Rycroft SD, Brake I, Scott B, Baker E, Livermore L, Blagoderov V, Roberts D (2011) Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 53–70. doi: 10.3897/zookeys.150.2193

Abstract

The Scratchpad Virtual Research Environment (http://scratchpads.eu/) is a flexible system for people to create their own research networks supporting natural history science. Here we describe Version 2 of the system characterised by the move to Drupal 7 as the Scratchpad core development framework and timed to coincide with the fifth year of the project's operation in late January 2012. The development of Scratchpad 2 reflects a combination of technical enhancements that make the project more sustainable, combined with new features intended to make the system more functional and easier to use. A roadmap outlining strategic plans for development of the Scratchpad project over the next two years concludes this article.

Keywords

Taxonomy, database, Virtual Research Environment, Biodiversity, e-infrastructure

Introduction

In recent years the value of data as a primary research output has been increasingly recognised (RIN 2011). New technology has made it possible to create, store and reuse datasets, either for new analysis or for combination with other data in order to

answer different questions. Such data were typically made available as supplementary files published alongside their respective papers or submitted to data repositories that are linked back to the supporting publication. Either way, the act of data preservation happened close to the time of publication, and usually some considerable period after the dataset was initiated. This time lag acts as a major barrier to the development of public archives for research data.

At this crucial time when researchers would rather be dealing with the final development of their paper and moving on to new projects, they are asked to deal with the considerable challenge of formatting and depositing data, often using complex data standards that may be unfamiliar to the contributors. In these circumstances identifying the correct metadata to describe versions of these data is a major challenge, particularly since research practices increasingly involve large multi-contributor datasets that have developed and evolved over a considerable period of time (Smith 2009). Coupled with concerns about the risk of exposing data before the originators have fully exploited it, and the lack of standard norms for citing data, all but the most committed researchers are likely to be unmoved by calls to publish their data. As a result, data deposition is usually something of an afterthought for most researchers, with current efforts arguably driven by mandates from research funders and journal editors, rather than self-motivated individuals (Costello 2009).

A solution to this problem is to embed the process of data creation, archival and storage into a system that supports the research practices of the contributor community, a process made easier by the steady migration away from paper-based note taking and into direct electronic capture. This must support the data management needs of a project from its inception through to publication and store the entire data workflow, taking into account methodological steps that alter the data (such as equations and processing algorithms) throughout. With this as a goal the collection of accurate metadata about the lifecycle of these data can be captured, with the final data suitably structured for archiving. This is especially important to researchers that would rather not hand off control of their data to remote strangers. When the time comes to deposit data (at publication or the end of funding), the relevant information could easily be transferred to a different, public storage repository, or made more widely accessible within the system in which it was created, for public access.

A general class of systems that support this process are Virtual Research Environments (VRE). Their purpose is to help researchers to work collaboratively by managing the increasingly complex range of tasks involved in carrying out research on both small and large scales (Carusi and Reimer 2010). The concept of VREs is still evolving, but the term can be understood as a shorthand for the tools and technologies needed by researchers to do their research, interact with other researchers (who may come from different disciplines, institutions or countries) and to make use of resources and technical infrastructures available at local, national, and sometimes international scales. Critically, a VRE must incorporate the context in which those tools and technologies are used. As a result the detailed design of a VRE will depend on many factors including the research discipline and security requirements.

Scratchpads (http://scratchpads.eu/) are an example of a VRE framework that has been constructed to support the needs of specialists interested in natural history (Smith et al. 2009). The system allows people to create their own website that supports the particular needs of their research community by selecting a personalised choice of features, visual design, and constituent data. Within any one Scratchpad network, users self-assemble their data and activities, often around user-defined or imported vocabularies (including biological classifications). These vocabularies provide a mechanism for navigating and structuring content. They can also provide a quality control framework for standardising certain types of data. Each Scratchpad includes service layers that provide integration, analytical and publication functions that add considerable value to the user. The original Scratchpad architecture is described in Smith et al. (2009), which details the motivation for the project as well as the original technical framework that supports the system. Two full time developers lead the technical development of the platform, which is presently hosted on a single virtual server at the Natural History Museum, London. Additional developers contributing software modules used by the Scratchpads are based at several other institutions in the UK, continental Europe and the US. Development proceeds according to an agile model with the overall vision and direction managed by a wider group of stakeholders that are closely connected to the user community.

In September 2011 there were over 300 Scratchpad community networks running on the Scratchpad platform (http://scratchpads.eu/scratchpads/stats). Thematically, these networks reflect the varied interests of natural historians, but can be broadly broken down into sites concerning specific groups of taxa, biogeographic regions or projects and societies. Networks range from 1 to 1,049 registered users (mean, 15, mode 1), and are composed of a mix of professional scientists and amateur naturalists. Just 17 Scratchpad networks have more than 50 contributors and almost half of all networks (129) have only one contributor. Contributor number is not necessarily indicative of quality or impact of a network, since two of the ten most visited Scratchpads have just two contributors each. Collectively the Scratchpad platform had over 4,400 registered and active users who have created 337,507 pages (nodes) of content between February 2007 and September 2011 (Figure 1). Scratchpad networks are free to all users. During January to September 2011 the Scratchpads received an average of 41,000 unique visitors per month across the platform.

February 2012 will mark the fifth anniversary of the Scratchpad project. It will also mark the planned release of a major new version of the software that incorporates many new features. This work is possible thanks to the EU FP7 funded ViBRANT project (http://vbrant.eu/), which is an e-Infrastructure initiative designed to support the development of virtual research communities. Additional support is provided by the NERC funded eMonocot project (http://e-monocot.org/). This paper provides a description of new features that will be released in Scratchpads 2, the motivation behind their development, and a roadmap for the future development of the Scratchpads over the next few years. As such it builds on the technical description of the Scratchpads provided in Smith et al. (2009) and does not duplicate descriptions there unless the concept or the functional component has changed substantially since originally being described.

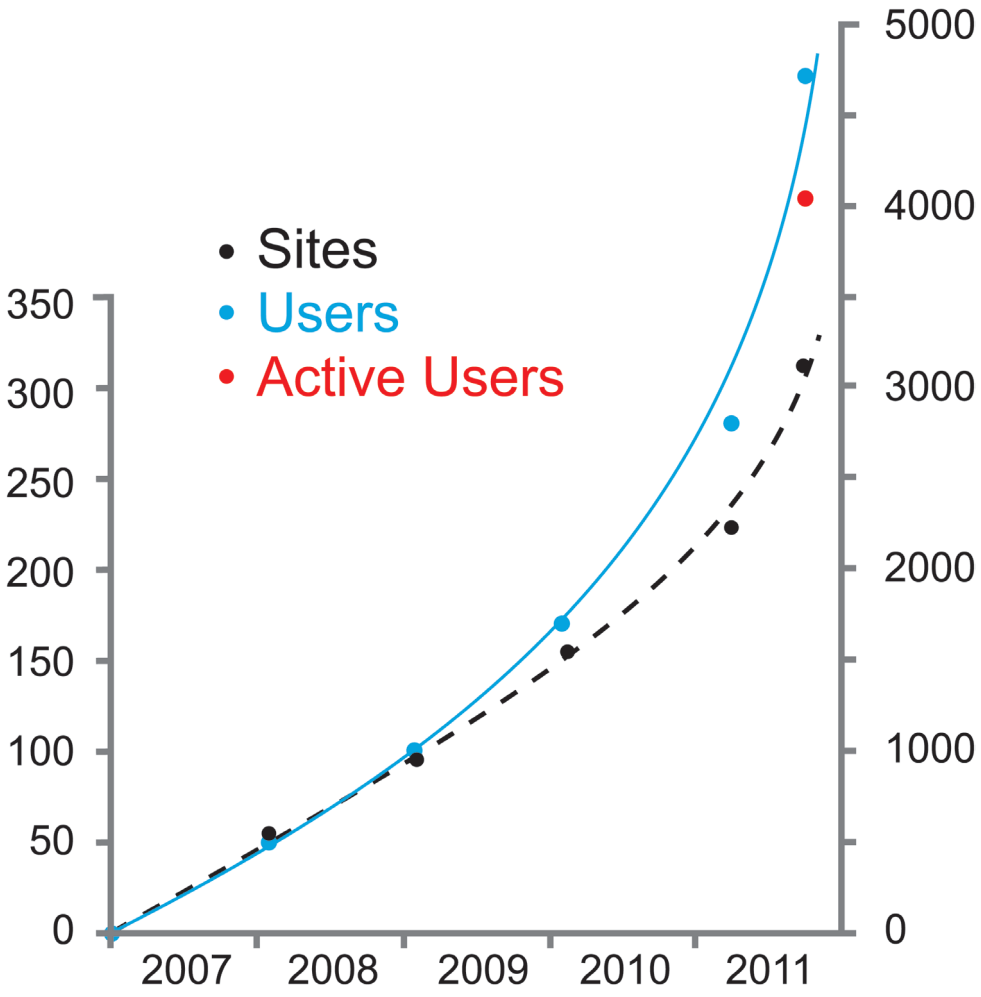


Figure 1. Scratchpad usage statistics from February 2007 to September 2011. The black dashed line represents the number of Scratchpad community sites (in hundreds) and the blue solid line represents the number of registered users (in thousands). As of September 2011 we have switched to recording the number of active users (currently 4424) since this figure provides a more accurate guide to usage.

Implementation

Development framework

Since their inception the Scratchpads have been developed using the Drupal (http://drupal.org/) Content Management System (CMS). Drupal offers a modular framework within which core functionalities can be readily extended through the development of new modules, or use of an extensive library of contributed modules. This approach means that the Scratchpads can make use of an extremely large community of

contributing developers that provide core functionalities common to many web-based applications (e.g. user management), in addition to a smaller pool of distributed developers providing niche functionality that have general applications within the system (e.g. bibliographic management). This makes the Scratchpad project more sustainable as it allows funding to be focused on the development of functionality specific to the biodiversity sector that is of direct application to the Scratchpads.

The Scratchpads were initially released in Drupal version 5 as part of the EU funded European Distributed Institute of Taxonomy project (EDIT, http://www.e-taxonomy.eu/). At the end of 2008 the Scratchpads were upgraded to Drupal version 6, and new modules have been constantly developed or modified since. Version 2 of the Scratchpads has been developed using Drupal 7, which offers significant benefits over previous versions (see Table 1).

Table 1. List of new Scratchpad 2 features. A list of features that are either new or significantly improved in Scratchpad 2, with short mention of major benefits and the previous Scratchpad 1 method.

New SP2 feature	Major Benefit	Previous SP1 method
Single primary units (entities)	Reuse of code, better linking & better normalizing	Three primary units (nodes, users & taxonomy)
Editing via overlay module (opaque editing environment)	More space, more intuitive link between editing and viewing of content	Editing in central area of webpage
Workflows	Easier navigation of tasks that involve multiple independent actions	Complex actions required to pursue a single goal (e.g. set up a site or import data)
More intuitive user interface	Easier navigation, more efficient use of space	Sometimes confusing and cluttered user interface
Consistent theming	More consistent and user-friendly layout of sites	Large choice of themes, colour schemes and layout
Profiles (project specific SP templates)	Specific configuration settings, choice of modules and theme for a set of project sites	Only one SP template for all sites
Integrated point and area maps	Display of specimen data, regional distribution and GBIF data in one map	Separate point and distribution maps, separate map with GBIF data
Extension of publication module	support of wider range of datasets and manuscripts; wider range of journals	Only prototype publication module available
Import from Excel files, dynamic templates generated directly from SP	Generation of import files much easier, data validated against controlled vocabularies before import	Import via comma or tab delimited UTF-8 files, only few templates available
Integration of a HTTP accelerator	Improved overall performance of Scratchpad platform	Application server accessed directly
Use of native RDF (Resource Description Framework) [planned for later instance of SP2]	Display of data embedded as RDF within the HTML allows content to be machine readable without the need for dedicated services	Only limited services for external data harvesting (e.g. harvesting of taxon descriptions by EOL)

Site management and distributed hosting

From April 2011 the Scratchpads adopted Ægir (Aegir, http://www.aegirproject.org/) as a site management tool. This provides a Drupal based hosting front end for the entire Scratchpad platform including all versions of the Scratchpads and Scratchpad training sites. Our configuration for Aegir allows sign up data to be automatically fed into the new site creation process, such that new sites can be set up in just a few clicks. To register for a new Scratchpad a user just has to complete a validated sign up form and the new Scratchpad is created automatically without any intervention by the Scratchpad development team. Backup and site upgrades are also managed by Aegir. Aegir also allows the Scratchpad team to deploy different Scratchpad profiles that have been developed to support sites with a subset of the full Scratchpad functionality (see below).

User feedback surveys have indicated a strong desire by more experienced users to host their own Scratchpads on a local server that is under their control. Until recently all production Scratchpads (i.e. publicly accessible sites in long term use) have been hosted at the NHM London. Attempts to host Scratchpads at other institutions have occurred, but none of these have gone beyond an experimental stage. As part of the ViBRANT project, technical development of the Scratchpads has enabled the existing NHM sites to be mirrored at the Botanic Garden and Botanical Museum Berlin (BGBM). In 2012, it will be possible to install new production sites on the BGBM server and we anticipate additional servers to come online in the near future. By distributing the hosting of the Scratchpads we hope to reduce the overall load on the NHM server that increasingly often reaches its performance limit when there are a high number of concurrent users. These distributed sites will also be centrally managed through the Scratchpad Aegir site (http://get.scratchpads.eu/).

Scratchpad project profiles

Interest in the Scratchpad project is more and more coming from project based initiatives in addition to individuals. The data-gathering needs of these projects usually map to a subset of the full functionality offered by the Scratchpads, but may require a high level of customisation and standardisation in order to support the efforts of a particular initiative. Using the same site model as the Scratchpads, these initiatives allow communities of users to construct data according to templates specific to an initiative, and often have particular branding requirements that identify that the sites are part of a common effort. As part of Scratchpad 2 we can now support this functionality through the development of dedicated Scratchpad *profiles*. These profiles contain configuration settings, a list of modules to install,

alternative themes and additional site setup settings that are specific to a particular initiative. Modifications to the Aegir site management have enabled us to deploy project specific profiles in the same way as regular Scratchpads. At present the only project to make use of this functionality is eMonocot (http://e-monocot.org/), which aims to create a global online resource for monocot plants by collating data provided by taxonomists working through dedicated eMonocot Scratchpads. There are, however, several potential applications for Scratchpad site profiles, including the GBIF (Global Biodiversity Information Facility) nodes portal toolkit, which is intended to be a mechanism for member countries to establish a web presence and view a subset of relevant species observation records from GBIF (http://www.gbif.org/). Another potential application of Scratchpad profiles are "LifeDesks" (http://www.lifedesks.org/). These are currently deployed in Drupal 6 by the Encyclopedia of Life (EOL) project (http://eol.org/) and are functionally very similar to the Scratchpads.

Code management

The Scratchpad project is Open Source and released under a GPL version 2 license. Originally the codebase was managed through a dedicated SVN repository. This was converted to a Git repository (https://git.scratchpads.eu/) in February 2011 to stay with the same system used by Drupal itself and to improve the development environment.

Within the repository there are two Scratchpad code branches. One (master) is used for development and contains the latest version of the code. This is inevitably unstable being the development environment, and it is less thoroughly tested than the second (stable) code branch. Code is released to the stable branch on an intermittent cycle, after it has been subjected to user acceptance testing by a trusted subsection of the Scratchpad user community.

Data services

A common criticism of version 1 of the Scratchpads was that each site was a data silo that lacked two-way connectivity to the wider landscape of biodiversity informatics initiatives (Page 2009). This criticism is partially justified. Scratchpad taxon pages provide significant inbound connectivity via the API's of a diverse collection of biodiversity projects and within the Scratchpads an increasing number of users are providing data via outbound connectivity to third party projects such as the EOL. Also users have long had the capability to create their own dynamic CSV or XML feeds on any data type present within the Scratchpads. Despite these functions, usage of the outbound connectivity from the Scratchpads is comparatively low. This problem will be

addressed within Scratchpad 2 by applying data services to all content by default, and more prominently advertising the presence of these functions.

Within Scratchpad 2 we will supply DwCA format, along with the appropriate extensions, for the majority of content. In some cases DwCA format is inappropriate or unsupported by external systems and services that are currently in use. For example, EOL species pages presently harvest Scratchpad content in a version of the Species Profile Model XML format. Likewise, the Scratchpad character project exports data in a variety of well-known formats for which there is no obvious DwCA extension. In these cases the present output formats (Structured Descriptive Data, Lucid format and Nexus format) will be maintained to keep interoperability with a wide array of third party applications.

DwCA files will be created at regular intervals for each site, as a background task, because building the archives is a comparatively slow process. We plan to drive this off the underlying database so that the archives dynamically reflect modifications to the structure of the site. Thus as new fields are added to the entity type, which define the appropriate DwCA extension field, their content will be dynamically mapped to the DwCA file when it is next created.

Consistent theming

For each current Scratchpad site the maintaining user (i.e. the site coordinator with administrative privileges) could choose between any of the default themes that came with Drupal 6. Some maintainers also selected themes from those on Drupal.org and requested that they be uploaded to their sites. Depending on the options that came with each theme, users could select to have menu-bars on the left, right or both sides of the page, customise the arrangement of content within these menu-bars, and alter the colour scheme. As a consequence some Scratchpad maintainers employed idiosyncratic layouts and colour schemes that did not make their site visually appealing to the widest possible audience.

As part of Scratchpad 2 this problem is addressed by the development of a new dedicated Scratchpad theme that provides less layout and colour scheme flexibility. This new theme will enforce compatibility with the Web Accessibility Initiative (WAI) Double-A standards (http://www.w3.org/WAI/). The theme will nevertheless offer a significant degree of customisation while allowing the Scratchpad development team to exploit a higher degree of layout standardisation. The goal is to present content in a more consistent and user-friendly way across all the sites. Dedicated themes will be developed for separate site profiles as these come on stream, allowing collections of sites to conform to the brands of commissioning initiatives. Note that this design decision will present certain challenges for existing sites, some of which may struggle to conform to the restrictions imposed by the new site theme.

Site administration

Users administrating version 1 of the Scratchpads found this a complex process because many administration functions are not intuitive, hard to physically find on the administrative interface, and when selected, their effect was often not immediately apparent. As part of the Scratchpad 2 release the administration back end has been completely redesigned with a new dedicated administration theme. This provides more intuitive grouping for the administration functions and makes the link between the cause and effect of each feature more obvious. For example, the options to configure menu-bar content are directly accessible from the menu-bar and altering these settings has an immediate visible effect. The administration functions also benefit from the full width display of the *overlay* module that provides a visual indication that the user is performing an administrative action.

Taxon pages

Scratchpad taxon pages provide a mechanism for users to dynamically construct and curate pages of information about any taxon selected from the site's biological taxonomy. These pages use taxonomic names as a search term to integrate tagged content in a Scratchpad with third party content external to the site. This third party content draws upon a variety of external data sources (e.g. *Biodiversity Heritage Library*, *flickr*, *GBIF* and *NCBI Genbank*), which have suitable APIs that support this type of integration.

The original implementation of taxon pages in Scratchpads version 1 suffered from a number of problems. These relate to the scientific accuracy of the third party content, the content selection interface, and the visual presentation of content, which may be poorly displayed and hard to organise for certain types of data. In consequence, many Scratchpad communities do not use the taxon page feature, or turn off the majority of third party content because the burden of curating these pages outweighs their perceived benefit. As part of Scratchpads 2 the taxon pages have been significantly re-engineered to address these issues, in part by making much greater use of EOL species page content. This is a close match to Scratchpad taxon page data. EOL provides a rich API that allows third party projects to access this information. To this end Scratchpads version 2 will use EOL as the primary provider for third party taxon page content. In addition we will work with EOL to support the rating and verification of source material through the API, such that registered Scratchpad users will be able to feed back to EOL content ratings and validate the status of content. EOL species page content will be integrated with existing Scratchpad taxon page content with the corresponding source clearly identified. A filter will allow Scratchpad users to choose whether to display just their Scratchpad Content, Scratchpad and trusted EOL content, or

Scratchpad and all EOL content. As in Scratchpads version 1, an on-demand citation can be generated for any taxon page that creates a permanent archived version of the page and a citation as well as a permanent URL for that page.

Mapping

Scratchpads version 1 supports three types of maps:

1) Point locality maps using the Google Maps API and the *gmaps* module, which are constructed dynamically from any content type containing geolocation data. Point locality maps are primarily used with Scratchpad specimen records but can also be applied to other appropriate content such as users.
2) The recording of taxon presence / absence distributions conforming to the TDWG level 4 geographical scheme. This is enabled by the *country maps* module.
3) Third party distribution maps dynamically obtained from GBIF via their API.

At present these maps are independent from each other and in consequence it is possible for a user to display a species page showing three, potentially conflicting, distribution maps for the same taxon. As part of Scratchpad 2 we will integrate these maps so that point information, and regional distributions can be displayed together. This will be implemented through an improved *Google Maps* module that incorporates version 3 of the Google Maps API. Feeds of georeferenced data from multiple sources (e.g. GBIF and FLICKR) can be displayed as points on a map, in addition to areas corresponding to TDWG level 4. As part of ongoing development work we plan to make these externally supplied map points and their metadata locally editable, such that individual records can be hidden, and point metadata edited locally within the Scratchpad.

Dynamic content templates and data import / export

Import mechanisms within Scratchpads version 1 operate on delimited text files for any content type (e.g. tab or comma delimited files, usually generated by users from spreadsheets). In addition, specific import mechanisms are provided for a limited number of additional data types including biological taxonomies. As part of the Scratchpad 2 development, data can now be imported directly into a site using an Excel template, omitting the need to convert the file into a delimited text file format. The template is dynamically constructed from the Scratchpad, ensuring that it reflects any underlying changes to the entity type, in much the same way that the DwCA and extension files do. Furthermore, this Excel template can incorporate validation directly from the user's Scratchpad. For example, a user may wish to import specimen records that directly link to a biological taxonomy that has already been embedded in the user's site. The template incorporates this taxonomy as a separate worksheet connected to the column containing the specimen records taxon name so that records are validated before the

import. The goal is to improve the user experience and reduce the number of errors that occur during data imports. The templates also contain embedded help text to guide users through the process of preparing their data. Technically this is made possible by the Drupal *feeds* module and the PHPExcel library.

Scratchpad workflows

Research on Scratchpads (Smith et al. 2010) and the Drupal CMS (http://drupalusability.org/) suggest that navigating tasks involving multiple independent actions (e.g. importing a biological taxonomy, or administrative tasks like adding new users) is the single greatest usability issue within the system. The problem has a significant effect on user retention because many users become frustrated when performing tasks that are infrequently required but have a profound impact on their site. Likewise, the need to perform complex actions, especially in the early stages of setting up a site, has been demonstrated to be one of the biggest barriers to entry for many new users (Smith et al. 2010).

In an attempt to address these issues the *form-flow* module has been developed by the Scratchpad team. This supports the construction of workflows, which are a mechanism to link together complex actions that would otherwise require the use of multiple forms, editing environments and menu selections in pursuit of a single goal. Form-flow allows the Scratchpad development team to integrate multiple-step forms into a single "flow". When a user complets the series of forms, they are collectively submitted as part of a single action. Error checks and validation are performed at every step, and users can navigate backwards and forwards between the component forms without loss of data. Within Scratchpad 2 form-flows exist for site setup functions; adding users and associated permissions; importing content including biological taxonomies; creating new entity types; publishing and exposing data through a service; and creating customised views of data. The entry point to these form-flows will replace the existing start point for these tasks, although maintainers will still have independent access to the underlying elements of a form-flow. In addition, maintainers can construct form-flows through the user interface.

Matrix editing

The matrix editor addresses the problem of how to edit multiple records for any entity in an intuitive editing environment while making efficient use of space within a webpage. The matrix editor emulates spreadsheet functionality in a web browser. The module (http://drupal.org/project/slickgrid) makes use of the jQuery slickgrid plugin (https://github.com/mleibman/SlickGrid) and defines a view-style in which all data can be handled within an editable grid. Features of the *slickgrid* module include grouping fields (to link logically connected fields); support for collapsible taxonomy fields (tree structures, such as those representing biological classifications);

tabs (to organise columns under tabs); deletion of multiple entities (e.g. rows) via the grid; multiple undo (to revert previous changes) and many more functions (see the module description at http://drupal.org/project/slickgrid for full details).

Character projects

The *character project* module is built on top of the *slickgrid* module and defines specialised plugins dedicated to describing the molecular and morphological phenotype of organisms. This enables users to manage complex collections of morphometric, text and DNA character states that are optionally controlled via selection of a limited number of predefined states. The data editor allows datasets to be entered, changed, and has numerous features for manipulating rows, columns, and blocks of data, and for recoding data. It supports the import and export of SDD (Structured Descriptive Data), Nexus and Lucid data files, and is intended to provide the framework for a more integrated suite of analytical and visualisation tools that will support the production of identification keys, phylogenetic trees and natural language descriptions of taxa. The c*haracter project* module also makes use of the advanced entity relationships possible in Drupal 7. These allow metadata to be recorded about the connection between one or more entities. For example, within the character project this provides a common method for states to be annotated with images, text and bibliographic references present within a Scratchpad database.

Publication module

A major long-term goal for the Scratchpads is to support users throughout the complete lifecycle of their data, from the inception of a project, through to its publication. As part of Scratchpads version 1 a prototype module was built that supported this functionality. This was outlined by Blagoderov et al. (2010) who described a method to publish nomenclatural acts via Scratchpads that are formally registered in the printed journal *Zookeys*. The workflow supports the generation of manuscripts directly from the Scratchpad database and is extended in Scratchpad 2 to support the construction of a wider range of datasets and manuscripts for submission to several additional endpoints. Within the first release of Scratchpads 2 these endpoints are limited to the major Pensoft series of journals (*Zookeys*, *Phytokeys*, and *Mycokeys*), as well as the construction of *Red List Threat Assessments* (Figure 2). The latter enable Scratchpad users to document the risk of extinction to species within a political management unit according to precise criteria defined by the International Union for Conservation of Nature (IUCN). Other publishers can implement software to handle the XML output from a Scratchpad, delivered in the open TaxPub schema, and, once available, their journals can be added to the list of possible endpoints.

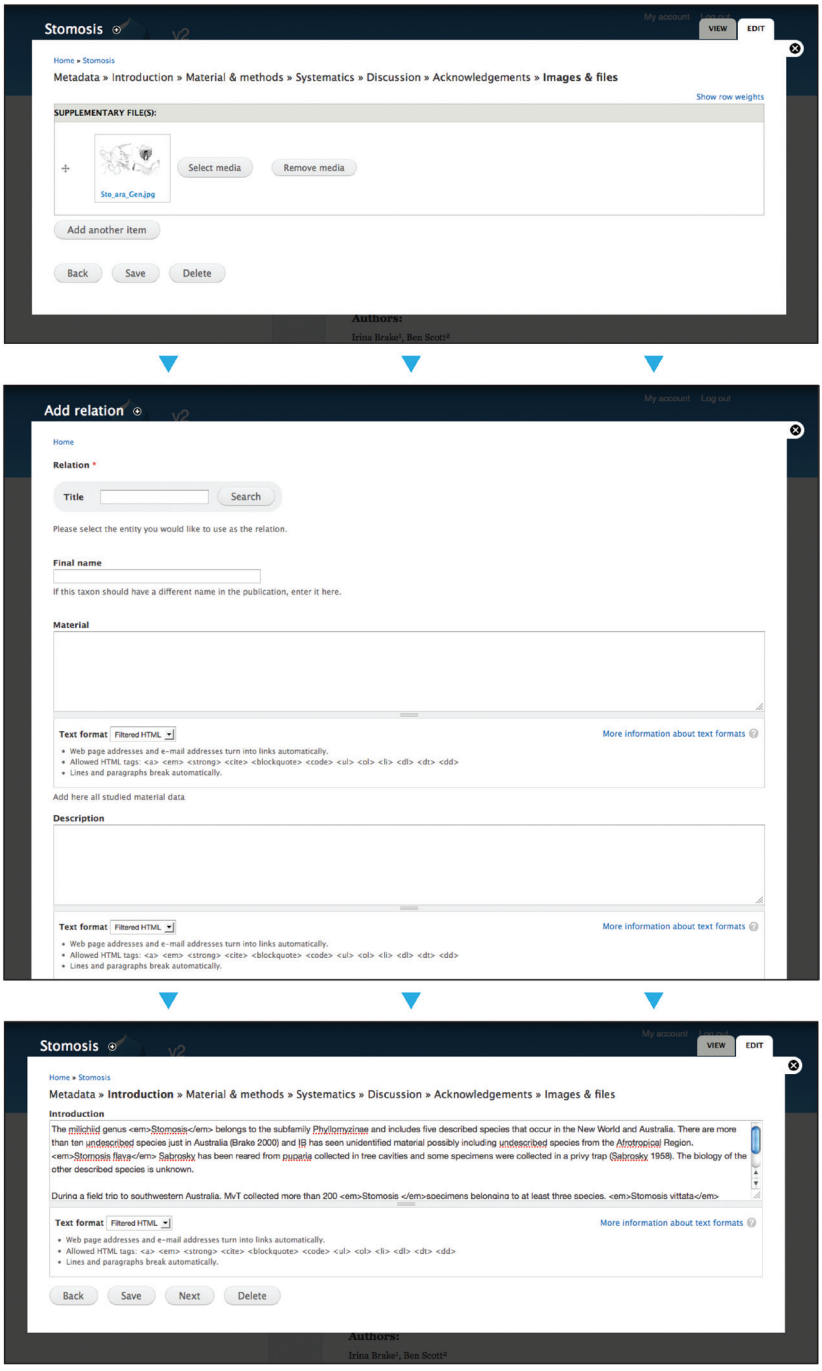


Figure 2. Screenshots of the Scratchpad 2 publication module showing an example workflow. Top, the section writing tool showing material and methods section; middle, the relationship selector that allows a taxon and additional materials to be associated with a section of the publication; and bottom, supplementary files such as illustrations, photos or graphs can be added to complete the publication.

When the publisher's API supports feedback mechanisms (such as comments received through peer review) the module will be further developed to automatically update the publication, with the goal of speeding up the process of editing the final document while maintaining an enduring link to the supporting data.

Help and support services

The Scratchpads have employed a variety of mechanisms over the past five years to support users (see Brake et al., this volume for a full review). Despite these advances, providing adequate support to a rapidly growing number of users remains an ongoing challenge. This is a particular problem with agile software development methods that can result in the rapid development of user interfaces, which occasionally require users to relearn tasks they previously performed by another method. To address this a help desk was formally established with the appointment of a dedicated user support manager in January 2010. The help desk deals with all the emails, issues, calls and meetings relating to user support. In September 2010 a custom-built issues tracker (http://dev.scratchpads.eu/project/issues) was developed that provided a mechanism for administrative users to report bugs, access support and make feature requests directly from their own Scratchpad, without the need to log into a separate system.

As part of the ViBRANT programme, basic and advanced training courses were organised to support and extend the Scratchpad userbase. These one-day courses are free of charge, paid for under the ViBRANT grant, and are intended to help current and prospective Scratchpad owners develop their site building skills, learn best practices and gain a better understanding of what Scratchpads can do for their research communities. A distance learning package has also been developed for those unable to attend a training course in person. Further, the help system has been extensively re-developed to become context-sensitive which helps novice users to control their Scratchpads. Throughout 2011 an extensive survey has been undertaken to further identify how the needs of users can be better supported. Full details of this are available in Brake et al. (2011) published in this volume.

Discussion

Prioritisation of these development activities for Scratchpad 2 have been conducted in close coordination with the user community, via feedback mechanisms that have been solicited and collated by a team within the ViBRANT project. This work has provided insight into the technical and social challenges faced by contributors when using the Scratchpads. Research into the motivation behind user engagement with the Scratchpads (Smith et al. 2010), has also led to the development of technical innovations designed to sustain engagement and expand the existing userbase. With these results in mind the development of Scratchpads 2 reflects a combination of backend enhancements intended to make the tech-

nical maintenance of the project more sustainable at a larger scale, coupled with new front-end features intended to make the Scratchpad system more functional and easier to use.

Based on the amount of time involved in the development of Scratchpads 2, the transition to Drupal 7 has proven harder than originally anticipated. We estimate that 13 person-months of developer time have been spent on transition of Scratchpads from Drupal 6 to Drupal 7. This compares to just 3 person-months on the Drupal 5 to Drupal 6 transition. However, the comparison of effort is not equal because subsequent developments to the Drupal 6 version of the Scratchpads have made the system much more complex and feature rich. For example, the initial Drupal 6 version of the Scratchpads contained fewer than half the number of Scratchpad specific modules than the Scratchpads contained just prior to the Drupal 7 redevelopment. In addition, the Drupal 7 transition has resulted in a complete redevelopment of the Scratchpad architecture.

Further complications to the development of Scratchpad 2 involve the transition to *entities* and *relations*, which are a defining feature of the Drupal 7 core architecture. These features were very poorly documented on Drupal 7's release in early January 2011. Drupal is an Open Source project and therefore dependent on volunteer contributions to upgrade; consequently, it has taken a very considerable period of time for Drupal developers to re-write their modules to take advantage of these functions in contributed modules relevant to the Scratchpads.

Despite these challenges, we expect the Drupal 7 transition to provide a much more sustainable platform on which to innovate and provide continued developments. Priority areas for development after the Scratchpad 2 release include:

– The production of a central registry for all the Scratchpad sites providing metadata on every entity type in each Scratchpad. This will also log user contributions, providing a mechanism to quantify activity that can be converted into a single contributor metric. In addition the registry will display statistics about non-contributing visitors. Registry functionality will replace the existing statistics pages at http://scratchpads.eu/scratchpads/stats and will be driven by enhancements to the *scratchpadify* module.
– Improved integration of polytomous keys and semi-automated construction of natural language taxonomic descriptions. These will be dynamically driven by the *character project* module that supports the documentation of taxon phenotypes, rather than statically creating keys from one-time exports, as is the case with the current Scratchpads.
– Integrated Single Sign On (SSO) across the Scratchpads, enabling users to access multiple Scratchpads with an existing login (such as a user's Google, Facebook, or Yahoo ID) rather than creating a new user login for each Scratchpad.
– Integration of Digital Object Identifiers (DOIs) for select content within Scratchpads. At present a number of communities are using the Scratchpads as a system for distributing specialist journal articles such as the *European Mosquito Bulletin* (http://e-m-b.org/) and Phasmid Studies (http://phasmid-study-group.org/content/Phasmid-Studies). Others are archiving datasets that have a persistent and lasting value to the wider community (e.g. the comprehensive citations of Milichiid flies at

http://milichiidae.info/content/citation). In an effort to formalise these outputs so that they are independently registered and citable we will be exploring the assignment of CrossRef DOIs to journal articles, and DataCite DOIs to datasets. Implementation of this function raises a challenge with respect to distributing the hosting of the Scratchpads and the maintenance of URL links. Nevertheless, this is an essential step for this output to become more readily accepted as formal scholarly content.

– The Scratchpad home site (http://scratchpads.eu) will be rebuilt with an emphasis on dynamically showcasing content from current Scratchpads, rather than emphasising the software.
– There will be greater integration of external analytical and vocabulary services into the Scratchpads. These will be driven by new developments from the ViBRANT programme, and include access to the catalogue of services available to the Oxford Batch Operation Engine (https://oboe.oerc.ox.ac.uk/) and developments to the GBIF controlled vocabularies server (http://vocabularies.gbif.org/).
– The Scratchpad training materials will be redeveloped with both botanical and zoological examples and will include support for training non-maintainer contributors from within a single taxonomic community (presently these materials focus on maintainers from multiple communities). As part of this redevelopment we will incorporate more standardised approaches to the training content that clarify the goal of a training task, alongside the prerequisites for its delivery, rather than just providing a set of step-by-step instructions and screenshots.

An ongoing issue with the Scratchpads and all e-infrastructure projects is finding an enduring model that secures their financial sustainability. In practice a mixed approach will be necessary for the Scratchpads, which relies on a combination of core support from institutions with a vested interest in the project, in addition to funds from external grant awarding bodies to drive innovation and new developments. As part of this mixed model we will be looking at opportunities to raise modest amounts of revenue from existing Scratchpad communities. This will take the form of value-added services such as priority technical support, maintenance of a persistent resolver for DOI identifiers on content, and data parsing services to facilitate the rapid construction of site.

Conclusions

We describe Scratchpads 2, a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. This represents a significant upgrade on the existing Scratchpad infrastructure. The original system has been in operation for five years demonstrating a clear demand for a structure of this type. The changes described here considerably expand the technical stability and functional capabilities of the system allowing the infrastructure to continue to grow at a sustainable cost. These changes include new tools to manage the distribution and hosting of sites, data services on all content, more consistent theming, new taxon pages, integrated mapping, dynamic

content templates, workflows, new data editing environments, a new publication module and improved user-support functions. The guiding principle used during the development of Scratchpads 2 has been to construct a scholarly communication system that closely resembles and is intertwined with the scholarly pursuit of natural history, rather than being its after-thought or annex. We would be the first to admit that Scratchpad 2 does not fully deliver this aspiration, but we believe that it lays sustainable groundwork towards this goal.

Availability and requirements

Project name: Scratchpads
Project home page: http://www.scratchpads.eu/
Operating system(s): Platform independent (Web application)
Programming language: PHP
Other requirements: none
License: Web application is freely accessible for all users. Source code is available under GNU General Public License version 2.
Content: remain the property of the contributors published under Creative Commons by-sa-nc licence.
Restrictions to use: none

Authors' contributions

VSS designs and leads the project, and coordinates the biological, sociological and technical insight that defines the Scratchpad program of work. SDR leads all aspects of the technical development, writing and integrating the package of software and providing the system administration. SDR also manages the additional technical developers including BS who has designed and constructed many of the complex editing and publication interfaces present within Scratchpad 2. EB has developed the profile functionality within the Scratchpads, in particular the eMonocot profile and associated training materials. EB alongside IB, LL, DR and VB provide selected testing and user support. IB leads the user support work and development of the training materials. DR realised, with VSS, the original Scratchpad implementation under the EU project EDIT (contract number 018340) and project manages the Scratchpads as part of the ViBRANT project, tests modules and develops selected functionality on selected sites. VSS wrote the manuscript. Other authors provided editorial comments and approved the final draft.

Acknowledgements

This work has been supported by the EU funded FP7 ViBRANT project (contract number RI-261532), and the NERC funded eMonocot project (grant reference NE/

H02185X/1). We would like to thank all members of the ViBRANT and eMonocot consortium, in addition to all contributing users of the Scratchpads for their enduring support for the project.

References

Blagoderov V, Brake I, Penev L, Roberts D, Rycroft S, Smith VS (2010) Blagoderov V, Brake I, Georgiev T, Penev L, Roberts D, Rycroft S, Scott B, Agosti D, Catapano T, Smith VS (2010) Streamlining taxonomic publication: a working example with Scratchpads and ZooKeys. ZooKeys 50: 17–28. doi: 10.3897/zookeys.50.539

Brake I, Duin D, Van de Velde I, Smith VS, Rycroft SD (2011) Who learns from whom? Supporting users and developers of a major biodiversity e-infrastructure. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 177–192. doi: 10.3897/zookeys.150.2191

Carusi A, Reimer T (2010) Virtual Research Environment collaborative landscape study. JISC, Bristol, 106 pp.

Costello MJ (2009) Motivating online publication of data. BioScience 59: 418–427. doi: 10.1525/bio.2009.59.5.9

Page RDM (2009) Wikis versus Scratchpads. iPhylo (Blog), http://iphylo.blogspot.com/2009/2001/wikis-versus-scratchpads.html

RIN (2011) Data centres: their use, value and impact. The Research Information Network, London, 60 pp.

Scollan B, Byrnes A, Nagle M, Coyle P, York C, Ingram M (2008) Drupal usability research report. Interaction design & information architecture, University of Baltimore, 23 pp.

Smith VS (2009) Data publication: towards a database of everything. BMC Research Notes 2: doi: 10.1186/1756-0500-1182-1113

Smith VS, Duin D, Self D, Brake I, Roberts D (2010) Motivating online publication of scholarly research through social networking tools. In: Webcentives: incentives and motivation for web-based collaboration at COOP2010, The 9th International Conference on the Design of Cooperative Systems. Aix-en-Provence, France, 1–9.

Smith VS, Rycroft SD, Harman KT, Scott B, Roberts D (2009) Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life BMC Bioinformatics 10(Suppl 14): S6: doi: 10.1186/1471-2105-1110-S1114-S1186

ZooKeys 150: 71–87 (2011)
doi: 10.3897/zookeys.150.2166
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Biodiversity information platforms: From standards to interoperability

W. G. Berendsohn, A. Güntsch, N. Hoffmann, A. Kohlbecker, K. Luther, A. Müller

Department of Biodiversity Informatics and Laboratories, Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin, Königin-Luise-Straße 6-8, 14195 Berlin, Germany

Corresponding author: *W. G. Berendsohn* (w.berendsohn@bgbm.org)

Academic editor: *V. Smith* | Received 29 September 2011 | Accepted 23 November 2011 | Published 28 November 2011

Citation: Berendsohn WG, Güntsch A, Hoffmann N, Kohlbecker A, Luther K, Müller A (2011) Biodiversity information platforms: From standards to interoperability. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 71–87. doi: 10.3897/zookeys.150.2166

Abstract

One of the most serious bottlenecks in the scientific workflows of biodiversity sciences is the need to integrate data from different sources, software applications, and services for analysis, visualisation and publication. For more than a quarter of a century the TDWG Biodiversity Information Standards organisation has a central role in defining and promoting data standards and protocols supporting interoperability between disparate and locally distributed systems. Although often not sufficiently recognized, TDWG standards are the foundation of many popular Biodiversity Informatics applications and infrastructures ranging from small desktop software solutions to large scale international data networks. However, individual scientists and groups of collaborating scientist have difficulties in fully exploiting the potential of standards that are often notoriously complex, lack non-technical documentations, and use different representations and underlying technologies. In the last few years, a series of initiatives such as Scratchpads, the EDIT Platform for Cybertaxonomy, and biowikifarm have started to implement and set up virtual work platforms for biodiversity sciences which shield their users from the complexity of the underlying standards. Apart from being practical work-horses for numerous working processes related to biodiversity sciences, they can be seen as information brokers mediating information between multiple data standards and protocols. The ViBRANT project will further strengthen the flexibility and power of virtual biodiversity working platforms by building software interfaces between them, thus facilitating essential information flows needed for comprehensive data exchange, data indexing, web-publication, and versioning. This work will make an important contribution to the shaping of an international, interoperable, and user-oriented biodiversity information infrastructure.

Keywords
EDIT, Common Data Model, CDM, Scratchpads, Standards, TDWG, biowikifarm, Taxonomy, Biodiversity, Biodiversity informatics

Introduction

In the last two to three decades there was a growing recognition that biological diversity is a global asset of tremendous value to present and future generations (Convention on Biological Diversity, UN 1992). This has led to a rising number of projects that gather data in the domain of biodiversity. The central component of biodiversity is organismic diversity, which is largely described in terms of systematics (the classification of organisms into taxonomic groups such as species), biogeography (the geographical distribution of the taxa in past and present), and synecology (the interaction of organisms in communities). "Biodiversity informatics" (Anon. 1999) focuses on this level of biodiversity, whilst the closely related and interacting of ecoinformatics and bioinformatics concentrate on ecosystems and on the molecular and related physiological level, respectively. Biodiversity informatics addresses data from preserved collections (natural history museums, herbaria), living collections (botanical and zoological gardens and culture collections), as well as from data collections from research (e.g. floristic and faunistic mapping, monitoring) and citizen science initiatives (e.g. bird watching). Another important data source is literature, especially taxonomic literature, which in its entirety (going back for more than 250 years) continues to be highly relevant for today's research. Last but not least, output from on-going research in systematics and synecology provides an ever-growing amount of data, extending into diverse new data types like cladograms, multimedia records of species, the specific data needed for new types of collections (e.g. DNA banks, Gemeinholzer et al. 2011), and a growing body of evidence about important functional attributes of organisms, such as a multitude of ecological traits, and also their potential to be invasive or serve as a vector for diseases.

Efforts to share these data soon led to the realisation that capture and storage of biodiversity data is not enough; although most of the attributes are shared across the entire domain, the data sets are not easily linked or integrated. The lack of shared vocabularies and the diversity of data structures used has impeded (and still impedes) the sharing of data. Data sharing is essential to facilitate the collaboration and large-scale analysis needed for a successful treatment of the pressing issues connected with biodiversity. Standards provide a consistent representation of the data to be shared enabling data from different sources to be combined, whilst minimising loss or duplication of data.

"Biodiversity Information Standards (TDWG)" is an organisation that works on defining such standards in the field of biodiversity informatics. TDWG was originally established as the "Taxonomic Databases Working Group" by major botanical institutions and projects from around the globe in 1985 (Anon. 2007). Task groups within TDWG initially worked on data dictionaries and exchange standards for botanical databases. Early examples for exchange standards are the "International

Transfer Format for Botanical Garden Data" (ITF, IUCN/WWF 1987) and the "Herbarium Information Standard and Protocol for Interchange of Data" (HISPID, Croft 1992). The "Descriptive Language for Taxonomy" (DELTA, Dallwitz 1980) was accepted for the encoding of taxonomic descriptions and identification keys. Standard works listing abbreviations for periodicals (Bridson and Smith 1991) and taxon authors (Brummitt and Powell 1992) as well as a newly devised standard scheme for geographical areas (Hollis and Brummitt 1992) were accepted as data standards. In the 1990s, the focus shifted to work on data models, which in turn revealed the high complexity of the domain (e.g. Berendsohn and Nimis 2000). Modelling efforts, albeit sometimes leading to extensive discussions of minute details, did serve to further stabilise the usage of terms and data format definitions in the domain (see Berendsohn 2005 for a compilation). The scope of TDWG was widened to include all organism groups and reached out beyond the taxonomic community, which recently also led to changing the organisation's name to "Biodiversity Information Standards (TDWG)". In the last decade, much discussion centred on community protocols for data exchange on the Internet, and the definition of appropriate XML schemas for data exchange. Based on all these developments, the discussion of how to achieve a joint semantic and structural description for domain specific data was recently revived (now under the term "ontology") and also included in the workplan of the ViBRANT project.

To be able to discuss the role of Biodiversity Information Platforms we need to have an exemplary look at some of the TDWG standards and other formats currently used in the field of taxonomy.

ABCD (Access to Biological Collection Data) and DwC (Darwin Core) are two standards intended to support the exchange of collection and observation data. Both have been ratified by TDWG as standard XML schemas. The ABCD standard (see Berendsohn 2007) set out to capture all data elements used in specimen and observation data collections that may be provided by collection information systems. It comprises nearly 1200 elements and attributes (including several hundred which are descriptors of elements, e.g. for language). No collection uses more than a fraction of the elements defined in ABCD, but the set of elements used varies greatly. The ABCD standard is directly used by the Global Biodiversity Information Facility (GBIF) and the Biological Collection Access Service (BioCASe). It has been extended to support the DNA Bank Network, the GeoCASE portal ("ABCDEFG", the "extension for geosciences") and the latest version of HISPID.

The DwC standard (Wieczorek et al. 2009) describes the occurrence of species and the existence of specimens in collections. It is a smaller set of data element definitions also designed to support the sharing and integration of primary biodiversity data. Efforts were made to keep DwC and ABCD largely compatible on the element level. DwC draft 1.4 is under discussion but already used in GBIF. Version 1.2 is used e.g. in the MaNIS (Mammal Networked Information System) and ORNIS projects (Stein and Wieczorek 2004). A variety of DwC is also used in the Ocean Biogeographic Information System (OBIS, Halpin et al. 2009).

TCS (Taxonomic Concept Transfer Schema, Anon. 2005) was developed for exchanging taxonomic data. However, TCS defines only the structure of the taxonomic backbone. For a broader export/import of taxonomic data other formats have to be used in addition (e.g. ABCD or DwC).

SDD (Structured Descriptive Data, Hagedorn et al. 2005) is the current TDWG standard for descriptive data. Many of the existing descriptive data managing tools, e.g. Lucid (Anon. 2010), Xper² (Ung et al. 2010), and DiversityDescriptions (Weiss et al. 2008) already support import and export of SDD conformant data, allowing their users to exchange highly structured descriptive data. See Hagedorn (2007) for references and an in-depth analysis of descriptive data and tools.

DwC-A (Darwin Core Archives, Robertson et al. 2009) is an updated and extended version of DwC. It is developed by GBIF in the context of the Global Names Architecture (GNA, Anon. 2011). DwC-A is based on the DwC terms and the DwC text guidelines, however, the extended version is not limited to occurrence data but also covers organism names, taxonomies, species information, factual data, distributions, media, and literature.

Taxonomy relies on the results of more than 250 years of research laid down in scientific publications. Digitisation of this content is well under way, but to become truly useful the content needs to be converted into structured databases. Efforts are under way to standardise the markup for the content of taxonomic literature as a prerequisite for this process. TaxPub is an extension of NLM/NCBI Journal Publishing DTD (Version 3.0) that adds elements and attributes relevant to taxonomic descriptions to the already included elements for document features (Catapano 2010). From within the community, the TaxonX schema (Sautter et al. 2007) was developed to streamline the process of text markup. See also Penev et al. (2011) for further information.

The work done has led to a comprehensive overview of the data in the highly complex domain of biodiversity informatics. But all these modelling efforts and resulting standards have no effect if the applications the researchers use cannot import and export standardised data. This is only starting to happen. For example, tools for descriptive data can exchange data using SDD, and some formats that are not (yet) TDWG standards such as Species Profile (Anon. 2009) and Plinian Core (Anon. 2007) are in practical use for data sharing by a number of applications (LifeDesks, Scratchpads, content partners of the Encyclopedia of Life and a variety of Spanish-language tools).

There is a need for workflow-based approaches for converting and integrating data and shielding the user from the complexity of the standards and data structures. Focusing on this problem, the European Distributed Institute of Taxonomy (EDIT) created the EDIT Platform for Cybertaxonomy. The EDIT Platform supports the entire taxonomic workflow, therefore it provides possibilities to import and export data in a standardised way (ABCD, DwC, SDD). Additionally, the EDIT-funded Scratchpads provide a scalable data publishing framework with flexible data models that can be modified by its users.

Biodiversity information platforms as information brokers

Lack of interoperability is one of the major obstacles to establishment of efficient workflows that help scientists and other users and user groups of the Biodiversity Informatics Infrastructures to improve quality and efficiency of their working processes. Advanced workflow management systems such as Taverna (Hull et al. 2006) and Kepler (Altintas et al. 2004) can greatly improve the execution of service-driven processes. However, there are still considerable technical barriers to overcome for users who wish to compose or re-use workflows from disparate services and data standards. Moreover, workflow management systems do not attempt to be comprehensive and to provide a complete working environment for entire research areas. Rather, they offer the means to streamline very specific sequences of data operations, which are time consuming and occur often in the day-to-day work processes.

In contrast, the emerging biodiversity information platforms implement a different and complementary approach by trying to cover many different scientific and other working activities and hiding underlying data models and access protocols completely from their users. These platforms are usually centred around a local or distributed data store based on a comprehensive information model providing a unified instance of all data needed for scientific activities ranging from field work to data publication on paper and in web portals. Moreover, biodiversity information platforms provide the necessary interfaces to deploy external software tools and services in a way that users can still work with often highly specialised software applications they are used to. Data from external applications can be seamlessly integrated and further processed in the platform environment. In this way, biodiversity information platforms exploit their potential as information brokers and help users to benefit from information standards, which they would be unable to deploy otherwise.

The ViBRANT project work package 4 (standardisation) aims to improve interoperability between biodiversity information platforms and focuses on three emerging systems: Scratchpads (Anon. 2006), EDIT Platform for Cybertaxonomy (Anon. 2007), and biowikifarm (Metawiki contributors 2011), which are briefly outlined in this section.

Scratchpads

The software platform Scratchpads (Smith et al. 2009) has been initiated by the European Distributed Institute of Taxonomy (EDIT) and is based at the Natural History Museum in London. The key aim of Scratchpads is to provide a scalable data publishing framework with flexible data models that can be modified or added to by its users. Automated integration of third party content and automated semantic enrichment of contributed and third party content are further key features of this platform. The principle design decisions for this platform are founded on the insight that the effort

(transaction cost) required by users to sufficiently structure (or restructure) their data is too high, relative to their perceived benefit from using the system. Thus it provides users with a system that allows assembling data quickly in a semi structured way.

Scratchpads are build on the content management system Drupal (2001), originally using version 5; at the time of writing it is being transitioned to Drupal 7. Making use of the data structuring principles provided by Drupal, data is organised around term vocabularies, such as biological classifications of taxon names. These vocabularies can be associated with various content types. Content is managed in so-called nodes, which can contain structured or semi structured data depending on the given content type. Structured content types are provided by specific modules like the biblio module (Jerome 2006) that allows users to manage and display lists of scholarly publications. The character node type allows users to build and manage structured descriptions of taxa in a controlled matrix. The set of predefined content types can be complemented by custom content types which users can define to adapt their scratchpad to their needs. This approach provides flexibility to accommodate use cases that were not originally envisaged, but at the cost of heterogenic data structures between the various scratchpad instances.

The content entities are related to each other by tagging them with terms from the associated vocabularies. In that sense taxonomic names provide a central link between diverse items of information about a taxon. Scratchpad taxon pages allow users to dynamically construct and curate pages of information (e.g. phenotypic, genomic, images, specimens, geographic distribution). External data from some third party data services (bhl, flickr, wikipedia, yahooimages) can also be dynamically aggregated into these taxon pages. Data provided by web services, however, in general can be placed only into taxon pages; it is not possible to integrate and process them in the local data structure. The only exception is taxonomic classifications which can be obtained in form of uBio ClassificationBank for Species 2000, ITIS and NCBI Genbank.

File based imports exists for classification terms, locations and specimen data which are all based on the CSV (Comma-Separated Values) file format. Following the principle of high flexibility none of these imports requires the data fields to be ordered in a predefined structure, thus these imports always involve user interaction and cannot be automated. Structured data in standardised data exchange formats only exist for bibliographic data. The Scratchpads can import Tagged EndNote, RIS, MARC, EndNote 7 XML, EndNote 8+ XML and BibTeX formatted bibliographies.

Scratchpads provide a limited range of services to expose data to other software systems. At present these are restricted to specimen and bibliographic data (Smith et al. 2009). Specimen data is provided by TapirLink software (DeGiovanni et al. 2007) external to the Scratchpads. TapirLink uses each set of Scratchpad database specimen records as a data source. These data fit the DarwinCore v1.2.1 standard (Wieczorek and al. 2009). Bibliographic data are currently available from the Scratchpads in BibTeX or Endnote format. Bibliographic data is also exposed using the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) module (OAI undated). In addition, Scratchpad users can create views of their data in arbitrary XML formats that

can be accessed by others. There is a special module to export descriptive data to EOL (C. Parr, in litt.).

The flexibility which allows adapting scratchpads to individual needs leads to semi-structured heterogenic data, which often hinders their integration into service-oriented software environments. A major task in achieving better platform interoperability will be to implement web service APIs, which communicate data in commonly accepted exchange formats.

EDIT Platform for Cybertaxonomy

The EDIT Platform for Cybertaxonomy (Berendsohn 2010), henceforth called "EDIT Platform" provides researchers with a set of coupled tools for: full, customised access to taxonomic data; editing and management of data; collaborative work in teams; and efficient publishing to both the web and in printed form. The EDIT Platform has been funded through the EDIT (European Distributed Institute of Taxonomy) project. Development of the EDIT Platform is coordinated by the Dept. of Biodiversity Informatics at the Botanic Garden and Botanical Museum Berlin-Dahlem, and its various components are being evolved by a team of software developers and architects from institutions all over Europe.

Establishing interoperability between various existing applications and data standards was a major aim in developing the EDIT Platform. A central data repository and information broker application has been created to achieve interoperability with and between existing applications and web based data providers. It allows other software to exchange data, via import and export functionality in major data formats, or via web services.

This data repository as well as the core components "EDIT Taxonomic Editor" and "EDIT DataPortal" are based on the EDIT Common Data Model (CDM), which comprehensively covers the information domain, including nomenclature, taxonomy, descriptive data, media, geographic information, literature, specimens, and persons. Wherever possible, the CDM has been made compatible with existing community data standards. This model as a base allows managing data consistently in highly structured form. A Java (Oracle 2011) application programming interface (API) for the CDM makes it easy to develop new CDM applications and to integrate existing applications. An example for the latter is the integration of Xper² (Ung et al. 2010) with the CDM. The CDM library provides an import and export package for taxonomic classifications, descriptive data, specimens and observations, and media in many standardised or quasi standardised data formats such as SDD, DarwinCore, TCS/RDF, TCS/XML, TaxonX and several MS Excel formats especially developed and in use for biodiversity data. In addition to that a generic XML export exists which allows dumping the entire CDM data base into a file and reimporting it.

The import and export functionalities are complemented by web services exposed by the CDM Community Server (EDIT 2011), a standalone server application which

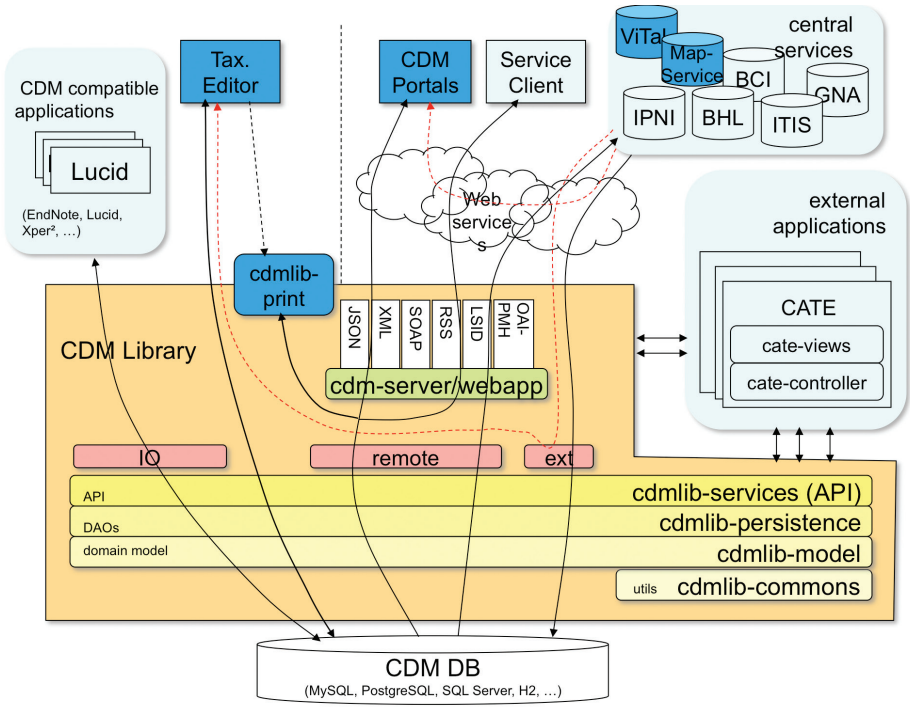


Figure 1. Architecture of the CDM Library and the EDIT Platform

can be connected to any CDM database. The major web service is the CDM REST (representational state transfer) API (EDIT 2011a), a RESTful (Fielding 2000) interface to all resources stored in the CDM. This web service exposes data items as XML or JSON serialisations. For example the EDIT DataPortal extensively uses the access points provided by this generic web service API; the same is true for the print publisher tool built into the EDIT Taxonomic Editor software. In perspective, this web service API is an excellent base for the future integration of EDIT Platform functionality into the above mentioned workflow environments; it needs only minor extensions in order to fully conform to these environments.

Another web service implements the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting, OAI undated) specification and thus allows aggregators like the Biodiversity Heritage Library (BHL), GBIF, and the Encyclopedia of Life (EOL) to harvest reference and taxon items selectively. However, for such large-scale aggregators who wish to harvest entire datasets and keep indices local and fresh, Darwin Core Archive is a better option.

EDIT Platform components like the EDIT Taxonomic Editor can directly use external data providers. This is made possible by service wrappers allowing querying, retrieving and integrating of external data into a CDM data store, where these remote objects can be reused, without losing the information on their origin. Service wrappers already exist for specific data providers like the International Plant Names Index

(IPNI 2008) and the Biodiversity Collections Index (BCI 2007). Other services implement widespread web service search protocols like OpenURL (OCLC 2003) and SRU (Search/Retrieve via URL, Library of Congress 2011) and thus enable all CDM based EDIT Platform components to find and integrate data from any data provider which exposes its data through these search interfaces.

The EDIT Platform architecture is mainly service oriented, thus all data flows between different applications are established through web services. The EDIT Map Services, for example, produce distribution and occurrence maps based on data coming from a CDM Store. The communication between both components is effected through a URI based web service API.

biowikifarm

Biowikifarm.net (Hagedorn et al. 2010) is a shared technical platform supporting a number of MediaWiki installations used by a diverse array of projects in biology, and especially in biodiversity research. The primary purpose of the shared platform is to enable long-term maintenance of the published data and to work more efficiently by distributing administrative and maintenance work among several partners. Furthermore, the biowikifarm operates a shared media repository, enabling synergies in reusing media content.

Using MediaWiki as technological basis allows biowikifarm to focus on the "long tail" of scientific information (Heydorn 2008). Supporting integration and preservation of this specific kind of unstructured or low structured data is a key feature by which it distinguishes itself from Scratchpads and the EDIT platform.

The biowikifarm is part of the activities of Plazi (Anon. 2008a). It is maintained through the Julius Kühn Institute (JKI; programming and management) and the Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM; technical support and hosting). The IT Centre of the Bavarian Natural History Collections (SNSB) is guaranteeing long-term online availability should dedicated project funds run out. In addition, users of the biowikifarm provide a significant contribution to the management of the farm.

Through the MediaWiki API, the software can be used as a service-oriented architecture, providing services to obtain page, file and relation objects (links, categories, semantic properties, data records) in a wide variety of data formats, including xml, rdf, json, html, and plain text.

Each of the MediaWiki installations contributing to this shared repository has different data structures, from simple arbitrarily structured wiki text pages to wikis like the "Offene Naturführer" (Anon. 2011a), which collects nature handbooks and determination guides in semi structured wiki pages. Even if all of them are sharing a common web service API the heterogenic and often unstructured nature of the content makes it hard to integrate the biowikifarm into workflow environments. This is a task, which has to be accomplished for each partner's MediaWiki individually.

Bringing it all together

None of the described platforms existed 5 years ago nor was there any commonly used tool available to edit and share biodiversity data in general and taxonomic core data in particular. At that time most applications designed for explicit handling of taxonomic core data (names, concepts, classifications) were in-house products, covering only the restricted requirements of local users and not supporting import and export of data in any standard format, perhaps with the sole exception of the databases providing collection and observation data in GBIF and related networks. User driven export of data to share it with other applications or projects was either not possible or ended up in user defined formats. An example is provided by the Global Compositae Checklist (Flann 2009), a project that aims to build up a global Compositae checklist based on local checklists. According to the coordinator (C. Flann, pers. comm.) they had to digest 55 different formats for a total of 67 data sets coming from 57 different sources, with only 1 dataset fully compliant with TCS the official standard for taxon classification data (several more were at least using TCS data definitions).

Obviously there was a considerable need for applications providing a joint platform for such projects. However, building them is more challenging then generally expected. To mention only the main obstacles: (1) a very complex and broad data domain covering the major fields of taxonomy and nomenclature, specimen and observations, descriptive data, literature, media, molecular data, and more; (2) high demands on the usability of user interfaces which may cover all the complexity of the domain; (3) the absence of a standard that covers all the domain – existing standards cover only parts and often overlap; and (4) the huge number of use cases to cover. The development is further complicated by the prerequisites for sustainable interoperability, namely (5) the demand for a generic open architecture that allows users with IT skills to adapt the software to their needs and allows participation in development (open source approach), (6) the demand for independence from hardware and operating system and database management platforms, and (7) the demand for web services, which make data and functions machine accessible and thus allow integration with other applications and with automated workflows.

Over the past years the described platforms first concentrated on the implementation of their core functionality, enabling users to do their every-day work of compiling, editing and integrating data, and publish the results on the web or as a print publication. At the same time, the basis for more advanced features was laid by building the systems using flexible and generic (though very different) architectures, as described in the previous section.

At present, all platforms have left the prototype status and are used to create content of high value. Although the list of demanded improvements and additional features is still long, it is now time to take a step back and reconsider how to integrate this content into the larger biodiversity e-infrastructure and how to connect the platforms in a way that creates additional value. All three platforms as well as other platforms like the emerging GBIF checklist bank (GBIF 2010) have specific characteristics that

make them attractive for certain users and certain use cases. For some of these use cases one may want to transfer data from one platform to another either manually or in an automated way using web service infrastructures.

As an example, we want to use the capability of the EDIT Platform to act as a data warehouse handling multiple classifications within one database for complex high level queries on several datasets compiled using Scratchpads and CDM implementations. For this, periodic import of all relevant data of the respective Scratchpads and CDM Data Stores into a CDM based database will be needed. An automated procedure using a service producing DwC Archive as the transfer format is being devised for this purpose. It is also envisioned to use the result as a contribution to the Global Names Architecture, once its setup becomes clear.

There are a number of other use cases for data exchange between platforms. Single users or user groups may want to compile data within one system but synchronise them with a repository based on another system to use it for other reasons. This use case is comparable with the handling of contact data. Present-day users do not necessarily hold their contact data within only one system but synchronise them among systems and tools each of them having their own purpose (direct calls, exchanging v-cards, sending FAXes, creating serial letters, advanced backup, synchronisation, etc.). Also with emerging requirements and growing software functionality users may want to switch systems without losing data or having to re-enter it manually, just as they may change their preferred mail client or word processing tool from time to time.

Moreover, other platforms can be used for backing up or versioning data. This may be a preferred use of biowikifarm, taking advantage of the highly developed versioning technology of MediaWiki. Exporting data from one platform to a MediaWiki may serve as a perfect way to fulfil the requirement of providing stable and accessible versions within a constantly changing data environment.

As described in the sections above, the Scratchpads as well as the EDIT Platform do already support a number of available and commonly used standards for data exchange. However, as most existing TDWG standards and other commonly used formats handle only a subset of the full data domain managed by the platforms these standards have only limited value for inter-platform data exchange. They are preferably to be used for the initial import or to enrich existing data. For example, ABCD and Darwin Core imports are used to add specimen data to the existing taxonomy data. SDD can be used to enrich taxon records by supplementing them with highly structured descriptive data. Also the various literature formats are very helpful for enriching a community site with a commonly shared literature repository. Users of platform software can communicate with, for example, the Biodiversity Heritage Library, both to use the indexed and digitised taxonomic literature, and to provide information on missing titles. On the export side existing standards like ABCD and Darwin Core are used to expose data subsets like specimen data to data aggregators like GBIF by using existing wrapper technologies such as BioCASe (Holetschek 2005) or TapirLink (DeGiovanni et al. 2007).

However, most of the standards or standard implementations have drawbacks as they cover only a certain slice of the data, or are not widely accepted by the community, and/or are not yet fully implemented or are implemented only for either import or export, not both. Also, many of the implementations are file-based, and do not provide the web services needed for use within automated workflows. This may be circumvented by implementing further software similar to the BioCASE software that is used for web service data exchange of ABCD data.

An interesting question in the context of web services is the problem of data rights and licenses. Where the schema for the resulting document does not contain a space for licensing information, presently there is no way for a user (especially a machine) to discern the licensing terms under which the data is provided. Care has thus to be taken to include this information, where possible (e.g. in the metadata for Darwin Core Archive or in the metadata section of the ABCD schema). Where not, appropriate service extensions must be defined. Of course, in an ideal world the data would be provided under a Creative Commons 0 license (see Hagedorn et al. 2011), with no rights reserved; but even that has to be made explicit.

For taxonomy-centred datasets TCS - the official TDWG standard for exchanging taxonomic data - should be the preferred format. However, TCS defines only the structure of the taxonomic backbone. Other data types such as specimen, literature or descriptive data need to be explicitly implemented using another format. This causes problems when trying to exchange broad and rich data like those stored in the Scratchpads or the EDIT Platform. Both sides need to support not only TCS but also all extensions used by the other side to fill in the gaps. As this requires considerable coordination efforts TCS export has not yet been fully implemented by both platforms and TCS imports still have limitations.

Darwin Core Archive (DwC-A), a new format developed by GBIF and others tries to address the described problems by offering a more comprehensive data format that covers all major areas of biodiversity data. It currently comes in two different flavours either taxonomy centred or specimen centred – but other implementations are possible. Although DwC-A has its limitations it is already much more widely used than TCS due to its ease of use and relative unambiguousness. However, all three platforms currently support DwC-A only in parts or not at all. Within ViBRANT this will be addressed. DwC-A import and export functionality will be implemented for the Scratchpads and the EDIT Platform; for MediaWiki it is probably sufficient to implement import functionality.

As DwC-A is primarily a file based exchange format a harvesting mechanism will be implemented to complement the export functionality, which allows automated harvesting of DwC-A data via web services. Within ViBRANT, this technology will be used in particular to integrate all Scratchpad data within one large EDIT Platform based database to allow visualisation and advanced querying across-Scratchpad (and EDIT Platform) data. The service implementation will enable users to provide access to selected slices of the dataset (e.g. to exclude access to data on research in progress). Adequate filter mechanisms need to be implemented that allow definition of exactly

which data should be exported. This may become a challenging task due to the complexity of the respective data models.

In contrast to the Scratchpads and especially the EDIT Platform, the data in MediaWikis are often unstructured or at most semi-structured. Creating generic export functionality for them will thus be very difficult, involving potentially extensive data curation measures, rendering it impractical in most cases. However, importing data will be straightforward and will enable users to use this platform as a repository for versioning and publishing. DwC-A could be used here as an exchange format, but as MediaWiki is a mainly text based system it may be better suited to export formats used for text publications. For example the emerging publication format TaxPub (Anon. 2008b) maybe more appropriate for this purpose. Within ViBRANT the format best suited for exporting data to MediaWiki will be investigated and a data flow based on the selected format will be implemented.

Conclusion

Biodiversity informatics faces an increasing need for integrated working environments facilitating efficient and streamlined data capture, processing, and publishing based on community standards. The EDIT Platform for Cybertaxonomy, Scratchpads, and biowikifarm each provide practical innovative software solutions which help their users who wish to organise their data in a standardised and networked manner. Further integration will be achieved in the course of the ViBRANT project by designing and implementing interfaces between the technologies. In work package 4 ("Standardisation"), the development of several data exchange modules will contribute to an improved overall interoperability between the EDIT Platform, Scratchpads and the biowikifarm as well as facilitate external connectivity. Based on a new DwC-A export module for Scratchpads and a corresponding import function built into the Java-API of the EDIT Platform for Cybertaxonomy, a comprehensive data index across all Scratchpad and CDM Datastore instances can be realised for the first time. The index will serve both (human) users wishing to perform cross-platform searches and software systems that need machine readable access to the "ViBRANT universe".

Connectivity between CDM stores and Scratchpads as well as CDM stores and the biowikifarm platform will be realised with XSL transformation of CDM XML publishing output. Based on these pipelines, data managed by an EDIT Platform installation can be further processed in a Scratchpad. Stable versions can be created with exports into the biowiki platform providing a semi-structured and addressable snapshot of dynamic taxonomic databases.

The processing of descriptive data will be handled as a complementary mechanism using the Xper^2 system. For this, SDD-based interfaces between the EDIT-Platform and Xper^2 will be implemented and optimised for the transfer of high volumes of descriptive information. Collaborative compilation and development of new character- and character state lists will be enabled through the biowikifarm system. A service for

the generation of interactive keys based on SDD-documents will greatly improve the user-friendliness of portal systems across platforms.

With this, the different platforms will for the first time be able to mutually benefit from their respective strengths. The new development will represent an important cornerstone for the establishment of a harmonised and consistent international biodiversity information infrastructure.

Achnowledgements

EDIT was funded within the European Union's Sixt Framework Programme under grant agreement n°018340.ViBRANT is funded within the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°261532. We thank the three reviewers (Cynthia Parr, Vladimir Blagoderov and Mike Sadka) as well as Gregor Hagedorn for their most helpful comments on the manuscript.

References

Altintas I, Berkley C, Jaeger E, Jones M, Ludascher B, Mock S (2004) Kepler: an extensible system for design and execution of scientific workflows. In: Scientific and Statistical Database Management, Proceedings. 16th International Conference on, San Diego, 423–424.

Anonymous (1999) "Biodiversity Informatics" – The Term. http://www.bgbm.org/BioDivInf/TheTerm.htm [accessed 20 Sept 2011]

Anonymous (2005) Taxonomic Concept Transfer Schema (Cover Page). Biodiversity Information Standards (TDWG), http://www.tdwg.org/standards/117/ [accessed 22 Sept 2011]

Anonymous (2006) Scratchpads. http://scratchpads.eu/ [accessed 22 Nov 2011]

Anonymous (2007) Plinian Core. http://www.pliniancore.org/en/inicio.htm [accessed 21 Nov 2011]

Anonymous (2007) EDIT Platform for Cybertaxonomy. http://wp5.e-taxonomy.eu/ [accessed 22 Nov 2011]

Anonymous (2008a) Plazi. http://www.plazi.org/ [accessed 22 Nov 2011]

Anonymous (2008b) TaxPub. http://taxpub.sourceforge.net/ [accessed 22 Nov 2011]

Anonymous (2009) Species Profile Model (SPM). http://wiki.tdwg.org/twiki/bin/view/SPM/WebHome [accessed 21 Nov 2011]

Anonymous (2010) Lucidcentral. Centre for Biological Information Technology,

The University of Queensland, Brisbane. http://www.lucidcentral.org/ [accessed 21 Nov 2011]

Anonymous (2011) Global Names - managing names, serving biology; the Global Names Infrastructure. NSF. http://www.globalnames.org/ [accessed 21 Nov 2011]

Anonymous (2011a) Offene Naturführer. http://offene-naturfuehrer.de/web/ [accessed 21 Nov 2011]

ASC (1993) An information model for biological collections (draft). Report of the Biological Collections Data Standards Workshop, August 18–24, 1992. Association of Systematic Collections, Committee on Computerization and Networking.

BCI (2007) Biodiversity Collections Index. http://www.biodiversitycollectionsindex.org/ [accessed 21 Nov 2011]

Berendsohn WG (1997) A taxonomic information model for botanical databases: the IOPI model, Taxon 46, 283–309. [Reprint:] http://www.bgbm.fu-berlin.de/biodivinf/docs/IOPI_Model/ [accessed 21 Nov 2011] doi: 10.2307/1224098

Berendsohn WG, Anagnostopoulos A, Hagedorn G, Jakupovic J, Nimis PL, Valdés B, Güntsch A, Pankhurst RJ, White RJ (1999) A comprehensive reference model for biological collections and surveys. Taxon 48, 511–562. [Reprint:] http://www.bgbm.fu-berlin.de/biodivinf/docs/CollectionModel/, accessed 21 Nov 2011.

Berendsohn WG (2005) Standards, Information Models, and Data Dictionaries for Biological Collections. TDWG Subgroup on Accession Data, Berlin. http://www.bgbm.org/TDWG/acc/Referenc.htm [accessed 20 Sept 2011]

Berendsohn WG (2007) (Ed) Access to Biological Collection Data. ABCD Schema 2.06 - ratified TDWG Standard. TDWG Task Group on Access to Biological Collection Data, BGBM, Berlin. http://www.bgbm.org/TDWG/CODATA/Schema/default.htm [accessed 20 Sept 2011]

Berendsohn WG (2010) Devising the EDIT Platform for Cybertaxonomy. In: Nimis PL, Vignes Lebbe R (Ed) Tools for Identifying Biodiversity: Progress and Problems, ISBN 978-88-8303-295-0, EUT, Trieste, 1–6. http://www.openstarts.units.it/dspace/bitstream/10077/3737/1/Berendsohn,%20bioidentify.pdf [accessed 21 Nov 2011]

Berendsohn WG, Nimis PL (2000) The complexity of collection information. In: Berendsohn WG (Ed) Resource Identification for a Biological Collection Information Service in Europe (BioCISE). BGBM, Berlin. http://www.bgbm.org/biocise/Publications/Results/ [accessed 21 Nov 2011]

Bridson GDR, Smith ER (1991) Botanico-Periodicum-Huntianum/supplementum. Hunt Institute for Botanical Documentation, Pittsburgh.

Brummitt RK, Powell CE (1992) Authors of Plant Names. Royal Botanic Gardens, Kew, 732.

Catapano T (2010) TaxPub: An Extension of the NLM/NCBI Journal Publishing DTD for Taxonomic Descriptions. Proceedings of the Journal Article Tag Suite Conference 2010 [Internet]. National Center for Biotechnology Information, Bethesda (MD). http://www.ncbi.nlm.nih.gov/books/NBK47081/ [accessed 21 Nov 2011]

Croft JR (1992) (compiler): HISPID - Herbarium Information Standards and Protocols for Interchange of Data. Summary paper and data dictionary. Australian National Botanic Gardens, Canberra.

Dallwitz MJ (1980) A general system for coding taxonomic descriptions. Taxon 29, 41–46. http://delta-intkey.com [accessed 21 Nov 2011] doi: 10.2307/1219595

DeGiovanni R, Vieglais D, Wieczorek C, Richards K, Hyam R (2007) TapirLink. http://wiki.tdwg.org/twiki/bin/view/TAPIR/TapirLink [accessed 21 Nov 2011]

Drupal (2001) Drupal homepage. http://drupal.org [accessed 21 Nov 2011]

EDIT (2011) CDM Community Standalone Server. http://wp5.e-taxonomy.eu/cdm-server/ [accessed 21 Nov 2011]

EDIT (2011a) CDM Library REST API. http://wp5.e-taxonomy.eu/cdmlib/rest-api.html [accessed 21 Nov 2011]

Fielding RT (2000) Architectural styles and the design of network-based software architectures. Doctoral dissertation, University of California, Irvine.

Flann C (Ed.) (2009) Global Compositae Checklist. http://www.compositae.org/checklist [accessed 21 Nov 2011]

GBIF (2010) GBIF Checklist Bank. http://www.gbif.org/informatics/name-services/checklist-bank/ [accessed 22 Nov 2011]

Hagedorn G, Thiele K, Morris R, Heidorn PB (2005) The Structured Descriptive Data (SDD) w3c-xml-schema, version 1.0, Biodiversity Information Standards (TDWG), http://www.tdwg.org/standards/117/ [accessed 21 Nov 2011]

Hagedorn G (2007) Structuring Descriptive Data of Organisms - Requirement Analysis and Information Models. Universität Bayreuth, Bayreuth. urn:nbn:de:bvb:703-opus-7273, http://opus.ub.uni-bayreuth.de/volltexte/2010/727/ [accessed 21 Nov 2011]

Hagedorn G, Weber G, Plank A, Giurgiu M, Homodi A, Veja C, Schmidt G, Mihnev P, Roujinov M, Triebel D, Morris RA, Zelazny B, van Spronsen E, Schalk P, Kittl C, Brandner R, Martellos S, Nimis PL (2010) An online authoring and publishing platform for field guides and identification tools. In: Nimis PL, Vignes Lebbe R (Eds) Tools for Identifying Biodiversity: Progress and Problems, ISBN 978-88-8303-295-0, EUT, Trieste, 13–18. http://dbiodbs1.units.it/bioidentify/files/volume_bioidentify_low.pdf [accessed 23 Sept 2011]

Hagedorn G, Mietchen D, Morris RA, Agosti D, Penev L, Berendsohn WG, Hobern D (2011) Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 127–149. doi: 10.3897/zookeys.150.2189

Halpin PN, Read AJ, Fujioka E, Best BD, Donnelly B, Hazen LJ, Kot C, Urian K, LaBrecque E, Dimatteo A, Cleary J, Good C, Crowder LB, Hyrenbach KD (2009) OBIS-SEAMAP: The world data center for marine mammal, sea bird, and sea turtle distributions. Oceanography 22(2): 104–115. doi: 10.5670/oceanog.2009.42

Heydorn B (2008) Curating the Dark Data in the Long Tail of Science. Google Tech Talks. http://www.youtube.com/watch?v=mgN74bR57i0 [accessed 21 Nov 2011]

Holetschek J (2005) (Ed) Biological Collection Access Services (BioCASe). http://www.biocase.org/ [accessed 21 Nov 2011]

Hollis S, Brummitt R (1992) World Geographical Scheme for Recording Plant Distributions. Plant Taxonomic Database Standards No. 2, International Working Group on Taxonomic Databases for Plant Sciences (TDWG). Hunt Institute for Botanical Documentation, Pittsburgh.

Hull D, Wolstencroft K, Stevens R, Goble C, Pocock M, Li P, Oinn T (2006) Taverna: a tool for building and running workflows of services. In: Nucleic Acids Research 34, iss. Web Server issue: 729–732.

IUCN/WWF (1987) The International Transfer Format (ITF) for Botanic Garden Plant Records. Plant Taxonomic Database Standards No. 1. Hunt Institute for Botanical Documentation, Pittsburgh.

Jerome R (2006) Drupal bibliography module. http://drupal.org/project/biblio [accessed 21 Nov 2011]

Library of Congress (2011) Search/Retrieval via URL. http://www.loc.gov/standards/sru/ [accessed 21 Nov 2011]

Metawiki contributors (2011) Biowikifarm. http://biowikifarm.net [accessed 22 Nov 2011]

OAI (undated) Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/pmh/ [accessed 21 Nov 2011]

Oracle (2011) Oracle Technology Network: Java. http://www.oracle.com/technetwork/java [accessed 21 Nov 2011]

OCLC (2003) OpenURL. http://www.oclc.org/research/activities/openurl/ [accessed 21 Nov 2011]

Penev L, Lyal CHC, Weitzman A, Morse DR, King D, Sautter G, Georgiev T, Morris RA, Catapano T, Agosti D (2011) XML schemas and mark-up practices of taxonomic literature. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 89–116. doi: 10.3897/zookeys.150.2213

Robertson T, Döring M, Wieczorek J, De Giovanni R, Vieglais D (2009) Darwin Core Text Guide. Biodiversity Information Standards (TDWG). http://rs.tdwg.org/dwc/terms/guides/text/index.htm [accessed 26 Sept 2011]

Smith VS, Rycroft SD, Harman KT, Scott B, Roberts D (2009) Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life, BMC Bioinformatics 10 (14): 6. doi: 10.1186/1471-2105-10-S14-S6

Sautter G, Böhm K, Agosti D, Klingenberg C (2009) Creating digital resources from legacy documents: An experience report from the biosystematics domain. Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications Heraklion, Crete.

Stein BR, Wieczorek J (2004) Mammals of the World: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics 1: 14–22. https://journals.ku.edu/index.php/jbi/article/view/7/5 [accessed 21 Nov 2011]

IPNI (2008) The International Plant Name Index. http://www.ipni.org/ [accessed 21 Nov 2011]

UN (1992) United Nations convention on biological diversity. Rio de Janeiro, 5 June 1992. http://treaties.un.org/doc/Treaties/1992/06/19920605%2008-44%20PM/Ch_XXVII_08p.pdf [accessed 21 Nov 2011]

Ung V, Dubus G, Zaragüeta-Bagils R, Vignes Lebbe R (2010) Xper²: introducing e-Taxonomy. Bioinformatics 26(5): 703–704. http://bioinformatics.oxfordjournals.org/cgi/reprint/btp715v1.pdf [accessed 21 Nov 2011] doi: 10.1093/bioinformatics/btp715

Weiss M, Hagedorn G, Triebel D (2008) DiversityDescriptions. http://www.diversityworkbench.net/Portal/DiversityDescriptions [accessed 21 Nov 2011]

Wieczorek J, Döring M, De Giovanni R, Robertson T, Vieglais D (2009) Darwin Core, Biodiversity Information Standards (TDWG) http://rs.tdwg.org/dwc/ [accessed 21 Nov 2011]

ZooKeys 150: 89–116 (2011)
doi: 10.3897/zookeys.150.2213
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

XML schemas and mark-up practices of taxonomic literature

Lyubomir Penev[1], Christopher HC Lyal[2], Anna Weitzman[3], David R. Morse[4], David King[4], Guido Sautter[5], Teodor Georgiev[6], Robert A. Morris[7], Terry Catapano[8], Donat Agosti[8]

1 *Bulgarian Academy of Sciences & Pensoft Publishers, Sofia, Bulgaria* **2** *The Natural History Museum, Cromwell Road, London, UK* **3** *Smithsonian Institution, Washington, DC, USA* **4** *The Open University, Milton Keynes, UK* **5** *IPD Böhm, Karlsruhe Institute of Technology, Germany & Plazi, Bern, Switzerland* **6** *Pensoft Publishers, Sofia, Bulgaria* **7** *University of Massachusetts, & Harvard University Herbaria, Boston, USA* **8** *Plazi, Bern, Switzerland*

Corresponding author: *Lyubomir Penev* (info@pensoft.net)

Academic editor: *V. Smith* | Received 7 October 2011 | Accepted 23 November 2011 | Published 28 November 2011

Citation: Penev L, Lyal CHC, Weitzman A, Morse DR, King D, Sautter G, Georgiev T, Morris RA, Catapano T, Agosti D (2011) XML schemas and mark-up practices of taxonomic literature. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 89–116. doi: 10.3897/zookeys.150.2213

Abstract

We review the three most widely used XML schemas used to mark-up taxonomic texts, TaxonX, TaxPub and taXMLit. These are described from the viewpoint of their development history, current status, implementation, and use cases. The concept of "taxon treatment" from the viewpoint of taxonomy mark-up into XML is discussed. TaxonX and taXMLit are primarily designed for legacy literature, the former being more lightweight and with a focus on recovery of taxon treatments, the latter providing a much more detailed set of tags to facilitate data extraction and analysis. TaxPub is an extension of the National Library of Medicine Document Type Definition (NLM DTD) for taxonomy focussed on layout and recovery and, as such, is best suited for mark-up of new publications and their archiving in PubMedCentral. All three schemas have their advantages and shortcomings and can be used for different purposes.

Keywords

mark-up, XML schema, taxonomy, TaxonX, TaxPub, taXMLit

Introduction

Traditional taxonomic publication has led to a vast quantity of valuable data effectively trapped in paper publications. Recent developments in transferring these to digital media,

particularly using PDF format and placing them on the web, have increased overall access to publications dramatically but not taken taxonomic publication to a format appropriate to today's methodologies of accessing and re-purposing data. Although simple searches of single or multiple documents may lead to the user finding the search terms in context, this context may not be what the user sought or, if the search is successful, the information sought (e.g. taxon treatments, specimen data) are not retrieved in a format suitable for repurposing (such as analysis of specimen data). To allow more precise searching for prioritised components of publications and retrieval of data in a format that is repurposable, taxonomic papers are being marked-up in XML and interfaces for queries being developed (Kirkup et al. 2005; Curry and Connor 2007, 2008; Agosti et al. 2007; Lyal and Weitzman 2008; Penev et al. 2010a; Willis et al. 2010).

The XML format has been identified as an important means of extending access to data from scientific papers (Murray-Rust and Rzepa 2002; Cui 2008b). Standards in XML for a range of taxonomic data have been developed through Biodiversity Information Standards (TDWG) such as for taxonomic names (Taxonomic Concept Transfer Schema), specimen data (ABCD, Darwin Core) and taxonomic descriptions (Structured Descriptive Data, SDD) (Hagedorn et al. 2005), for example, although so far there is no agreed-upon standard for taxonomic literature. An alternative to XML may be RDF, but there is less work done on RDF in the context of taxonomic literature; the relative merits and demerits of each will not be explored here, although it is worth noting that XML can be used as a stage in conversion to RDF where desired and appropriate (Cui 2008a; Cui et al. 2010a, and see below). XML mark-ups are currently being used both for new papers which are 'born-digital' and legacy literature, whose very varied structure poses much greater problems.

There are currently several different XML schemas and Document Type Definitions (DTD) (in the text, schema refers to both, unless specifically mentioned) being used for the mark-up of taxonomic literature, of which the three most widely used ones are discussed in this publication. The different schema designs reflect different priorities and consequently criteria for development. One distinction is whether the focus of the mark-up is on structure of the document as a whole (document-centric) or some part of the content of the document (content-centric). Another is the extent to which the marked-up text is potentially interoperable with (or using common elements with) other implementations. Notably, even with these distinctions, there are developing convergences between different approaches. An example of the content-centric approach is a focus on morphological descriptions (Heidorn et al. 2002; Cui and Heidorn 2007; Cui 2008a, b). In their work the publication is viewed more as metadata and the emphasis placed on the detail of morphological terms and the potential or repurposing the content. In this, the mark-up approaches SDD (Hagedorn et al. 2005), a schema produced explicitly for descriptive data. At the other extreme, some projects have employed a very generic schema to contain the document and structural information (i.e., pages, paragraphs, lines, headings, etc.) and used particular elements of taxonomic texts to assist in mark-up, relying on repeatable structural components of taxonomic descriptions (for example distributions, taxon names, morphological de-

scriptions, stratigraphic detail, etc.) (Kirkup et al. 2005; Curry and Connor 2007, 2008). Weitzman and Lyal (2004) used a version of the TEI-Lite schema (http://www.tei-c.org/Guidelines/Customization/) with some taxonomy tags as an interim mark-up standard in the INOTAXA project. This is a very generic solution to properly model the complexity of taxonomic texts and, while the broader TEI tag set can certainly be customized for retrospective conversion of legacy taxonomic literature, TEI-Lite *per se* is not an ideal fit; the version of TEI-Lite created has not been used outside the INOTAXA project.

More elaborate schemas have been designed to have a wide application to legacy taxonomic literature, provide access to more detail, and incorporate bibliographic information about the publication that is at least compatible with standards used in other sectors (particularly libraries). TaxonX (http://www.taxonx.org, http://sourceforge.net/projects/taxonx) was created by an interdisciplinary group around Plazi (http://www.plazi.org, see also http://en.wikipedia.org/wiki/Plazi) (Agosti et al. 2007; Agosti and Egloff 2009). The goal of TaxonX is to model taxon treatments in publications to provide a basis for data mining and extraction, while generic textual features are given marginal importance. A further schema,taXMLit, (Weitzman and Lyal 2004) (http://www.sil.si.edu/digitalcollections/bca/documentation/taxmlitv1-3intro.pdf; http://wiki.tdwg.org/twiki/bin/viewfile/Literature/WebHome?rev=1;filename=taXMLit_v5-04.xsd) has been developed as part of the INOTAXA project (www.inotaxa.org). It was seen as a step towards developing an interoperable system allowing simultaneous access to both literature content and other data types such as specimen data and names. The goal is to provide very flexible possibilities for data mining though tagging a wide range of components within the taxonomic papers.

TaxonX and taXMLit are mark-up XML schemas developed primarily to encode historical (legacy) taxonomic literature (implying any text post-publication including modern texts, although neither has been used by publishers as a vehicle to deliver new publications). In contrast, the TaxPub DTD (http://sourceforge.net/projects/taxpub), an extension of the DTD of the US National Library of Medicine (NLM, http://dtd.nlm.nih.gov), has been developed specifically to facilitate mark-up of new, "born digital" taxonomic publications as part of the publication process. While TaxonX has been developed primarily to model treatments but model the entire publication at a very generic level, taXMLit and TaxPub provide an extensive tag set (in TaxPub's case inherited from the base NLM DTD) for mark-up of generic (i.e., non Taxonomy-specific) document features, enabling location of relevant content throughout the document.

Once a document is marked-up into XML the full potential of that transformation can only be achieved through the creation of queries tailored to the schema elements. These can be incorporated into a portal for ease of human use, as well as built into web services. For TaxonX the portal is Plazi (http://www.plazi.org), for taXMLit the portal is INOTAXA (http://www.inotaxa.org).

An important aspect for use of a schema is the ease with which text may be parsed into it. A mark-up tool, GoldenGATE, was developed by Plazi (together with IPD Böhm at the Karlsruhe Institute of Technology, Germany) to facilitate this process

(http://plazi.org/?q=GoldenGATE). Pensoft Publishers have developed the Pensoft Mark-up Tool (PMT) based on TaxPub for routine use in their publishing practices (Penev et al. 2010a; b). Cui (2008a) and Cui et al. (2010b) discussed a mark-up tool for species descriptions.

Sautter et al. (2007a) compared seven different schemas for mark-up of taxonomic publications: ABCD, SDD/UBIF, TaxonX, taXMLit, Linnaean Core, Darwin Core and NCD (Natural Collection Description). The authors concluded that only four of them – ABCD, TaxonX , taXMLit and SDD/UBIF, were appropriate for mark-up of taxonomic documents; the first three of them have been evaluated as more "document-centric" and the last one as clearly "data-centric", the former being more optimal for mark-up of variously and inconsistently structured documents in the legacy literature than the latter. TaxonX and taXMLit have been analysed comparatively in order to investigate the possibility of mapping between them (Catapano and Weitzman 2007).

In this paper two schemas reviewed by Sautter et al. (2007a), ABCD (designed for specimen data) and SDD (designed for morphological descriptive data) are not considered further, as we assess them as much less appropriate for full mark-up of publications than the others. However, in the near future the relationships of the schemas designed for literature to more data-centric schemas, such as SDD and Darwin Core, should certainly be explored as being of primary interest for integration of "data-centric" and "document-centric" schemas.

The present paper aims at understanding the prioritized functions and scope of the three schemas most widely used for mark-up of taxonomic literature, namely TaxonX, taXMLit and TaxPub, and summarizes the experience and use cases accumulated during the four years following the analysis by Sautter et al. (2007a). In the context of an EU-funded project to support the development of virtual research communities involved in biodiversity science, ViBRANT, it is important to increase the compatibility of these schemas and this paper is a first step towards this.

The concept of "taxon treatment"

Perhaps the most significant component of taxonomic literature is the 'taxon treatment': information about a single taxon, typically headed by the taxon name and including morphological, distributional, taxonomic and other information about that taxon. Taxonomic treatments are important because they permit labelling and delimiting a dedicated piece of information describing a taxon within a document from other similar pieces of information, describing other taxa. The retrieval of this content type has been identified as valuable to users of marked up text through formal and informal assessment (Parr and Lyal 2007), and the importance of enabling the user to retrieve a digitized taxon treatment as a core element has been recognised by most projects employing XML for taxonomic publications (e.g., Weitzman and Lyal 2004; Kirkup et al. 2005; Lyal and Weitzman 2008; Agosti et al. 2007; Sautter et al. 2007a). Subsequent usages

of the marked-up paper, for example dissemination of content to various aggregators, can in some cases be performed at the level of treatments. In addition, marked-up text or data can be retrieved by machine from either within or outside treatments. Inevitably the concept of the taxon treatment is incorporated in most if not all schemas developed for taxonomic literature, both in the mark-up process and to inform user queries.

Determining the boundaries of taxon treatments in the mark-up process can be problematic and require manual intervention. Curry and Connor (2008) described the automatic identification and tagging of elements that typically occur within treatments, using stylistic rules to parse the text; they seem to have identified treatment boundaries *a priori*. More extensive algorithms also based on publication-specific stylistic rules (but not requiring *a priori* identification of treatment boundaries) were employed in a trial mark-up of a large single volume of the *Biologia Centrali-Americana* into taXMLit (Weitzman and Lyal 2006; Lyal and Weitzman 2008). The Plazi project atomises the publication into taxon treatments and, seek to maximize the number and consistency of tags by machine (either before or after publication) (Agosti et al. 2007; Catapano 2010; Penev et al. 2010a). The concept of taxon treatments from the viewpoint of their mark-up in taxonomic literature has been described by Catapano (2010) and Penev et al. (2010a). Therefore, we shall only briefly summarize the main features of treatments.

According to a definition by Norman Johnson (pers. commun.) adopted by Catapano (2010), a taxon treatment is a "publication or (more frequently) section of a publication documenting the features or distribution of a related group of organisms (called a "taxon", plural "taxa") in ways adhering to highly formalized conventions". Some of these conventions (those pertaining to a subset of the treatment dealing with nomenclature) are maintained by scientific commissions accepted by the taxonomic profession, including the *International Code for Zoological Nomenclature* (ICZN) for animals, and the *International Code of Nomenclature for algae, fungi, and plants (ICNafp)*.

There is considerable structural diversity in taxon treatments across taxonomic literature, the main sources of variation being historical differences in the approach to treatments between different groups of taxonomists and across time, and different editorial and publishers' formats. Nevertheless, it is possible to identify a few key features commonly found in treatments, such as the "Nomenclature" section, containing names and synonyms, "Material examined", containing data on the studied specimens, "Type designation" (for new or revised taxa), "Morphological description", "Etymology", on the origin of the newly proposed names, "Differential diagnosis" separating the taxon from similar taxa, as well as data on biology, ecology, or conservation status, etc.

Penev et al. (2010a) listed the following cases in which a logically delimited block of text within a taxonomy paper can be regarded as a taxon treatment:

1. New taxon description or re-description of a known taxon
2. Change of a nomenclatorial status of a taxon (a nomenclatural act)

3. Summary of some or all previous knowledge on a taxon from literature sources, usually structured in logical pieces, e.g., nomenclature, morphological description, distribution, ecology, biology
4. Summary of some or all previous knowledge plus newly published data on the same taxon, e.g., localities, ecological/biological observations
5. Summary of newly published data on an already known taxon
6. Summary of treatments of subordinated taxa, for instance a revision or catalogue of a genus listing treatments of ALL or SOME of its species is a treatment of that genus
7. Listing of subordinated taxa, e.g., a checklist of a family from a region forms a treatment of that family.

At the same time, the following cases do not usually constitute a treatment:

1. A citation of a taxon name within a text, although such a citation usually holds information linked to the particular taxon. For instance, listing of a species within a "plain" checklist cannot usually be a treatment of that species (in early literature under the ICZN such an instance must be considered a treatment in certain circumstances); a sentence within a text paragraph stating that "taxon X is parasitic on taxon Y" is neither a treatment of taxon X nor of taxon Y.
2. An identification key, because in some cases keys are constructed for related taxa that do not form a taxon (they may form a "species-group" or "taxa-group", but this is not a taxon unless a name is given to that group). Identification keys, even they are exhaustive for a named taxon, are usually tagged separately from taxon treatments. However, some keys include all of the information within a publication about a given taxon, and the practice may be to consider them treatments. In some cases keys include taxon treatments, including those of new taxa, or synonymies. How keys are tagged is probably an editorial matter.
3. A single picture or group of pictures of a taxon. In some early publications, however, a taxon is based exclusively on an image and its caption, a source which is available under the relevant Code, and therefore the picture and caption have to be regarded as a treatment.
4. A single map or group of maps of the occurrences of a taxon.
5. Gene sequence(s) of a taxon.
6. SDD (Structured Descriptive Data) (or any) matrices, or raw data, or databases. Treatments can be relatively easily generated from databases, however, and information on a taxon can be considered as becoming a treatment when (a) it is published, and (b) corresponds to the aforementioned description of a taxon treatment.

A publication may consist of one or many treatments of different taxa of different taxonomic ranks. One taxon may have more than one treatment within a publication, although the tradition of systematics publishing usually assumes one "core" treatment per taxon within a document. One treatment can include nested treatments, e.g., a genus and its species.

Implementation of the TaxonX schema and the TaxPub DTD largely follow the above restrictions. Implementation of taXMLit has been less restrictive in marking up complete papers, encompassing the less usual formats discussed above where appropriate, since more open-ended concepts of what makes a treatment have proven necessary, authors having been found to publish nomenclatural and taxonomic changes and treatments in a much wider variety of ways than listed in the more restricted list above. In the electronic era, broader notions of a treatment can easily be added to the electronic forms by simple extension of the schema or DTD.

Descriptions of schemas

1. TaxonX

1.1. Sources:

http://sourceforge.net/projects/taxonx/; http://www.taxonx.org/schema/v1/taxonx1.xsd; www.plazi.org, Sautter et al. 2007a

1.2. Description

TaxonX is an XML schema for encoding taxonomic literature in order to:

- Create open, stable, persistent, full text digital surrogates of taxonomic treatments
- Identify taxonomic treatments and their major structural components to enable networked reference and citation
- Identify lower level textual data such as scientific names and localities (Darwin Core or any other relevant schema may be used), morphological characters, and bibliographic citations in order to facilitate their extraction by, and integration with, external applications and resources
- Study and describe the structure of systematics publications by creating few typical corpora of literature, such as entire journals (e.g., AMNH Novitates, Zootaxa), taxa (e.g., all ant systematics papers post 1995), or faunistic studies (e.g. all ant systematics paper covering Madagascar ranging from 1758 to 2011)

TaxonX is a lightweight (with only 30+ elements) and flexible schema for mark-up of treatments which can be quickly learned and may be applied to the wide variety of formatting present in legacy documents as well as new publications. In many cases it relies on use of external schemas for modelling certain kinds of information [e.g., the use of MODS (Metadata Object Description Schema: http://www.loc.gov/standards/mods/) for file level bibliographical metadata; Darwin Core for observation

data: http://rs.tdwg.org/dwc/]. It has loose content requirements that allow for a wide variety of instances to be encoded over time and at many levels of granularity, while maintaining validity through iterations. Additionally, TaxonX contains mechanisms for semantic normalization of the data contained in treatments.

1.3. Design and development

Development of TaxonX began at the American Museum of Natural History (AMNH) and continued through the duration of a subsequent NSF/DFG grant (see below). As the project was concluding, participants established Plazi, a Switzerland-based independent not-for-profit organization aiming to help remove technological, social, and legal barriers to the creation of and access to taxonomic literature. Among its many activities, Plazi maintains the TaxonX schema and a repository of XML-encoded publications, develops the semi-automatic mark-up tool GoldenGATE (Sautter et al. 2007a), and strenuously advocates open access to scientific literature (Agosti and Egloff 2009).

TaxonX provides for the encoding of taxon treatments, with elements for the major structural components of treatments (e.g., Nomenclature, Materials examined, Description, etc.) and phrase-level features of interest in taxonomy (e.g., scientific names, locality names, characters, etc.) as well as mechanisms for linking to external resources and the semantic normalization of terms mentioned in the source document. The TaxonX instances encoded by Plazi contain a moderate degree of mark-up. Bibliographic metadata for the source documents are provided in each instance. Other sections of treatments are identified and named when they occur, but are not always present due to the wide variability of the structure of the source documents. All scientific names are marked and associated with an LSID, but other features may not always be identified. The section "Materials examined" can be broken down to individual materials citations, which in turn may normalized and linked to external resources, such as a type specimen, through LSIDs or other links.

A special emphasis has been given to link data to external resources, such as Life Science Identifiers (LSIDs). Tools in GoldenGATE have been developed to communicate automatically with external sources such as nameservers to retrieve LSIDs to taxonomic names in case they have already been entered, or to enter them upon discovery in an article, create the record and subsequently retrieve the LSIDs (e.g., in collaboration with the Hymenoptera Name Server), or on a manual base with Zoobank.

1.4. Implementations

Use Case 1: The GoldenGATE (GG) software tool (http://plazi.org/?q=GoldenGATE). GG development is lead by Guido Sautter (Sautter et al. 2007b) to serve the mark-up of legacy literature. GG itself is highly flexible and integrates a set of tools and modules that allow highly automated large-scale output of documents marked in TaxonX or other XML schemas. The use cases listed below have been performed using

GG. In 2010, GG launched a web interface to integrate social networking elements like crowdsourcing in the mark-up process.

Use Case 2: Ants of Madagascar. In 2006-2008, all available literature on the ants of Madagascar was OCR-ed, marked-up to the treatment level and stored on Plazi's treatment repository; this comprises ca 4,000 treatments from ca. 2,500 pages extracted from 119 legacy publications with taxonomic descriptions. The project formed the basis for the subsequent development of Plazi's mark-up projects (see below).

Use Case 3: The Zootaxa-TaxonX-ZooBank Project. In 2007, GBIF approved a Seed Money Award project entitled "Extracting Nomenclatural Data, Species Descriptions and Collecting Events from Legacy Publications: The Zootaxa-TaxonX-ZooBank Project" (GBIF Tracking Number 2007-94). Within this project, a TAPIR protocol has been developed for first time to render to GBIF occurrence data that have been marked up in taxonomic publications (http://data.gbif.org/datasets/provider/241).

Use Case 4: SPM (Species Profile Model) export from Plazi to Encyclopedia of Life (EOL). Plazi has developed a web service providing treatments in Species Profile Model (SPM) format allowing EOL and other interested parties, such as GBIF and others, to automatically harvest and consume content. Plazi received a small grant from EOL (managed by GBIF) to implement a service based on the SPM for the provision of taxonomic descriptions to EOL to complement a previous GBIF Seed Money Award to Zootaxa and Plazi that mobilised species occurrence records for the GBIF network (Use Case 3). The data for the project were taxonomic publications related to ants (Use Case 1). An XSLT conversion to SPM RDF/XML was developed and deployed as a web service using the eXist XML database (www.exist-db.org) so that SPM files generated dynamically from the TaxonX files can be retrieved via an HTTP GET request. A documented Application Programming Interface (API) is provided for the service, which allows the client applications latitude on tailoring the service. Sufficient documentation is provided so that clients can use the service for processing of the underlying XML document. At the date of writing (September 2011), 5892 treatments have been made accessible to EOL, including fish, ant and platygasteroid wasps.

Use Case 5: Overall content in taxonX. At the date of writing, 1,012 articles from 131 different journals and books spanning a period from 1758 to 2011 have been converted into TaxonX resulting in 15,863 treatments accessible on plazi.org. Most of the taxa covered are animals with an increasing number on plants and fungi taxa, (Plazi.org, accessed November 21, 2011).

1.5. Problems encountered and lessons learned

Based on accumulated experience, the following success factors of TaxonX can be summarized:

- It is a lightweight and flexible schema which can be quickly learned and may be applied to a wide variety of formatting found in legacy documents

- It relies on use of external schemata (see use of MODS for file-level bibliographical metadata).
- Its loose content requirements allow for instances to be encoded over time and at many levels of granularity, while maintaining validity through iterations.
- It contains mechanisms for semantic normalization of the data contained in treatments. See TaxonX's use of Darwin Core (soon perhaps Linnaean Core, SDD, etc.) to normalize phrase level data, and xid elements for inclusion of LSID's, ITIS, HNS, or other external identifiers.

However, there are also some hurdles for the adoption of TaxonX, such as:

- The heterogeneity and structural looseness of the data contained in some legacy taxonomic treatments makes encoding and semantic normalization even by a lightweight and flexible schema difficult and requires substantial expert intervention.
- The flexibility of the schema may present difficulties both in authoring and in profiling the encoded data for use by external applications as well as in conversion into other schemas/DTDs, but not at a very basic level, that is treatment and nomenclature element.
- Dependence on external schemas requires vigilance and active maintenance of the schema; may complicate long-term validation of instances; namespace wrangling makes authoring difficult
- Mark-up, even in a light way, needs some domain specific expertise, namely specific quality controls to assure that the elements are properly identified, and therefore costs time.

Potential users of TaxonX could be:

- Biodiversity Heritage Library would become much more useful if at least treatment boundaries, nomenclatural elements and respective names were to be marked-up and linked to the respective scan on BHL.
- Ultimately, one could envision this to be an intermediary step to extract and store the treatments in more powerful structures, such as databases. All the treatments are primarily linked to genetic, distributional, nomenclatural and other data via the taxonomic name applied to the treatment. At Antbase/HNS, this link is in a simple form already implemented by a link from each citation to the respective PDF copy of the referring page.
- Future aggregators of treatments might be institutions like ZooBank, or essentially dedicated databases allowing specific applications, like iSpecies (http:/www.ispecies.org), or the Taxon Pensoft Profile (http://ptp.pensoft.eu), to collect the treatments and use them for specific purposes.
- All aggregators that will benefit from improved search, information retrieval, and information extraction.

2. TaxPub

2.1. Sources:

http://sourceforge.net/projects/taxpub/; Catapano 2010

2.2. Description

TaxPub was designed with the aim to enable the mark-up of new "born-digital" taxonomic literature that could forgo unnecessary variation in style and form and adhere to a limited set of data elements so as to lower costs of both authoring and processing. TaxPub is an extension of the Journal Publishing Tag Set of the U.S. National Library of Medicine's Journal Archiving Tag Suite (see http://dtd.nlm.nih.gov/). For more details see Catapano (2010).

2.3. Design and development

Starting in 2008, TaxPub was designed and developed by members of Plazi with the assistance of experts from the U.S. National Center for Biotechnology Information. The TaxPub extension is maintained as an open source project at SourceForge (http://sourceforge.net/projects/taxpub/) inheriting from the base DTD an extensive and robust set of elements for generic textual structures while adding a small number of elements relevant to taxonomy. These include elements for mark-up of taxon names, citations to specimens and other material, and statements describing morphology, as well as for treatments and treatment sections. Further semantics may be applied to many elements through use of terms in external vocabularies (such as Darwin Core) as values of attributes (more details in Catapano 2010 and http://species-id.net/wiki/TaxPub).

TaxPub, being part of the National Library of Medicines Journal Article Tag Suite (JATS), has the additional advantage that it can directly be archived in PubMedCentral, one of the most secure existing archives and, as a consequence, its content is cross-linked with the huge body of biomedical literature stored therein.

2.4. Implementations

The first TaxPub encoded treatments were provided from the Ohio State University based "vSysLab" (Virtual Systematics Laboratory) presentation of data on wasps (Platygastroidea) described as part of the US National Science Foundation's Planetary Biodiversity Inventories program (see http://vsyslab.osu.edu/home_page.html).

Soon after the initial release of TaxPub, Plazi was joined by Pensoft, the publisher of the online open access taxonomy journal ZooKeys, in a collaboration to integrate TaxPub into its publication workflow. The approach differed from OSU's in applying mark-up to submitted manuscripts. Pensoft faced a set of challenges similar to those in retrospective conversion. Among them was the identification and encoding of treat-

ments, scientific names, and bibliographic references. Developing their own software tool (Pensoft Mark-up Tool, PMT, see Penev et al. 2010a), in 2010 ZooKeys began to publish TaxPub versions of their articles. Although lacking a very fine-grained level of mark-up granularity (for example, <material-citation> is not used), the ZooKeys articles accomplish many of the goals of the TaxPub extension. Treatments are identified, and thus are directly and easily machine addressable, as are treatment sub-sections. All scientific names and name parts are tagged with <tp:taxon-name> elements. <tp:nomenclature-citation> elements include <tp:taxon-name> and link to full bibliographic entries, themselves marked up with <mixed-citation>. Specifically, because TaxPub motivated and enabled its use of the NLM DTD, ZooKeys and PhytoKeys articles are approved for display and archiving in PubMedCentral.

The ZooKeys' exemplar papers (Stoev et al. 2010; Blagoderov et al. 2010b; Brake and von Tschirnhaus 2010; Taekul et al. 2010) are entirely based on revision #123 available from the SVN trunk of TaxPub (http://sourceforge.net/projects/taxpub). In fact, the present exemplar papers are the first published TaxPub articles in biodiversity science, intended to demonstrate the advantages of the XML-based mark-up and editorial workflow in the way biodiversity information is being published and disseminated.

Use Case 1: Editorial use at Pensoft. TaxPub is used to mark-up taxonomic papers during the editorial process, using the Pensoft Mark-up Tool (PMT). As a result, through PMT and InDesign, 3 electronic versions of a paper are generated and routinely published: (1) PDF identical to the printed version; (2) HTML to provide links to external resources and semantic enhancements to published texts for interactive reading; (3) XML version compatible to PubMedCentral archiving NLM DTD TaxPub extension), thus providing a machine-readable copy to facilitate future data mining.

Currently TaxPub is used routinely in the editorial process of six journals published by Pensoft:

- ZooKeys – www.pensoft.net/journals/zookeys
- PhytoKeys – www.pensoft.net/journals/phytokeys
- MycoKeys – www.pensoft.net/journals/mycokeys
- International Journal for Hymenoptera Research – www.pensoft.net/journals/jhr
- International Journal of Myriapodology – www.pensoft.net/journals/ijm
- Comparative Cytogenetics – www.pensoft.net/journals/compcytogen

In addition, the TaxPub DTD and some of its phrase-level elements, such as taxon names, are used in Pensoft's ecology journals:

- BioRisk – www.pensoft.net/journals/biorisk
- NeoBiota – www.pensoft.net/journals/neobiota
- Nature Conservation – http://www.pensoft.net/journals/natureconservation

Use Case 2: Export of new taxa to EOL. All new species descriptions in Pensoft journals are exported to EOL on the day of publication through a tool that maps the

content to EOL elements; the file contains bibliographic metadata, taxonomic classification, species description and links to the species images. The exported XML file is harvested by EOL on a daily basis.

Use Case 3: Export of taxon treatments to Plazi. All taxon treatments identified within the XML file of a published paper are harvested by Plazi and uploaded to the Plazi Treatment Repository. Thereafter, treatments are available for use by various organizations and individuals, e.g., EOL.

Use Case 4: Export of taxon treatments to the Wiki environment Species-ID (http://species-id.net/wiki/Main_Page). All taxon treatments at the level of genera and species identified within the XML file of a published paper are exported to Species-ID through a special software tool, including images, keys and bibliographies. The citation template of the taxon's wiki page automatically includes the original source (article) to provide a permanent scientific record, as well as all consequent contributors to the respective wiki page (Penev et al. 2011).

Use Case 5: Archiving in PubMedCentral. ZooKeys was accepted for indexing and archiving in PubMedCentral in August 2010, followed by PhytoKeys. Since then TaxPub XML output of ZooKeys issues 50-54 has passed 4 rounds of testing at NLM. All suggestions have been implemented in the XML export and, where needed, corrections implemented in TaxPub.

Use Case 6: Use of TaxPub XML files to create a semantically enhanced HTML version of the publication. The process was described and exemplified in issue 50 of ZooKeys (Penev et al. 2010a, b); from then it has become routine practice for several of Pensoft's journals (list provided above in the text).

Use Case 7: Acceptance of manuscript in XML by Journal. In ZooKeys 50, Penev et al. (2010b) and Blagoderov et al. (2010a) piloted acceptance of manuscripts in XML format, generated from two independent sources: Scratchpads (sample papers: Blagoderov et al. 2010b; Brake and Tschirnhaus 2010) and the SysLab tool from the Hymenoptera Online database (Taekul et al. 2010). This process should become routine practice during the ViBRANT project.

3. taXMLit

3.1. Sources:

http://wiki.tdwg.org/twiki/bin/viewfile/Literature/WebHome?rev=1;filename=taXMLit_v5-04.xsd; Weitzman and Lyal (2004)

3.2. Description

The taXMLit schema is designed to accommodate taxonomic literature. It was developed particularly in the context of Zoological and Botanical publications and should also be applicable to publications on fungi and palaeontology, although this has yet to

be tested. The schema does not take into account the kinds of data needed for viral or bacterial publications. It covers all of the components of taxonomic publications and the taxon treatments contained within them, but does not encode individual character statements, which are dealt with by other projects such as SDD.

The schema is highly atomised, permitting both recovery of publication components (e.g. taxon treatments, diagnostic keys, images, bibliographic entries, discussion paragraphs) and of data within those components, such as specimen data, biological associations, atomised taxonomic names, and nomenclatural and taxonomic acts. It can be applied to the entire text of a publication and not only formal treatments as discussed above. The richness permits full application to any legacy format so far encountered. The full taXMLit contains data elements extracted from the text that permit detailed data querying, browsing, and download; a version that does not include the respective elements and is more document-centric has also been developed ('taXMLite': http://wiki.tdwg.org/twiki/bin/viewfile/Literature/WebHome?rev=1;filename=taXMLite_v5-04.xsd). This was developed to permit preliminary mark-up and subsequent upload access through the INOTAXA interface developed for taXMLit (see below); it is not discussed further here.

Implementation of the schema in an appropriate system ('INOTAXA' – http://www.inotaxa.org has been designed for this purpose) allows the text of marked-up taxonomic publications to be fully humanly searchable. In INOTAXA users may chose to view and download data (e.g. taxonomic names, specimen data, citations, biological association data, persons' names) for use in analysis or other applications, or access taxon treatments, keys, images, or other content components as reference resources. In conjunction with the appropriate system, the schema would also facilitate static links from the text to other data sources (e.g. specimen databases on the web, ZooBank). Use of the schema for multiple taxonomic works allows these to be searched or browsed simultaneously, and permits links between different works that cover the same taxa or their synonyms. Moreover, this paves the way for uses to create virtual compilations of taxon treatments, comprising components of more than one original work, e.g. checklists, faunas, and floras. These applications require that the schema should, in the appropriate parts, use elements the same as or similar to those in schemas used by other relevant systems, and be mappable to them.

3.3. Design and development

TaXMLit and INOTAXA were conceived in 2001 in a Mellon-funded meeting focussing on the potential for combining information, literature, and research data, and funded in 2001 by the Atherton Seidell Fund. The project initially selected the *Biologia Centrali-Americana* to use in trials (57 volumes, more than 50,000 taxon treatments) with a wide coverage of animals and plants and a variety of editorial styles applied. This provided a varied base for testing the schema and also developing a called-for resource. Subsequently a number of other texts published between 1758 to 2008 and including formal taxonomic publications, catalogues and other formats have also been marked

up to provide an even stronger test of applicability of the schema. Some of these are currently accessible through the INOTAXA.org pilot (currently accessible are two papers from Zootaxa, the more recent being Pyle et al. (2008) on *Chromis*, a paper that has been used by a number of initiatives to enable comparison); others will be made available in the near future.

An initial problem was source quality. Most legacy literature is not born digital, and incorporates outdated fonts, complex terms not easily resolvable by OCR, diacritic marks and other problematic aspects. Tests against BHL content in 2009 indicated a success rate for correct recognition of the scientific name of only 14-35% (Weitzman, unpublished; Freeland, unpublished). Morse et al. (2009) examined this problem and presented some solutions to be incorporated in the ViBRANT workflow. To date, mark-up to taXMLit has been undertaken through preliminary mark-up to TEI-Lite with a systematic 'flavour' (created by Weitzman and Lyal), and subsequent parsing to taXMLit, this being been either manual or automated through use of a purpose-written script based on stylistic features and landmarks introduced in the TEI-Lite mark-up.

Within taXMLit each text paragraph in the original publication (i.e., any text component terminated by the stroke of an 'Enter' key) is captured entire and given an ElementID, which run sequentially through the text. This facilitates later reconstruction of the order of the text components. In some cases, reconstruction will require a different order than the original. For example, polytomous keys, which have the structure of a tree, can be spread throughout the text with contrasting statements at the same level (called 'lugs' by taxonomists) separated by treatments or other complex elements, but need to be reconstructed without these interruptions. Individual paragraphs are then be parsed into more or less detailed elements as required. The ElementID allows the use of an IDREF attribute (a cross-reference within the mark-up). The full set of elements within taXMLit is large, designed to accommodate the atomisation of many elements (taxonomic names, for example, are fully atomised, with a rank assigned to each component) and provide the detail required for search, browse and download of identified components. While taXMLit uses elements that cover the same concepts as those used in other schemas (e.g. ABCD and Darwin Core, designed for specimen data), the individual elements are not all exactly the same, because the data as presented in the literature may be different in format from those recovered from specimen labels, for example, and may not be as easy to interpret. However, taXMLit is designed to permit mapping to ABCD and Darwin Core.

Much taxonomic literature employs abbreviations as standard (e.g. for genus names after the first use, or author names) and descriptors may be omitted (e.g. for suprageneric hierarchical ranks, or for repeated components of label data). While this information is simple to interpret for a human reader, it is less accessible to machine processing or amenable to database storage. For this reason taXMLit uses the attribute 'Explicit' with many elements to denote whether the information included is explicitly stated or implicit and derived either by programming code or by a human in the final mark-up verification.

The use of 'Implicit' is intended as a matter of project policy to accommodate only unequivocal interpretation (e.g. the abbreviation "A." in front of a species name

where the only genus name in context is "A-us" marked as that name). While editorial practice is to limit interpretations of the text drawing on information and knowledge from outside the text itself, the schema includes an element to accommodate alternative spellings of the same name included in a single text, to capture interpreted place names and, for added geographical coordinates, using a 'source' attribute. The facility for retrieving such interpretations is being developed.

Some of the original formatting is retained (e.g. underlining, italics, bold etc), although font and line indentation, for example, are not. Page numbers are retained.

As described by Parr and Lyal (2007) a formal assessment of user needs including taxonomists and some other groups was carried out as part of the development, and part of the testing of each phase of the INOTAXA build was carried out by taxonomists and others new to the system. The elements of taXMLit, the selection of elements to index in the INOTAXA database, and the query and browse functionalities of the INOTAXA interface, were designed in concert with this user assessment.

To support querying and browsing content, search speed is maximised by storing the marked-up texts in a relational database. Fifty-seven of the fields are indexed to permits Boolean searches. To date, upload to the database has been via individual scripts, but the database has recently been simplified and made scalable, and a generic upload tool is being built.

3.4. Implementations

Use Case 1: Mark-up of 'old' taxonomic literature. Literature used for this is primarily the *Biologia Centrali-Americana* (BCA) (http://www.sil.si.edu/digitalcollections/bca), but also employed other papers including parts of Linnaeus 1758 *Systema Naturae* and Linnaeus 1752 *Species Plantarum*. The lessons learned enabled the schema to be developed to deliver the flexibility required for older literature written before more modern standardization and the advent of the nomenclatural codes.

Use case 2: Mark-up of recent taxonomic zoological and botanical taxonomic literature. Fourteen texts in different formats were marked up spanning the dates 1992-2008, including 'standard' taxonomic papers from the Coleopterist's Bulletin, Mosquito Systematics, Proceedings of the Biological Society of Washington, Systematic Botany, Transactions of the American Entomological Society and Zootaxa, part of a synonymic catalogue and a book chapter. As with older pre-Code texts, lessons learned enabled the schema to be refined to accommodate variation in modern literature, and manage multiple publications including treatments of the same taxa under the same and different names and in different systematic placements.

Use case 3: Storage of mark-up. To enable rapid search and retrieval of marked up content in a scalable manner a database with selected (high-usage) fields indexed was constructed in MySQL. This permits much more rapid access and retrieval of simple and complex queries than would be possible from storage as simple XML documents.

Use case 4: Human search and browse of content. The INOTAXA interface to content in taXMLit was built in several phases with testing of each phase primarily

by taxonomists who were new to the system. The prototype includes three publications (a BCA volume on Coleoptera: Curculionidae, a Zootaxa paper on Curculionidae taxa also included in the BCA volume, and Pyle et al. (2008) on *Chromis* fish), together including more than 800 taxon treatments (Weitzman and Lyal 2006; Lyal and Weitzman, 2008). Two additional sources of information were added: the digitised contents of Vaurie and Selander (1971) (georeferenced localities for specimens in the BCA) and a list of person names in all possible formats (e.g. Smith, Smith, J., J. Smith, J Smith etc) – this allows expressing synonymy of different name strings representing the same individual without editing / changing the original text. A link between the treatment retrieved and the treatment in the original text in PDF or JPEG format is available through the interface, as are links to any original images. Further information on INOTAXA and the queries that it permits is available at http://www.inotaxa.org.

Use case 4: Availability of content to Encyclopedia of Life. Currently marked-up text is mapped to the EoL schema and delivered to EoL with associated images for display on their pages (866 pages). The process is automatic and will deliver further pages on the next data upload to INOTAXA.

3.5. Problems encountered and lessons learned

- Interoperability. So that the schema could potentially deliver data in a format usable by other applications two choices were available: to incorporate elements of the target schemas or develop new schema elements within taXMLit that could be mapped to others. The latter was selected with the logic that taXMLit could be versioned as a stand-alone entity and updated by users as appropriate, without having to accommodate independent changes by embedded schemas.
- GUIDs. Initially unique identifiers were not explicitly included; however, as biodiversity informatics has moved towards implementation, a placeholder for GUIDs has been included in many elements.
- Accommodating multiple formats of legacy literature. Although taxonomic literature is reputedly standardized in content, experience with many different papers and books has demonstrated the extreme variability of formatting and structure applied, even within single papers. To accommodate the observed variation most of the complex elements of taXMLit are optional and available in many different places within the schema.
- Implicit content. Much content is implicit in nature (see discussion above). Care must be taken in recognizing such content, but it is necessary to do so to facilitate searching and browsing functionality in the interface, and even to retrieve some taxon treatments. Such implicit content is indicated as such in display by the use of a different font colour and annotation.
- Policy on correction of errors. Because spelling errors and other infelicities in the original publication may have nomenclatural significance, and because correction relies on individual expertise, apparent errors are not changed in the current im-

plementation of taXMLit and INOTAXA. Such change or annotation must be explicitly authored, and the ability to do this will be introduced in a later implementation.

- Mark-up. Semi-automated mark-up has been achieved using a purpose-written script, incorporating rules developed to accommodate the structure of the individual publication. Even with this, there are places where specialist knowledge is required. To facilitate this, a SpecialistReview attribute has been introduced throughout the schema.
- Recovery of original formatting. Only some of the original formatting is retained, where this aids in understanding (e.g. italicisation). INOTAXA delivers content in a standardised format to aid comprehension, but allows (subject to copyright) access to the original text.
- Hierarchies. Each publication marked up in taXMLit inevitably has an independent taxonomic hierarchy, which is displayed in INOTAXA. Where a work is produced in multiple fascicles, it is assumed unless stated otherwise that the hierarchy does not change.

Evaluation, comparison and cross-points between taxonX, TaxPub and taXMLit

The three schemas discussed above serve different purposes, but to an extent have to address the same issues. One is the identity of the communities who will use the output, and an understanding of the uses to which this output will be put. Further user needs analysis would be valuable, including building on Parr and Lyal's (2007) analysis. So far, there has been no published study that explicitly makes use of marked up literature (although the number of views of content harvested by EOL from INOTAXA, Plazi and Pensoft indicate that this product at least is valued).

One question arising from a consdideration of meeting user needs relates to the size of the data 'packages' identified by elements within the schemas, a 'package' being a logical unit of information delivery enabling reuse. Packages discussed above include the taxon treatment, collection data for a single specimen, taxon name and publication citation, among others. The schemas discussed target different sizes of packages, taXMLit opting for the largest number and smallest packages, although these are nested within larger more encompassing packages (e.g. taxon treatments). Interoperability with non-literature schemas seems to require a high degree of atomisation (Lyal and Weitzman 2008). A related issue is that while data may be extracted from a publication (such as locality data for a specimen) the relevant metadata that are given elsewhere in the text (such as confidence limits in a georeference) may not be associated.

The complexity and atomisation of the mark-up (number, size and nesting of data packages) is likely to be proportional to the cost of mark-up, which will differ between the three schemas. A cost-benefit analysis may be helpful, although would need to be in the context of the uses planned for the marked up text (see below). There are pros-

pects to reduce costs through automation of the different phases (Curry and Connor 2007; Sautter et al. 2007b; Morse et al. 2009; Cui et al. 2010a).

One of the strategic goals of biodiversity informatics is an increase in accessibility, compatibility and interoperability of data originating from different sources. If elements of taxonomic information coming from different sources are compatible (and thus can be made interoperable) they can then be easily harvested, indexed, collated, used and reused. The format of the final output of the individual schemas – in a form of XSLT stylesheets for instance – will be determined by the expectations and needs of the end users (Fig. 1).

In the context of schemas for taxonomy mark-up, compatibility is understood here as the *ability of the schemas to identify, mark-up and export elements used in both legacy and prospective taxonomic literature and needed for data mining and reuse by users*. An important criterion of compatibility is that schemas can be mapped to a shared (TDWG) vocabulary, thus allowing conversion between both literature schemas and others. Table 1 presents a rough evaluation of the schemas under consideration here with regard to a set of criteria that might prevent or facilitate generating a unified output from different taxonomic sources (and marked up with different schemas).

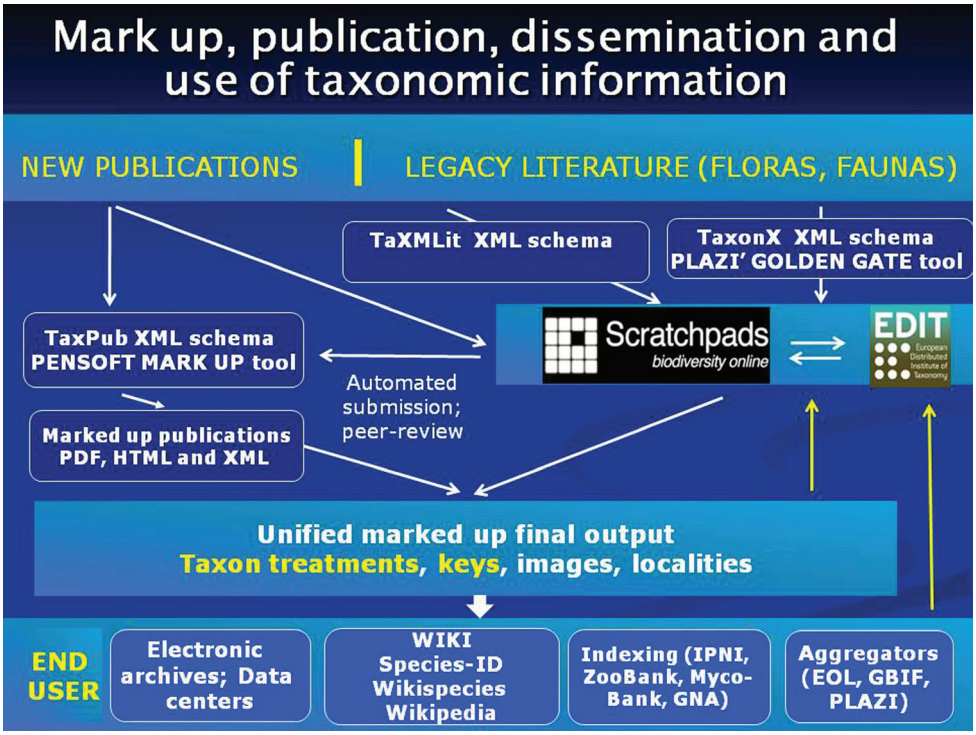


Figure 1. Flowchart of mark-up, publication, dissemination and use of taxonomic information. Scratchpads (http://scratchpads.eu/) and EDIT Cybertaxonomy Platform (http://wp5.e-taxonomy.eu/) stand for the community-based collaborative platforms for taxonomists developed by the EDIT FP6 project (http://www.e-taxonomy.eu/).

Table 1. Evaluation of the three most widely used schemas for taxonomy mark-up (taxonX, TaxPub, taXMLit) with regard to availability of key text structure elements. Legend: "-" absent; "+" present, but needs further development; ++ available. Notes in table: [1] TaxonXhas a focus on treatments; [2] taXMLIt bibliographic metadata can be mapped to MODs etc; [3] taXMLit recognises a more inclusive definition of taxon treatments and marks all the same way; [4]taXMLit marks citations, nomenclature, specimens, distributions and elements within these in detail, and identifies paragraph types; no further granularity is planned; [5]The intention is that it can be mapped; [6]Can be mapped to DC2; [7]Reference lists, in-text citations of bibliographic references, but only generic link between both.

Criteria	Taxon X	TaxPub	taXMLit
Overall structure of document captured	n/a[1]	++	++
Bibliographic metadata: uses / can be mapped to current widely adopted standards (NLM, BibTex, MARC, MODS, etc)	++	++	+[2]
Taxon name mark-up fully granular, including names, ranks and authorities	++	+	++
Nomenclatural acts: use of controlled vocabularies and normalized (standardized) tags for different acts	+	+	+
Taxon treatments (as defined in text) delimited within texts	++	++	++[3]
Internal structure of treatments – level of mark-up granularity	++	++	++[4]
Nomenclature section of treatments (names, authorities, synonyms separately tagged)	++	++	++
Species occurrence data (Localities): compliance to Darwin Core; formats for use of geographical coordinates	++	++[5]	++[6]
Reference lists, in-text citations and links between both	+[7]	++	++
Accommodates persistent identifiers (UUID, GUID, LSIDs, DOI etc.) to identify different elements (taxon names, publications, treatments, datasets, keys, phylogenetic trees etc.)	++	++	+
Permits annotation so original text and annotation both visible to user	+/-	n/a	+

TaxonX and TaxPub are largely interoperable since both have been developed by the same author and contributors, and also because both schemas have been used together in some of the cases mentioned above. The challenge will be to ensure output compatibility between taXMLit and the others, particularly with TaxonX.

The schemas themselves are only part of the necessary comparison with respect to mark-up. Given the various complexities and challenges faced in the process of retrospective mark-up, different teams are developing different protocols and editorial decisions. Some of these have been indicated above. Table 2 provides comparison of some of the critical decision areas.

Table 2. Evaluation of editorial and mark-up practices employed by main users of the three most widely used schemas (taxonX, TaxPub, taXMLit) in retrospective mark-up of historical taxonomic literature. Legend: "-" weak; "+" present, but needs further development; "++" good; "+++" very good.

Criteria	taxonX	TaxPub	taXMLit
1) What are the tolerances for text accuracy?	Generally high – structural mark-up generation is mostly independent of text, detail mark-up to some degree depends on text accuracy - text accuracy is checked during mark-up generation	n/a in prosepective publishing.	Text accuracy managed through pre-mark-up checking manually and through processes developed in ABLE project (Morse et al. 2009)
2) What are the editorial policies for, among others:			
a) corrections/retention of typos and other errors in the text	Retained if in original publication	n/a in prospective publishing	Retained If in original publication; some corrections marked as implicit if unequivocal.
b) interpretation of unclear text	What is "unclear text"? Abbreviated or omitted taxonomic epithets are disambiguated or filled in, respectively, during mark-up generation	n/a in prospective publishing	Use 'implicit' attribute for unequivocal clarification; if reliant on subjective interpretation not changed.
c) choice of "copy-text", i.e., the exemplar from which the digitized version of the text will be made. It is highly unlikely that every copy of any edition of a work will have exactly the same text	Probably not relevant to TaxonX – most documents marked are journals or journal articles, which extremely rarely have more than one edition	n/a in prospective publishing	Source copy text in large institutional libraries. Possibility for multiple copies of same work to be uploaded if marked up. Within texts treat cancels and cancellands separately.
3) What are the policies and practices for normalization and other annotation, such as:			
a) expansion of abbreviations	Abbreviations get tagged, data they imply stored in DwC children of dedicated tax:xmldata element, but not widely used as of yet	n/a in prospective publishing	Use of 'implicit' attribute for unequivocal expansions (e.g. generic names, author names)
b) normalization of taxon names, personal names, corporate names, etc.	Taxon names atomized, epithets expanded or filled in where abbreviated or missing, normalized epithets stored in DwC children of dedicated tax:xmldata element No normalization for person or corporate names as of yet	n/a in prospective publishing	Primarily reproduced as original; in some cases for both taxon names and person names, some normalization occurs in a separate part of the mark-up. Facility for linking synonyms of Parties / Agents outside text in place through INOTAXA.

Criteria	taxonX	TaxPub	taXMLit
c) modernization of archaic or changed place names (e.g., Rhodesia/Zimbabwe)	None as of yet	n/a in prospective publishing	Primarily reproduced as original; also has ability to capture an 'interpreted' place name. Facility for searching forms of changed place names being developed in INOTAXA.
d) annotation and other editorialization, as for example, correction of incorrect taxon names, assignment of coordinates to location names	Actual taxon name stays as in original publication, normalized epithets stored in DwC children of dedicated tax:xmldata element usually contain correct value	n/a in prospective publishing	Primarily reproduced as original; also has ability to capture corrections and additions as 'interpreted' data and, for added coordinates, using a 'source' attribute,
4) What are the textual objects of interest which will be encoded (i.e., do not aim to tag everything). What is in scope, and what is not? What has the highest priority?			
treatments	+++	+++	+++ (high)
keys	+	++	+++ (high)
Phylogenetic and other trees	+	+	++ (low)
Front and back matter	-	+++	+++ (medium)
Discussion paragraphs	++	+++	+++ (medium)
Names	+++	+++	+++ (high)
Specimen data	++	++	+++ (high)
Taxonomic and nomenclatural acts	+++	++	++ (medium)
bibliographies	++	+++	+++ (medium)
other			Front and Back Mattersection types, image legends, indexes
5) What are the purposes of the mark-up? One just cannot "tag everything", as no single encoding of a text is going to be equally suitable for all thinkable purposes. Three main categories can be seen as:			
a) rendition/representation of the text in HTML, PDF, ePub, or other formats	+++	+++	+++
b) archiving of the text for long term preservation	++	+++	++

Criteria	taxonX	TaxPub	taXMLit
c) analysis, data mining, and other processing	+++	++	+++
6) What are the policies and practices for the handing of non-textual features such as illustrations, inserted plates, fold-out maps, etc.?			
a) how should multi-column text be handled?	Multi-column text normalized into single column	According to NLM publishing and archiving Tag Suite.	Currently most layout elements such as this are ignored unless columns numbered separately in original, in which case each column is treated as if it were a page.
b) what are the policies and practices regarding overlapping hierarchies in the text (say, a significant section starts in one chapter and concludes in another chapter of a book)?	Not encountered so far, so no respective policy – most documents marked in TaxonX are journals or journal articles, which next to never exhibit overlapping hierarchies	n/a in prospective publishing	+ treatments are dated from the first date of publication; supplements handled separately.

Conclusions

There is a rich legacy of several hundred years of taxonomic literature. In addition to this, many new papers are published each year, driven by the recording of an estimated ca. 17,000 new taxa being described with some unknown number of re-descriptions. We can use technology to help us process this data overload, but only if we can first impose some form of structure on the data that facilitates machine-processing. Applying structure to a text is a remarkably challenging activity. This paper has considered three XML schemas devised to help address this problem.

Table 1 shows that the three currently most widely used schemas - TaxonX, TaxPub and taXMLit - cover the key text elements used by taxonomists fairly equally. This is encouraging as it suggests interoperability should be achievable among the schemas. Equally, it might lead one to ask why there should be three separate schemas.

The answer lies in Table 2. This table shows a greater range of answers to its questions, each schema with its own strengths, weaknesses and associated editorial practices. Reading this table in conjunction with the main text of this paper, we can see that each schema is focused on a different user need.

TaxonX addresses core data requirements of working taxonomists. Table 2 shows that it focuses on the three core elements required by taxonomists: treatments, names, and taxonomic and nomenclatural acts. Its use in Plazi has shown how it can successfully meet this basic user need. The focus on taxon treatments has led to some exciting developments towards making the data available as RDF triples in conjunction with GBIF or through the Species Model transfer (SPM) to the Encyclopedia of Life. This development will permit greater linking of taxonomic data across repositories.

TaxPub addresses the need to ensure that data in new publications is immediately accessible. Being specifically targeted at new literature, it can avoid many of the problems applicable only to historic literature (leading in Table 2 the frequent statement 'n/a in prospective publishing'). This focus has allowed TaxPub not only to be successfully piloted as a publication tool, but for systems using the schema to automatically populate other resources, such as Plazi and Species-ID, and the prospect of generating treatments from and uploading to databases. Hence, data in the new text are immediately available for other researchers. In addition, table 2 shows it is the schema most suited to archival use as befits a schema targeted at publication and derived from JATS (Journal Archiving and Interchange Tag Suite - http://dtd.nlm.nih.gov/).

TaXMLit provides the richest mark-up of the three schemas. Table 2 shows it handles a greater range of data and in more detail than the other schemas and, for example, is the only schema that can handle a change in geographic place name since publication. However, this richness comes at a cost, since to efficiently exploit the data within a series of taXMLit texts they are ideally converted to a searchable database (as exemplified by INOTAXA), and to create a fully marked-up text is both time consuming and requires expert input. The future development goals for taXMLit include

greater automation of mark-up, and a possible lightweight derivative taXMLite. The full taXMLit schema best serves the needs of a wide variety of researchers, and for those who wish to trawl the data as opposed to answer pre-defined questions, such as those working on the impact of climate change.

Therefore, we may conclude that having three solutions to the one problem of marking up taxonomic literature is appropriate because each schema addresses a different user need. TaxPub is the most suitable for born-digital literature, and the mark-up can be achieved at relatively low cost. There are few technical hurdles, for example, as the source material is already in digital format; and if there are ambiguities in the text they can be presented to the author(s) while they prepare their text. However, in focusing on new literature, TaxPub is not meant for handling historic texts (although there is an archival version in JATS that is designed for legacy literature, and might not only be used for mark-up but could also be submitted to PubMedCentral; this version, however, has not yet been customized for taxonomy). In contrast, both TaxonX and taXMLit can handle the issues that accompany historic texts. TaxonX focuses on taxon treatments, whereas taXMLit covers all data within a text. Hence, TaxonX is easier and cheaper to mark-up, but the results are not as widely usable as they would be had the original text been marked up in taXMLit. There is a clear need to understand the cost-benefit of marking up texts to assist users to decide which of the two schemas is more appropriate for them.

All three schemas have a role to play in ViBRANT. Both TaxonX and taXMLit could benefit from ViBRANT's investigations into the use of citizen scientists to review texts and the use of automatic tools for data mining historic literature. This aims to enhance the accuracy of the data extracted, and to reduce the cost and time required to produce the mark-up. TaxPub at the same time will allow ViBRANT to publish its content in a semantically enhanced and state of the art way that not only provides the already proven option for easy dissemination of its content as well as provide a stable archive of the valuable content created through ViBRANT's infrastructure.

This paper has discussed a means of achieving more use of the data in taxonomic literature by making that data easier to share, search, link, and combine, especially through semantic enhancement, and by exposing the data to new automated analytical techniques such as data mining. To achieve these goals, it is necessary to apply some form of structure to the literature. In the context of taxonomic literature mark-up we are fortunate to have seen the development of these three schemas to apply structure, for each addresses a particular user need. In addition, the schemas' common coverage assures us that the core data they contain can be converted from one schema to another, and so could be equally accessible to any tool-sets developed to exploit each schema. This is true now of marked up taxonomic literature and is also true of future marked up taxonomic literature, whether newly written born-digital texts or digitised historic texts. These are the tools to support our advance towards liberating the data stored in taxonomic literature or to prevent their confinement from begin with.

Acknowledgements

The current work is funded in part by the ViBRANT (Virtual Biodiversity Research and Access Network for Taxonomy, http://vbrant.eu) FP7 project. We also thank all the numerous colleagues who helped in the establishment of the schemas in one way or another, mostly by comments, testing or providing working examples.

References

ABCD – Access to Biological Collection Data – a joint CODATA and TDWG initiative. http://www.bgbm.org/TDWG/CODATA/

Agosti D, Egloff W (2009) Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2: 53. http://www.biomedcentral.com/1756-0500/2/53 doi: 10.1186/1756-0500-2-53

Agosti D, Klingenberg C, Sautter G, Johnson N, Stephenson C, Catapano T (2007) Why not let the computer save you time by reading the taxonomic papers for you? Biológico, São Paulo 69 (suplemento 2): 545–548. http://hdl.handle.net/10199/15441

Blagoderov V, Brake I, Georgiev T, Penev L, Roberts D, Rycroft S, Scott B, Agosti D, Catapano T, Smith VS (2010a) Streamlining taxonomic publication: a working example with Scratchpads and ZooKeys. ZooKeys 50: 17–28. doi: 10.3897/zookeys.50.539

Blagoderov V, Hippa H, Nel A (2010b) *Parisognoriste*, a new genus of Lygistorrhinidae (Diptera, Sciaroidea) from the Oise amber with redescription of Palaeognoriste Meunier. ZooKeys 50: 79–90. doi: 10.3897/zookeys.50.506

Brake I, von Tschirnhaus M (2010) Stomosis arachnophila sp. n., a new kleptoparasitic species of freeloader flies (Diptera, Milichiidae). ZooKeys 50: 91–96. doi: 10.3897/zookeys.50.505

Catapano T (2010) TaxPub: An extension of the NLM/NCBI Journal Publishing DTD for taxonomic descriptions. Proceedings of the Journal Article Tag Suite Conference 2010. http://www.ncbi.nlm.nih.gov/books/NBK47081/#ref2

Catapano T, Weitzman AL (2007) Progress in making literature easily accessible: schemas and marking up TaxonX / Goldengate & taXMLit / INOTAXA. TDWG Annual Meeting 2007. http://wiki.tdwg.org/twiki/pub/Literature/WebHome/Catapano_Weitzman_Markup_Final.pdf

Cui H (2008a) Converting Taxonomic Descriptions To New Digital Formats. Biodiversity Informatics 5: 20–40 https://journals.ku.edu/index.php/jbi/article/view/46/1551

Cui H (2008b) Approaches to Semantic Mark-up for Natural Heritage Literature. Proceedings of the iConference 2008. http://ischools.org/conference08/pc/PA5-2_iconf08.doc

Cui H, Heidorn PB (2007) The reusability of induced knowledge for the automatic semantic mark-up of taxonomic descriptions. Journal of the American Society for Information Science and Technology. 58(1): 133–149. http://www3.interscience.wiley.com/cgi-bin/fulltext/113466052/PDFSTART doi: 10.1002/asi.20463

Cui H, Jiang Y, Sanyal PP (2010a) From Text to RDF Triple Store: An Application for Biodiversity Literature[Demo]. Proceedings of the 73rd ASIS&T Annual Meeting v. 47. Oct 22–27, 2010. Pittsburg, PA. http://www.asis.org/asist2010/proceedings/proceedings/ASIST_AM10/submissions/415_Final_Submission.pdf

Cui H, Boufford D, Selden P (2010b) Semantic Annotation of Biosystematics Literature without Training Examples. Journal of American Society of Information Science and Technology. 61 (3): 522–542. http://harvard.academia.edu/DavidBoufford/Papers/740926/Semantic_annotation_of_biosystematics_literature_without_training_examples

Curry GB, Connor, RJ (2007) Automated extraction of biodiversity data from taxonomic descriptions. In: Curry GB, Humphries CJ (Eds) Biodiversity databases: Techniques, politics, and applications: Systematics Association Special Volume 73: Boca Raton, Florida, CRC Press, Chapter 6: 63–81.

Curry GB, Connor RCH (2008) Automated extraction of data from text using an XML parser: An earth science example using fossil descriptions. Geosphere 4(1): 159–169. doi: 10.1130/GES00140.1

Darwin Core (2008) http://rs.tdwg.org/dwc/

Hagedorn G, Thiele K, Morris R, Heidorn PB (2005) The Structured Descriptive Data (SDD) w3c-xml-schema, version 1.0. http://www.tdwg.org/standards/116/

Heidorn PB, Cui H, Yu B, Wu J, Zhang H (2002) Taxonomic description creation, search and display in XML. Abstract. Botany 2002. http://www.isrl.illinois.edu/~pheidorn/papers/Botany2002Abstract.pdf

Kirkup D, Malcolm P, Christian G, Paton A (2005) Towards a digital African Flora. Taxon 5(2): 457–466. doi: 10.2307/25065373

Lyal CHC, Weitzman L (2008) Releasing the content of taxonomic papers: solutions to access and data mining. Proceedings of the BNCOD Workshop "Biodiversity Informatics: challenges in modelling and managing biodiversity knowledge" http://biodiversity.cs.cf.ac.uk/bncod/LyalAndWeitzman.pdf

Morse D, Dil A, King D, Willis A, Roberts D, Lyal C (2009) Improving search in scanned documents: Looking for OCR mismatches. In: Bernardi R, Chambers S, Gottfried B (Eds) Proceedings of the workshop on advanced technologies for digital libraries (AT-4DL 2009): 58–61 http://www.unibz.it/en/public/universitypress/publications/all/Documents/9788860460301.pdf

Murray-Rust P, Rzepa HS (2002) Scientific publications in XML - towards a global knowledge base. Data Science 1: 84–98. doi: 10.2481/dsj.1.84

Parr CS, Lyal CHC (2007) Use cases for online taxonomic literature from taxonomists, conservationists, and others. TDWG Annual Conference, Slovakia http://www.tdwg.org/proceedings/article/view/269.

Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith V, Brake I, Ryrcroft S, Scott B, Johnson N, Morris R, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress W, Thompson C, Erwin T (2010a) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1–16. doi: 10.3897/zookeys.50.538

Penev L, Roberts D, Smith V, Agosti D, Erwin T (2010b) Taxonomy shifts up a gear: New publishing tools to accelerate biodiversity research. ZooKeys 50: 1–4. doi: 10.3897/zookeys.50.543

Penev L, Hagedorn G, Mietchen D, Georgiev T, Stoev P, Sautter G, Agosti D, Plank A, Balke M, Hendrich L, Erwin T (2011) Interlinking journal and wiki publications through joint citation: Working examples from ZooKeys and Plazi on Species-ID. ZooKeys 90: 1–12. doi: 10.3897/zookeys.90.1369

Pyle RL, Earle JL, Greene BD (2008) Five new species of the damselfish genus *Chromis* (Perciformes: Labroidei: Pomacentridae) from deep coral reefs in the tropical western Pacific. Zootaxa 1671:3–31.

Sautter G, Böhm K, Agosti D (2007a) A Quantitative Comparison of XML Schemas for Taxonomic Publications. Biodiversity Informatics 4: 1–13. https://journals.ku.edu/index.php/jbi/article/view/36

Sautter G, Agosti D, Böhm K (2007b) Semi-Automated XML Markup of Biosystematics Legacy Literature with the GoldenGATE Editor. Proceedings of PSB 2007, Wailea, HI, USA, 2007. http://psb.stanford.edu/psb-online/proceedings/psb07/sautter.pdf

Stoev P, Akkari N, Zapparoli M, Porco D, Enghoff H, Edgecombe GD, Georgiev T, Penev L (2010) The centipede genus *Eupolybothrus* Verhoeff, 1907 (Chilopoda: Lithobiomorpha: Lithobiidae) in North Africa, a cybertaxonomic revision, with a key to all species in the genus and the first use of DNA barcoding for the group. ZooKeys 50: 29–77. doi: 10.3897/zookeys.50.504

Taekul C, Johnson NF, Masner L, Polaszek A, Rajmohana K (2010) World species of the genus Platyscelio Kieffer (Hymenoptera, Platygastridae). ZooKeys 50: 97–126. doi: 10.3897/zookeys.50.485

TDWG (2007 onwards) TDWG: standards. Biodiversity Information Standards. http://www.tdwg.org/standards/

Weitzman AL, Lyal CHC (2004) An XML schema for taxonomic literature – taXMLit - http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf

Weitzman AL, Lyal CHC (2006) INOTAXA — INtegrated Open TAXonomic Access and the "*Biologia Centrali-Americana*". Proceedings Of The Contributed Papers Sessions Biomedical And Life Sciences Division, SLA. 8pp. http://units.sla.org/division/dbio/Baltimore/index.html

Willis A, King D, Morse D, Dil A, Lyal C, Roberts D (2010) From XML to XML: The Why and How of Making the Biodiversity Literature Accessible to Researchers. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/787_Paper.pdf

ZooKeys 150: 117–126 (2011)
doi: 10.3897/zookeys.150.2109
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool

Steven Bachman[1], Justin Moat[1], Andrew W. Hill[2], Javier de la Torre[2], Ben Scott[3]

1 *Royal Botanic Gardens, Kew, UK* **2** *Vizzuality, Spain* **3** *Natural History Museum, UK*

Corresponding author: *Steven Bachman* (s.bachman@kew.org)

Academic editor: *V. Smith* | Received 16 September 2011 | Accepted 18 November 2011 | Published 28 November 2011

Citation: Bachman S, Moat J, Hill AW, de la Torre J, Scott B (2011) Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 117–126. doi: 10.3897/zookeys.150.2109

Abstract

GeoCAT is an open source, browser based tool that performs rapid geospatial analysis to ease the process of Red Listing taxa. Developed to utilise spatially referenced primary occurrence data, the analysis focuses on two aspects of the geographic range of a taxon: the extent of occurrence (EOO) and the area of occupancy (AOO). These metrics form part of the IUCN Red List categories and criteria and have often proved challenging to obtain in an accurate, consistent and repeatable way. Within a familiar Google Maps environment, GeoCAT users can quickly and easily combine data from multiple sources such as GBIF, Flickr and Scratchpads as well as user generated occurrence data. Analysis is done with the click of a button and is visualised instantly, providing an indication of the Red List threat rating, subject to meeting the full requirements of the criteria. Outputs including the results, data and parameters used for analysis are stored in a GeoCAT file that can be easily reloaded or shared with collaborators. GeoCAT is a first step toward automating the data handling process of Red List assessing and provides a valuable hub from which further developments and enhancements can be spawned.

Keywords

Red List, Conservation, Open Source, Biodiversity, Mapping, IUCN, GBIF, Flickr, Geospatial, Google maps, HTML5, JSON, AJAX

Introduction

Recent estimates suggest there could be 8.7 million (± 1.3 million) species on the planet (Mora et al. 2011). Even at the lowest estimate, less than 1% (61,914, IUCN 2011)

of those species have been formally assessed using the Red List system to determine their conservation status i.e. an assessment of the risk that they will become extinct. A key factor in the lack of progress in the production of species conservation assessments is the scarcity of user friendly, but powerful, analytical tools which are readily available to scientists and communities to carry out these assessments. Furthermore, large amounts of primary biodiversity data are now available via services such as the Global Biodiversity Information Facility (GBIF), but have yet to be fully utilised for conservation action. With the trend in biodiversity loss increasing across the globe (Secretariat of the Convention on Biological Diversity 2010) it is essential that we speed up the production of assessments. This will enable us to more quickly identify species and regions at greatest risk so that it may guide conservation action. To scale up the production of conservation assessments to the level of mega-diverse groups such as plants and insects, there needs to be significant progress in the development of automated and semi-automated techniques that scientists and other experts can harness. Here, we present the Geospatial Conservation Assessment Tool (GeoCAT - http://geocat.kew.org), which is a first step towards that goal.

Methods

Analysis using GeoCAT

GeoCAT can be accessed from the following URL: http://geocat.kew.org/. The tool was developed to utilise spatially referenced primary occurrence data to analyse two aspects of the geographic range of a taxon: the extent of occurrence (EOO) and the area of occupancy (AOO). These two measures are the foundation of the 'B' criterion of the IUCN Red List system (IUCN 2001) - see '*Technology and algorithms*' section below for full definition of EOO and AOO. Figure 1 illustrates how GeoCAT users

Figure 1. GeoCAT workflow; Start a new project and add data to the map via the three options. Existing data may be derived from an output of an existing database or from an online source such as GBIF, Flickr or Scratchpads. Alternatively, click directly on the map to create markers to signify the occurrence of the taxon you wish to assess.

The intuitive mapping interface allows interaction with the data to delete, move or hide points from analysis. The metadata window exposes the attributes of the occurrences e.g. date of collection, collector, location and provides a direct link to the raw data.

After editing the data the analysis can be enabled and the results are displayed as grpahics on the map and through a report window. The EOO/AOO values, preliminary IUCN categories and parameters are shown. AOO cell size can be adjusted.

Statistics generated from the analysis and a basic map can be downloaded as a report. Occurrence data used in the analysis can be downloaded as a kml file for integration with Google Earth or as a CSV file. In addition, a single geocat. file encompassing all analysis results, parameters, map settings and occurrence data can be saved for later use, or to pass to collaborators for additional work.


Start new project
Add data
Import occurrence data from an online source e.g. GBIF, Flickr, Scratchpads
Upload from existing data
Manually add occurrences to the map
GBIF
www.gbif.org
Scratchpads
biodiversity online
flickr
Review/Edit
Delete, hide, edit (move points)
Examine metadata
Analyse
Generate EOO/ AOO values
Adjust analysis as necessary e.g. AOO cell size
EOO/AOO analysis
Extent of Occurrence
59,063.78 km2
LC
Area of Occupancy
52.00 km2
EN
AOO based on user defined cell width (2 km), change
Print complete report
Save/Export
Download summary report
Download data and report parameters (geocat, csv, kml)
kml
(Google Earth)
CSV
geocat

can quickly and easily add, review, edit and analyse data and finally save and export the results.

Technology and algorithms

GeoCAT is built using the latest web-technologies based in JavaScript and HTML5. The result is a responsive and intuitive environment for web-based GIS and conservation analysis algorithms. The tool was built to combine private data provided by the user, public resources such as Flickr, and scientific resources such as GBIF. GeoCAT makes importing geospatial species data simple, by either searching and loading data from the online sources or importing and mapping CSV files. The Google Maps API and the custom user interface provide a high quality map environment to perform geographic analysis of data location and its quality; the user can delete or move data individually or through filters (e.g. drawing bounding boxes) also defining thresholds for common components of the data such as coordinate precision. Algorithms for measuring species threat are implemented directly in the browser, avoiding any need to move data to desktop applications or to send the data for server-side processing. The GeoCAT file format streamlines the process of restarting a project by encoding all data, including algorithm parameters, outputs, and application state, into a web syntax called JSON. The file can then be stored by a user for sharing or later use.

The inclusion of external data from GBIF and Flickr was an important feature for bringing a robust species assessment tool to the web. To achieve this functionality, GeoCAT relies on cross domain AJAX requests, where the application in the user's browser directly queries, receives, and parses data from the external sources. Therefore the application relies heavily on consistent data standards, where the data received will be in a predictable format. For example, GBIF provides a REST API where data can be queried and downloaded, the data standards are encoded in their web-services documentation, http://data.gbif.org/ws/. Georeferenced images from Flickr are queried using machine tags and a keyword search.

GeoCAT presently uses two algorithms to calculate EOO and the AOO (after Willis et al. 2003). These were originally developed in the Avenue scripting language for ArcView 3.3 within the Conservation Assessment Tools (CAT) extension (developed at the Royal Botanic Gardens, Kew and downloadable from: http:// www.kew.org/gis/projects/cats (Moat 2007)), these were reprogrammed in JavaScript for GeoCAT.

EOO is a measure of the geographic range size of a species. One of the simplest methods to calculate this is a convex hull which is defined as the smallest polygon in which no internal angle exceeds 180˚ and contains all sites of occurrence (see Figure 2). There are many algorthims developed to calculate the convex hull from a set of points, but within GeoCAT we use a quickhull (Eddy 1977 and Bykat 1978) with code developed from Echo 2 (http://blogs.infoecho.net/echo/2007/03/) and Eriestuff (http://eriestuff.blogspot.com/2008/03/google-maps-convex-hull-of-point-set-or.html)

AOO is a measure of the area in which a species occurs. One of the more straightforward ways of measuring this is to sum the area of square grids the species occupies.

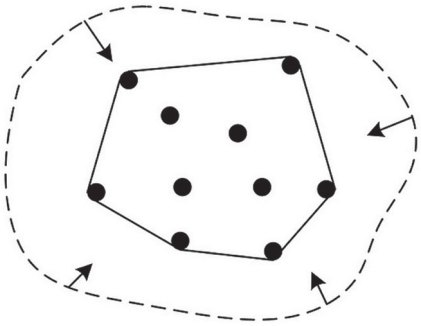


Figure 2. Illustration of a convex hull of a set of points. Imagine stretching a rubber band so that all points are inside it, then releasing it; when it becomes tight, the area enclosed is the convex hull.

There is much discussion on the influence of scale and what cell size is appropriate (Kunin and Hartley 2003, Willis et al. 2003, Callmander et al. 2007). IUCN states that: "the most appropriate scale will depend on the taxon in question, and the origin and comprehensiveness of the distribution data" (IUCN Standards and Petitions Subcommittee 2010). Within GeoCAT we have allowed the user to choose the cell size using three methods. The default is 2km^2 cell size (as recommended in the IUCN guidelines - IUCN 2010), user defined cell size and finally 1/10th of the maximum distance between the most distance pair of points (Willis et al. 2003). The last method uses a factor of 10 as this reflects the relationship between EOO and AOO in the IUCN criteria and gives a size of the grid reflecting the geographic scale of the species distribution. Cells are calculated using simple maths to degrade each point to the lower left corner of the cell ((Floor ((x or y)/cellwidth) * cellwidth)), cells are then constructed from this lower left corner. In addition, the number of points within the cell are recorded and used to colour the cell on the map to give an indication of density of collections.

Open source

GeoCAT was developed as an open source tool. This means that the methods and contributions of the code itself can help inform the informatics community in the future. Open source also aids in the transparency of decision making, by allowing anyone to see and audit algorithms. The code is accessible to anyone from the project's Github repository (https://github.com/Vizzuality/GeoCAT). We hope that this will help lower the cost of attracting new algorithms and community developed solutions with the tool.

Scratchpad integration

GeoCAT relies on primary occurrence data to drive the analysis. One of the major new platforms for primary biodiversity data is the Scratchpads project (Smith et al. 2009). With 281 sites across a broad spectrum of natural history science (including the lesser known groups where their conservation status is poorly known) and thousands of primary data records, the Scratchpads project is an obvious choice for integration with GeoCAT. Scratchpad users will be able to access specimen or occurrence data directly from GeoCAT. Similar to the GBIF and Flickr 'source data' options it is possible to

query data from a specific Scratchpad site to directly access and plot specimen data for analysis. In addition, from within a Scratchpad site, users will be able to open GeoCAT directly from a Scratchpad page where structured specimen or occurrence data exists. This will instantly display the data, assuming it contains georeferenced records. Finally, users will be able to upload a .geocat report file to a Scratchpad page. From here a URL link can take the user back to GeoCAT site where further analysis can be performed. In summary:

- GeoCAT will be able to access and import Scratchpad specimens via a new web service to output structured data in Darwin Core format.
- GeoCAT can be opened directly from within a scratchpad page, using a link encoded with URL parameters to retrieve the structured data source.
- Users should be able to upload a .geocat file report in a scratchpad page and the page will offer the option to open it in GeoCAT.

Other systems containing large amounts of primary data such as BRAHMS (http://dps.plants.ox.ac.uk/bol/BRAHMS/Home/Index) have also integrated with GeoCAT by supporting the export of a compatible CSV file.

Caveats

It is intended that the tool is utilised primarily by those wishing to carry out Red List conservation assessments, although it also functions well as a simple web mapping tool for other uses such as georeference checking. It is expected that the user has a good understanding of the taxa being assessed, the quality of the underlying data and a good knowledge of the Red List criteria. GeoCAT can provide metrics that partially fulfil criterion B assessments and allow a preliminary rating to be obtained. In order to complete a full Red List assessment a number of additional sub-criteria must be met and a minimum set of data are required to accompany the assessment. For further information see the IUCN Red List technical documents: (http://www.iucnredlist.org/technical-documents/data-organization).

It is not within the scope of this paper to discuss the use of EOO and AOO for Red List assessments as this has been considered elsewhere (see IUCN Standards and Petitions Subcommitte 2010 and references therein). It is hoped that complementary algorithms such as Alpha hulls (α-hulls - generalisations of convex hulls) can be incorporated into later versions of GeoCAT to provide the user with a wider range of options for a more robust analysis. The use of α-hulls may be a more appropriate method for investigating reductions or continuing declines in EOO (IUCN 2010).

At present there are some limitations on number of occurrence records that can be displayed from both GBIF (500) and Flickr (250). Users will be informed if their query returns more points than the display limit. Initial performance tests suggest the map display can handle many thousands of points, but further testing is needed. The

records displayed are the first to be queried, but it is hoped that later versions of GeoCAT will provide further refinements to this query e.g. most recently collected records first or those with highest georeference precision.

Conclusion

Future directions

It is anticipated that significant improvements will be made for later versions of GeoCAT. An obvious shortcoming is that the tool only deals with one aspect of the IUCN criteria and can only report a preliminary assessment based on the two range measures extent of occurrence (EOO) and area of occupancy (AOO). An obvious extension is to incorporate additional range based analysis that can inform other aspects of the IUCN criteria such as number of locations, sub-populations and degree of fragmentation.

Although the tool is geospatial in its focus, there is the opportunity to extend analysis into the temporal elements of museum specimen data, such as the date of collection. Statistical approaches have already been investigated (Solow and Roberts 2003, McPherson and Myer 2009, Collen et al. 2010) and could simply be modified for inclusion in GeoCAT as an additional module. Examining occurrence data through time could open up other parts of the Red List criteria such as Criterion A that deals with 'reduction' or decline in population size. For example within GeoCAT historical specimens can be removed when they occur in areas known to have been subject to recent habitat loss. Reductions in EOO and AOO can then be recorded and potentially applied to Criterion A.

At present assessments can only be carried out one at a time. In order to scale up the production of assessments a batch option is needed whereby a single file of occurrence data for multiple species can be uploaded and processed. This would allow hundreds of assessments to be processed in a matter of seconds.

Further enhancements can also be made with regard to the handling of point data through the GBIF portal. The added bonus of the slick mapping interface makes GeoCAT a useful tool for georeference checking and cleaning. Querying the raw data from GBIF can often reveal obvious georeferencing mistakes such as outliers or swapped latitude and longitude pairs. The easy click and drag editing of points means they can be accurately placed on the map to ensure the most precise analysis. GeoCAT allows you to track which points have been edited, but at present there is no easy mechanism for feeding back this information to the original data provider – this could be a service integrated into the GBIF portal. Until this feedback loop is established the erroneously georeferenced records from data providers will continue to be served up by GBIF.

Harvesting of GBIF data also provides an opportunity to put the occurrence data of your target species in the context of the background collecting rate in a region.

Presence of your target species i.e. the one you wish to assess is easy to determine with a verified record, but absence is more difficult. GBIF data can be used to determine a background collecting rate for your target group e.g. plants. Absence of your target species in an area with a high intensity of background sampling provides evidence that your target species may be absent.

An exciting potential extension of GeoCAT is to provide better integration with cloud based data such as Google Fusion tables. This could work in three ways: i) linking GeoCAT to specimen data stored in the cloud thereby allowing on-the-fly editing ii) exporting assessment results to tables in the cloud and iii) linking to custom layer data in kml/kmz format. This could lead to the first entirely automated cloud based conservation assessments.

The functionality of adding user generated kml/kmz files also offers significant potential. Threat datasets from fires to land cover change and deforestation can be added. At present the layer files can be visualised, but it is not possible to interact with the layers via spatial queries in the same was as a GIS. Adding this kind of functionality would instantly allow more rigorous data driven assessments.

Benefits

GeoCAT provides a mechanism for data driven conservation assessments in a transparent, repeatable and rapid way through a user friendly environment. The benefits can be summarised as the following:

- Data driven assessments, giving an auditable data trail i.e. complete transparency of data used for assessments
- A simple, modern and easy to use interface.
- Accessible - opening up to assessors across the world - only an Internet connection is needed.
- Standardised, automated and repeatable analysis.
- Single-click analysis of Extent of Occurrence (EOO) and Area of Occupancy (AOO)
- Ability to import occurrence data from online sources such as GBIF or Flickr and other systems such as Brahms and Scratchpads. GeoCAT also allows export and reporting to other formats for further analysis or storage.
- Quick to use and easy to distribute data which can only accelerate the production of Red List assessments.
- Code is open source and development of algorithms are encouraged so the tool can develop towards a powerful automated assessment tool and for other geographic analysis.

GeoCAT responds directly to the growing need for more data driven analytical tools to aid the process of assessing species against the Red List criteria. The tool is

intended to be a platform from which enhancements can be made and we encourage the developer community to engage with the GeoCAT project. We believe there are many exciting possibilities for the future development of GeoCAT. We hope GeoCAT can be utilised for the assessment of taxa at any spatial scale and across any taxonomic group, but especially those that are poorly represented on the Red List at present.

If you wish to acknowledge use of GeoCAT please use the following citation:

Bachman S, Moat J, Hill AW, de la Torre J, Scott B (2011) Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 117–126. (version XX).

References

Bykat A (1978) Convex hull of a finite set of points in two dimensions. Information Processing Letters. 7: 296–298. doi: 10.1016/0020-0190(78)90021-2

Callmander MW, Schatz GE, Lowry PP II, Laivao MO, Raharimampionona J, Andriambololonera S, Raminosoa T, Consiglio T (2007) Identification of priority areas for plant conservation in Madagascar using Red List criteria: rare and threatened Pandanaceae indicate sites in need of protection. Oryx 41, 2: 168–176. doi: 10.1017/S0030605307001731

Collen B, Purvis A, Mace G (2010) When is a species really extinct? Testing extinction inference from a sighting record to inform conservation assessment. Diversity and Distributions 16: 755–764. doi: 10.1111/j.1472-4642.2010.00689.x

Eddy W (1977) A new convex hull algorithm for planar sets. ACM Transactions on Mathematical Software 3: 398–403. doi: 10.1145/355759.355766

Hartley S, Kunin WE (2003) Scale dependence of rarity, extinction risk, and conservation priority. Conservation Biology 17: 1559–1570. doi: 10.1111/j.1523-1739.2003.00015.x

IUCN (2001) IUCN Red List Categories and Criteria: Version 3.1. IUCN Species Survival Commission. IUCN, Gland, Switzerland and Cambridge, UK: 30 pp.

IUCN (2011) IUCN Red List of Threatened Species. Version 2011.2 Available here: http://www.iucnredlist.org/

IUCN Standards and Petitions Subcommitte (2010) Guidelines for using the IUCN Red List Categories and Criteria. Version 8.1 (August 2010) Prepared by the Standards and Petitions Subcommittee in March 2010. http://intranet.iucn.org/webfiles/doc/SSC/RedList/RedListGuidelines.pdf.

McPherson JM, Myer RA (2009) How to infer population trends in sparse data: examples with opportunistic sighting records for great white shark. Diversity and Distributions 15: 880–890. doi: 10.1111/j.1472-4642.2009.00596.x

Moat J (2007) Conservation Assessment Tools Extension for ArcView 3.x, Version 1.0. GIS Unit. Royal Botanic Gardens, Kew. http:// www.kew.org/gis/projects/cats

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How Many Species Are There on Earth and in the Ocean? PLoS Biol 9(8):e1001127. doi: 10.1371/journal.pbio.1001127

Secretariat of the Convention on Biological Diversity (2010) Global Biodiversity Outlook 3. Montréal, 94 pp.

Smith VS, Rycroft SD, Harman KT, Scott B, Roberts D (2009) Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. BMC Bioinformatics 10(Suppl 14):S6 doi: 10.1186/1471-2105-10-S14-S6

Solow AR, Roberts DL (2003) A nonparametric test for extinction based on a sighting record. Ecology 84: 1329–1332. doi: 10.1890/0012-9658(2003)084[1329:ANTFEB]2.0.CO;2

Willis F, Moat J, Paton A (2003) Defining a role for herbarium data in Red List assessments: a case study of Plectranthus from eastern and southern tropical Africa. Biodiversity and Conservation 12: 1537–1552. doi: 10.1023/A:1023679329093

ZooKeys 150: 127–149 (2011)
doi: 10.3897/zookeys.150.2189
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information

Gregor Hagedorn[1], Daniel Mietchen[2], Robert A. Morris[3], Donat Agosti[4], Lyubomir Penev[5], Walter G. Berendsohn[6], Donald Hobern[7]

1 *Julius Kühn-Institute, Federal Research Centre for Cultivated Plants, Königin-Luise-Str. 19, 14195 Berlin, Germany* **2** *EvoMRI Communications, Zwätzengasse 10, 07743 Jena, Germany; Open Knowledge Foundation Deutschland, Prenzlauer Allee 217, 10405 Berlin; Germany, and Pensoft Publishers, 13a Geo Milev Str., Sofia, Bulgaria* **3** *Harvard University Herbaria and University of Massachusetts at Boston* **4** *Plazi, Zinggstr. 16, 3007 Bern, Switzerland* **5** *Bulgarian Academy of Sciences & Pensoft Publishers, Sofia, Bulgaria* **6** *Botanischer Garten und Botanisches Museum, Freie Universität Berlin, Königin-Luise-Straße 6-8, 14195 Berlin, Germany* **7** *Atlas of Living Australia, CSIRO Ecosystem Sciences, GPO Box 1700, Canberra, ACT 2601, Australia*

Corresponding author: *Gregor Hagedorn* (g.m.hagedorn@gmail.com)

Academic editor: *V. Smith* | Received 3 October 2011 | Accepted 22 November 2011 | Published 28 November 2011

Citation: Hagedorn G, Mietchen D, Morris RA, Agosti D, Penev L, Berendsohn WG, Hobern D (2011) Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 127–149. doi: 10.3897/zookeys.150.2189

Abstract

The Creative Commons (CC) licenses are a suite of copyright-based licenses defining terms for the distribution and re-use of creative works. CC provides licenses for different use cases and includes open content licenses such as the Attribution license (CC BY, used by many Open Access scientific publishers) and the Attribution Share Alike license (CC BY-SA, used by Wikipedia, for example). However, the license suite also contains non-free and non-open licenses like those containing a "non-commercial" (NC) condition. Although many people identify "non-commercial" with "non-profit", detailed analysis reveals that significant differences exist and that the license may impose some unexpected re-use limitations on works thus licensed. After providing background information on the concepts of Creative Commons licenses in general, this contribution focuses on the NC condition, its advantages, disadvantages and appropriate scope. Specifically, it contributes material towards a risk analysis for potential re-users of NC-licensed works.

Keywords

Creative Commons, Open Access, Open Content, Licensing, Non-profit, Open Educational Resources, Data Sharing, Software Licenses, Europeana

Copyright, Science, and Education

Copyright is a state-guaranteed right given to creators of "literary and artistic works" (see Art. 2 (1) of the Berne Convention, World Intellectual Property Organisation 1979) to control the reproduction, distribution, adaptation or translation of their works. The term "work" includes a wide array of forms of intellectual creations, including text, photographs, diagrams, maps, movies, etc. To be eligible for copyright, a work must be original, individual, singular and new (see, e.g., Agosti and Egloff 2009).

Copyright typically lasts 50 to 70 years after the death of the last contributor of a work. Under the standard term in the European Union ("life plus 70 years") even very old works may require individual negotiations with the rights owners before they can be made accessible digitally or parts of them re-used in a new context. If contractual arrangements between contributors of a work and the death year of at least one potential rights owner are unknown, only works before perhaps 1871 can reasonably be assumed to be in the public domain.

The well-known Biodiversity Heritage Library (BHL, biodiversitylibrary.org), providing access to the published knowledge about the species of world, has somewhat more favorable conditions. Operating under U.S.A. legislation, most works published before 1923 as well as a significant number of works published between 1923 and 1978 (see Hirtle 2011) are no longer under copyright control. This makes the public domain in the U.S.A. exceptionally rich. After 1978, however, copyright duration has largely been increased to life plus 70 years. A continuing exception is, e.g., that works created by an officer or employee of the United States Government – including, e.g., the Food and Drug Administration and the National Institutes of Health – as part of that person's official duties are in the public domain, irrespective of publication year. However, the success of BHL cannot be simply projected into the future.

For singular cultural works such as poems, novels, paintings, or musical compositions, and where a reproduction concerns major parts of the work, the balance offered by copyright law between the rights of creators and the rights of the public for creativity and innovation is widely considered reasonable. However, the balance may already be questionable when it comes to the creative or even unavoidable (background music or company logos visible in documentary movies) inclusions of fragments of copyrighted works. Increasing IPR management and risk avoidance by companies may create a stifling and suppressive environment (Aoki et al. 2006). In some jurisdictions exceptions for educational activities (Nabhan 2009; Seng 2009; Xalabarder 2009) or fair-use doctrines allow a limited re-use of small excerpts of copyrighted materials (see, e.g., the English Wikipedia). In many other jurisdictions this is, however, not permitted (see, e.g., the German or Japanese Wikipedias), severely limiting non-commercial efforts to provide educational materials.

Ideas, knowledge, inventions, information, or data are intentionally not copyright-protected (see, e.g., World Intellectual Property Organisation 1979). The public interest, e.g., to talk about "$e = mc^2$" prior to 2025 (i.e. 70 years after the death of Einstein) is considered to outweigh the interest of scientists to be rewarded for their work. For the most parts, scien-

tists have traditionally been satisfied with the moral right to be cited for their original work. In addition, technical inventions may also be protected by patents (with a much briefer protection period of mostly 20 years and explicit provisions for knowledge dissemination).

Unfortunately, in science and education, knowledge and data are often intermingled with copyrighted expressions of the same. Many publishers establish barriers to knowledge sharing by asserting copyright on non-copyrightable plain or formal expressions of that knowledge. In the area of biodiversity, often dealing with textually expressed data or data expressed in images, this is a major obstacle. Attempts are under way to establish special procedures to extract and disseminate the – non-copyrightable – knowledge that is included within protected works. One such procedure for textual expressions, operating under principles of Swiss law, is described in Agosti and Egloff (2009). Even more difficult is the situation for drawings, photographs or diagrams. Documentary photographs or drawings are widely granted full copyright status even if they depict entirely factual information that might not itself be copyrightable. In the case of biodiversity knowledge, where much data and knowledge is expressed in these documentary forms, this can be a substantial barrier to knowledge dissemination. Finally, copyright on diagrams may severely diminish the ability of teachers and educators to disseminate knowledge efficiently. For example, the results of publicly funded research on climate change may be published in a scientific journal that does not allow educators or teachers to re-use graphs or other materials for their teaching purposes.

Such problems could best be reduced by implementing appropriate restrictions to copyright protection or by establishing legal licenses for the use of works that serve primarily as expressions of knowledge (including diagrams or documentary photography) for educational or research purposes. Unfortunately, such restrictions can only be introduced by legislative acts and depend therefore on political bargaining.

Open

Fortunately, many individuals and organizations in the scientific field, while legally entitled to complete and century-long copyright control, see advantages in less restrictive terms-of-use. In contrast to cultural works, the primary intent of scientific works is most often the dissemination of knowledge. Increasing this dissemination may thus be a principal goal. Alternatively, it may be seen a secondary goal because it improves the researchers' reputation and the brand recognition of their institutions or research areas – factors that influence future chances of obtaining research funds.

Furthermore, society often funds research to foster innovation and the general welfare, or to address problems of critical societal interest such as climate change. The vast majority of costs for such research are paid up-front by research grants or institutional funds. Trying to generate relatively minute additional income by preventing public access to the results of publicly funded research is considered inappropriate by many.

One approach to increasing dissemination and public access is found in the Open Access (OA) movement, which aims to provide readers with free and unrestricted access to

the scientific literature (Suber 2004–2010). Publications freely available to the reader are more frequently cited than publications behind a pay wall (Swan 2010). Furthermore, a growing number of funding agencies (see, e.g., Wellcome Trust 2003–2011, or the OA pilot of the Seventh Research Framework Programme, European Commission 2011), as well as the public, are asking for free and open access to the results of publicly-funded research.

The original Bethesda Open Access declaration (Brown et al. 2003) includes the right to create and distribute derivative works. Some OA journals, however, do not grant such rights. Parts of the research community, e.g. some authors choosing an OA journal, consider read-access sufficient. Based on anecdotal evidence, this may in part be because they erroneously believe that permissions for the most typical re-use scenarios (e.g., disseminating lecture materials that include other researcher's published graphs or images on a conference website) are already granted. The majority of Open Access journals today, however, provide licenses that do indeed confer broad re-use rights. This gives the journals significant advantages in dissemination and broadens the researchers' legal re-use options. It greatly simplifies the ways in which future works may build on existing ones (see, e.g., Bourne et al. 2008; MacCallum 2007).

Closely related to the Open Access movement is the Open Educational Resources movement (OER, see, e.g., Wilson-Strydom 2009; Butcher 2011). Clearly, re-use and adaptation rights are vital for educational resources (Keller and Mossink 2008).

Individuals or organizations desiring to grant Open Access and defined community re-use of their creative works can do so by creating their own individual terms-of-use. For the copyright owner and license giver (licensor), this may often appear to be the simplest and safest way to proceed. For re-users (licensees), however, legal advice may be required to assess whether individual terms-of-use permit their intended use in their own jurisdiction. Not doing so carries the risk of unwitting copyright violations that may result in legal action. The administrative and legal cost of handling a large diversity of such terms-of-use can become an impediment to re-use. At the same time, there is a substantial risk to the copyright holder that self-created terms-of-use licenses may have undesired outcomes, particularly when interpreted under multiple jurisdictions on a global scale.

Creative Commons (abbreviated CC) licenses have been created to address these problems. These licenses provide standardized terms-of-use definitions. Amongst the ca. 2.2 million articles deposited in PubMed Central, already about 10% are CC-licensed (U.S. National Library of Medicine 2011b; Björk et al. 2010).

Creative Commons

Creative Commons is a US-based non-profit organization that authors, reviews, and publishes a suite of licenses defining standard options for the distribution and re-use of creative, copyrightable works. Together with its worldwide affiliate organizations and partners, it provides standardized and scrutinized license texts, translation into languages and adaptations for various jurisdictions

Creative Commons Logo, CC BY, Creative Commons

(presently over 50 adaptations are provided). Creative Commons is never a party in the contract of such a license (other than in connection with materials for which it holds copyright itself).

The roots of the organization go back several years before the actual foundation. In 1999 Lawrence Lessig argued that the balance between public and private interests, and between the free flow of expressions of ideas and knowledge and state-guaranteed control and monopolies must at all times be carefully crafted, in the interest of both the society and the economy of a country, and that present tendencies favor monopolies and control too much (Lessig 1999). Subsequently, Creative Commons was founded in 2001 by Lawrence Lessig, Hal Abelson, and Eric Eldred, with the first licenses issued in 2002 (Creative Commons 2011a). After years of continuous growth, an important milestone was the migration of the Wikipedia and other Wikimedia Foundation projects to use CC BY-SA in 2009.

CC licenses have since been upheld by courts in several cases (District Court of Amsterdam 2006, Badajoz Sixth Court 2006, Tribunal de première instance de Nivelles 2009; Landgericht Berlin 2010). Using standardized licenses thus affords both licensor and licensee a certain degree of legal safety. Furthermore, many individuals and organizations world-wide are now using these licenses. This provides significant efficiency advantages in license management when creating works that are partially based on or integrate previous works (e.g., re-use of illustrations).

Creative Commons realizes that no single license is adequate for all purposes and provides a set of licenses to cover a wide range of use cases.

A special case is the "no rights reserved" or "rights-release" license, CC0 or CCZero. Not all jurisdictions have the concept of a public domain and in some jurisdictions specific rights cannot be released. CC0, as a license, describes a degree of freedom that is as close as possible to the concept of a public domain. The CC0 license is useful when the re-use of works shall be made as easy as possible. Some copyright owners release valuable works in this way, but the majority of applications are works where the eligibility for copyright is doubtful anyways (data containing occasional free-form text comments, simple graphics, icons, etc.).

Logo of the CC Zero or CC0 Public Domain Dedication License – "No Rights Reserved" (CC BY, Creative Commons)

All other current CC licenses are combinations of four conditions, each represented by a concise summary plus either specific clauses or modifications in the full legal code.

All current CC licenses (except CC0) contain the "Attribution" condition (abbreviated "BY"), requiring appropriate attribution of the creators of a work. In the case of derived works, this condition also requires clear statements or methods to enable the user to understand the kind of changes made during a derivation. The Attribution condition is similar to the scientific community standard that all information sources must be appropriately cited. The community standard,

Icon of the Attribution (= BY) condition (CC BY, Creative Commons)

however, refers to the attribution of ideas or data, whereas the CC BY license refers to the creative form.

In biodiversity, the attribution rule is widely accepted. An exception is the case of data or metadata publishing where it is controversial whether a legally binding attribution requirement may be an obstacle to re-use. For example, the Europeana Foundation recently changed their Data Access policy for textual metadata describing multimedia objects published through the Europeana portal from something resembling CC BY-NC to CC0 (Europeana Foundation 2011). Those favoring CC0 or Open Data Licenses see problems with any legal approach to enforce norms of Attribution, Share Alike, or other terms on data or metadata (Science Commons 2011). For example, data aggregation and inheritance may lead to unmanageable attribution and licensing stacks. Furthermore, the borderline between non-copyrightable factual information and copyrightable works (e.g., sufficiently originally creative prose inside a database, or the information model as a whole) is blurred. Conversely, a license requiring attribution greatly increases the willingness of individuals and organizations to release textual data that are research results and should be properly attributed. Without it, large volumes of valuable data may remain unpublished.

The "Share Alike" condition (abbreviated "SA") allows the distribution of derivative works, but requires that all such works must also be shared under the same conditions: "If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one." The effect is that licenses with the SA condition "spread" into derivative works; the biological metaphor of a "viral license" is often used here.

Icon of the "Share Alike" (= SA) condition (CC BY, Creative Commons)

In contrast to "Share Alike", a creator may prevent the distribution of derivatives of a work by adding the "No Derivative Works" condition: "You may not alter, transform, or build upon this work" (abbreviated "ND"). This condition does not prevent simple format changes: It is the creative work that is protected, not a particular digital representation. The license clarifies this in "... rights may be exercised in all media and formats whether now known or hereafter devised … include the right to make such modifications as are technically necessary to exercise the rights in other media and formats, but otherwise you have no rights to make Adaptations." This clarification allows for the creation of lossy conversions (e.g., creating smaller images as needed for small-screen media, or converting a lossless png-image into a lossy jpg-image as needed for mobile media), as long as the new representations remain truthful to the original work. In the case of tiny image thumbnails, the latter may no longer hold and the thumbnail may have to be classified an abridgment or condensation.

Icon of the "No Derivative Works" (= ND) condition (CC BY, Creative Commons)

The definition of "Adaptation" in the CC license is very far-reaching and the ND condition prevents many desirable uses: Cropping images or videos, adding arrows or lettering, translating into another language, creating time-synched relations between a

video and other media, and (with the exception of collections of works) the use of images in a new context. Even displaying ND licensed images in lectures or presentations may be prohibited. The license explicitly names synching of audio with video as an example of prohibited use of ND licensed works and a presentation is synching one media (e.g., an ND-licensed graph from a publication) with a live audio stream. While the ND condition may have useful applications for some works of art, we recommend avoiding it for copyrightable forms of biodiversity research documentation. To describe it by a biological metaphor: the ND license is sterile and cannot spread (Katz 2006).

The final condition is the "Non-Commercial" condition (abbreviated "NC"). This condition is widely used, but also often inadequately understood. The main topic of this article is the analysis of the implications of using this condition.

Icon of the "Non-commercial" (= NC) condition (CC BY, Creative Commons)

The "non-commercial" condition

The short description of the non-commercial condition of Creative Commons licenses is that one "may not use this work for commercial purposes". The full license text is:

"You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file-sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works." (http://creativecommons.org/licenses/by-nc/3.0/legalcode)

The full text no longer uses the term "commercial purposes", but only the concepts of "intent or direction" and "commercial advantages". To our knowledge, the concept of "commercial advantages" is at present neither defined by CC nor in the law of most countries (Keller and Mossink 2008; Wilson-Strydom 2009 p. 15). Between 2006 and 2008, a document "Proposed Best Practice Guidelines to Clarify the Meaning of Noncommercial" was developed on the CC wiki under "DiscussionDraftNonCommercial_Guidelines". This was replaced in 2008 (http://wiki.creativecommons.org/index.php?action=historysubmit&diff=19061&oldid=18887) with a link to a report "Defining Noncommercial" (Creative Commons Corporation & Netpop Research 2009). This report surveys the frequency distributions of various interpretations of the terms "commercial use" and "non-commercial use", mainly by U.S.A. Internet users. The survey confirms that significantly differing interpretations of "non-commercial" exist. The majority of users tend to identify "commercial" with "for profit". However, the study also shows that "uses by organizations, by individuals, or for charitable purposes are less commercial but not decidedly [considered as] noncommercial" (ibid., p. 73). Furthermore, the use of works surrounded by or connected with advertisements is largely considered commercial (score 82.6 to 84.6 on a scale of 0 = non-commercial and 100 = commercial). Many people will interpret it as

acceptable to use a work licensed as non-commercial in combination with advertisement for cost-recovery, while others will not. A major implication from this study is that the definition given in the CC license is ambiguous, since both sides believe that the CC NC license term is "essentially the same as" or "compatible with" their definition (ibid., p. 11).

In practice, the interpretations range from considering editorial use of images in a for-profit journal as non-commercial (e.g. the interpretation by Wired magazine, see Benton 2011a) to disallowing any use where money is exchanged, regardless whether for cost recovery or not. The question as to how “non-commercial” will be interpreted in court is largely unresolved. Given the large number of potentially contentious licensing cases (e.g., Prodromou 2005; Benton 2011a), a similarly large number of court decisions in relevant jurisdictions will be required. Until this is achieved, any long-term project that considers the use of CC NC licenses will require a careful assessment of legal risks. We present here some insights we have gained in our own risk management analysis, so as to inform the decisions of others.

Formally, the word “commercial” means “referring to commerce”, which in turn may be defined as, for example: “1. the activity embracing all forms of the purchase and sale of goods and services” (Collins 2003). The term “commercial” is thus not directly linked to the concept of making profits. A non-profit enterprise that buys and sells services is a commercial enterprise according to this and many other definitions. It can consequently obtain commercial advantages, e.g., by using images for a public awareness campaign under a free license rather than paying for them on the market. Different interpretations exist: “non-commercial” may be identified with “non-profit” (summarized, e.g., by Wilson-Strydom 2009) or it may be identified with “directly making money” (Kleinman 2008, ignoring commercial advantages that only later lead to monetary profits). However, licensors that intend to apply permissive interpretations of the NC license often feel obliged to clarify their point in a license interpretation statement (e.g., Massachusetts Institute of Technology 2011; Smith 2011; or examples given in Keller and Mossink 2008).

Importantly, the NC license does not refer to the status of potential users at all; focusing solely on the manner in which a work is used. Both for-profit and non-profit organizations may use NC licenses. However, non-profit organizations probably need to rely on factors other than their status to decide whether they may use NC-licensed works.

Monetary compensation and commercial advantages

The CC NC condition distinguishes between (1) a general definition of activities allowed under the license and (2) the special case of “the exchange of the work for other copyrighted works”. In the first case, “non-commercial” is defined in the NC condition by two elements:

a) "no private monetary compensation" (i.e., any kind of payment to the licensee by a third party) and

b) "no commercial advantage" to the licensee (i.e., any direct or indirect non-cash-profit, potentially including profits in reputation [e.g. through sponsoring] or savings of expenses [one does not have to buy a copy of the work in the shop...]).

The second case of exchanging copyrighted work does allow commercial advantages, focusing only on monetary compensation. The introduction of a special case stresses that (a) absence of "monetary compensation" is a core principle that is upheld in all cases, and (b) that any form of "commercial advantage" is a binding principle for all other activities than exchanging copyrighted works.

The authors further believe it reasonable that "compensation" includes both full and partial cost recovery.

Primary or secondary intent

All evaluations of intent only concern the user (licensee), not the copyright owner (licensor). The latter may well have commercial motives when releasing material under an NC license (see, e.g., Benton 2011a).

With respect to the licensee, the availability of the license does not depend on the type of legal entity, but on the context and goal of the activity in which the work is reproduced or re-used. The license specifies that it excludes activities that are "primarily intended for or directed toward commercial advantage". Deciding which "intention" or "direction" is the primary one is the main focus of controversy.

For example, a charitable non-profit organization may sell a calendar with CC-NC-licensed images as a means to raise funds. This is considered to be commercial use even by permissive interpretations of the NC-clause (e.g., Kleinman 2008), despite the fact that the ultimate intention for the funds is a charitable cause. But what about a general brochure, distributed free-of-charge? Increases in the membership base or in public recognition translate into a commercial advantage in the form of higher income through membership fees or voluntary contributions. To some extent, non-profit organizations compete with each other for donations and funds that the members of the public are willing to spend on membership fees. If a non-profit nature conservation organization uses an NC-licensed image in an advertisement brochure and the paid membership increases, it could be argued – similarly to the case of the calendar – that this use of the licensed work was primarily intended and directed toward commercial advantage.

In the case of for-profit companies, a commercial advantage can be assumed to be the primary goal in the majority of cases. Still, for example a for-profit journal, university or hospital may have a charter or mission statement that establishes charitable purposes as its ultimate goal, making the assessment of primary intent a non-trivial one.

The principle of primary intent does help with the question of cost recovery. Rutledge (2008) argues that the NC license allows for all forms of monetary compensation

that relate to recovery of costs, such as printing, postage, and even salaries, since cost recovery cannot reasonably be assumed to be a primary commercial motive. While this is a reasonable position in connection with monetary cost recovery, it remains doubtful whether it also eliminates concerns about gained or lost non-monetary commercial advantages.

In general, "intent" can be problematic to assess. Legal case history for assessing the non-commercial (or non-profit) status of individual actions in which an NC-licensed work is used is probably limited to District Court of Amsterdam (2006). However, a rich case history is available in most jurisdictions for the analogous case of assessing the non-profit or charity status of organizations for taxation purposes. Similarly to the CC NC licenses, such assessment goes beyond a simple calculation of profits. A non-profit organization may make losses in one year and profits in another without threatening its non-profit status, and a for-profit organization making losses several years in a row cannot simply claim a "non-profit" status for taxation purposes. Taxation status is typically assessed by a complex set of rules, governed by law, but in detail often defined by individual taxation authorities. Despite a long case history and detailed assessment rules, it is possible that an organization achieves non-profit status in one taxation district, and fails to do so in another. Assessing the non-commercial intent of individual actions in court may be vastly more complicated.

Re-use options of NC-licensed works

The CC NC clause defines wide-ranging limitations to protect the commercial interests of the creator or copyright owner of a work. In our understanding, the following conditions determine whether an NC-licensed work may or may not be re-used:

1. Any natural or legal person or organization, including commercial enterprises, may exercise licensed rights over an NC-licensed work. The ability to re-use, copy, or derive from a work depends on the context and goal of the activity, not on the type of legal entity exercising the rights.

2. Charging money for the work as a means to obtain a profit is clearly prohibited; there will be little doubt that this has been the primary intent when exercising the rights granted by an NC-license.

3. Charging money for the work as a means to recover cost seems initially prohibited. The license text uses the term "compensation" rather than "profit" or "gain" and stresses that this principle is to be upheld even in the case of exchanging works. However, cost recovery is likely permitted if a different primary intent and direction can be demonstrated.

4. Regardless of profit or cost recovery, the use of a licensed work may lead to (non-monetary) commercial advantages. Arguably, most uncharged uses of a work can be interpreted as an advertisement, and increased public recognition is generally seen as an advantage for any legal entity participating in commerce. Users of NC-licensed works must thus demonstrate that the use is neither primarily intended for, nor directed towards such increased recognition.

5. One might perhaps be in doubt whether the concept of "commercial advantages" might be applicable to private individuals as well: For someone working as a gardener or professional biologist, the action of re-publishing a biodiversity-related NC-licensed work could be assumed to be directed towards financial advantages (e.g., self-advertisement to improve the chances of finding new employment). However, the mentioning of "private monetary compensation" may be interpreted to implicitly clarify that the (broader) concept of "commercial advantages" is not to be applied to private individuals.

6. In most cases, the allowed use of an NC-licensed work therefore hinges on the question of whether the advertisement effects are primary or not. The following thought examples may demonstrate that the legitimacy of using NC-licensed content may be difficult to decide. Assume that an NC-licensed image is used in these contexts:

a) A large for-profit soft-drink producer runs an advertisement campaign "better drinks for a more joyful life".

b) A large for-profit company advertises their products with "50 cents from each purchase buys and preserves a piece of Amazonian rain forest".

c) A large non-profit nature conservation organization runs an advertisement campaign to increase its paying membership base, with the ultimate goal to increase its financial and political abilities to serve the cause of nature protection. However, by doing so, it is competing with other nature conservation organizations.

Most readers would probably consider cases a) and b) a license violation, but formally all organizations might claim that this particular action is primarily intended for and directed toward a public benefit. Thus, with different degree of likelihood, in each of these cases, a court might or might not decide that the advertisement is directed towards commercial advantages, making the use of the work a violation of the license terms.

Software

Software programs are copyrighted works and can in principle be released under CC licenses. This is, however, not recommended (Creative Commons 2011b). Unlike most other copyrighted works, software can be used as a tool to create other works. With respect to NC licenses, the condition "You may not exercise any of the rights granted to You in Section 3 above in any manner that …" implies that software licensed under such a license (e.g., xper2, Ung et al. 2010; FRIDA, Martellos et al. 2010; or OpenKeyEditor, van Spronsen et al. 2010) may not be used to produce creative works or non-copyrightable data sets for commercial purposes.

This is not dependent on the presence of a Share Alike condition. A work created with the help of a software application is normally an independent creation. The cases where software generates derivative works are fairly limited, e.g., where software-created works are primarily derivatives of copyright materials embedded in the software (i.e. materials other than software algorithms or source code) or where the arrangement and formatting applied by the software to non-copyrightable data is actually the primary

copyrightable creative component. This may indeed occur in biodiversity where data are formatted as software-generated "species pages".

This is, however, not the primary concern with NC-licensed software. The creator of such work created using NC-licensed software may have full ownership and copyright to it, but is limited by the contractual obligations which arise from using the NC license. The critical question is perhaps: Which level of diligence in preventing commercial use of such works or data sets is required? Is it sufficient that no commercial use was *intended* at the time of creation (but may the work later be sold)? May the author give it as a present to a third party, which may then put it to commercial use? Or is the author required to prevent this from happening for all times, including binding future copyright heirs?

Following the recommendations of Creative Commons, we advise that the only Creative Commons license suitable for software is the CC0 rights release license. Dedicated software licenses should be used in all other cases.

License compatibility

Works licensed under CC licenses that do not include the NC condition are naturally available for non-commercial use. However, a common misconception is that such works and those licensed with an NC condition can always be mixed in a derivative work, creating a new work under the more restrictive license.

While it is possible, e.g., to combine works licensed under CC BY-NC with works licensed under CC BY content, it is not possible to do so with works under licenses containing the Share Alike condition (e.g., the CC BY-SA license on Wikipedia text and most images). Share Alike prevents the use of a work under a more restrictive license – specifically in this case under an NC license. A derived work that combines NC-SA and other licenses must be shared under an NC license. This would be incompatible with the Share Alike license terms for an included CC BY-SA work (Katz 2006). License compatibility can be checked, e.g., with the Creative Commons Licenses Compatibility Wizard (Creative Commons Taiwan 2011).

The problem of license incompatibility may also arise when licensors, recognizing the problems with the CC NC license, amend it with their own definitions (see, e.g., Massachusetts Institute of Technology 2011, Smith 2011, or examples given in Keller and Mossink 2008). In the case of the CC BY-NC-SA license, if two licensors annotate a license in contradictory ways, these two licenses may be incompatible with each other (while each may remain compatible with unmodified and unspecified CC BY-NC-SA licenses).

License incompatibility problems also surface in relation to license models outside Creative Commons. Only the CC BY and CC BY-SA licenses (but not CC BY-ND, CC BY-NC, or CC BY-NC-SA) meet the criteria of openness that are used to determine compatibility with, e.g., software licenses laid out in each of:

- Open Knowledge Definition (Open Knowledge Foundation 2006),
- OSI Open Source Definition (Open Source Initiative 2004),
- Definition of Free Cultural Works (Möller 2008; Möller and Anonymous 2007ff),
- the GNU Free Software Definition (Free Software Foundation 2010),

One option to avoid such license incompatibility is to remove the "Share Alike" clause, insisting on attribution alone (CC BY, e.g., Benenson 2008). This further increases the dissemination and reusability of a work. However, this also allows the possibility that derived works may not be "given back", i.e. that works derived from free and open content may not themselves be open.

In light of the incompatibility between the most frequently used CC licenses (CC BY-SA and CC BY-NC-SA), a highly relevant question for biodiversity information dissemination is: Which combinations of works under different licenses result in a "collection" (in which cases the above CC licenses may be mixed) and which create an illicit derivative work or adaptation? In our experience, the (unambiguously incompatible) case of combining two texts seamlessly into a new work, such that the borders between the original works can no longer be traced, is not very relevant for the biodiversity domain. Typically, original works remain delimited and authorship and license of the parts documented. A web page with a gallery of images where the license and creators of each image is annotated will certainly be a collection. The same should apply for similarly clearly separated blocks of text, or combinations of text and image blocks.

Further, copyright law does not refer to digital representations but to abstract works. Thus, whether an image gallery is composed of separate files bound together by a web page, or whether the elements have been combined into a single file (e.g., because of the need for non-rectangular cropping or connecting elements) should not change the status as an "image collection", provided the parts remain individually recognizable and attribution and license individually documented.

However, the ways in which media (sound, images, or video) or text are combined in many biodiversity projects go significantly beyond image galleries or the traditional collection examples ("encyclopedias and anthologies") mentioned in the Creative Commons license text. Images and other media are often closely embedded and integrated with corresponding text. The CC licenses do anticipate creative arrangements. Collections may "by reason of the selection and arrangement of their contents, constitute intellectual creations" (http://creativecommons.org/licenses/by-sa/3.0/legalcode). Within biodiversity, the Encyclopedia of Life (EoL, http://eol.org) uses complex combinations of CC BY-NC-SA and CC BY-SA material. However, EoL has license agreements with its contributors allowing for use on EoL independent of the Creative Commons licenses. A more relevant example may thus be the complex ways in which Wikipedia occasionally combines text under CC BY-SA with images under various open content licenses share-alike-licenses, e.g., some images being licensed exclusively under the GNU Free Documentation License.

Licensing patterns

The majority of large-scale global collaborative projects promote the use of "free" or "open content" licenses. Free and open are often used interchangeably, but we will use them here in the sense that free just means that accessing the information does not involve costs beyond those of accessing the web, whereas open shall refer to the absence of "non-commercial" and "no-derivative" conditions.

The distribution of the various CC licenses depends on the cultural and commercial context of the various communities. Statistics maintained by Creative Commons to record various license uses show that 60% of all CC-licensed works in 2010 (primarily from Flickr and Yahoo, Linksvayer 2011a) are under non-free CC licenses (with ND, NC condition). The proportion of open licenses is slowly increasing over the years, however (Linksvayer 2011a, b).

Within the context of biodiversity, the proportions of non-open licenses are similar. A quality control web service (Morris 2009) showed that 76% of nearly 95 000 CC-licensed images in the Flickr EoL Images Group (Flickr 2011) had NC licenses on them. However, the average for EoL may be different, since EoL had other media sources in addition to Flickr. For the Atlas of Living Australia (ALA), 34 out of 58 CC licensed data sets include a non-commercial term (58,6%; 28 CC BY-NC, 6 CC BY-NC-SA, pers. comm. Miles Nicholls).

By contrast, a Google search reveals that among the PubMed Central corpus (U.S. National Library of Medicine 2011a), the open content CC BY license was chosen nearly three times as often as all NC licenses combined (Mietchen 2011).

The "Defining Noncommercial" report (Creative Commons Corporation & Netpop Research 2009) shows that the vast majority of copyright holders publishing works under a non-commercial license are willing to interpret the license in a liberal sense, e.g., accepting the use in combination with cost compensation or as advertisement by educational or non-profit organizations (see also Dobusch 2011). However, organizations planning to re-use NC-licensed works are a) forced to accept a legal litigation risk and b) are restricted due to license compatibility issues in the case of licenses containing the Share Alike condition. As a result, many public education projects like Wikipedia, OpenStreetMap, Wikibooks, Wikiversity, Connexions, Encyclopedia of Earth Citizendium, WikiEducator, Appropedia, etc. have decided that NC licenses are not suitable for them. Non-open licenses like CC BY-NC-SA seem to dominate in terms of number of published items, whereas open content licenses (CC BY, CC BY-SA) may dominate in terms of re-use.

By their very nature, the severe constraints on NC-licensed works reduce the societal benefits arising from those works (Möller and Anonymous 2007ff). Non-commercial licenses do not create the same kind of synergistic, agile, collaborative environment or re-use and continuous improvement that open content licenses create.

At the Creative Commons Global Summit 2011, CC representative Mike Linksvayer stated (Linksvayer 2011b; Dobusch 2011): "… the NC condition still sounds very appealing to many creators and is thus probably overused by those without exist-

ing revenue streams to a project. This could in turn lead to an under-use of non-NC licenses, which realize far more value since there are projects (e.g., Wikipedia) that rely on free licenses to exist."

The EU project ViBRANT (Virtual Biodiversity Research and Access Network for Taxonomy) is based on a combination of multiple platforms (Berendsohn et al. 2011). In its first years it recommended the use of a CC BY-NC-SA license on its Scratchpads web publication platform (Smith et al. 2009; Smith et al. this volume), the CC BY-SA license on the MediaWiki (biowikifarm, Hagedorn et al. 2010) platform, and CC BY (with major contributors choosing CC BY-NC-SA, however) on the CDM platform (Berendsohn 2010). The present paper is partly motivated by observing the resultant incompatibilities. For the future, contributors employing the ViBRANT Scratchpad 2 platform to be deployed in 2012 will be encouraged to use an open license. A CC BY license will be the default for new content, although users may choose other licenses, including those with a non-commercial clause.

Summary and conclusions

Creative Commons licenses are not antagonistic to copyright – they are based on it. A violation of a CC license is a copyright violation. CC licenses replace individual contracts (that the copyright owner and the user of a work may negotiate) with a standardized license. Managing individual licenses incurs a high legal and management overhead (which induces many publishers not to negotiate licenses, but rather to demand total transfer of copyright). The availability of a set of such standard contracts for a spectrum of use cases is an important feature of CC licenses.

The Creative Commons Non-Commercial (CC NC) licenses exclude re-use scenarios leading to monetary profits or other commercial advantages (increased notability, etc.). It thus effectively protects copyright owners whose income depends on commercially licensing their works. NC licenses therefore are an important instrument to contribute a work to causes in which a third party's gain does not diminish the revenue of the copyright holder. Contributing marketable works under an NC license is a laudable act.

Nevertheless, the NC licenses are also deceptive. The phrases "creative commons" and "non-commercial", together with the strong tendency in colloquial language to (incorrectly) identify "commercial" with "profit" and "non-commercial" with "non-profit", may suggest that releasing works under this license contributes to a "non-commercial commons" that is easily re-usable for all non-profit-minded entities. This, however, is not the case. NC licenses come at a high societal cost: they provide a broad protection for the copyright owner, but strongly limit the potential for re-use, collaboration and sharing in ways unexpected by many users:

1. While some interpretations plausibly argue that in public perception non-commercial and non-profit are widely seen as closely related, a public misconception is likely to be irrelevant in a court case. Most non-profit organizations or charities

engage in commercial activities like buying and selling goods and services. They are potential buyers of copyrighted works; allowing them to re-use a work free of cost potentially diminishes the commercial revenues of the copyright owner. Copyright owners licensing their works under an NC license might well intend to apply a strict interpretation of non-commercial, so as to not lose potential profits from this market sector.

2. The phrase "commercial advantages" covers a very broad spectrum of activities, including advertisements, sponsoring, fund-raising, or any other activity that improves brand recognition or public relations of an organization or individual. The fact that this is widely ignored (Wild 2011; Benton 2011b) does not make it legal.

3. The CC Attribution-NC-Share-Alike license is incompatible with the CC Attribution-Share-Alike license. NC licenses therefore cannot be used on major collaboration platforms like Wikipedia or Wikimedia Commons (Möller 2007; Wikimedia Commons 2009).

4. The decision whether an activity is "primary" or "secondary" will be difficult to argue and decide in courts. For example, fundraising will primarily be directed towards monetary gain. This, however, may ultimately be intended to hire a person to work in nature conservation. Risk management will require careful documentation of intent and actions while running a project involving the re-use of NC licenses.

In conclusion, the licensing concepts "commercial advantages", "primarily", and "intent" are difficult to define and assess, resulting in a significant risk of litigation to private persons as well as organizations that use works supplied under an NC license. Being an educational or non-profit organization does reduce the likelihood of litigation in terms of frequency (because many licensors accept such use). For a given litigation, however, we fear that a substantial risk exists of losing the case.

Individual claims of license violation brought forward by copyright owners can often be settled out of court. In some countries an internet platform may further be covered by some form of a copyright infringement liability limitation privilege (e.g., requiring a take-down-notice). However, another threat to project sustainability may come from competitors in the publishing business which may consider a particular use of NC licensed works illicit. Depending on the specifications of unfair competition laws in a given country, they may attempt to acquire an injunction stopping any "license violations" that lead to unfair competitive advantages. Should they succeed, this would then require to remove all NC-licensed materials from a project.

In addition to managing legal risk, projects considering to re-use, disseminate, or create derived works under NC licenses may also need to evaluate their future project development options. For example, collaboration needs and cost-compensation schemes for the provision of content on an Internet platform may differ from needs and schemes for the provision of works in print, on offline media like CDs or as smartphone applications. Creative Commons recommends seeking individual permissions for any use of NC licensed content that may be controversial as to whether it is commercial or not (Linksvayer 2009). Especially if content is created and re-used collaboratively on a platform that leads to tight integration of the contributions, it may not be practical to later reverse

the choice of license: All contributors would have to be contacted for a new negotiation. In our experience, the proportion of contributions which cannot be reached, have lost interest, never meant to market their contribution (having misunderstood the purpose of NC), or are unwilling to consent is relatively high. The contributions of these, which may be intermediate revisions of a text, then have to be laboriously removed.

Creative Commons is aware of the problems with NC licenses. Within the context of the upcoming version 4.0 of Creative Commons licenses (Peters 2011), it considers various options of reform (Linksvayer 2011b; Dobusch 2011):

- hiding the NC option from the license chooser in the future, thus formally retiring the NC condition
- dropping the BY-NC-SA and BY-NC-ND variant, leaving BY-NC the only non-commercial option
- rebranding NC licenses as something other than CC; perhaps moving to a "non-creativecommons.org" domain as a bold statement
- clarifying the definition of NC

The authors of this article view NC licenses as a valid choice. Without them, many works would not be publicly licensed at all. However, NC licenses should no longer be presented as an obvious or easy choice. Rather than abandoning NC licenses, we would prefer Creative Commons to rename and rebrand them, reducing the mismatch between the actual consequences and the expectations generated by terms like "non-commercial" and "creative commons". A combination of: 1) a name like "Non-Open Commons: Attribution-Commercial Rights Reserved, NCC BY-CR", 2) explanations on the license chooser highlighting potential misunderstanding, and 3) a visual design change in the license display of the short and long license texts (e.g. red-gray striped instead of yellow) might better communicate the actual consequences. Independently, a clarification of the license terms, stating that uses of NC-licensed works by organizations certified as charities or non-profit organizations for the purpose of taxation in their country of residence are always appropriate, might help to reduce the risk of using NC-licensed works. Such a clarification should not change the NC license by making the use of NC licensed words dependent on the status of the user. It should only clarify that certification by taxation authorities is a sufficient test to evaluate primary versus secondary intentions. Finally, a license update should attempt to clarify the borders of collections, and contain guidelines how to document the license status of collections containing a mixture of incompatible licenses.

Given such changes, we hope that the preference for NC-licensing by publicly funded organization who can afford to provide materials into an Open Content Commons is waning. We believe NC licenses should not be used for the dissemination of results from publicly-funded organizations or research projects. The public rightfully expects a return for its investments in the form of re-usable digital content. This is the new digital infrastructure of the information era.

With respect to individual users, the major providers of collaborative biodiversity platforms could immediately start to make the choice of NC licenses less deceptive. A choice of licensing options should be given and the NC license should be present in ways that avoid raising false assumptions. "All Commercial Rights Reserved, most use by for-profit as well as non-profit organizations prohibited" is a better representation of the effect of the license.

Open content licenses such as CC BY (used by many Open Access publishers) or CC BY-SA (used, e.g., by Wikipedia) will enable a much wider re-use of a contribution and increase the efficiency of non-profit organizations in informing and educating others about biodiversity and nature conservation. We therefore recommend copyright owners to balance the negative impact of the non-commercial restriction on open knowledge dissemination, collaboration and ease of re-use against income which may be lost. In many cases, the potential profits from commercial use are comparatively low or irrelevant.

However, a publisher may indeed, with appropriate citation of the authorship, use an openly-licensed work in a book that generates a profit. The resulting dissemination of knowledge on biodiversity, regardless of profits, may well be in the interest of biodiversity education and society in general. Open licenses like CC0, CC BY, or CC BY-SA allow the commercial and private sector to collaborate and to develop businesses based on and contributing to the digital commons (Keller and Mossink 2008, Fletcher 2011). Furthermore, open licenses will help small companies or local non-profit initiatives more than big companies. Large companies can afford to buy works and can bear high management overhead, the cost of legal advice, or the risk of litigation much better than small organizations and initiatives.

Each creator of a work considering licensing options is therefore encouraged to balance the potentially lost income against the increased benefit to society. Within our own field of biodiversity, we hope that more organizations and publishers encourage their contributors to avoid NC licenses. The "commons" of CC NC licenses is available to a few, but not to the many.

Acknowledgments

The authors thank S. Baskauf, W. Egloff, I. Kuchma, M. Linksvayer, R. Page, G. Riccardi, D. Roberts, V. Smith, as well as one anonymous reviewer for review and helpful criticism.

References

Agosti D, Egloff W (2009) Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2: 53. doi: 10.1186/1756-0500-2-53

Aoki K, Boyle J, Jenkins J (2006) Bound by Law. Center for the Study of the Public Domain, Duke Law School, ISBN 0974155314, http://www.law.duke.edu/cspd/comics/digital.php

and http://www.law.duke.edu/cspd/comics/pdf/cspdcomicscreen.pdf – archived at webcitation.org/637UJkHAQ

Badajoz Sixth Court (2006) Ordinary procedure 761/2005. Translated by Felipe L, Ambía S, Ringenbach J, Ruiz Gallardo C, Riquelme C, Paiva M http://mirrors.creativecommons.org/judgements/SGAE-Fernandez-English.pdf – archived at webcitation.org/6294snvKA

Benenson F (2008) Moving on from Copyleft. http://fredbenenson.com/blog/2008/10/22/moving-on-from-copyleft/ – archived at webcitation.org/6294snvKK

Benton J (2011a) Wired releases images via Creative Commons, but reopens a debate on what "noncommercial" means. Nieman Journalism Lab, Nieman Foundation for Journalism at Harvard University. http://www.niemanlab.org/2011/11/wired-releases-images-via-creative-commons-but-reopens-a-debate-on-what-noncommercial-means/ – archived at webcitation.org/63AbPFbCh

Benton J (2011b) "Consumers of Creative Commons licenses do not understand them": A little more context to Wired's use of CC. Nieman Journalism Lab, Nieman Foundation for Journalism at Harvard University. http://www.niemanlab.org/2011/11/consumers-of-creative-commons-licenses-do-not-understand-them-a-little-more-context-to-wireds-use-of-cc/ – archived at webcitation.org/63AALdKgr

Berendsohn WG (2010) Devising the EDIT Platform for Cybertaxonomy. In: Nimis PL, Vignes Lebbe R (eds) Tools for identifying biodiversity: Progress and problems. ISBN 978-88-8303-295-0, Trieste, 1–6. http://www.openstarts.units.it/dspace/bitstream/10077/3737/1/Berendsohn,%20bioidentify.pdf – archived at webcitation.org/62fHpaatU

Berendsohn WG, Güntsch A, Hoffmann N, Kohlbecker A, Luther K, Müller A (2011) Biodiversity information platforms: From standards to interoperability. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 71–87. doi: 10.3897/zookeys.150.2166

Björk B, Welling P, Laakso M, Majlender P, Hedlund T, Guðnason G (2010) Open Access to the scientific journal literature: Situation 2009. PLoS ONE 5(6): e11273. doi: 10.1371/journal.pone.0011273

Bourne PE, Fink JL, Gerstein M (2008) Open Access: Taking full advantage of the content. PLoS Comput Biol 4(3): e1000037. doi: 10.1371/journal.pcbi.1000037

Brown PO and 23 further participants (2003) Bethesda Statement on Open Access Publishing. http://www.earlham.edu/~peters/fos/bethesda.htm – archived at webcitation.org/637NXihls

Butcher N (2011) A Basic Guide to Open Educational Resources (OER). Commonwealth of Learning, Vancouver and UNESCO, Paris. ISBN 978-1-894975-41-4. http://www.col.org/PublicationDocuments/Basic-Guide-To-OER.pdf – archived at webcitation.org/62NtU3C16

Collins (2003) Collins English Dictionary, complete and unabridged. 6th edition, HarperCollins, New York

Creative Commons (2011a) History. http://creativecommons.org/about/history – archived at webcitation.org/62JIvzFQ0

Creative Commons (2011b) Can I apply a Creative Commons license to software? http://wiki.creativecommons.org/Frequently_Asked_Questions#Can_I_use_a_Creative_Commons_license_for_software.3F – archived at webcitation.org/62fEmvAyy

Creative Commons Corporation & Netpop Research (2009) Defining "Noncommercial": A study of how the online population understands "noncommercial use". http://wiki.creativecommons.org/Defining_Noncommercial and http://mirrors.creativecommons.org/defining-noncommercial/Defining_Noncommercial_fullreport.pdf – archived at webcitation.org/6294snvKb and webcitation.org/6295fKwmQ

Creative Commons Taiwan (2011) Creative Commons licenses compatibility wizard. http://creativecommons.org.tw/licwiz/english.html – archived at webcitation.org/6294snvKk

District Court of Amsterdam (2006) Curry v. Audax. Translated by Steijger L, Hendriks N http://mirrors.creativecommons.org/judgements/Curry-Audax-English.pdf – archived at webcitation.org/6294snvKt

Dobusch L (2011) CC Global Summit 2011, Pt. III: Discussing the non-commercial module. http://governancexborders.com/2011/09/17/cc-global-summit-2011-pt-iii-discussing-the-non-commmercial-module/ – archived at webcitation.org/6294snvL1

European Commission (2011) Policy initiatives: Open Access. http://ec.europa.eu/research/science-society/open_access/ – archived at webcitation.org/62NzYihSy

Europeana Foundation (2011) Europeana data exchange agreement. http://www.version1.europeana.eu/c/document_library/get_file?uuid=deb216a5-24a9-4259-9d7c-b76262e4ce55&groupId=10602 – archived at webcitation.org/62fZRR5b3

Fletcher K (2011) Why Not NC (Non Commercial) http://kefletcher.blogspot.com/2011/10/why-not-nc-non-commercial.html – archived at webcitation.org/639WxRt7W

Flickr (2011) Encyclopedia of Life Images Group. http://www.flickr.com/groups/encyclopedia_of_life/

Free Software Foundation (2010) The Free Software definition. Version 1.104, updated 2010/11/12. http://www.gnu.org/philosophy/free-sw.html – archived at webcitation.org/6296Nwgqb

Hagedorn G, Press B, Hetzner S, Plank A, Weber G, von Mering S, Martellos S, Nimis PL (2010) A MediaWiki implementation of single-access keys. In: Nimis PL, Vignes Lebbe R (eds) Tools for identifying biodiversity: Progress and problems. ISBN 978-88-8303-295-0, Trieste, 77-82. http://www.openstarts.units.it/dspace/bitstream/10077/3753/1/Hagedorn%20et%20alii%20bioidentify.pdf – archived at webcitation.org/62dagDhyV

Hirtle PB (2011) Copyright term and the public domain in the United States. http://www.copyright.cornell.edu/resources/publicdomain.cfm – archived at webcitation.org/62EGfHGOb

Katz Z (2006) Pitfalls of open licensing: an analysis of Creative Commons licensing. IDEA—The Intellectual Property Law Review, Vol. 46 (3): 391-413. http://law.unh.edu/assets/pdf/idea-vol46-no3-katz.pdf – archived at webcitation.org/62O4CcyrQ

Keller P, Mossink W (2008) Reuse of material in the context of education and research. SURFdirect, Utrecht. http://www.surffoundation.nl/SiteCollectionDocuments/Report_SURFCC_Reuse%20of%20material_Eng_DEF.doc – archived at webcitation.org/62KX892td

Kleinman M (2008) CC HOWTO #2: How to use a work with a NonCommercial license. http://mollykleinman.com/2008/08/21/cc-howto-2-how-to-use-a-work-with-a-noncommercial-license/ – archived at webcitation.org/62YsJP8wu

Landgericht Berlin (2010) Aktenzeichen 16 O 458/10. http://www.ifross.org/Fremdartikel/LG%20Berlin%20CC-Lizenz.pdf – archived at webcitation.org/6296M6cZH

Lessig L (1999) Reclaiming a Commons. Draft 1.01, Keynote address at The Berkman Center's "Building a Digital Commons". In: Lessig L, Nesson C, Zittrain J (editors): Open Code · Open Content · Open Law, Building a Digital Commons, Harvard Law School, Cambridge, Massachusetts. http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/opencode.session.pdf – archived at webcitation.org/62NtOPJfW

Linksvayer M (2009) Defining Noncommercial report published. http://creativecommons.org/weblog/entry/17127 – archived at webcitation.org/63Ap1S1Su

Linksvayer M (2011a) Notes on CC adoption metrics from "The Power of Open". http://labs.creativecommons.org/2011/06/27/powerofopen-metrics/ – archived at webcitation.org/637jlkG5e

Linksvayer M (2011b) The definition and future of noncommercial. http://wiki.creativecommons.org/images/c/c2/20110917-noncommercial.pdf – archived at webcitation.org/639Vz5aLr

MacCallum CJ (2007) When is Open Access not Open Access? PLoS Biol 5(10): e285. doi: 10.1371/journal.pbio.0050285

Martellos S, van Spronsen E, Seijts D, Torrescasana Aloy N, Schalk P, Nimis PL, (2010) User-generated content in the digital identification of organisms: the KeyToNature approach. Int. J. Information and Operations Management Education, 3 (3): 272–283 doi: 10.1504/IJIOME.2010.033550

Massachusetts Institute of Technology (2011) Open Course Ware, privacy and terms of use: MIT interpretation of "non-commercial". http://ocw.mit.edu/terms/#noncomm – archived at webcitation.org/62IkUt5CG

Mietchen D (2011) Comment on the blog post "A wiki approach to Open Access and Open Science", http://wir.okfn.org/2011/07/14/a-wiki-approach-to-open-access-and-open-science/#comment-6 – archived at webcitation.org/62BFpwGfJ

Möller E (2008) Definition of Free Cultural Works Vers. 1.1. Permalink http://freedomdefined.org/index.php?oldid=5437 – archived at http://www.webcitation.org/62Na6T89T

Möller E & Anonymous co-authors (2007ff) The case for free use: Reasons not to use a Creative-Commons-NC-License. Permalink http://freedomdefined.org/index.php?oldid=10561 – archived at webcitation.org/62Na5LaP3

Morris PJ (2009) Quality control for Flickr images in the group Encyclopedia of Life Images. http://www.aa3sd.net/qc_test/index.php

Nabhan V (2009) Study on limitations and exceptions for copyright for educational purposes in the Arab countries. http://www.wipo.int/edocs/mdocs/copyright/en/sccr_19/sccr_19_6.pdf – archived at webcitation.org/62JijyZZR

Open Knowledge Foundation (2006) Open Definition – conformant licenses. http://opendefinition.org/licenses/ – archived at webcitation.org/6294snvLB

Open Source Initiative (2004) The Open Source Definition (annotated) Version 1.9. http://opensource.org/docs/definition.php – archived at webcitation.org/6296IHziS

Peters D (2011) Copyright Experts Discuss CC License Version 4.0 at the Global Summit. https://creativecommons.org/weblog/entry/29639 – archived at webcitation.org/639WZpRsr

Prodromou E (2005) Use cases for NonCommercial license clause. http://lists.ibiblio.org/pipermail/cc-licenses/2005-April/002215.html – archived at webcitation.org/639X0WNck

Rutledge V (2008) Towards a better understanding of NC licenses. Connections 13 (1): 13. col.org/SiteCollectionDocuments/Connections_FEB2008.pdf and col.org/news/Connections/2008feb/Pages/fairComment.aspx – archived at webcitation.org/62JZrDOBC

Science Commons (2011) Protocol for implementing open access data. http://sciencecommons.org/projects/publishing/open-access-data-protocol/ – archived at webcitation.org/62fZlJ7nY

Seng D (2009) WIPO study on the copyright exceptions for the benefit of educational activities for Asia and Australia. http://www.wipo.int/edocs/mdocs/copyright/en/sccr_19/sccr_19_7.pdf – archived at webcitation.org/62JiiIIBn

Smith VS, Rycroft SD, Brake I, Scott B, Baker E, Livermore L, Blagoderov V, Roberts D (2011) Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 53–70. doi: 10.3897/zookeys.150.2193

Smith VS, Rycroft SD, Harman KT, Scott B, Roberts D (2009) Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. BMC Bioinformatics. 2009, 10 (Suppl 14): S6 doi: 10.1186/1471-2105-10-S14-S6. http://www.biomedcentral.com/1471-2105/10/S14/S6

Smith VS (2011) ViBRANT virtual biodiversity: Creative Commons Non-Commercial licences. http://vbrant.eu/content/creative-commons-non-commercial-licences – archived at webcitation.org/62Ng8tIuS

Suber P (2004-2010) Open Access overview. http://www.earlham.edu/~peters/fos/overview.htm – archived at webcitation.org/62BFuimPI

Swan A (2010) The open access citation advantage: studies and results to date (preprint) http://eprints.ecs.soton.ac.uk/18516/ – archived at webcitation.org/6294snvLK

Tribunal de première instance de Nivelles (2009) Cause no. 09-1684-A. http://www.turre.com/wp-content/uploads/2010-10-26-Décision-trib.-Nivelles-Lichôdmapwa.pdf [Belgium] – archived at webcitation.org/6296ABsmS

Ung V, Dubus G, Zaragüeta-Bagils R, Vignes Lebbe R (2010) Xper²: introducing e-taxonomy. Bioinformatics 26 (5), 703-704. http://bioinformatics.oxfordjournals.org/cgi/reprint/btp715v1.pdf – archived at webcitation.org/62fUFOGK0

U.S. National Library of Medicine (2011a) PMC. http://www.ncbi.nlm.nih.gov/pmc/ – archived at webcitation.org/6294snvLS

U.S. National Library of Medicine (2011b) PMC Open Access Subset. http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ – archived at webcitation.org/6294snvLb

van Spronsen E, Martellos S, Seijts D, Schalk P, Nimis PL (2010) Modifiable digital identification keys. In: Nimis PL, Vignes Lebbe R (eds) Tools for identifying biodiversity: Progress and problems. ISBN 978-88-8303-295-0, Trieste, 127-131. http://www.openstarts.units.it/dspace/bitstream/10077/3762/1/van%20Sprosen%20et%20al,%20bioidentify.pdf – archived at webcitation.org/62fUVPw4x

Wellcome Trust (2003-2011) Open access policy. Position statement in support of open and unrestricted access to published research. http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002766.htm – archived at webcitation.org/62BQ50b5U

Wikimedia Commons (2009) Licensing Justifications. Permalink: http://commons.wikimedia.org/w/index.php?oldid=25264121

Wild A (2011) Creative Commons Is Not Public Domain. http://blogs.scientificamerican.com/compound-eye/2011/08/28/creative-commons-is-not-public-domain/ – archived at webcitation.org/63AC0xaVH

Wilson-Strydom M (2009) The Potential of Open Educational Resources. Concept Paper Prepared by OER Africa. South African Institute for Distance Education, Johannesburg, ISBN 978-0-620-45936-5. http://www.oerafrica.org/ResourceDownload.aspx?assetid=281 – archived at webcitation.org/62NjCKCWX

World Intellectual Property Organisation (1979) Berne convention for the protection of literary and artistic works. http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html – archived at webcitation.org/62966pAfg

Xalabarder R (2009) Study on copyright limitations and exceptions for educational activities in North America, Europe, Caucasus, Central Asia and Israel. http://www.wipo.int/edocs/mdocs/copyright/en/sccr_19/sccr_19_8.pdf – archived at webcitation.org/62JiZesw7

ZooKeys 150: 151–166 (2011)
doi: 10.3897/zookeys.150.2167
www.zookeys.org

FORUM PAPER

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Towards the bibliography of life

David King, David R. Morse, Alistair Willis, Anton Dil

Department of Computing, The Open University, Milton Keynes, MK7 6AA, United Kingdom

Corresponding author: *David King* (d.j.king@open.ac.uk)

Academic editor: *V. Smith* | Received 29 September 2011 | Accepted 24 November 2011 | Published 28 November 2011

Citation: King D, Morse DR, Willis A, Dil A (2011) Towards the bibliography of life. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 151–166. doi: 10.3897/zookeys.150.2167

Abstract

This paper discusses how we intend to take forward the vision of a Bibliography of Life in the ViBRANT project. The underlying principle of the Bibliography is to provide taxonomists and others with a freely accessible bibliography covering the whole of life. Such a bibliography has been achieved for specific study areas within taxonomy, but not for "life" as a whole.

The creation of such a comprehensive tool has been hindered by various social and technical issues. The social concerns focus on the willingness of users to contribute to the Bibliography. The technical concerns relate to the architecture required to deliver the Bibliography. These issues are discussed in the paper and approaches to addressing them within the ViBRANT project are described, to demonstrate how we can now seriously consider building a Bibliography of Life. We are particularly interested in the potential of the resulting tool to improve the quality of bibliographic references. Through analysing the large number of references in the Bibliography we will be able to add metadata by resolving known issues such as geographical name variations. This should result in a tool that will assist taxonomists in two ways. Firstly, it will be easier for them to discover relevant literature, especially pre-digital literature; and secondly, it will be easier for them to identify the canonical form for a citation.

The paper also covers related issues relevant to building the tool in ViBRANT, including implementation and copyright, with suggestions as to how we could address them.

Keywords

bibliography, citation, reference manager

What is a Bibliography of Life?

At the time of writing, the first result when searching for "Bibliography of Life" is Rod Page's blog post from October 2010, *Mendeley, BHL and the "Bibliography of Life"* (Page 2010). In his post, Rod offers this definition:

> "bibliography of life," a freely accessible bibliography of every taxonomic paper ever published.

The principle of *freely accessible bibliographies* already exists in taxonomy, albeit focused in particular domains, such as ants (e.g., Antbase, http://antbase.org/) or fish (e.g., Fishbase, http://www.fishbase.org/). The aim of the Bibliography of Life is to employ the same approach as these existing bibliographies, but on a far more ambitious scale. The domain covered by this bibliography is to be the whole of taxonomy.

There is a precedent for this ambition. In the domain of Computer Science, the *Digital Bibliography & Library Project* (DBLP, http://www.informatik.uni-trier.de/~ley/db/) evolved from a small specialized bibliography to a digital library covering most sub-domains of computer science (Ley 2009). The increase in scope was driven by the library's users. From small beginnings, the bibliography now lists more than 1,700,000 publications (as at September 2011). At a larger scale and in a different discipline, biomedical science, PubMed (http://www.ncbi.nlm.nih.gov/pubmed) is a well-known database that provides free access to the MEDLINE database of references and abstracts. Both of these databases are maintained by publicly funded institutions rather than commercial organisations. The DBLP is hosted by the Universität Trier, in Germany and the PubMed database is maintained by the United States National Library of Medicine (NLM, http://dtd.nlm.nih.gov).

There is a similar drive in taxonomy to produce a comprehensive library and matching bibliography. We do not see commercial organisations rising to this challenge. For while there are excellent resources, such as Thomson Reuters' BIOSIS (http://thomsonreuters.com/products_services/science/science_products/a-z/biosis/), the focus in extending these resources is generally on modern, born-digital material, which is both relatively easy to process and potentially commercially profitable through copyright access charges. Taxonomic research is informed by the full history of publications in the subject, and so compared to many other sciences, the historical taxonomic literature remains relevant to current research. In general, commercial organisations do not appear to be actively extending their coverage of the historic literature. Hence, a number of digitisation projects exist, such as the Biodiversity Heritage Library (BHL, http://www.biodiversitylibrary.org/), that attempt to bring old paper documents into the digital age. There remains the problem, however, of producing a comprehensive bibliography of the newly digitised documents. We suggest that while the concept of a *bibliography of life* might be easy to define, the simple fact that it does not exist indicates there are practical difficulties with the idea. This article explores some of these difficulties, and a possible solution.

Creating the Bibliography of Life

There are two aspects to the creation of the Bibliography of Life. The first is the social aspect, which involves collecting the references and the second is the technical aspect, which involves providing the infrastructure to hold the references. The two aspects are shown in Figure 1 as *populate* and *build* respectively. Other boxes in the figures show how the issues discussed in the this paper relate to these two aspects that are involved in creating the Bibliography of Life.

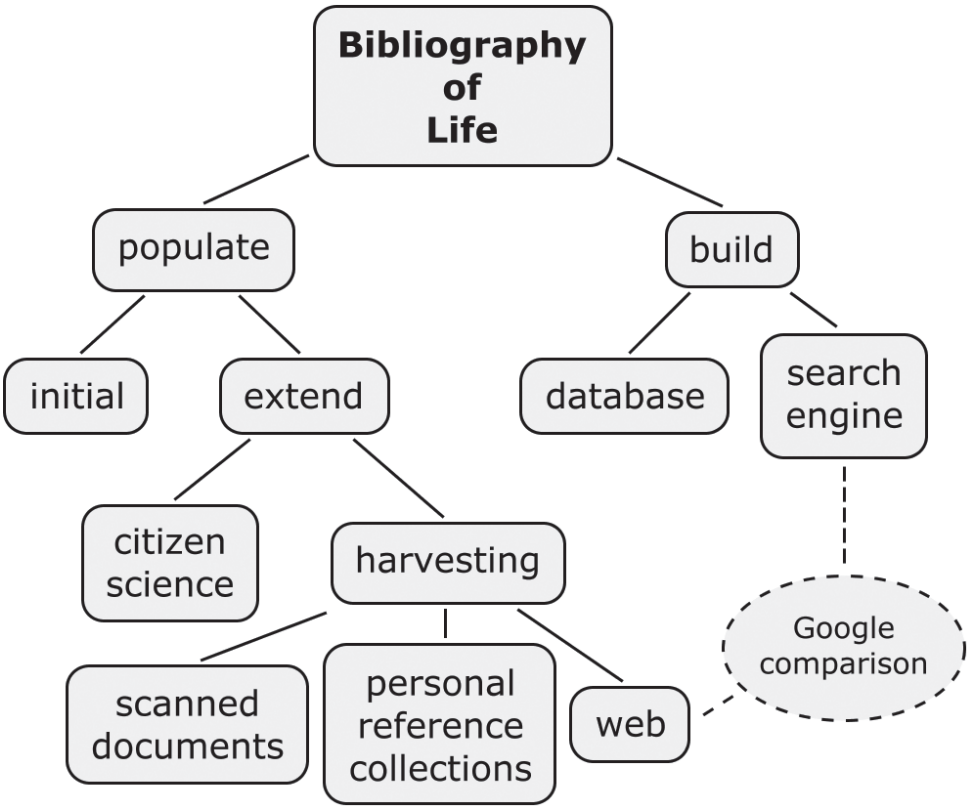


Figure 1. Social and Technical Aspects of the Bibliography of Life

We intend to *populate* the bibliography with references in two stages. There is the *initial* load from currently available sources to achieve critical mass and to prove the infrastructure. For sustainability we will provide the ability to *extend* the reference collection in the Bibliography of Life. This will be achieved by *harvesting* more resources, including those not generally accessible such as *scanned documents* and *personal reference collections*, and by harvesting *web* hosted resources including Scratchpads. This will be augmented by contributions through *citizen science*, such as the manual addition of references, as well as enabling all users to edit and refine references.

To support the Bibliography of Life infrastructure we intend to build two components. A *database* to hold the references, and a *search engine* to exploit the rich data available to us through holding our own copy of the references, including our own keyword lists and cross-links to original documents.

Hence, the bibliography of life will provide more support for the working taxonomist than existing web-based search engines, such as Google or Google Scholar. Figure 1 shows the points of comparison between a web-based search engine and the proposed Bibliography; in addition to data that can be harvested from the web, the Bibliography of Life must also harvest scanned documents and personal reference collections. The Bibliography will also provide its own database system that stores the key taxonomic facts and allows search to be optimised across these. The rest of this paper considers these steps towards delivering the Bibliography of Life in more detail.

Loading the initial set of references

The initial set of references for the Bibliography of Life's can be gleaned from existing resources. Biostor (http://biostor.org/) has demonstrated that a number of references – 63,873 as at November 2011 – can be accumulated relatively easily. However, this number is still relatively small. There has been some discussion (Hull 2010) around the notion that there are some fifty million published journal articles alone. Though this number covers all domains, it does suggest the scale of task in building a comprehensive Bibliography.

Owing to funding patterns there are many smaller bibliographic resources available to provide the initial set of references for the Bibliography of Life. In general, funding is predicated on breaking a big problem into smaller, manageable chunks. In consequence, there has been a multiplicity of databases built. In the absence of large-scale funding a *cottage industry* approach has taken hold, with those researchers interested in the technology and problems of bibliographic reference management building systems in their own personal time. This has meant that opportunities for added value are often missed, while large-scale challenges such as de-duplication and automatic validation are not addressed. The resulting resources are useful, but limited in their scope. They are, however, available for harvesting to populate the Bibliography of Life.

There are a variety of tools we can exploit or extend to harvest references. One such specifically designed for the taxonomic domain is FaLX, developed as part of the European Distributed Institute of Taxonomy project (EDIT, http://www.e-taxonomy.eu/). It could aggregate references from Connotea (http://www.connotea.org/), Scratchpads and CiteULike (http://www.citeulike.org/). We have not yet determined which harvesting tool will best serve our needs, or if we will need to develop our own.

The added value of a large-scale tool

This section discusses the added value we seek to achieve with the creation of a large-scale Bibliography. We intend it to represent something more than the sum of the content of existing, specialist bibliographic resources.

De-duplication

Ideally each target article should have a unique reference. However, multiple references can arise from the import of the same accurate reference into a bibliography from different sources, and also by the existence of near identical references to the same document. How to reduce duplication of bibliographic references remains an open problem in digital libraries research (Kan and Tan 2008). When a search retrieves many identical references to the same article the duplicates are easily ignored and only one copy of the reference is retained. A good cue for this is to check the Digital Object Identifier (DOI http://www.doi.org/) first. However, even if the DOI is the same, sometimes other data can be contradictory or incomplete. It is resolving these *near* identical references that can be difficult. A variety of resolution techniques are required because the problems can come from a variety of sources, such as using different journal abbreviations or a mismatch between fascicle and article page numbers.

The problem of reference de-duplication in bibliographic databases is more formally known as ***citation matching*** (Lee et al. 2007, Kan and Tan 2008), and improving on existing techniques will form one of the core areas of research for Work Package 7 in the ViBRANT project. A preliminary review of the landscape suggests that de-duplication techniques developed in information extraction and database management, and applied in other domains are not yet widely used in digital library curation. For example, we have found examples of citation tools being used to detect plagiarism (Plagiarism Today 2011), which might have transferable techniques we can exploit.

Internationalisation

Internationalisation is a common cause of near identical matches. This can occur when there are multiple names for the same entity such as place names or person names. Also problems arise with the transliteration of entities into Latin script. A topical example is that of the name "Gaddafi", which is also frequently transcribed as "Kadafi" or "Qaddafi". There are many variations of the name in Latin script, a problem compounded by the choice of formal Arabic pronunciation of the name or the Libyan dialect, and whether the name is transliterated for an English or French speaking audience (Time:Gaddafi 2011). Even equipped with this knowledge, however, no consensus has emerged on a unique Latin rendering (Yahoo:Gaddafi 2011).

The personal name problem is compounded by cultural differences, affecting such characteristics as name order. This can give rise to further variations depending on whether the name order is amended to match the typical Western style of given name first when the name is transliterated. The World Wide Web Consortium (W3C, www.w3.org/) has produced advice on handling this aspect of internationalisation (W3C:personal names) and other aspects of internationalisation too (W3C:internationalisation). Personal name variations are currently addressed by a variety of techniques including data mining (Phua et al. 2006), while Biostor implements Feitelson's (Feitelson 2004) weighted clique algorithm for finding equivalent names. These techniques achieve at most 85–90% accuracy, so there is room for further improvement in addressing this difficult problem. In addition, automatic matching techniques do not allow for the occasions when a researcher may deliberately use a different name for different publications, such as to distance themselves from their early work (McKay et al. 2010). As we can expect to encounter variations in author names stored in the Bibliography of Life, we expect to complement the automatic resolution services with an internal look up table to reconcile variations in the spelling of author names. This look-up table could be provided as a separate resource that could be queried via a web service.

Geographical names constitute a similar problem for the Bibliography of Life. For example, Lusaka, the capital of Zambia has been known in the past as Lusaaka, Lusaakas, Lusakas, Lusaka's and Lusaaka's. The general problem is compounded by the fact that spellings tend to be less codified in older sources.

Similarly, in the authors' previous work on the ABLE project (Automatic Biodiversity Literature Enhancement, http://able.myspecies.info/) we encountered an issue with the Anglicised spelling of central American locations in the *Biologia Centrali-Americana*: there was a consistent pattern of replacing an 'i' with a 'y'. Successful data mining of the literature identified by the Bibliography could allow us to build another look up table to help taxonomists resolve these name differences.

Journal abbreviations

A second common cause of mismatches is the varied abbreviations of journal names. Modern titles tend to follow the ISO 4 standard for abbreviating words and draw on the words in the ISSN's "List of Title Word Abbreviations" (http://www.issn.org/2-22660-LTWA.php). However, this does not apply to historic literature, with references to titles abbreviated before the international standard was codified. Similar techniques to resolving personal name variations can be applied to journal abbreviations. This collated list of variations could also be provided as a separate resource, which could be queried via a web look-up service.

Data quality

The question of data quality is not a new one, and it has many dimensions such as completeness, accuracy, correctness, currency and consistency of data (Redman 1996). Data quality can arise whether the reference is user submitted or harvested from an online library. There is no guarantee in either case that the input is validated. It would be a disservice to its users if the Bibliography of Life permitted the propagation of bad data.

Manual validation of the data is possible, and a Bibliography of Life requires an editing facility so that users can amend references. Such a service will be developed in ViBRANT by extending the functionality of the GoldenGATE editor so that it can commit the changes back into the Bibliography of Life. However, care must be taken by users editing bibliographic details since this could allow the introduction of new errors, typically through miskeying the intended change.

For the automatic addressing of quality issues, Ley and Reuther (2006) suggest two broad approaches.

The first approach to data validation they call *database bashing*. In this approach the data are checked against other databases. Unfortunately, this is not a foolproof approach because it is possible that both databases contain wrong data derived from a common source, and so an error can be propagated without detection. However, we will, where possible, check against external databases, although it is our ultimate goal that the Bibliography of Life will itself become the authoritative database for taxonomic references.

The second approach to data validation suggested by Ley and Reuther (2006) is *data edits*. This is the application of rules to highlight/resolve discrepancies. This can help address issues such as the Hungarian and Japanese use of family name first when giving names, which may or may not be amended to given name first in the reference. This approach is clearly limited to addressing known issues and common mistakes made when citing references.

We will use both approaches: referring to external resources and applying rule based corrections, to enhance data quality.

Thus far in the Bibliography of Life we have taken existing data and applied some initial steps to ensure the quality of the data. However, this alone will not ensure that the Bibliography of Life is a success.

Sustainability: extending the set of references

It is necessary that the Bibliography of Life adds sufficient value to working taxonomists so that they continue to engage with it. This is the critical success factor we see in delivering the Bibliography of Life. The initial set of references is unlikely to achieve

this, despite the advantages of data quality and quantity that it offers compared to smaller, more specific reference databases. We have the social challenge of building a community of users for whom it is worth their time and effort to contribute to the Bibliography of Life. This problem is potentially self-resolving once there are enough users and enough references to make it a truly useful resource. The question, of course, is how to achieve that desirable critical mass?

This is where building the Bibliography of Life through a larger project such as ViBRANT will be crucial, for ViBRANT gives users another reason to engage with the environment in which the Bibliography of Life is hosted.

How the Bibliography of Life would be used

We recognise that for the successful uptake of the Bibliography, it must integrate easily into the taxonomist's daily workflow. If interacting with the Bibliography becomes an onerous additional task, then the Bibliography will not be used. A possible workflow is shown in Figure 2.

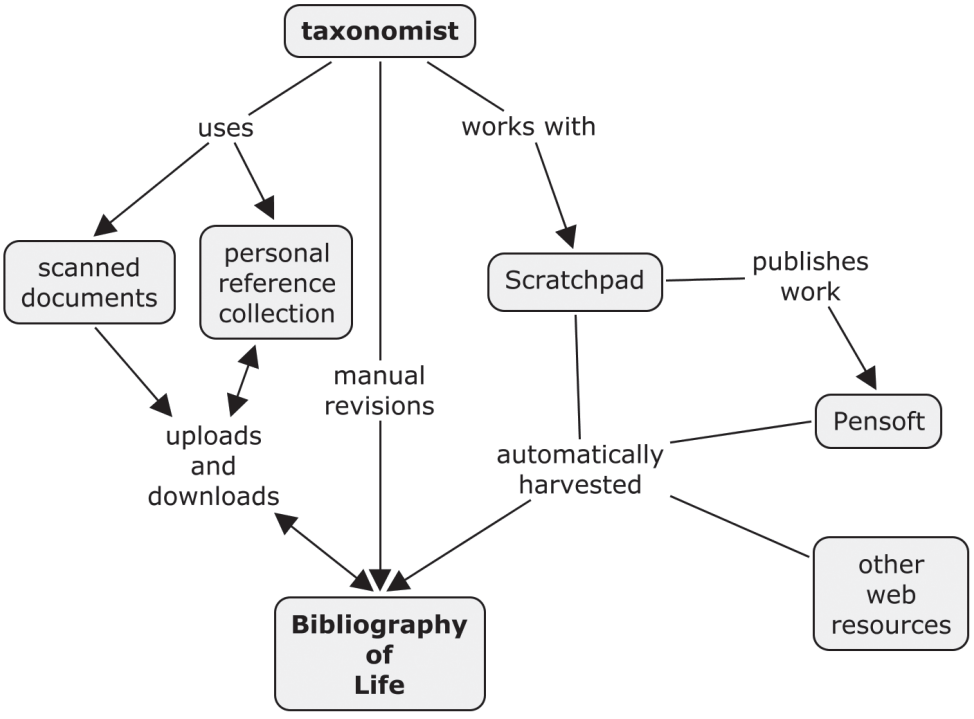


Figure 2. Interactions between a taxonomist and the Bibliography of Life

The ViBRANT environment provides Scratchpads (http://scratchpads.eu/), an on-line tool for taxonomists, encompassing open science and open publication in conjunction with social networking. The feature of Scratchpads most relevant to the Bibliography is the ability for users to store and share their bibliographies. Potentially then, Scratchpads will provide an important resource for a Bibliography of Life. To make this as simple as possible for Scratchpad users, references entered into their Scratchpad will be automatically validated against the Bibliography of Life and added to it, if necessary. In addition, new material published through Scratchpads and ViBRANT partner Pensoft (http://www.pensoft.net/) will automatically be added (Figure 2).

Complementing these two sources of new references, we will continue to revisit periodically the specialist databases used to provide the initial set of references. This will be supplemented by an extended web harvester, to access other less specialised web-hosted resources that contain relevant data. We can endeavour to test our coverage against that of generic search tools such as Google, so that there are not major gaps in our coverage of readily accessible references. There is, however, yet another source of smaller academic databases we wish to access.

Researchers maintain personal databases of domain relevant academic literature. These may be in formal personal reference management tools or simply as *ad hoc* Word documents. We intend that the Bibliography will accept data in all the common bibliographic reference styles, such as BibTeX and Endnote, as well as text strings in, for example, Word documents. To ensure that the Bibliography of Life is relevant to our users we will also have to provide matching export formats.

A related source of data is to parse literature directly for references, such as that held by the individual taxonomist. Parsing literature is a difficult problem, even for major commercial concerns such as Mendeley (Mendeley:reference extraction, 2010). One simple technique is to look for the isolated word "References" in the body of the text and examine the subsequent text. This is one of the methods used by open source tools such as ParaCite (http://paracite.eprints.org/) and can be effective on born-digital literature and on well-scanned historic literature. However, as a technique such keyword searches are limited in scope and depend on references being in a dedicated section within a document. Greater problems of automated extraction are provided by embedded references or, worse still, references in an endnote or footnote. Research into reference extraction from across the wide variety of historic taxonomic literature is one of our research goals within ViBRANT.

A further source of references, but one which brings another set of complications, are micro-citations. This is the minimal citation style peculiar to taxonomy, used by nomenclators. By their incomplete nature, satisfactorily resolving the citation is difficult (Gupta et al. 2009) though there are some examples we can build on to address the issues (Page 2011b). If the Bibliography of Life is to be the comprehensive tool envisaged, then we will need to incorporate micro-citation capture. This too, is the subject of one of our research goals.

Automatic extraction can be complemented by supported user input, as exemplified by GoldenGATE (http://plazi.org/?q=GoldenGATE), in which the user first identifies the reference which can then be parsed by the software extraction routines (Sautter et al. 2007). This is a useful facility for a user to add references as they read and review a document. This facility will be available to the Bibliography of Life.

Using other people's data: the issue of copyright

"Could it be true that laws designed more than three centuries ago, with the express purpose of creating economic incentives for innovation by protecting creators' rights, are today obstructing innovation and economic growth? The short answer is: yes." (Hargreaves 2011)

For the Bibliography of Life, copyright is an issue because current law prevents automated text processing for purposes such as harvesting texts for references. Although it is possible to negotiate a licence to do such processing with the rights holder (usually the academic publisher) on a case by case basis, this is impractical in general, and impossible in the case of orphan works, where the copyright holder is not known.

Some organisations choose to avoid working with potentially copyrighted materials simply to avoid the risk of copyright infringement. In our domain, BHL generally follows this approach, though working with information aggregators such as BioOne (http://www.bioone.org/) has enabled BHL to expand access to more recent, copyrighted publications (Rinaldo and Norton 2010). However, we do not have the option to ignore copyrighted material if we are to build a truly comprehensive Bibliography of Life that includes the modern literature.

Swiss-based Plazi (http://plazi.org/) have used the copyright laws particular to Switzerland to automatically extract taxonomic information from texts. However, these laws do not apply outside the Swiss jurisdiction and in any case, Plazi also argue (Agosti and Egloff 2009) that a system based on legal licensing is more desirable.

Without a resolution to this problem of licensing, the Bibliography of Life might be left with a gap in its records that undermines its sustainability. However, the Bibliography is not intended solely for the professional taxonomist. In other target user groups, some of the problems identified above may not arise.

The Bibliography is not intended solely for the professional taxonomist. In another target user group we may be able to circumvent some of the problems identified above.

Not just for professional taxonomists

The Bibliography of Life could also facilitate the work of citizen scientists. We expect such individuals to be competent taxonomists, being, for example, retired professional researchers or highly motivated amateurs. We do not envisage a role for more casual citizen scientists such as secondary school students in using and managing bibliograph-

ic references. We anticipate that citizen scientists will interact with the Bibliography of Life in a similar manner to the professional taxonomist. However, they will not have the same access to other professional tools so we must ensure that the Bibliography can adapt to their more *ad hoc* use of it. Following the lead of other domains of research, we hope that the citizen scientist will be particularly helpful with quality control by manually reviewing ambiguous data and by engaging in other manual processing of documents to, for example, identify taxon names. We will need to co-operate with the outreach partners within ViBRANT to encourage this behaviour in our users.

Underpinning this expected use of the Bibliography is the technical infrastructure to deliver it.

How to build it

There are two possible architectures for a Bibliography of Life: one is a dedicated database and the other is a search portal.

The first option is to build a database, for which there are two approaches. Either we can build our own database to store references or we can use an existing database. Building our own database gives us complete control over what we build so we can tailor it to meet our users' needs. While the first option sounds desirable, it does have to be built and carries the risk, through being yet another tool, of not achieving a critical mass of users.

The alternative is to build on another's database, leaving us only to ensure the sustainability of our taxonomic specific software enhancements. Of the currently available storage solutions, there are three front runners, in the commercial sector, Mendeley (http://www.mendeley.com/) and Papers (http://www.mekentosj.com/papers/), and in the public sector, CiteBank (http://citebank.org/).

Mendeley and Papers are both tools for an individual to organise their bibliographies. Both offer social network enhancements to enable papers to be shared among groups; though both restrict the number and size of groups and storage of references, that are available for free. If we were to work with either organisation then we will need to enter into a contractual relationship with them. Concerns over either organisation are their long term business plans and viability. The two named organisations represent the current leading on-line reference manager tools suitable for our use. There have been other earlier tools that rose, and then fell from prominence, such as CiteULike and Connotea. In a similar vein there is the publicly funded Zotero (http://www.zotero.org/), which has found a niche in the social sciences, but which would also require a commercial arrangement to handle the volumes of data a Bibliography of Life would generate.

Of the publicly funded bibliographic databases only CiteBank has the ambition to match the Bibliography of Life. Other databases are focused on a sub-domain of taxonomy and lack the scope to expand in line with the potential size of the Bibliography of Life. CiteBank is the bibliographic offshoot of the Biodiversity Heritage Library,

which has achieved sustained funding (BHL:funding). However, its current vision is to continue as an index to BHL content only, and so is not suitable for building the Bibliography of Life that we envisage (Freeland, pers. comm.).

An alternative approach is not to build a Bibliography of Life database at all, but a functionally equivalent portal offering a federated search across existing taxonomic bibliographic resources. Hence, our task in ViBRANT would be to build a user interface to a global search of these existing data stores, complemented by an index to speed up query results. The latter would be necessary because we would have to do additional processing such as de-duplication on the fly to consolidate the results. The leading, proven indexing technology applicable to this task is Apache SOLR (http://lucene.apache.org/solr/). It offers many advantages if used in ViBRANT, not the least being its integration with Drupal, the foundation for Scratchpads. However, to build the index would still require that we address the same issues as if we were to populate our own reference database. Given the potential performance penalty, there seems to be no advantage in adopting a purely search portal approach over populating a searchable database.

Therefore, for performance reasons, and the ease with which we can offer additional benefits, we propose to build a database in place of a portal. Further, to ensure continuity of service, we will follow the lead set by DBLP and host the database within an academic institution. For the immediate delivery of the service we intend to host the Bibliography within our employing institution, the Open University. Longer term, we will explore the other hosting options made possible by the ViBRANT environment.

Searching and extracting references

Having developed a database infrastructure, the second technical aspect to building a Bibliography of Life is extracting references from the database. For this we propose several approaches, including building our own dedicated search engine. However, we also intend to make use of existing services too, principally Mendeley.

There are several on-line tools for storing and sharing references. For the Bibliography of Life we intend to expose the references to Mendeley because it is the tool with the greatest coverage currently of taxonomic literature. This exposure will allow users to search the Bibliography of Life using a familiar tool, and should they wish, exploit the social networking aspects of Mendeley too. Note, the use of such tools is not without complications. For example, there are seven groups in Mendeley related to *ants* (Mendeley:ants), suggesting a fragmented approach to the researchers use of that tool.

These existing tools, however, do not deliver the full capability of a bibliography of Life. In particular, they will search primarily on published references and keywords. An advantage of hosting our own database is the extra value we can add by automatically reconciling author and journal names and extracting complementary metadata. Another possibility, if we can access the source document too, is for us to data mine it for additional keywords such as taxon names. These data can be added to the Bibli-

ography of Life because we control its design, and we can provide a search engine to exploit this additional data.

What it is not

The Bibliography of Life is not simply another search engine. Google (http://www.google.com/) is seemingly all-conquering in terms of popular search on the Internet. Its specialist academic derivative, Google Scholar (http://scholar.google.com/), is very popular too, based on informal, unscientific surveys. Yet these two search engines are not the solution to providing a Bibliography of Life.

Google and Google Scholar only search what is publicly available on the web. Private and personal bibliographies are not included in their results, neither in terms of breadth of coverage nor accuracy of information. These bibliographies are often a rich index to the pre-digital literature, which is not otherwise easily found even though the papers referenced are important in taxonomy. A Bibliography of Life can address this exposure, particularly for historic, taxonomic literature, which is only now being digitised and becoming publicly referenced on-line. Though it should be noted that contemporary, born-digital literature is well covered by these search engines.

A further complication arises from the different purpose of on-line search. For example, Google Scholar is aimed at helping researchers find articles, or related papers such as patent applications. Searches are based on authors or expected key words. If searching for keywords in the article itself, an overwhelming number of results can be returned. Defining a discriminating search query can be an arduous task. This could be made easier by the addition of appropriate metadata available to the search tool. A Bibliography of Life provides the opportunity to develop domain specific metadata to support searches. The relevance of the results is also affected by the granularity of the reference returned, especially when dealing with books or journal volumes. It would be far more productive to the taxonomist if the results referred directly to the relevant article, say, rather than the volume in which the article is found. This can be problematical in taxonomy, and other disciplines using scanned historic documents, because these are often indexed at the level of the scanned document rather than at the level of a meaningful search result (Page, 2011a). The whole scanned document might not be the most appropriate level of reference.

Hence, we argue for the creation of specific taxonomic reference tool to assist the taxonomist locate and manage accurate references as being preferable to relying solely on generic search engines.

Conclusion

This paper has outlined our intended approach to delivering a Bibliography of Life within the ViBRANT project. The Bibliography is specifically intended to benefit the

professional and expert citizen scientist working in taxonomy. We have set out the social and technical issues that have prevented its creation before.

The social concerns focus on the willingness of users to contribute to the Bibliography. This can be addressed initially by automatically collating existing references. This will also allow us to begin exploiting these data for the benefit of our users, and enhancing the quality of the data. Sustainability will be achieved through making the Bibliography an integral part of a taxonomists' workflow, and minimising any additional effort on their part to engage with it. We have shown how we intend to use Scratchpads to deliver this goal.

The technical concerns relate to the architecture required to deliver the Bibliography. We have argued that maximum benefit, in terms of being able to exploit the data, and greatest security of long term availability, is for us to build our own database. We recognise that users may wish to engage with the references using a variety of tools. We intend to expose the references to such new tools as Mendeley. In addition, to realise the maximum benefit from the data and the metadata we can extract from it, we will provide a dedicated search engine.

The ambitious vision of a comprehensive Bibliography of Life has not been realised before. In ViBRANT we have the commitment of a sufficiently large amount of time and resource to achieve a tool that can deliver more benefit to a taxonomist than existing smaller scale taxonomic bibliographic resources. In this, we will progress the vision of a "freely accessible bibliography of every taxonomic paper ever published" (Page 2010).

Acknowledgements

The authors would like to thank Vince Smith and Chris Freeland for valuable discussions about the Bibliography of Life.

The authors would also like to thank the reviewers and editors for their constructive suggestions which led to improvements to this paper.

ViBRANT (grant number 261532) is funded by the European Union 7th Framework Programme within the Research Infrastructures group.

All the authors are researchers at the Open University, which is the lead institution for Work Package 7 on biodiversity literature access and data mining in the ViBRANT project. David Morse is the Work Package leader.

References

Agosti D, Egloff W (2009) Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2: 53. doi: 10.1186/1756-0500-2-53

BHL:funding http://biodivlib.wikispaces.com/Funding+Sources

Feitelson DG (2004) On identifying name equivalences in digital libraries. Information Research 9(4): 192.

Gupta D, Morris B, Catapano T, Sautter G (2009) A new approach towards bibliographic reference identification, parsing and inline citation matching. In: Proceedings of the International Conference on Contemporary Computing, Noida (India), August 2009.

Hall D (2010) How many journal articles have been published (ever)? http://duncan.hull.name/2010/07/15/fifty-million/

Hargreaves I (2011) Digital Opportunity: A review of Intellectual Property and Growth. Crown copyright. http://www.ipo.gov.uk/ipreview.htm?intcmp=239

Kan M-Y, Tan YF (2008) Record matching in digital library metadata. Communications of the ACM 51: 91–94. doi: 10.1145/1314215.1340938

Lee D, Kang J, Mitra P, Giles CL, On B-W (2007) Are your citations clean? Communications of the ACM 50: 33–38. doi: 10.1145/1323688.1323690

Ley M (2009) DBLP - Some Lessons Learned. In: Proceedings of the Very Large Databases Conference, Lyon (France), August 2009.

Ley M, Reuther P (2006) Maintaining an online bibliographical database: The problem of data quality. In: Actes des sixièmes journées Extraction et Gestion des Connaissances, Lille (France), January 2006.

McKay D, Sanchez S, Parker R (2010) What's My Name Again? Sociotechnical Considerations for Author Name Management in Research Database. In: Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction - OZCHI '10, Brisbane (Australia), November 2010. doi: 10.1145/1952222.1952274

Mendeley: ants http://www.mendeley.com/groups/search/?query=ants

Mendeley: reference extraction http://feedback.mendeley.com/forums/4941-mendeley-feedback/suggestions/834313-version-0-9-7-does-not-extract-references-from-the-Mendeley

Page R (2010) Mendeley, BHL, and the "Bibliography of Life". http://iphylo.blogspot.com/2010/10/mendeley-bhl-and-of-life.html

Page R (2011a) Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. BMC Bioinformatics 12: 187. doi: 10.1186/1471-2105-12-187

Page R (2011b) Microcitations: linking nomenclators to BHL. http://iphylo.blogspot.com/2011/03/microcitations-linking-nomenclators-to.html

Phua C, Lee V, Smith K (2006) The Personal Name Problem and a Recommended Data Mining Solution. In: Wang J (Ed) Encyclopedia of Data Warehousing and Mining. Idea Group, London.

Plagiarism Today (2011) http://www.plagiarismtoday.com/2011/08/08/using-citations-to-detect-plagiarism/

Redman TC (1996) Data Quality for the Information Age. Artech House, London.

Rinaldo C, Norton CN (2010) The Biodiversity Heritage Library: an expanding international collaboration. In: Proceedings of the 36th International Association of Aquatic and Marine Science Libraries and Information Centers Conference, Mar del Plata (Argentina), October 2010.

Sautter G, Böhm K, Agosti D (2007) Semi-automated XML markup of biosystematic legacy literature with the GoldenGATE editor. In: Pacific Symposium on Biocomputing 2007, Maui, Hawaii (USA), January 2007.

Time: Gaddafi (2011) http://newsfeed.time.com/2011/02/23/how-do-you-spell-gaddafi-the-linguistics-behind-libyas-leader/

W3C: Internationalisation http://www.w3.org/International/

W3C: Personal names http://www.w3.org/International/questions/qa-personal-names

Yahoo: Gaddafi (2011) http://uk.news.yahoo.com/how-should-you-spell-gaddafi%E2%80%99s-name-.html

ZooKeys 150: 167–176 (2011)
doi: 10.3897/zookeys.150.2248
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Data standards, sense and stability: Scratchpads, the ICZN and ZooBank

Edward Baker[1,2], Ellinor Michel[1]

1 *International Commission on Zoological Nomenclature, London, United Kingdom* **2** *Department of Entomology, Natural History Museum, London, United Kingdom*

Corresponding author: *Edward Baker* (edwbaker@gmail.com)

Academic editor: *V. Smith* | Received 14 October 2011 | Accepted 22 November 2011 | Published 28 November 2011

Citation: Baker E, Michel E (2011) Data standards, sense and stability: Scratchpads, the ICZN and ZooBank. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 167–176. doi: 10.3897/zookeys.150.2248

Abstract

The International Commission of Zoological Nomenclature has used the Scratchpads platform (currently being developed and maintained by ViBRANT) as the foundation for its redesigned website and as a platform for engaging with its users. The existing Scratchpad tools, with extensions to provide additional functions, have allowed for a major transformation in presentation of linked nomenclatural tools. Continued development of the new website will act as a springboard for the ICZN to participate more fully in the wider community of biodiversity informatics.

Keywords

International Commission on Zoological Nomenclature, Scratchpads, Bulletin of Zoological Nomenclature, BioStor, BioCode, ZooBank, Global Names Architecture

Introduction

The International Commission on Zoological Nomenclature (ICZN: http://iczn.org) aims to provide 'standards, sense and stability for animal names in science' by acting as an advisor and arbiter for the zoological community. The ICZN produces the *International Code of Zoological Nomenclature* ('the Code') - a set of rules for the naming of animals and the resolution of nomenclatural problems. In addition it publishes the

Bulletin of Zoological Nomenclature (BZN) containing applications (Cases) made to the ICZN, Comments on these Cases, and the rulings of the Commission (Opinions). The ICZN is also responsible for creating ZooBank (http://zoobank.org), an online repository of nomenclatural acts intended to be the official registry of zoological nomenclature, currently under consideration as a prerequisite for electronic-only publication under the Code (details of this discussion with links to associated publications can be found at http://iczn.org/content/availability-electronic-publication).

Work began in December 2009 to move the ICZN's website to the Scratchpads (http://scratchpads.eu; Smith et al. 2011) platform. Prior to this the website was made of individual, hand-written HTML pages; with an increasing volume of content it was becoming difficult to maintain. The lack of a content management system (CMS) prevented many improvements to the site, including its visual appearance, use of metadata and development of new functionality.

Currently the ICZN finds itself at the centre of a biodiversity information crisis (e.g. Godfray 2002, 2007). The number of unrecognised taxa is estimated to be an order of magnitude more than currently described (e.g. Mora et al. 2011), legacy information is hidden behind barriers to access and shifting frames of reference, and fields such as molecular biology, not in existence at the founding of the ICZN in 1895, are presenting masses of data that is sometimes poorly contextualised and lacking a taxonomic framework. Maintaining a correct and stable nomenclature is important to all concerned with the living world, including those working in policy, public health and customs enforcement. The solution to the biodiversity information crisis will come as much from improved informatics and computer science as it will from biology - it is for this reason that the ICZN is investing considerable time in making functional and stable resources for nomenclature, and the sciences that it supports.

Another set of problems are sociological rather than technological in nature; as useful as the Code and ICZN rulings are to those who make use of zoological nomenclature, they require the acceptance of, and adoption by, the entire zoological community - from taxonomists to journal editors. A recent example of journal editors not fully understanding the requirements of the Code was the publication of *Darwinius masillae* by Franzen et al. (2009) in the (normally) electronic-only journal PLoS One. Since electronic-only publication is not allowed under the 4th Code, had the name been published as planned in e-only format, it would have been unavailable. In this instance the ICZN worked with the journal editors to create an interim solution of hard copy production of the journal for this nomenclatural act (http://iczn.org/plos).

The ICZN's response to these challenges is to develop ZooBank as an online repository for names, and the new website (using Scratchpads) as a platform for delivering not only the BZN but for outreach to the biological (and other) communities.

The initial transfer of content to the new website was completed at the end of March 2010 and the website has continued to evolve since then. The use of Scratch-

pads and the underlying Drupal CMS has allowed the ICZN to create a larger, more functional online presence and begin to create, organise and disseminate its outputs in a way that is standards-compliant, scalable and allows integration with other online services (e.g. BioStor). Outreach to the zoological community has been improved by online Frequently Asked Questions, guidelines of editors of journals publishing taxonomic papers, translations of the Code into foreign languages and providing a forum for discussing the draft BioCode.

How the ICZN uses Scratchpads

The aim of the ICZN site is to provide information about the Commission, its supporting body (the International Trust for Zoological Nomenclature), and provide access to the the Code and BZN. In this respect it differs from the majority of Scratchpad sites, which generally have a taxonomic focus. Unlike other Scratchpads, the ICZN website is not a resource built directly by a community (although Cases and Comments are written by the zoological community, only the ICZN Secretariat can add content to the site). The use of the Scratchpads platform has however allowed parts of the site to be used as a tool for community engagement (e.g. the draft BioCode).

Customisation overview

The ICZN site builds on the functionality of a standard Scratchpads installation in a number of ways including novel use of existing tools. The most obvious of these is a new theme designed for the ICZN, as well as a number of modules that are either not enabled in a standard Scratchpad (contemplate, views_Accordion), or that have been written for the ICZN website (iczn_aker, icznblocks) - Table 1.

Table 1. Additional modules used by the ICZN site over a standard Scratchpads installation

Module	Functionality
contemplate	http://drupal.org/project/contemplate Allows individual content types to be templated easily.
views_accordion	http://drupal.org/project/views_accordion Extends the Views module functionality to provide expanadable/collapsible displays of contents. (see e.g. http://iczn.org/category/faqs/frequently-asked-questions)
iczn_aker	https://github.com/edwbaker/ICZN-Aker Used to include (server-side) content from the private Case management Scratchpad onto iczn.org
icznblocks	https://github.com/edwbaker/ICZN-Blocks Provides the tabbed block on the iczn.org home page. Makes use of jQuery UI to provide transitioning effects.

Bulletin of Zoological Nomenclature

The Bulletin of Zoological Nomenclature (BZN: http://iczn.org/bzn) publishes Cases sent to the ICZN, Comments on these Cases and the rulings of the Commission (Opinions). Information relating to an individual Case is therefore spread out over several issues of the BZN. This, combined with the fact that taxonomists are generally interested in particular taxa, means that the traditional journal browsing structure of volume/issue/article is not necessarily the best way for visitors to find content. Previously BZN content was displayed as a series of Tables of Content for individual issues, and individual articles did not have their own page or associated metadata.

BZN: Improvements

Browsing

Visitors are able to browse the content of the BZN by major taxonomic group (a restricted vocabulary the ICZN has used for many years) and Case number in addition to the standard volume and issue. Browsing by Case allows the entire published history of a Case to be accessible on a single page (e.g. http://iczn.org/case/3455). This is the first time this has been achieved and demonstrates the clear advantage of digital management of distributed information such as nomenclatural cases and judgements.

Communication

Having the BZN online in a structured form for the first time has allowed the ICZN to automatically alert users automatically to new content by e-mail or RSS feed (http://iczn.org/content/notification-cases-comments-and-opinions).

Creating an account on the site allows the user to subscribe to e-mail notifications for all BZN content, or a subset defined by the ICZN's taxonomic groups.

Digitisation and metadata creation

The ICZN website has full bibliographic data for all BZN papers from Volume 63 (2006) to the present. In addition, the full text of Comments is also available (Cases and Opinions only have abstracts available).

The Biodiversity Heritage Library (BHL: http://www.biodiversitylibrary.org/bibliography/51603) has scans of Volumes 1-67, although it has no article-level metadata for these scans. The ICZN is using Rod Page's BioStor (http://biostor.org/issn/0007-5167) tool to collect metadata for those volumes for which we have no data for at present. Once data collection is completed, an export from BioStor will be used to populate the missing volumes on the ICZN website. These articles will have a link back to the article on the BioStor site (e.g. http://biostor.org/reference/66840) where

visitors will be able to view the relevant pages from BHL, or download the article in PDF format.

In the near future it is planned to release a set of simple instructions for people who would like to contribute to BZN metadata creation on BioStor; this will expediate collection of these data through crowd-sourcing.

BZN: Technical implementation

BZN papers belong to one of four categories; (General) Articles, Cases, Comments or Opinions. All papers are entered into the standard Drupal Biblio module (http://drupal.org/project/biblio).

Several other Scratchpads are used as an online platform for journals. The European Mosquito Bulletin (http://e-m-b.org) has many articles online, however there is no article-level metadata which is an essential requirement for the ICZN, and easily achieved with the Drupal biblio module which is part of the standard Scratchpads profile. A more scalable and functionally robust system has been used for the journal Phasmid Studies (http://phasmid-study-group.org/category/PSG-Publications/1165), storing article data using the biblio module, and creating a browsable volume/issue hierarchy using a standard Drupal vocabulary.

In an extension of the method used by Phasmid Studies the browsing of the BZN by volume/issue, taxonomic group and Case number is facilitated by the use of three separate vocabularies. These vocabularies are browsed using the Scratchpads TinyTax taxonomy browser (originally intended to navigate biological taxonomies). Pages relating to terms in these taxonomies are generated dynamically using the Mado module (originally designed to display species pages) and a small number of custom views.

BZN: Future plans

The use of web technologies can bring three key improvements to the BZN in the future:

1. easier submission of Cases
2. shorter time between a Case being submitted and an Opinion being issued
3. wider reach and community involvement

Online submission of Cases and Comments is a priority for the ICZN. The method we use to implement this functionality must integrate closely with the website, allowing papers submitted and edited online to be published to the site with a single click once they have passed the review. The Scratchpads platform already has support for the creation of papers and their electronic submission to a journal (Blagoderov et al. 2010) and it is hoped that we can adapt this process to fit our existing editorial and publication protocols.

It is hoped that the amount of time between a Case being published and an Opinion being issued will be reduced by allowing the pre-publication of Comments ac-

cepted for publication on the ICZN website. This is already being trialled for selected Comments (http://iczn.org/preprints).

Expanding the reach of the ICZN and increasing community involvement can be partially achieved by extending the automatic notification of new BZN papers by RSS or e-mail to mailing lists (e.g. Taxacom, ICZN-List). By expanding the at-present crude (although functional for a print journal) classification to order or family level for new Cases it will be possible to customise these alerts to contain only notifications about a given taxonomic group. This customisation would allow the editors of taxon-specific journals to easily publish details of new Cases, and for individual scientists to be made aware of Cases that might impact the nomenclature of 'their' group. Linking these feeds to social media platforms such as Facebook and Twitter will allow further dissemination of information to interested parties and help consolidate the ICZN's current social media presence.

The online BZN will be enhanced with additional XML metadata to allow nomenclatural acts to be automatically included in ZooBank. The XML schema to be used is currently under consideration by the ZooBank developers in consultation with Pensoft and others.

Case management

Case management: technical implementation

The ICZN uses a separate and private Scratchpad (Aker: http://aker.iczn.org) to manage data about Cases. A custom content type holds basic information about each Case and it's progress from submission, through publication, voting by the Commission and finally the publication of a Commission ruling (Opinion).

Customised views in Aker provide HTML content to the ICZN website. Currently this information is limited to a list of Case submissions that have yet to be published (new Cases) and Cases currently accepting Comments (open Cases). The use of the matrix editor allows batches of Cases to be edited simultaneously, which is particularly useful when it comes to advancing Cases through the system on BZN publication dates.

Aker provides information on new and open Cases in HTML format to the ICZN website. Using XHTML as the transfer format makes it easier for other people to re-use this content on their own sites either using an HTML iframe or a server-side solution.

The HTML is generated by using the 'XML data document' style in the Views module. The style options used to generate the HTML document are as follows:

Root element name	ul
Top-level child element name	li
Content-type	text/html

These settings result in a standard HTML unordered (bulleted) list.

A custom module, iczn_aker, is used on the ICZN website to provide a server-side solution to the display of these views.

Case management: future plans

We are planning to develop a system for online submission that integrates with the private Scratchpad, allowing for seamless online management of Cases. Similar technology already exists in the Scratchpad's publication module, but is likely to need extensive customisation as we need not only to allow for the creation and submission of articles, but also to see them through review and publication.

The ability to filter the XHTML output of Case information by the ICZN's taxonomic grouping would allow external websites to include details of relevant new and open Cases dynamically. Adding the ability to search by taxon name would allow details to be displayed on dynamically created taxon pages, such as those used generated by Scratchpads and SpeciesFile.

The Code

The International Code of Zoological Nomenclature is a set of rules and protocols for the naming of animals.

The Code: future plans

The Code is a long and technical document that can be challenging or intimidating to first-time users. The ICZN Secretariat has written concise instructions for journal editors publishing nomenclatural acts that explains what they must do to meet the Code's requirements. It is hoped that this will be the first of several sets of guidelines. Although these documents will stand in their own right, they must also be integrated with both the ICZN Scratchpad and the online Code.

At present, foreign-language versions of the Code are not presented via the same interface as the official code. As Drupal has in-built support for translations in the long-term there is a possibility that this could function as a platform for both the transcription and display of future translations.

There is currently an Editorial Committee charged with writing a new edition of the Code. Discussions are active on whether the revised Code could be streamlined and simplified by more dynamic, linked structure. Development of the new Code in conjunction with the Scratchpads structure could present technical improvements that make the Code a more widely accessible tool.

Official Lists & Indexes

The Official Lists and Indexes of Names and Works in Zoology (ICZN 1987; Supplement: ICZN 2001) give details of the names and works (publications) ruled upon by the ICZN. Although when dealing with paper publications such lists and indexes provide a useful, if not essential, entry point to the BZN, with an improved and metadata-rich online presentation of the journal the need for separately maintained entry points is eliminated. The expanded metadata required to allow harvesting of BZN papers by ZooBank will provide the information required to create an automatically updated, searchable version of the print publication that can also be filtered by taxonomic group and that links directly to the Case history and associated Opinion.

Draft BioCode

The 2011 draft BioCode aims to "provide an over-arching common framework" (http://www.bionomenclature.net/biocode2011.html) for biological nomenclature, working alongside the existing (or special) Codes. The 2011 draft BioCode was published in various journals and websites (e.g. Greuter et al, 2011). In order to make it clearer how the BioCode relates to the special Codes an article level treatment was created on the ICZN Scratchpad (http://iczn.org/biocode) showing the BioCode article alongside relevant articles from, and links to, the special Codes.

Although the draft BioCode is not an output of the ICZN it was decided that providing a forum for its discussion would benefit the zoological community and improve response. People wishing to contribute to the discussion can create a free account (http://iczn.org/user) and leave comments on individual articles. It is hoped that grouping articles from the draft BioCode and special Codes together in a thematic format will provide a useful resource not only for putting the draft BioCode in context but also for comparisons of style, language, structure and methodology between the existing Codes.

Draft BioCode: technical implementation

BioCode articles are linked to articles form the special Codes via a standard nodereference field.

The collapsible/expandable views are created using the Views Accordion plugin for the Drupal Views module. These are included on BioCode article pages using the Contemplate module and the views_embed_view() function. This functionality is not enabled by default on Scratchpad sites as it would allow site owners to run arbitrary PHP code.

Outreach

In addition to the technical procedures outlined above the use of a CMS has allowed more time for additional content to be made available online. The bulk of this content has focused on bridging the sociological gap between working biologists and the nomenclatural community. Examples of this work include a series of Frequently Asked Questions (http://iczn.org/faqs), some educational resources about nomenclature (http://iczn.org/education_outreach), videos of ICZN sponsored events (http://iczn.org/video) and even a PodCast (http://iczn.org/podcast).

The future

"The Linnean Enterprise has persisted for two and a half centuries, and the ICZN Code is itself more than a century old" (Pyle and Michel 2010).

The ICZN has responsibilities that go far beyond serving the current generations of zoologists. It must also honor and preserve the "robust historical legacy" (Pyle and Michel 2010) of existing zoological nomenclature whilst also ensuring a stable platform for its long-term future. This can only happen by working with the zoological community in its entirety.

Through ZooBank the ICZN will help to create the Global Names Architecture - allowing taxonomic and other biological resources across the web to connect to each other and create a resource far greater than the sum of its parts. The ICZN website will continue to expand, not just as a destination for people to find information, but as an active platform and arena for the zoological and other nomenclatural communities to converse with the Commission and each other.

The Scratchpads platform has enabled the ICZN to make large strides towards its goals, both technical and sociological. In the future we wish to build upon this relationship, both as a user of, and contributor to, the Scratchpads platform and via ZooBank as an engaged participant of the global biodiversity informatics landscape.

Acknowledgements

The ICZN would like to thank NHM volunteers Mike Higginson and Alex Panagiotopoulos who have contributed to digitisation and metadata creation for the BZN as well as Tom Ensom who has helped in the compilation of articles from the special Codes for the BioCode comparison project. The authors would like to thank Rod Page for facilitating our use of BioStor as a tool for BZN digitisation and for sharing his ideas on the future of both the BZN and the ICZN.

This work uses infrastructure that has been developed by the EU funded ViBRANT project (Contract no. RI-261532).

References

Blagoderov V, Brake I, Georgiev, T, Penev L, Roberts D, Rycroft S, Scott B, Agosti D, Catapano T, Smith VS (2010) Streamlining taxonomic publication: a working example with Scratchpads and ZooKeys. ZooKeys 50: 17-28. doi: 10.3897/zookeys.50.539

Franzen JL, Gingerich PD, Habersetzer J, Hurum JH, von Koenigswald W, Smith BH (2009) Complete Primate Skeleton from the Middle Eocene of Messel in Germany: Morphology and Paleobiology. PLoS One 4(5): e5723. doi: 10.1371/journal.pone.0005723

Greuter W, Garrity G, Hawksworth DL, Jahn R, Kirk P, Knapp S, McNeill J, Michel E, Patterson DJ, Pyle R, Tindall BJ (2011) Draft BioCode (2011): Principles and Rules Regulating the Naming of Organisms. Bulletin of Zoological Nomenclature 68: 10-28. http://iczn.org/node/7216

Godfray C (2002) Challenges for taxonomy. Nature 417: 17–19. doi: 10.1038/417017a

Godfray C (2007) Linnaeus in the information age. Nature 446: 259–260. doi: 10.1038/446259a

ICZN (1987) Official Lists and Indexes of Names and Works in Zoology. London: International Trust for Zoological Nomenclature.

ICZN (2001) Official Lists and Indexes of Names and Works in Zoology. Supplement: 1986-200. London: International Trust for Zoological Nomenclature.

Mora C, Pittensor DP, Adl S, Simpson AGB, Worm B (2011) How Many Species Are There on Earth and in the Ocean?. PLoS One 9(8): e1001127. doi: 10.1371/journal.pbio.1001127

Pyle RL & Michel E (2010) ZooBank: Revieweing the First Year and Preparing for the Next 250. In: Polaszek A (Ed) Systema Naturae 250 - The Linnean Ark. CRC Press, Boca Raton, 173-184. doi: 10.1201/EBK1420095012-c16

Smith VS, Rycroft SD, Brake I, Scott B, Baker E, Livermore L, Blagoderov V, Roberts D (2011) Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 53–70. doi: 10.3897/zookeys.150.2193

ZooKeys 150: 177–192 (2011)
doi: 10.3897/zookeys.150.2191
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Who learns from whom? Supporting users and developers of a major biodiversity e-infrastructure

Irina Brake[1], Daphne Duin[2], Isabella Van de Velde[3], Vincent S. Smith[1], Simon D. Rycroft[1]

1 *Department of Entomology, Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom* **2** *VU-University Amsterdam, Department of Organization Sciences, De Boelelaan 1081, 1081 HV, Amsterdam, The Netherlands* **3** *Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussels, Belgium*

Corresponding author: *Irina Brake* (i.brake@nhm.ac.uk)

Academic editor: *L. Penev* | Received 3 October 2011 | Accepted 23 November 2011 | Published 28 November 2011

Citation: Brake I, Duin D, Van de Velde I, Smith VS, Rycroft SD (2011) Who learns from whom? Supporting users and developers of a major biodiversity e-infrastructure. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 177–192. doi: 10.3897/zookeys.150.2191

Abstract

Support systems play an important role for the communication between users and developers of software. We studied two support systems, an issues tracker and an email service available for Scratchpads, a Web 2.0 social networking tool that enables communities to build, share, manage and publish biodiversity information on the Web. Our aim was to identify co-learning opportunities between users and developers of the Scratchpad system by asking which support system was used by whom and for what type of questions. Our results show that issues tracker and emails cater to different user mentalities as well as different kind of questions and suggest ways to improve the support system as part of the development under the EU funded ViBRANT programme.

Keywords

Shared knowledge, computer-supported cooperative work, issue tracking, software engineering, e-infrastructures

Introduction

Recently, many large research projects have developed e-infrastructures that are used by scientists with varying degrees of IT skills and by developers with sometimes little knowledge of the needs of the users. The key for large user uptake of an e-infrastructure

is to address this knowledge gap by encouraging the two groups to talk to each other. Ideally the communication is bidirectional and instructive for both groups. An important question is how to support this type of communication?

With the emergence of the interactive web, Web 2.0, a range of computer supported communication systems have been developed that facilitate learning from and between users and development teams. The present paper investigates to what extent the Scratchpad support services provide learning opportunities for both groups by asking: which support systems are used, by whom and for what type of questions? Additionally, we will reflect on the pros and cons of the different systems. The Scratchpad project has a variety of support services and we will focus on the use of two particular support services, the "help emails" and the "issues tracker", for which we have access to the usage data.

The results of our study aim to i) increase knowledge on users' and developers' needs for information; ii) further improve communication between developers and users; iii) improve the support system performance of Scratchpads and similar initiatives.

In the following paragraphs we give a short background of our research setting, followed by a description of the data and methods used, and conclude with a discussion of the results and formulate recommendations for project management and further research.

What are Scratchpads?

Scratchpads (http://scratchpads.eu) are a Web 2.0 social networking tool that enables communities to build, share, manage and publish biodiversity information on the Web. Scratchpad sites range in function from supporting the work of societies and conservation efforts to the production and dissemination of species pages and peer reviewed journal articles.

Scratchpads are free and rely on the open source content management system Drupal (http://drupal.org/). The system allows individuals or groups of people to create their own networks supporting their research communities on the Web. The tool is flexible and scalable enough to support hundreds of networks each with their community's choice of features, visual design, and data. A detailed description of the system architecture and template design of Scratchpads can be found in Smith et al. (2009) and Smith et al. (2011) in this volume. Scratchpads are further developed as part of the EU FP7 funded ViBRANT project (http://vbrant.eu/) and additional support is provided by the NERC funded eMonocot project (http://e-monocot.org/).

As of 7 September 2011, Scratchpads serve 4,299 registered users across 283 sites (see Fig. 1 in Smith et al. in this volume), ranging from academic to citizen-science audiences. These users have generated 374,770 pages of content since Scratchpads were founded in 2007.

Co-learning and the Web

The very nature of a Web 2.0 environment like Scratchpads makes it possible and imperative that users and developers collaboratively use and build on information systems. Although both still have their own roles and expertise these are highly entangled and benefit from open communication flows. Simply put, users and developers teach and learn from each other about what they need, know and experience when using the system and internalise this knowledge in their day to day work. In this paper we call this co-learning.

The added value of involving users in product design has been widely reported on by Von Hippel et al. (1994, 1995, 2005). They use the concept of "sticky information" to describe value and challenges of integrating local (user) knowledge in product design. Crowston et al. (2008) state that a buffer of active users is a desirable feature in Open Software projects. According to them "active users create a rich support structure and their archived answers form a valuable knowledge base" (p.70). Inspired by Wagner's (1997) perspective on co-learning, we argue in this paper that this knowledge base could be of use for users as well as for developers. Wagner (1997) formulates co-learning as an agreement between two parties (in his case researchers and practitioners). In a co-learning agreement he states:

Both (parties) are engaged in action and reflection. By working together, each might learn something about the world of the other. Of equal importance, however, each may learn something more about his or her own world and its connections to institutions and schooling (1997, p.16).

With the Web 1.0 e-learning was introduced. Quickly e-courses and e-conferences were made available by institutions that before specialised in offline teaching, very much a one way direction of learning from teacher to student. With Web 2.0 and its integration of interactive technologies, Wagner's definition can now be applied to offline and online learning settings. Colazzo et al. (2008) describe how the introduction of "virtual-communities" has not only changed the relation between people involved in learning activities but also the technical approach to e-learning. Their argument is that e-learning has evolved into co-learning with "co" referring to "collaborative" and the "community" element of the interactive Web.

Hence support services can have multiple functions for different actors. This may all sound clear-cut but is co-learning an easy process? Perhaps not. For instance Schuler et al. (1993) describe how software development can significantly benefit from genuine communication between developers and users. However, they stress this is not a straightforward process to set up. Potential barriers such as different values, work styles, even languages may hinder the communication (p. 107). Also we know from Bratitsis et al. (2008) that simply making support service available does not ensure that they will be used. In short, co-learning in a Web 2.0 setting appears to have much to offer for innovation and usability of a system but only bears fruit under the right conditions.

In our case study we deal with: a research e-infrastructure; multiple technologies that facilitate learning; and two parties (users and developers) which engage in a co-learning agreement. The learning technologies we refer to are smart services for communication. By analysing the usage data of the Scratchpad support services we aim to measure the presence of co-learning opportunities in the Scratchpad environment. Additionally, we will explore the process of co-learning to better understand mechanisms behind the use. Based on Wagner's concept of co-learning (1997) we argue that in our case a co-learning opportunity appears every time a message is posted in one of the support services, either by a user or a developer.

For the purpose of the argument that we make in this paper the users and developers are portrayed as distinct communities, while in reality the line between user and developer is often fluid. Some users have a developer's background and some developers have a research background in the field.

Support structure

The Scratchpad platform offers a number of support systems. In this paper we will focus on the two support systems most relevant to our research question on co-learning: the request emails ('contact us' email and direct mailing to the Scratchpad development team) and issues tracker. The complete Scratchpad support structure is detailed in Appendix 1.

The Scratchpad 'contact us' email (scratchpad@nhm.ac.uk) has been active since about August 2008. The emails cover general enquiries about the project, specific help requests, feature requests and bug reports. They are received by the whole Scratchpad development team and are answered by the team member best suited to the task. After an initial contact via the 'contact us' email or during training sessions, many requests are sent directly to the personal email address of team members thus making them more difficult to track.

To overcome this lack of overview, the issues tracker (http://dev.scratchpads.eu/project/issues) was implemented in September 2010. This tracker uses a Drupal module and is integrated into the Scratchpad system. Users access the issues tracker via their individual Scratchpad and are automatically logged in with their username. The user can view existing issues or create a new issue for which he/she needs to select whether it is a bug report, feature request or support request. The issue is added to the list and an email is sent to alert the Scratchpad development team. Each time the request is updated an email is sent to the user as well as to the Scratchpad development team. Issues are picked up by the developer responsible for this kind of requests or can be delegated to a certain developer. New issues are marked as "active" and as the issues are dealt with, the status is changed to other values, like "fixed", "duplicate", "postponed", etc.

Users access the issues tracker via a tab on their Scratchpad. This tab also gives the titles of the last ten issues, so that the user can check whether for example a recent bug has been filed already. If this is the case, the user can subscribe to an issues to receive

notification about any updates to this issue, thus for example learning how a specific problem can be solved.

As we highlighted above the Scratchpad support systems facilitate a two-way flow of information between users and developers. Although from the outside it might look as if the services cater first of all for the information needs of the users, they help developers in their work as well. Apart from being alerted to bugs, developers use the information gained from requests to improve the usability of the Scratchpad system as well as the support system itself. Additionally, requests for new features influence the decision process for the future development of the system. For example, several users asked for a quick way to simultaneously edit multiple content which led to the development of the matrix editor.

Data & methods

In the present study, all issues that were raised via the issues tracker (296) in a nine months period between October 2010 and June 2011 were evaluated. This period closely succeeds the start of the issues tracker in September 2010. Additionally, the email help requests sent to the general Scratchpad 'contact us' email address (58 requests), or directly to the lead developer (56 requests) and the user support manager (127 requests) were evaluated as well as some of the messages (10 requests) sent directly to other Scratchpad team members (see Appendix 1 for Scratchpad development team roles).

For the issues a matrix was exported from the issues tracker that included the issue ID, date of creation, user name, Scratchpad URL, request category, number of comments, and date of first reply.

Emails were exported from the respective software into a matrix that included the email address of the sender as well as the receiver, date of creation, and the subject and content of the email. In order to be able to compare emails with issues, all emails were sorted into initial request emails (equalling an issue) and replies to these initial requests (equalling issue comments). Each initial request email was given an ID and the Scratchpad URL, request category, number of comments, and date of first reply was was deducted from the text and date of the request email and its replies.

For both systems the number of days until an issue was first replied to was calculated. Additionally, all emails and issues were labeled as posted either by a "user" or a "developer". For the purpose of this paper all Scratchpad team members, including those involved in support roles, are regarded as "developers" and all Scratchpad users that are not part of the team as "users" regardless of their professional background.

There are three request categories: bug reports are posted if certain features of a Scratchpad don't work the way they are supposed to work; support requests are posted if a user does not know how to proceed, if he/she would like help in setting up a site, or if he/she would like changes in the deeper structure of his/her own Scratchpad for which he/she does not have the permission; and feature requests are posted if a user would like additional features or functionality added to the Scratchpads as a whole.

Results

From October 2010 to June 2011, the email service and the issues tracker together facilitated 547 co-learning opportunities between users and developers. Persons who posted issues worked on 43 different Scratchpads, which is 17.3% of all Scratchpads (249 on 30 June 2011). Persons who sent email requests came from 72 different Scratchpads, which is 28.9 % of all Scratchpads. 27 requests were sent from persons without a Scratchpad at the time of emailing.

Request categories

Both support systems taken together, about half of the requests were support requests, followed by about a quarter bug notifications and one fifth feature requests (Tab. 1).

Table 1. Overview of requests by category and support system. Number of requests posted by users and developers, by request category and by support system (October 2010–June 2011).

Request category	Issues	Emails	total
bug	116 (39.2%)	32 (12.7%)	148 (27.1%)
support	77 (26.0%)	211 (84.1%)	288 (52.7%)
feature	103 (34.8%)	8 (3.2%)	111 (20.3%)
total	296	251	547

There is a significant difference in which system was used for which kind or request. The issues tracker was clearly the preferred system for bugs (79.4% of bugs were posted as issues) and features (92.8%), but not for support requests (36.5%). However, there is some overlap between the two systems as sometimes requests moved from one to the other support system: Five emails were follow ups from the issues tracker (all support requests) and 15 issues were posted as a result of email requests (11 support requests, 3 bugs and 1 feature request).

Pattern of requests over time

In the analysed period 296 issues were raised via the issues tracker (Fig. 1). That is 32.9 per month with a peak in November 2010 due to the follow up for a training course that resulted in many new feature requests, but also in bugs and support requests. In the last three months of the analysed period, the number of issues posted was less, partly because of a drop of issues posted by the developers.

In the same time period a total of 251 email requests were posted meaning an average of 27.9 requests per month (Fig. 2).

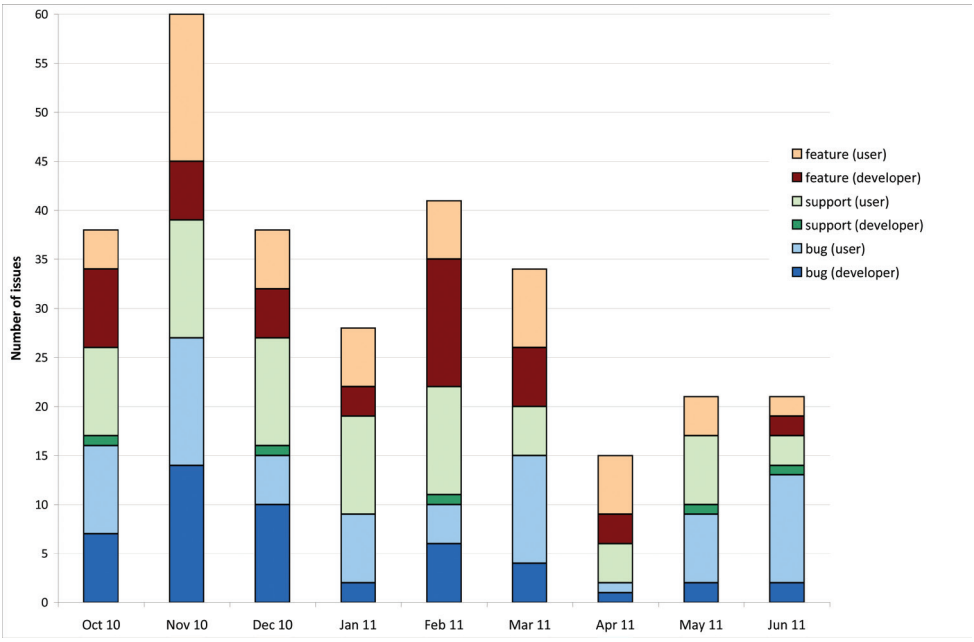


Figure 1. Pattern of issues over time. Number of issues per month divided into request category and each category divided into issues posted by users versus developers (October 2010–June 2011).

The number of all requests posted to both support systems per month in the analysed period is not related to the number of Scratchpads (see Fig. 1 in Smith et al. in this volume) nor to the number of (active) users. However, the number of support requests per month seems to reflect the number of new Scratchpads in the latter part of the studied period (March-June 2011), though not in the earlier part (Fig. 3).

Pattern of requests by Scratchpad

On average users posted 5.5 requests per Scratchpad (417 requests, 76 Scratchpads). For 22 Scratchpads, five or more requests were posted. Half of these sites were created more than a year before the requests were posted and eight were created shortly before the requests were posted. This pattern is the same if only support requests are considered. So most requests including most support requests are posted by more experienced users.

The pattern of requests over time for individual Scratchpads shows that requests are posted in phases, with periods of high activity alternating with periods of little or no activity (Fig. 4). If requests would have been constant over time, the graph would have depicted straight parallel lines. Instead, large areas in one colour indicate a high

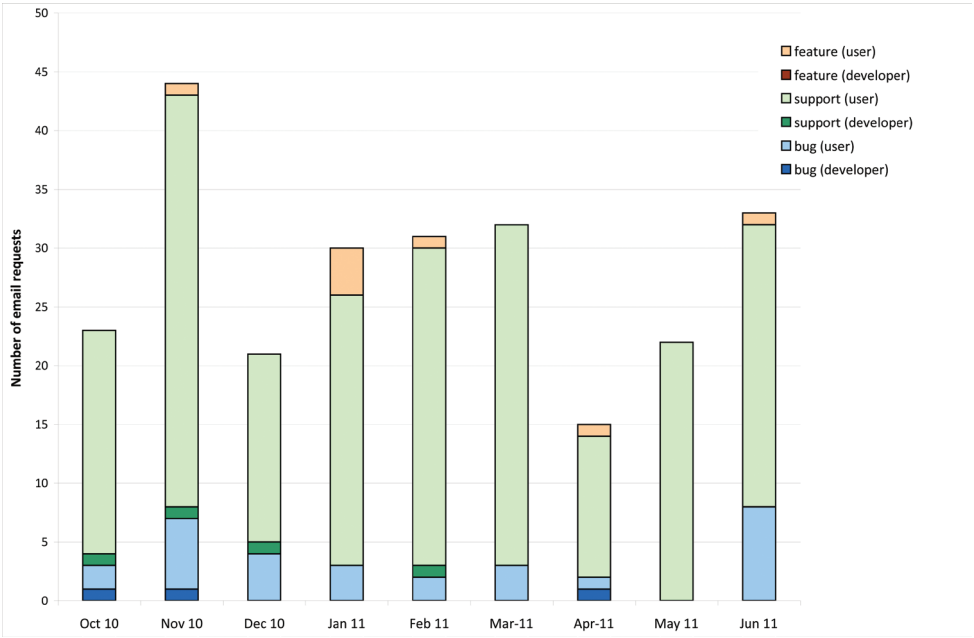


Figure 2. Pattern of email requests over time. Number of email requests per month divided into request category and each category divided into emails sent by users versus developers (October 2010–June 2011).

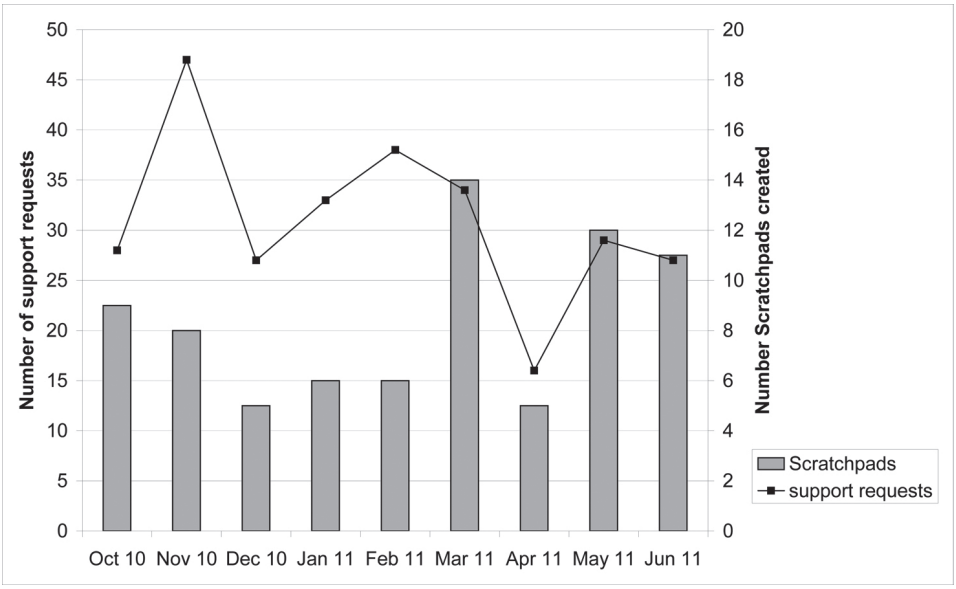


Figure 3. Pattern of support requests and of new Scratchpads over time. Number of support requests posted to both support systems by users and number of new Scratchpads created per month (October 2010–June 2011).

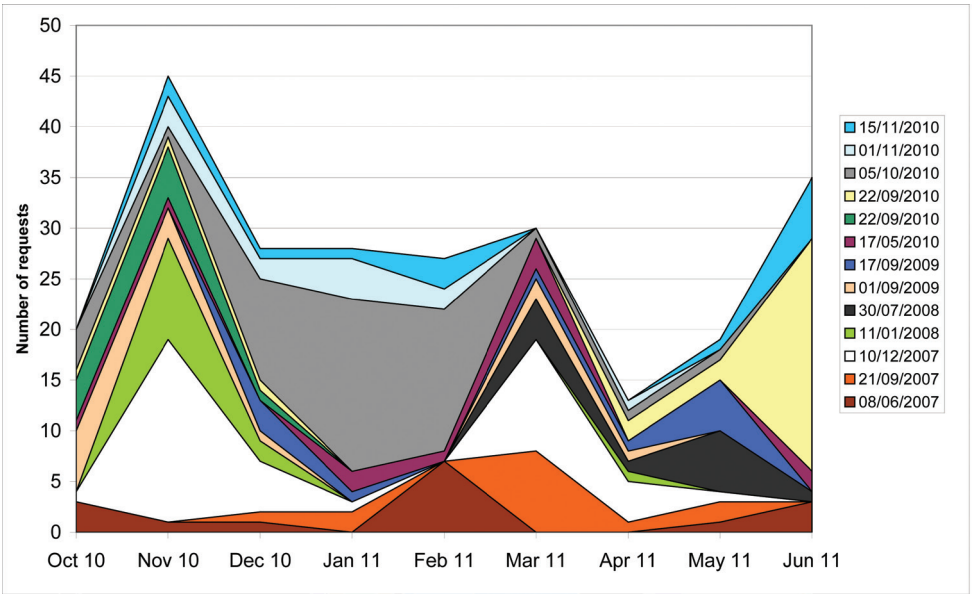


Figure 4. Pattern of requests over time by Scratchpad. Number of requests posted per month (October 2010–June 2011) for Scratchpads for which ten or more requests were posted during the analysed period. In the legend the creation date of the individual Scratchpads is cited.

request activity on the respective Scratchpad for that time period and small or missing areas indicate low or absent activity.

User support system preference

When evaluating the 22 users that posted at least five requests, the results show that 15 (68.2%) created more emails than issues, whereas only 7 (17.8) created more issues. Most users used both systems, but 5 (22.7%) clearly prefer using the issues tracker (more than 80% of their requests are posted as issues) and 5 clearly prefer to send emails. Three of the latter did not post a single issue even after having been encouraged to do so.

Difference between the use of the different support systems by users versus developers

Issues were raised by 49 different persons with an average of 6.0 issues per person. A third (17) of the persons raised only one issue. Six of the persons are developers. However, these six developers posted a significantly higher number of 99 (33.4%) issues (16.5 issues/developer), though it has to be taken into consideration that some of these

issues were originally raised by users via email and later on posted by developers to the issues tracker. Out of 296 issues, only 197 were raised by users.

Users and developers also differ in the number of issues posted as different request categories. Support requests are nearly exclusively (93.5%) posted by users, whereas developers posted slightly more bug reports (58.6%) and feature requests (55.3%).

Email requests were sent by 95 different persons with an average of 2.6 emails per person. More than half (57) of the persons sent only one email request. Three of the persons are developers. Only 7 (2.8%) email requests were sent by developers whereas 244 emails were sent by users.

Request processing amount

On average 3.0 comments were posted per issue and 3.1 per email request (Tab. 2). Comments are posted by developers as well as users and often represent a discussion thread. In both support systems most comments were posted for support requests (4.0 for issues, 3.3 for emails) and least for feature requests (2.0 for issues as well as emails).

Table 2. Number of comments by support category. Number of comments posted by developers and users to the two different support systems by request category (October 2010–June 2011).

request category	number of requests [issues/emails]	range of comments [issues/emails]	number of comments [issues/emails]	average number of comments [issues/emails]
bug	116/32	0-12/1-9	367/76	3.2/2.4
support	77/211	1-14/0-29	308/686	4.0/3.3
feature	103/8	0-12/1-5	208/16	2.0/2.0
total	296/251	0-14/0/29	883/778	3.0/3.1

Request processing time

182 issues were replied to the same or the following day. 58 issues were replied to within 2-7 days, 25 within 8 to 30 days, 14 after 30 days and 17 have not had any replies by the end of the analysed period (Fig. 5). With "days", week days are meant, not work days, so within the 2–7 days range issues are included that were replied to the following work day.

Comparing the response rate to issues posted by users versus developers, it becomes obvious that user issues are replied to faster, which is especially true for feature requests and bugs. The major part of the requests that have not been replied to within the analysed period were feature requests posted by developers.

220 email requests were replied to the same or the following day (Fig. 6). 25 requests were replied to within 2-7 days, 5 within 8 to 30 days, 1 after 30 days and none have not had any replies by the end of the analysed period.

Comparing issues versus emails, only 61.5% of issues were responded to the same day but 87.6% of emails. Therefore emails are replied to faster than issues.

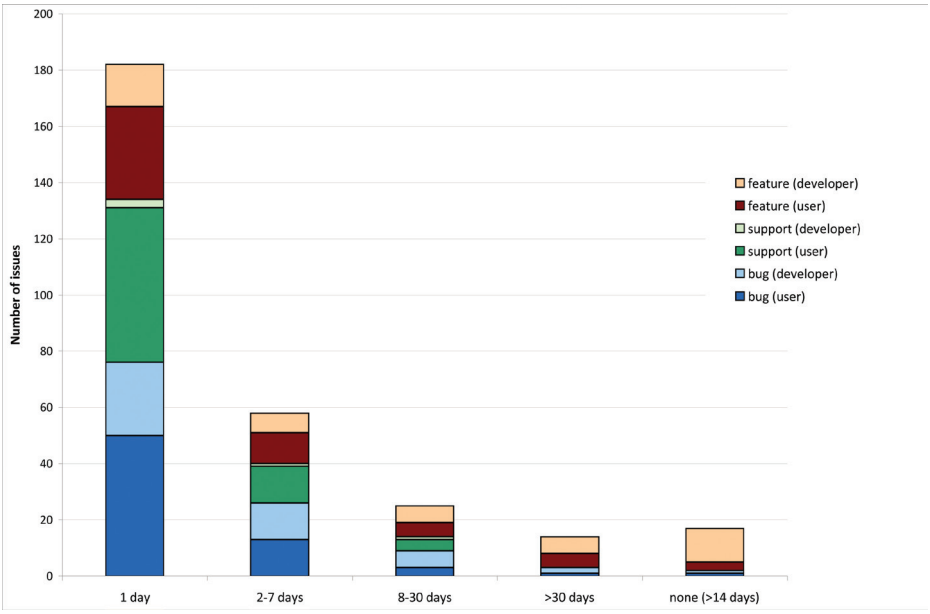


Figure 5. Request processing time for issues. Time lapse between posting of issues and the first reply to this issue divided into request category and each category divided into issues posted by users versus developers (October 2010–June 2011).

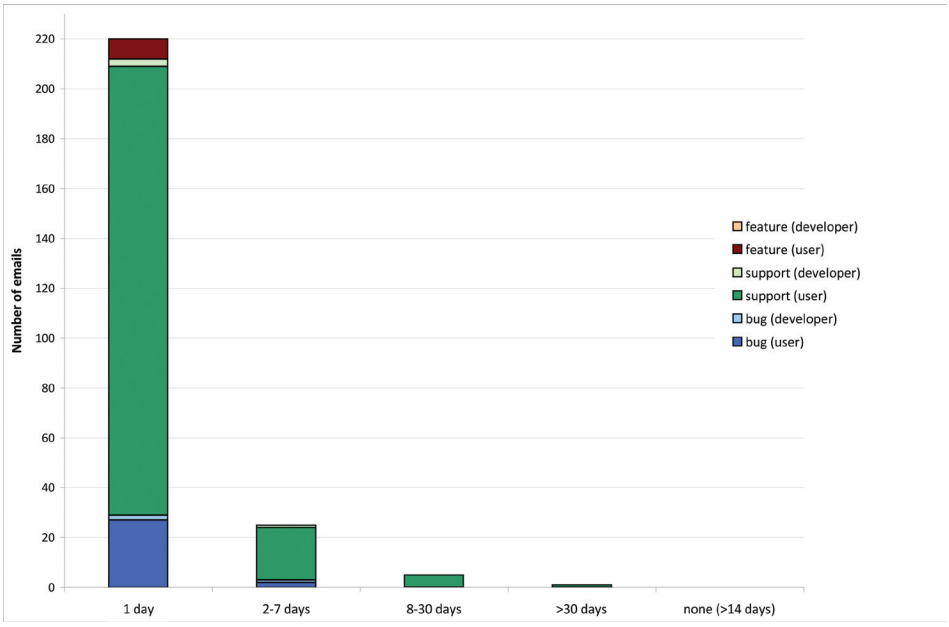


Figure 6. Request processing time for email requests. Time lapse between posting of an email and the first reply to this email divided into request category and each category divided into emails sent by users versus developers (October 2010–June 2011).

Conclusions

In this paper we explored the presence of co-learning opportunities for users and developers in the Scratchpad environment with its various support services and aimed to get a better understanding of the process and mechanisms behind their use. We analysed the usage data of the "request emails" and the "issues tracker". The results show that the support email service and the issues tracker facilitated 547 co-learning opportunities between users and developers. We think that each request offers a co-learning opportunity because on one hand questions asked by the users are answered by the developers and ways to solve a problem are explained. On the other hand the developers learn about problems with the system and how users work with the system, thus enabling them to improve the Scratchpad system. As a consequence, in early 2012 a new Scratchpad version will be released featuring many enhancements.

Request categories: The two support systems are used for different kinds of requests. The issues tracker is the preferred system for bugs and feature requests whereas sending an email is the preferred way for support requests. This reflects the more private nature of support requests, which concern only one Scratchpad, whereas bugs and feature requests usually concern all Scratchpads.

Pattern of requests over time: There is a relationship between the number of support requests posted and the number of new Scratchpads per month in the latter part of the analysed period. Assuming that users need more support starting a Scratchpad than later, this would explain the seeming correlation between the number of support requests and number of new Scratchpads in the latter part of the analysed period. The discrepancy in the first part of the analysed period can be explained by the presence of training courses in November (2x), December, January and February, which generated additional support requests due to renewed Scratchpad activity of the training participants. Although we can speculate why these discrepancies happen, further research is needed to better understand what exactly triggers these fluctuations.

Pattern of requests by Scratchpad: Requests are posted at various times during the life time of a Scratchpad, not just directly after it was created, even though users often need help in the first months after registering for a new Scratchpad. Usually requests are posted in phases of high activity alternating with low or no activity. Thus it is difficult to predict when periods of higher activity can be expected. The vast majority of requests usually occurred when funding for a person to work on the Scratchpad was available resulting in an extensive use and development of the respective site.

User support service preference: The decision on which system to use depends not only on the kind of request (see above), but also on the personality of the user. Most users prefer to write emails, though the number of users that very clearly prefer one system over the other are the same for issues and emails. The reason behind the preference of emails could be that the emails are not published and therefore the bar-

rier to pose what is perceived as "stupid" questions is lower. After the initial contact via the 'contact us' email, when a personal contact has been established to a developer it is also for many users more natural to ask questions of this person than to post a request to the more anonymous issues tracker.

Difference between the use of the different support systems by users versus developers: An additional factor that influences the way the two support systems are used for different kinds of requests is the role of the person posting the request. Support requests were mostly posted by users, whereas bug reports and feature requests were posted both by users and developers. This reflects the fact that the developers use the issues tracker to keep track of bugs and ideas for new features. Also, often the results of testing of the Scratchpad system as well as problems that arise in other parts of the user support (e.g. training courses, demos) are transferred to the issues tracker. Further research is needed to better understand under what conditions users and developers decide to use the email service and when the issues tracker.

Request processing amount: There is a wide range in the number of comments/replies posted in answer to a request. Some requests can be easily fixed and therefore only require one comment notifying of the fix. However, in many cases a request needs to be discussed. Support requests require the most comments/replies because it is often necessary to first get a clear picture of the problem and then to develop customised solutions for which more engagement with the user is needed. This process could be abridged especially for support requests and bug reports by saving the page the user was viewing when entering the issues tracker. This would enable the developer dealing with the request to grasp the problem quicker.

Request processing time by system: The time until a request is taken up by one of the developers is different for the two support systems: Emails are replied to much faster than issues. This is partly due to the fact that it takes up to one hour for notifications about new issues to reach the email account of the developers. Emails can also be answered quicker, because it is only necessary to hit the reply button, whereas if a developer receives a notification for an issue, he/she needs to log into a Scratchpad first, go from there to the issues tracker and find and open the correct issue. The process of replying to an issue could be made faster by sending out the notification immediately after a request has been posted and by improving the log in options for developers. Another reason why emails are replied to faster is because most are support requests, which are easier to fix because they usually don't involve any changes to the Scratchpad system, but just changes to the structure or layout of individual sites.

Request processing time by role: An additional factor that influences the time until a request is taken up by one of the developers is the role of the person posting the request. Requests from users are replied to faster than those from developers. This is mostly due to the fact that developers post bug and feature request on the issues tracker for archiving purposes and these requests don't require immediate attention because they already have been discussed in developer meetings.

Summary

Based on this study we now have a better view of how two support systems are used by users and developers of Scratchpads and can develop several recommendations for further improvements of the support services itself and the way they are used.

The results underline the importance of offering two support systems, a public system (issues tracker) as well as a private one (emails), to cater for different user mentalities as well as for different categories or requests. A possible advantage of email, privacy of communication, might be important for certain users, but emails are difficult to track for the Scratchpad developers team. Therefore a system should be created whereby emails can be logged into an area of the issues tracker that is private to the Scratchpad team and reply messages should be sent from this area.

Storing the issues tracker and the request emails in one place is also important because they hold a wealth of information on the Scratchpad system. Currently, this information is distributed over several different email archives and the issues system and thereby not accessible to all developers. Having only one archive and tagging all items with keywords would facilitate later tapping of all data. For consistency this tagging should be done by the team.

Although we now have a better view of the presence and process of co-learning opportunities our data did not tell us if actual learning between the two parties has occurred because we did not analyse the content of comments and replies to the requests. Further research using e.g. using survey methods is needed to explore this matter in more depth.

Within the ViBRANT project, networking activities such as workshops, peer-based training courses and cascade training are designed to enhance the use of Scratchpads and to develop a network that will foster long-term sustainability of the user community. Sociological studies of the Scratchpads' user-base will underpin software development priorities and maximise engagement in the Scratchpads' community.

References

Bratitsis T, Dimitracopoulou A, Martínez-Monés A, Marcos J, Dimitriadis Y (2008) Supporting members of a learning community using Interaction Analysis tools: The example of the Kaleidoscope NoE scientific network. International. Conference ICALT2008, Santader, Spain.

Colazzo L, Molinari A, Villa N (2008) From e-learning to "co-learning": the role of virtual communities. In: Kendall M, Samways B (Eds) IFIP International Federation for Information Processing 281: 329–338.

Crowston K, Howison J (2006) Hierarchy and centralization in free and Open Source software team communications. Knowledge, Technology, & Polio 18(4): 65-85.

Schuler D, Namioka A (1993) Participatory design: principles and practices. L. Erlbaum Associates Inc. Hillsdale, NJ, USA.

Smith VS, Rycroft SD, Harman KT, Scott B, Roberts D (2009) Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. BMC Bioinformatics 10 (Suppl 14): S6 doi: 10.1186/1471-2105-10-S14-S6

Smith VS, Rycroft SD, Brake I, Scott B, Baker E, Livermore L, Blagoderov V, Roberts D (2011) Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 53–70. doi: 10.3897/zookeys.150.2193

Von Hippel E (1994) „Sticky information" and the locus of problem solving: implications for innovation. Management Science 40(4): 429-439.

Von Hippel E, Tyre MJ (1995) How learning by doing is done: problem identification in novel process equipment. Research Policy 24(1): 1-12, doi: 10.1016/0048-7333(93)00747-H

Von Hippel E (2005) Horizontal innovation networks - by and for users. Industrial and Corporate Change 16(2): 293-315, doi: 10.1093/icc/dtm005

Wagner J (1997) The unavoidable intervention of educational research: a framework for reconsidering researcher-practitioner cooperation. Educational Researcher 26(7): 13-22, doi: 10.3102/0013189X026007013

Appendix 1: Scratchpad support structure

Scratchpad development team

After a conceptual period of several month, the development of the Scratchpad system started with the hiring of a full-time developer in December 2006 and the first Scratchpads were created in January 2007. For the first two years, user support was provided solely by the developer and the project leader, but gradually more experienced maintainers also provided support to colleagues. In January 2010 a user support manager, who is responsible for the help desk, training and the help system, became part of the Scratchpad team. At the time of writing (July 2011) the team consists of the project leader, three developers, and three user support staff (some only part time).

Support systems

To help with the use of Scratchpads, the first **screencasts** were published in July 2007, followed by **FAQ** in May 2008. Both were available on www.scratchpads.eu. The screencasts were difficult to keep updated and had to be taken down in 2010. In January 2011, the FAQ were replaced by a **help system** that is integrated into the Scratchpads, so that users can find help pages directly on their own site. This help system also contains the manuals for the training courses (see below), providing a more task centered approach than the help pages, which are mostly feature specific.

A mostly empty Scratchpad template, the **sandbox** (http://sandbox.scratchpads.eu/), has been in use since October 2007 to allow Scratchpad maintainers and users to practice using Scratchpads. The sandbox is rebuilt every 6 hours.

The **help desk** was formally started with the appointment of a user support manager in January 2010, but the Scratchpad 'contact us' email (scratchpad@nhm.ac.uk) has been active since about August 2008. The help desk deals with all the emails, issues, calls and meetings relating to user support.

As the first feedback system, **UserVoice** (www.uservoice.com) was used together with the EOL Lifedesks (http://www.lifedesks.org) from August 2009 to about April 2010 but was not embraced by the users. This was followed in September 2010 by the current **issues tracker** (see main text).

In order to support and extend the user communities working with Scratchpads, basic and advanced **training courses** are organised. These one-day courses are free of charge and are intended to help current and prospective Scratchpad owners to develop their site building skills, to learn best practices and gain a better understanding of what Scratchpads can do to support research communities.

The training **manuals** are available on the Scratchpads website and have recently been integrated into the help system on each Scratchpad. This allows users to follow the instructions on their own Scratchpad. Additionally, the opportunity is given to do self-training on a **home training site** that can be provided by the Scratchpad development team.

To inform users about new features, bug fixes, training courses, etc. a regular **blog** has been running since January 2010 on the Scratchpads website.

ZooKeys 150: 193–210 (2011)
doi: 10.3897/zookeys.150.2164
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Studying the effects of virtual biodiversity research infrastructures

Daphne Duin, Peter van den Besselaar

VU-University Amsterdam, Dep. of Organization Sciences. & Network Institute De Boelelaan 1081, 1081 HV, Amsterdam, The Netherlands

Corresponding author: *Daphne Duin* (d.duin@vu.nl)

Academic editor: *V. Smith* | Received 29 September 2011 | Accepted 25 November 2011 | Published 28 November 2011

Citation: Duin D, van den Besselaar P (2011) Studying the effects of virtual biodiversity research infrastructures. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 193–210. doi: 10.3897/zookeys.150.2164

Abstract

The research environment of scholars is increasingly web-based. This makes it urgent to study the effects of moving to the Web on research practices, scholarly output and innovation. We propose a theoretical framework and a methodology to study these effects. In a pilot study, we apply theory and method on an online community in biodiversity research, to demonstrate the feasibility of the approach. We also indicate the practical relevance of this kind of analysis for improving the quality of virtual research environments. In the last section, directions for further research are suggested.

Keywords

Knowledge creation, users, scientific collaboration, Webscience, impact assessment, biodiversity and taxonomic research infrastructures, virtual communities of practice

Introduction

Moving science to the (social) Web has generated excitement for its potential to support knowledge creating activities in research environments (eResearch2020 2010). Core ideas behind social web applications and services are: to make the Web a place for user generated content; to harness the power of crowds; to provide access to data on a large scale; to offer an architecture for participation and to create network effects and openness (Tim O'Reilly in: Anderson 2007 p.14). The social Web, also called Web

2.0, brings people, ideas, tools and information resources together and is in this way a promising means to accelerate scientific developments and to disseminate scientific information for policy and education.

In the field of biodiversity research numerous Web based tools are currently available. The tools facilitate knowledge creation within the global expert community of biodiversity researchers and bioinformaticians. The tools allow users to do collaborative work on the Web. Users can create content, share data and have access to knowledge that was once only available to individual researchers, whether in paper achieves, on stand-alone computers or in difficult to access data systems of their institutions. Several of these kind of tools are supported under the 7th Frame work Programme ViBRANT. A sum of such tools and online services is referred to as Virtual Research Environments (VREs), a cyberinfrastructures or e-infrastructures (Fraser 2005). These concepts are continuously evolving and often used interchangeably. The different terminologies have in common that they comprise digital infrastructures and services which enable research to take place (idem). Even though the expectations on the impact of Web-based science are high, most virtual research environments, being relatively new, struggle with engaging user communities and with the implementation of a sustainable model.

Within the context of a larger trend to move biodiversity to the Web (see also: Global Biodiversity Information Facility GBiF; Encyclopedia of Life EOL; Biodiversity Heritage Library BHL; ViBRANT), we are interested in the effects that the move to the Web has on researchers' work environment, research practices, scholarly output and the changing needs for support (see also JISC Virtual Research Environment programme). A better understanding of the effects will contribute to: i) a design and management that better fits the needs of the users; ii) improving sustainability; iii) and more generally to research on infrastructure policy.

In this paper we will zoom-in on the questions mentioned above. Our main aim is methodological. We will elaborate a method for studying the effects of web-based biodiversity research infrastructures on scientific collaboration, innovation, and performance. In what follows we will put forward a theoretical framework, discuss empirical data and a methodology - which we think will help in answering the question. To illustrate the possibilities and limitations of the methodology suggested we will discuss empirical data that we collected for a pilot study on one online community of the Scratchpad platform [http://scratchpads.eu/] and conclude with recommendations for further research. Scratchpads are an online platform for collaborative and distributed work in biodiversity research. The Scratchpad environment is currently one of the more established services that is coordinated under the ViBRANT FP7 umbrella and is in the air since 2007.

Here we stressed why it is important to examine the effects of moving science to the Web. In the following paragraph we bring together previous research on the organisation of knowledge creation and discuss how we think we can use these findings in our own work.

Knowledge creation

Knowledge creation is at the heart of the academic profession. Influencing the creation of new knowledge is a challenge for organisations as knowledge flourishes best when it is enabled, not managed (Sveiby 2001). Key concepts for understanding knowledge creation are "implicit knowledge" and "explicit knowledge" as put forward by Nonaka et al. (1994, 1995). Implicit knowledge is experience based and context specific knowledge that cannot be expressed in words, sentences, numbers or formulas. This also includes cognitive skills such as beliefs, images, intuition and mental models as well as technical skills such as craft and knowhow. Explicit knowledge is codified, general knowledge that can be expressed in words, sentences, numbers or formulas. It includes theoretical approaches, problem solving, manuals and databases (Nonaka 1997). According to the authors, the answer to mobilisation and creation of knowledge is to enable interaction and the exchange of implicit (tacit) and explicit (codified) knowledge (Nonaka et al. 2000). Woo et al. (2003) and Herschel et al. (2001) emphasise that converting implicit knowledge to explicit knowledge is often seen a problematic task, labour intensive and expensive. One solution to overcome this problem is the creation of Communities of Practice (CoP). These CoPs bring together knowledgeable experts to work on complex problems (Andriessen 2005; Wenger 1998). This relates also to Sveiby's (2001) observation that implicit knowledge is best kept in knowledgeable people and is achieved by making knowledgeable people communicate. According to him: "knowledge shared is knowledge doubled" (p. 347). McFayden et al. (2009) add to this the importance of combining diverse and overlapping knowledge inputs between exchange partners for the creation of new knowledge. Overlapping knowledge allows for greater specialisation and support in CoPs because a common knowledge base (e.g. mental frames, shared knowhow) eases communication (Demsetz 1991). On the other hand, heterogeneous or sparse networks provide more opportunities to secure access to new information and diverse perspectives (Burt 2001). In other words, a CoP needs a common basis of implicit and explicit knowledge for good communication flows and stability but also diversity in order to be innovative and flexible.

Next to the contributions from knowledge management studies on innovation, also social network studies have also contributed important insights to our understanding of the conditions and constrains for knowledge creation. Scientists, like other professionals, bring more to work than skills and experience, "they also bring the assets they can procure through their social networks" (Gargiulo et al. 2000: p. 183). This is often referred to as "social capital" (Bourdieu 1980; Coleman 1988 in: Gargiulo et al. 2000 p. 183; Burt 2001, 2007). Burt demonstrates that "compensation, positive performance evaluations, promotions, and good ideas are disproportionately in the hands of people whose networks span structural holes" (2004, p. 349). Structural holes are non-redundant connections between actors in a network. In other words, structural holes are ties to people that are themselves not connected. People in an organisation who connect not connected groups are called "brokers".

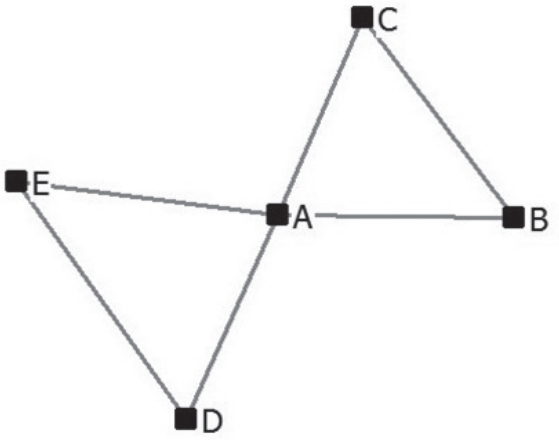


Figure 1. Brokerage and structural holes.

Social network studies make use of sociograms to support their analysis. These are graphic representations of social links that a person has. Figure 1 is an example of a sociogram of structural holes that are linked by of one actor in a network, represented by node A at the centre of the graph.

The nodes are actors, the ties their connections. Actor A is a broker in this network because connects two groups that are otherwise unconnected (spans structural holes).

Social network studies show that brokers are valuable individuals for organisations. Brokers are people who have the capabilities to “”translate, coordinate and align between different perspectives (…) and address conflicting interest” (Wenger 1998: p. 109). Moreover brokers are more likely to express new ideas and to have them judged valuable (Burt 2004). This idea of “selection and synthesis across structural holes and between groups is not new” (Burt 2004: p. 350). Hence, most structural holes studies were carried out among local based workers (e.g assembly line workers of the same factory). What we aim to study is how this functions in online (distributed) research communities.

Virtual research environments like Scratchpads aim not to replace existing data and communication systems but, rather offer additional ways of working with existing facilities. They add an additional organisational and network layer to the traditional work environment of a researcher. Researchers already participate in multiple professional and personal networks such as: at the level of their department; the institutions; national/international projects: alumni networks; advisory boards etc. Becoming a member of an online work group would add another network layer to their organisation of work. We would like to argue that to be able to study the effects of moving biodiversity online we have to take into consideration already existing structures of the researchers work environment and investigate to what extent these change when a new way of working is adapted. Hence, instead of looking at uniplex networks a study of multiplex networks will be helpful in getting a deeper

understanding of how the introduction of a new network layer might change exiting organisational structures. As Lee et al. demonstrate "multiplex networks involve multiple relations that create multiple ties in one network have been shown to influence the formation or dissolution of ties in other networks" (2011, p. 759).

Today, online networks are important vehicles for knowledge sharing and learning in the workplace (Ardichvili 2008). The expectation is that the social Web provides enabling conditions for knowledge creation as mentioned above. The social Web overcomes a number of barriers for knowledge exchange and interaction - by giving distributed communities the tools to control the level of openness of their communication and tools to simulate a face-to-face setting with help of online instruction videos, VOIP, document- /image- / biography- sharing tools, forums and other layouts of online communities. Triggered by the developments of Web 2.0 tools, the playing field of CoPs moves to the Web, which turns them into Virtual Communities of Practice (Samarah et al. 2008). The claims about the usefulness of Web 2.0 tools for knowledge creation are often made. However, there is a surprising dearth of empirical studies that show the impact of Web 2.0 tools on knowledge creation in virtual communities, with the exception of work by Samarah et al. (2008).

In summary, knowledge flourishes when knowledgeable people are brought together and interact. Especially the exchange of different but partly overlapping knowledge enables the creation of new knowledge within expert communities. Another important enabling condition that arises from the literature is the amount of social capital of individual actors as well as social capital kept within collective working groups (teams, labs, departments). The open question to be studied is whether Web 2.0 tools, such as Scratchpads, do provide these conditions. This is something to be studied.

In this paper we will discuss a pilot study that examines the possibilities of a social network approach to study co-authorship and Scratchpad membership. But before we come to discuss our pilot study we will investigate the challenges of studying online social settings. Web data are still a relatively new empirical data source in the social sciences and there is some debate on how to collect and interpret data sets collected from the Web. In the following paragraph we will discuss some of the pros en cons of the use of web data to study organisation(s).

Virtual communities of practice

The Web has become a major medium for communication in science. ViBRANT products and services, currently being developed under the 7th Frame Work programme, mirror a trend within biodiversity research moving science to the Web. In this paper we concentrate on one of the products of ViBRANT, the Scratchpads. Scratchpads are an online platform for biodiversity research where virtual communities of academic experts link remote resources together (people, biographies, images) and offer an environment for learning and knowledge creation that before was only possible in geographical proximity (Smith et al. 2009). The platform has today a global user

community of > 3000 people, which is steadily growing. The communities are created around different biodiversity research topics, such as around a particular group of organisms, around a project, or bioinformatics topics. Scratchpad communities are managed by individual researchers that apply for a site and can invite /or make it open to fellow researchers to register and participate in content sharing and analysis. Some sites are communities-of-one, other sites have more than 200 registered users. Also the level of activity among users of one Scratchpad may vary significantly. The content creation of some sites is the effort of a single researcher, while the other users of the site take a more "passive" role. For other Scratchpad sites the whole community is actively engaged in the creation of content. What all users have in common is that at one point they decide to register as a user of a community. They either saw an interest in connecting to the other users or to the content of the site, they identified with the people, the content or with both.

From a previous study that we did among the users we know that Scratchpads are used among biodiversity researchers mainly: to disseminate research results; to share data; to collaborate in the writing of project proposals and papers; and for preparing meetings (Smith et al. 2010 p.4). In the field of organisational knowledge creation Scratchpads can be coined Virtual Communities of Practice.

Virtual Communities of Practice are a type of knowledge based social network whose members rely primarily on networked ICT's in order to 1) discuss problems and issues associated with their day to day activities 2) collaborate on projects 3) share documents, solutions or good and bad practices, plan for face to face meetings or continue face to face relationships and work beyond face-to-face events (Anandarajan and Anandarajan 2010, p.154)

The move of science to the Web leads to new questions regarding the impact of the online environment on scholars' behaviour, relations, and scholarly output. It also provides us with new types of data and methodologies. The Web is constituted by a myriad of socio-technical interactions which often leave digital traces. Users of the Web leave digital footprints of their behavior and network relations. Their footprints can be found in web server log file data or in the information that is stored on institutional web pages and social network sites. Consequently "it forms an interesting, modern site for research ethnography" (Beaulieu 2005, p. 183) and for quantitative studies of these digital traces (cf. Thelwall 2010). Today also offline activities can be studied from the Web as personal information about researchers' work is disseminated widely online, in publication databases, conference websites and sharing tools for presentations and images, just to name some examples.

The use of such data sets for social research, like the digital footprints of researcher's online activity, has several advantages. Firstly, the scale on which we can do research becomes much larger, as one can collect large datasets covering the actions of many users and over long periods of time and geographical distances. Secondly, research using such data is unobtrusive as the actors under study are not interrupted in their work by data collection activities. Thirdly, the data are observational, and not based on opinions

only, such as in surveys and interviews. Fourthly, costs are potentially lower, as web data can be collected from behind a desk and are often freely available (cf. Johns et al. 2004). Hine (2005) describes the use of the Internet data for social research as a trend where excitement and anxiety come together. Some of the advantages are mentioned above. The disadvantages of the use of web data relate to lower responses rates of online surveys; non-representative of the sample, a decreasing quality of the data, and privacy issues. As a consequence, it is often argued that (secondary) web data is best used in combination with other, primary data, to control for issues such as representativeness (Buckman 2006). Something we plan to do in future research. In the next section we discuss the methods that we used for analyzing web data.

Pilot study, data and methods

As argued, the creation of new knowledge can be enabled by bringing a variety of knowledgeable people together in an environment that facilitates the interaction and exchange of heterogeneous and overlapping knowledge inputs. We carried out a pilot study on one Scratchpad community to explore this question and test our approach. The research questions are: 1) to what extent do Scratchpads connect people that were not connected before (as co-author)? 2) To what extent do Scratchpads create new links between different bodies of knowledge (structural holes) and reinforce existing links?

For both questions we build on ideas and techniques stemming from bibliometrics (cf. Glänzel 2002) and Social Network Analysis (cf. Wasserman and Faust 1994). In the literature section above we explained why Social Network Analysis offers a useful framework. In order to answer the first question we compare the offline, traditional collaborative network connections of the Scratchpad users, their co-author relations, with Scratchpad membership. In other words, we are interest to know who connects to whom because of membership who was not connected before by co-author relations. Or, were all users already connected before they joined and is the Scratchpad only a different media to continue to work with people one already used to collaborate with? Co-author relations are valuable measure for academic collaboration but should be handled with care (Glänzel 2002). Also, co-authorship is certainly not the only form of collaboration in science. People are part of multiplex social networks (e.g department, institutions, editorial boards). Each network has its own type of interactions (drinking coffee, talking in meetings, peer reviewing on the same journal). The combined number and type of network connections and interactions has an effect on someone's social capital. Sometimes networks overlap, you meet the same people in different settings taking on different roles. But sometimes new networks do not overlap and fill "missing links" in one's social capital. In our data example we stack two networks on top each other: the co-author network and the Scratchpad membership network, and study to what extent the Scratchpad membership connects researchers that were not already connected by co-authorship ties.

The second question deals with the potential of the Scratchpad community to create new knowledge (span structural holes) and create favorable conditions to continue to exist over a long period of time (redundancy). For this question the Scratchpad community is taken as an analogy for a research team where every member brings in social capital in the form of their co-author network. This time we did not look at the co-author relations among the 11 members but to what extent their ego, co-author relations overlap. Do the Scratchpad members co-author with the same peers, or do they bring in their personal, unique co-author contacts?

Our case is a single Scratchpad which we give here the fictional name Livingcreatures.info. User registration coming from automated bots, so called spam-signs ups, were excluded from the analysis. The member list that we used included members' personal details such as their affiliation and was used as the starting point of studying co-author relations. For each member we collected their publications over a period of 10 years, preceding their online collaboration in Livingcreatures.info. Publications were searched for and downloaded from the ISI Web of Science database and combined with publications from Google Scholar (using Publish or Perish). The Web of Science is a much more structured database, with for instance better name ambiguity filters than Publish or Perish. However the combination of both was thought important as biodiversity research is underrepresented in the Web of Science (cf. Krell 2002). Therefore we needed to complement this with publication information from additional data sources. Publish or Perish uses Google Scholar data, with a much wider coverage. The resulting publication lists cover journal articles, books chapters, series and peer reviewed and non-peer reviewed papers. In the next step we retrieved the co-author relations of each Scratchpad member and this enables us to study to what extent the co-author relations of the members overlap. This indicates the degree of differences and similarities between the types of knowledge represented in Livingcreatures.info.

The Scratchpad under study was launched early 2011. As for August 2011, this Scratchpad has 11 registered members, all male. Ten of the members were in the period of our analysis affiliated to one of the natural history institutions in the world, number 11 is mentioned in the acknowledgements as a private taxonomic specialist. Their institutional addresses are located in five different continents (3 in Europe, 3 in Asia, 2 in Africa, 1 in South America, 1 in Oceania). Together, these 11 Scratchpad members have 187 co-authors (including inter-group relations) with whom they collaborated in the period from 2001-2010. Table 1 gives the breakdown of the publications of the group. Together they contributed to 135 publications in ten years. Four Scratchpad members have no co-author relations. The information that we found during our web search suggests that this may be explained by individual characteristics. From the web data we learnt that two of them are early career researchers, number three is in a non-research position in a research institute, and number four is a volunteer researcher. Table 1 shows the details about the publications and co-authors of the 11 members.

We collected and analyzed the co-author data of the members of one Scratchpad and applied a social network approach to the data in order to get a better understand-

Table 1. Scratchpad members†, number of papers and their co-authors. (2001-2010).

Scratchpad members	number of publications	number of unique co-authors	number papers with one author	number of papers with co-authors	number papers with > 2 authors	max number of co-authors on 1 paper
Group total	**135**	**180**	**18**	**117**	**80**	-
member 1	30	66	3	27	18	19
member 2	18	40	4	14	13	9
member 3	0	0	0	0	0	0
member 4	16	21	1	15	14	6
member 5	0	0	0	0	0	0
member 6	1	3	0	1	1	2
member 7	17	52	1	16	11	16
member 8	70	40	9	61	36	4
member 9	0	0	0	0	0	0
member 10	0	0	0	0	0	0
member 11	3	4	0	7	1	2

† Members from 1 Scratchpad site. Group total is not the sum of the cells, several members have collaborated on the same publication

ing of the effects of moving biodiversity research to the Web. In the next paragraph we discuss the results of the analysis.

First results

We operationalised the two research questions in the following way: Does Scratchpad membership: i) connect people that were otherwise not connected; ii) provide network conditions that are beneficial for the creation of new knowledge and conditions for stability, that is, does the Scratchpad link researchers from different but not too different fields?

Do Scratchpad connect?

We used UCINET6 (Borgatti et al.2002), a software tool for social network analysis, to construct the co-author matrices and to visualise the co-author relations within the Scratchpad community and with authors outside the community. Figure 2 is a visualisation of co-author ties between Scratchpad members (inter-group relations). The graph is based on a (symmetrical) adjacency matrix with 11 rows and 11 columns. If member 1 has published with member 2 the cell contains a 1 if they did not publish together the cell entry is 0. The nodes represent the 11 Scratchpad members, the lines between the nodes their co-author relations.

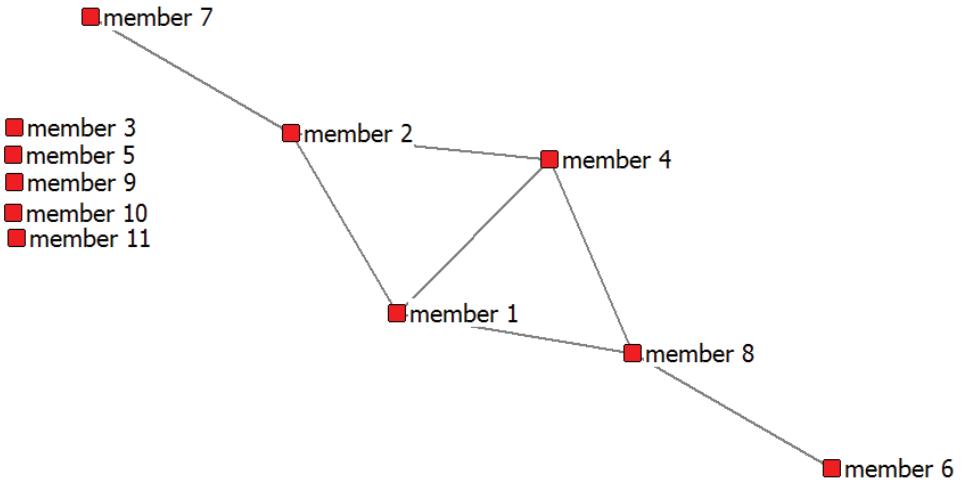


Figure 2. Graph of co-author ties† between the members of the Scratchpad Livingcreatures.info‡. (2001-2010).

† Data sources: Web of Science and Publish or Perish.

‡ For privacy reasons we use a fictional name.

What does the network show? The graph shows that six of the members are connected through co-author relations. They do not form a dense clique as they do not connect all co-authors with each other. Of these six, four have published with three others in the group, the two members positioned at the tips of the graph have published with only one other group member. The four members (1, 2, 4, 8) in the center of the graph (with each three links) already were acquainted with one of the co-authors in the co-author network of their fellow Scratchpad member. On the other hand the two members in the tips seem more "peripheral players" in *this* network.

Figure 2 also shows that the Scratchpad connects the five isolated members (3,5,9,10,11) with each other and with the members that already co-authored before they joined the Scratchpad. In other words, the five isolates are each connected to 10 potential "new" peers. If we compute in a similar way a sociogram of the Scratchpad this would look as followed, see figure 3. In the Scratchpad the 11 members are all linked to each other by *membership* of the same community. Note that a membership tie is different from as a co-author tie. A membership tie refers to sharing common interests and resources, a co-author tie refers to jointly producing a publication.

We studied the network of a particular Scratchpad community in isolation, not taking into consideration a possible overlap between different Scratchpads (currently more than 200 communities are active). Figure 3 looks trivial but serves to demonstrate the different structure as opposed to the co-author network (Fig. 2). Figure 4 is a fictional example of a network structure of how the Scratchpad under study (this one is

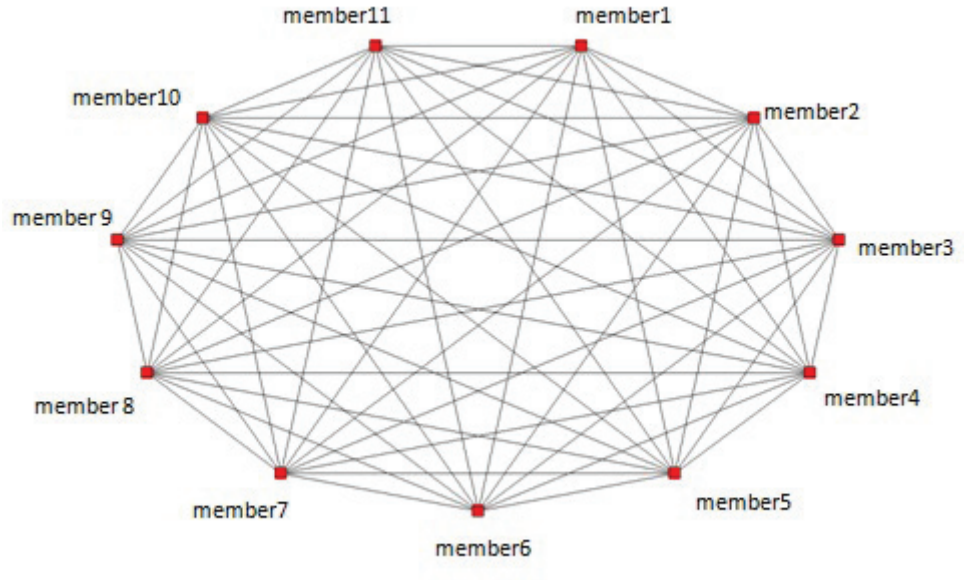


Figure 3. Graph of membership ties among Scratchpad members Livingcreatures.org

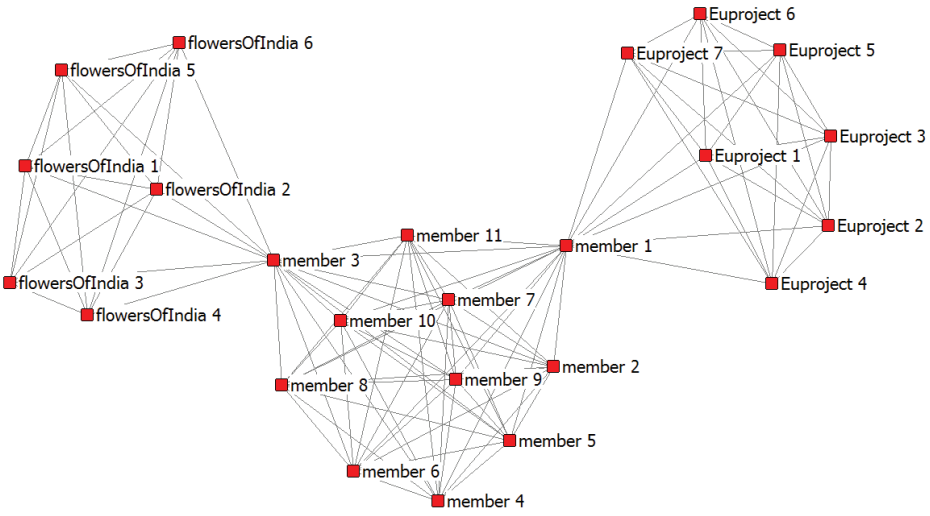


Figure 4. Livingcreatures.info and ties with two other Scratchpads. This graph shows a fictional situation.

real, only has a fictional name!) might be embedded in a lager structure of Scratchpad communities. In this fictional example two members of Livingcreatures.info are also members in other Scratchpads, one around an EU project the other one of Flowers of India. This is a graph of a fictional situation to show possible complexity in the larger Scratchpad structure.

In the next step, we extend the co-author network of the members with the co-author relations from outside the Scratchpad. The question is whether including these links changes the network topology, and whether the isolates are still isolates in the larger co-author network.

Overlapping and diverse knowledge

The literature discussed above concludes that a main enabling condition for knowledge creation and innovation is the prevalence of a mix of overlapping and diverse types of knowledge inputs that are exchanged. Scratchpads do so, if the membership is scholarly heterogeneous. Another conclusion is that if stability is important, a certain level of redundancy in the network is important (Gargiulo et al. 2000). Figure 5 is the visualization of these connections. This graph is based on a symmetrical 192 × 192 matrix representing the 192 authors and their "external co-author relations".

In red we still see the Scratchpad members. The layout is similar to Figure 2 to facilitate comparison. In blue we have the authors that are not in the Scratchpad. The red circles indicate those non-members that co-author with more than one Scratchpad member. Adding the external co-authors does not change much how the Scratchpad members are linked. The four core members have several indirect relations, in contrast to the more marginal members who lack these indirect links. Member 9 has his own small network. Adding the external authors did not link him to the large component. The other four members are the isolates, nodes without any co-author relations with other nodes.

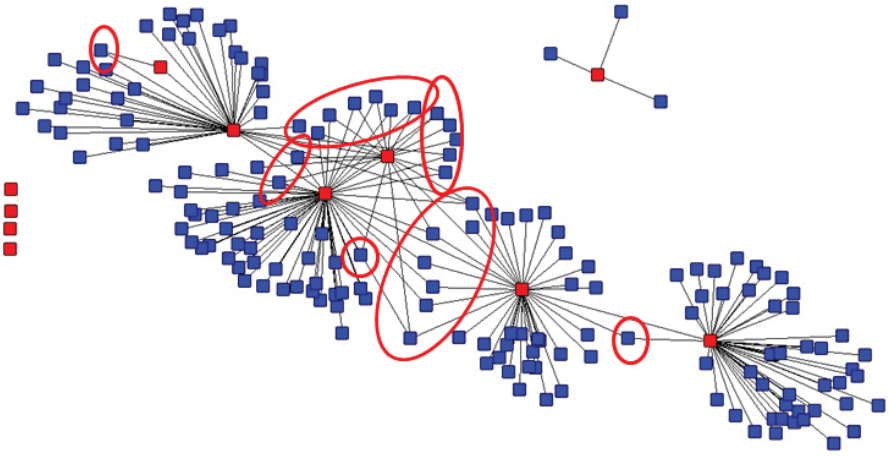


Figure 5. Graph of co-author ties† between the members of the Scratchpadlivingcreatures.info‡. (2001-2010).

† Data sources: Web of Science and Publish or Perish.

‡ For privacy reasons we use a fictional name.

Table 2 shows the number of the shared co-authors for each of the members. Note that the co-author relations between Scratchpad members are not included. Seven Scratchpad members have co-authors who are also co-authors of other members. The number of 'overlapping connections' (redundancy) ranges between 1 (for members 6, 7, and 11) and 3 (members 1, 2, 4, and 8).

Table 2 shows us that there are only a few shared co-authors and most of the 180 co-authors (see Table 1) are not shared by the Scratchpad members. The average number shared co-authors do not differ much among them. Total number of redundant connections that are brought in by the members is much lower than their contribution to bringing in new co-authors and so span structural holes. Redundancy as mentioned here refers to the Scratchpad level, meaning that overlapping co-authorships are redundant for the social capital of the Scratchpad community as a whole. At the level of the individual actor however the connection might have an added value. Redundancy at the network level will rise when Scratchpad members increase their collaboration with the same co-authors from outside the Scratchpad or when co-authors would decide to join the Scratchpad. Redundancy in the Scratchpad network will secure "access" to a specific co-author or group of authors. The goals that the Scratchpad members have set will define if more redundancy (more overlapping connections) at the network level is useful for the group.

The shared co-authors are also interesting because they are part of the professional networks of several of the Scratchpad members, and therefore they may have an interest to join the Scratchpad. The network suggests that they would contribute to intensifying interaction and to the exchange of knowledge. Adding them to the network will add a second type of redundancy and therefore stability. However, from an innovation point of view, adding redundant actors to a network is wasted energy.

We conclude that the Scratchpad under study is a rather globally distributed Virtual Community of Practice in biodiversity research, some members have a long

Table 2. Scratchpad members and number of co-authors they share with fellow members. (2001-2010).

Scratchpad members Livingcreatures.info	Number co-authors † shared with fellow members
member1	3
member2	3
member3	0
member4	3
member5	0
member6	1
member7	1
member8	3
member9	0
member10	0
member11	1

† Data sources: Web of Science and Publish or Perish.

record of publications and others have a much shorter list (see Table 1), possibly because they just starting their academic career. Also from Table 1 we conclude that the Scratchpad members of this community have a collaborative attitude. Their publication behavior demonstrates that they are used to collaborate with several authors on one paper, running from two authors up to 19 authors on a paper. When comparing co-author relations and Scratchpad membership we see that signing up for the Scratchpad has created new ties for every one of the 11 members, though some gained more new connections than others. Scratchpad members do share co-authors from outside with their fellow Scratchpad members, showing that the two knowledge networks (author network, Scratchpad network) partially overlap. However, most co-authors of every Scratchpad member are new to the other members, suggesting that members bring in not only similar but also different knowledge sources and skills. Depending on the goal of the Scratchpad the members could try to increase either the overlapping - either the diverse types of knowledge inputs at the network level (redundancy versus spanning structural holes).

Of course, co-author ties and Scratchpad networks only form two of many types of networks of researchers. Other networks (e.g. based on organisation membership, committee membership etc.), may change the network configuration, and therefore may show a different role and effect of Scratchpads in the total network of biodiversity research. This is something to consider in future investigation. A second issue that needs further research is what Scratchpad members actually do in the Scratchpad, as this may teach us about the nature of the Scratchpad relations.

Discussion

This paper aimed to explore what theoretical framework, method and data can be used to study the effects of the increasing role of web-based research environments on the practice, innovation and performance of biodiversity researchers.

We used the rich body of theories on organisational dimensions of knowledge creation, which suggests enabling conditions for knowledge production and innovation. The design of Scratchpads is partly based on the criteria in "bringing together knowledgeable people in communities of practice". Social network theory teaches us how to study and assess the network configuration of these knowledgeable people, as the patterns of links determines the added value for individual actors as well as for the network as a whole, in terms of social capital, knowledge creating power and stability. The concept of multiplex networks reflects that scientists work in a 'multi layered' research environment: e-scientists are active in a variety of professional networks in and outside their organisation, real and virtual.

As science moves to the Web, the behavioral footprints of scientific work practice are more and more available as (secondary) web data. Web data are an inexpensive way

to 'observe' the behavior of large groups of people. From Buckman (2006) we took that the there is a challenge to compile a representative data set from the Web and that therefore the data has to be controlled for with use of a primary data set.

From our pilot study we learnt that through computing of relatively simple graphs, we get a better understanding of the effects of Scratchpad membership on scholarly networks. It enables us to compare characteristics of Scratchpad networks with e.g. co-authors. Our analysis suggests that Scratchpads do create links between researchers that do not exist in the co-author network, and therefore fill structural holes in the network. This is one of the enabling conditions for the creation of new knowledge.

Analyzing the number of shared co-authors among members of Scratchpad Livingcreatures.info indicates that the members form a loosely collaborating group of researchers. However, if we also take into consideration the collaboration between the members, the network seems denser. Some Scratchpad members were already collaborating before joining the Scratchpad, however most co-author relations are from outside the Scratchpad community. In other works, the Scratchpad partly reinforces already existing relations, but also creates new links for those members that were not yet included in the co-author network.

Our pilot study shows that the selected approach is promising. In the next phase we will extend the study in several ways. Firstly, we used data on only one Scratchpad. We plan to repeat the analysis for a large set of Scratchpads, which will enable us to test whether the level of variety correlates with knowledge production and innovation, as the theory suggests. Secondly, in the current pilot study we treated all co-author relations as having the same importance, which does not well reflect real world relationships. This is also something to take into account in future research. Thirdly, we used only two different networks of the researchers (Scratchpads; co-authorships) while neglecting many others, such as organisational proximity, professor-student relations, project membership, and scientific specialisation. In order to get the full picture of the role of Scratchpads in scholarly networks, the analysis should be extended with the kind of networks mentioned. Fourthly, it is crucial to compare Scratchpad members with non-members, in order to test if changes in research practice and performance of members are different from eventual changes in the field at large. Finally, the Scratchpad we studied in this paper was launched in 2011. In order to assess the effects of the deployment of virtual environments, we suggest using a longitudinal research approach: have co-author networks and the thematic orientation of Scratchpads users changed over time, and is this change different from other researchers in the field?

These questions are not only theoretical relevant, but may also be useful in the practice of organising Scratchpads and other virtual research environments. It may also help to identify potential interesting new Scratchpad members that might be actively invited to participate. Moreover, these lines of research could contribute to sustain user engagement and to general research infrastructure policy.

References

Anandarajan M, Anandarajan A (2010) e-research collaboration. theory, tools and techniques. Springer-Verlag, Germany.

Anderson P (2007) What is Web 2.0? Ideas, technologies and implications for education. JISC Technology & Standards Watch. http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf

Andriessen JHE (2005) Archetypes of knowledge communities. In: Van den Besselaar P, De Michelis G, Preece J and Simone C (Eds) Communities and technologies. Proceedings of the second communities and technologies conference Milano,(Italy) June 13–16, 2005: 191–213. doi: 10.1007/1-4020-3591-8_11

Ardichvili A (2008) Learning and knowledge sharing in virtual communities of practice: motivators, barriers, and enablers. Advances in developing human resources 10 (4): 541–554. doi: 10.1177/1523422308319536

Beaulieu A (2005) Sociable hyperlinks. An ethnographic approach to connectivity. In: Hine C (Ed) Virtual methods. Berg, New York, 183–199.

BHL- Biodiversity Heritage Library: http://www.biodiversitylibrary.org/

Borgatti SP, Everett MG, Freeman LC (2002) UCINET for Windows. Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Buckman A (2006) Analysis of log file data to understand behavior and learning in an online community. In: Weiss J, Nolan J, Hunsinger, J, Trifonas P (Eds) The International Handbook of virtual learning environments. Springer, Netherlands, 1449–1465. doi: 10.1007/978-1-4020-3803-7_58

Burt RS (2001) Structural holes versus network closure as social capital. In: Lin N, Cook K, Burt RS (Eds) Social capital. Aldinede Gruyter, New York, 31–56.

Burt RS (2004) Structural holes and good ideas. American Journal of Sociology 110 (2): 349–99. doi: 10.1086/421787

Burt RS (2007) Brokerage and closure: An introduction to social capital. European Sociological Review 23(5): 666–667. doi: 10.1093/esr/jcm030

Demsetz H (1991) The theory of the firm revisited. In: Williamson OE, Winter SG, Coase RH (Eds) The Nature of the Firm. Origins, Evolution, and Development. Oxford University Press, New York, 159–179.

eResearch2020 (2010) The role of e-infrastructures in the creation of global virtual research communities. http://www.eresearch2020.eu/eResearch2020%20Final%20Report.pdf

Fraser M (2005) Virtual Research Environments. Overview and activity: Ariadne 44. http://www.ariadne.ac.uk/issue44/fraser/

Gargiulo M, Benassidoi M (2000) Trapped in your own net? Network cohesion, structural holes, and the adaptation of social capital. Organization Science 11 (2): 183–196. doi 10.1287/orsc.11.2.183.12514

GBiF- Global Biodiversity Information Facility: http://www.gbif.org/

Glänzel W (2002) Coauthorship patterns and trends in the sciences (1980–1998) :a bibliometric study with implications for database indexing and search strategies. Library trends 50 (3): 461–473.

Herschel RT, Nemati H, Steiger D (2001) Tacit to explicit knowledge conversion. Knowledge exchange protocols. Journal of Knowledge Management 1: 107–116.

Hine C (2005). Introduction. In: C Hine (Ed) Virtual methods. Issues in social research on the internet. Sage, Berg, 109–113.

JISC Virtual research environment programme http://www.jisc.ac.uk/whatwedo/programmes/vre.aspx

Johns MD, Shing-Ling CL, Hall GJ (Eds) (2004) Online Social Research: Methods, Issues and Ethics. Peter Lang, New York.

Krell FT (2002) Why impact factors don't work for taxonomy. Nature 415 (6875): 957. doi: 10.1038/415957a

Lee S, Monge P (2011)The coevolution of multiplex. Communication networks in organizational communities. Journal of Communication 61: 758–779. doi: 10.1111/j.1460–2466.2011.01566.x

McFayden A, Semadeni M, Cannella Jr. AA (2009) Value of strong ties to disconnected others: examining knowledge creation in biomedicine. Organization Science 20 (3): 552–564. doi: 10.1287/orsc.1080.0388

Nonaka I (1994) A dynamic theory of organizational knowledge creation. Organization Science 5 (1): 14–37. doi: 10.1287/orsc.5.1.14

Nonaka I, Takeushi H (1995) The knowledge-creating company. Oxford University Press, New York.

Nonaka I (1997) Organizational knowledge creation. Presentation at the Knowledge Advantage Conference held November 11–12. http://www.knowledge-nurture.com/downloads/NONAKA.pdf

Nonaka I, Toyama R, Konno N (2000) SECI, Ba and Leadership. A unified model of dynamic knowledge creation. Long Range Planning 33 (1): 5–34. doi: 10.1016/S0024–6301(99)00115–6

Samarah I, Paul S, Tadisina S (2008) Knowledge conversion in GSS-aided virtual teams. An empirical study. In: Proceedings of the 41st Hawaii International Conference on System Sciences (HICSS 2008): 344.

Scratchpads: http://scratchpads.eu/

Smith VS, Duin D, Self D, Brake I, Roberts D (2010). Motivating online publication of scholarly research through social networking tools. Conference Proceedings paper delivered at COOP2010, the 9th International Conference on the Design of Cooperative Systems on 18 May, 2010 as part of a workshop titled Incentives and Motivation for Web-Based Collaboration: 329–340. http://vsmith.info/files/Webincentives8.pdf

Smith VS, Rycroft SD, Harman KT, Scott B, Roberts D (2009) Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. BMC Bioinformatics 10 (Suppl 14): S6. doi: 10.1186/1471–2105–10-S14-S6

Sveiby KE (2001) A knowledge-based theory of the firm to guide strategy formulation. journal of intellectual capital 2 (4): 344–358. doi: 10.1108/14691930110409651

Thelwall M, Klitkou A, Verbeek A, Stuart D, Vincent C (2010) Policy-relevant webometrics for individual scientific fields. Journal of the American Society for Information Science and Technology 61 (7): 1464–1475 doi: 10.1002/asi.21345

ViBRANT- Virtual Biodiversity Research and Access Network for Taxonomy: http://vbrant.eu/

Wasserman S, Faust, K (1994) Social network analysis. Cambridge University Press, Cambridge.

Wenger E (1998) Communities of practice. Learning meaning, identity. Cambridge University Press, New York.

Woo J-H, Clayton MJ, Johnson RE, Flores BE, Ellis C (2003) Dynamic knowledge map. Reusing experts' tacit knowledge in the AEC industry. Automation in Construction 13: 203- 207. doi: 10.1016/j.autcon.2003.09.003

ZooKeys 150: 211–229 (2011)
doi: 10.3897/zookeys.150.2149
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Engaging the broader community in biodiversity research: the concept of the COMBER pilot project for divers in ViBRANT

Christos Arvanitidis[1], Sarah Faulwetter[1,2], Georgios Chatzigeorgiou[1,3], Lyubomir Penev[4], Olaf Bánki[5], Thanos Dailianis[1], Evangelos Pafilis[1], Michail Kouratoras[6], Eva Chatzinikolaou[1], Lucia Fanini[1], Aikaterini Vasileiadou[1,7], Christina Pavloudi[1,3], Panagiotis Vavilis[6], Panayota Koulouri[1], Costas Dounas[1]

1 *Institute of Marine Biology and Genetics, Hellenic Centre for Marine Research, 71003 Heraklion, Crete, Greece* **2** *Department of Zoology-Marine Biology, Faculty of Biology, National and Kapodestrian University of Athens, Panepistimiopolis, 15784, Athens, Greece* **3** *Department of Biology, University of Crete, 71409 Heraklion, Crete, Greece* **4** *Pensoft Publishers, Geo Milev Street 13a 1111 Sofia, Bulgaria* **5** *Global Biodiversity Information Facility, Universitetsparken 15, DK-2100 Copenhagen, Denmark* **6** *Hellenic Centre for Marine Research, 71003 Heraklion, Crete, Greece* **7** *Department of Biology, University of Patras, 26504 Rio, Patras, Greece*

Corresponding author: *Christos Arvanitidis* (arvanitidis@hcmr.gr)

Academic editor: *V. Smith* | Received 27 September 2011 | Accepted 22 November 2011 | Published 28 November 2011

Citation: Arvanitidis C, Faulwetter S, Chatzigeorgiou G, Penev L, Bánki O, Dailianis T, Pafilis E, Kouratoras M, Chatzinikolaou E, Fanini L, Vasileiadou A, Pavloudi C, Vavilis P, Koulouri P, Dounas C (2011) Engaging the broader community in biodiversity research: the concept of the COMBER pilot project for divers in ViBRANT. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 211–229. doi: 10.3897/zookeys.150.2149

Abstract

This paper discusses the design and implementation of a citizen science pilot project, COMBER (Citizens' Network for the Observation of Marine BiodivERsity, http://www.comber.hcmr.gr), which has been initiated under the ViBRANT EU e-infrastructure. It is designed and implemented for divers and snorkelers who are interested in participating in marine biodiversity citizen science projects. It shows the necessity of engaging the broader community in the marine biodiversity monitoring and research projects, networks and initiatives. It analyses the stakeholders, the industry and the relevant markets involved in diving activities and their potential to sustain these activities. The principles, including data policy and rewards for the participating divers through their own data, upon which this project is based are thoroughly discussed. The results of the users analysis and lessons learned so far are presented. Future plans include promotion,

links with citizen science web developments, data publishing tools, and development of new scientific hypotheses to be tested by the data collected so far.

Keywords
Citizen science, marine biodiversity, SCUBA diving, data collection and publication, sustainability

Introduction

The interdisciplinary nature of biodiversity science and current problems

The Rio Earth Summit (1992) drew international concern to the global biological diversity loss and transformed the concept of biodiversity into a matter of public awareness and into an important issue in the political arena (Magurran 2004). The extent to which changes in biodiversity may induce reduction of ecosystem performance and of its potential to provide humankind with products and services still remains the focus of much scientific effort (Worm et al. 2006). The effects of these changes on the ecosystem's goods and services may imply losses of several trillions of dollars forever (e.g. Costanza et al. 1997). These calculations have been made, however, without taking into account either those ecosystem functions to which no value was assigned (e.g. their ability to perform the biogeochemical cycles), nor the societal consequences caused for example by the lost jobs, especially in the current volatile global economy.

Perhaps, the major achievement after the Rio Summit was that it changed scientists' views on ecosystem theory. The CBD (Convention on Biological Diversity 1993) forced scientists to consider multiple levels of biological organisation (e.g. genes, species, ecosystems) and an extended range of geographical or any other type of observational scales (e.g. from local to global) in which alterations may occur. These changes in scientific thinking brought to researchers, environmental managers, and policy makers the issue of the vast amount of data and information required to meet the CBD's goals, such as monitoring and conservation of biodiversity at a global scale. However, there are two fundamental problems which seriously impede our efficiency in the collection of the datasets required to achieve the targets set by the CBD: the biodiversity crisis (e.g. Singh 2002) and the taxonomic impediment (e.g. Agnarsson and Kuntner 2007). The former problem refers to the decline of biodiversity resources and has emerged as one of the major economic issues of this century. Quantifying the change in biodiversity and the resulting impact on ecosystems' goods and services for humankind is seriously hampered by the latter problem, that is, by the major gaps in our taxonomic knowledge (Lyal and Weitzmann 2004, Wheeler et al. 2004, Carvalho et al. 2005). A recent study by Mora et al. (2011) has estimated that ~8.7 million eukaryotic species exist globally, of which ~2.2 million are characterised as marine. As only 1.2 million species have hitherto been catalogued, this means that some 86% of the existing species on Earth and 91% of the species in the ocean still await description. Although the term "taxonomic impediment" refers to the discipline of taxonomy, the multidisciplinary nature of biodi-

versity implies that it adversely affects other disciplines such as ecology: the inability to accurately classify the organisms into species (/taxa) results in poor ecological datasets and conclusions based on them. Another dimension of this problem is that the population of the professional data collectors (e.g. taxonomists) is diminishing. Consequently, solutions should be sought along two directions: (a) to find ways to increase taxonomic efficiency and (b) to establish data collection programmes and networks.

From conventional taxonomy to web-based "cybertaxonomy"

Descriptive taxonomy and classification of living organisms has its origins in Ancient Greece (Aristotle) and in its modern format dates back nearly 250 years, when Linnaeus introduced the binomial classification system still in use today. After almost 200 years of flourishing, the discipline is confronted by serious problems primarily because of the aged system used for its administration: The rules and conventions for descriptive taxonomy date back to the nineteenth century and the corresponding nomenclatural codes (e.g. zoological, botanical) that were developed in the mid 20th century, have not been updated to embrace modern information technology. Only very recently, the old tradition of communicating taxonomic acts through printed paper has started being replaced by approaches allowing electronic means (such as online-only journals) to publish scientific findings, as decided for example by the International Botanical Congress in Melbourne in July 2011 (Knapp et al. 2011). However, so far only the International Code of Botanical Nomenclature has incorporated these changes; for taxonomic acts in zoology, printed versions are still required. Crucial taxonomic information for the active functioning of the discipline, the type-material of each species, is still made available only through formal loans from museums and academic zoological/botanical repositories (Causey et al. 2004). In the twentieth century, taxonomy expanded towards modern disciplines such as genetics and phylogeny (Godfray 2002). The phenomenal explosion of sequence, genomic, transcriptomic, proteomic, metabolomic and other molecular disciplines, has largely been assisted by the achievements of computer science and internet technology (e.g. Johnson and Browman 2007). As a consequence, the rules for their functioning and the potential for their further development resulted in worldwide information facilities and projects/initiatives (e.g. the Consortium for the Barcode of Life – CBOL, http://www.barcoding.si.edu (Herbert et al. 2003, Stoeckle 2003), or the Global Biodiversity Information Facility – GBIF, http://www.gbif.org), launched as an international platform to aggregate and index occurrence data worldwide. Molecular classification has inevitably utilised computing power for the development of robust phylogenies and resulted in initiatives, such as the Assembling the Tree of Life initiative – ATOL, http://www.phylo.org/atol (Cracraft and Donoghue 2004). At the same time, taxonomy publishing has also been experiencing major developments in the past few years. Several important components of the Semantic Web, such as cross-linking, semantic tagging, data publication, data sharing, data aggregation, etc., have become ordinary components in the vocabulary of the biodiversity scientists (Penev et al. 2010a, b). Therefore, internet and web developments can profoundly assist current science to overcome the taxonomic impediment.

Engaging a broader community in marine biodiversity research

Most of the ecological information and data are collected in the framework of temporally limited projects, simply because the collection costs are covered by the project funds. This trend commonly results in series of datasets that are predominately discontinuous or unevenly spread, geographically, temporally or ecologically. The latter becomes more obvious in the marine environment in which the collection costs are much higher than in the terrestrial realm due to the diverse and expensive floating means as well as the specific sampling gears and methods used. Several international projects which are targeted at continuous data collection from specific habitats have been launched in the last couple of decades. An exemplar project of this category is the NaGISA project (National Geography in Shore Areas; http://www.nagisa.coml.org/) which operates under the umbrella of CoML (Census of Marine Life, http://www.coml.org/). As the population of the professional taxonomists is diminishing, the mobilization of citizen scientists has become a key element to the success of the information and data collection process (e.g. Delaney et al. 2007, Hand 2010, Silvertown 2009, Trumbull et al. 2000). The implementation of citizen science in the marine environment currently faces two difficulties: (a) only the tidal zone can be approached by all citizens, and (b) the maximal depth safely reachable by recreational SCUBA divers is limited to 40 m. In the latter case, expensive diving equipment and certified training are required.

Community development in web-based biodiversity data systems and the role of COMBER

COMBER (Citizens' Network for the Observation of Marine BiodivERsity, http://www.comber.hcmr.gr) is a pilot project which has been initiated under the ViBRANT e-infrastructure and as part of this it taps into a suite of developments aimed at supporting virtual research communities in biodiversity science. ViBRANT is a European funded FP7 project (2010-2013) with the goal to provide an integrated framework of existing and newly developed services for managing biodiversity data. Scratchpads are the platform for these developments, and this platform is based on Drupal. Within ViBRANT the necessary links will be constructed to enable a free and usable data flow between Scratchpads and existing standardized taxonomic infrastructures (e.g. CBOL, EDIT platform, EOL, GBIF).

COMBER aims at engaging citizen scientists – that is, all persons interested in nature– in a coastal marine biodiversity observation network. It is currently operating in the Cretan (Greece) coastal environment with the potential to expand to the whole Mediterranean basin or any other European region. The activities have also been demonstrated in a few other coastal areas of the southern Aegean Sea. The basic characteristics of this pilot project are: (a) a web site which has been developed and functions as the main communication and promotion vehicle of the network, offering data-entry tools for collecting information which, at a later stage, are channeled to large data

aggregators (e.g. GBIF) and publication media (e.g. PENSOFT); (b) a well-defined scientific hypothesis which has been formulated to be tested with the collected data; (c) a focus on fish species; (d) a suite of tools, such as a waterproof identification guide (see below), on-the-spot professional introductory lectures, underwater training, and demonstration of web site usage as well as data entry which are used to facilitate *in vivo* identifications by participating divers; (e) collaboration with two commercial diving centres in order to ensure operational safety and to explore the market development potential for the sustainable continuation of the initiative after the end of the project; (f) exploration of new services and tools to enhance the SCUBA diving and snorkeling services which are targeted towards the tourism industry.

Material and methods

Users, stakeholders, industry and market approach

The different categories of all the interested parties were identified during the design phase of the project: (a) a user is any person interested in participating in the activities of the project; this category includes people skilled to dive with a mask and a snorkel or certified SCUBA divers; (b) the main stakeholders identified so far are the diving centre instructors and owners, the directors of the tourist offices and the director of the Cretaquarium (HCMR); they were all approached and informed about the project, its activities and the potential it may create for the tourist industry and local markets; (c) the only industry involved is the tourist industry and its relevant markets which in this case are the services offered by the diving centres and by the Cretaquarium.

Potential participants were informed about the project through: (a) the website of the project; (b) an information desk in the Cretaquarium; (c) posters and leaflets which were distributed in the participating diving clubs and in the tourist information offices. Often, divers were approached directly before their dives in the diving centres and usually expressed interest in participation.

Training and data collection

Fish species were chosen as a target taxon for the implementation of the pilot project since they are abundant and most frequently attract the attention and interest of the wide audience. The species observation and data collection was facilitated by usage of the commercial BIOWATCH underwater fish card (http://www.bio-watch.com). The underwater fish card (Dounas 2009, Dounas and Koulouri 2011) includes the forty most common fish species of the Mediterranean coastal environment and it differentiates them on the basis of morphological characteristics (e.g. body shape, fin morphology), colour pattern, and habitat. During the dive, each participant was equipped with a fish card which was used both to identify species and directly note

down observations during the dive. For convenience, it was suitably modified to be attached on the diver's buoyancy control device (BCD) with a rope and clips. In addition, small circles were drawn next to each species figure to assist the divers to quickly and accurately record their observations. Four abundance classes were assigned, following a geometric scale: (a) absence, indicated by a blank field; (b) 1–3 individuals, marked by a single bar; (c) 4–10 individuals, marked by two bars; (d) more than ten individuals, marked by three bars.

Training of participants in data collection and data entry was implemented as short seminars given by marine scientists. The seminars were divided into three parts: (a) Before the dive, participants followed a short (~15min) introduction on the data collection protocol, including how to distinguish target fish species using the underwater fish card and correctly record the observations; (b) During the dive, each scientist accompanied maximally 3 participants to continue training in fish identification and data recording, thus ensuring maximally possible accuracy. During the first 10–15 minutes of each dive the scientists pointed out various fish species and helped the participants in correctly identifying them. After this initial period, participants were encouraged to continue the data collection by themselves, however, the scientists were available for help all the time; (c) After the dive, a short de-briefing and discussion of possible questions followed. Participants were then introduced to the website, created an account, completed their diving profile (e.g. diving level, number of total dives), logged dive information (e.g. location, depth, visibility, air consumption) and recorded the observed species. Finally, participants were asked to complete a questionnaire targeted at the experiences gained through participation and the perception of the project. The questions included can be roughly divided into five categories: (a) motivation to participate (5 questions), (b) perception on the continuation of the project (1 question), (c) willingness to pay for a similar service in the future (1 question), (d) project design and implementation (4 questions), and (e) suggestions and comments (4 questions).

Web developments and data management

COMBER uses Drupal (http://www.drupal.org), a free and open source Content Management System (CMS) as a software to perform all underlying functionality of the system. This allows full interoperability with ViBRANT and Scratchpads which are based on the same software. Many elements of the site, such as user management, profile creation, image galleries and discussion fora have been created using built-in features or readily available Drupal modules. Users can log into the site with their Facebook account, a valuable feature to strongly facilitate the registration process on the site. Registered users can continue to contribute data after participation in the seminars, use the diving log to keep track of their dives and species observations, upload photos of fish species and discuss various topics in the discussion fora. A competitive element is introduced by a five-star ranking system indicating the activity level of the user – the more dives with fish observations are contributed to the system, the higher

the user ranks in a "Top contributors" list, thus providing a playful incentive to contribute (see relevant paragraph below).

Results

Principles and implementation

Principles

The COMBER pilot project has been designed according to five fundamental principles: (a) Diving safety, ensured by involving two certified diving centres in the project which were responsible for the strict adherence to safety rules; (b) Simplicity: this principle refers to the underwater observation protocol and is extremely important, especially for non-professional recreational divers, because the diving process itself contains many elements requiring the divers' concentration (buoyancy control, pressure equalising, air consumption, adjusting to swimming underwater, monitoring depth and dive time to calculate the dive profile and avoid dangers of decompression sickness and control of diving equipment). Therefore, an additional activity such as the observation and recording of the fish species and their relative abundance on the fish card definitely introduces an additional concern which may easily turn into stress. The data collection protocol has thus been designed in a very straightforward way to require as little effort from the divers as possible; (c) Efficiency: this principle refers to the accuracy of the data collected by SCUBA divers without experience in fish identification. The fishcard focuses on easily recognisable characteristics to identify fish species. Colour and patterns are the most easily used characteristics. However, due to the progressive absorption of wave lengths of the light with depth, most of the colours except for green and blue tones tend to disappear after ca. ten metres depth. Therefore, the briefing before diving focuses on body shape and colour patterns which are not lost, and the training is continued underwater by observing living animals. This transition is very important to train the divers in how to work most accurately and also to provide them with some sense where to search and in which habitats certain species are to be found; (d) Interdisciplinarity: many scientific disciplines are actively involved and interrelated in this experiment: taxonomy, ecology, statistics, sociology, economics, education; (e)Sustainability: all of the above interrelated disciplines serve the same dual goal: to involve citizen scientists in order to produce reliable data and information and to sustain these activities for as long as possible through the development of the relevant network, goods, and services.

Rewarding for all involved parties

The users/contributors of the COMBER activities and the project infrastructure are rewarded by: (a) a free BIOWATCH fishcard after their participation to the project;

(b) the COMBER website, which – besides offering tools to keep an electronic dive log – provides facilities to upload annotated photos and discuss with other divers in a social networking environment and automatically accredits "contribution stars" to the divers according to their activity level (number of dives); (c) the association of their name with the information and data from the moment they submit their data, ensuring full credits for their work in any upcoming publication which uses these data.

Data policies and management

The pilot project closely follows ViBRANT's policy on the management of intellectual property. The concept of "Open Science" is adopted by COMBER as an overarching principle. In short, this concept implies the free/open software use under the Creative Commons movement. Clear documentation of the methodology used and of the data and results extracted is centrally placed in this concept. The intellectual rights of the information and data submitted by the user always stay with the user and allow him/her to get flexible rights for reuse. All the relevant statements and legal conditions regulating this policy are published on the web page of the pilot project. Any application, including software, source code, is free for use (GNU General Public License). Any other content uploaded on the COMBER web page, such as training courses, literature references and resources, images, videos, etc., are also distributed under a Creative Commons license and hence free for use by any user, provided that credits are given upon re-use of data.

From concept to implementation

The concept of the basic components of COMBER as well as the activities and information flow is shown in Figure 1. The central component of the project is the COMBER web infrastructure, which consists of a web-accessible front end for dissemination of information and data entry interfaces, as well as data management and storage services on the back end. Contrarily to these virtual tools, the component of tools and services currently refers to those provided on the spot, such as the underwater fish card, the SCUBA diving equipment, and the training by professional scientists. However, this part will eventually include commercial services to raise funds for the sustainability of the project after the end of the ViBRANT funding. Citizen scientists make direct use of the latter component during their dives and they are closely linked to the former component through the use of the web infrastructure and the virtual tools and services, including the reward system. The component of the "observations" comprises the actual species observations by the divers, which are recorded during their dive time. This data collection is an essential step in the process and therefore an important component of the project. The species identity data, as well as information on the diving profile of the diver, answers to the questionnaire, diving location, accompanying HCMR scientists and dive masters, weather conditions and typical diving information (tank charge, depth, duration), are then all uploaded to the electronic in-

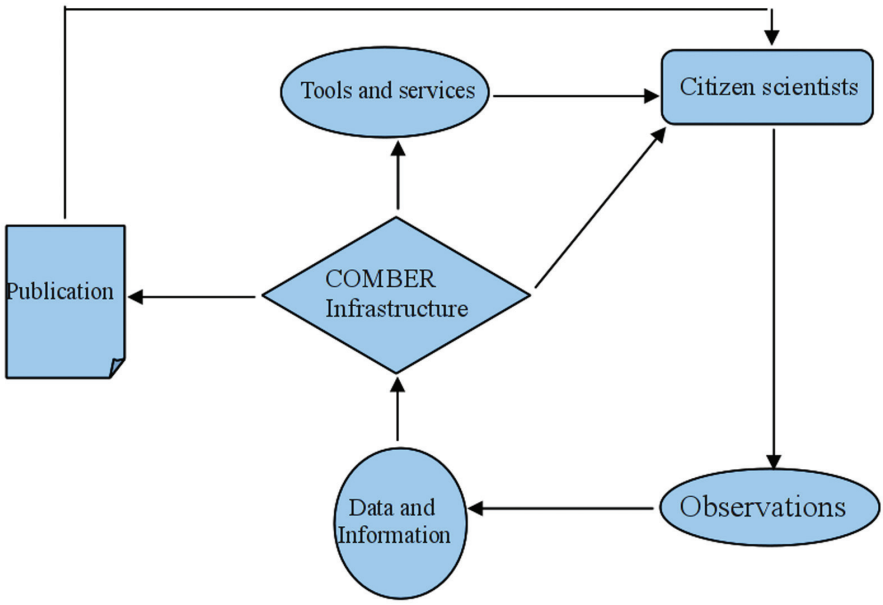


Figure 1. Schematic representation of the basic components of the COMBER project.

frastructure of COMBER and are always associated to the diver's name. The entire set of the submitted data are the intellectual property of the contributor (but free for use under a Creative Commons license, see below); by associating the name to the data it is ensured that the contributor receives full credits for his work in future publications.

User analysis

Identity of the participants

During the two months of the project (July–August 2011), 48 users (excluding the four supervising scientists) participated in the project. Twenty of the users contributed data from more than one dive or snorkeling trip and thus expanded the sampling area to several other locations in Greece (Figure 2). In total, 1,879 species observations were recorded during 95 dives and 39 snorkelling trips.

Participants came from ten countries, with the majority (42%) coming from Greece, followed by the United Kingdom and the Netherlands (12% each). The majority (70%) of the participants held a basic-level diving certificate (PADI Open Water / Advanced Open Water, CMAS *), 12% held an advanced certificate (PADI Rescue Diver, CMAS **) and 16% held a professional diving license. However, half of the divers had an advanced diving experience (>30 dives), independent of their certificate. Most of the participants already had certain knowledge about marine organisms (72% declared they had advanced (36%) or basic (36%) knowledge about

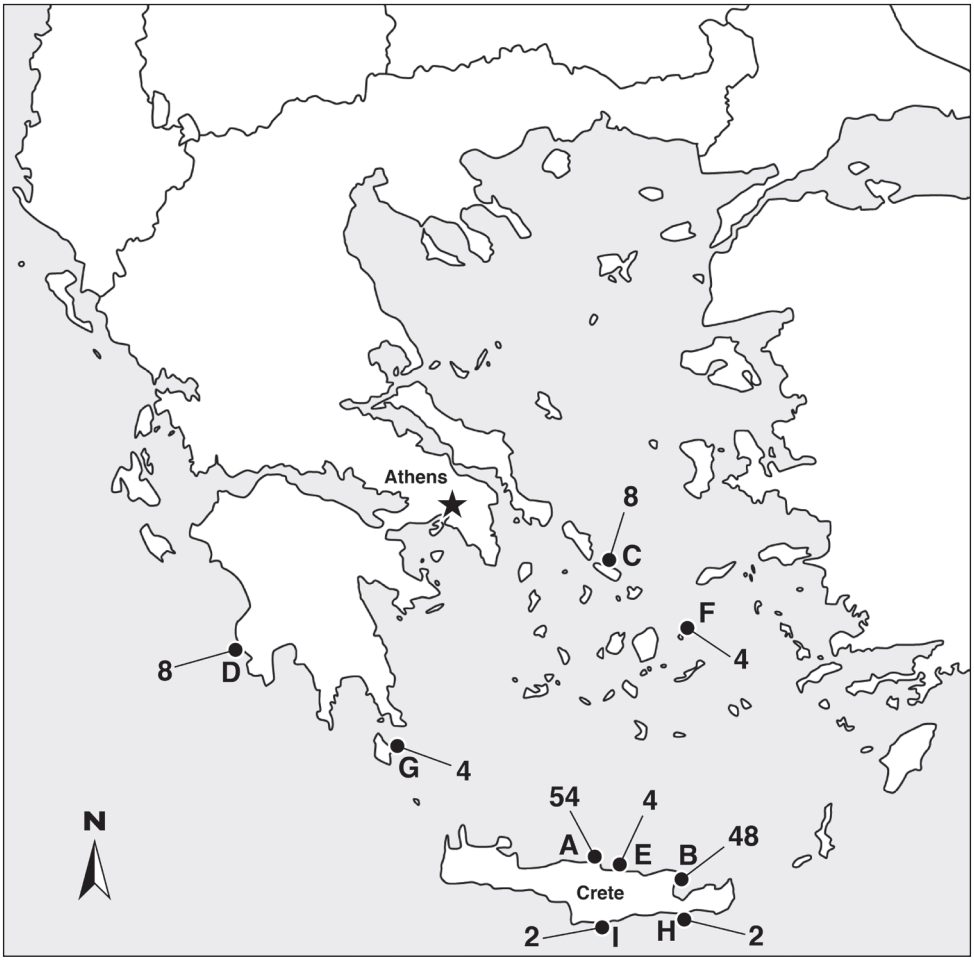


Figure 2. Map of observation sites: **A** = Lygaria **B** = Agios Nikolaos **C** = Tinos **D** = Pylos **E** = Hersonissos **F** = Donousa **G** = Kythira **H** = Ierapetra **I** = Tripiti. The two diving clubs where the project was conducted under supervision of the scientists are based in Lygaria and Agios Nikolaos (A, B). Numbers refer to dives or snorkel trips.

marine organisms, while 28% declared they had no knowledge at all). The genders were unevenly distributed (64% male, 36% female), but all age groups were present (21–55 years), with a slight dominance of 20–30 year old (39%) and 40–50 year old (30%) persons (30–40 years old: 18%, >50 years old: 13%). Two (independent) criteria can be applied to the profile data, each one separating the participants into two equally sized groups: I. age/profession: (a) a group of young (<30 years old) local participants, most of them biology students; (b) persons over 35 years old, mostly male, none of them pursuing a profession related to biology, but almost all of them with an academic education. They originate from various countries (thus many being tourists). Both the diving level and the knowledge of marine organisms were heterogeneously distributed in both groups; II. diving skills/knowledge of

marine organisms: (a) a group which had little diving experience (<30 dives); 82% of these participants had basic or no knowledge of marine organisms; (b) a group of experienced divers (>30 dives); here 64% claimed they had an advanced knowledge of marine organisms.

Behaviour, motivation and perception of the project

During the dives, the experience in diving was a major factor contributing to underwater behaviour and data collection. Inexperienced divers (with <30 dives) moved slowly, needed more time to observe and identify the fish, asked questions more frequently and needed more supervising during their dives. The major factor which generally influenced their behaviour was the effort spent to control their buoyancy and equipment, thus it was harder for them to focus on diving and observing at the same time. Experienced divers moved in a more efficient way, they needed less time to observe and identify species and to collect information, they observed fish in different habitats (e.g. under rocks, in the water column above them) and needed much less supervising attention by the scientists. From the observations of the accompanying scientists, a general trend for increased quantity and quality of data with an increasing number of dives could be discerned, the validity of which is currently tested in ongoing analyses of all data.

The results from the questionnaires concerning the motivation for participation and participants' perception of the project (answered by 25 users) can be divided into three broad categories: a) Identification process: The majority of the divers (64%) declared that some fish were easy to recognise but they had doubts about the validity of their results, while the remaining persons had no difficulties in identifying species. However, 90% of the participants found the short seminars before the diving helpful and claimed that by using the fish card only, they would have had problems to identify the species; b) Motivation: A large part of the participants (64%) had never participated before in any kind of volunteering work concerning nature conservation or observation. However, 28% are actively engaged in volunteer projects and 8% had already participated in similar projects but are not regularly engaged. Most of the divers participated because they appreciated the feeling of contributing and thus being useful for science and being part of an international network (48%) and because they like gaining new knowledge about nature (20%). Only a small percentage of them participated because their friends or dive buddies wanted to participate (8%) or simply out of curiosity (14%). The majority (84%) claimed they would continue contributing data on future dives, a minority stated that they were not interested, or they would like to but would probably lack motivation without instructors around (16%); c) Overall perception of the project: Both the project idea and its implementation were generally judged positively. On a scale from 1 ("did not like it") to 5 ("liked it very much"), 96% rated both the project idea and its implementation to be good or very good, however, the implementation part was not always scored with full marks and participants provided valuable suggestions for improvement, most of them asking the organisers to: (a)

offer more detailed introductory seminars about marine biodiversity and to make identification underwater easier; (b) provide online material (presentations, photos, videos, quizzes); (c) include more fish species and other taxa (e.g. sponges, mollusks) and (d) to better promote the website (through higher ranking in search engines, Facebook and Twitter). The project had a strong impact on the participants' perception of biodiversity: 84% declared that they now see the underwater world with different eyes, only 16% claimed that the participation left no impression on them. This is reflected in the answers to the free-text questions concerning what participants liked most or what left an impression on them: 72% stated that they appreciated learning more about the marine life and that being able to differentiate species (and thus the greater diversity) made diving a richer experience. The actual diversity of life that they were not aware of before participation left a strong impression on many participants, but there was also a positive perception of experience of citizen science: divers were impressed by the difficulties of identifying species and data collection and thus the difficulties of conducting science and they felt a personal reward through their contribution to data collection. Overall, the project was highly appreciated and 80% of the participants declared they would be willing to even pay for a similar commercial course (e.g. a "marine biodiversity diver" course).

The major groups that were identified among participants were also reflected in their answers to the questions concerning the perception of the project. Of the persons who had no problems with the identification of fish, 63% had a good diving experience (>30 dives), while within the group of persons doubting their results, the experienced divers accounted for only 26%. Generally, the experienced divers also showed a higher willingness both to continue observations on their own (100% of the experienced divers and 71% of the non-experienced divers would like to continue data contribution), and to pay for a commercial offer (88% of the experienced, 78% of the non-experienced divers). Furthermore, people with an existing knowledge of marine life found the identifications easy, often had previously participated in volunteering projects and appreciated the ability to become a part of a scientific network and to contribute to science and knowledge creation. This group consisted of many local people, often young biology students; they expressed interest in more detailed seminars, in expanding the functionality of the website and in continuing the observations.

Discussion

Lessons learned

Particular features of the industry and its associated markets

Tourism is among the most prominent economic sectors in Greece, with an average annual contribution of more than 15% to the GDP which shows a constantly increasing rate in the recent years, approaching the 20% in 2011. Greece welcomed over 19.3

million tourists in 2009, a number which was further raised in the following years (http://en.wikipedia.org/wiki/Economy_of_Greece). This sector is very important for the country's labour force and particularly for the island of Crete. A number of markets are associated with the tourism industry such as accommodation, transport, and recreation, which can potentially be positively affected by the proposed pilot project. However, there is still much uncertainty whether these markets follow the general trend for the industry. The recreation market, to which SCUBA diving services belong, is not directly associated with the tourism trend since it appears to have its own idiosyncratic dynamics. For example, during the current year in which tourism has been raised in Crete by 15% over the high season in comparison to last year, the SCUBA diving services sales dropped by a factor which reached 30%, at least as reflected in the accounting books of the collaborating diving clubs. This might relate to the fact that tourists visiting the island increasingly prefer to book their holidays in hotels offering "all-inclusive" accommodation and rarely participate in recreational activities not included in the pre-paid packages. This uncertainty has to be taken into account particularly when projections are made in a volatile economic environment.

Homogeneity of the provided services and heterogeneity of the users

Since international diving safety regulations do not allow for much variation in diving protocols, the diving process during the data collection is relatively homogeneous, despite a large variation in locations, habitats and species communities. On the other hand, there is a remarkable heterogeneity in divers' attributes such as their skills, interests, expected rewarding, and repetitiveness of the dive, to cite a few among others. This mismatch between the diving process and the divers' attributes may discourage many recreational divers, especially those who are at the beginners' stage. The pilot project on the other hand, offers some positive arguments which, if correctly communicated, can be instrumental in increasing the number and frequency of the dives. This is simply because COMBER provides an alternative diving approach through which the divers can: (a) learn about the marine environment and its life; (b) contribute to the internationally recognised goal of marine biodiversity monitoring and conservation; (c) be rewarded for their involvement in the pilot project in multiple ways; (d) have fun in a team of other divers.

Necessity of the "guided" approach and correction plans

One of the most important lessons learned so far is that the supervising and guidance of the COMBER dives is instrumental for the success of the project. This guidance is implemented at all the three stages of the dive (before, during, after). The divers need some initial information on the pilot project before they start working underwater, such as the aim, the means, the expected results, the effort required by them, the target organisms, the way they have to work, the responsible bodies and people, and extra safety measures. All participants welcomed the guidance provided during the first ten

to fifteen minutes of the dive in order to be introduced to fish identification and data collection in the field and to get an initial feedback on the accuracy of their observations. After this short period the divers generally seemed more confident with their identifications during the remaining dive time, although several of them usually kept requiring assistance. During the debriefing stage, the divers posed additional questions on some doubtful observations, on data entry through the web interface and on the continuation of their effort in the future. In most cases, discussions with the scientists and consultation of field guides allowed divers to critically assess and correct their own observations before entering them into the system, thus entering the "questioning" phase of their observation which is at the core of the scientific approach: seeking for the truth in their observations by using certain scientific criteria which in this case are taxonomic and, to a lesser degree, ecological characters. The latter has been specifically designed in order to avoid mis-observations leading to failure in the collection of reliable data, as has been observed in similar recent attempts (Goffredo et al. 2010).

The way forward

Future plans and promotion

The engagement of the broader community is a big challenge not only for the project itself but also for the marine biodiversity discipline in general. It can be broadly regarded as a significant trade zone between science, on the one hand, and society, industry, and markets, on the other. The cornerstone on which this trading zone must be built is the sustainability of the activities to both these ends. Economically healthy and sustainable activities may also serve the production and publication of reliable datasets (see also next paragraphs) needed for the study, monitoring, and conservation of the marine biodiversity while the latter also raises the concern of society for healthy and productive ecosystems.

The project has initiated the efforts in order to identify the major stakeholders and the industry and relevant markets involved. However, for the sustainability of the activities it is also important to identify the relevant target groups that may play a crucial role in the project. Taking into account the results of the questionnaire, the future expansion of the project should be developed into two different directions, aiming at two major target groups: a) a more commercially-oriented offer for experienced divers (both tourists and locals), with more comprehensive and detailed seminars, allowing them to obtain an internationally recognised diving certificate ("marine biodiversity diver"), and b) focusing on the development of local "nature clubs" which are targeted at motivated, nature-loving persons living in the area, allowing them to regularly contribute, to engage themselves in nature conservation and to meet other people with similar interests. This target group could include (biology) students, local (amateur) divers and members of other nature clubs (such as hiking or photography groups) or any other interested person.

How to use the data (from pre-treatment to scientific hypothesis testing including cleaning)

Information and data must be corrected before they are subjected to scientific analysis and hypothesis testing. This process must also follow certain criteria based on specific assumptions. The basic assumption is that the fish species recorded by the professional scientist who is supervising the dive can be used as the first criterion to identify outliers in data collected by the divers. Additional criteria may be: (a) species which are not recorded by the scientists in any of the dives at a specific location should not be included in the datasets collected by the divers; (b) broad categories of depth or habitat (e.g. hard and soft substrates, seagrass meadows) can be another criterion following the same approach as above; (c) the same criteria apply also for the abundance classes records.

The next step after data cleaning is their use (and re-use) in testing the scientific hypotheses. This is still open to discussions within the ViBRANT consortium. However, the aim of the pilot project is to examine whether the data collected by the divers are suitable for biodiversity monitoring needs. Recent biodiversity measures, based on species relatedness such as the taxonomic distinctness (e.g. Warwick and Clarke 2001), could provide the concept to formulate and test the scientific hypothesis: whether the fish species lists collected by the divers are random samples from the regional species inventory. The relevant indices of the average taxonomic distinctness (Δ^+) and variation in taxonomic distinctness (Λ^+) can be used as the statistics to test the hypothesis.

Data publishing horizons

One of the key general concepts of the ViBRANT project is to provide an e-infrastructure to facilitate maximum possible automation of the whole process of handling taxonomic data, from the collection through data management and analyses, to the stage of publication, indexing and preservation. The ultimate goal of the pilot project is to create the network of the marine biodiversity citizen scientists and also the electronic infrastructure needed for the uploaded datasets to be channeled to all interested parties, such as global biodiversity species registries (e.g. GBIF, OBIS, etc.), and published by electronic publication media, using advanced data publishing technologies. Such a technology was currently launched by the "data paper" project by GBIF and PENSOFT Publishers. According to the concept (see Chavan and Penev, in press, and Penev et al. 2011 for a detailed description), occurrence datasets and/or taxon checklists can be uploaded through the Integrated Publishing Toolkit of GBIF (IPT) (http://ipt.pensoft.net/ipt/) in accordance with the Darwin Core mapping standards. During the upload the data author is requested to fill in extended metadata descriptions, based on the Ecological Metadata Language (EML). Metadata files include such important elements such as data authors, taxonomic and geographical coverage, project description, institutional support, data storage and software management, intellectual prop-

erty rights and so on. After metadata are described, the author can generate a "data paper" manuscript from them, just by pushing a button. The manuscript is submitted to a scholarly journal and undergoes standard peer-review process. In case of acceptance, the author inserts the necessary corrections or additions recommended by the reviewers in the metadata on the IPT and then generate the revised manuscript again by pressing a button.

The data paper concept and associated tools were launched to provide incentives for data collectors to publish their data in a proper way, that is: (a) through enriched metadata description, and (b) indexing and collation of the data themselves within large international infrastructure, in this case, the GBIF data portal. The data paper will provide an opportunity for data collectors to be credited for their efforts and will open perspectives for a future collaboration with data authors having published similar types of data.

One important feature of the IPT with far-reaching consequences for biodiversity data publishing is the option for an easy creation of Darwin Core archives. The Darwin Core Archive (DwC-A) is an international biodiversity informatics data standard and the preferred format for publishing data through the (GBIF) network. The format is defined in the Darwin Core Text Guidelines. Darwin Core is no longer restricted to occurrence data, and together with the more generic Dublin Core metadata standard (on which its ideas are based), it is used by GBIF and others to encode metadata about organism names, taxonomies and species information. In addition, the whole set of data associated with the occurrence dataset, such as environmental measurement, habitat descriptions etc., can be deposited at the Dryad Data Repository (http://www.datadryad.org). Dryad provides a simplified metadata interface, however it assigns DOI numbers to each data file within a data package and to the data package as a whole. In addition to preservation and storage, Dryad also provides a workflow and standards that allow data to be cited in case they are used in future analyses, alone or with other data.

The current volume offers two exemplar papers that demonstrate the data publishing workflow described above (Faulwetter et al. 2011; Lambkin and Bartlett 2011). Both papers published data through (a) PENSOFT's GBIF IPT, (b) Dryad Data Repository and (c) DwC-A supplementary files associated with the articles and downloadable from the journal's website.

Relevant web infrastructure developments

GBIF has initiated a community driven project called the 'Nodes Portal Toolkit' that should enable communities to deploy, maintain, and extend biodiversity data portals. The project should provide an easy way for communities to start web based biodiversity data information systems with a link to the GBIF infrastructure. The GBIF Nodes Portal Toolkit will be Drupal-based, as this will allow for the integration of already existing modules. This informatics platform will also allow community development of new modules with extended functionalities for web-based biodiversity data infor-

mation systems. The first version of the Nodes Portal Toolkit will be built around Scratchpads, linking well with developments in ViBRANT. A second version will have extended functionalities, such as a tool for displaying geographical distribution maps of species, similar to what is currently displayed in the OBIS data portal. We expect COMBER to become in the coming years fully integrated with the developments in ViBRANT and the GBIF Nodes Portal Toolkit, offering interested parties a ready-made installation file allowing them to set up and deploy their own citizen-science portals without prior technical knowledge.

Acknowledgements

The authors are much indebted to the owners of the following diving clubs: (a) Happy Divers, Agios Nikolaos (Mr. Nikolaos Koutoulakis) and, (b) European Diving Institute, Lygaria (Mr. Michalis Kanakakis). Mr Michalis Papadakis (director of the Cretaquarium) is specially acknowledged for his courtesy to endorse part of the COMBER's promotion and communication activity in the Cretaquarium. Finally, the divers who participated in this first phase of the project are specially thanked for their efforts and enthusiasm: M. Beekhuyzen, Donut Berrens, Lena Chatzigeorgiou, Alex Coxon, Michael Dahlmeyer, Chris Dekker, Elodie Delva, Felix Elbaz, Konstantina Evagelou, Klitos Giannakopoulos, Giorgos Gkourogiannis, Alexis Glaropoulos, Jaap Gouverneur, Kerry Gruendel, Clare Gruendel, Chris Hill, Cheryl Horton, Serina Kapsoritaki, Demosthenes Kartsakis, Artemis Katsadoura, Nikos Kazantzakis, Simon Kerslake, Maria Kourepini, Litsa Lambrini, Anastasia Lemetti, Charalambos Malimoglou, Manolis Maragakis, Giorgos Milios, Marco Molteni, Matteo Molteni, Virginia Moutlia, Sofia Petraki, Jason Petroutsos, Virpi Roponen, Elena Sarropoulou, Patricia Schneider, Michal Szpernal, Matevsz Szpevnal, Konstantinos Tsiboukas, Thanos Vasileiadis, Anita Westen, Michael Widmer, Shaun Wilson, Katrin Zdragka, mikee, janssen. This work was supported by the EU's infrastructure project ViBRANT (Contract no. RI-261532).

References

Agnarsson I, Kuntner M, (2007) Taxonomy in a Changing World: Seeking Solutions for a Science in Crisis. Systematic Biology 56: 531–539. doi: 10.1080/10635150701424546

Carvalho MR de, Bockmann FA, Amorim DS, Vivo M de, Toledo-Pisa M de, Menezes NA, Figueiredo JL de, Castro RMC, Gill AC, McEachran JD, Compagno LJV, Schelly RC, Britz R, Lundberg JG, Vari RP, Nelson G (2005) Revisiting the taxonomic impediment. Science 307: 353. doi: 10.1126/science.307.5708.353b

Causey D, Janzen DH, Peterson AT, Vieglais V, Krishtalka L, Beach JH, Wiley EO (2004) Museum collections and taxonomy. Science 305: 1106–1107. doi: 10.1126/science.305.5687.1106b

Chavan V, Penev L (in press) Data Paper: Mechanism to incentivize discovery of biodiversity data resources. BMC Bioinformatics.

Convention on biological diversity (1993) United Nations, Treaty Series 1760: 142-383. http://treaties.un.org/doc/Publication/MTDSG/Volume%20II/Chapter%20XXVII/XXVII-8.en.pdf

Costanza R, d'Arge R, de Groot R, Farberk S, Grasso M, Hannon B, Limburg K, Naeem S, O'Neill R, Paruelo J, Raskin RG, Suttonkk P, van den Belt M (1997) The value of the world's ecosystem services and natural capital. Nature 387: 253–260. doi: 10.1038/387253a0

Cracraft J, Donoghue MJ (2004) Assembling the Tree of Life. Oxford University Press, New York, 576 pp.

Delaney GD, Sperling CD, Adams CS, Leung B (2007) Marine invasive species: validation of citizen science and implications for national monitoring networks. Biological Invasions 10: 117–128. doi: 10.1007/s10530-007-9114-0

Dounas C (2009) Illustrated guide for the identification of organisms in the field. International patent, WO 2009/144516, World Intellectual Property Organisation, International Bureau, 3 December 2009.

Dounas C, Koulouri P (2011) Mediterranean coastal fishes: an illustrated snorkeler's guide. Kaleidoskope – BIOWATCH Editors, Heraklion, Crete, Greece, 100+x pp (In Greek).

Faulwetter S, Chatzigeorgiou G, Galil BS, Arvanitidis C (2011) An account of the taxonomy and distribution of Syllidae (Annelida: Polychaetes) in the eastern Mediterranean, with notes on the genus *Prosphaerosyllis* San Martín, 1984 in the Mediterranean. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 281–326. doi: 10.3897/zookeys.150.2146

Godfray HCJ (2002) Challenges for taxonomy. Nature 417: 17–19. doi: 10.1038/417017a

Goffredo S, Pensa F, Neri P, Orlandi A, Scola Gagliardi MS, Velardi A, Piccinetti C, Zaccanti F (2010) Unite research with what citizens do for fun: "recreational monitoring" of marine biodiversity. Ecological Applications 20: 2170–2187. doi: 10.1890/09-1546.1

Hand E (2010) Citizen science: People power. Nature 466: 685–687. doi: 10.1038/466685a

Herbert PDN, Cywinska A, Ball SL, Waard JR de (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society B 270: 313–322. doi: 10.1098/rspb.2002.2218

Johnson SC, Browman HI (2007) Introducing genomics, proteomics and metabolomics in marine ecology. Marine Ecology Progress Series 332: 247–248. http://www.int-res.com/articles/theme/m332_TS.pdf

Knapp S, McNeill J, Turland NJ (2011) Changes to publication requirements made at the XVIII International Botanical Congress in Melbourne - what does e-publication mean for you? Phytokeys 6: 5–11. doi: 10.3897/phytokeys.6.1960

Lambkin CL, Bartlett JS (2011) Bush Blitz aids description of three new species and a new genus of Australian beeflies (Diptera, Bombyliidae, Exoprosopini). In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 231–280. doi: 10.3897/zookeys.150.1881

Lyal CHC, Weitzmann AL (2004) Taxonomy: exploring the impediment. Science 305: 1106. doi: 10.1126/science.305.5687.1106a

Magurran AE (2004) Measuring Biological Diversity. Blackwell Publishing, Madlen, Oxford, Carleton, 260 pp. http://books.google.com/books?id=tUqzLSUzXxcC

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How Many Species Are There on Earth and in the Ocean? PLoS Biology 9: e1001127. doi: 10.1371/journal.pbio.1001127

Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith VS, Brake I, Ryrcroft S, Scott B, Johnson NF, Morris RA, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress WJ, Thompson FC, Erwin T (2010a) Semantic tagging of and semantic enhancements to systematics papers. ZooKeys working example. ZooKeys 50: 1–16. doi: 10.3897/zookeys.50.538

Penev L, Kress W, Knapp S, Li DZ, Renner S (2010b) Fast, linked, and open – the future of taxonomic publishing for plants: launching the journal PhytoKeys. PhytoKeys 1: 1–14. doi: 10.3897/phytokeys.1.642

Penev L, Mietchen D, Chavan V, Hagedorn G, Remsen D, Smith V, Shotton D (2011). Pensoft Data Publishing Policies and Guidelines for Biodiversity Data. Pensoft Publishers, http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf

Silvertown J (2009) A new dawn for citizen science. Trends in Ecology & Evolution 24: 467–471. doi: 10.1016/j.tree.2009.03.017

Singh JS, (2002) The biodiversity crisis: A multifaceted review. Current Science 82: 638-647.

Stoeckle M (2003) Taxonomy, DNA, and the Bar Code of Life. BioScience 53: 796–797. doi: 10.1641/0006-3568(2003)053[0796:TDATBC]2.0.CO;2

Trumbull, DJ, Bonney R, Bascom D, Cabral A (1999) Thinking scientifically during participation in a Citizen-Science project. Science Education 84: 265–75.

Worm B, Barbier EB, Beaumont N, Emmett Duffy J, Folke C, Halpern BS, Jackson JBC, Lotze HK, Micheli F, Palumbi SR, Sala E, Selkoe KA, Stachowicz JJ, Watson R (2006) Impacts of biodiversity loss on ocean ecosystem services. Science 314: 787–790. doi: 10.1126/science.1132294

Warwick RM, Clarke KR (2001) Practical measures of marine biodiversity based on relatedness of species. Oceanography and Marine Biology: Annual Review 39: 207–231.

Wheeler QD, Krell FT (2007) Codes must be updated so that names are known to all. Nature 447: 142. doi: 10.1038/447142c

ZooKeys 150: 231–280 (2011)
doi: 10.3897/zookeys.150.1881
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Bush Blitz aids description of three new species and a new genus of Australian beeflies (Diptera, Bombyliidae, Exoprosopini)

Christine L. Lambkin[1,†], Justin S. Bartlett[2,‡]

1 *Entomology, Queensland Museum, PO Box 3300, South Brisbane, Queensland, Australia 4101* **2** *DEEDI Entomology, Ecosciences Precinct, GPO Box 46, Brisbane, Queensland, 4001*

† *urn:lsid:zoobank.org:author:73892CE4-985E-4A69-8C41-EC9E2458751B*
‡ *urn:lsid:zoobank.org:author:369DC37F-3E93-4EEE-934A-60A44156F153*

Corresponding author: *Christine L. Lambkin* (christine.lambkin@qm.qld.gov.au)

Academic editor: *T. Dikow* | Received 4 August 2011 | Accepted 27 September 2011 | Published 28 November 2011

urn:lsid:zoobank.org:pub:90084A21-1988-455F-AB04-D6BFE3E38D5C

Citation: Lambkin CL, Bartlett JS (2011) Bush Blitz aids description of three new species and a new genus of Australian beeflies (Diptera, Bombyliidae, Exoprosopini). In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 231–280. doi: 10.3897/zookeys.150.1881

Abstract

Bush Blitz is a three-year multimillion dollar program to document the plants and animals in hundreds of properties across Australia's National Reserve System. The core focus is on nature discovery – identifying and describing new species of plants and animals. The Bush Blitz program has enabled the collection and description of beeflies (Diptera, Bombyliidae) from surveys in Western Australia and Queensland. Three new species of Australian beeflies belonging to the Exoprosopini are described; *Palirika mackenziei* Lambkin, **sp. n.**, *Palirika culgoafloodplainensis* Lambkin, **sp. n.**, and *Larrpana bushblitz* Lambkin, **sp. n.** Phylogenetic analysis of 40 Australian exoprosopine species belonging to the *Balaana* generic-group Lambkin & Yeates, 2003 supports the placement of the three new species into existing genera, and the erection and description of the new genus *Ngalki* Lambkin, **gen. n.** for *Ngalki trigonium* (Lambkin & Yeates, 2003), **comb. n.** Revised keys are provided for the genera of the Australian *Balaana* genus-group and the species of *Palirika* Lambkin & Yeates, 2003 and *Larrpana* Lambkin & Yeates, 2003. With the description of the three new species and the transferral of *Munjua trigona* Lambkin & Yeates, 2003 into the new genus *Ngalki* Lambkin, **gen. n.**, three genera are rediagnosed; *Munjua* Lambkin & Yeates, 2003, *Palirika* and *Larrpana*.

Keywords

Ngalki, *Palirika*, *Larrpana*, *Munjua, Balaana* genus-group, phylogenetic analysis, cybertaxonomy, Scratchpads, Morphbank

Introduction

While there are more than 140,000 published species in Australia, more than 40 per cent of continental Australia has never been comprehensively surveyed by scientists. This research was supported through funding from the Bush Blitz species discovery program, a partnership between the Australian Government, BHP Billiton and Earthwatch Australia. This innovative partnership harnesses the expertise of many of Australia's top scientists from museums, herbaria, universities, and other institutions and organisations across the country. Bush Blitz is expected to uncover hundreds of new species and provide baseline scientific data that will help us protect our biodiversity for generations to come.

This paper describes three species of beeflies from the Exoprosopini (Diptera, Bombyliidae, Anthracinae); two captured during Bush Blitz surveys and a third species collected from south-western Queensland (Qld). All three species belong to genera recently described (*Palirika* Lambkin & Yeates, 2003 and *Larrpana* Lambkin & Yeates, 2003) in a large revisionary monograph (Lambkin et al. 2003) and therefore can be described reasonably easily as all collected material has been examined recently, and the context for their description is in place.

The beeflies belong to the Family Bombyliidae, a very large, cosmopolitan family of stoutly built flies, mostly with very characteristic venation. Almost 5000 species have been described worldwide (Evenhuis and Greathead 1999) and around 370 have been described from Australia, with many more species awaiting description (Yeates and Lambkin 2006). Nine of the 15 recognised subfamilies (Yeates 1994) are found in Australia, and a key to these subfamilies is available (Lambkin et al. 2003). Most Australian species belong to the subfamilies Bombyliinae, Anthracinae and Lomatiinae. The Anthracinae are well represented in Australia, mainly by the cosmopolitan *Anthrax* Scopoli, 1763, *Ligyra* Newman, 1841, *Villa* Lioy, 1864, and a number of endemic genera including *Palirika* and *Larrpana* (Yeates and Lambkin 1998; Lambkin et al. 2003). Of the seven anthracine tribes, three (Villoestrini, Prorostomatini, Aphoebantini) are not found in Australia. Keys to the four tribes of the Anthracini occurring in Australia are available (Lambkin et al. 2003). The tribe Xeramoebini is represented in Australia by only two, still undescribed, species of *Petrorossia* Bezzi, 1908. There are 28 species of Australian Villini in the genera *Villa*, *Exechohypopion* Evenhuis, 1991 and *Lepidanthrax* Osten Sacken, 1877 (Evenhuis and Greathead, 1999). The Anthracini is represented by 34 described species in the genera *Anthrax*, *Brachyanax* Evenhuis, 1981, and *Thraxan* Yeates & Lambkin, 1998 (Yeates and Lambkin 1998). Based on the phylogenetic analyses the Australian Exoprosopini was expanded to ten genera containing 65 species, including seven new genera for 42 species in the *Balaana* genus-group Yeates & Lambkin (Lambkin et al. 2003).

Australian exoprosopines are large beeflies of diverse and striking appearance (Figs 4C, 6D, 7B) with wings usually bearing distinct hyaline and black patterns. Like most bombyliids, adult Australian exoprosopines are well covered in long, dense, coloured hairs arranged in patterns, often in stripes across the dorsal surface of the abdomen, leading to their common name of beeflies. In Australian exoprosopines, like other

members of the Anthracinae, many of the long hairs, especially on the dorsal surface, are modified into short, broad, flattened scales, often in contrasting stripes. The scales may be erect or upstanding, producing a "fluffy" appearance as in *Larrpana bushblitz* Lambkin, sp. n. (Fig. 7A). Sometimes the dorsal scales are tightly adpressed, producing a smooth, often shiny appearance as in *Palirika* (Fig. 6D). Some beeflies have some scales or hairs that are reflective, appearing shining gold or brilliantly silver as on the terminal tergites of the male anthracine *Anthrax maculatus* Macquart, 1846 (Yeates and Lambkin 1998). While many beeflies have vestiture (hairs or scales) that is shiny, only the endemic Australian genus *Palirika* has metallic, reflective scales for which the colour of the reflected light is different from the colour of the scales. In this genus black scales on the dorsal surface of the face, thorax, and abdomen may be iridescent and refractive, and reflect green, blue, maroon or purple colours (Fig. 6C, D). The reflectivity may be very dull, almost dark as in *Palirika mackenziei* Lambkin, sp. n. (Fig. 6C, D), or highly reflective and bright (Lambkin et al. 2003).

Adult Australian exoprosopines favour warm, sunny localities, especially in the more arid regions. Most have a strong, hovering flight, and are commonly taken from blossom, or sitting on patches of bare earth. Adults are pollen and nectar feeders, and many are important pollinators of native plants. Many species can be collected congregating on hilltops, demonstrating a landmark-based mating system (Lambkin et al. 2003). Very little is known about the life histories of Australian exoprosopines, but some larvae are hyperparasites, parasitising prepupal instars of Hymenoptera that, in turn, are parasitising Coleoptera (Yeates et al. 1999).

This paper describes three new species of exoprosopine beeflies; two captured during Bush Blitz surveys and a third species collected from south-western Qld.

Palirika mackenziei sp. n. was collected from the large grazing property, Plevna Downs, owned by the Mackenzie family, 63 km west of Eromanga, in extremely arid south-western in late December 2007. While accompanied by Noel Starick (QM volunteer) and Robyn Mackenzie, CLL hand netted a single female specimen (Fig. 6C, D) hill-topping on the summit of Tompilly Hill (Fig. 1A, B), a jump up on Plevna Downs. This species was unlike any other *Palirika* collected; smaller and darker in both body and wing infuscation.

Four male specimens (Fig. 7A, B) of *Larrpana bushblitz* sp. n. were hand netted by CLL from Karara Pastoral Lease in Western Australia, hill-topping on Forrest lookout (Fig. 1D), 24.4km SE Boiada Camp and on a nearby hilltop 23.5km ESE Boiada Camp during the Bush Blitz survey co-organised by WAM on Charles Darwin Reserve, Karara, Lochada and Kadji Kadji Pastoral Leases, 213 km ESE of Geraldton, in September 2009. This species appeared similar to the two male specimens of *Larrpana zwicki* Lambkin & Yeates, 2003 collected only near Windorah (Lambkin et al. 2003).

Palirika culgoafloodplainensis Lambkin sp. n. was collected from Culgoa Floodplains National Park (NP) on the Queensland/New South Wales Border, 134 km WSW Dirranbandi, during the Bush Blitz survey of Culgoa Floodplains NP Qld, Culgoa NP and Ledknapper Nature Reserve (NR) NSW (NSW) organised by CLL and Noel Starick from QM between November 2009 and June 2010. A single male specimen

(Fig. 4C, D) was sorted by QM volunteer John Purdie from a Malaise trap sample from 7 km NNW Toulby Gate (Fig. 1C) on Culgoa Floodplains National Park (NP). Malaise and Pitfall traps had been set at four sites on Culgoa Floodplains by CLL, Noel Starick and NP Ranger Cheryn Kelly in November 2009 as part of the Bush Blitz survey. The rangers had agreed to take monthly samples until we could return. This specimen was from a Malaise trap that had been reset on the 20th January 2010 by Ranger-In-Charge (RIC) Andy (Keith) Coward. Because of significant rain, the rangers were unable to return to take another sample until the 19th March. Subsequent flooding in March and April 2010 prevented access to the survey areas until mid-May when CLL, Noel, Rhys Smith (QM volunteer) and rangers Andy and Megan Simpson retrieved the Culgoa Floodplains NP traps. This species was similar to *Palirika bouchardi* Lambkin & Yeates, 2003 that has been extensively collected from arid areas of central and western Australia from all states except South Australia.

Previous phylogenetic analysis of the worldwide Exoprosopini showed that the Australian bombyliids that were previously placed in *Exoprosopa* Macquart 1840, belonged to the monophyletic *Balaana* group of genera, sister to the Australian *Ligyra* (Lambkin et al. 2003). Phylogenetic analysis of 207 morphological characters of the *Balaana* group of genera led to the description of seven new genera for 42 species in that genus-group in Lambkin et al. (2003). Phylogenetic analysis of the same 207 morphological characters scored for two *Ligyra* outgroup taxa and 40 Australian species belonging to the *Balaana* generic-group supports the placement of the three new species into existing genera, and the erection of the new genus *Ngalki* Lambkin gen. n. for *Ngalki trigonium* (Lambkin & Yeates, 2003), comb. n. (Figs 2, 3).

Revised keys are provided for the genera of the Australian *Balaana* genus-group and the species of *Palirika* and *Larrpana*. The three new species are fully described; with diagnoses, distribution maps, and images of both external characters and dissected genitalia. The new genus *Ngalki* is described with diagnosis, and images of both external characters and dissected genitalia. With the description of three new species and the transferral of *Munjua trigona* Lambkin & Yeates, 2003 into the new genus *Ngalki*, three genera are rediagnosed; *Munjua* Lambkin & Yeates, *Palirika* and *Larrpana*.

We attempted to use cybertaxonomic tools to produce this paper as had been used to streamline taxonomic publication of new fly species by Winterton (2009), Brake and von Tschirnhaus (2010), Blagoderov et al. (2010b), and Winterton and Gaimari (2011). Attempts to use the morphological phylogenetic matrix to produce natural language descriptions provided only clumsy, inadequate descriptions. Using Blagoderov et al. (2010a) as a guide we completed automatic generation of the manuscript within a Virtual Research Environment (Scratchpads). As the publication module in Scratchpads is still under development, semantic enhancements, and parallel release of the publication on paper and on-line accompanied with registration of new taxa with ZooBank (http://www.zoobank.org/) as per the recent proposed amendment to the *International Code of Zoological* nomenclature for a universal register for animal names (Polaszek et al. 2005a, 2005b; ICZN 2008) were completed through submission of a Microsoft Office Word 2003 document to ZooKeys.

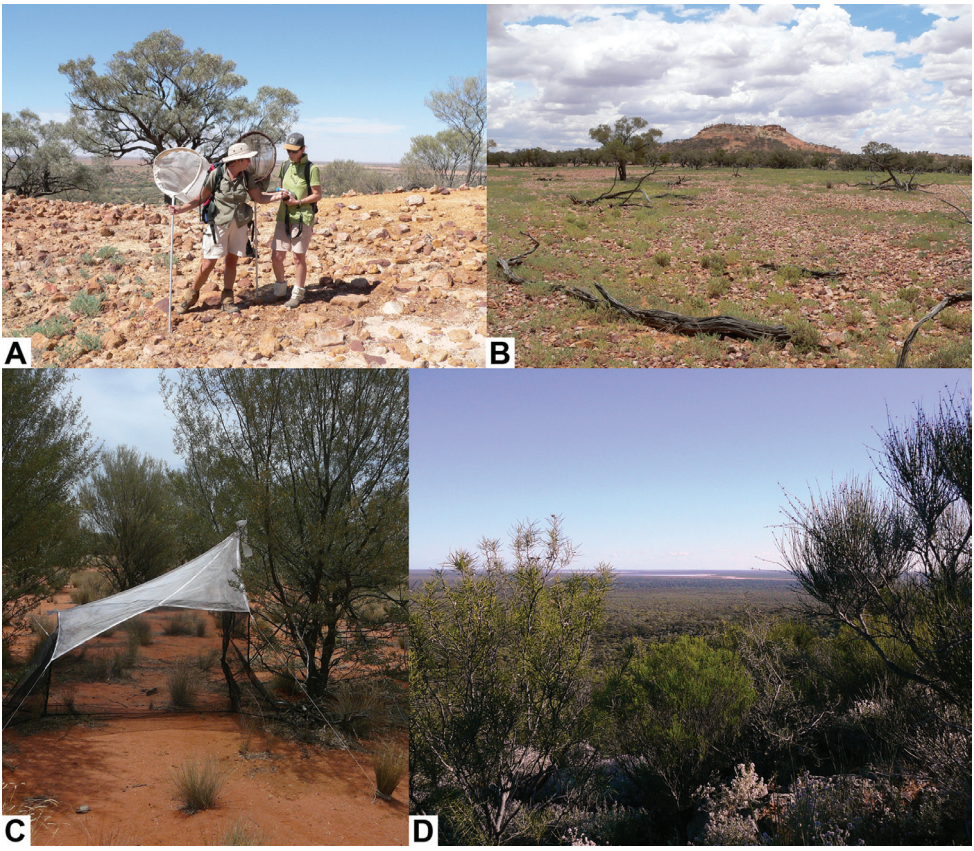


Figure 1. Collection sites. **A** CLL showing Robyn Mackenzie the single female specimen of *Palirika mackenziei* sp. n. collected hill-topping on the summit of Tompilly Hill in late December 2007 **B** Tompilly Hill, a jump up on Plevna Downs, in extremely arid south-western Queensland **C** A single male specimen of *Palirika culgoafloodplainensis* sp. n. was collected during a Bush Blitz survey from this Malaise trap, 7 km NNW Toulby Gate on Culgoa Floodplains National Park (NP) on the Queensland/New South Wales Border, 134 km WSW Dirranbandi **D** Forrest lookout on Karara Pastoral Lease 213 km ESE of Geraldton in Western Australia, where two male specimens of *Larrpana bushblitz* sp. n. were hand netted hill-topping by CLL in September 2009 during a Bush Blitz survey. Photographs A and B by N. Starick, QM.

Methods

Taxonomic Methods

The following collection acronyms are used in the text: Australian Museum, Sydney, New South Wales, Australia (AM); Queensland Museum, South Brisbane, Queensland (QM); Western Australian Museum, Perth, Western Australia (WAM). Numbers quoted with individual specimens are unique identifiers (e.g. WAM 82396, T152479 (QM), K 253702 (AM)) from the respective institutions database and are attached under each specimen on a white label. A single hind leg was removed from one specimen of each species and placed into absolute ethanol for frozen tissue storage at QM for future DNA

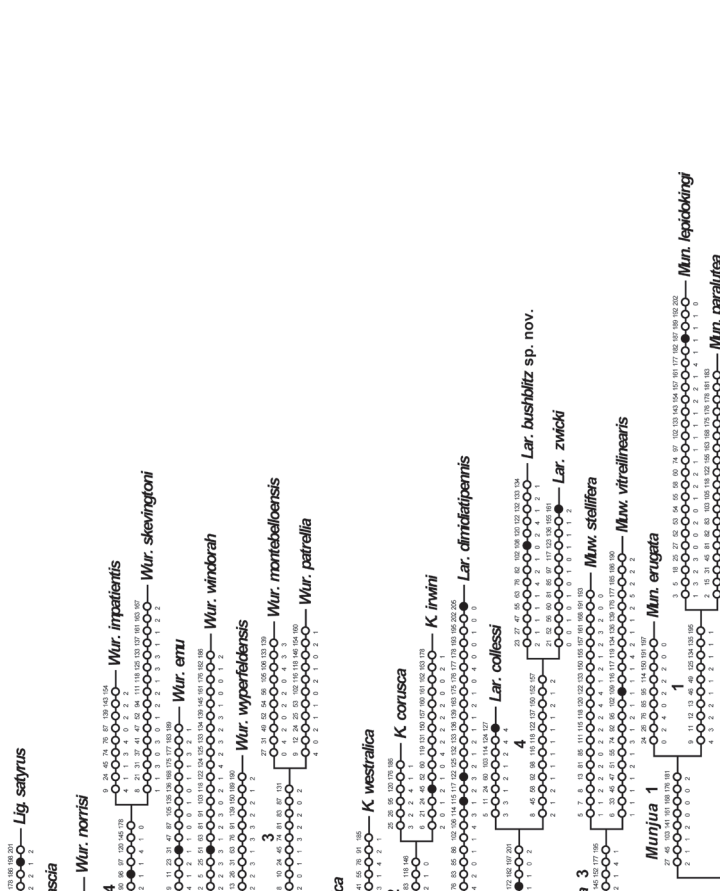


Figure 2. One of five most parsimonious cladograms (931 steps, CI = 0.24, RI =0.47). Part 1. Black circles = unique character changes, open circles = homoplasious changes. Bremer supports over branches.

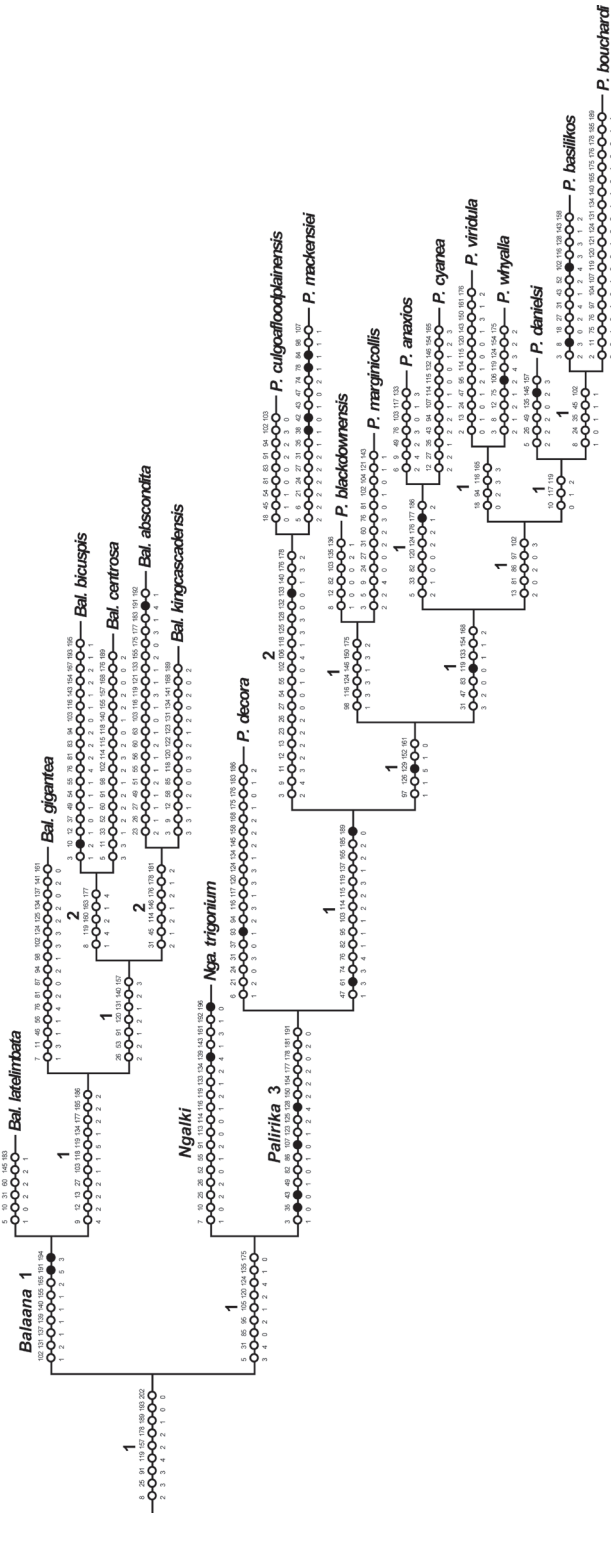


Figure 3. One of five most parsimonious cladograms (931 steps, CI = 0.24, RI =0.47). Part 2. Black circles = unique character changes, open circles = homoplasious changes. Bremer supports over branches.

extraction. Those samples were given a tissue number (e.g. A007534) that was entered into the QM Vernon database and attached under each specimen on a yellow label.

For explanation of morphological abbreviations, see Appendix 1.

The genitalia of each species were prepared by dissecting the terminal abdominal segments and then placing in cool 10% KOH overnight. Following maceration the specimen was washed, and then dissected in distilled water. Dissected genitalia were placed in alcohol for microscopic examination and into K–Y® Jelly for photography. All dissected parts from a specimen were placed in a genitalia vial containing glycerine which was pinned beneath the identification label.

Images were taken of the whole fly, external features, and dissected genitalia. A series of multiple-focal-depth digital images were taken using a Canon EOS 500D digital camera fitted, via a Leica 10446175 1x SLR Projection Lens, to a Leica MZ6 stereo dissecting microscope, and combined into a high resolution serial montage image using Helicon Focus v.5.2 Pro (Kozub 2011) or Zerene Stacker v.1.02 (Littlefield 2011). Higher-resolution digital images were deposited in Morphbank (www.morphbank.net). Separate collections of images were created for each species in Morphbank where each collection receives a unique identifier and associated URL. The URL links to the Morphbank collections have been embedded within the descriptions for each species. Images were assembled into plates using Adobe Photoshop C S5 version 13.0.3 (Adobe Systems, 2010b) and Adobe Illustrator C S5 version 15.0.2 (Adobe Systems, 2010a). Those samples were given a photograph number (e.g. PS1714) that was entered into the QM Vernon database and attached under each specimen on a purple label.

Distribution maps were produced using ArcView GIS version 3.1 (ESRI, 1998).

We intended to use cybertaxonomic methods to document these newly discovered Australian beeflies, enabling descriptions of the three new species to be generated using web resources to populate electronic documents through links to Morphbank, Life Science Identifiers, and Zoobank as had been done by Winterton (2009) whose revision serves as an example for making taxonomic description and key development more efficient by avoiding redundancy in data handling and using digital media. We hoped to complete taxonomic descriptions using a character matrix in Structured Descriptive Data format developed in Lucid Builder to simultaneously generate natural language descriptions and a key. However we encountered problems transferring the compiled phylogenetic data matrix to Lucid. Instead MacClade 4 (Maddison and Maddison 2003) was used to generate natural language descriptions based on a phylogenetic matrix including 413 phylogenetic (morphological) and phenetic (colour) characters. The resultant descriptions were clumsy and inadequate. Instead, we developed descriptions in Microsoft Office Word 2003 based on the electronic versions of closely related described species from Lambkin et al. (2003).

Several initiatives around the world have been developing tools to bring revisionary taxonomy to the web. Recent examples include software produced through the CATE (Creating a taxonomic e-science, http://www.cate-project.org), EDIT (European Distributed Institute of Taxonomy, http://www.e-taxonomy.eu) and the Australian TRIN (Taxonomy Research & Information Network, http://www.taxonomy.org.au/) projects.

These efforts support the compilation of large distributed datasets, descriptions and identification of biota. One of the tools developed in association with the EDIT initiative are the Scratchpads (http://scratchpads.eu), a Web 2.0 Virtual Research Environment, that enable taxonomists to collaborate in the production of websites documenting the diversity of life. Using Blagoderov et al. (2010a) as a guide we set up the Australasian Asiloidea Online Scratchpad (http://australasianasiloidea.myspecies.info/). We initially included for public view the published diagnoses of genera and species of the Exoprosopini (Bombyliidae: Anthracinae) and the *Taenogera* genus-group (Therevidae: Agapophytinae). Pages including images, diagnoses, and descriptions were established for each of the undescribed species in the Australasian Asiloidea Online Scratchpad, but hidden from public view until publication.

The paper has been semantically tagged and enhanced using the Pensoft Mark Up Tool (PMT) which is based on the US National Library of Medicine's DTD (Document Type Definitions) TaxPub extension http://sourceforge.net/projects/taxpub). We intend parallel release of the publication on paper and on-line accompanied by a) links to archived images on Morphbank, and b) with registration of authors, publications, taxon names and other nomenclatural acts in Zoobank, with assignment of Life Science Identifiers (LSIDs) for each new taxa as per the recent proposed amendment to the *International Code of Zoological* nomenclature for a universal register for animal names (Polaszek et al. 2005a; Polaszek et al. 2005b; ICZN 2008). The final XML output of the paper will be archived in PubMedCentral, a PDF uploaded in the Biodiversity Heritage Library (BHL), and all revised species registered in ZooBank (Penev et al. 2010).

Data resources

The nomenclatural and distributional information will be included in the Australian Faunal Directory (AFD), an open-access online catalogue of taxonomic and biological information on all animal species known to occur within Australia (ABRS, 2009), and the Australian Natural Heritage Assessment Tool (ANHAT), an open-access online map-supported database developed by the Australia Heritage Division of the Department of Sustainability, Environment, Water, Population and Communities that helps identify and prioritise areas for their natural heritage significance, focusing on biodiversity (NHAS 2009). The occurrence data has been uploaded as a Darwin Core Archive (DwC-A), to the Global Biodiversity Information Facility (GBIF) via the Pensoft Data Hosting Center at the GBIF's Integrated Publishing Toolkit (IPT) (http://ipt.pensoft.net/ipt/). The data underpinning the analysis reported in this paper including the data matrix and a most parsimonious tree, together with matrices and trees from Lambkin et al. (2003), were deposited in the Dryad Data Repository (http://datadryad.org/) at doi: 10.5061/dryad.5j64k, the TREEBASE Repository (www.treebase.org/) at http://purl.org/phylo/treebase/phylows/study/TB2:S12050, and at GBIF, the Global Biodiversity Information Facility, http://ipt.pensoft.net/ipt/resource.do?r=bushblitz

Phylogenetic analysis

Phylogenetic analysis was based on 207 morphological characters from Lambkin et al. (2003) (Appendix 1). The three new taxa were added to the data matrix from Lambkin et al. (2003) and scored for external morphology including wing venation, and internal morphology of male and female genitalia to produce a matrix for two *Ligyra* outgroup taxa and 40 Australian species belonging to the *Balaana* generic-group in Mesquite version 2.74 (Maddison and Maddison 2010) (Appendix 2 & Appendix 3 LambkinOzBombs2011.nex).

Multistate characters used for phylogenetic analyses have been treated as unordered (non-additive Mickevich and Mitter 1981; Mickevich and Weller 1990). All synapomorphies were weighted equally (Farris 1990). Character polarity was determined by comparison with the outgroups. Variation in morphology between specimens of a taxon was scored as polymorphism and interpreted in the cladistic analyses as "partial uncertainty" (Swofford and Begle 1993) where PAUP* chooses a state from the set of available states that allows minimization of the tree length. There are 66 constant characters in the analysis as the morphological data matrix was based on coding of a much broader taxon sample of 107 worldwide exoprosopine taxa for 207 morphological characters used in Lambkin et al. (2003).

Phylogenetic analyses completed 100 random step-wise addition searches, with tree-bisection-reconnection (TBR) branch swapping, MULPARS, and branches having maximum length zero collapsed to yield polytomies in effect using PAUP*4.0b10 (Swofford, 2002).

We used Bremer support (Bremer 1992; Källersjö et al. 1992; Bremer 1994) to measure the strength of evidence for nodes. Bremer support of a group is the difference in length between the tree under consideration and the shortest tree lacking that group. Bremer support values were calculated with TreeRot v.2 (Sorenson 1999) with 20 heuristic searches of the data.

Cladograms and character distribution were analysed in WinClada version 1.00.08 (Nixon 2002) and edited in Adobe Illustrator C S5 version 15.0.2 (Adobe Systems 2010a).

Results

Cladistic analysis of the 42 taxa of 141 non-constant characters produced five most parsimonious trees (MPTs) of length =931, CI = 0.243, CI excluding uninformative characters = 0.231, and RI = 0.468. The five trees differ only in the placement of *Larrpana dimidiatipennis* (Bowden, 1971); as sister to the remaining *Larrpana*, sister to *Muwarna* Lambkin & Yeates, 2003, or sister to the *Balaana* genus-group excluding *Wurda* Lambkin & Yeates, 2003 and *Kapu* (Lambkin & Yeates, 2003). Most parsimonious tree 5 was chosen with reference to the majority-rule consensus tree (Margush and McMorris 1981) as the MPT included those nodes that were found most often in the remaining MPTs. Most parsimonious tree 5 is shown in two parts

in Figures 2 and 3 with unambiguous changes on the branches, generic names and Bremer Supports above the branches.

Previous phylogenetic analysis of 207 morphological characters for the worldwide Exoprosopini showed that the Australian bombyliids that were previously placed in *Exoprosopa*, belonged to the monophyletic *Balaana* group of genera, sister to the Australian *Ligyra.* Phylogenetic analysis of characters of the *Balaana* group of genera then led to the description of seven new genera for 42 species in that genus-group (Lambkin et al. 2003). Phylogenetic analysis of the same 207 morphological characters for two *Ligyra* outgroup taxa and 40 Australian species supports the placement of the three new species into existing genera in the *Balaana* generic-group. *Palirika mackenziei* sp. n. and *Palirika culgoafloodplainensis* sp. n. form a clade within the well supported genus, *Palirika* and *Larrpana bushblitz* sp. n. forms a clade with *Larrpana zwicki* within the genus *Larrpana* (Figs 2, 3).

In Lambkin et al. (2003), *Munjua* was erected for three unusual species for which there were few apparent similarities. In that phylogenetic analysis, another particularly aberrant fly, *Munjua trigona* (Fig. 9A, B), was sister to the clade of *Munjua* and the two well-supported terminal clades of *Palirika* and *Balaana* Lambkin & Yeates, 2003. This fly clearly did not belong to either *Palirika* or *Balaana* as it possessed none of their diagnostic characters, and was therefore placed in the already heterogeneous *Munjua* rather than creating a monotypic genus (Lambkin et al. 2003).

In this phylogenetic analysis, *Munjua trigona* falls between *Palirika* and *Balaana* as sister to *Palirika* (Figs 2, 3). As this species clearly does not belong to *Palirika,* a new genus *Ngalki* for *Ngalki trigonium* is created.

With the description of the three new species and the transferral of *Munjua trigona* into the new genus *Ngalki*, the three genera *Munjua*, *Palirika*, and *Larrpana* require rediagnoses.

Taxonomy

Palirika Lambkin & Yeates, 2003

urn:lsid:catalogueoflife.org:taxon:d916e5f0-29c1-102b-9a4a-00304854f820:col20110201
http://species-id.net/wiki/Palirika

Type species: *Palirika decora*, Lambkin & Yeates, 2003: 812.

Rediagnosis. Small black, rounded, dense, adpressed metallic scales dorsally on thorax and abdomen (Fig. 6D); no abdominal white scales, sternal vestiture black, not metallic. Epandrium rounded, strongly curved, red, extended smoothly basolaterally (Fig. 4E, F). Gonocoxae deeply narrowed medially, with thickened setae ventromedially, tuft of 6–8 very long, basally-directed, thick setae medially; H projecting in lateral view; EP without lateral lobes, medial projection laterally; LAEA large, convex, extending to G margin; EJA racquet-shaped, longer than the length of G (Fig. 5). Female T_8 A little more than marginal thickening (Fig. 6F), spermathecal tube more than 8 × length of SP, clear thick-walled ring joining clear thick-walled BB and pigmented subquadrate SR.

Included species: *Palirika anaxios* Lambkin & Yeates, 2003, *Palirika basilikos* Lambkin & Yeates, 2003, *Palirika blackdownensis* Lambkin & Yeates, 2003, *Palirika bouchardi*, *Palirika culgoafloodplainensis* sp. n., *Palirika cyanea* Lambkin & Yeates, 2003, *Palirika danielsi* Lambkin & Yeates, 2003, *Palirika decora*, *Palirika mackenziei* sp. n., *Palirika marginicollis* (Gray, 1883), *Palirika viridula* Lambkin & Yeates, 2003, *Palirika whyalla* Lambkin & Yeates, 2003.

***Palirika culgoafloodplainensis* Lambkin, sp. n.**
urn:lsid:zoobank.org:act:85D57E72-0396-4994-8CF6-7C22B5BAC978
http://species-id.net/wiki/Palirika_culgoafloodplainensis
Figs 1C, 2–3, 4–5, 11; Morphbank Collection 692336

Material examined. *Holotype.* **Queensland:** ♂, 28.94°Sx146.918°E, Culgoa Floodplain NP, 7km NNW Toulby Gate, 160m, (CG4AM), Malaise, 20Jan-19Mar2010, C. Kelly, A. Coward, 19273, [dissected], PS1937, A006859, T165704 (QM). Condition: Fair (see remarks below).

Diagnosis. Wing length 20.0 mm

Large dark flies with distinct triangular basal infuscation on the wings wings. Face and frons with transparent scales. Occiput with white scales broadly filling indentation. Collar whitish-cream. Broad laterothoracic stripe of dense white flattened scales. Scutum black with lime-green metallic scales except pink metallic scales anterolaterally to PR bristles and posterolaterally anterior to APA. Scutellum with lime-green metallic scales; very long, white, flattened-scale fringe on posterior margin. Widened base of costa with reddish-brown scales, white scales posteriorly. Wing pattern dimidiate (Fig. 4C); with distinct indentation base of first r_{2+3}; extension along R_{4+5}, covering basal 1/3 of first r_{2+3} and r_5; indistinct mottling base of m_1 along m-m; no infuscated band; anal and posterior cells with apically notched hyaline area, infuscation extending along CuA_2; cup infuscated basal 4/5; anal infuscated basal 2/3. Squama edged with dense white scales admixed with some reddish-brown scales. T_1 with Ma white dorsally, black medially and ventrally, dense very long flattened white scales posterolaterally. Abdominal tergites black with bluish-green scales. Epandrium with long setae grouped loosely apically. Epiphallus with short, medial projection. G with long black setae medially, directed basally, longest on weak ventral ridge; LAEA very large, extending well beyond G margins.

Description. *Male. Head* (Figs 4A–D). Face red with transparent scales, frons brown with transparent scales; setae black, frontal depression distinct. Antennal scape red, 3 × length of pedicel, with long black setae dense laterally and ventrally; pedicel red; PP black, conical, 3 × length of pedicel, distinct apical joint; BSM rod-like, black, 3 × length of pedicel; ASM black, conical, length at least width of BSM (Fig. 4B). Occiput with white scales broadly filling indentation (Fig. 4A).

Thorax. (Figs 4C–D). Collar whitish-cream. Broad laterothoracic stripe of dense white flattened scales (Fig. 4D). Scutum black with lime-green metallic scales except

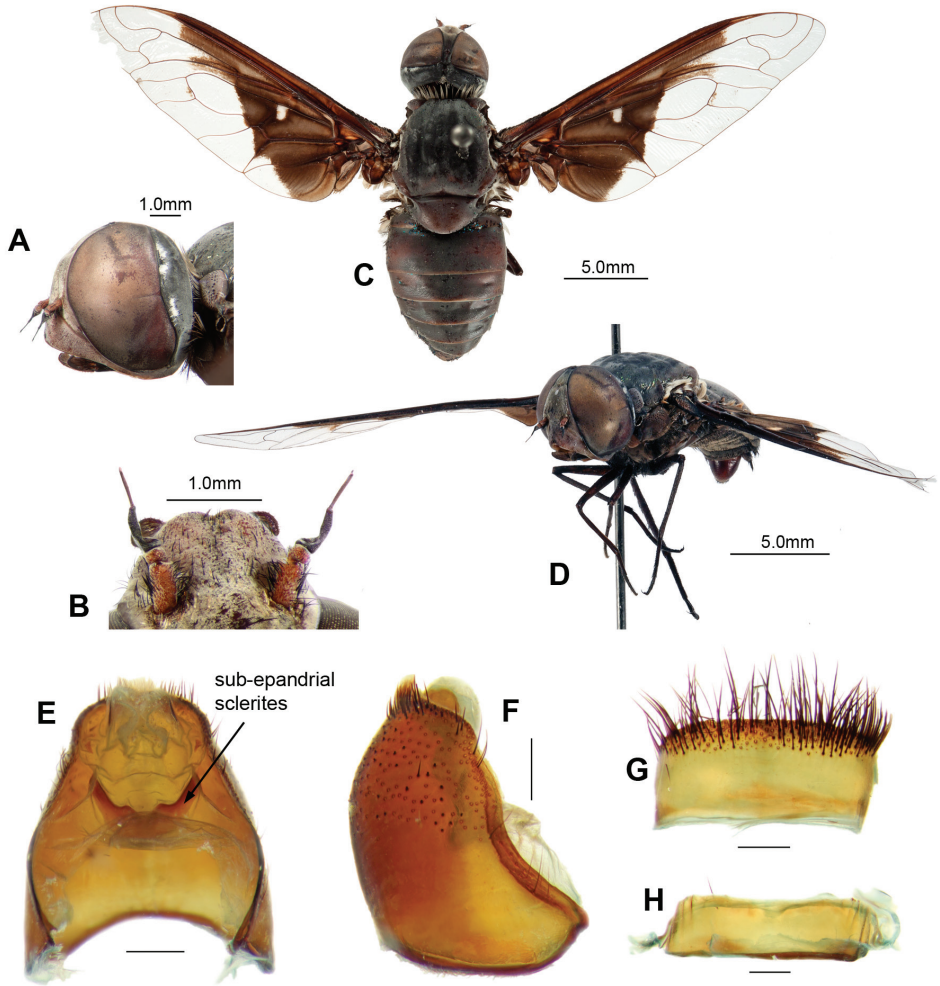


Figure 4. *Palirika culgoafloodplainensis* sp. n., Male holotype. **A** Head lateral **B** Antennae dorsal **C** Adult, dorsal **D** Adult, antero-lateral; Male genitalia: **E** Epandrium ventral with sub-epandrial sclerites **F** Epandrium lateral **G** T_8, dorsal **H** S_8, ventral. Scale line (E–H) = 0.5 mm.

pink metallic scales anterolaterally to PR bristles and posterolaterally anterior to APA; black setae. Pleural hairs black with reddish-brown iridescence. AN with black Ma; long, lightly iridescent scales at base of wing reddish-brown; long flat broad pale brown scales posteromedially. K with very long fine reddish-brown scales medially. Ma on LT black with reddish-brown iridescence. Tympanal ridge and PL with dense very long fine white flattened scales. Scutellum red, darker basally with lime-green metallic scales; very long, white, flattened-scale fringe on posterior margin. *Legs*. Legs reddish-brown, darkening apically, with black scales and setae, tarsi dark reddish-brown to black; fore-tarsi with straight microtrichia. Pulvilli sharp, curved, 1/3 length of mid- and hind-tarsal claws. Halter knob reddish-brown with apical margin yellow. *Wing* (Fig. 4C), cup nar-

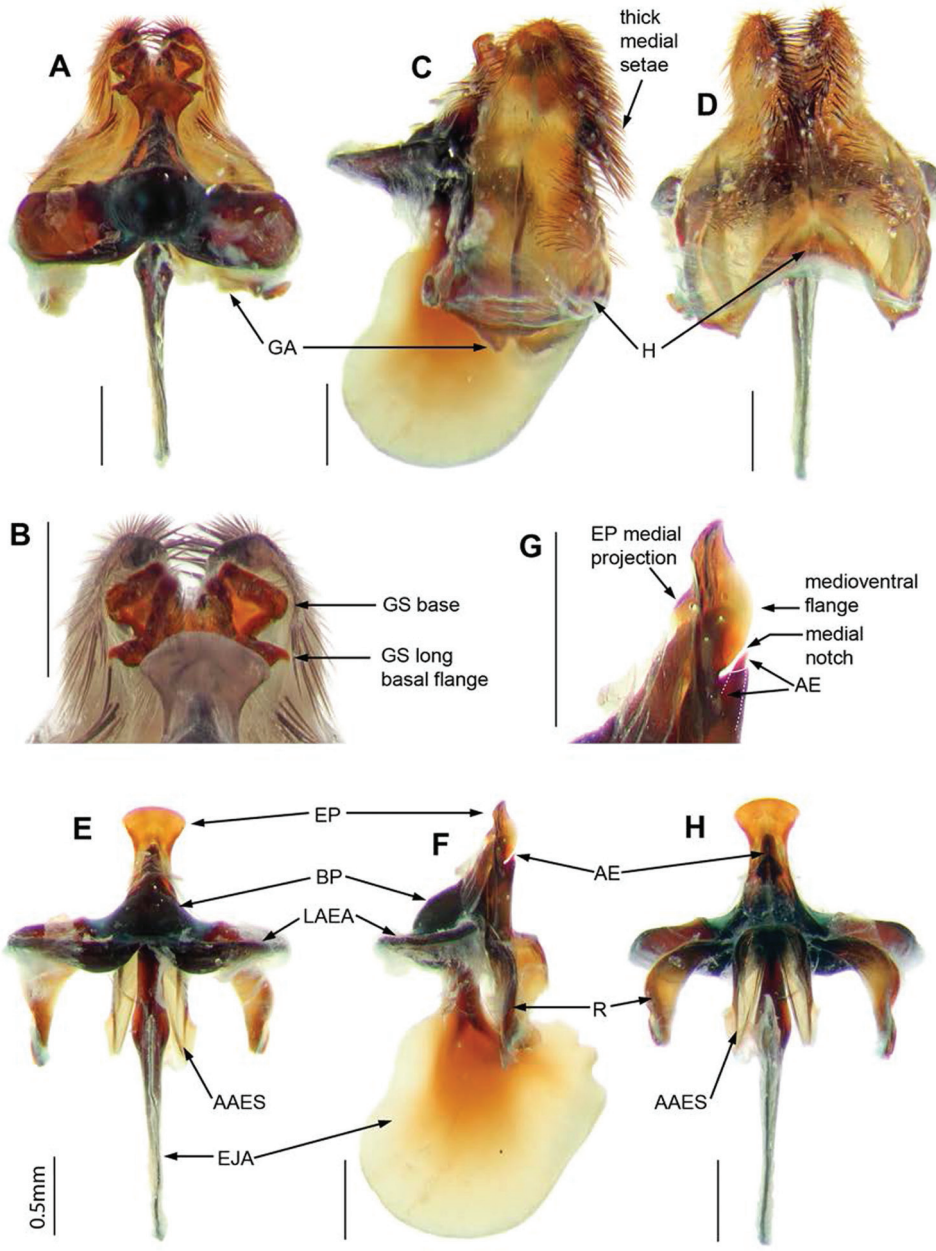


Figure 5. *Palirika culgoafloodplainensis* sp. n., Male holotype genitalia: **A** Gonocoxal complex dorsal **B** Gonocoxal complex lateral **C** Gonocoxal complex ventral **D** Gonostyli **E** Adeagal complex dorsal **F** Adeagal complex lateral **G** Epiphallus lateral **H** Adeagal complex ventral. Scale line = 0.5 mm.

rowly open or closed only at wing margin. Patagium distinct with dense white long flat scales. Widened base of costa with reddish-brown scales, white scales posteriorly. Wing pattern dimidiate (Fig. 4C); with distinct indentation base of first r_{2+3}; extension along R_{4+5}, covering basal 1/3 of first r_{2+3} and r_5; indistinct mottling base of m_1 along m-m; no infuscated band; anal and posterior cells with apically notched hyaline area, infuscation extending along CuA_2; cup infuscated basal 4/5; anal infuscated basal 2/3. Anal basal edge with dense black scales; alula edged with dense reddish-brown scales; squama edged with dense white scales admixed with some reddish-brown scales.

Abdomen. Black, $T_{1–4}$ dark reddish-brown posterolaterally; tergites with bluish-green scales; T_1 with Ma white dorsally, black medially and ventrally, dense very long flattened white scales posterolaterally; $T_{2–7}$ with tufts of long, black setae laterally and posteriorly. Sternites black with dark reddish-brown scales and hairs. *Genitalia* (Figs 4E–H, 5A–H). Epandrium strongly convex, red with convex apical margin; tapering basal flange; long, black setae loosely grouped apically; SES large, fused medially (see Fig. 4E). Gonocoxae red, narrowed apically; GA short, triangular; thick tufts of long black setae medially, directed basally, longest on weak ventral ridge (Fig. 5C); EJA very large, extending well beyond gonocoxal margins, racquet-shaped; LAEA very large, extending well beyond G margins, deeply convex (Fig. 5A); AAES strong wedges (Fig. 5E, H); GS (Fig. 5B) cupped within G margins, large subquadrate base projecting apically; EP long, expanded slightly apically, without lateral lobes, short medial projection; medioventral flange above AE present (Fig. 5G); large recurved R (Fig. 5F, H); H triangular, projecting slightly in lateral view (Fig. 5C).

Female. Unknown.

Etymology. This species is named *culgoafloodplainensis* after the remote Queensland Culgoa Floodplain National Park where the type specimen was collected, and where CLL and Noel Starick received so much hospitality, enthusiasm, and encouragement over the years from all the staff, but especially RIC Andy Coward.

Distribution. (Fig. 11). This species has only been collected from the type locality in central south-western Queensland.

Remarks. Due to extended storage in propylene glycol as retrieval of sample was prevented by extensive and prolonged flooding the specimen bears few setae, hairs or scales, therefore colour patterns referred to in the description are based on those remaining, usually at junctions of sclerites.

***Palirika mackenziei* Lambkin, sp. n.**
urn:lsid:zoobank.org:act:CD50600C-901E-46CA-840F-D4FEA863AF0E
http://species-id.net/wiki/Palirika_mackenziei
Figs 1A–B, 2–3, 6, 11; Morphbank Collection 692335

Material examined. *Holotype.* **Queensland:** ♀, 26°43.7'S, 142°39.1'E, Plevna Downs, Tompilly Hill summit, 13 Dec 2007, C. Lambkin, N. Starick & R. Mackenzie, 15454, sweep net, hilltopping, 220m, [dissected], PS1893, A007533, T152481 (QM). Condition: Good.

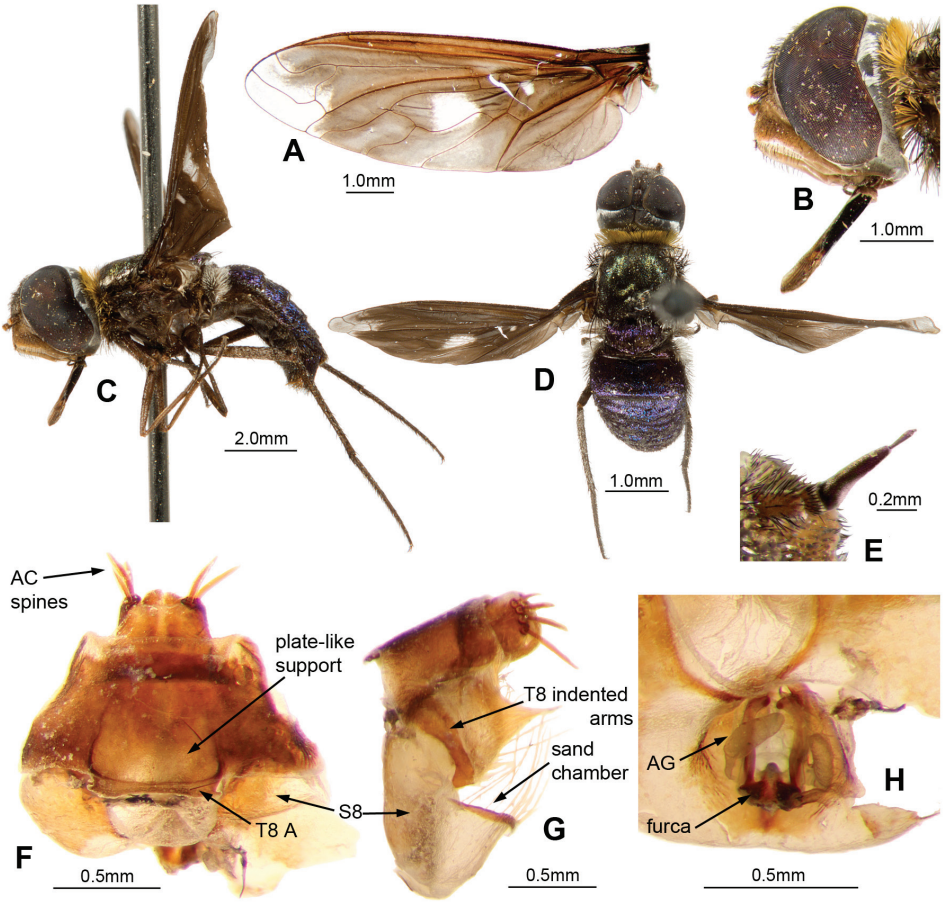


Figure 6. *Palirika mackenziei* sp. n., Female holotype. **A** Wing **B** Head lateral **C** Adult, lateral **D** Adult, dorsal; Male genitalia: **E** Antennae lateral; Genitalia: **F** Dorsal **G** Lateral **H** Dorsal, furca.

Diagnosis. Wing length 9.0 mm.

Small dark flies with heavily infuscated wings, hyaline only apically and medial spot. Face orange with shiny reddish-brown scales, frons black with shiny black scales. Collar yellow. Narrow laterothoracic stripe of whitish scales. Scutum black with dull lime-green metallic scales except pinkish metallic scales anterolaterally and posteromedially. Scutellum dark brown, darker basally with royal-blue metallic scales, purple metallic scales laterally and posteriorly. Widened base of costa with shiny reddish-brown scales, no paler scales posteriorly. Wing pattern broadly dimidiate (Fig. 6A); black with hyaline areas, apically and medially. Apical hyaline area covering extreme apex of r_1, apex of first r_{2+3}, apical half of second r_{2+3}, all r_4, and extreme apex r_5. Medial hyaline area covering middle of dc extending from M_1 across m-cu and into m_2. Paler prediscoidal opaque area distinct. Alula and squama edged with long broad grey scales. T_1 with white Ma;

long white flattened scales posterolaterally. Tergites black with royal-blue metallic scales that reflect purple (Fig. 6D). Female T_8 A short, plate-like support distinct.

Description. *Female. Head.* (Figs 6B–E). Face orange with shiny reddish-brown scales, frons black with shiny black scales; setae black, longest below distinct frontal depression. Antennal scape red, 3 × length of pedicel, with long black setae dense laterally and ventrally; pedicel red with black setae shorter and sparser dorsally; PP conical, 5 × length of pedicel, black with silvery pruinescence, distinct apical joint; BSM rod-like, expanded apically, reddish-brown, 2 × length of pedicel; ASM reddish, conical, length less than width of BSM (Fig. 6E). Occiput with shiny black scales broadly filling indentation.

Thorax. (Figs 6B–D). Collar yellow. Narrow laterothoracic stripe of whitish scales. Long white flattened scales posteromedially. Scutum black with dull lime-green metallic scales except pinkish metallic scales anterolaterally and posteromedially; black setae. AN with black Ma, long yellow setae anteriorly; long, lightly iridescent scales at base of wing black. Prealar bristles strong, black and long. Postalar bristles strong, black and reaching almost apex of scutellum. Pleural hairs black with reddish iridescence. Ma of LT black, reddish-brown dorsally. Tympanal ridge and PL with yellowish flattened scales. Scutellum dark brown, darker basally with royal-blue metallic scales, purple metallic scales laterally and posteriorly; strong, black apical bristles. *Legs*. Reddish-brown, tarsi darker; scales and setae black. Pulvilli sharp, curved, half length of mid- and hind-tarsal claws. Halter knob dark reddish-brown, posteromedial edge broadly yellow. *Wing* (Fig. 6A), cup open. Widened base of costa with shiny reddish-brown scales, no paler scales posteriorly. Wing pattern broadly dimidiate; black with hyaline areas, apically and medially. Apical hyaline area covering extreme apex of r_1, apex of first r_{2+3}, apical half of second r_{2+3}, all r_4, and extreme apex r_5. Medial hyaline area covering middle of dc extending from M_1 across m-cu and into m_2. Paler prediscoidal opaque area distinct. Anal edged with long black scales basally. Alula edged with long broad grey scales. Squama edged with dense overlapping long grey scales.

*Abdomen. (*Figs 6C–D). T_1 with white Ma; long white flattened scales posterolaterally. Tergites black with royal-blue metallic scales that reflect purple when not viewed dorsally, pleura with lateral tufts of long black setae on $T_{2–7}$. Sternites black, with black scales and setae. *Genitalia* (Fig. 6F–H). Dorsal T_8 A short, plate-like support distinct; T_{10} with 4 pairs of stout AC spines. Furca with 2 long broad posteriorly directed arms with small recurved hook-like dorsal extensions apically.

Male. Unknown.

Etymology. This species is named *mackenziei* to acknowledge the enthusiasm and interest in all kinds of natural history by the Mackenzie family of Plevna Downs Station where the type, and only, specimen was collected. Since 2007, following the discovery of dinosaurs on their property, together with a large undescribed spider, CLL and Noel Starick have been welcomed by the Mackenzie family. Robyn Mackenzie was thrilled to be helping catch hill-topping beeflies on the summit of Tompilly Hill when the only female specimen of this unusual *Palirika* was captured (Fig. 1A). We have happily instructed the family, the local Natural History Society, students, teachers,

regional property owners and community members on the ins and outs of biodiversity of arid areas, especially the insects.

Distribution. (Fig. 11). This species has only been collected from the type locality in remote far south-western Queensland.

Larrpana Lambkin & Yeates, 2003

urn:lsid:catalogueoflife.org:taxon:d9155668-29c1-102b-9a4a-00304854f820:col20110201
http://species-id.net/wiki/Larrpana

Type species: *Exoprosopa dimidiatipennis* Bowden, 1971: 64.

Rediagnosis. Dimidiate wing pattern as in Figure 7A–B, dark basally, hyaline apically; infuscation forming a distinctly separated, basal triangle leaving the apex of the posterior cubital and anal cells broadly hyaline. Cream laterothoracic stripes, white scale bands on T_3, sparse white scales on $T_{6–7}$. Epandrial basolateral flange longer and broader than the length of the epandrial base. Gonocoxae deeply narrowed medially, tufts of thickened setae on distinct ventral flange that projects basally; EP with rounded, projecting, lateral lobes; short, wedge-shaped AAES; EJA long. Sperm pump long with unpigmented papillae, collar or clear ring surrounding join between BB and thick-walled round SR.

Included species: *Larrpana bushblitz* sp. n., *Larrpana dimidiatipennis*, *Larrpana collessi* Lambkin & Yeates, 2003, *Larrpana zwicki*.

Larrpana bushblitz Lambkin sp. n.

urn:lsid:zoobank.org:act:1CFF61E9-10C6-4185-8135-CDA4DAEB69A9
http://species-id.net/wiki/Larrpana_bushblitz
Figs 1D, 2–3, 7–8, 11; Morphbank Collection 692334

Material examined. *Holotype.* **Western Australia:** ♂, 29.302°S × 116.725°E, Karara, 23.5km ESE Boiada Camp, 356m, 17 Sep 2009, Lambkin, sweeping, 18402, rocky hilltop, hilltopping, PS1714, WAM 82396 (WAM). Condition: Good.

Paratypes. *Western Australia:* 1♂, same data as holotype, T152479 (QM); 1♂, 29.309°S × 116.731°E, Karara, Forest lookout, 24.4km SE Boiada Camp, 17 Sep 2009, 18405, Lambkin, sweeping, 410m, rocky hilltop, hilltopping, [dissected], PS1894, WAM 82397 (WAM); 1♂, same data Forest lookout, A007534, T152480 (QM).

Diagnosis. Wing length 14 mm

Medium, dark, densely setose, flies with black, dimidiate wings with five indistinct yellowish spots (Fig. 7A, B); infuscation indented in 1st r_{2+3}; no lobe or medial band; dc infuscated except for rectangular hyaline area at junction of m-cu and m-m. Thoracic collar yellow. Dorsal surface of thorax, scutellum and abdomen covered with long upstanding setae, producing distinct fluffy appearance. T_3 with uninterrupted white band of upstanding scales narrowing medially, laterally spanning entire tergite. Alula and squama edged with dense long cream scales; proximal 1/3 of anal cell edged with

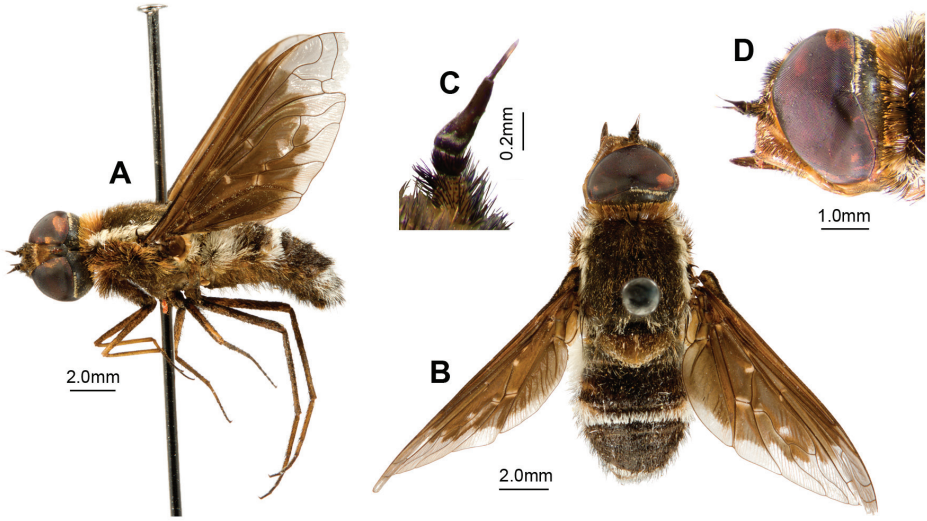


Figure 7. *Larrpana bushblitz* sp. n.. **A** Adult, lateral **B** Adult, dorsal **C** Antennae dorsal **D** Head lateral.

black scales, longest basally. Male (Fig. 8) E with basal flange very long, broad, extending basally, apically recurved. H large, subquadrate in lateral view, distinctly projecting.

Description. *Male. Head* (Fig. 7A–D). Frons reddish-brown, face red, face and frons with transparent scales, black setae longest below shallow frontal depression. Antenna (Fig. 7C). Scape 2.5 × length of pedicel, red; pedicel red; PP long, 3–4 × length of pedicel, as long as scape and pedicel combined, dark reddish-brown, with reddish pruinescence; BSM dark reddish-brown, 2 × length of pedicel, not expanded apically; ASM minute blunt cone, length less than width of BSM. Narrow band of cream scales at posterior margin of eye medially.

Thorax (Fig. 7A, B). Collar yellow. Very broad distinct laterothoracic stripe of dense long white scales. Scutum black; scales long, reddish-brown, white posteriorly; long dense black setae, longest anteriorly and posteriorly. AN and PN with Ma admixed black and reddish-brown; long, slightly iridescent, reddish-brown scales at base of wing. Pleural hairs black, with reddish-brown iridescence. Scales on APA white. Laterotergite with dark reddish-brown Ma ventrally, white dorsally and red medially. Plumula with dense long white scales and TR with dense long yellow scales. Scutellum dark reddish-brown, black basally; scales black basally, transparent pale-brown medially and posteriorly, posterior scales longest; long dense, black setae. Legs reddish-yellow with black scales. Microchaetae on fore-tarsi curved apically. Pulvilli straight sharp cones, more than 1/3 length of mid- and hind-tarsal claws. Halter knob red with pale whitish apical edge. *Wing* (Fig. 7A, B). Widened base of C with black scales with pale band posteriorly. Spur-veins present on base of R_4 extending into r_4 and on apex of m-cu extending into m_2 in some specimens; bump at basal bend of m-cu. Wing pattern (Fig. 7A, B) black, dimidiate, broad basal infuscation following R proximal to i-r_1 to wing margin in apical 4/5 anal cell; indented in 1st r_{2+3}; no lobe or medial band;

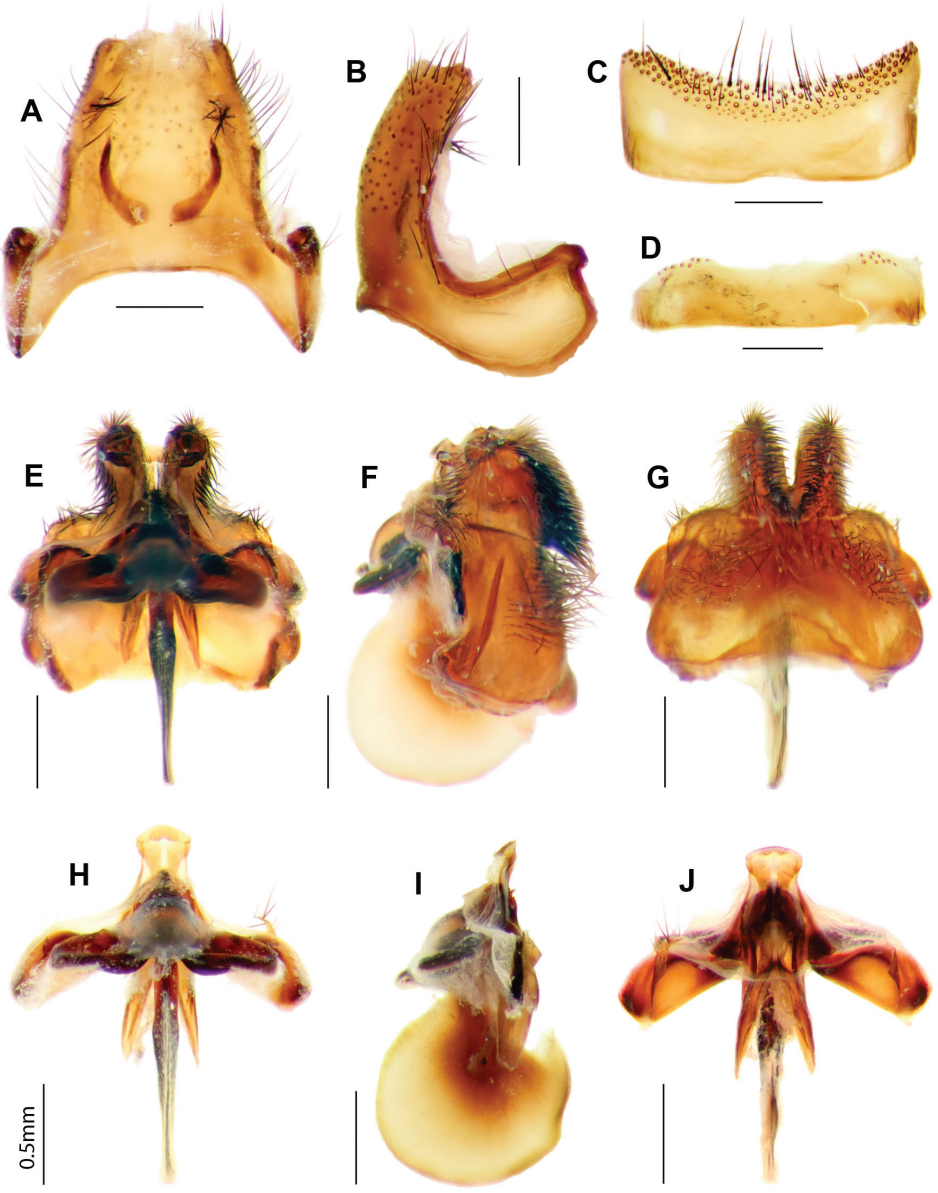


Figure 8. *Larrpana bushblitz* sp. n. Male genitalia: A) Epandrium ventral with sub-epandrial sclerites **B** Epandrium lateral **C** T_8, dorsal **D** S_8, ventral; Male genitalia: **E** Gonocoxal complex dorsal **F** Gonocoxal complex lateral **G** Gonocoxal complex ventral **H** Adeagal complex dorsal **I** Adeagal complex lateral **J** Adeagal complex ventral. Scale line = 0.5 mm.

dc infuscated except for rectangular hyaline area at junction of m-cu and m-m; apex hyaline. Dark yellowish-brown areas bordering base of CuA_1, join of R_1 and R_s, r-m continuing onto base of R_{2+3}, and base m-cu; together with prediscoidal opaque area

forming indistinct pentagonal pattern of spots within infuscation. Anal and cup infuscated for basal 4/5. Alula and squama edged with dense long cream scales; proximal 1/3 of anal cell edged with black scales, longest basally.

Abdomen. (Fig. 7A, B). Tergites black with red anterolateral areas medially rounded on $T_{2–3}$ < 1/4 width of tergite. Scales dense, black except: T_3 with uninterrupted white band of upstanding scales narrowing medially, laterally spanning entire tergite; $T_{1–3}$ with dense long white upstanding lateral scales; $T_{6–7}$ with white scales. T_1 with Ma white dorsally and laterally, yellow ventrally. $T_{1–3}$ with long dense white setae laterally, pale brown anteriorly, and black posteromedially; $T_{4–7}$ with long dense thick black setae. Sternites red with sparse long pale reddish scales, dense long black setae. *Genitalia* (Fig. 8). Epandrium red with distinct anterolateral flange bearing cluster of long black setae; basal flange very large, long, broad, extending basally and apically upcurved; setae black, loosely grouped anterolaterally; SES very long, linear, broadened basally. Gonocoxae red, strongly narrowed medially; ventral ridge projecting; LAEA deeply convex; GS cupped within G margins, subquadrate base projecting apically; very large recurved rami extending beyond G margins; setae long black, not short apically, dense tufts of long, thickened setae medially, directed basally, very long thin setae continuing laterally; H large, subquadrate in lateral view, distinctly projecting. Epiphallus 1.4 × neck width; with apical margins inturned forming projecting rounded lobes.

Female. Unknown.

Etymology. This species is named as a noun in apposition after the three-year, multimillion dollar Bush Blitz program that organised and funded the survey on Charles Darwin Reserve, Karara, Lochada and Kadji Kadji Pastoral Leases in Western Australia on which this species was collected. The core focus of the Bush Blitz program is to document the plants and animals in hundreds of properties across Australia's National Reserve System, and on nature discovery – identifying and describing new species of plants and animals. The Bush Blitz program also funded the survey in western New South Wales and Queensland on which *Palirika culgoafloodplainensis* sp. n. was collected (ABRS BB 2009/23887) and funded the description of these three species (ABRS BB TTG209-06).

Distribution. (Fig. 11). *Larrpana bushblitz* sp. n. has only been collected from Karara Pastoral Lease, 213 km ESE of Geraldton in Western Australia.

Comments. On collection, this species appeared similar to the two male specimens of *Larrpana zwicki* collected only near Windorah (Lambkin et al. 2003) and phylogenetic analysis (Figs 2–3) indicates a close relationship between the two.

Munjua Lambkin & Yeates, 2003

urn:lsid:catalogueoflife.org:taxon:d916e848-29c1-102b-9a4a-00304854f820:col20110201
http://species-id.net/wiki/Munjua

Type species: *Munjua erugata* Lambkin & Yeates, 2003: 795.

Rediagnosis. Wing with medial hyaline band not linear and narrowing apically, apical infuscated band not meeting posterior wing margin more broadly than medial

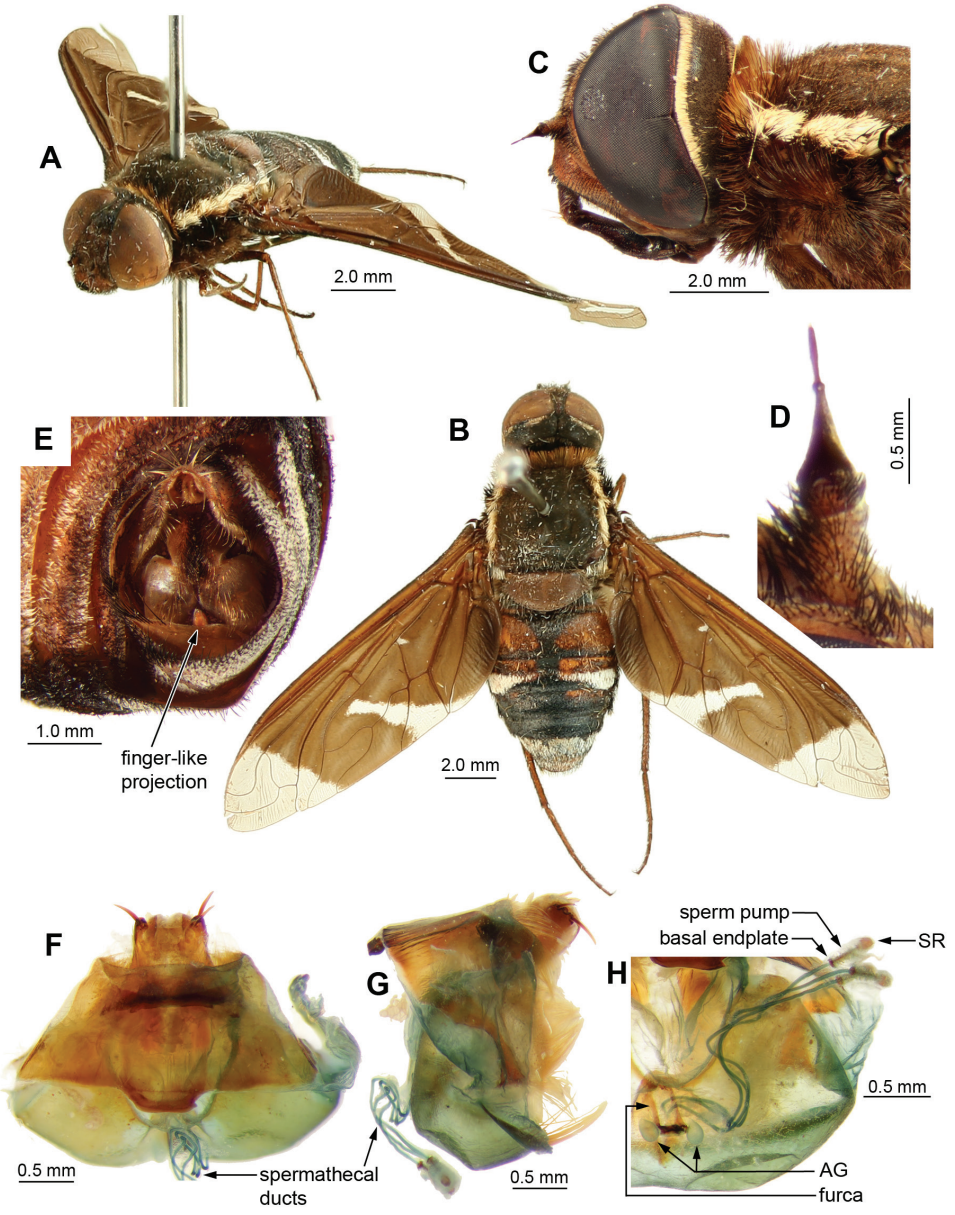


Figure 9. *Ngalki trigonium*. **A** Adult, antero-lateral **B** Adult, dorsal **C** Head and thorax lateral **D** Antennae lateral **E** Male genitalic complex showing diagnostic finger-like projection on hypandrium, clearly visible *in situ*. Female genitalia: **F** Dorsal **G** Lateral **H** Dorsal, furca and spermathecal complex.

hyaline band. Gonocoxae deeply narrowed medially, broadly indented basally, with tufts of thickened setae ventromedially, H projecting but not forming a finger-like extension; AE short; EP with medioventral process above AE; very long AAES reach-

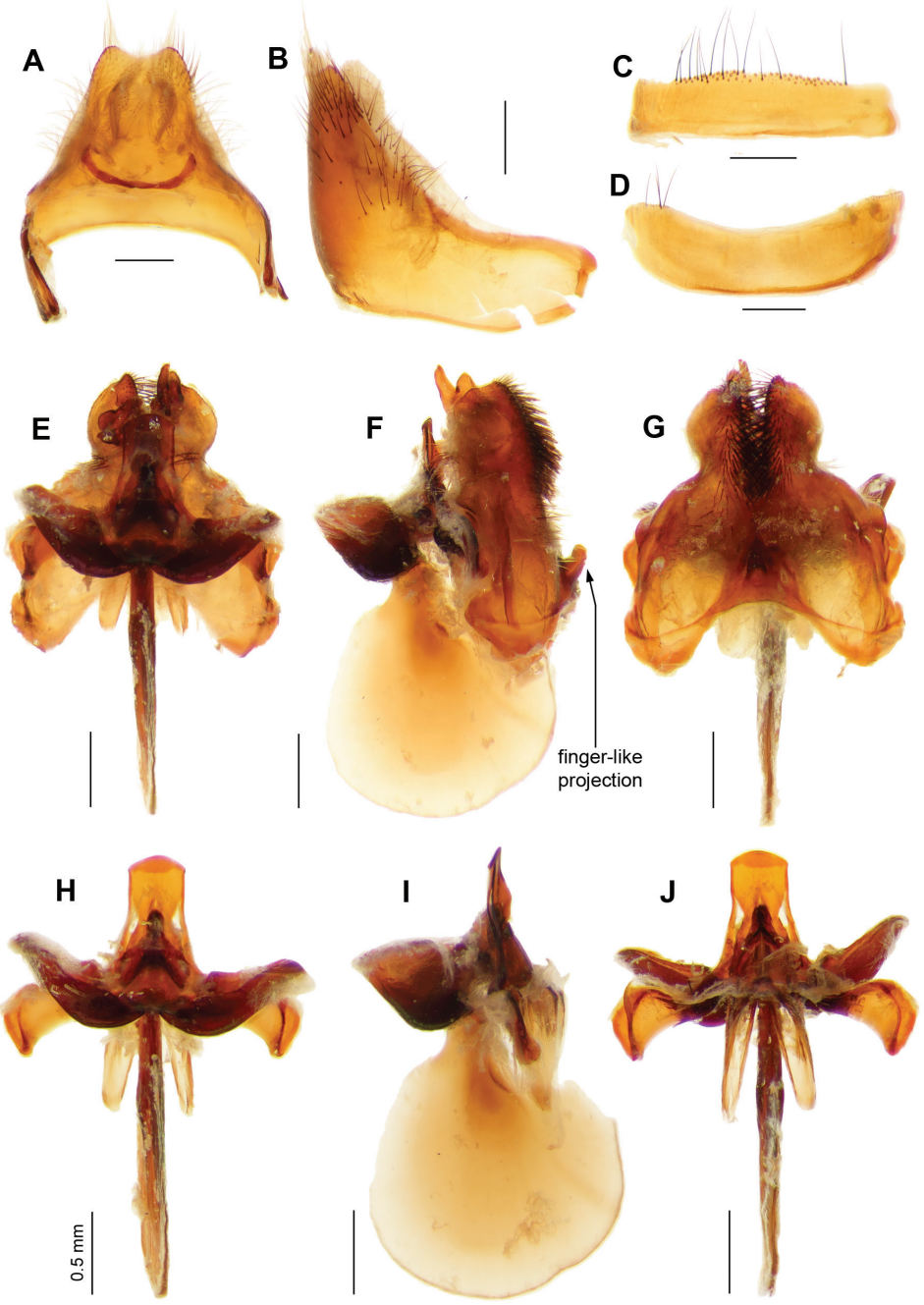


Figure 10. *Ngalki trigonium*. Male genitalia: **A** Epandrium ventral with sub-epandrial sclerites **B** Epandrium lateral **C** T_8, dorsal **D** S_8, ventral **E** Gonocoxal complex dorsal **F** Gonocoxal complex lateral showing diagnostic finger-like projection on hypandrium **G** Gonocoxal complex ventral **H** Adeagal complex dorsal **I** Adeagal complex lateral **J** Adeagal complex ventral. Scale line = 0.5 mm.

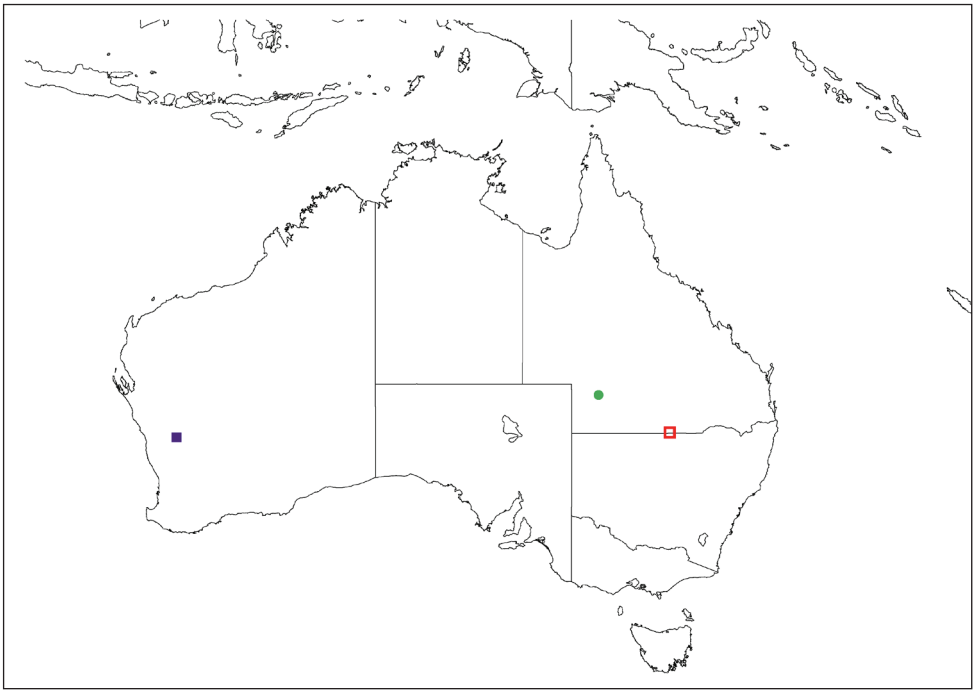


Figure 11. Map of distribution. Closed circle - *Palirika mackenziei* sp. n. Open square - *Palirika culgoafloodplainensis* sp. n. Closed square - *Larrpana bushblitz* sp. n.

ing G margins; EJA racquet-shaped, very long. Sperm pump short with unpigmented papillae, apical endplate simple with thin processes; thick-walled round SR with distinct basal bulb.

Included species: *Munjua erugata* Lambkin & Yeates, 2003, *Munjua lepidokingi* Lambkin & Yeates, 2003, *Munjua paralutea* Lambkin & Paramonov, 2003.

Comments. See reference to the rediagnosis of the genus *Munjua* in the phylogenetic results.

Ngalki Lambkin, gen. n.

urn:lsid:zoobank.org:act:77AACA67-1FA6-4CD9-871E-D95C34FDE977
http://species-id.net/wiki/Ngalki

Type species: *Munjua trigona* Lambkin & Yeates, 2003: 804.

Diagnosis. Wing with medial hyaline band linear and narrowing apically, apical infuscated band meeting posterior wing margin twice breadth of medial hyaline band (Fig. 9A, B). Gonocoxae deeply narrowed medially, broadly indented basally, with tufts of thickened setae ventromedially, H projecting forming finger-like extension (Figs 9E, 10F); AE short; EP with medioventral process above AE; very long AAES reaching G margins; EJA racquet-shaped, very long (Fig. 10). Sperm pump short with

unpigmented papillae, apical endplate simple with thin processes; thick-walled round SR with no basal bulb (Fig. 9F–H).

Etymology. The name for the genus *Ngalki* is from the aboriginal term *ngalki* for "little finger" from the Ngiyampaa language spoken in much of central New South Wales (Donaldson, 1994), referring to the diagnostic character of the male genitalia for this genus, and is treated as neutral. This follows the tradition set in Lambkin et al. (2003), of using appropriate aboriginal terms for the names of new genera of Australian exoprosopines.

Included species: *Munjua trigona* Lambkin & Yeates, 2003

Comments. See reference to the erection of the genus *Ngalki* in the phylogenetic results.

***Ngalki trigonium* (Lambkin & Yeates), comb. n.**

urn:lsid:catalogueoflife.org:taxon:db706730-2dc5-11e0-98c6-2ce70255a436:col20110201
http://species-id.net/wiki/Ngalki_trigonium
Figs 2–3, 9–10; Morphbank Collection 692333

Munjua trigona Lambkin & Yeates, 2003: 804.

Material examined. *Paratypes.* **New South Wales:** 1♂, Round Hill Nature Reserve, 27 Dec 1976, G. Daniels, GDCB Reg # 14199, K 253709; 1♀, Round Hill Nature Reserve, same, GDCB Reg # 17925, K 253717 (AM). **Victoria:** 1♀, Wyperfield Nat Park, 7 Dec 1976, G. Daniels, GDCB Reg # 14163, K 253707; 3♂, Wyperfield Nat Park, 8 Dec 1976, G. Daniels, GDCB Reg # 17924, #14164, # 17923, K 253720, (PS1936) K 253702, K 253712 (AM).

Other material. New South Wales: 1♀, Round Hill area, 24–25 Nov 1991, A. Sundholm, [dissected], PS1935, K 289927 (AM). **Western Australia:** 1♂, Fraser Range, 8 Nov 1977, A. Atkins, [dissected], GDCB Reg # 14165, PS1934, K 253698 (AM).

Rediagnosis. Large dark flies (wing length 15 - 20 mm), wings as in Figure 9B, dark with narrow, linear, medial hyaline band; long, finger-like apically-directed projection on hypandrium (Figs 9E, 10F). Laterothoracic stripe creamy-white. Broad white scale band on T_3, $T_{4–5}$ with black scales, $T_{6–7}$ with white scales. Epandrium with long golden setae; SES joined medially. Epiphallus with short medial projection. Female with no BB between pump and pale, square SR.

Redescription. *Male. Head* (Fig. 9C). Face and frons red with reddish-yellow scales, black setae longest below deep frontal depression. Antennal scape 3 × length of pedicel, red; pedicel red; PP 3.5 × length of pedicel, black with reddish pruinescence; BSM dark reddish-brown, 3.5 × length of pedicel, expanded apically; ASM short, blunt (Fig. 9D). Narrow line of creamy-white scales on posterior margin of eye.

Thorax. Collar yellow, with tips of Ma darker, reddish. Laterothoracic stripe broad creamy-white, distinct. Scutum black with long hair-like reddish-brown scales, sparse white flattened scales posteromedially. Pleural vestiture dark-red with mauve iridescence; admixed dark-red and black Ma on PE and AN; long scales at base of wing with mauve iridescence. Anepimeral setae admixed dark-red and black. Scales

on APA yellow. Plumula and TR hairs creamy yellow. LT with black Ma, red dorsally. Scutellum red with scales yellowish-red, setae black, sparse long white scale fringe on posterior margin. Legs red with dark reddish-brown scales. Pulvilli chisel-like wedges, less than a half length of mid- and hind-tarsal claws. Halter knob dark reddish-brown, with yellow apical edge. *Wing* (Fig. 9B). R_{4+5} with abrupt bend basally, cup closed at wing margin, or narrowly open. Widened base of C with dark, reddish-brown scales, paler brown scales posteriorly. M_2 very sinuous, width of m_1 at wing margin < 1/2 width of m_2. Wing pattern (Fig. 9A, B) black, with broad infuscated band following R proximal to i-r_1, obliquely through 1st r_{2+3}, r_5 and m_1 to meet wing margin in m_2; no infuscation in 2nd r_{2+3}. Hyaline band very narrow, linear, from dc through m_2 basally, apex of cua, cup and anal cells. Squama with reddish-yellow scales.

Abdomen (Fig. 9B). Integument with large red areas laterally, leaving black medial band; T_2 with black medial band broad basally, width > 1/3 width of the tergite, tapering sharply apically; $T_{3–4}$ with black medial band < 1/4 width, black apical band; $T_{5–7}$ with red areas laterally < 1/4 width. T_1 with Ma white, reddish-brown scales medially. T_2 with dense long cream hairs anterolaterally, dark reddish-brown scales, some white scales laterally. Broad white scale band on T_3, interrupted medially, dark reddish-brown scales anteromedially, black scales posteromedially. $T_{4–5}$ with black scales, $T_{6–7}$ with white scales. Sternites red; $S_{2–3}$ and $S_{4–5}$ basally with dense white scales and dense long, fine white hairs; $S_{4–5}$ apically with dark reddish-brown scales, $S_{6–7}$ with black scales, black setae. *Genitalia* (Figs 9E, 10). Epandrium red with basal flange short and broad; setae black, long golden setae apically; SES joined medially. Gonocoxae red, strongly narrowed medially; setae black, short apically, medially with thick tufts of basally directed setae, long setae laterally around apex of H; distinct ventral ridge; LAEA deeply convex, extending past G margins; GS cupped within G margins, large subquadrate base with slight projection apically; large recurved R; H crescent-shaped, laterally delimited by swollen and expanded G, laterally subrectangular, with distinct large, blunt finger-like apical projection. Epiphallus long, not expanded apically; with medial projection.

Female. Same as male. *Genitalia* (Fig. 9F–H). Dorsal T_8 A short, entire; T_{10} with 4 pairs of short, thick AC spines; apical endplate with long thin processes; basal endplate with thick processes; long unpigmented papillae, no BB, narrow ring between pump and lightly pigmented, square SR.

Etymology. In Lambkin et al. (2003), the name *trigona* given to this species was derived from the Greek *trigonas* "triangular" This was the name the late Sergei Paramonov gave to this species in his unpublished manuscript, and was used to honour his extensive work on Australian bombyliids. With the transfer to *Ngalki*, the specific emendation requires adjustment, and becomes *trigonium* to reflect the neutral gender of the new genus-group name.

Distribution. This species has been collected in the southern Australian Bassian region, from semi-arid and arid mallee areas.

Comments. The finger-like projection on the H (Figs 9E, 10F) in the males of *Ngalki trigonium* is apparent without dissection and, together with the unusual wing pattern, allows easy identification.

Keys

Key to the Genera of the Australian *Balaana* Group

1 Metallic scales on body, black reflecting blue, bluish-black, green or maroon, no white or yellow scales on $T_{2–7}$ or on $S_{2–7}$ (Fig. 6D)..... .....***Palirika*** **Lambkin & Yeates**

– No metallic reflecting black scales; white or yellow scales present $T_{2–7}$, usually distinct bands or lateral triangles on T_3 (Fig. 9B) **2**

2 (1) Wing dimidiate and at least apical half of anal cell margin hyaline, at most short narrow lobe following m-m into m_2, no medial hyaline band (Fig. 7B) ..... ***Larrpana*** **Lambkin & Yeates**

– Wing not dimidiate or anal cell fully infuscated; distinct medial hyaline band usually present (Fig. 9B) **3**

3 (2) Female spermathecal reservoir a long cylinder; T_2 with black scales; $T_{6–7}$ with white scales; proboscis extending beyond oral cavity, not longer than head ... . ***Balaana*** **Lambkin & Yeates**

– Female spermathecal reservoir round to subquadrate never a long cylinder; T_2 with some yellow scales unless T_6 or T_7 with black scales or proboscis longer than head **4**

4 (3) Male with no medioventral process on epiphallus above aedeagus, anterior arms of aedeagal sheath long, reaching gonocoxal margins; quadrate sub-epandrial sclerites in epandrium. EITHER Deeply infuscated wings, only apex hyaline; paler yellowish spots at base of R_{2+3}, at base of CuA_1, join of R_1 and R_s, r-m and base of m-cu; T_6 black scales; OR medial hyaline band a narrow line; black scales forming median circle apex of T_2 and base of T_3, yellow scales anteriorly and laterally on T_2, medially and laterally on T_3 ....***Muwarna*** **Lambkin & Yeates**

– Male with medioventral process on epiphallus above aedeagus (Figs 5G, 10I), linear or single fused (Fig. 10A) sub-epandrial sclerites in epandrium. Wings less infuscated with broad medial hyaline band broader posteriorly; IF deeply infuscated with medial hyaline band a narrow line (Fig. 9B) no median circle of black scales on $T_{2–3}$ **5**

5 (4) Male with anterior arms of aedeagal sheath long, reaching gonocoxal margins (Fig. 10E). Ventral ridge on gonocoxae small or absent, not projecting basally AND hypandrium projecting. Epiphallus without lateral lobes; ejaculatory apodeme extending beyond gonocoxae by more than length of gonostylus (Fig. 10E). EITHER yellow vestiture with wing infuscation distinctly variegated, bright yellow basally and medial band dark brown to black; OR no yellow scales on $T_{2–7}$ (Fig. 9B); OR only yellow scales anteromedially on $T_{2–3}$, $S_{2–3}$ with dense, white scales and setae, $S_{5–7}$ with dense, black scales and setae.....**6**

– Male with anterior arms of aedeagal sheath short, not reaching gonocoxal margins; if ventral ridge on gonocoxae very small or absent then hypandrium

not projecting. Abdominal yellow scales at least anteriorly T_2, $S_{2–3}$ with dense, white scales; hemispherical tufts of macrochaetae laterally on T_1 white or yellow, not dark reddish-brown or black; wing infuscation not distinctly variegated, IF yellow scales only anteromedially T_2 then scales on $S_{5–7}$ not black, at most reddish-brown **7**

6 (5) Medial hyaline band linear, narrowing anteriorly, with apex of anal cell and cup hyaline (Fig. 9B); male gonocoxae with long, finger-like projection from hypandrium (Figs 9E, 10F) ..... ***Ngalki*** **Lambkin, gen. n.**

– Medial hyaline band not linear, not narrowing anteriorly; male gonocoxae with no finger-like projection from hypandrium ..... ***Munjua*** **Lambkin & Yeates**

7 (5) Male epandrium with strongly grouped setae on anterolateral flange; ventral ridge on gonocoxae large, distinctly projecting basally, hypandrium not projecting, epiphallus with rounded projecting lateral lobes, ejaculatory apodeme short, extending beyond gonocoxae by less than length of gonostylus, hind-tibial scales not protruding, dark flies, T_2 and T_4 mostly black scales ... . ***Kapu*** **(Lambkin & Yeates)**

– Male epandrium with loose setae, without an anterolateral flange; hind-tibial scales protruding, pale yellowish flies with striped abdominal vestiture, T_2 and T_4 mostly yellow scales ..... ***Wurda*** **Lambkin & Yeates**

Key to Species of *Palirika*

1 No pre-apical infuscated band on wing (Figs 4C, 6A) **2**

– Pre-apical infuscated band on wing present **3**

2 (1) Infuscation of wing blade almost complete except for hyaline apical area and isolated spot over dc (Fig. 6A) ..... ***Palirika mackenziei*** **Lambkin, sp. n.**

– Infuscation of wing blade only extending over half wing area, indistinct extension along R_{4+5} and isolated mottled area along m-m (Fig. 4C) .... ***Palirika culgoafloodplainensis*** **Lambkin, sp. n.**

3 (1) Anal and posterior cells with notched hyaline area, infuscation extending along CuA_2; apically-directed spur-vein on i-r_1 cross vein..... ..... ***Palirika bouchardi***

– Anal and posterior cells without extension along CuA_2, rarely spur-vein on i-r_1 cross vein **4**

4 (3) Hyaline medial band continues anteriorly through entire r_5; dark thorax; abdomen: males dark prussian-blue, almost black; females bluish-green .... ***Palirika cyanea***

– Hyaline band usually through dc anteriorly, not entirely through r_5, or absent; thorax and abdomen not as above **5**

5 (4) Collar white with contrasting tuft of black lateral Ma at base of pronotal lobe, bright green thorax, anterolaterally dark maroon; bright, dark blue abdomen; anal and cup fully infuscated ..... ***Palirika decora***

– Collar entirely white or yellow, at most 4 reddish Ma above postpronotal lobe; thorax, abdomen, anal and cup not as above **6**

6 (5) Face yellow, most facial setae shiny gold; blue-green thorax and abdomen .... ***Palirika viridula***

– Face orange-red to reddish-brown; if yellow, most facial setae black; thorax and abdomen not as above **7**

7 (6) Thorax green **8**

– Thorax dark, not green **11**

8 (7) Thorax anterolaterally with dark maroon scales; anal and cup infuscation various **9**

– Thorax entirely green; anal and cup hyaline apically **10**

9 (8) Abdomen bluish-green; brown wing infuscation, anal and cup hyaline apically; blue face scales ..... ***Palirika whyalla***

– Abdomen entirely purple; black wing infuscation, anal and cup fully infuscated; purple face scales ..... ***Palirika anaxios***

10 (8) Thorax bright yellowish-green; abdomen blue to bluish-green metallic scales; infuscated wing band short, much narrower than hyaline band..... ..... ***Palirika marginicollis***

– Thorax dark bluish-green; abdomen T_2 blue-green, $T_{3–7}$ blue, $T_{4–6}$ admixed maroon at least laterally; infuscated wing band broader than hyaline band ... . ***Palirika blackdownensis***

11 (7) Abdomen purple with blue scales on $T_{4–6}$, Queensland ..... ***Palirika danielsi***

– Abdomen entirely purple, no blue scales on $T_{4–6}$, Western Australia..... ..... ***Palirika basilikos***

Key to Species of *Larrpana*

1 No yellow scales on $T_{2–7}$ (Fig. 7A, B), Ma on T_1 black, white or yellow..... **2**

– Abdominal yellow scales at least anteriorly T_2; Ma on T_1 white, not black .... ***Larrpana collessi***

2 (1) Wing without small paler yellowish spots in infuscation, m-m without infuscation..... ***Larrpana dimidiatipennis***

– Wing with small paler yellowish spots in infuscation; m-m with infuscation (Fig. 7A, B)..... **3**

3 (2) Wing without short narrow lobe following m-m into m_2 (Fig. 7A, B)..... ..... ***Larrpana bushblitz* Lambkin, sp. n.**

– Wing with short narrow lobe following m-m in m_1 into m_2..... ..... ***Larrpana zwicki***

Acknowledgments

Firstly, we thank the Bush Blitz program partners and the program managers from the Australian Biological Resources Study (ABRS), a section within the Parks Division of the Department of Sustainability, Environment, Water, Population and Communities (especially Brooke Glasser, Annabel Wheeler, and Kate Gillespie) for organising and funding the Bush Blitz field work on Charles Darwin Reserve, Karara, Lochada and Kadji Kadji Pastoral Leases in WA, funding the Bush Blitz survey of Culgoa Floodplains NP Qld, Culgoa NP and Ledknapper NR NSW (ABRS BB 2009/23887), and funding this descriptive work (ABRS BB TTG209-06). We thank the Council of Heads of Australian Faunal Collections (CHAFC) for funding the employment of Rhys Smith, Kathy Ebert, Kathleen Nugent, Wendy Hebron, and Karin Koch at QM for sorting, databasing and curating of collected specimens from the Bush Blitz survey of Culgoa Floodplains NP, Culgoa NP and Ledknapper NR. We thank the QM and especially John Hooper (Head Biodiversity & Geosciences Programs) for supporting our participation in the Bush Blitz program. CLL thanks Geoff Monteith and Noel Starick (QM, Brisbane), Catherine Young (TMAG, Hobart), Celia Symonds (UNSW, Sydney), Remko Leijs (SAM, Adelaide), Ray Mjadwesch (Mjadwesch Environmental Service, Bathurst) for their company and help in the field. We thank the numerous QM volunteers who have willingly worked on sorting the massive amount of material collected from the Bush Blitz program: Rhys Smith, Noel Starick, Jackie Chan, and John Purdie. CLL and Noel Starick thank all the National Park staff who have always encouraged our work, often accompanied us in the field, taken numerous samples under difficult circumstances ranging from intense dust storms to record breaking floods, and provided accommodation, great company and good advice: Andy Coward, Megan Simpson, Cheryn Kelly, and Stephen Peck (Culgoa Floodplains NP Qld); Rick Ohlsen and Bart Schiebaan (Culgoa NP NSW); and Shayne OSullivan (Ledknapper NR, NSW).

CLL and Noel Starick thank the Mackenzie family of Plevna Downs Station for the enthusiasm and interest they always show in all forms of natural history, the welcome and hospitality they have given to us and all staff from the Queensland Museum, and the encouragement they provide to the local Natural History Society, regional property owners and community members on development of a knowledge base of biodiversity of arid areas. For providing collection permits to collect and take samples from National Reserves, we thank Jacqui Brock (Scientific Permit W1TK05498008: Queensland Parks and Wildlife Service, DERM, QLD) and Brendon Neilly (Scientific Research Licence S10016: NPWS, OEH, NSW). We thank David Britton and Jacqui Recsei (AM, Sydney) for supplying specimens of *Ngalki trigonium* for dissection and photography.

We thank Neal Evenhuis and an anonymous reviewer for their positive comments on the manuscript. Lastly, but not least, we gratefully acknowledge the work of Geoff Thompson (QM) for photographing specimens and guidance in the use of the imaging systems, Paul Avern (QM) for help with writing the occurrence data in Darwin Core format, Federica Turco (QM) for production of the distribution map and help with the

GBIF and Dryad uploads, Vladimir Blagoderov, (Natural History Museum, London) for help with Scratchpads, Debbie Paul (School of Computational Science, Florida State University Tallahassee) for assistance with Morphbank, and Lyubo Penev and Teodor Georgiev (Pensoft Publishers, Sofia, Bulgaria) for their support and assistance with data publication through GBIF and Dryad.

References

ABRS (2009) Australian Faunal Directory. In 'Australian Faunal Directory', Australian Biological Resources Study, Canberra.

Adobe Systems (2010a) Adobe Illustrator C S5 version 15.0.2.

Adobe Systems (2010b) Adobe Photoshop C S5 version 13.0.3 × 32.

Bezzi M (1908) Eine neue Aphoebantus-Art aus den palaearktischen Faunengebiete (Dipt.). Zeitschrift fÅr Systematische Hymenopterologie und Dipterologie. Mecklenberg 8.

Blagoderov V, Brake I, Georgiev T, Penev L, Roberts D, Rycroft S, Scott B, Agosti D, Catapano T, Smith VS (2010a) Streamlining taxonomic publication: a working example with Scratchpads and ZooKeys. ZooKeys 50: 17–28. doi: 10.3897/zookeys.50.539

Blagoderov V, Hippa H, Nel A (2010b) *Parisognoriste*, a new genus of Lygistorrhinidae (Diptera: Sciaroidea) from the Oise amber with redescription of *Palaeognoriste* Meunier. ZooKeys 50: 79–90. doi: 10.3897/zookeys.50.506

Bowden J (1971) A Note on the name *Exoprosopa dimidiata* Roberts (Diptera: Bombyliidae). Journal of the Australian Entomological Society 10: 64. doi: 10.1111/j.1440-6055.1971.tb00012.x

Brake I, von Tschirnhaus M (2010) *Stomosis arachnophila* sp. n., a new kleptoparasitic species of freeloader flies (Diptera, Milichiidae). ZooKeys 50: 91–96. doi: 10.3897/zookeys.50.505

Bremer K (1992) Ancestral areas: a cladistic reinterpretation of the center of origin concept. Systematic Biology 41: 436–445.

Bremer K (1994) Branch support and tree stability. Cladistics 10: 295–304. doi: 10.1111/j.1096-0031.1994.tb00179.x

Chapman AD (2009) Numbers of living species in Australia and the World. Australian Biological Resources Study, Canberra, 61 pp.

Donaldson T (1994) Ngiyampaa. In: Thieberger N, McGregor W (Eds) Macquarie Aboriginal words : a dictionary of words from Australian Aboriginal and Torres Strait Islander languages, Macquarie Library, Macquarie University, North Ryde, NSW, 23–40.

ESRI (1998) ArcView GIS version 3.1. Environmental Systems Research Institute, Inc., Redlands, California.

Evenhuis NL (1981) Studies in Pacific Bombyliidae (Diptera) VI. Description of a new anthracine genus from the Western Pacific, with notes on some of Matsumura's *Anthrax* types. Pacific Insects 23: 189–200.

Evenhuis NL (1991) Studies in Pacific Bombyliidae (Diptera). 10. Bombyliidae of New Caledonia. In: Chazeau J, Tillier S (Eds) Zoologia Neocaledonica Volume 2, Mémoires du Muséum national d'Histoire naturelle. Zoologie, Tome 149, Paris, 279–288.

Evenhuis NL, Greathead DJ (1999) World Catalog of Bee Flies (Diptera: Bombyliidae). Backhuys, Leiden, Netherlands.

Farris JS (1990) Phenetics in camouflage. Cladistics 6: 91–100. doi: 10.1111/j.1096-0031.1990.tb00528.x

Gray G (1883) Notices of new genera and species. In: Cuvier B (Ed) The Animal Kingdom arranged in conformity with its organization. Whittaker, Treacher, and Co., London, 780 pp.

ICZN (2008) Proposed amendment of the International Code of Zoological Nomenclature to expand and refine methods of publication. Zootaxa 1908: 57–67.

Källersjö M, Farris JS, Kluge AG, Bult C (1992) Skewness and permutation. Cladistics 8: 275–287. doi: 10.1111/j.1096-0031.1992.tb00071.x

Kozub D (2011) Helicon Focus version 5.2 Pro (Helicon Soft Ltd.: Kharkov, Ukraine.)

Lambkin CL, Yeates DK, Greathead DJ (2003) An evolutionary radiation of bee flies in semi-arid Australia: Systematics of the Exoprosopini (Diptera: Bombyliidae). Invertebrate Systematics 17: 735–891. doi: 10.1071/IS03020

Lioy P (1864) I ditteri distribuiti secundo un nuovo metodo di classificazione naturale [part]. Atti dell' I.R. Istituto Veneto di Scienze Lettere ed Arti (3) 9: 719–71.

Littlefield R (2011) Zerene Stacker LLC, version 1.02. (Zerene Systems: Richland, Washington.)

Macquart J (1840) Diptères exotiques nouveaux ou peu connus. Roret, Paris. 21plates, 5–135 pp.

Macquart J (1846) Diptères exotiques nouveaux ou peu connus. Supplément. Mémoirs de la Société Royale des Sciences, de l'Agriculture et des Arts, de Lille 1844: 133–364.

Maddison DR, Maddison WP (2003) MacClade 4. Analysis of Phylogeny and Character Evolution. Version 4.06 for OS X. (Sinauer Associates, Inc.: Sunderland, Massachusetts.)

Maddison WP, Maddison DR (2010) Mesquite: A modular system for evolutionary analysis. Vers. 2.74. (http://mesquiteproject.org.)

Margush T, McMorris F (1981) Consensus n-trees. Bulletin of Mathematical Biology 43: 239–244.

McAlpine J (1981) Morphology and terminology: Adults. In: McAlpine JF, Peterson BV, Shewell GE, Teskey HJ, Vockeroth JRWood DM (Eds) Manual of Nearctic Diptera. Research Branch Agriculture Monograph No. 27. Canadian Government Publishing Centre: Ottawa, 9–63.

Mickevich MF, Mitter C (1981) Treating polymorphic characters in systematics: A phylogenetic treatment of electrophoretic data. In: Funk V, Brooks D (Eds) Advances in cladistics. Proceedings of the first meeting of the Willi Hennig Society. New York Botanical Garden, Bronx, 45–58.

Mickevich MF, Weller SJ (1990) Evolutionary character analysis: tracing character change on a cladogram. Cladistics 6: 137–170. doi: 10.1111/j.1096-0031.1990.tb00533.x

Newman E (1841) Entomological notes [part]. Entomologist 1: 220–223.

NHAS (2009) Australian Natural Heritage Assessment Tool. In: Australian Natural Heritage Assessment Tool. Natural Heritage Assessment Section, Department of Sustainability, Environment, Water, Population and Communities, Australia.

Nixon KC (2002) WinClada version 1.00.08. L.H. Bailey Hortorium, Cornell University, Ithaca, New York. [http://www.cladistics.com/about_winc.htm]

Osten Sacken CR (1877) Art. XIII. Western Diptera: Descriptions of new genera andspecies of Diptera from the region west of the Mississippi and especially from California. Bulletin United States Geological Survey: 189–354.

Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith VS, Brake I, Rysrcroft S, Scott B, Johnson NF, Morris RA, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress J, Erwin T (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1–16. doi: 10.3897/zookeys.50.538

Polaszek A, Agosti D, Alonso-Zarazaga M, Beccaloni G, de Place Bjørn P, Bouchet P, Brothers DJ, Cranbrook Eo, Evenhuis NL, Godfray HCJ, Johnson NF, Krell FT, Lipscomb D, Lyal CHC, Mace GM, Mawatari SF, Miller SE, Minelli A, Morris S, Ng PKL, Patterson DJ, Pyle RL, Robinson N, Rogo L, Taverne J, Thompson FC, van Tol J, Q.D. W, Wilson EO (2005a) Commentary: A universal register for animal names. Nature 437: 477.

Polaszek A, Alonso-Zarazaga M, Bouchet P, Brothers DJ, Evenhuis NL, Krell FT, Lyal CHC, Minelli A, Pyle RL, Robinson N, Thompson FC, van Tol J (2005b) ZooBank: the open-access register for zoological taxonomy: technical discussion paper. Bulletin of Zoological Nomenclature 62: 210–220.

Scopoli IA (1763) Entomologia Carniolica exhibens Insecta Carnioliae Indigena et Distributa in Ordines, Genera, Species, Varietates. Methodo Linneana.

Sorenson MD (1999) TreeRot, version 2. (Boston University: Boston, MA.)

Swofford DL, Begle DP (1993) PAUP: Phylogenetic Analysis Using Parsimony Version 3.1.1 User's Manual. Illinois Natural History Survey, Champaign.

Swofford DL (2002) PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b.10. Sinauer Associates, Sunderland, Massachusetts.

Winterton SL (2009) Revision of the stiletto fly genus *Neodialineura* Mann (Diptera: Therevidae): an empirical example of cybertaxonomy. Zootaxa 2157: 1–33.

Winterton SL, Gaimari SD (2011) Revision of the South American window fly genus *Heteromphrale* Kröber, 1937 (Diptera, Scenopinidae). ZooKeys 84: 39–57. doi: 10.3897/zookeys.84.774

Yeates DK (1994) Cladistics and classification of the Bombyliidae (Diptera: Asiloidea). Bulletin of the American Museum of Natural History 219: 1–191.

Yeates DK, Lambkin CL (1998) Cryptic species diversity and character congruence: Review of the tribe Anthracini (Diptera: Bombyliidae) in Australia. Invertebrate Taxonomy 12: 977–1078. doi: 10.1071/IT97019

Yeates DK, Logan D, Lambkin CL (1999) Life history of *Ligyra satyrus* (Diptera: Bombyliidae). Journal of the Australian Entomological Society. 38: 300–304. doi: 10.1046/j.1440-6055.1999.00127.x

Yeates DK, Lambkin CL (2006) Family Bombyliidae. On The Fly: The Interactive Atlas and Key to Australia Fly Families. Australian Biological Resources Study, Canberra & Centre for Biological Information Technology, St. Lucia, Brisbane.

Appendix 1

Morphological Characters

For a full description of these characters see Appendix 2 in Lambkin et al. (2003).

Morphological terminology follows that of McAlpine (1981), Yeates (1994), and Lambkin et al. (2003).

Head

Frontal

1. Head scales: (0) absent; (1) present
2. Face scale density: (0) absent; (1) sparse; (2) overlapping; (3) carpet
3. Frons scale density: (0) absent; (1) sparse; (2) overlapping; (3) carpet; (4) distinct dense medial patch
4. Frons with vertical groove: (0) absent; (1) depression; (2) groove
5. Frons with horizontal depression: (0) absent; (1) shallow; (2) distinct; (3) deep

Dorsal

6. W head/W thorax: (0) <; (1) =; (2) >
7. L antennae to compound eye/L scape: (0) <; (1) ≥; (2) ≥ 2×; (3) ≥ 4×
8. L antennal separation/L scape: (0) <; (1) ≥; (2) ≥ 2×; (3) ≥ 3×
9. Male compound eye separation /W OT: (0) meet; (1) <; (2) =; (3) ≤ 2×; (4) ≤ 3×; (5) > 3×
10. Female compound eye separation /W OT: (0) ≤ 2×; (1) ≤ 3×; (2) > 3×
11. OT to posterior margin of compound eye/ L OT: (0) ≤ OT; (1) > OT; (2) ≥ 2×; (3) ≥ 3×
12. L occiput/L OT: (0) ≤ 2×; (1) < 3×; (2) occiput long, well developed ≥ 3×
13. L vertex; i.e. L OT to occipital groove/ L OT: (0) ≤ L OT; (1) ≤ 2×; (2) wide > 2×
14. Depth occipital foveal depression: (0) no vertex; (1) not depressed; (2) shallow; (3) deep, slopes posteriorly at 45° to a short OG
15. W occipital foveal depression: (0) no vertex; (1) not depressed; (2) narrower than compound eye separation; (3) wider than compound eye separation
16. Apical occipital groove: (0) narrow; (1) wide rounded

Morphological Abbreviations			
A	apodeme (Fig. 6F)	Ma	macrochaetae
AAES	anterior arms of aedeagal sheath (Fig. 5E)	MB	membranous base
AC	acanthophorite (Fig. 6F)	mcu	section of wing vein CuA1 between vein
AE	aedeagus (Fig. 5F–H)		M3 and crossvein m-cu
AG	accessory glands (Fig. 9H)	m-m	basal section of wing vein M2
AN	anepisternum (mesopleuron)	MT	mediotergite (metanotum)
APA	anterior postalar ridge	OG	occipital groove
ASM	antennal apical stylomere	OT	ocellar triangle
BSM	basal stylomere	PA	postalar
BB	spermathecal basal bulb	PE	proepimeron (prosternum)
BP	basiphallus (Fig. 5E, F)	PL	plumula
BSM	antennal basal stylomere	PN	postpronotal lobe (humeral callus)
C	costa	PP	postpedicel
Cu	cubital vein	PR	prealar
dc	discal cell	R	radial vein
E	epandrium	1st r2+3	wing cell r2+3 basal to crossvein i-r1
EJA	ejaculatory apodeme (Fig. 5E, F)	2nd r2+3	wing cell r2+3 apical to crossvein i-r1
EP	epiphallus (Fig. 5E, F)	RM	ramus (Fig. 5F, H)
F	flagellomere	S	sternite (Fig. 6F, G)
G	gonocoxa	SES	subepandrial sclerites (Fig. 4E)
GS	gonostylus (Fig. 5B)	Sc	subcostal vein
H	hypandrium (Fig. 5C, D)	Scm	scutum
i-r1	wing inter-radial crossvein between veins	Scu	scutellum
	R2+3 and R4;	SP	sperm pump (Fig. 9H)
i-r2	wing inter-radial crossvein between veins	SR	spermathecal reservoir (Fig. 9H)
	R4 and R5.	ST	spermathecal
K	katepisternum (sternopleuron)	SS	scutoscutellar suture
L	length	T	tergite (Fig. 6F, G)
LAEA	lateral aedeagal apodemes (Fig. 5E, F)	TR	tympanal ridge
LT	laterotergite (metapleuron)	W	width
M	medial vein	WR	wing root

Lateral

17. Shape of head laterally: (0) round; (1) protruding but rounded, blunt; (2) conical
18. L proboscis: (0) < L oral cavity; (1) > L oral cavity; (2) > L head; (3) > 1.5× L head; (4) > 2× L head
19. L palps/ L proboscis: (0) rudimentary; (1) < 0.25×, short; (2) < 1, long
20. Genae setae surrounding oral cavity: (0) not grouped; (1) grouped laterally; (2) grouped apically, small tuft
21. L setae below antennae/L scape: (0) >; (1) ≤; (2) ≤ 0.5

22. Horizontal depression between antennae: (0) absent or shallow; (1) distinct; (2) deep
23. W of face projection/W compound eye including indentation posterior margin of eye: (0) < 1/4; (1) ≤ 1/2; (2) ≤ 1
24. W of the indentation on the posterior margin of the compound eye/L OT: (0) ≤ L OT; (1) > L OT; (2) ≥ 2× L OT
25. L of the line from the posterior margin of compound eye bisecting the compound eye facets/L OT: (0) absent; (1) <; (2) ≥; (3) ≥ 2×

Antennae

26. L scape /L pedicel: (0) ≤; (1) ≤ 3×; (2) > 3×
27. L PP/ L pedicel: (0) ≤ 3×; (1) ≤ 4×; (2) > 4×
28. PP base shape: (0) broad not round base; (1) onion-like, round base abruptly narrowed; (2) medially divided, laterally; (3) conical, broad base, gradually narrowed
29. L PP rod/L base: (0) long > 2×; (1) short < 2×; (2) no thin rod
30. Distinct joint between PP and BSM: (0) absent; (1) present
31. L BSM/L pedicel: (0) absent; (1) ≤; (2) ≤ 2×; (3) ≤ 3×; (4) > 3×
32. BSM apical hairs: (0) absent; (1) present
33. L ASM/W BSM: (0) < W BSM, minute spine; (1) > W BSM; (2) > 2×; (3) conical twisted hat

Thorax

34. Collar Ma: (0) pointed; (1) midstyle; (2) pectinate
35. Scm reflective vestiture: (0) bright; (1) very dull dark; (2) not reflective
36. Ma AN: (0) pointed; (1) midstyle; (2) pectinate
37. L PR bristles/ L PN: (0) > 2×; (1) >; (2) <; (3) absent
38. LT vestiture: (0) bare; (1) some hair; (2) dense hair
39. Ma LT: (0) absent; (1) pointed; (2) midstyle; (3) pectinate
40. MT vestiture: (0) bare; (1) some hair; (2) dense hair
41. L PA bristles/Scu: (0) absent; (1) < 0.5; (2) ≤; (3) >
42. L thorax/Scu: (0) ≤ 2×; (1) ≤ 3×; (2) > 3×
43. Scu vestiture reflective: (0) bright; (1) very dull dark; (2) absent

Legs

44. C_1 very long setae: (0) absent; (1) some; (2) dense
45. L forefemur/ L coxa: (0) ≤ 1.5×; (1) ≤ 2×; (2) ≤ 2.5×; (3) > 2.5×

46. Forefemoral spines: (0) absent; (1) short < W femur; (2) long
47. Forefemoral long hairs: (0) absent; (1) some; (2) dense
48. Foretibial spicules: (0) absent; (1) some; (2) dense
49. L foretarsus/ L foretibia: (0) ≥ 1; (1) ≥ 0.75; (2) > 0.5
50. Foretarsal microchaetae: (0) absent; (1) very few < 10; (2) present
51. Foretarsal microchaetae: (0) absent; (1) ends bulbous; (2) ends slightly bent; (3) ends distinctly bent
52. L foreclaw/ L midclaw: (0) < claw; (1) ≤ half; (2) ≤ third
53. Midfemoral spines: (0) absent; (1) short < W femur; (2) some long
54. Midfemoral long hairs: (0) absent; (1) some; (2) dense
55. Midtibial spicules: (0) absent; (1) some; (2) dense
56. L midpulvilli/ L claw: (0) ≥; (1) <; (2) ≤ half; (3) < 0.2
57. Midpulvilli: (0) large, flattened, membranous; (1) small rounded setose; (2) chisel-conical
58. Hindfemoral spines: (0) absent; (1) short < 0.5 W femur; (2) some long
59. Hindfemur long hairs: (0) absent; (1) some; (2) dense
60. Long hindtibial scales: (0) no long scales; (1) some long scales; (2) fluffy - protruding; (3) feathery; (4) very long, dense, feathered fringes
61. Hindtibial spicules: (0) absent; (1) some spicules; (2) apical patch; (3) dense spicules
62. Hindtibial spicules L: (0) absent; (1) same; (2) inner row longer
63. L hindpulvilli/L claw: (0) ≥; (1) <; (2) ≤ half; (3) ≤ 0.2
64. Hindpulvilli: (0) large, flattened, membranous; (1) small rounded setose; (2) chisel -conical

Wing

65. Patagium: (0) absent hairs only; (1) present scales
66. Basicosta: (0) absent; (1) blunt; (2) sharp

Wing venation

67. Crossvein forming an extra anterior apical submarginal cell: (0) absent; (1) extra apical submarginal cell
68. R_{2+3} join R_{4+5}: (0) basal to r-m > L r-m; (1) at r-m
69. R_{2+3} rises from R_{4+5}: (0) acutely; (1) at right angles
70. 2nd r-m crossvein: (0) absent; (1) present
71. Spurvein base R_{2+3}: (0) absent; (1) bump or present
72. R_{2+3} apical loop: (0) long apical loop; (1) loop > 180°; (2) at least 90° bend; (3) absent
73. i-r_1, R_{2+3} to R_4: (0) absent; (1) present

74. i-r_1 crossvein: (0) absent; (1) straight; (2) slightly sinuous; (3) distinctly sinuous
75. Spurvein i-r: (0) absent; (1) bump or present
76. L i-r_1/L r-m: (0) absent; (1) ≤; (2) ≤ 2×; (3) < 3×; (4) ≥ 3×
77. Spur-vein base R_{4+5}: (0) absent; (1) bump or present
78. Spur-vein R_4: (0) absent; (1) bump or present
79. R_5/R_1 meet wing: (0) R_5 distal to R_1; (1) equal; (2) R_5 basal to R_1
80. i-r_2, R_4 to R_5: (0) absent; (1) present
81. M_1: (0) straight; (1) slightly sinuous; (2) sinuous
82. L m-m/r-m: (0) ≤ 2×; (1) ≤ 3×; (2) > 3×
83. m-m: (0) straight; (1) slightly sinuous; (2) sinuous
84. m-m spurvein: (0) absent; (1) into m_2; (2) into discal; (3) crossvein form basal cell
85. m-m to hind wing margin: (0) oblique; (1) parallel; (2) horizontal
86. M_2: (0) straight; (1) slightly sinuous; (2) sinuous
87. r_5 open: (0) open; (1) narrow < r-m; (2) just closed; (3) closed and stalked - acute; (4) closed and stalked -obtuse
88. m_1 open: (0) open; (1) closed and stalked
89. m_2 open: (0) open; (1) closed and stalked - acute
90. L mcu/r-m: (0) > 3×; (1) < 3×; (2) ≤ 2×; (3) ≤
91. m-cu: (0) straight; (1) slightly sinuous; (2) 90° basally; (3) sinuous
92. Spur-vein m-cu: (0) absent; (1) spur into discal cell
93. Spur-vein into M_2: (0) absent; (1) spur-vein into m_2; (2) cross-vein to CuA_1
94. W anal/ W posterior cubital: (0) ≤; (1) ≤ 1.5×; (2) ≤ 2×; (3) > 2×
95. cup open: (0) open; (1) narrow < r-m; (2) closed at wing margin (Fig. 9B); (3) closed and stalked - acute
96. Anal lobe margin: (0) hairs; (1) some scales; (2) scales dense
97. Anal cell: (0) broad rounded; (1) rounded; (2) thin linear; (3) very reduced
98. Alula reduced: (0) not reduced; (1) reduced, L < 4× W
99. Alula margin: (0) hairs; (1) some scales; (2) scales dense
100. Squamal margin: (0) hairs; (1) some scales; (2) scales dense
101. Squama reduced: (0) not reduced; (1) reduced, L < 4× W
102. wing L: (0) ≤ 10; (1) 10-15; (2) 16-20; (3) 21-25; (4) > 25
103. wing L/W: (0) ≤ 3×; (1) long > 3×
104. L wing/ L abdomen: (0) ≥ 3×; (1) ≥ 2 ×; (2) > 1.5 ×; (3) >

Abdomen

105. Abdomen apically: (0) rounded; (1) narrowed; (2) truncate, parallel sided
106. Abdomen L/ W T_2: (0) > 2.5×; (1) < 2.5×; (2) < 2×; (3) ≤ 1.5×; (4) ≤
107. Abdominal vestiture reflective: (0) bright; (1) very dull dark; (2) not reflective

108. Abdominal bristles/hair: (0) dorsally and laterally; (1) lateral and apically T_7; (2) apically; (3) absent
109. Long lateral hairs > T_1: (0) dense tufts; (1) some; (2) absent
110. Ma T_1: (0) pointed; (1) midstyle; (2) pectinate
111. Scales: (0) absent; (1) adpressed scales; (2) upstanding scales; (3) long upstanding scale tufts

Male genitalia

112. Male genitalia twisted: (0) no twisting, gonocoxae ventral; (1) 90°; (2) 180°, gonocoxae dorsal
113. E setae grouping: (0) not grouped; (1) medioapically; (2) lateroapically; (3) laterally
114. E setae group: (0) not grouped; (1) loose; (2) strong; (3) dense tufts
115. Epandrial spines: (0) absent; (1) short and broad; (2) long
116. E apically: (0) deeply indented; (1) concave-indented; (2) truncate; (3) convex rounded; (4) convex pointed
117. E apical flange: (0) absent; (1) < quarter base; (2) < third base; (3) < half base; (4) > half rest base
118. E medial flange: (0) absent; (1) slight; (2) distinct < quarter base
119. L E basal flange: (0) absent; (1) < quarter base; (2) < third base (Fig. 10B); (3) < half base; (4) ≥ half rest base; (5) > base
120. Mid W/L basal flange: (0) absent; (1) < quarter length base; (2) < half length; (3) > half length (4) > length
121. E posterolateral flange: (0) absent; (1) < quarter length base
122. E basal flange recurved: (0) absent (Fig. 10B); (1) < quarter base; (2) < third base; (3) < half base; (4) > half rest base; (5) > base
123. E basally extended: (0) absent; (1) present
124. SES: (0) absent; (1) linear; (2) triangular; (3) quadrate; (4) single (Fig. 10A)
125. L SES/G W: (0) absent; (1) < eighth; (2) < quarter; (3) > quarter
126. G setae: (0) some; (1) dense; (2) tufts
127. G setae group: (0) absent; (1) not grouped; (2) apically; (3) medially; (4) laterally; (5) basally
128. Thick G setae number: (0) absent; (1) no thick setae; (2) some; (3) many; (4) 6-8 long
129. Thick G setae position: (0) absent; (1) no thick setae; (2) apically; (3) medially; (4) laterally; (5) basally
130. G subapical indentation: (0) absent; (1) slight < third; (2) narrowed apically > third
131. W G medial indentation: (0) absent; (1) < third; (2) > third; (3) > half
132. G medial weakness: (0) absent; (1) desclerotised line; (2) lines of weakness; (3) division medially

133. G ventral division: (0) line fusion basally; (1) line fusion entire; (2) fused medially; (3) fused basally; (4) fused
134. G medioventrally: (0) deeply indented; (1) indented medially; (2) flat; (3) convex shell
135. G ventral ridge: (0) absent; (1) slight; (2) distinct
136. G basal projection of the ventral ridge: (0) absent; (1) slight; (2) distinct; (3) recurved hook
137. G basomedial margin: (0) deeply indented; (1) indented, concave; (2) smooth, linear; (3) convex
138. H: (0) absent; (1) present
139. H laterally: (0) absent; (1) indented between G, smooth; (2) projecting; (3) with spur; (4) with finger (Fig. 10F)
140. RM: (0) small; (1) > L GS; (2) large recurved
141. L G A: (0) absent; (1) < GS; (2) > GS
142. G plates dorsoapically: (0) medially parallel; (1) angled basomedial plates diverge dorsally
143. G dorsoapical plates extension apically: (0) not extended apically; (1) small apical extension; (2) long apical extension beyond the base of the gonostyli
144. GS: (0) large base, long pointed flange; (1) laterally bifid; (2) simple curved hook; (3) medial hook; (4) lateral hook
145. GS basal projection: (0) absent; (1) small; (2) < GS
146. L AE: (0) < GS; (1) = GS; (2) > GS
147. AE EP separate: (0) absent; (1) present
148. EP: (0) absent; (1) present
149. EP deep ventral notch: (0) no EP; (1) absent; (2) medial
150. EP expanded apically: (0) no EP; (1) not expanded (Fig. 10H); (2) ≤ 2× neck; (3) < 3× neck; (4) > 3 × neck
151. EP apical plate: (0) no EP; (1) absent; (2) apical plate
152. EP medioventral projection: (0) no EP; (1) absent; (2) above AE
153. EP lateroapical lobes: (0) no E; (1) absent; (2) rounded dorsally; (3) pointed dorsally
154. EP lateral projection laterally: (0) no EP; (1) absent (2) lateral
155. EP medial projection laterally: (0) no EP; (1) absent; (2) medial
156. EP pair ventral projections: (0) no EP; (1) absent; (2) medial below AE
157. L EP: (0) no EP; (1) < GS base; (2) > GS base; (3) > G margin
158. EP recurved apicomedial projection: (0) no EP; (1) absent; (2) present dorsally
159. EP recurved apically: (0) no E; (1) absent; (2) apex recurved lateral view
160. EP apical setae: (0) no EP; (1) absent; (2) present
161. BP expanded: (0) not expanded; (1) round; (2) swollen spherical; (3) bilobed
162. LAEA: (0) spoon convex up; (1) spoon concave; (2) linear; (3) absent
163. L lateral AE A: (0) < L GS; (1) < G margin; (2) = G margin; (3) absent
164. AAES: (0) spoon convex; (1) spoon concave; (2) narrow wedge; (3) linear
165. AAES: (0) < B; (1) = L GS; (2) < G margin; (3) to G margin

166. EJA: (0) racquet - round; (1) linear
167. L EJA: (0) within G; (1) = G; (2) > G < L GS; (3) > G > L GS

Female genitalia

168. AC spines: (0) 3 prs; (1) 4 prs; (2) 5 prs; (3) > 5 prs; (4) > 10 prs
169. AC spines apically: (0) thin tapering; (1) broader spoon shaped
170. T_{9+10} sclerites: (0) 1 sclerite; (1) 3 sclerites
171. T_9 dorsal medioapical unsclerotised lacuna: (0) absent; (1) present
172. T_8 dorsal medioapical unsclerotised lacuna: (0) absent; (1) present
173. T_8 hair: (0) apical half; (1) apical half bare medially; (2) apical edge
174. T_8 laterally: (0) not indented; (1) indented arms
175. T_8 A divided medially: (0) not divided; (1) slightly; (2) distinctly ≥ half width
176. T_8 A lateral projections: (0) absent; (1) slight; (2) not linear; (3) linear
177. T_8 A L/medial W: (0) margin thickened, sclerotised; (1) ≤ quarter; (2) ≤ half; (3) <; (4) ≥; (5) > 2×; (6) ≥ 3×
178. T_8 A internal structure: (0) absent; (1) linear; (2) quadrate plate
179. S_8 sclerites: (0) 1 linear 2 round; (1) 1 linear 2 round 1 medial; (2) U-shaped 2 triangular; (3) 2 round; (4) sheet
180. Furca: (0) U-shaped; (1) 3 separate sclerites; (2) 2 rods
181. ST tube: (0) short < third L pump; (1) present; (2) long > × 8 L pump
182. Basal endplate: (0) absent; (1) small; (2) present (Fig. 9H); (3) large
183. Basal endplate: (0) absent; (1) simple-thin processes; (2) thick processes; (3) funnel
184. Sperm pump: (0) short pump; (1) very long pump
185. Long pump papillae: (0) no long papillae; (1) unpigmented; (2) pigmented
186. Pump processes basally: (0) no processes; (1) short processes; (2) long papillae
187. Pump processes medially: (0) no processes; (1) short processes; (2) long papillae
188. Pump processes apically: (0) no processes; (1) short processes
189. Apical endplate: (0) large; (1) present; (2) small; (3) absent
190. Apical endplate: (0) absent; (1) simple- thin processes; (2) thick processes; (3) funnel; (4) double
191. SR: (0) L ≤ W; (1) L > W; (2) L > 2 W; (3) L > 4 W; (4) L > 6 W; (5) L > 8 W; (6) L > 30 W
192. ST basal sclerotised plate: (0) absent; (1) basal sclerotised plate
193. Tube ST to pump: (0) absent; (1) present; (2) long
194. SR shape: (0) round square; (1) oval; (2) pear, expanded apically; (3) long; (4) expanded basally
195. SR apically: (0) rounded blunt; (1) nipple; (2) narrowed; (3) knob
196. ST round basal bulb: (0) no round BB; (1) round BB
197. SR pigmented: (0) unpigmented; (1) pigmented; (2) basally unpigmented
198. ST long medial tube: (0) no medial tube; (1) tube medially
199. Long membranous base: (0) no long base; (1) long MB

200. ST long MB basally swollen: (0) no long base; (1) symmetrically; (2) asymmetrically
201. ST clear rings: (0) no rings; (1) clear ring; (2) long striated collar
202. SR to pump: (0) symmetrical; (1) asymmetrical
203. SR medially bent: (0) absent; (1) bent reservoir
204. SR apically bent: (0) absent; (1) tip only
205. Tubules: (0) absent; (1) present
206. SR walls: (0) thin unsclerotised; (1) thick sclerotised
207. SR walls: (0) no dimples; (1) with dimples thin unsclerotised

Appendix 2. Matrix of 207 characters for the Australian *Balaana* genus-group.

Taxa/Character	1									10									20								
Lig. satyrus	1	0	2	0	3	0	0	1	4	1	2	2	1	3	3	1	2	1	2	2	1	0	1	1	3	1	1	3
Lig. sinuatifascia	1	1	1	0	2	0	0	1	3	1	2	1	1	3	2	1	2	1	2	2	1	0	1	0	2	2	1	3
Bal. abscondita	1	1	2	0	2	0	0	2	4	1	2	2	2	3	3	1	2	1	2	2	1	0	2	1	3	1	1	3
Bal. bicuspis	1	1	1	0	2	0	0	1	4	2	2	1	2	3	3	1	2	1	2	2	1	0	1	1	3	2	2	3
Bal. centrosa	1	1	2	0	3	0	0	1	4	1	3	2	2	3	3	1	2	1	2	2	1	0	1	1	3	2	2	3
Bal. gigantea	1	1	2	0	2	0	1	2	4	1	3	2	2	3	3	1	2	1	2	2	1	0	1	1	3	1	2	3
Bal. kingcascadensis	1	1	3	0	2	0	0	2	3	?	2	1	2	3	3	1	2	1	2	2	1	0	1	1	3	2	2	3
Bal. latelimbata	1	1	2	0	1	0	0	2	3	0	2	1	1	3	3	1	2	1	2	2	1	0	1	1	3	1	1	3
K. adelaidica	1	1	2	0	3	0	0	1	4	1	2	1	2	3	3	1	2	1	2	2	1	0	1	0	2	1	0	3
K. corusca	1	1	2	0	3	0	0	2	3	1	2	1	2	3	3	1	2	1	2	2	1	0	1	0	3	2	1	3
K. irwini	1	1	2	0	3	2	0	2	3	1	2	1	2	3	3	1	2	1	2	2	0	0	1	1	2	1	1	3
K. westralica	1	1	2	0	3	0	1	2	4	1	2	1	1	3	3	1	2	0	2	2	1	0	1	0	2	1	1	3
Lar. bushblitzi	1	1	2	0	1	0	0	2	3	?	2	2	2	3	3	1	2	1	2	2	1	0	2	0	3	1	1	3
Lar. collessi	1	1	2	0	3	0	0	1	3	?	3	2	2	3	3	1	2	1	2	2	1	0	1	1	3	1	2	3
Lar. dimidiatipennis	1	1	2	0	1	0	0	1	3	1	1	0	1	3	3	1	2	1	2	2	2	0	2	0	3	1	1	3
Lar. zwicki	1	1	2	0	1	0	0	2	3	?	2	2	2	3	3	1	2	1	2	2	0	0	1	0	3	1	2	3
Mun. erugata	1	1	2	0	2	0	0	1	3	1	2	1	1	3	3	1	2	1	2	2	1	0	1	0	2	2	2	3
Mun. lepidokingi	1	1	1	0	3	0	0	1	4	1	3	2	2	3	3	1	2	2	2	2	1	0	1	1	3	1	0	3
Mun. paralutea	1	2	2	0	2	0	0	1	4	1	3	2	2	3	2	1	2	1	2	2	1	0	1	1	2	1	2	3
Muw. stellifera	1	1	2	0	1	0	1	2	3	0	2	1	2	3	2	1	2	1	2	2	1	0	1	1	2	1	0	3
Muw. vitreilinearis	1	1	2	0	2	1	0	1	3	0	2	1	1	3	2	1	2	1	2	2	1	0	1	1	2	1	0	3
Nga. trigonium	1	1	2	0	3	0	1	2	3	0	2	1	1	3	3	1	2	1	2	2	1	0	1	1	2	2	1	3
P. anaxios	1	1	1	0	2	2	0	2	4	?	2	1	1	3	3	1	2	1	2	2	1	0	1	1	3	1	1	3
P. basilikos	1	1	2	0	3	0	0	3	3	0	2	1	2	3	3	1	2	0	2	2	1	0	1	0	3	1	2	3
P. blackdownensis	1	1	1	0	3	0	0	1	3	?	2	0	1	3	3	1	2	1	2	2	1	0	1	1	3	1	1	3
P. bouchardi	1	2	1	0	3	0	0	1	3	0	3	1	2	3	3	1	2	1	2	2	1	0	1	0	3	1	1	3
P. cyanea	1	1	1	0	2	0	0	2	3	1	2	2	1	3	3	1	2	1	2	2	1	0	1	1	3	1	2	3
P. culgoafloodplainensis	1	1	2	0	3	0	0	2	4	?	3	2	2	3	3	1	2	0	2	2	1	0	2	1	3	2	0	3
P danielsi	1	1	1	0	2	0	0	2	3	?	2	1	2	3	3	1	2	1	2	2	1	0	1	1	3	2	1	3
P decora	1	1	1	0	3	1	0	2	3	1	2	1	1	3	3	1	2	1	2	2	2	0	1	0	3	1	1	3
P mackensiei	1	1	2	0	2	2	0	2	?	1	3	2	2	3	3	1	2	1	2	2	2	0	2	2	3	2	2	3
P marginicollis	1	1	2	0	2	0	0	2	4	1	2	1	1	3	3	1	2	1	2	2	1	0	1	0	3	1	0	3
P viridula	1	0	1	0	3	0	0	2	3	1	2	1	1	3	3	1	2	0	2	2	1	0	1	0	3	1	1	3
P whyalla	1	1	2	0	3	0	0	1	3	1	2	2	2	3	3	1	2	0	2	2	1	0	1	1	3	1	1	3
Wur. emu	1	1	2	0	3	0	0	1	4	1	1	0	1	3	2	1	2	2	2	2	1	0	2	0	2	1	2	3
Wur. impatientis	1	3	3	0	2	0	0	2	4	1	2	1	2	3	3	1	2	1	2	1	0	0	1	1	1	1	0	3
Wur. montebelloensis	1	1	3	0	3	0	0	2	3	0	2	1	1	3	2	1	2	1	2	2	1	0	1	1	2	1	0	3
Wur. norrisi	1	1	2	0	2	0	0	2	3	0	2	2	1	3	3	1	2	1	2	2	0	0	1	0	2	1	1	3
Wur. patrellia	1	1	3	0	3	0	0	2	4	?	2	0	1	3	2	1	2	1	2	2	1	0	1	2	1	1	1	3
Wur. skevingtoni	1	3	3	0	2	0	0	1	3	?	2	1	2	3	3	1	2	1	2	1	1	0	1	0	1	1	0	3
Wur. windorah	1	2	2	0	2	0	0	1	3	1	2	0	1	3	2	1	2	2	2	2	1	0	1	0	3	1	2	3
Wur. wyperfeldensis	1	1	2	0	3	?	0	1	3	1	0	1	2	3	2	1	2	1	2	2	1	0	1	0	2	2	1	3

Taxa/Character	30										40										50							
Lig. satyrus	1	4	0	0	1	2	2	0	2	2	2	2	1	2	2	0	0	2	2	2	2	1	0	2	2	1	3	2
Lig. sinuatifascia	1	4	0	0	1	2	2	0	2	2	2	3	1	2	2	0	0	2	2	2	2	1	0	2	2	0	2	2
Bal. abscondita	1	2	0	0	1	2	2	1	2	2	2	2	1	2	2	1	0	2	2	1	2	2	1	2	2	1	1	2
Bal. bicuspis	1	3	0	0	1	2	2	0	2	2	2	2	1	2	2	0	0	2	2	1	2	1	1	2	1	1	2	2
Bal. centrosa	1	3	0	1	1	2	2	1	2	2	2	2	1	2	2	0	0	2	2	2	2	1	2	2	2	0	2	2
Bal. gigantea	1	3	0	0	1	2	2	1	2	2	2	2	1	2	2	0	1	2	2	2	2	1	1	1	2	0	1	2
Bal. kingcascadensis	1	2	0	0	1	2	2	1	2	2	2	2	1	2	2	1	0	2	2	2	2	1	1	2	2	0	2	2
Bal. latelimbata	1	2	0	0	1	2	2	1	2	2	2	2	1	2	2	0	0	2	2	2	2	1	1	1	2	0	2	2
K. adelaidica	1	2	0	0	1	2	2	1	2	2	2	2	1	2	2	1	0	2	2	2	2	1	0	2	2	0	1	2
K. corusca	1	3	0	0	1	2	2	1	2	2	2	2	1	2	2	1	0	2	2	1	2	1	0	2	2	0	2	2
K. irwini	1	3	0	0	1	2	2	1	2	2	2	2	1	2	2	2	0	2	2	1	2	1	1	2	2	0	2	2
K. westralica	1	2	0	0	1	2	2	0	2	2	2	3	1	2	2	1	0	2	2	1	2	1	0	2	2	1	2	2
Lar. bushblitzi	1	3	0	0	1	2	2	1	2	2	2	2	1	2	2	1	0	1	2	1	2	1	1	2	2	1	2	2
Lar. collessi	1	3	0	0	1	2	2	1	2	2	2	2	1	2	2	0	0	2	2	1	2	1	1	2	2	0	2	2
Lar. dimidiatipennis	?	?	?	?	1	2	2	1	2	2	2	3	1	2	2	0	1	2	2	2	2	1	0	2	2	0	1	2
Lar. zwicki	1	3	0	0	1	2	2	1	2	2	2	2	1	2	2	1	0	2	2	1	2	1	0	2	2	0	1	2
Mun. erugata	1	3	0	0	1	2	2	1	2	2	2	2	1	2	2	1	0	2	2	2	2	1	1	1	2	0	2	2
Mun. lepidokingi	1	3	0	0	1	2	2	1	2	2	2	2	1	2	2	1	1	2	2	1	2	1	0	2	0	1	2	2
Mun. paralutea	1	2	0	0	1	2	2	1	2	2	2	2	1	2	2	0	1	2	2	1	2	1	1	1	2	0	2	2
Muw. stellifera	1	3	0	0	1	2	2	1	2	2	2	2	1	2	2	0	0	2	2	2	2	1	1	1	2	0	2	2
Muw. vitreilinearis	1	3	0	1	1	2	2	1	2	2	2	2	1	2	2	1	0	1	2	2	2	2	1	1	2	1	2	2
Nga. trigonium	1	4	0	0	1	2	2	1	2	2	2	2	1	2	2	0	0	2	2	2	2	1	0	1	2	1	2	2
P. anaxios	1	3	0	1	1	0	2	1	2	2	2	2	1	0	2	0	0	2	2	2	2	1	1	1	2	0	2	2
P. basilikos	1	4	0	0	1	1	2	1	2	2	2	2	1	1	2	1	0	2	2	1	2	1	2	1	2	0	2	2
P. blackdownensis	1	4	0	0	1	0	2	1	2	2	2	2	1	0	2	0	0	1	2	1	2	1	1	1	2	0	2	2
P. bouchardi	1	3	0	0	1	1	2	1	2	2	2	2	1	0	2	1	0	2	2	1	2	1	1	1	2	0	2	2
P. cyanea	1	3	0	1	1	1	2	1	2	2	2	2	1	2	2	0	0	2	2	1	2	1	1	1	2	0	2	2
P. culgoafloodplainensis	1	4	0	0	1	0	?	?	2	2	2	?	1	0	2	1	0	1	2	1	2	1	1	1	1	1	2	2
P danielsi	1	3	0	0	1	0	2	1	2	2	2	2	1	0	2	0	0	2	2	2	2	1	1	1	2	0	2	2
P decora	1	3	0	0	1	0	2	0	2	2	2	2	1	0	2	0	0	2	2	1	2	1	1	1	2	0	2	2
P mackensiei	1	2	0	0	1	1	2	1	1	2	2	2	0	1	2	0	0	0	2	1	2	1	1	1	0	1	2	2
P marginicollis	1	2	0	0	1	0	2	1	2	2	2	2	1	0	2	0	0	1	2	1	2	1	1	1	2	0	2	2
P viridula	1	3	0	0	1	0	2	1	2	2	2	2	1	0	2	0	0	1	2	1	2	1	1	1	2	0	2	2
P whyalla	1	3	0	0	1	0	2	1	2	2	2	2	1	0	2	0	0	2	2	1	2	1	1	1	2	0	2	2
Wur. emu	1	1	0	0	1	2	2	1	2	2	2	2	1	2	2	0	0	1	2	1	2	1	2	1	2	0	2	2
Wur. impatientis	1	2	0	0	1	2	2	1	2	2	2	2	1	2	2	1	0	2	2	1	2	1	0	2	2	0	2	2
Wur. montebelloensis	1	4	0	0	1	2	2	1	2	2	2	2	1	2	2	1	0	2	2	2	2	1	0	2	2	0	2	2
Wur. norrisi	1	4	0	0	1	2	2	1	2	2	2	2	1	2	2	0	0	1	2	1	2	1	0	2	2	0	2	2
Wur. patrellia	1	2	0	0	1	2	2	1	2	2	2	2	1	2	2	1	0	2	2	1	2	1	2	1	0	0	1	2
Wur. skevingtoni	1	3	0	0	1	2	2	0	2	2	2	3	1	2	2	0	0	0	2	1	2	1	1	2	2	0	2	2
Wur. windorah	1	2	0	0	1	2	2	1	2	2	2	2	1	2	2	0	0	2	2	1	2	3	2	1	2	0	2	2
Wur. wyperfeldensis	1	3	0	0	1	2	2	1	2	2	2	2	1	2	2	0	0	2	2	1	2	1	2	2	0	0	1	2

Taxa/Character	**60**										**70**										**80**							
Lig. satyrus	1	2	2	3	2	1	2	0	1	1	0	0	2	1	1	0	2	0	0	2	1	1	2	2	0	1	2	1
Lig. sinuatifascia	2	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	2	0	0	2	1	1	1	1	0	1	2	1
Bal. abscondita	2	2	2	1	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	1	2	1	0	1	2	1
Bal. bicuspis	1	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	4	0	0	2	0	2	2	2	0	1	2	1
Bal. centrosa	2	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	1	2	1	0	1	2	1
Bal. gigantea	1	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	4	0	0	2	0	2	2	1	0	1	2	0
Bal. kingcascadensis	1	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	1	2	1	0	0	2	1
Bal. latelimbata	2	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	1	2	1	0	1	2	1
K. adelaidica	0	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	1	2	1	0	1	2	0
K. corusca	1	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	2	2	2	0	1	2	0
K. irwini	0	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	2	2	2	0	1	2	0
K. westralica	1	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	4	0	0	2	0	1	2	1	0	1	2	0
Lar. bushblitzi	1	2	2	1	2	1	2	0	1	1	0	0	2	1	2	0	4	0	0&1	2	0	2	2	1	0	1	2	1
Lar. collessi	2	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	2	1	1	0	1	2	1
Lar. dimidiatipennis	1	2	2	?	?	1	2	0	1	1	0	0	2	1	2	0	4	0	0	2	0	1	1	2	0	0	1	1
Lar. zwicki	0	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	1	1	1	0	0	2	1
Mun. erugata	1	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	4	0	0	2	0	1	2	1	0	0	2	1
Mun. lepidokingi	2	2	2	2	2	1	2	0	1	1	0	0	2	1	1	0	3	0	0	2	0	1	2	1	0	1	2	1
Mun. paralutea	1	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	2	1	0	0	1	2	1
Muw. stellifera	1	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	4	0	0	2	0	2	2	2	0	0	2	1
Muw. vitreilinearis	1	2	2	2	2	1	2	0	1	1	0	0	2	1	3	0	4	0	0	2	0	1	2	2	0	1	2	1
Nga. trigonium	1	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	1	2	1	0	0	2	1
P. anaxios	1	3	2	2	2	1	2	0	1	1	0	0	2	1	3	0	3	0	0	2	0	1	0	0	0	0	1	1
P. basilikos	1	3	2	2	2	1	2	0	1	1	0	0	2	1	3	0	4	0	0&1	2	0	0	1	0	0	0	2	1
P. blackdownensis	1	3	2	2	2	1	2	0	1	1	0	0	2	1	3	0	4	0	0	2	0	1	0	1	0	0	1	1
P. bouchardi	1	3	2	2	2	1	2	0	1	1	0	0	2	1	3	1	3	0	0	2	0	0	1	0	0	0	2	1
P. cyanea	1	3	2	2	2	1	2	0	1	1	0	0	2	1	3	0	4	0	0	2	0	1	0	0	0	0	1	1
P. culgoafloodplainensis	1	3	2	2	2	1	2	0	1	1	0	0	2	1	3	0	4	0	0	2	0	0	1	0	0	0	1	1
P danielsi	1	3	2	2	2	1	2	0	1	1	0	0	2	1	3	0	4	0	0	2	0	0	1	0	0	0	2	1
P decora	1	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	1	0	1	0	0	1	1
P mackensiei	1	3	2	2	2	1	2	0	1	1	0	0	2	1	2	0	4	0	1	2	0	1	1	1	1	0	1	1
P marginicollis	2	3	2	2	2	1	2	0	1	1	0	0	2	1	3	0	3	0	0	2	0	0	1	1	0	0	1	1
P viridula	1	3	2	2	2	1	2	0	1	1	0	0	2	1	3	0	4	0	0	2	0	0	1	0	0	0	2	1
P whyalla	1	3	2	2	2	1	2	0	1	1	0	0	2	1	3	1	4	0	0	2	0	0	1	0	0	0	2	1
Wur. emu	2	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	4	0	0	2	0	1	2	1	0	1	2	0
Wur. impatientis	2	2	2	2	2	1	2	0	1	1	0	0	2	1	3	0	4	0	0	2	0	1	2	1	0	1	2	0
Wur. montebelloensis	2	2	2	2	2	1	2	0	1	1	0	0	2	1	3	0	4	0	0	2	0	2	2	2	0	1	2	0
Wur. norrisi	2	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	2	0	0	2	0	1	2	1	0	1	2	1
Wur. patrellia	2	2	2	2	2	1	2	0	1	1	0	0	2	1	3	0	4	0	0	2	0	2	2	2	0	1	2	0
Wur. skevingtoni	2	2	2	2	2	1	2	0	1	1	0	0	2	1	2	0	2	0	0	2	0	1	2	1	0	1	2	1
Wur. windorah	2	2	2	1	2	1	2	0	1	1	0	0	2	1	2	0	4	0	0	2	0	2	2	1	0	1	2	1
Wur. wyperfeldensis	2	2	2	1	2	1	2	0	1	1	0	0	2	1	2	0	3	0	0	2	0	1	2	1	0	1	2	1

Taxa/Character	90										100										110							
Lig. satyrus	2	1	0	0	2	1	2	1	0	2	2	0	2	1	1	1	3	2	1	0	1	1	1	2	3	2	0	0
Lig. sinuatifascia	3	1	0	0	3	1	2	0	0	2	2	0	1	1	1	1	3	2	1	0	1	1	1	2	3	2	2	2
Bal. abscondita	3	1	0	0	1	1	2	0	0	2	2	0	1	0	1	0	3	2	1	0	1	1	1	2	2	0	1	0
Bal. bicuspis	3	1	0	0	2	1	2	0	0	2	2	0	1	0	1	0	3	2	1	0	1	1	1	2	1	0	1	0
Bal. centrosa	3	2	0	0	1	1	2	0	1	2	2	0	2	1	1	0	3	2	1	0	1	1	1	2	3	2	2	0
Bal. gigantea	3	3	0	0	2	1	2	0	1	2	2	0	3	1	1	0	3	2	1	0	1	1	1	2	1	0	2	0
Bal. kingcascadensis	3	1	0	0	1	1	2	0	0	2	2	0	1	1	1	0	3	2	1	0	1	1	1	2	2	0	2	0
Bal. latelimbata	3	3	0	0	1	1	2	0	0	2	2	0	1	0	1	0	3	2	1	0	1	1	1	2	1	0	2	0
K. adelaidica	3	1	0	0	1	1	2	0	1	2	2	0	1	0	1	0	3	2	1	0	1	1	1	2	2	0	1	1
K. corusca	3	1	0	0	1	2	2	0	1	2	2	0	1	0	1	0	3	2	1	0	1	1	1	2	2	0	2	1
K. irwini	3	1	0	0	1	1	2	0	1	2	2	0	1	0	1	0	3	2	1	0	1	1	1	2	2	0	2	1
K. westralica	3	2	0	0	1	1	2	0	1	2	2	0	1	0	1	0	3	2	1	0	1	1	1	2	2	0	2	1
Lar. bushblitzi	3	2	1	0&1	1	1	2	0	1	2	2	0	1	0	1	1	3	2	0	0	1	2	1	2	1	0	1	1
Lar. collessi	3	2	0	0	1	1	2	0	0	2	2	0	2	1	1	1	3	2	1	0	1	2	1	2	2	0	2	1
Lar. dimidiatipennis	3	2	0	0	1	1	2	0	0	2	2	0	1	0	1	0	4	2	1	0	1	1	1	2	3	1	2	2
Lar. zwicki	3	2	1	0	1	1	2	1	1	2	2	0	2	0	1	1	3	2	1	0	1	2	1	2	1	0	1	0
Mun. erugata	3	1	0	0	1	2	2	0	0	2	2	0	2	1	1	0	3	2	1	0	1	1	1	2	2	0	2	0
Mun. lepidokingi	3	1	0	0	1	1	2	1	0	2	2	0	1	1	1	0	3	2	1	0	1	1	1	2	1	0	2	0
Mun. paralutea	3	1	0	0	1	1	2	0	0	2	2	0	2	0	1	1	3	2	1	0	1	1	1	2	1	0	2	0
Muw. stellifera	3	2	0	0	1	1	2	0	0	2	2	0	2	0	1	0	3	2	1	0	1	2	1	2	1	2	2	0
Muw. vitreilinearis	3	2	1	0	1	2	2	0	0	2	2	0	1	0	1	0	3	2	1	1	1	1	1	2	1	0	1	1
Nga. trigonium	3	2	0	0	1	2	2	0	0	2	2	0	2	0	1	1	3	2	1	0	1	1	1	0	0	0	1	0
P. anaxios	3	3	0	0	1	1	2	1	0	2	2	0	2	0	1	1	3	0	1	0	1	1	1	2	2	2	2	1
P. basilikos	3	3	0	0	1	1&2&3	2	0	0	2	2	0	4	1	1	1	3	0	1	0	1	1	1	2	2	2	3	1
P. blackdownensis	3	3	0	0	1	1	2	1	1	2	2	0	2	0	1	1	3	0	1	0	1	1	1	2	2	2	3	0
P. bouchardi	3	3	0	0	1	1	2	1	0	2	2	0	1	1	0	1	3	1	1	0	1	1	1	2	2	2	2	1
P. cyanea	3	3	0	0	2	1	2	1	0	2	2	0	2	1	1	1	3	1	1	0	1	1	1	2	1	0	2	0
P. culgoafloodplainensis	3	2	0	0	2	1&2	2	0	0	2	2	0	3	0	1	1	4	0	1	0	1	1	1	2	2	2	2	0
P danielsi	3	3	0	0	1	1	2	0	0	2	2	0	3	1	1	1	3	0	1	0	1	1	1	2	2	2	2	1
P decora	3	3	0	1	2	2	2	0	0	2	2	0	2	0	1	1	3	0	1	0	1	1	1	2	1	0	3	1
P mackensiei	3	3	0	0	1	1	2	0	1	2	2	0	0	1	1	1	4	1	1	0	1	1	?	?	?	?	?	?
P marginicollis	3	3	0	0	1	1	2	1	1	2	2	0	1	1	0	1	3	0	1	0	1	1	1	2	2	2	3	0
P viridula	3	3	0	0	2	2	2	0	0	2	2	0	3	1	1	1	3	0	1	0	1	1	1	2	1	0	3	0
P whyalla	3	3	0	0	2	1	2	0	0	2	2	0	3	1	1	1	2	0	1	0	1	1	1	2	2	2	3	0
Wur. emu	3	1	0	0	1	1	2	0	0	2	2	0	1	1	1	0	3	2	1	0	1	1	1	2	1	0	1	0
Wur. impatientis	2	1	0	0	1	1	1	1	0	2	2	0	1	1	1	1	3	2	1	0	1	1	1	0	0	0	1	0
Wur. montebelloensis	3	1	0	0	1	1	2	0	0	2	2	0	1	1	1	0	4	2	1	0	1	1	1	2	1	0	1	0
Wur. norrisi	3	1	0	0	1	1	2	0	0	2	2	0	1	1	1	0	3	2	1	0	1	1	1	0	0	0	1	0
Wur. patrellia	3	1	0	0	1	1	2	0	0	2	2	0	0	1	1	1	3	2	1	0	1	1	1	2	1	0	2	0
Wur. skevingtoni	2	1	0	0	2	1	1	1	0	2	2	0	1	1	1	1	3	2	1	0	1	2	1	0	0	0	1	0
Wur. windorah	3	3	0	0	1	1	2	0	0	2	2	0	1	0	1	1	3	2	1	0	1	1	1	2	1	0	1	0
Wur. wyperfeldensis	3	3	0	0	1	1	2	0	0	2	2	0	1	1	1	1	3	2	1	0	1	1	1	2	1	0	1	0

Taxa/Character	**120**										**130**										**140**							
Lig. satyrus	2	0	2	0	1	2	1	3	3	3	0	1	3	3	0	0	0	1	1	1	1	1	0	2	1	1	1	0
Lig. sinuatifascia	3	0	5	0	3	3	2	3&4&5	3	3&4&5	0	3	3	3	0	2	3	1	1	1	2	1	0	2	1	2	0	0
Bal. abscondita	2	1	0	0	1	3	2	3	3	3	0	1	3	1	1	0	0	1	1	1	2	1	0	2	1	1	1	0
Bal. bicuspis	2	0	0	0	1	3	2	3	3	3	0	1	3	3	1	0	0	1	1	1	2	1	0	1	1	1	0	0
Bal. centrosa	2	0	0	0	1	3	2	3	3	3	0	1	3	3	1	0	0	1	1	1	1	1	0	2	1	1	0	0
Bal. gigantea	3	0	0	0	3	2	2	3	3	3	0	2	3	3	2	0	0	0	1	1	1	2	0	2	1	1	0	0
Bal. kingcascadensis	3	0	2	1	1	3	2	3	3	3	0	2	3	3	0	0	0	1	1	1	2	2	0	2	1	1	1	0
Bal. latelimbata	3	0	0	0	1	3	2	3	3	3	0	2	3	3	0	0	0	1	1	1	1	1	0	2	1	2	0	0
K. adelaidica	3	0	0	0	1	3	2	3	3	3	0	3	3	3	2	2	2	0	1	1	2	1	0	2	1	1	1	0
K. corusca	4	0	0	0	1	3	2	3	3	3	0	3	3	3	2	2	2	0	1	1	2	1	0	2	1	1	0	0
K. irwini	3	0	0	0	1	3	2	3	3	3	0	2	3	3	2	2	2	0	1	1	2	1	0	2	1	1	0	0
K. westralica	3	0	0	0	1	3	2	3	3	3	0	3	3	3	2	2	2	0	1	1	2	1	0	2	1	1	1	0
Lar. bushblitzi	2	0	4	0	1	3	2	3	3	3	0	3	1	2	1	2	2	1	1	2	2	1	0	2	1	1	0	0
Lar. collessi	4	0	0	0	4	3	2	4	3	3	0	3	3	3	0	2	2	0	1	2	2	1	0	2	1	1	0	0
Lar. dimidiatipennis	4	0	3	0	1	2	2	3	3	3	0	3	1	2	0	2	1	0	1	1	2	1	0	2	1	1	0	0
Lar. zwicki	4	0	1	1	1	3	2	3	3	3	0	3	3	3	0	2	1	1	1	2	2	1	0	2	1	1	0	0
Mun. erugata	3	0	0	0	1	3	2	3	3	3	0	3	3	3	0	0	0	0	1	2	2	2	0	2	1	1	0	0
Mun. lepidokingi	3	0	0	0	1	2	2	3	3	3	0	3	3	1	1	1	0	0	1	2	2	2	0	1	1	1	0	0
Mun. paralutea	3	0	1	0	1	2	2	3	3	3	0	3	3	3	1	1	0	0	1	2	2	2	0	2	1	1	0	0
Muw. stellifera	4	0	4	0	3	3	2	3	3	3	0	3	3	1	0	2	2	0	1	2	2	1	0	2	1	2	0	0
Muw. vitreilinearis	3	0	1	0	3	3	2	3	3	3	0	3	3	3	2	2	1	0	1	1	2	1	0	2	1	2	0	0
Nga. trigonium	2	0	0	0	4	3	2	3	3	3	0	3	3	1	2	1	0	0	1	4	2	1	0	1	1	1	0	0
P. anaxios	0	0	0	1	2	2	1	3	4	5	0	3	3	3	0	1	0	1	1	2	2	1	0	2	1	1	0	0
P. basilikos	2	0	0	1	4	2	1	3	3	5	0	3	3	1	0	1	0	1	1	2	2	1	0	1	1	1	0	0
P. blackdownensis	2	0	0	1	3	2	1	3	4	5	0	3	3	3	0	2	1	1	1	2	2	1	0	2	1	1	1	0
P. bouchardi	3	1	0	1	2	2	1	3	4	5	0	2	3	1	1	1	0	1	1	2	1	1	0	2	1	1	0	0
P. cyanea	0	0	0	1	2	2	1	3	4	5	0	3	0	1	0	1	0	1	1	2	2	1	0	2	1	1	1	0
P. culgoafloodplainensis	2	0	0	1	4	3	2	3	3	3	0	3	0	0	0	1	0	1	1	2	1	1	0	2	1	1	0	0
P danielsi	2	0	0	1	4	2	1	3	4	5	0	3	3	1	0	0	0	1	1	2	2	1	0	2	1	1	2	0
P decora	3	0	0	1	3	2	2	3&5	4	3&5	0	3	3	3	1	1	0	0	1	2	2	1	0	2	1	2	0	0
P mackensiei	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
P marginicollis	2	1	0	1	3	2	1	3	4	5	0	3	3	3	0	1	0	1	1	2	2	1	0	1	1	1	1	0
P viridula	0	0	0	1	4	2	1	3	4	5	0	3	3	1	0	1	0	1	1	2	2	1	0	1	1	1	0	0
P whyalla	2	0	0	1	3	2	1	3	4	5	0	3	3	1	0	1	0	1	1	2	2	1	0	2	1	1	0	0
Wur. emu	3	0	0	0	1	3	2	3	3	3	0	3	3	1	0	1	0	0	1	1	2	1	0	1	1	2	0	0
Wur. impatientis	4	0	0	0	1	2	2	3	3	3	0	3	3	1	0	2	2	0	1	2	2	1	0	2	1	1	1	0
Wur. montebelloensis	3	0	0	0	1	3	2	3	3	3	0	2	3	3	0	2	1	0	1	3	2	1	0	2	1	2	1	0
Wur. norrisi	3	0	0	0	1	2	2	3	3	3	0	3	3	1	0	2	2	0	1	1	2	1	0	1	1	2	1	0
Wur. patrellia	3	0	0	0	1	3	2	3	3	3	0	2	3	1	0	2	1	0	1	1	2	1	0	2	1	2	0	0
Wur. skevingtoni	4	0	0	0	1	3	2	3	3	3	0	3	3	3	0	2	2	1	1	1	2	1	0	1	1	1	1	0
Wur. windorah	3	0	1	0	4	2	2	3	3	3	0	3	3	3	2	2	2	0	1	3	2	1	0	1	1	1	0	0
Wur. wyperfeldensis	3	0	0	0	1	3	2	3	3	3	0	3	3	1	0	2	1	0	1	2	2	1	0	2	1	2	1	0

Taxa/Character	150										160										170									
Lig. satyrus	3	1	2	1	2	1	1	3	1	1	1	1	0	1	2	3	0	3	3	0	0	0	0	2	1	1	0	7
Lig. sinuatifascia	2	2	2	1	2	1	1	2	1	1	2	1	0	1	2	3	0	3	2	0	0	0	0	2	1	0	1	3
Bal. abscondita	1	1	2	1	1	2	1	3	1	1	1	1	0	2	2	2	0	3	1	0	0	0	0	2	1	0	2	3
Bal. bicuspis	1	1	2	1	2	1	1	3	1	1	2	1	0	1	2	2	0	2	1	0	0	0	0	2	1	1	1	4
Bal. centrosa	1	1	2	1	1	2	1	2	1	1	2	1	0	1	2	2	0	3	0	0	0	0	0	2	1	1	0	4
Bal. gigantea	1	1	2	1	1	1	1	2	1	1	1	0	0	2	2	2	0	3	1	0	0	0	0	2	1	1	1	2
Bal. kingcascadensis	1	1	2	1	1	1	1	3	1	1	1	1	0	2	2	2	0	3	0	0	0	0	0	2	1	1	2	2
Bal. latelimbata	1	1	2	1	1	1	1	2	1	1	1	1	0	2	2	2	0	3	1	0	0	0	0	2	1	1	1	3
K. adelaidica	1	1	2	2	2	2	1	3	1	1	1	1	0	1	2	2	0	2	0	0	0	0	0	2	1	0	1	3
K. corusca	1	1	2	2	2	2	1	3	1	1	1	1	1	1	2	2	0	2	0	0	0	0	0	2	1	1	1	3
K. irwini	2	1	2	2	2	2	1	2	1	1	2	0	0	2	2	2	0	2	0	0	0	0	0	2	1	1	0	3
K. westralica	1	1	2	2	2	2	1	3	1	1	1	1	1	1	2	2	0	2	0	0	0	0	0	2	1	1	0	3
Lar. bushblitzi	2	1	1	2	2	2	1	2	1	1	1	1	0	1	2	2	0	3	?	?	?	?	?	?	?	?	?	?
Lar. collessi	1	1	2	2	2	2	1	3	1	1	1	1	0	1	2	2	0	3	1	0	0	0	1	2	1	0	1	3
Lar. dimidiatipennis	1	1	2	2	2	2	1	3	1	1	1	1	0	2	2	2	0	3	1	0	0	0	0	2	1	1	0	4
Lar. zwicki	2	1	1	2	2	1	1	2	1	1	1	2	0	1	2	2	0	3	?	?	?	?	?	?	?	?	?	?
Mun. erugata	2	1	2	1	1	2	1	3	1	1	1	0	0	2	2	3	0	3	0	0	0	0	0	2	1	1	0	3
Mun. lepidokingi	1	1	2	1	2	2	1	2	1	1	1	1	0	2	2	3	0	3	0	0	0	0	0	2	1	1	0	4
Mun. paralutea	1	1	2	1	1	1	1	3	1	1	1	0	0	1	2	3	0	3	1	0	0	0	0	2	1	2	1	3
Muw. stellifera	2	1	1	2	2	1	1	2	1	1	1	3	0	1	2	3	0	3	2	0	0	0	0	2	1	0	1	4
Muw. vitreilinearis	1	1	1	2	2	2	1	3	1	1	1	1	0	1	2	3	0	3	1	0	0	0	0	2	1	0	2	5
Nga. trigonium	1	1	2	1	1	2	1	2	1	1	1	3	0	2	2	3	0	3	1	0	0	0	0	2	1	0	1	3
P. anaxios	2	1	1	1	1	2	1	2	1	1	1	0	0	2	2	2	0	3	?	?	?	?	?	?	?	?	?	?
P. basilikos	2	1	1	1	1	2	1	2	2	1	1	0	0	2	2	2	0	3	2	0	0	0	0	2	1	0	1	2
P. blackdownensis	3	1	1	1	2	2	1	2	1	1	1	0	0	2	2	2	0	3	?	?	?	?	?	?	?	?	?	?
P. bouchardi	2	1	1	1	1	2	1	2	1	1	1	0	0	2	2	3	0	3	2	0	0	0	0	2	1	1	3	2
P. cyanea	2	1	1	1	2	2	1	2	1	1	1	0	0	2	2	3	0	3	2	0	0	0	0	2	1	0	2	1
P. culgoafloodplainensis	2	1	2	1	2	2	1	2	1	1	1	1	0	2	2	2	0	3	?	?	?	?	?	?	?	?	?	?
P danielsi	2	1	1	1	1	2	1	3	1	1	1	0	0	2	2	2	0	3	?	?	?	?	?	?	?	?	?	?
P decora	2	1	2	1	2	2	1	2	2	1	1	1	0	2	2	3	0	3	2	0	0	0	0	2	1	1	0	2
P mackensiei	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	0	0	0	2	1	0	3	2
P marginicollis	3	1	1	1	2	2	1	2	1	1	1	0	0	2	2	2	0	3	1	0	0	0	0	2	1	2	1	2
P viridula	3	1	1	1	1	2	1	2	1	1	1	1	0	2	2	3	0	3	2	0	0	0	0	2	1	0	2	2
P whyalla	2	1	1	1	2	2	1	2	1	1	1	0	0	2	2	3	0	3	2	0	0	0	0	2	1	2	1	2
Wur. emu	1	1	2	2	2	2	1	2	1	1	1	1	0	2	2	2	0	3	1	0	0	0	0	2	1	1	1	3
Wur. impatientis	3	1	2	1	2	2	1	3	1	1	1	0	0	1	2	2	0	3	0	0	0	0	0	2	1	1	2	5
Wur. montebelloensis	1	1	2	2	1	2	1	3	1	1	2	1	0	2	2	2	0	3	0	0	0	0	0	2	1	0	1	4
Wur. norrisi	3	1	2	1	1	2	1	3	1	1	1	0	0	1	2	2	0	3	0	0	0	0	0	2	1	1	2	6
Wur. patrellia	1	1	2	2	2	2	1	3	1	1	1	1	0	2	2	2	0	3	?	?	?	?	?	?	?	?	?	?
Wur. skevingtoni	3	1	2	1	1	2	1	3	1	1	1	1	0	2	2	2	0	2	?	?	?	?	?	?	?	?	?	?
Wur. windorah	1	1	2	2	2	2	1	2	1	1	1	0	0	2	2	2	0	3	0	0	0	0	0	2	1	0	0	5
Wur. wyperfeldensis	2	1	2	2	1	2	1	3	1	1	2	1	0	2	2	2	0	3	0	0	0	0	0	2	1	0	1	4

Taxa/Character	**180**										**190**										**200**							**207**
Lig. satyrus	0	1	2	2	0	2	2	2	1	1	1	0	0	1	0	0	1	1	1	0	0	2	1	0	0	1	1	0
Lig. sinuatifascia	0	1	2	2	0	1	1	2	1	1	1	0	0	0	0	0	1	1	0	0	0	1	1	0	0	1	1	0
Bal. abscondita	0	2	2	1	0	2	2	2	1	1	1	4	1	0	3	0	1	1	0	0	0	1	0	0	0	1	1	0
Bal. bicuspis	0	1	2	2	0	2	2	2	1	1	1	5	0	1	3	1	1	1	0	0	0	1	0	0	0	1	1	0
Bal. centrosa	0	1	2	2	0	2	2	2	1	2	1	5	0	0	3	0	1	1	0	0	0	1	0	0	0	1	1	0
Bal. gigantea	0	1	2	2	0	2	2	2	1	1	1	5	0	0	3	0	1	1	0	0	0	1	0	0	0	1	1	0
Bal. kingcascadensis	0	2	2	2	0	2	2	2	1	2	1	5	0	0	3	0	1	1	0	0	0	1	0	0	0	1	1	0
Bal. latelimbata	0	1	2	1	0	1	1	2	1	1	1	5	0	0	3	0	1	1	0	0	0	1	0	0	0	1	1	0
K. adelaidica	0	2	2	2	0	2	2	2	1	2	1	1	0	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0
K. corusca	0	1	2	2	0	2	1	2	1	2	1	0	0	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0
K. irwini	0	1	2	2	0	2	2	2	1	2	1	0	0	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0
K. westralica	0	1	2	2	0	1	2	2	1	2	1	0	0	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0
Lar. bushblitzi	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Lar. collessi	0	1	1	2	0	1	1	2	1	2	1	1	0	1	0	0	1	0	0	0	0	2	1	0	0	1	1	0
Lar. dimidiatipennis	0	1	2	2	0	1	1	2	1	2	1	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	1	0
Lar. zwicki	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Mun. erugata	0	2	2	2	0	1	1	2	1	2	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	1	1	0
Mun. lepidokingi	0	2	1	2	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	1	0	0	0	1	1	0
Mun. paralutea	0	1	2	1	0	1	1	2	1	2	1	1	0	1	0	1	1	1	0	0	0	1	1	0	0	1	1	0
Muw. stellifera	0	1	2	2	0	1	1	2	1	2	1	0	0	0	0	1	1	1	0	0	0	1	1	0	0	1	1	0
Muw. vitreilinearis	0	1	2	2	0	2	2	2	1	2	2	1	0	1	0	1	1	1	0	0	0	1	1	0	0	1	1	0
Nga. trigonium	0	1	2	2	0	1	1	2	1	1	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
P. anaxios	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
P. basilikos	0	2	2	2	0	2	1	2	1	0	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0
P. blackdownensis	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
P. bouchardi	0	2	2	2	0	1	1	2	1	1	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0
P. cyanea	0	2	2	2	0	2	2	2	1	0	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0
P. culgoafloodplainensis	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
P danielsi	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
P decora	0	2	2	1	0	1	2	2	1	1	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0
P mackensiei	0	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
P marginicollis	0	2	2	2	0	2	1	2	1	0	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0
P viridula	0	2	2	2	0	2	1	2	1	0	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0
P whyalla	0	2	2	2	0	2	1	2	1	0	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0
Wur. emu	0	1	2	2	0	2	1	2	1	1	1	0	0	0	0	0	1	1	0	0	0	1	1	0	0	1	1	0
Wur. impatientis	0	1	1	2	0	2	1	2	1	2	1	1	0	1	0	1	1	1	0	0	0	1	1	0	0	1	1	0
Wur. montebelloensis	0	1	2	1	0	1	1	2	1	2	1	1	0	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0
Wur. norrisi	0	1	1	2	0	2	1	2	1	2	1	1	0	1	0	1	1	1	0	0	0	1	1	0	0	1	1	0
Wur. patrellia	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Wur. skevingtoni	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Wur. windorah	0	1	1	1	0	2	2	2	1	2	1	0	0	0	0	0	1	1	0	0	0	1	1	0	0	1	1	0
Wur. wyperfeldensis	0	1	2	1	0	1	1	2	1	1	2	1	0	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0

Appendix 3

Matrix of 207 characters for the Australian *Balaana* genus-group. (doi: 10.3897/zookeys.150.1881.app) File format: NEXUS matrix file.

Explanation note: This NEXUS phylogenetic matrix of 207 morphological characters for 42 Australian exoprospine beeflies was created in Mesquite v2.74. The file also contains the most parsimonious tree 5 file from the PAUP* maximum parsimony phylogenetic analysis. The data is also deposited in the Dryad Repository: doi: 10.5061/dryad.5j64k and in the TREEBASE Repository at http://purl.org/phylo/treebase/phylows/study/TB2:S12050

Citation: Lambkin CL, Bartlett JS (2011) Bush Blitz aids description of three new species and a new genus of Australian beeflies (Diptera, Bombyliidae, Exoprosopini). In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 231–280. doi: 10.3897/zookeys.150.1881.app

ZooKeys 150: 281–326 (2011)
doi: 10.3897/zookeys.150.2146
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

An account of the taxonomy and distribution of Syllidae (Annelida, Polychaetes) in the eastern Mediterranean, with notes on the genus *Prosphaerosyllis* San Martín, 1984 in the Mediterranean

Sarah Faulwetter[1,4], Georgios Chatzigeorgiou[2,4], Bella S. Galil[3], Christos Arvanitidis[4]

1 *Department of Zoology – Marine Biology, Faculty of Biology, National and Kapodestrian University of Athens, Panepistimiopolis, 15784, Athens, Greece* **2** *Department of Biology, University of Crete, 71409 Heraklion, Crete, Greece* **3** *National Institute of Oceanography, Israel Oceanographic and Limnological Research, POB 8030, Haifa 31080, Israel* **4** *Institute of Marine Biology and Genetics, Hellenic Centre for Marine Research, 71003 Heraklion, Crete, Greece*

Corresponding author: *Sarah Faulwetter* (sarifa@hcmr.gr)

Academic editor: *V. Smith* | Received 27 September 2011 | Accepted 23 November 2011 | Published 28 November 2011

Citation: Faulwetter S, Chatzigeorgiou G, Galil BS, Arvanitidis C (2011) An account of the taxonomy and distribution of *Syllidae* (Annelida: Polychaetes) in the eastern Mediterranean, with notes on the genus *Prosphaerosyllis* San Martín, 1984 in the Mediterranean. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 281–326. doi: 10.3897/zookeys.150.2146

Abstract

The syllid fauna of three locations in Crete and Israel (eastern Mediterranean Sea) was studied, yielding 82 syllid species, many of which were found for the first time in the respective areas: Seventeen species were recorded for the first time on the Israeli coasts and 20 in Greek waters. *Perkinsyllis augeneri* (Hartmann-Schröder, 1979) and *Prosphaerosyllis chauseyensis* Olivier et al., 2011 are new records for the Mediterranean Sea. Detailed information is given on the morphology, ecology and distribution of the species recorded for the first time in the studied areas. In addition, an update on the distribution of the genus *Prosphaerosyllis* San Martín, 1984 in the Mediterranean is given and an identification key to the Mediterranean species is provided.

Keywords
Polychaetes, Syllidae, eastern Mediterranean Sea, taxonomy, distribution, new records, alien species

Introduction

The Syllidae are a highly diverse family of polychaetes with currently around 900 valid species belonging to over 80 genera (pers. obs.) and have recently received considerable taxonomic and phylogenetic research effort, including a high number of new taxon descriptions (e.g. Aguado et al. 2007, Aguado and San Martín 2009, De Matos Nogueira et al. 2001, San Martín 2005, 2008, San Martín and Hutchings 2006, San Martín et al. 2009). Syllids are (usually) small-sized polychaetes with a high diversity of morphological and ecological features and are found globally on all types of substrates from the intertidal to the abyss (San Martín 2003).

The present study contributes to the current knowledge of the syllid fauna of three different locations in the eastern Mediterranean Sea: two in Crete, one in Israel. The material has been collected in the framework of two different research programmes and from two different habitats (Fig. 1, Table 1): a) hard-bottom samples from Crete have been obtained within the NaGISA project (Natural Geography in Shore Areas, http://www.nagisa.coml.org), a field project of the Census of Marine Life (COML, http://www.coml.org); b) soft-sediment samples from the Israeli coast have been obtained in the framework of a project focusing on the soft bottom benthos of Haifa Bay. In all samples, Syllidae were highly abundant and yielded many species recorded for the first time in the respective area, as well as a species new to science (Faulwetter et al. 2011).

In the Mediterranean Sea, syllids have been studied by numerous authors in extensive taxonomic and biogeographic works (e.g. Ben-Eliahu 1977a, 1977b, Campoy 1982, Çinar 1999, San Martín 1984b, 2003, Musco and Giangrande 2005), however, most research on the taxon is being carried out in the western Mediterranean basin, whereas the syllid fauna of the eastern Mediterranean has only recently started to be investigated more intensely (e.g. Ben-Eliahu 1977a, 1977b, Çinar 1999, Çinar and Ergen 2002, 2003, Çinar et al. 2003, Aguado and San Martín 2007, Abd-Elnaby and San Martín 2010, 2011). In Greece, polychaetes have been studied by various authors (e.g. Bellan 1964, Fassari 1982, Arvanitidis 1994, 2000, Simboura 1996, Simboura and Nicolaidou 2001, Antoniadou et al. 2004). However, the only studies in the Aegean Sea focussing specifically on Syllidae are those of Çinar (1999) and Çinar and Ergen (2002) from the Turkish Aegean coasts. Polychaetes of the Mediterranean coast of Israel have been studied by Monro (1937), Tebble (1959), Fauvel (1955, 1957), Ben-Eliahu (1976a, 1976b), Ben-Eliahu and Golani (1990) and Ben-Eliahu and Fiege (1995) and syllids in particular by Harlock and Laubier (1966) and Ben-Eliahu (1977a, 1977b).

Table 1. Sampling stations and their characteristics

Station Code	Location	Coordinates	Depth	Habitat
ALA-IL-1	Haifa Bay, Israel	32°53.792'N, 35°03.928'E	13.1 m	Fine to medium sand
ALA-IL-2	Haifa Bay, Israel	32°54.052'N, 35°03.905'E	13.9 m	Sand of mixed grain sizes
ALA-IL-5	Haifa Bay, Israel	32°54.259'N, 35°04.160'E	11.4 m	Silty sand
ALA-IL-7	Haifa Bay, Israel	32°54.544'N, 35°04.093'E	10.5 m	Sand of mixed grain sizes with silt
ALA-IL-8	Haifa Bay, Israel	32°55'N, 35°04.239'E	7.8 m	Coarse sand with silt
ALA-IL-9	Haifa Bay, Israel	32°54.518'N, 35°03.950'E	8.7 m	Coarse sand
ALA-IL-10	Haifa Bay, Israel	32°52.509'N, 35°03.520'E	12.8 m	Medium to coarse sand
CALA-1, CALB-1	Alykes, Crete, Greece	35°24.95'N, 24°59.25'E	1 m	*Cystoseira* spp., *Fucus virsoides*
CALA-5, CALB-5	Alykes, Crete, Greece	35°24.95'N, 24°59.25'E	5 m	Filamentous Chlorophyceae, *Amphiroa* sp., *Padina pavonica*
CALA-10, CALB-10	Alykes, Crete, Greece	35°24.95'N, 24°59.25'E	10 m	*Cystoseira* spp., filamentous Chlorophyceae
CALA-15, CALB-15	Alykes, Crete, Greece	35°24.95'N, 24°59.25'E	15 m	Filamentous Chlorophyceae, filamentous Phaeophyceae
CALA-20, CALB-20	Alykes, Crete, Greece	35°24.95'N, 24°59.25'E	20 m	Filamentous Phaeophyceae, *Bryopsis* sp., *Caulerpa* spp.
CELA-1, CELB-1	Elounda, Crete, Greece	35°15.1'N, 25°45.5'E	1 m	*Jania* sp., *Dasycladus clavaeformis*, Porifera spp., *Litophyllum* sp.
CELA-5, CELB-5	Elounda, Crete, Greece	35°15.1'N, 25°45.5'E	5 m	*Jania* sp., *Dasycladus clavaeformis, Litophyllum* sp., *Amphiroa* sp.
CELA-10, CELB-10	Elounda, Crete, Greece	35°15.1'N, 25°45.5'E	10 m	Filamentous Phaeophyceae, *Jania* sp., Porifera spp., *Bryopsis* sp.
CELA-15, CELB-15	Elounda, Crete, Greece	35°15.1'N, 25°45.5'E	15 m	Filamentous Phaeophyceae, *Jania* sp., *Peyssonellia* sp., filamentous Chlorophyceae
CELA-20, CELB-20	Elounda, Crete, Greece	35°15.1'N, 25°45.5'E	20 m	*Padina pavonica*, filamentous Chlorophyceae, *Amphiroa* sp.

This paper gives an account of the syllid species encountered in the three sampling locations and provides detailed information on the morphology, distribution and ecology of those species recorded for the first time in the respective area. Furthermore, during this study it became clear that the distribution range of the genus *Prosphaerosyllis* San Martín, 1984 in the Mediterranean is outdated or confused. In addition, since several new species have recently been described in this genus (Çinar et al. 2011, Olivier et al. 2011) and were also identified in the present material, an update on the distribution of the genus *Prosphaerosyllis* in the Mediterranean and an updated identification key are provided at the end of this paper.

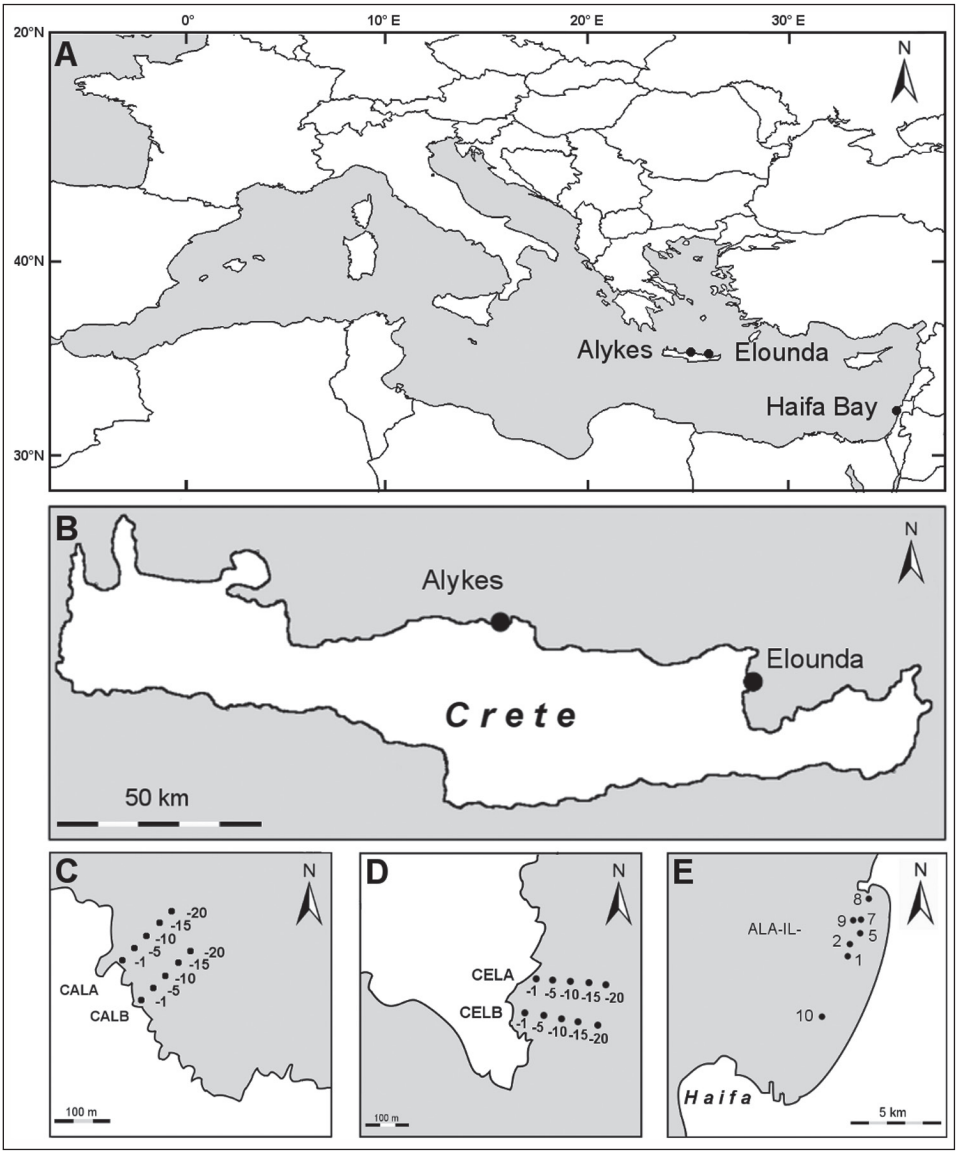


Figure 1. Map of the sampling stations **A** Location of the stations in the Mediterranean **B** Locations of the two sampling stations in Crete **C** Alykes **D** Elounda **E** Haifa Bay.

Material and methods

Specimen collection and processing

Specimens from Israel were collected on 31 May 2009 and 11 Oct 2009 in Haifa Bay, (Israel, eastern Mediterranean Sea) from soft sediments of mixed grain sizes in shallow waters (Table 1). Sediment samples were taken with a Van-Veen grab (KAHLSICO,

model WA265/SS214) 32×35cm, volume 20 l, penetration 20 cm. The sediment was preserved in buffered formalin 10% for 3–7 days, then sieved through a 250 μm mesh sieve and subsequently stored in 70% ethanol. In this study, only a subset of the collected material is presented.

Specimens from Crete were collected in September 2007 and June 2008 from two sites in northern Crete characterized by a continuous hard bottom habitat with dense algal coverage and a moderate wave exposure (Table 1). At each site, two vertical transects with sampling depths at 1 m, 5 m, 10 m, 15 m and 20 m were defined and five replicates were taken from each transect and depth. Samples were collected by means of SCUBA diving according to the NaGISA protocol (Iken and Konar 2003). A plexiglas frame (25 × 25 cm) with a net of 0.5 mm mesh size attached to its top opening was placed onto the rock and the surface within the frame was scraped off. The sample was collected by a manually operated suction device, supplied by air from an extra scuba tank. Large particles (>2 cm) were collected manually after suction. The samples were subsequently washed through a 0.5 mm mesh sieve, fixed and preserved in 99% ethanol.

Specimens were examined under an Olympus SZx12 stereomicroscope and an Olympus BX50 microscope and identified by employing the most recent literature on Syllidae (e.g. Nygren 2004, San Martín 2003, 2005, San Martín and Hutchings 2006). Illustrations in pencil were made by means of a drawing tube, subsequently scanned, imported into a graphic program (GIMP), re-drawn and saved as a vector graphic. All specimens are deposited in the invertebrate collection of the Institute of Marine Biology and Genetics, Hellenic Centre for Marine Research. Comparative material has been loaned by the Zoologisches Museum and Institut, Universität Hamburg, Germany, Ege University, Izmir, Turkey and the Muséum National d'Histoire Naturelle, Paris, France.

Information on habitat and global distribution of species was adopted from San Martín (2003), unless indicated otherwise, and updated with findings from this study. Information on species distribution among Mediterranean regions was adopted from Musco and Giangrande (2005) and updated according to recent literature and to findings from this study. Abbreviations for biogeographic regions used in the text are: MED (Mediterranean), WB (Western Basin), EB (Eastern Basin), CB (Central Basin), AD (Adriatic Sea), AS (Aegean Sea), BS (Black Sea), LB (Levantine Basin), following Arvanitidis et al. 2002 who modified Por's (1989) system.

Electronic publication

This manuscript was prepared in a Virtual Research Environment (Scratchpads) allowing for rapid and simultaneous publication of the results in print as well as electronically in a semantically enhanced form (Blagoderov et al. 2010, Penev et al. 2010). This publication and all supplementary data (tables, figures, taxon information) are

also available under a Creative Commons license on the Polychaete Scratchpads (http://polychaetes.marbigen.org).

The underlying dataset of this study has been published under a Creative Commons license according to the Pensoft Data Publishing Policies and Guidelines for Biodiversity Data (Penev et al. 2011) and are available through the GBIF Integrated Publishing Toolkit hosted by Pensoft (http://ipt.pensoft.net/ipt/resource.do?r=easternmedsyllids). The data are furthermore available in Darwin Core Archive format, a simple and extensible schema for sharing biodiversity data which has been developed by the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/informatics/standards-and-tools/publishing-data/data-standards/darwin-core-archives/) to allow easy and rapid mobilisation of species occurrence data through the internet. Darwin Core Archives are essentially a set of text files stored together with an XML descriptor file which describes the structure of the data files. Data are described through the Darwin Core schema, allowing for their usage within the semantic web. This new type of data publishing allows data to be indexed and discoverable through global biodiversity infrastructures such as GBIF or other data repositories, allows data to be integrated and compared with other datasets and ensures proper accreditation of the data provider (Penev et al. 2011). Additionally, the data have been deposited in the Dryad Data Repository (http://www.datadryad.org) and can be accessed at doi: 10.5061/dryad.4b7k408g.

Results

Examination of a total of 111 samples yielded 82 syllid species (Table 2), of which 49 were found in Alykes (Crete), 62 in Elounda (Crete) and 23 in Haifa Bay (Israel). Species of all subfamilies have been found in the stations in Crete, with the majority (80%) of species belonging to Syllinae and Exogoninae, whereas the samples from Israel did not contain any specimens of Anoplosyllinae or Autolytinae, and 73% of the examined species belong to the small-sized Exogoninae (Fig. 2). The material yielded a number of species reported for the first time in the studied areas: Twenty species are reported for the first time in Greek waters, of these, six are new additions to the Aegean fauna. Seventeen species are newly reported for the Israeli coast, of these, 4 are also new records for the Levantine Basin. The studied material yielded also 4 species which are new additions to the eastern Mediterranean and 2 to the Mediterranean fauna (Table 2, Fig. 3). Information on morphology, distribution and ecology of the newly recorded species are given below.

Table 2. Species occurences at sampling stations. Years and replicates have been pooled. Transects sampled in Alykes and Elounda have been combined. †= new record for Greece, ‡= new record for the Aegean, §= new record for Israel, |= new record for the Levantine Basin, ¶= new record for the eastern Mediterranean, #= new record for the Mediterranean.

Species	CALA-1 CALB-1	CALA-5 CALB-5	CALA-10 CALB-10	CALA-15 CALB-15	CALA-20 CALB-20	CELA-1 CELB-1	CELA-5 CELB-5	CELA-10 CELB-10	CELA-15 CELB-15	CELA-20 CELB-20	ALA-IL-1	ALA-IL-2	ALA-IL-5	ALA-IL-7	ALA-IL-8	ALA-IL-9	ALA-IL-10
Branchiosyllis exilis (Gravier, 1900)		+					+	+						+			
Brania arminii (Langerhans, 1881)		+			+												
Brania pusilla (Dujardin, 1851)						+		+									
Eurysyllis tuberculata Ehlers, 1864	+			+		+	+	+	+	+				+			
Eusyllis assimilis Marenzeller, 1875			+			+	+	+	+								
Eusyllis lamelligera Marion & Bobretzky, 1875					+												
Exogone dispar (Webster, 1879)	+				+		+	+	+								
Exogone naidina Örsted, 1845			+	+	+	+	+	+	+	+							
Exogone rostrata Naville, 1933				+													
Exogone verugera (Claparède, 1868)		+		+	+		+	+	+								
Haplosyllis spongicola (Grube, 1855)	+	+	+	+		+	+	+	+								
Myrianida convoluta (Cognetti, 1953)			+			+		+	+								
Myrianida edwarsi (Saint-Joseph, 1887)						+											

Species	CALA-1 CALB-1	CALA-5 CALB-5	CALA-10 CALB-10	CALA-15 CALB-15	CALA-20 CALB-20	CELA-1 CELB-1	CELA-5 CELB-5	CELA-10 CELB-10	CELA-15 CELB-15	CELA-20 CELB-20	ALA-IL-1	ALA-IL-2	ALA-IL-5	ALA-IL-7	ALA-IL-8	ALA-IL-9	ALA-IL-10
Myrianida inermis (Saint-Joseph, 1887) †, ‡						+											
Myrianida prolifera (O.F. Müller, 1788)						+		+									
Myrianida quindecimdentata (Langerhans, 1884) †	+		+			+	+	+									
Nudisyllis divaricata (Keferstein, 1862)						+		+	+								
Odontosyllis ctenostoma Claparède, 1868			+			+	+	+	+								
Odontosyllis fulgurans (Audouin & Milne Edwards, 1834)		+	+	+	+		+	+	+	+							
Odontosyllis gibba Claparède, 1863		+	+				+	+	+	+							
Opisthosyllis brunnea Langerhans, 1879 †						+	+										
Paraehlersia ferrugina (Langerhans, 1881)	+	+	+	+	+		+	+	+	+							
Parapionosyllis brevicirra Day, 1954							+							+			
Parapionosyllis elegans (Pierantoni, 1903) §														+			+
Parapionosyllis minuta (Pierantoni, 1903)														+			+
Parexogone hebes Cognetti, 1955 §											+			+			
Perkinsyllis augeneri (Hartmann-Schröder, 1979) §, |, ¶, #														+			

Species	CALA-1 CALB-1	CALA-5 CALB-5	CALA-10 CALB-10	CALA-15 CALB-15	CALA-20 CALB-20	CELA-1 CELB-1	CELA-5 CELB-5	CELA-10 CELB-10	CELA-15 CELB-15	CELA-20 CELB-20	ALA-IL-1	ALA-IL-2	ALA-IL-5	ALA-IL-7	ALA-IL-8	ALA-IL-9	ALA-IL-10
Plakosyllis brevipes Hartmann-Schröder, 1956			+		+				+								
Prosphaerosyllis adelae San Martín, 1984 §, |, ¶												+		+	+		+
Prosphaerosyllis campoyi (San Martín, Acero, Contonente & Gómez, 1982)†								+									
Prosphaerosyllis chauseyensis Olivier et al. 2011 §, |, ¶, #											+	+	+	+	+	+	+
Prosphaerosyllis longipapillata (Hartmann-Schröder, 1979) §														+			
Prosphaerosyllis marmarae Çinar et al. 2011 §												+		+	+		
Prosphaerosyllis xarifae (Hartmann-Schröder, 1960) †,§							+	+									+
Salvatoria alvaradoi (San Martín, 1984) †, ‡		+	+			+	+	+	+	+							
Salvatoria clavata (Claparède, 1863)	+	+	+	+	+	+	+	+	+	+							
Salvatoria euritmica (Sardà, 1984) †	+				+	+		+	+	+							
Salvatoria limbata (Claparède, 1868)	+	+	+	+	+	+	+	+	+	+							
Salvatoria neapolitana (Goodrich, 1930) †						+			+	+							

Species	CALA-1 CALB-1	CALA-5 CALB-5	CALA-10 CALB-10	CALA-15 CALB-15	CALA-20 CALB-20	CELA-1 CELB-1	CELA-5 CELB-5	CELA-10 CELB-10	CELA-15 CELB-15	CELA-20 CELB-20	ALA-IL-1	ALA-IL-2	ALA-IL-5	ALA-IL-7	ALA-IL-8	ALA-IL-9	ALA-IL-10
Salvatoria vieitezi (San Martín, 1984) †	+	+	+	+	+	+	+	+	+	+							
Salvatoria yraidae (San Martín, 1984) †			+	+	+		+	+	+	+							
Sphaerosyllis austriaca Banse, 1959	+			+			+	+									
Sphaerosyllis bulbosa Southern, 1914 §														+			+
Sphaerosyllis glandulata Perkins, 1981 †, §									+					+	+		+
Sphaerosyllis gravinae Somaschini & San Martín, 1994 §, |, ¶															+		
Sphaerosyllis hystrix Claparède, 1863		+				+		+						+			
Sphaerosyllis levantina Faulwetter et al. 2011														+			
Sphaerosyllis pirifera Claparède, 1868	+	+	+	+	+	+	+	+	+	+							
Sphaerosyllis sp. [San Martín, 2003]														+			
Sphaerosyllis taylori Perkins, 1981 §											+	+		+	+		+
Sphaerosyllis thomasi San Martín, 1984 §														+			
Syllides edentatus Westheide, 1974 †							+	+									
Syllides fulvus (Marion & Bobretzy, 1875)					+	+	+	+	+								
Syllides japonicus Imajima, 1966 ‡							+		+								

Species	CALA-1 CALB-1	CALA-5 CALB-5	CALA-10 CALB-10	CALA-15 CALB-15	CALA-20 CALB-20	CELA-1 CELB-1	CELA-5 CELB-5	CELA-10 CELB-10	CELA-15 CELB-15	CELA-20 CELB-20	ALA-IL-1	ALA-IL-2	ALA-IL-5	ALA-IL-7	ALA-IL-8	ALA-IL-9	ALA-IL-10
Syllis alternata Moore, 1908	+		+	+	+	+	+	+	+	+							
Syllis armillaris (O.F. Müller, 1771)	+	+	+	+		+	+	+	+								
Syllis beneliahuae (Campoy & Alquézar, 1982)	+	+	+	+	+	+		+		+							
Syllis columbretensis (Campoy, 1982)	+	+	+	+	+	+	+	+	+	+							
Syllis compacta Gravier, 1900 †	+	+		+	+		+	+	+	+							
Syllis corallicola Verrill, 1900	+	+	+	+	+	+	+	+	+	+							
Syllis cruzi Núnez & San Martín, 1991 †, ‡					+			+									
Syllis ferrani Alós & San Martín, 1987						+	+	+		+							
Syllis garciai (Campoy, 1982)	+	+	+	+	+	+		+	+	+		+		+			
Syllis gerlachi (Hartmann-Schröder, 1960)	+	+	+	+	+	+	+	+	+	+							
Syllis gerundensis (Alós & Campoy, 1981) †, ‡		+			+	+	+										
Syllis gracilis Grube, 1840	+					+		+									
Syllis hyalina Grube, 1863	+	+		+	+	+	+	+	+	+							
Syllis jorgei San Martín & López, 2000 §	+				+	+	+							+			
Syllis krohnii Ehlers, 1864		+			+	+	+	+	+	+							

Species	CALA-1 CALB-1	CALA-5 CALB-5	CALA-10 CALB-10	CALA-15 CALB-15	CALA-20 CALB-20	CELA-1 CELB-1	CELA-5 CELB-5	CELA-10 CELB-10	CELA-15 CELB-15	CELA-20 CELB-20	ALA-IL-1	ALA-IL-2	ALA-IL-5	ALA-IL-7	ALA-IL-8	ALA-IL-9	ALA-IL-10
Syllis parapari San Martín & López, 2000					+												
Syllis prolifera Krohn, 1852	+	+	+	+	+	+	+	+	+	+							
Syllis pulvinata (Langerhans, 1881) †, ‡						+	+		+								
Syllis rosea (Langerhans, 1879)	+																
Syllis tyrrhena (Licher & Kuper, 1998) †, ‡, ¶								+									
Syllis variegata Grube, 1860						+	+		+								
Syllis westheidei San Martín, 1984 †				+													
Synmerosyllis lamelligera (Saint-Joseph, 1887)	+			+		+			+								
Trypanosyllis aeolis Langerhans, 1879									+								
Trypanosyllis coeliaca Claparède, 1868 §	+	+	+			+	+	+	+					+			
Trypanosyllis zebra (Grube, 1860)	+					+		+									
Virchowia clavata Langerhans, 1879						+											
Xenosyllis scabra (Ehlers, 1864)	+			+	+	+	+	+	+								

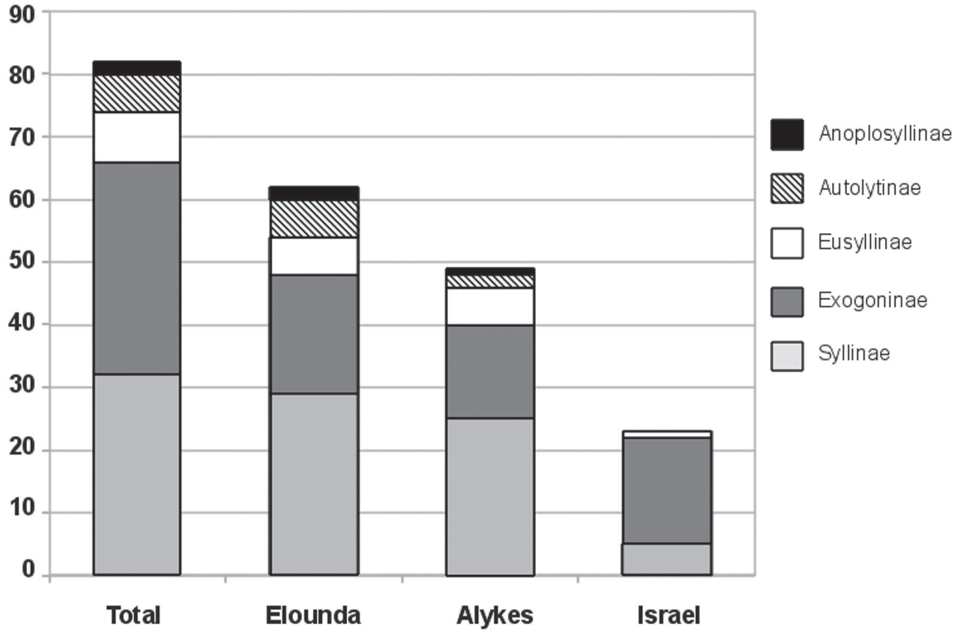


Figure 2. Numbers of species per subfamily at the three locations and in total.

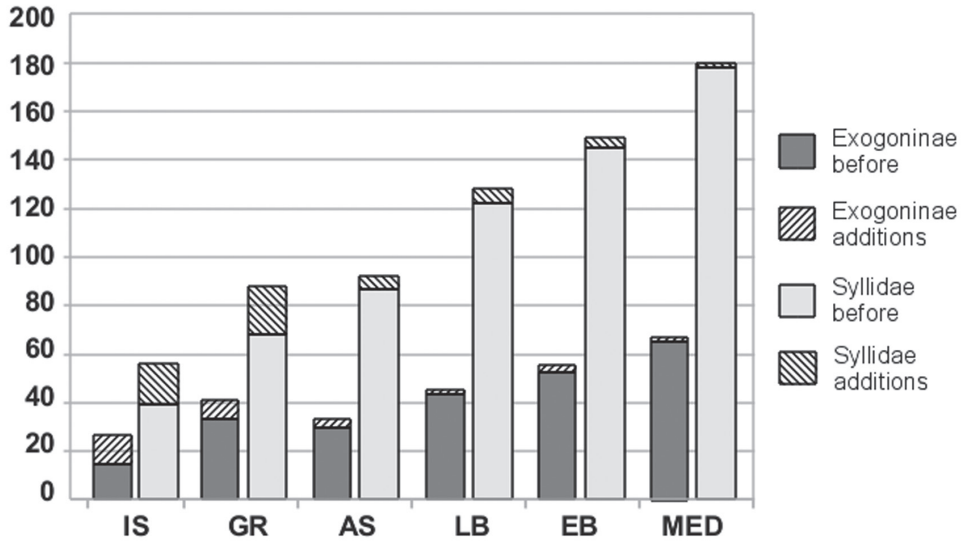


Figure 3. Numbers of additions of Syllidae and Exogoninae to various regions of the Mediterranean. IS=Israel, GR=Greece, AS=Aegean Sea, LB=Levantine Basin, EB=Eastern Basin, MED=Mediterranean.

New records

Subfamily Anoplosyllinae Aguado & San Martín, 2009

Genus *Syllides* Ørsted, 1845

Type species. *Syllides longocirrata* Ørsted, 1845

Syllides edentatus Westheide, 1974

http://species-id.net/wiki/Syllides_edentatus

Syllides japonica edentata Westheide, 1974a: 81, figs 36e, 37; Campoy 1982: 320; San Martín et al. 1985: 32.
Syllides edentatus: San Martín 1984b: 143, fig. 27; 2003: 143, fig. 70; Çinar 1999: 211, fig. 4.86; Çinar and Gambi 2005: 753.

Material examined. Elounda, Crete, Greece: CELA-5b-08 (2 ind.), CELA-5d-08 (2 ind.) [coll. 12.6.2008]; CELB-10c-07 (1 ind.) [coll. 27.9.2007].

Type locality. Galápagos Islands (Pacific Ocean).

Distribution. Galápagos Islands, north-east Pacific, Atlantic, Mediterranean Sea: WB, AS. New record for the Greek coast.

Habitat. Shallow subtidal depths, in sandy and muddy sediments, among photophilic algae and *Zostera* beds, in vermetid reefs.

Syllides japonicus Imajima, 1966

http://species-id.net/wiki/Syllides_japonicus

Syllides japonicus Imajima, 1966: 112, figs 36a–h; Banse 1971: 1477, fig. 5; San Martín 2003: 142, fig. 69; San Martín and Hutchings 2006: 360, figs 86c–f, 87a–e.
Syllides cf. *japonicus*: San Martín 1984b: 139, fig. 26.

Material examined. Elounda, Crete, Greece: CELA-15a-07 (1 ind.) [coll. 26.9.2007]; CELA-5d-08 (1 ind.), CELB-15d-08 (1 ind.) [coll. 12.6.2008].

Type locality. Japan (Pacific Ocean).

Distribution. Japan, Australia (San Martín and Hutchings 2006), Mediterranean Sea: WB, AS, LB (Abd-Elnaby and San Martín 2010). New record for the Aegean Sea.

Habitat. Shallow subtidal depths, in sandy and muddy sediments, on rocks with algal cover, among *Posidonia oceanica* rhizomes.

Subfamily Autolytinae Langerhans, 1879

Genus *Myrianida* Milne Edwards, 1845

Type species. *Myrianida fasciata* Milne Edwards, 1845

***Myrianida inermis* (Saint-Joseph, 1887)**
http://species-id.net/wiki/Myrianida_inermis

Autolytus inermis Saint-Joseph, 1887: 237, pl. 12, fig. 117; Gidholm 1967: 193, fig. 22; Campoy 1982: 235; San Martín 1994: 274, fig. 4; 2003: 487, figs 267a, c–e; Hartmann-Schröder 1996: 182.
Autolytus (Autolytides) inermis: Fauvel 1923: 322, figs 123h–k.
Myrianida inermis: Nygren 2004: 135, figs 65a–e.

Material examined. Elounda, Crete, Greece: CELB-1e-07 (1 ind.) [coll. 29.9.2007].

Type locality. Dinard, France (north-east Atlantic Ocean).

Distribution. North-east Atlantic, north-west Atlantic (San Martín 1994), north-east Pacific (Nygren 2004), Arctic (Ramos et al. 2010). Mediterranean Sea: WB, AS. New record for the Aegean Sea.

Habitat. Until 100m depth, on rocks among algae and hydrozooans, in coralligenous substrates (Nygren 2004, San Martín 2003).

***Myrianida quindecimdentata* (Langerhans, 1884)**
http://species-id.net/wiki/Myrianida_quindecimdentata

Autolytus quindecimdentatus Langerhans, 1884: 249, pl. 15, figs 3a–b; Gidholm 1967: 195, fig. 23; Ben-Eliahu 1977a: 86, fig. 13; Campoy 1982: 241; San Martín 1984b: 417, fig. 113; 2003: 494, figs 272a–d, 273a–b; Núñez and San Martín 1996: 213, figs 5k–m; Hartmann-Schröder 1996: 185; Çinar 1999: 63, fig. 4.8; Çinar et al. 2003: 747.
Autolytus lugens Saint-Joseph, 1887: 234, pl. 12, fig. 116; Fauvel, 1923: 318, fig. 122g; Cognetti 1961: 304.
Odontosyllis longicornis Hartmann-Schröder, 1960: 98, figs 101–104.
Myrianida quindecimdentata: Nygren 2004: 135, figs 77a–e.

Material examined. Alykes, Crete, Greece: CALA-10c-08 (4 ind.) [coll. 17.6.2008]; CALA-1b-08 (1 ind.), CALB-1c-08 (1 ind.), CALB-1d-08 (1 ind.) [coll. 18.6.2008]. Elounda, Crete, Greece: CELB-5e-07 (1 ind.) [coll. 27.9.2007]; CELB-1a-07 (2 ind.), CELA-1d-07 (2 ind.), CELB-1e-07 (5 ind.) [coll. 29.9.2007]; CELA-10b-08 (1 ind.) [coll. 11.6.2008]; CELB-1a-08 (1 ind.), CELB-1b-08 (1 ind.), CELA-5d-08 (1 ind.) [coll. 12.6.2008].

Type locality. Madeira (Atlantic Ocean).

Distribution. East and west Atlantic (European and African coasts, Cuba), northeast Pacific, Red Sea (San Martin 1994, Nygren 2004). Mediterranean Sea: WB, CB, AD, AS, LB. New record for the Greek coast.

Habitat. Subtidal depths, on biogenic calcareous substrates, among photophilic and sciaphilic algae and *Posidonia oceanica* rhizomes, endobiontic in sponges (Nygren 2004, San Martín 2003).

Subfamily Eusyllinae Malaquin, 1893

Genus *Perkinsyllis* San Martín, López & Aguado, 2009

Type species. *Pionosyllis longisetosa* Hartmann-Schröder, 1965

Perkinsyllis augeneri (Hartmann-Schröder, 1979)

http://species-id.net/wiki/Perkinsyllis_augeneri

Pionosyllis augeneri Hartmann-Schröder, 1979: 98, figs 119–125; 1980a: 52; 1981: 32, fig. 52 (Non Hartmann-Schröder 1991: 35); San Martín and Hutchings 2006: 326. figs 57a–j, 58a–f.

Perkinsyllis augeneri: San Martín et al. 2009: 26.

Material examined. Haifa Bay, Israel: ALA-IL-7 (7 ind.) [coll. 11.10.2009].

Type locality. Boone, west Australia.

Distribution. Australia, New Zealand. Mediterranean Sea: LB. New record for the Mediterranean Sea.

Habitat. Intertidal and shallow subtidal depths, in coarse coralline sand, in muddy sand and seagrass beds (San Martín and Hutchings 2006).

Taxonomic characters. Prostomium pentagonal with 4 eyes in trapezoidal arrangement, posterior pair closer together than anterior one. Palps longer than prostomium, basally fused. Antennae cylindrical, smooth, longer than prostomium and palps. Tentacular cirri similar to antennae but slightly longer. Dorsal cirri of some anterior segments slender, longer than body width, some shorter, in midbody alternating short and long cirri, posteriorly all shorter than body width. Parapodia with 9–10 falcigers per fascicle anteriorly, 6–7 posteriorly. Shafts smooth or slightly serrated. Blades with marked dorso-ventral gradation (dorsal ones 3 times longer than ventral ones), coarsely serrated, with small subdistal tooth. After proventriculum, dorsal blades unidentate, elongated, spiniger-like, twice as long as anteriorly, ventral blades stout, with strong serration, especially basally. Dorsal simple chaeta first appearing on midbody, blunt, subdistally serrated. Ventral simple chaetae posteriorly, bidentate, equally sized teeth forming a right angle, some long spines subdistally. Paired aciculae anteriorly, single ones posteriorly, with rounded, slightly enlarged tip. Pharynx through 4 chaetigers, pharyngeal tooth located anteriorly. Proventricle through 5 chaetigers with ca. 20–22 muscle cell rows.

Remarks. The subfamilial affiliation of *Perkinsyllis augeneri* has not yet been fully resolved. In recent molecular phylogenies the species groups either within Exogoninae or as a sister group, and forms a sister clade of Eusyllinae in all analyses (Aguado and Bleidorn 2010, Aguado et al. 2007).

The morphological characters of the Mediterranean individuals agree well with the description of San Martín and Hutchings (2006) from Australia. Therefore, a detailed description of the specimens is unnecessary here. The Mediterranean specimens show slight differences from the description of the Australian ones in the length of the pharynx (6–7 chaetigers in Australian specimens vs 5 in Mediterranean ones), and the number of falcigers per bundle in anterior chaetigers (ca. 15 in Australian specimens vs ca. 10 in Mediterranean ones). These differences might however be attributed to fixation and / or individual variation.

Until now, the species had been known only from north-west Australia and New Zealand, while the record from the Carribean Sea (Hartmann-Schröder 1980a) is assumed to be a different species (San Martín et al. 2009). The present findings thus extend the distribution range of the species to the eastern Mediterranean Sea. Since there are no intermediate records of the species from the Indian Ocean or Red Sea, this disjunct distribution suggests a potential human-induced introduction of the species to the Mediterranean Sea by vectors such as ballast water or fouling fauna on the hulls of ships. However, since the polychaete fauna of the Indian Ocean, Red Sea and eastern Mediterranean Sea is understudied, the species might have a truly circumtropical distribution. This is the second record of an Australian syllid species for the Mediterranean Sea (after *Prosphaerosyllis longipapillata* (Hartmann-Schröder, 1979), recorded for the first time in 2003 in Cyprus (Çinar et al. 2003)).

Subfamily Exogoninae Langerhans, 1879

Genus *Parapionosyllis* Fauvel, 1923

Type species. *Pionosyllis gestans* Pierantoni, 1903

***Parapionosyllis elegans* (Pierantoni, 1903)**
http://species-id.net/wiki/Parapionosyllis_elegans

Pionosyllis elegans Pierantoni, 1903: 236, pl. X, fig. 2: pl. XI, fig. 27.
Parapionosyllis elegans: Fauvel 1923: 291, figs 111d–e; San Martín 1984b: 194, figs 42–43; 2003: 285, fig. 156; Çinar 1999: 127, fig. 4.40; Çinar et al. 2003: 755.

Material examined. Haifa Bay, Israel: ALA-IL-7 (11 ind.), ALA-IL-10 (45 ind.) [coll. 11.10.2009].

Type locality. Gulf of Naples (western Mediterranean Sea).

Distribution. North-east Atlantic (Iberian Peninsula). Mediterranean Sea: WB, CB, AD, AS, LB. New record for the Israeli coast.

Habitat. Until 30 m depth (San Martín 1984b), in medium to coarse sands.

Genus *Prosphaerosyllis* San Martín, 1984

Type species. *Sphaerosyllis xarifae* Hartmann-Schröder, 1960

***Prosphaerosyllis adelae* San Martín, 1984**
http://species-id.net/wiki/Prosphaerosyllis_adelae

Sphaerosyllis (Prosphaerosyllis) adelae San Martín, 1984a: 376, figs 1–4.
Prosphaerosyllis adelae: San Martín: 2003: 220, fig. 116.

Material examined. Haifa Bay, Israel: ALA-IL-7 (11 ind.) [coll. 31.5.2009]; ALA-IL-7 (6 ind.), ALA-IL-10 (2 ind.) [coll. 11.10.2009].

Type locality. Balearic Islands (western Mediterranean Sea).

Distribution. Mediterranean Sea: WB, LB. New record for the eastern Mediterranean Sea.

Habitat. Until 13 m depth, in coarse sands, among *Posidonia oceanica* rhizomes.

***Prosphaerosyllis campoyi* (San Martín, Acero, Contonente & Gomez, 1982)**
http://species-id.net/wiki/Prosphaerosyllis_campoyi

Sphaerosyllis campoyi San Martín Acero, Contonente and Gomez, 1982: 175, fig. 2; San Martín et al. 1985: 30, figs 3c–d; Çinar 1999: 146, fig. 4.50; Çinar et al. 2003: 756.
Sphaerosyllis (Prosphaerosyllis) campoyi: Núñez et al. 1992: 51.
Prosphaerosyllis campoyi: San Martín, 2003: 222, figs 117–118.

Material examined. Elounda, Crete, Greece: CELA-10a-07 (1 ind.) [coll. 27.9.2009]; CELA-10b-08 (1 ind.) [coll. 11.6.2008].

Type locality. Andalusia, Spain (western Mediterranean Sea).

Distribution. North-east Atlantic (Iberian Peninsula, Canary Islands), Mediterranean Sea: WB, AS, LB. New record for the Greek coast.

Habitat. Until 70 m depth (Çinar et al. 2003), on rocks among algae, on coralligenous substrates, in medium to coarse sands with organic material.

Remarks. The specimens agree well with the description of San Martín (2003), except for having longer dorsal papillae (15 μm), especially posteriorly.

***Prosphaerosyllis chauseyensis* Olivier, Grant, San Martín, Archambault & McKindsey, 2011**
http://species-id.net/wiki/Prosphaerosyllis_chauseyensis
Figs 4–5

Prosphaerosyllis chauseyensis Olivier et al., 2011, figs 1–3a, b.

Material examined. Haifa Bay, Israel: ALA-IL-8 (12 ind.) [coll. 31.5.2009]; ALA-IL-1 (23 ind.), ALA-IL-2 (4 ind.), ALA-IL-5 (1 ind.), ALA-IL-7 (73 ind.), ALA-IL-8 (24 ind.), ALA-IL-9 (68 ind.), ALA-IL-10 (99 ind.) [coll. 11.10.2009].

Comparative material examined. *Sphaerosyllis brevicirra* Hartmann-Schröder, 1960 (Zoological Museum Hamburg, Holotype P-17566, Ghardaqa, Red Sea: 1 individual [Label: *Sphaerosyllis brevicirra* n. sp., Ghardaqa (Rot. Meer) (Typ), 29.3.56, coll. Remane/Schulz]); *Prosphaerosyllis chauseyensis* (Muséum National d'Histoire Naturelle, Paris, Holotype MNHN POLY TYPE 1524, Chausey Islands, France: 1 individual [Label: HOLOTYPE MNHN Paris 1524, Chausey, *Prosphaerosyllis* sp. A, (5 ind. for SEM +Holotype), C1AM et C3AV]).

Type locality. Chausey Islands, Normandy (north-east Atlantic).

Distribution. North-east Atlantic (Normandy), Mediterranean Sea: LB. New record for the Mediterranean Sea.

Habitat. Until 13 m depth, in medium to very coarse sand.

Reproduction. Three specimens collected at Station ALA-IL-8 on 31 May 2009 with egg capsules attached near dorsal cirri on midbody chaetigers.

Remarks. The specimens from Israel agree well with the specimens from Normandy, however, the Mediterranean specimens differ from the Holotype in: a) Papillation pattern: each segment with one papilla between dorsal cirri and four papillae, situated dorso-laterally and ventro-laterally on each side of parapodium, most developed in posterior chaetigers, from mid-body additional papillae arranged in two very irregular lines along middle of dorsum, increasing in length towards posterior end (ca. 20 μm posteriorly). Ventrally 2 smaller (about half the size of dorsal papillae) papillae in middle of ventrum at posterior end of each segment. Specimens from Normandy have an irregular papillation pattern, but papillation is more distinct laterally, as in specimens from Israel; b) Length of anal cirri: about 125 μm, ca. 2.5–3 times length of posterior dorsal cirri (Fig. 4) (the anal cirri are broken in the holotype and the large lateral anal papillae might have been erroneously regarded as anal cirri in the original description). Specimens from both locations have anterior dorsal cirri with two papillae (dorsal and ventral) and posterior dorsal cirri only with dorsal papilla (Fig. 5, not reported by Olivier et al. 2011).

Individuals identified by Ben-Eliahu (1977a) as *Sphaerosyllis tetralix* Eliason, 1920, from the Gulf of Elat and the Mediterranean Sea might in fact belong to *P. chauseyensis*. The description and illustrations agree with many characteristics of *P. chauseyensis*, including the characteristic papilla on the dorsal cirri. However, Ben-Eliahu reports the species to have palps widely separated anteriorly (fused in *P. chauseyensis*), dor-

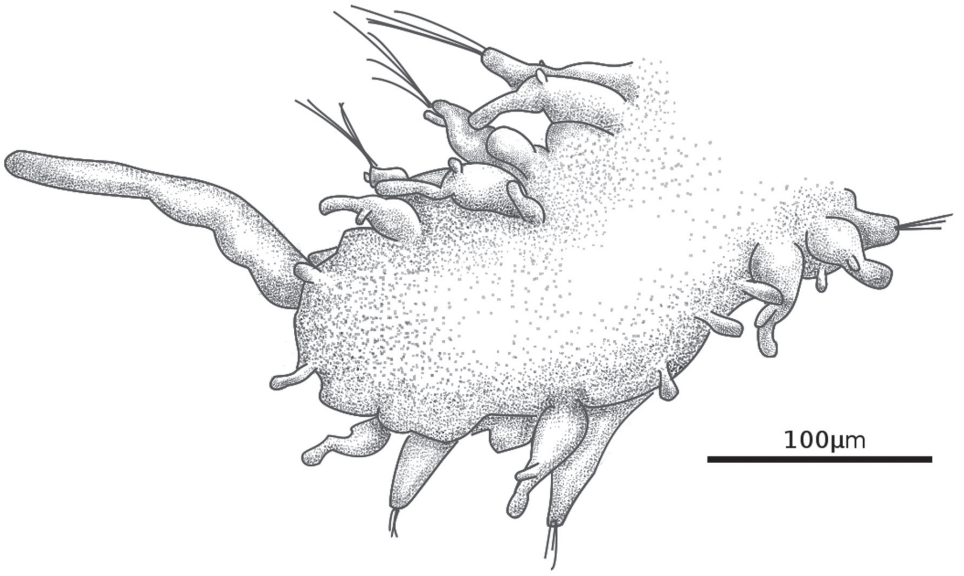


Figure 4. *Prosphaerosyllis chauseyensis*, pygidium (Israeli material).

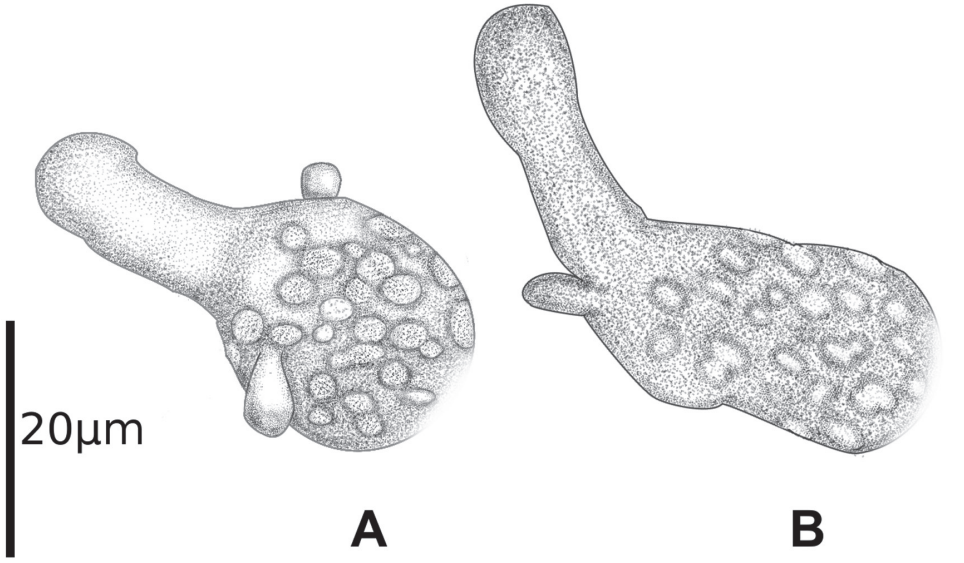


Figure 5. *Prosphaerosyllis chauseyensis*, anterior (A) and posterior (B) dorsal cirri (Israeli material).

sum with four longitudinal rows of papillae (irregular rows in *P. chauseyensis*) and the proventriculum stretching through 4 chaetigers (5 in *P. chauseyensis*). The material of the species described by Ben-Eliahu could not be examined during this study, therefore it can only tentatively be assigned to *P. chauseyensis*.

***Prosphaerosyllis longipapillata* (Hartmann-Schröder, 1979)**
http://species-id.net/wiki/Prosphaerosyllis_longipapillata

Sphaerosyllis longipapillata Hartmann-Schröder, 1979: 106, figs 148–150; 1982: 71; 1984: 23; 1985: 71; 1986: 43; 1991: 40; Çinar et al. 2003: 757, fig. 5.
Prosphaerosyllis longipapillata: San Martín 2005: 61, figs 17a–g, 18a–h.

Material examined. Haifa Bay, Israel: ALA-IL-7 (2 ind.) [coll. 11.10.2009].

Comparative Material examined. *Prosphaerosyllis longipapillata* (Hartmann-Schröder, 1979) (Department of Hydrobiology, Ege University, Izmir, Turkey, specimen reported in Çinar et al. 2003, Cyprus, Station D13: 1 individual [Label: *P. longipapillata*, Cyprus]).

Type locality. Broome, north-west Australia.

Distribution. Australia, Mediterranean Sea: LB. New record for the Israeli coast.

Habitat. Intertidal to 466 m depth (San Martín 2005), euryoceous, found on hard substrates with *Sargassum vulgare* (Çinar et al. 2003, San Martín 2005).

Remarks. The specimens from Israel agree well with the material and description of Çinar et al., 2003). However, both the material from Cyprus and Israel, as well as the description and illustrations of San Martín (2005), differ from Hartmann-Schröder's (1979) original description by the presence of dorsal papillae on the anterior chaetigers. Hartmann-Schröder (1979) reports "four long, threadlike papillae at the height of the parapodia and from chaetiger 7 onwards in pairs in a dorsal row between the parapodia". Furthermore, the Mediterranean material differs from the original description of *P. longipapillata* and from San Martín's (2005) description by having alternating rows of long and short papillae on the dorsum (Çinar et al. 2003, fig. 5). These two characteristics are reported however for *Prosphaerosyllis bilineata* (Kudenov and Harris, 1995) from California. To determine the identity of the Mediterranean material and whether *P. bilineata* and *P. longipapillata* are different species or not, careful examination of all type material is needed.

***Prosphaerosyllis marmarae* Çinar, Dagli & Açik, 2011**
http://species-id.net/wiki/Prosphaerosyllis_marmarae

Prosphaerosyllis marmarae Çinar et al. 2011: 2118, figs 2–4.

Material examined. Haifa Bay, Israel: ALA-IL-2 (3 ind.), ALA-IL-8 (12 ind.) [coll. 31.5.2009]; ALA-IL-7 (4 ind.) [coll. 11.10.2009].

Comparative material examined. *Prosphaerosyllis marmarae* (Department of Hydrobiology, Ege University, Izmir, Turkey, Paratype: 1 individual [Label: *P. marmarae*, Paratype]). *Prosphaerosyllis laubieri* (Muséum National d'Histoire Naturelle, Paris, Holotype MNHN POLY TYPE 1525, Chausey Islands, France: 1 individual [Label: HOLOTYPE MNHN Paris 1525, Chausey B1 AM12, *Prosphaerosyllis* sp. B, Holotype et SEM]).

Type locality. Erdek, Marmara Sea (eastern Mediterranean).

Distribution. Mediterranean Sea: LB, Marmara Sea. New record for the Israeli coast.

Habitat. Until 17 m depth, in muddy sand (Çinar et al. 2011), in coarse and mixed sand (this study).

Remarks. The specimens from Israel agree with the material of Çinar et al. (2011), except for the absence of eyespots (might be de-colourised due to fixation). The recently described *P. laubieri* Olivier et al. 2011 is very similar to *P. marmarae*. Both species have eyespots, strongly papillated palps, short, retractile antennae and dorsal cirri, pharynx and proventriculum each through 4 segments and short (8–10 μm) blades of falcigers. These two species differ however in the following characteristics: a) *P. laubieri* has small, scattered papillae all over the dorsum, in *P. marmarae* they are restricted to the lateral margins, near the dorsal cirri; b) cirrostyles of antennae and dorsal cirri of *P. marmarae* are much shorter (1/4 of total length) than those of *P. laubieri* (1/3 of total length) and appear as small, retracted caps; c) dorsal cirri of *P. laubieri* possess a small papilla at distal end of cirrophore (not reported by Olivier et al. 2011); d) falcigerous blades of *P. marmarae* are stouter than those of *P. laubieri* and serrated only at their bases (serrated all along cutting edge in *P. laubieri*). *P. riseri* Perkins, 1981 from Florida shares with *P. marmarae* the shape of the dorsal cirri and antennae (short and strongly retracted), however, its palps are less densely papillated. *Prosphaerosyllis* sp. A (San Martín 1991b) from Cuba has strongly papillated palps, but no cirri on chaetiger 2 and longer dorsal cirri.

Specimens from the Red Sea described by Ben-Eliahu (1977a) as *Sphaerosyllis brevicirra* Hartmann-Schröder, 1960 do not belong to this species (see Discussion section), but might in fact belong to *P. marmarae*. The morphological characteristics of her specimens agree very well wth those of *P. marmarae* (papillated palps, presence of eyespots, minute (19.5 μm), retractile cirri, falcigerous blades short (7.8 μm), proventriculum longer than proboscis (through 4 segments), no discernible dorsal papillation). Differences can be found in the cutting edge of the falcigerous blades which are smooth in the Red Sea specimens, whereas those of *P. marmarae* are serrated. However, due to the size of the blades (8 μm) this is a feature difficult to observe under an optical microscope and might have been overlooked. The material of the species described by Ben-Eliahu was not examined during this study, therefore it can only tentatively proposed to be assigned to *P. marmarae*.

Prosphaerosyllis xarifae **(Hartmann-Schröder, 1960)**

http://species-id.net/wiki/Prosphaerosyllis_xarifae

Sphaerosyllis xarifae Hartmann-Schröder, 1960: 103, figs 121–124; 1979: 103, figs 139–140; 1980b: 56; 1981: 37; 1984: 25; San Martín 1984b: 236, fig. 54; Çinar 1999: 166, fig. 4.62; Çinar et al. 2003: 760, fig. 6.

Sphaerosyllis sp.: San Martín and Alvarado 1981: 224, fig. 3.
Sphaerosyllis cf. xarifae: Campoy 1982: 279.
Sphaerosyllis (Prosphaerosyllis) xarifae: Núñez et al. 1992: 51.
Prosphaerosyllis xarifae: San Martín 2003: 225, figs 119–120; 2005: 60, figs 15a–f, 16a–f; Böggemann and Westheide 2004: 435; Fukuda et al. 2009: 1448, fig. 3.

Material examined. Haifa Bay, Israel: ALA-IL-10 (5 ind.) [coll. 11.10.2009]. Elounda, Crete, Greece: CELA-10b-08 (1 ind.) [coll. 11.6.2008]; CELA-5c-08 (1 ind.) [coll. 12.6.2008].

Type locality. Sarso, Red Sea.

Distribution. Circumtropical, Mediterranean Sea: WB, CB, AS, LB. New record for both the Israeli and Greek coasts.

Habitat. Until 40 m depth, euryoceous, among photophilic algae, in sand, mud, seagrasses, calcareous substrates (San Martín 2005).

Remarks. Specimens from Israel agree well with the description of San Martín (2003) and Hartmann-Schröder (1960) except for having more elongated dorsal papillae, especially posteriorly (20 μm, Cretan specimens: 8 μm).

Genus *Salvatoria* McIntosh, 1885

Type species. *Salvatoria kerguelensis* McIntosh, 1885

Salvatoria alvaradoi (San Martín, 1984)

http://species-id.net/wiki/Salvatoria_alvaradoi

Pseudobrania alvaradoi San Martín 1984b: 152, figs 28–29.
Salvatoria alvaradoi: San Martín 2003: 173, figs 87–88.

Material examined. Alykes, Crete, Greece: CALB-10b-08 (5 ind.), CALB-10d-08 (2 ind.) [coll. 17.6.2008]; CALB-5a-08 (2 ind.) [coll 18.6.2008]. Elounda, Crete, Greece: CELA-15a-07 (3 ind.), CELB-20e-07 (1 ind.) [coll. 26.9.2007], CELA-10a-07 (1 ind.) [coll. 27.9.2007]; CELB-1a-07 (1 ind.) [coll. 29.9.2007]; CELB-10a-08 (1 ind.), CELA-10b-08 (4 ind.), CELB-10b-08 (3 ind.), CELB-10c-08 (1 ind.), CELA-20a-08 (1 ind.), CELA-20d-08 (3 ind.) [coll. 11.6.2008]; CELA-5a-08 (8 ind.), CELA-5c-08 (18 ind.), CELA-5d-08 (1 ind.), CELB-15a-08 (1 ind.), CELB-15c-08 (9 ind.) [coll. 12.6.2008].

Type locality. Balearic Islands (western Mediterranean Sea).

Distribution. Mediterranean Sea: WB, CB, AS, Sea of Marmara (Karhan et al. 2008). New record for the Aegean Sea.

Habitat. Until 20 m depth, among algae with much sediment, among *Posidonia oceanica* rhizomes, in sediments with much organic material.

***Salvatoria euritmica* Sardá, 1984**
http://species-id.net/wiki/Salvatoria_euritmica

Pseudobrania euritmica Sardá, 1984: 10, fig. 1.
Grubeosyllis euritmica: San Martín 1991a: 718, figs 2c–d; Çinar 1999: 115, fig. 4.34; Çinar et al. 2003: 754.
Salvatoria euritmica: San Martín 2003: 169, figs 84–86; 2005: 53, figs 8a–g.
Pionosyllis yambaensis Hartmann-Schröder, 1990: 52, figs 18–22.

Material examined. Alykes, Crete, Greece: CALB-20b-08 (1 ind.) [coll. 17.6.2008]; CALB-1d-08 (4 ind.) [coll. 18.6.2008]. Elounda, Crete, Greece: CELA-15c-07 (2 ind.) [coll. 27.9.2007]; CELB-1b-07 (4 ind.), CELA-1d-07 (1 ind.) [coll. 29.9.2007]; CELA-10b-08 (1 ind.), CELA-20c-08 (1 ind.) [coll. 11.6.2008]; CELB-15d-08 (1 ind.) [coll. 12.6.2008].

Type locality. Strait of Gibraltar (western Mediterranean Sea).

Distribution. Caribbean Sea, Australia, north-east Atlantic (Iberian Peninsula, Canary Islands), Mediterranean Sea: WB, AS, LB. New record for the Greek coast.

Habitat. Until 20 m depth, on hard substrates between algae, in seagrass beds, on coralligenous substrates.

Remarks. *Pionosyllis yambaensis* was synonymized with *Salvatoria euritmica* by San Martín (2005) based on examination of type material.

***Salvatoria neapolitana* (Goodrich, 1930)**
http://species-id.net/wiki/Salvatoria_neapolitana

Pionosyllis neapolitana Goodrich, 1930: 651, figs 1–12.
Pseudobrania neapolitana San Martín 1984b: 160, figs 31–32.
Grubeosyllis neapolitana: Jiménez et al. 1994: 52 figs 1–2; Böggemann and Westheide 2004: 430.
Salvatoria neapolitana: San Martín 2003: 182, fig. 94.
Pionosyllis subterranea Hartmann-Schröder, 1956: 89 figs 6–9.
Brania subterranea: Westheide 1974a: 10, fig. 6; 1974b: 87, figs 10, 42d–f.
Grubeosyllis subterranea: Núñez et al. 1992: 45.

Material examined. Elounda, Crete, Greece: CELA-15a-07 (2 ind.), CELB-20c-07 (2 ind.) [coll. 26.9.2007]; CELB-15a-08: (5 ind.), CELB-15c-08 (1 ind.) [coll. 11.6.2008]; CELB-1d-08 (1 ind.) [coll. 12.6.2008].

Type locality. Bay of Naples, Italy (western Mediterranean Sea).

Distribution. Circumtropical, Mediterranean Sea: WB, AS (Çinar et al. 2008). New record for the Greek coast.

Habitat. Until 20 m depth, in coarse sand, among photophilic algae.

Remarks. *Pionosyllis subterranea* was synonymized with *P. neapolitana* and transferred to *Grubeosyllis* by Jiménez et al. (1994). San Martín (2003) subsequently replaced the name *Grubeosyllis* with *Salvatoria*, which has priority over the former.

***Salvatoria vieitezi* (San Martín, 1984)**
http://species-id.net/wiki/Salvatoria_vieitezi

Pseudobrania vieitezi San Martín, 1984b: 160, figs 31–32.
Grubeosyllis vieitezi: San Martin 1991a: 718, fig. 2e–f; Çinar 1999: 117, fig. 4.35; Çinar et al. 2003: 754; López and San Martín 1997: 105, fig 3.
Salvatoria vieitezi: San Martín 2003: 184, figs 95–96.

Material examined. Alykes, Crete, Greece: CALA-10d-08 (1 ind.), CALA-15c-08 (1 ind.), CALA-20c-08 (3 ind.,), CALB-20c-08 (1 ind.), CALB-20b-08 (1 ind.) [coll. 17.6.2008]; CALA-1b-08 (2 ind.), CALB-1b-08 (1 ind.), CALB-5a-08 (1 ind.) [coll. 18.6.2008]; CALB-20e-07 (1 ind.) [coll. 18.9.2007]; CALA-5c-07 (1 ind.) [coll. 19.9.2007]. Elounda, Crete, Greece: CELA-20d-07 (3 ind.) [coll. 26.9.2007]; CELA-10b-07 (1 ind.) [coll. 27.9.2007]; CELA-20c-08 (1 ind.), CELA-20d-08 (7 ind.) [coll. 11.6.2008]; CELA-5d-08 (1 ind.), CELB-15a-08 (1 ind.), CELB-1b-08 (5 ind.) [coll. 12.6.2008].

Type locality. Balearic Islands (western Mediterranean Sea).

Distribution. North-east Atlantic (Iberian Peninsula, Canary Islands), Caribbean, Mediterranean Sea: WB, CB, AS. New record for the Greek coast.

Habitat. Until 30m depth, on rocky substrates among photophilic algae, as endobiont of sponges, among *Posidonia oceanica* rhizomes.

***Salvatoria yraidae* (San Martín, 1984)**
http://species-id.net/wiki/Salvatoria_yraidae

Pseudobrania yraidae San Martín, 1984b: 156, fig. 30.
Grubeosyllis yraidae: Çinar 1999: 121, fig. 4.37.
Salvatoria yraidae: San Martín 2003: 163, figs 80–81.

Material examined. Alykes, Crete, Greece: CALB-10b-08 (1 ind.), CALB-15a-08 (1 ind.), CALB-20b-08 (3 ind.), CALB-20d-08 (1 ind.) [coll. 17.6.2008]. Elounda, Crete, Greece: CELA-15b-07 (1 ind.), CELA-15e-07 (2 ind.) [coll. 26.9.2007]; CELA-5c-07 (4 ind.) [coll. 27.9.2007]; CELA-10b-08 (3 ind.), CELB-10b-08 (8 ind.), CELB-10c-08 (1 ind.), CELA-20a-08 (1 ind.), CELA-20b-08 (1 ind.) [coll. 11.6.2008]; CELB-15a-08 (6 ind.), CELB-15c-08 (4 ind.), CELA-15d-08 (5 ind.), CELB-15d-08 (5 ind.) [coll. 12.6.2008].

Type locality. Balearic Islands (western Mediterranean Sea).

Distribution. Mediterranean Sea: WB, CB, AD, AS. New record for the Greek coast.

Habitat. Until 20 m depth, in sandy substrates, on rocks among algae.

Genus *Sphaerosyllis* Claparède, 1863

Type species. *Sphaerosyllis hystrix* Claparède, 1863

Sphaerosyllis bulbosa Southern, 1914

http://species-id.net/wiki/Sphaerosyllis_bulbosa

Sphaerosyllis bulbosa Southern, 1914: 20, plates I–II, figs 2a–g; Fauvel, 1923: 304, figs. 116h–r; Cognetti 1961: 30; Rullier 1972: 69; Campoy 1982: 276; Parapar et al. 1994: 98, fig. 4; Çinar et al. 2003: 756; San Martin 2003: 191, figs 98–99.
Sphaerosyllis (Sphaerosyllis) bulbosa: Hartmann-Schröder 1996: 175.

Material examined. Haifa Bay, Israel: ALA-IL-7 (4 ind.), ALA-IL-10 (51 ind.) [coll. 11.10.2009].

Type locality. Ireland (Atlantic Ocean).

Distribution. North-east Atlantic, Arctic Sea (Ramos et al. 2010), New Caledonia (Rullier 1972). Mediterranean Sea: WB, CB, AD, AS, LB, BS (Surugiu 2005). New record for the Israeli coast.

Habitat. Until 70 m depth, in sandy or muddy sediments, on calcareous substrates.

Remarks. The examined material differs from the description of San Martín (2003) in having papillated palps.

Sphaerosyllis glandulata Perkins, 1981

http://species-id.net/wiki/Sphaerosyllis_glandulata

Sphaerosyllis glandulata Perkins, 1981: 1123, figs 18–19; Uebelacker 1984: 33, figs 25–26; San Martín 1991a: 232; 2003: 193, fig. 100; Men et al. 1993: 31, fig. 8; Somaschini and San Martín 1994: 361, fig. 3; Çinar 1999: 152, fig. 4.53; San Martín and Bone 2001: 613.
Sphaerosyllis cf. glandulata: Ding and Westheide 2008: 131, figs. 5a–h.

Material examined. Haifa Bay, Israel: ALA-IL-7 (1 ind.) [coll. 31.5.2009]; ALA-IL-7 (47 ind.), ALA-IL-10 (19 ind.) [coll. 11.10.2009]. Elounda, Crete, Greece: CELA-15d-08 (1 ind.) [coll. 12.6.2008].

Type locality. Florida, Hutchinson Island.

Distribution. West Atlantic (Florida, Caribbean Sea), China (Ding and Westheide 2008) Mediterranean Sea: WB, AD, AS, LB (Abd-Elnaby and San Martín 2010). New record for both the Israeli and Greek coasts.

Habitat. Until 120 m depth, in calcareous habitats and fine to coarse sands, among photophilic algae.

Remarks. The specimens from Israel differ from San Martín's (2003) description in having papillated palps and a longer proventriculum (3–4 chaetigers vs 2 chaetigers in the Iberian material). Other characteristics, especially chaetal ones, agree well with former descriptions of *S. glandulata.*

Sphaerosyllis gravinae Somaschini & San Martín, 1994

http://species-id.net/wiki/Sphaerosyllis_gravinae

Sphaerosyllis gravinae Somaschini and San Martín, 1994: 358, figs 1–2; San Martín 2003: 188, fig. 97.

Material examined. Haifa Bay, Israel: ALA-IL-8 (4 ind.) [coll. 31.5.2009].

Type locality. Zannone Island, Italy (western Mediterranean Sea).

Distribution. Mediterranean Sea: WB, AD, LB. New record for the eastern Mediterranean Sea.

Habitat. Shallow subtidal depths, in medium to coarse sands, among algae.

Sphaerosyllis taylori Perkins, 1981

http://species-id.net/wiki/Sphaerosyllis_taylori

Sphaerosyllis taylori Perkins, 1981: 1140, fig. 26; Uebelacker 1984: 29, figs 21–22; San Martín 1984b: 247, fig. 58; 2003: 206, fig. 108; Russell 1991: 71; Núñez et al. 1992: 49; Parapar et al. 1994: 99; Simboura 1996: 53, fig. 6; San Martín and Bone 2001: 614; Çinar 1999: 161, fig. 4.58; Ruíz-Ramírez and Salazar-Vallejo 2001: 131, fig. 6 (115–122); Çinar et al. 2003: 759; Liñero-Arana and Díaz-Díaz 2011: 9, figs 2.1–2.5 in online material.

Material examined. Haifa Bay, Israel: ALA-IL-1 (1 ind.); ALA-IL-2 (33 ind.) [coll. 31.5.2009]; ALA-IL-7 (103 ind.), ALA-IL-10 (14 ind.) [coll. 11.10.2009].

Type locality. Florida, Hutchinson Island.

Distribution. North-east and north-west Atlantic (North Sea to Canary Islands, east coast of the U.S. to Venezuela), Pacific Ocean (Galápagos Islands) (Liñero-Arana and Díaz-Díaz 2011), Arctic Sea (Ramos et al. 2010), Mediterranean Sea: WB, CB, AD, AS, BS, LB (Abd-Elnaby and San Martín 2010). New record for the Israeli coast.

Habitat. Shallow subtidal depths, in muddy to coarse sands with organic material, on rocks among photophilic or calcareous algae, among *Posidonia oceanica* rhizomes.

***Sphaerosyllis thomasi* San Martín 1984**
http://species-id.net/wiki/Sphaerosyllis_thomasi

Sphaerosyllis thomasi San Martín, 1984b: 250, fig. 59; 2003: 199, figs 103–104; Arvanitidis 1994: 80; Çinar 1999: 163, fig. 4.60.

Material examined. Haifa Bay, Israel: ALA-IL-7 (2 ind.) [coll. 11.10.2009].

Type locality. Balearic Islands (western Mediterranean Sea).

Distribution. Mediterranean Sea: WB, CB, AD, AS, LB. New record for the Israeli coast.

Habitat. Shallow subtidal depths, in muddy to coarse sands, among *Posidonia oceanica* rhizomes.

Remarks. The examined specimens agree well with the description of San Martín (2003), especially in the chaeteal structures, but in the Israeli specimens the dorsal cirri are as long as parapodial lobes in posterior and midbody chaetigers and only slightly shorter than parapodial lobe in anterior chaetigers (dorsal cirri shorter than parapodial lobe in San Martín's (2003) description).

Subfamily Syllinae Grube, 1850

Genus *Opisthosyllis* Langerhans, 1879

Type species. *Opisthosyllis brunnea* Langerhans, 1879

***Opisthosyllis brunnea* Langerhans, 1879**
http://species-id.net/wiki/Opisthosyllis_brunnea

Opisthosyllis brunnea Langerhans, 1879: 541, fig. 7; Augener 1918: 274, text-fig. 25; Tebble 1956: 90, figs 5d–e; Day 1967: 253, figs 12.5 c–e. Hartmann-Schröder 1979: 86; 1980b: 48; 1981: 24; 1982: 58; 1991: 25, fig. 19; Fauchald 1977:20, fig. 5; San Martín 1984b: 311, figs 75–76; 2003: 330, fig. 183; Çinar 1999: 237, fig. 4.99; Amaral et al. 2005: 164, figs a–e on same page; Abd-Elnaby 2009: 15, plate 3-16, figs 3g–h.

Material examined. Elounda, Crete, Greece: CELA-1d-07 (1 ind.) [coll. 29.9.2007], CELA-5d-08 (1 ind.) [coll. 12.6.2008].

Type locality. Madeira (Atlantic Ocean).

Distribution. Circumtropical. Mediterranean Sea: WB, CB, AS, LB. New record for the Greek coast.

Habitat. Intertidal to shallow subtidal, on hard substrates (vermetid reefs, among photophilic algae), endobiont of sponges.

Genus *Syllis* Lamarck, 1818

Type species. *Syllis monilaris* Lamarck, 1818

***Syllis alternata* Moore, 1908**
http://species-id.net/wiki/Syllis_alternata

Syllis alternata Moore, 1908: 323; 1909: 321; Çinar 1999: 246, fig. 4.102; Çinar and Gambi 2005: 754; Çinar and Ergen 2003: 777.
Typosyllis alternata: Kudenov and Harris 1995: 83, fig. 1.32; Licher 2000: 253, figs 17p, 106; Imajima 2003: 163.

Material examined. Alykes, Crete, Greece: CALB-15c-07 (1 ind.) [coll. 18.9.2007]; CALB-1a-07 (1 ind.) [coll. 19.9.2007]; CALA-10d-08 (2 ind.), CALA-15d-08 (1 ind.), CALB-20b-08 (1 ind.), CALA-20b-08 (2 ind.), CALA-20c-08 (5 ind.), CALB-20d-08 (6 ind.) [coll. 17.6.2008]. Elounda, Crete, Greece: CELB-20c-07 (1 ind.) [coll. 26.9.2007]; CELB-1a-07 (4 ind.) [coll. 29.9.2007]; CELA-10b-08 (1 ind.), CELB-10b-08 (1 ind.), CELA-10c-08 (1 ind.), CELB-10c-08 (1 ind.), CELA-10d-08 (2 ind.), [coll. 11.6.2008]; CELB-1a-08 (1 ind.), CELA-5b-08 (1 ind.), CELA-5d-08 (2 ind.), CELB-15c-08 (5 ind.) [coll. 12.6.2008].

Type locality. Alaska (Pacific Ocean).

Distribution. East Pacific (Alaska to Panama), west Atlantic (North Carolina to Cuba) (Capa et al. 2001), Japan (Imajima 2003), Indonesia (Aguado et al. 2008), Mediterranean: WB, CB, AS, LB. New record for the Greek coast.

Habitat. Until 2500 m depth (Moore 1909), among *Posidonia oceanica* rhizomes, calcareous algae, corals and photophilic algae (San Martín 2003), in sandy and muddy sediments (Moore 1909).

***Syllis compacta* Gravier, 1900**
http://species-id.net/wiki/Syllis_compacta

Syllis (Typosyllis) compacta Gravier, 1900: 165, pl. 9, fig. 11, text-figs 35–38.
Syllis compacta: López et al. 1996: 110, fig 3; Çinar 1999: 263, fig. 4.113; San Martín 2003: 433, figs 238–239.
Syllis golfonovensis: San Martín 1984b: 395, fig. 104 (Non *Syllis golfonovensis* Hartmann-Schröder, 1962.

Material examined. Alykes, Crete, Greece: CALB-1e-07 (1 ind.), CALA-5e-07 (1 ind.) [coll. 19.9.2007]; CALA-15c-08 (1 ind.), CALA-20b-08 (1 ind.), CALB-20b-08 (1 ind.), CALA-20c-08 (1 ind.) [coll. 17.6.2008]. Elounda, Crete, Greece: CELA-15b-07 (1 ind.), CELA-15e-07 (3 ind.), CELB-20a-07 (2 ind.) [coll. 26.9.2007];

CELA-10a-07 (1 ind.), CELA-10d-07 (1 ind.) [coll. 27.9.2007]; CELA-5c-07 (1 ind.), CELB-5d-07 (1 ind.) [coll. 29.9.2007]; CELA-5b-08 (1 ind.), CELA-5d-08 (2 ind.), CELB-15d-08 (3 ind.) [coll. 12.6.2008].

Type locality. Red Sea.

Distribution. Red Sea. Mediterranean Sea: WB, CB, AD, AS. New record for the Greek coast.

Habitat. Shallow subtidal depths, on biogenic calcareous substrates, among photophilic algae and *Posidonia oceanica* rhizomes.

Remarks. The species is regarded by many authors (e.g. Augener 1913, Fauvel 1919, Licher 2000) as a synonym of *Syllis variegata* Grube, 1860. Recent works (e.g. San Martín 2003, Çinar 2005) however, regard the two species as distinct, which is also supported by molecular analyses (Aguado et al. 2007).

Syllis cruzi Núñez & San Martín, 1991

http://species-id.net/wiki/Syllis_cruzi

Syllis cruzi Núñez and San Martín, 1991: 238, figs 2a–j; Çinar and Ergen 2003: 780, fig. 2.
Typosyllis cruzi: Licher 2000: 169, fig. 75.

Material examined. Alykes, Crete, Greece: CALB-20d-08 (1 ind.) [coll. 17.6.2008]. Elounda, Crete, Greece: CELB-10a-08 (1 ind.) [coll. 11.6.2008].

Type locality. Canary Islands (Atlantic Ocean).

Distribution. North-east Atlantic (Canary Islands), Mediterranean Sea: WB, CB, AD, AS, LB. New record for the Aegean Sea.

Habitat. Until 115 m depth, on coralligenous substrates, among photophilic algae, endobiont of sponges.

Syllis gerundensis (Alós & Campoy, 1981)

http://species-id.net/wiki/Syllis_gerundensis

Typosyllis gerundensis Alós and Campoy, 1981*:* 21, figs 1–3; Campoy 1982: 446, figs 55–56; Licher 2000: 171, fig. 77.
Syllis gerundensis: Çinar and Ergen 2003: 783; San Martín 2003: 419, figs 230–231.

Material examined. Alykes, Crete, Greece: CALA-20b-08 (1 ind.), CALB-20b-08 (1 ind.) [coll. 17.6.2008]; CALA-5d-08 (2 ind.) [[coll. 18.6.2008]. Elounda, Crete, Greece: CELB-1e-07 (1 ind.) [coll. 29.9.2007]; CELB-1d-08 (1 ind.), CELA-5d-08 (3 ind.) [coll. 12.6.2008].

Type locality. Columbretes Islands, Spain (western Mediterranean Sea).

Distribution. Mediterranean Sea: WB, CB, AD, AS, LB. New record for the Aegean Sea.

Habitat. Shallow subtidal depths, on calcareous grounds, sandy bottoms, among *Posidonia oceanica* rhizomes and photophilic algae, endobiont of sponges.

***Syllis jorgei* San Martín & López, 2000**
http://species-id.net/wiki/Syllis_jorgei

Syllis jorgei San Martín and López, 2000: 430, figs 4–6; San Martín 2003: 382, figs 208–210; Çinar and Ergen 2003: 785.
Typosyllis lutea: Campoy 1982: 428.
Syllis lutea: San Martín 1984b: 370, figs 94–95; Arvanitidis 1994: 101 (Non *Syllis lutea* Hartmann-Schröder, 1960).

Material examined. Haifa Bay, Israel: ALA-IL-7 (3 ind.) [coll. 11.10.2009]. Alykes, Crete, Greece: CALA-20c-07 (1 ind.) [coll. 18.9.2007], CALB-1a-08 (1 ind.) [coll. 18.6.2008]. Elounda, Crete, Greece: CELB-1a-08 (1 ind.), CELA-1c-08 (1 ind.), CELB-5d-08 (1 ind.) [coll. 12.6.2008].

Type locality. Columbretes Islands, Spain (western Mediterranean Sea).

Distribution. East Atlantic (Canary Islands), Mediterranean Sea: WB, CB, AD, AS, LB. New record for the Israeli coast.

Habitat. Until 145 m depth (Çinar and Ergen 2003), on biogenic calcareous structures, among *Posidonia oceanica* rhizomes and photophilic algae.

***Syllis pulvinata* (Langerhans, 1881)**
http://species-id.net/wiki/Syllis_pulvinata

Typosyllis pulvinata Langerhans, 1881: 97, 104; Licher 2000: 158, fig. 70.
Syllis pulvinata: Çinar and Ergen 2003: 787; San Martín 2003: 372, figs 202–204.
Syllis (Typosyllis) truncata mediterranea Ben-Eliahu, 1977a: 10, fig. 2.
Syllis mediterranea: San Martín, 1984b; 209, fig. 8.

Material examined. Elounda, Crete, Greece: CELA-1b-08 (1 ind.), CELA-5c-08 (2 ind.), CELB-15d-08 (1 ind.) [coll. 12.6.2008].

Type locality. Canary Islands (Atlantic Ocean).

Distribution. North-east Atlantic (Cantabrian Sea to Canary Islands), Red Sea, Mediterranean: WB, CB, AD, AS, LB. New record for the Aegean Sea.

Habitat. Shallow subtidal depths, on calcareous substrates (vermetid reefs), among photophilic algae, endobiont of sponges.

***Syllis tyrrhena* (Licher & Kuper, 1998)**
http://species-id.net/wiki/Syllis_tyrrhena

Typosyllis tyrrhena Licher and Kuper, 1998: 228, figs 1–4; Licher 2000: 140, figs 2, 14–16, 62–63; Kuper 2001: 58, figs 1a–b, 20–24. Amaral et al. 2005: 162, figs a–f on same page.
Syllis tyrrhena: San Martín 2003: 379, fig. 207.

Material examined. Elounda, Crete, Greece: CELB-10b-08 (1 ind.) [coll. 11.6.2008].

Type locality. Island of Elba, Italy (western Mediterranean Sea).

Distribution. Brazil (Amaral et al. 2005), Mediterranean Sea: WB, AS. New record for the eastern Mediterranean Sea.

Habitat. Until 13 m depth, in sandy substrates of mixed grain sizes (Licher and Kuper 1998), on rocks among algae (this study).

***Syllis westheidei* San Martín, 1984**
http://species-id.net/wiki/Syllis_westheidei

Syllis westheidei San Martín, 1984b: 403, figs 108–109; 2003: 436, figs 240–241; Çinar 1999: 310, fig. 4.141.
Typosyllis westheidei: Licher 2000: 111, fig. 51; Böggemann and Westheide 2004: 418.
Typosyllis variegata: Westheide 1974a: 51, figs 21–22. (Non *Syllis variegata* Grube, 1860).

Material examined. Alykes, Crete, Greece: CALB-15d-08 (1 ind.) [coll. 17.6.2008].

Type locality. Balearic Islands (western Mediterranean Sea).

Distribution. Pacific Ocean (Galápagos Islands), Red Sea, Mediterranean: WB, CB, AD, AS. New record for the Greek coast.

Habitat. Shallow subtidal depths, on hard substrates, among photophilic algae, in *Posidonia oceanica* rhizomes and vermetid reefs.

Genus *Trypanosyllis* Claparède 1864

Type species. *Syllis zebra* Grube, 1860

***Trypanosyllis coeliaca* Claparède, 1868**
http://species-id.net/wiki/Trypanosyllis_coeliaca

Trypanosyllis coeliaca Claparède 1868: 513, pl. 13, fig. 3; Fauvel 1923: 270, figs 101f–h; Cognetti 1957: 27, fig. 5a; 1961: 296, Hartmann-Schröder 1979: 78; Perkins 1981: 1155, figs 33–34; Campoy 1982: 354; Uebelacker 1984: 93, fig. 88; San

Martín 1984b: 274, fig. 63; 2003: 308, figs 169–170; Arvanitidis 1994: 109; Çinar 1999: 316, fig. 4.144; Çinar and Ergen 2003: 789.

Pseudosyllis brevipennis Grube, 1863: 44, pl. 4, fig. 5.

Material examined. Haifa Bay, Israel, eastern Mediterranean Sea, Station ALA-IL-7 (1 ind.) [coll. 11.10.2009]. Alykes, Crete, Greece: CALA-10b-08 (1 ind.), CALB-10c-08 (1 ind.) [coll. 17.6.2008]; CALA-5a-08 (1 ind.), CALB-1d-08 (2 ind.), CALB-5a-08 (1 ind.) [coll. 18.6.2008]. Elounda, Crete, Greece: CELA-15b-07 (1 ind.), CELA-15c-07 (1 ind.) [coll. 26.9.2007]; CELB-5c-07 (1 ind.), CELA-10a-07 (1 ind.), CELB-10c-07 (1 ind.) [coll. 27.9.2007]; CELB-1a-07 (2 ind.), CELB-1e-07 (1 ind.) [coll. 29.9.2007]; CELB-10b-08 (1 ind.), CELA-15a-08 (1 ind.) [coll. 17.6.2008]; CELB-1b-08 (1 ind.), CELA-5b-08 (1 ind.), CELA-5c-08 (1 ind.), CELB-5c-08 (1 ind.), CELA-5d-08 (2 ind.) [coll. 18.6.2008].

Type locality. Gulf of Naples (western Mediterranean Sea).

Distribution. Circumtropical. Mediterranean Sea: WB, CB, AD, AS, LB. New record for the Israeli coast.

Habitat. From infralitoral depths to 760 m, on hard substrates, among algae, corals, hydrozoans, sponges and *Posidonia oceanica* rhizomes, in vermetid reefs, in coarse sand.

Remarks. Specimens from Greece have a faint or no visible trepan. Individuals without trepan but otherwise identical to *T. coeliaca* have in the past been identified as *Pseudosyllis brevipennis* Grube, 1863, but according to San Martín (2003) the absence of the trepan can be attributed to a number of reasons, including loss, and *P. brevipennis* is regarded as a synonym of *T. coeliaca.*

Discussion

The present study yielded a number of species reported for the first time in the respective areas, and a high number of the new additions belong to the subfamily Exogoninae (Fig. 3). This could be explained by the fact that the small-sized individuals of this subfamily might have been overlooked in earlier works on the syllid fauna of the area which report only very few or no Exogoninae species at all (e.g. Fauvel 1957, Tebble 1959, Bellan 1964, Ergen 1976). The Exogoninae genus *Prosphaerosyllis*, which has recently been raised from subgeneric to generic level by San Martín (2005), has a difficult and confused taxonomy and several species have recently been described or transferred to the genus (Olivier et al. 2011, Çinar et al. 2011). Currently, 31 species of the genus are considered valid (including an unnamed one, see San Martín 2003), of which 11 have so far been reported to occur in the Mediterranean Sea (Table 3). However, several of the reported species in the area do in fact belong to other species, the identity of which can only be determined through thorough examination of the material in question. The presence of the Red Sea species *P. brevicirra* (Hartmann-Schröder, 1960) in the Mediterranean belongs to these doubtful records. Records of *Sphaerosyllis brevicirra* from the western Mediterranean

Table 3. Reported distribution records of *Prosphaerosyllis* species in the Mediterranean. †= doubtful record, identity unknown. ‡= doubtful record, probably *Prosphaerosyllis* sp. [unnamed, San Martín 2003], §= doubtful record, probably *P. marmarae.* References: 1= this study, 2= San Martín 1984, 3= San Martín 2003, 4= Gambi et al. 1995, 5= Alós 1989, 6= Somaschini et al. 1994, 7= Lanera et al. 1990, 8= Zenetos et al. 1997, 9= Simboura 1996, 10= Çinar 1999, 11= San Martín et al. 1982, 12= Çinar and Ergen 2002, 13= Çinar et al. 2003, 14= Somaschini and San Martín 1994, 15= Çinar et al. 2011, 16= Katzmann, 1983, 17= Ben-Eliahu 1977a. Literature-based works (e.g. Musco and Giangrande 2005, Simboura and Nicolaidou 2001) are not included to avoid repetition of records.

Species	Type locality	WB	AD	CB	AS	LB
P. adelae San Martín, 1984	Balearic Islands, Spain, west Mediterranean	2, 3				1
P. brandhorsti (Hartmann-Schröder, 1965)	Isla Mocha, Chile, Pacific Ocean	4†				
P. brevicirra (Hartmann-Schröder, 1960)	Ghardaqa, Egypt, Red Sea	4‡, 5‡, 6‡, 7‡		8‡	9‡, 10‡	
P. campoyi (San Martín et al., 1982)	Andalusia, Spain, western Mediterranean	3, 11			1, 12	13
P. chauseyensis Olivier et al., 2011	Normandy, France, north-east Atlantic					1
P. giandoi (Somaschini and San Martín, 1994)	Tyrrenian Sea, Italy, western Mediterranean	14				
P. longipapillata (Hartmann-Schröder, 1979)	Broome, north-west Australia					3, 1
P. marmarae Çinar et al., 2011	Marmara Sea, Turkey, eastern Mediterranean				15	1
Prosphaerosyllis sp. [unnamed, San Martín 2003]	Cabo de Creus, Spain, western Mediterranean	3				
P. tetralix (Eliason, 1920)	Öresund, Sweden	3	16	8		17§
P. xarifae (Hartmann-Schröder, 1960)	Sarso, Egypt, Red Sea	3			12	

Sea by Alós (1989) and from the Aegean Sea (Simboura 1996, Çinar 1999) belong to an undescribed *Prosphaerosyllis* species (San Martín 2003). These differ from *P. brevicirra* by the absence of dorsal cirri on chaetiger 2 (reported as present in Alós' (1989) description but in fact absent (San Martín 2003)), the absence of the conspicious papilla on the dorsal cirrus and by thicker aciculae. Hartmann-Schröder (1960) does not mention the papilla on the dorsal cirrus in her description of the species (only visible in the illustrations, but confirmed through examination of type material); instead she focuses on the reduced length of the dorsal cirri as a character to distinguish the species from its congeners. This fact might have lead to confusion of *P. brevicirra* with other species possessing short dorsal cirri. Two other reports of the species from adjacent areas (Red Sea, Atlantic) likewise do probably not belong to *P. brevicirra:* Ben-Eliahu's (1977a) redescription of the species based on material from the Gulf of Elat (Red Sea) differs in several aspects from Hartmann-Schröder's (1960) descrip-

tion and from the type material. In particular, Ben-Eliahu does not mention or illustrate the papilla on the dorsal cirrus, her specimens have four eyes and one anterior pair of eyespots (eyespots, a character considered as invariable within species (Riser 1991), are absent in *P. brevicirra*) and the proventriculum occupies 4 chaetigers (3 in *P. brevicirra*). According to the description and illustrations, the species might in fact belong to *P. marmarae* (see remarks for this species above). The record of *Sphaerosyllis brevicirra* from the Spanish Atlantic coast (Parapar et al. 1994), though described as similar to Alós' (1989) specimens, differs in fact from these by the presence of dorsal cirri on chaetiger 2 and much longer dorsal cirri. It also differs from *P. brevicirra* in having falcigers with serrated blades in anterior chaetigers, no papilla on the dorsal cirrus and much longer dorsal cirri anteriorly (140 µm vs ca. 20 µm in *P. brevicirra*). The species *P. brandhorsti* (Hartmann-Schröder, 1965) has been recorded in Italy by Gambi et al. (1995). However, the only other records of the species apart from its type locality (Isla Mocha, Chile) are from the northern Pacific (Banse 1972) and belong possibly to possibly *P. ranunculus* (Kudenov and Harris, 1995). The presence of *P. brandhorsti* in the Mediterranean Sea has thus to be considered as doubtful. An identification key to the currently valid Mediterranean species of *Prosphaerosyllis* can be found below.

Key to the Mediterranean species of *Prosphaerosyllis*

1 Dorsal cirri on chaetiger 2 present **2**
– Dorsal cirri on chaetiger 2 absent... ***Prosphaerosyllis* sp. [San Martín 2003]**
2 Dorsal cirri and antennae with conspicuous papilla..... ***P. chauseyensis***
– Dorsal cirri and antennae without conspicuous papilla **3**
3 Papillae on dorsum arranged in regular longitudinal rows **4**
– Papillae on dorsum arranged irregularly **5**
4 Pharynx through 4–5 chaetigers, pharyngeal tooth on midline of pharynx ... . ***P. longipapillata***
– Pharynx through 3 chaetigers, pharyngeal tooth in anterior third of pharynx..... ..... ***P. tetralix***
5 Dorsal cirri papilliform **6**
– Dorsal cirri with bulbous cirrophore and rounded or elongated cirrostyle ... **7**
6 Prostomium retracted under posterior chaetigers. Antennae and dorsal cirri distally truncated. Aciculae subdistally with a crown of spines ..... ***P. adelae***
– Prostomium not retracted under posterior chaetigers. Antennae and dorsal cirri distally rounded. Aciculae with subdistal swelling..... ***P. giandoi***
7 Palps densely papillated. Dorsal papillation inconspicious..... ***P. marmarae***
– Palps with few or no papillae. Dorsum with distinct papillation **8**
8 Blades of falcigers in midbody with strong serration ..... ***P. campoyi***
– Blades of falcigers finely serrated ..... ***P. xarifae***

Acknowledgements

The authors acknowledge assistance from the following colleagues: Dr Melih Ertan Çinar (Ege University, Turkey), Dr Angelika Brandt (Zoological Museum, University of Hamburg, Germany) and Dr Tarik Meziane (Muséum National d'Histoire Naturelle) for loans of comparative material of *Sphaerosyllis longipapillata, Prosphaerosyllis marmarae, Sphaerosyllis brevicirra, Prosphaerosyllis chauseyensis* and *Prosphaerosyllis laubieri,* Dr Guillermo San Martín (Universidad Autónoma de Madrid) for the provision of some of the literature resources and Dr Lyubomir Penev (Pensoft Publishers, Bulgaria) for support with the publication of data as a Darwin Core Archive. Financial support was provided by ViBRANT (FP7, EU, Contract no. RI-261532). This study forms part of the core biodiversity project of the Institute of Marine Biology and Genetics (HCMR).

References

Abd-Elnaby FA (2009) New Records of Polychaetes from the South Part of Suez Canal, Egypt. World Journal of Fish and Marine Sciences 1: 7–19. http://idosi.org/wjfms/wjfms1%281%2909/2.pdf

Abd-Elnaby FA, San Martín G (2010) Eusyllinae, Anoplosyllinae, and Exogoninae (Polychaeta: Syllidae) for the Mediterranean Coasts of Egypt, Together the Description of One New Species. Life Science Journal 7: 132–139. http://www.sciencepub.net/life/life0704/20_4043life0704_132_139.pdf

Abd-Elnaby FA, San Martín G (2011) Syllinae (Syllidae: polychaeta) from the Mediterranean coast of Egypt with the description of two new species. Mediterranean Marine Science 12: 43–52. http://www.medit-mar-sc.net/files/201101/12-104153421_MMS_v12n1_ABD_ELNABY.pdf

Aguado MT, Bleidorn C (2010) Conflicting signal within a single gene confounds syllid phylogeny (Syllidae, Annelida). Molecular Phylogenetics and Evolution 55: 1128–1138. doi: 10.1016/j.ympev.2010.01.012

Aguado MT, Nygren A, Siddall ME (2007) Phylogeny of Syllidae (Polychaeta) based on combined molecular analysis of nuclear and mitochondrial genes. Cladistics 23: 552–564. doi: 10.1111/1096-0031.2007.00163.x

Aguado MT, San Martín G (2007) Syllidae (Polychaeta) from Lebanon with two new reports for the Mediterranean Sea. Cahiers de Biologie Marine 48: 207–224. http://www.sb-roscoff.fr/cbm/cbm.htm?execution=e1s2&_eventId=viewarticledetails&articleId=2476

Aguado MT, San Martín G (2009) Phylogeny of Syllidae (Polychaeta) based on morphological data. Zoologica Scripta 38: 379–402. doi: 10.1111/j.1463-6409.2008.00380.x

Aguado MT, San Martín G, ten Hove HA (2008) Syllidae (Annelida: Polychaeta) from Indonesia collected by the Siboga (1899–1900) and Snellius II (1984) expeditions. Zootaxa 1673: 1–48.

Alós C (1989) Adiciones a la fauna de anelidos poliquetos de la peninsula iberica: familia Syllidae. Cahiers de Biologie Marine 30: 329–338.

Alós C, Campoy A (1981) *Typosyllis gerundensis* n. sp.: nuevo Syllidae (Annelida, Polychaeta) del Mediterraneo. Publicaciones del Departamento de Zoologia, Universidad de Barcelona, Facultad de Biologia 7: 21–27.

Amaral ACZ, Rizzo AE, Arruda EP (2005) Manual de Identificação dos Invertebrados Marinhos da Região Sudeste-Sul do Brasil, volume 1. São Paolo: Edusp, 288 pp. http://books.google.de/books?id=SUOjxxkSIysC

Antoniadou C, Nicolaidou A, Chintiroglou C (2004) Polychaetes associated with the sciaphilic alga community in the northern Aegean Sea: spatial and temporal variability. Helgoland Marine Research 58: 168–182. doi: 10.1007/s10152-004-0182-6

Arvanitidis C (1994) Systematic and bionomic study of the macrobenthic polychaetes (Annelida, Polychaeta) of the North Aegean. PhD Thesis, Thessaloniki, Greece: Aristotle University of Thessaloniki. http://phdtheses.ekt.gr/eadd/handle/10442/5045

Arvanitidis C (2000) Polychaete fauna of the Aegean Sea: inventory and new information. Bulletin of Marine Science 66: 73–96. http://www.ingentaconnect.com/content/umrsmas/bullmar/2000/00000066/00000001/art00010

Arvanitidis C, Bellan G, Drakopoulos P, Valavanis VD, Dounas C, Koukouras A, Eleftheriou A (2002) Seascape biodiversity patterns along the Mediterranean and the Black Sea: lessons from the biogeography of benthic polychaetes. Marine Ecology Progress Series 244: 139–152. doi: 10.3354/meps244139

Augener H (1913) Polychaeta I. Errantia. Die Fauna Südwest-Australiens. Ergebnisse der Hamburger südwest-australischen Forschungsreise 1905 4: 65–304. http://www.biodiversitylibrary.org/item/31515

Augener H (1918) Polychaeta. Beiträge zur Kenntnis der Meeresfauna Westafrikas 2: 67–625. http://www.biodiversitylibrary.org/ia/beitrgezurkenn02mich

Banse K (1971) A new species, and additions to the descriptions of six other species of *Syllides* Örsted (Syllidae: Polychaeta). Journal of the Fisheries Research Board of Canada 28: 1469–1481. doi: 10.1139/f71-226

Banse K (1972) On some species of Phyllodocidae, Syllidae, Nephtyidae, Goniadidae, Apistobranchidae, and Spionidae (Polychaeta) from the Northeast Pacific Ocean. Pacific Science 26: 191–222. http://hdl.handle.net/10125/400

Bellan G (1964) Campagne de la Calypso: Méditerranée Nord-Orientale. 6. Annélides Polychètes. Annales de l'Institut océanographique 41: 271–288.

Ben-Eliahu MN (1976a) Errant polychaete cryptofauna (excluding Syllidae and Nereidae) from rims of similar intertidal vermetid reefs on the Mediterranean coast of Israel and in the Gulf of Elat. Israel Journal of Zoology 25: 156–177.

Ben-Eliahu MN (1976b) Polychaete cryptofauna from rims of similar intertidal vermetid reefs on the Mediterranean coast of Israel and in the Gulf of Elat: Sedentaria. Israel Journal of Zoology 25: 121–156.

Ben-Eliahu MN (1977a) Polychaete cryptofauna from rims of similar intertidal vermetid reefs on the Mediterranean coast of Israel and in the Gulf of Elat: Exogoninae and Autolytinae (Polychaeta Errantia: Syllidae). Israel Journal of Zoology 26: 59–99.

Ben-Eliahu MN (1977b) Polychaete cryptofauna from rims of similar intertidal vermetid reefs on the Mediterranean coast of Israel and in the Gulf of Elat: Syllinae and Eusyllinae (Polychaeta Errantia: Syllidae). Israel Journal of Zoology 26: 1–58.

Ben-Eliahu MN, Fiege D (1995) Polychaeta from the continental shelf and slope of Israel collected by "Meteor" 5 Expedition (1987). Senckenbergiana maritima 25: 85–105.

Ben-Eliahu MN, Golani D (1990) Polychaetes (Annelida) in the gut contents of goatfishes (Mullidae), with new polychaete records for the Mediterranean coast of Israel and the Gulf of Elat (Red Sea). Marine Ecology 11: 193–205. doi: 10.1111/j.1439-0485.1990.tb00239.x

Blagoderov V, Brake I, Georgiev T, Penev L, Roberts D, Ryrcroft S, Scott B, Agosti D, Catapano T, Smith VS (2010) Streamlining taxonomic publication: a working example with Scratchpads and ZooKeys. ZooKeys 50: 17–28. doi: 10.3897/zookeys.50.539

Böggemann M, Westheide W (2004) Interstitial Syllidae (Annelida: Polychaeta) from Mahe (Seychelles). Journal of Natural History 38: 403–446.

Campoy A (1982) Fauna de Espana de Anelidos Poliquetos de la Peninsula Iberica. Publicaciones de Biologia de la Universidad de Navarra 7: 1–780. http://hdl.handle.net/10171/11773, http://hdl.handle.net/10171/11777

Capa M, San Martín G, López E (2001) Syllinae (Syllidae, Polychaeta) del Parque Nacional de Coiba, Panamá. Revista de Biología Tropical 49: 1–18. http://www.ots.ac.cr/tropiweb/attachments/volumes/vol49-1/15-Capa-Syllinae-101-114.pdf

Çinar ME (1999) Türkiye'nin Ege Denizi sahillerinde dağılım gösteren syllidae (Polychaeta-Annelida) türlerinin taksonomisi ve ekolojisi. Phd Thesis, Izmir, Turkey: Ege University. http://www.belgeler.com/blg/26fl/turkiye-nin-ege-denizi-sahillerinde-dagilim-gosteren-syllidae-polychaeta-annelido-turlerinin-taksonomisi-ve-ekolojisi-taxonomy-and-ecology-of-syllidae-polychaeta-annelido-distrubuted-along-the-turkish-aegeon-coast

Çinar ME (2005) *Syllis ergeni*: a new species of Syllidae (Annelida: Polychaeta) from Izmir Bay (Aegean Sea, eastern Mediterranean Sea). Zootaxa 1036: 43–53.

Çinar ME, Dagli E, Açik S (2011) Annelids (Polychaeta and Oligochaeta) from the Sea of Marmara, with descriptions of five new species. Journal of Natural History 45: 2105–2143. doi: 10.1080/00222933.2011.582966

Çinar ME, Ergen Z (2002) Faunistic analysis of Syllidae (Annelida: Polychaeta) from the Aegean Sea. Cahiers de Biologie Marine 43: 171–178. http://www.sb-roscoff.fr/cbm/cbm.htm?execution=e1s5&_eventId=viewarticledetails&articleId=3650

Çinar ME, Ergen Z (2003) Eusyllinae and Syllinae (Annelida: Polychaeta) from northern Cyprus (Eastern Mediterranean Sea) with a checklist of species reported from the Levant Sea. Bulletin of Marine Science 72: 769–793. http://www.ingentaconnect.com/content/umrsmas/bullmar/2003/00000072/00000003/art00010

Çinar ME, Ergen Z, Benli HA (2003) Autolytinae and Exogoninae (Polychaeta: Syllidae) from northern Cyprus (Eastern Mediterranean sea) with a checklist of species reported from the Levant sea. Bulletin of Marine Science 72: 741–767. http://www.ingentaconnect.com/content/umrsmas/bullmar/2003/00000072/00000003/art00009

Çinar ME, Gambi MC (2005) Cognetti's syllid collection (Polychaeta: Syllidae) deposited at the Museum of the Stazione Zoologica "Anton Dohrn" (Naples, Italy), with descrip-

tions of two new species of *Autolytus*. Journal of Natural History 39: 725–762. doi: 10.1080/00222930400001327

Çinar ME, Katağan T, Koçak F, Öztürk B, Ergen Z, Kocatas A, Önen M, Kirkim F, Bakir K, Kurt Şahin G, Dağli E, Açik S, Doğan A, Özcan T (2008) Faunal assemblages of the mussel *Mytilus galloprovincialis* in and around Alsancak Harbour (Izmir Bay, eastern Mediterranean) with special emphasis on alien species. Journal of Marine Systems 71: 1–17. doi: 10.1016/j.jmarsys.2007.05.004

Claparède É (1863) Beobachtungen über Anatomie und Entwicklungsgeschichte wirbelloser Thiere an der Küste von Normandie angestellt. Wilhelm Engelmann Verlag, Leipzig, 172 pp. http://biodiversitylibrary.org/item/40310.

Claparède É (1864) Glanures Zootomiques parmi les Annélides de Port-Vendres (Pyrenées Orientales). Mémoires de la Société de physique et d'histoire naturelle de Genève 17: 463–600. http://www.biodiversitylibrary.org/ia/glanureszootomiq00clapare

Claparède É (1868) Les annélides chétopodes du Golfe de Naples. Mémoires de la Société de physique et d'histoire naturelle de Genève 20: 313–584. http://biodiversitylibrary.org/item/18576

Cognetti G (1957) I Sillidi del Golfo di Napoli. Pubblicazioni della Stazione Zoologica di Napoli 30: 1–100.

Cognetti G (1961) Les Syllidiens des côtes de Bretagne. Cahiers de Biologie Marine 2: 291–312.

De Matos Nogueira JM, Amaral ACZ, San Martín G (2001) Description of five new species of Exogoninae Rioja, 1925 (Polychaeta: Syllidae) associated with the stony coral *Mussismilia hispida* (Verrill, 1868) in Sao Paulo State, Brazil. Journal of Natural History 35: 1773–1794. doi: 10.1080/00222930152667096

Day JH (1967) A monograph on the Polychaeta of Southern Africa. Part 1. Errantia. Trustees of the British Museum (Natural History), London, 498 pp. http://www.biodiversitylibrary.org/ia/monographonpolyc01dayj

Ding Z-hu, Westheide W (2008) Interstitial Exogoninae from the Chinese coast (Polychaeta, Syllidae). Senckenbergiana Biologica 88: 125–159.

Eliason A (1920) Polychaeta. Biologisch-faunistische Untersuchungen aus dem Öresund. Acta Universitatis Lundensis 16: 1–103. http://www.biodiversitylibrary.org/ia/biologischfaunis00elia

Ergen Z (1976) Investigations on the taxonomy and ecology of Polychaeta from Izmir Bay and its adjacent area. Scientific Reports of the Faculty of Science Ege University 209: 1–66.

Fassari G (1982) Anellidi Policheti del mar Egeo. Animalia 9: 109–121.

Fauchald K (1977) Polychaetes from intertidal areas in Panama, with a review of previous shallow-water records. Smithsonian Contributions to Zoology 221: 1-81. http://hdl.handle.net/10088/5511 doi: 10.5479/si.00810282.221

Faulwetter S, Chatzigeorgiou G, Galil BS, Nicolaidou A, Arvanitidis C (2011) Sphaerosyllis levantina sp. n. (Annelida) from the eastern Mediterranean, with notes on character variation in Sphaerosyllis hystrix Claparède, 1863. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 327–345. doi: 10.3897/zookeys.150.1877

Fauvel P (1919) Annélides Polychètes de Madagascar, de Djibouti, et du Golfe Persique. Archives de zoologie expérimentale et générale 58: 315–473. http://www.biodiversitylibrary.org/page/6316667

Fauvel P (1923) Polychètes errantes. Faune de France, Paris, 488 pp. http://www.faunedefrance.org/bibliotheque/docs/P.FAUVEL(FdeFr05)Polychetes-errantes.pdf

Fauvel P (1955) Contribution a la faune des Annelides Polychetes des côtes d'Israel. I. Bulletin of the Sea Fisheries Research Station, Haifa 10: 3–12.

Fauvel P (1957) Contribution a la faune des Annelides Polychetes des côtes d'Israel. II. Bulletin of the Research Council of Israel 6B: 213–219.

Fukuda MV, Yunda–Guarin G, Nogueira JMM (2009) The genus *Prosphaerosyllis* (Polychaeta: Syllidae: Exogoninae) in Brazil, with description of a new species. Journal of the Marine Biological Association of the United Kingdom 89: 1443–1454. doi: 10.1017/S0025315409000095

Gambi MC, Giangrande A, Martinelli M, Chessa LA (1995) Polychaetes of a *Posidonia oceanica* bed off Sardinia (Italy): Spatio-temporal distribution and feeding guild analysis. Scientia Marina 59: 129–141. http://www.icm.csic.es/scimar/index.php/secId/6/IdArt/2722

Gidholm L (1967) A revision of Autolytinae (Syllidae, Polychaeta) with special reference to Scandinavian species, and with notes on external and internal morphology, reproduction and ecology. Arkiv för Zoologi 19: 157–213.

Goodrich ES (1930) On a new hermaphrodite Syllid. Quarterly Journal of Microscopical Science, London 73: 651–666. http://jcs.biologists.org/content/s2-73/292/651.full.pdf

Gravier CJ (1900) Contribution à l'étude des Annélides Polychètes de la Mer Rouge. Première partie. Nouvelles Archives du Museum d'Histoire Naturelle Paris 2: 137–282.

Grube AE (1850) Die Familien der Anneliden. Archiv für Naturgeschichte 16: 249–364. http://www.biodiversitylibrary.org/ia/diefamilienderan00grub

Grube AE (1860) Beschreibung neuer oder wenig bekannter Anneliden. Beitrag: Zahlreiche Gattungen. Archiv für Naturgeschichte 26: 71–118. http://www.biodiversitylibrary.org/page/7153453

Grube AE (1863) Beschreibung neuer oder wenig bekannter Anneliden. Beitrag: Zahlreiche Gattungen. Archiv für Naturgeschichte 29: 37–69. http://www.biodiversitylibrary.org/page/7071934

Harlock R, Laubier L (1966) Notes on *Branchiosyllis uncinigera* (Hartmann-Schröder, 1960) new to the Mediterranean. Israel Journal of Zoology 15: 18–25.

Hartmann-Schröder G (1956) Polychaeten-Studien I. Zoologischer Anzeiger 157: 87–91.

Hartmann-Schröder G (1960) Polychaeten aus dem Roten Meer. Kieler Meeresforschungen 16: 69–125.

Hartmann-Schröder G (1962) Zur Kenntnis des Eulitorals der chilenischen Pazifikküste und der argentinischen Küste Südpatagoniens unter besonderer Berücksichtigung der Polychaeten und Ostracoden. Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 60: 57–270.

Hartmann-Schröder G (1965) Zur Kenntnis des Sublitorals der chilenischen Küste unter besonderer Berücksichtigung der Polychaeten und Ostracoden. Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 62: 59–305.

Hartmann-Schröder G (1979) Die Polychaeten der tropischen Nordwestküste Australiens (zwischen Derby im Norden und Port Hedland im Süden). Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 76: 77–218.

Hartmann-Schröder G (1980a) Die Polychaeten der Amsterdam-Expeditionen nach Westindien. Bijdragen tot de dierkunde 50: 387–401.

Hartmann-Schröder G (1980b) Die Polychaeten der tropischen Nordwestküste Australiens (zwischen Port Samson im Norden und Exmouth im Süden). Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 77: 41–110.

Hartmann-Schröder G (1981) Die Polychaeten der tropisch-subtropischen Westküste Australiens (zwischen Exmouth im Norden und Cervantes im Süden). Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 78: 19–96.

Hartmann-Schröder G(1982) Die Polychaeten der subtropisch-antiborealen Westküste Australiens (zwischen Cervantes im Norden und Cape Naturaliste im Süden). Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 79: 51–118.

Hartmann-Schröder G (1984) Die Polychaeten der antiborealen Südküste Australiens (zwischenAlbany im Westen und Ceduna im Osten). Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 81: 7–62.

Hartmann-Schröder G (1985) Die Polychaeten der antiborealen Südküste Australiens (zwischen Port Lincoln im Westen und Port Augusta im Osten). Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 82: 61–99.

Hartmann-Schröder G (1986) Die Polychaeten der antiborealen Südküste Australiens (zwischen Wallaroo im Westen und Port MacDonnell im Osten). Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 83: 31–70.

Hartmann-Schröder G (1990) Die Polychaeten der subtropisch-tropischen und tropischen Ostküste Australiens zwischen Lake Macquarie (New South Wales) im Süden und Gladstone (Queensland) im Norden. Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 87: 41–87.

Hartmann-Schröder G (1991) Die Polychaeten der subtropisch-tropischen bis tropischen Ostküste Australiens zwischen Maclean (New South Wales) und Gladstone (Queensland) sowie von Heron Island (Grosses Barrier Riff). Mitteilungen aus dem Hamburgischen zoologischen Museum und Institut 88: 17–71.

Hartmann-Schröder G (1996) Annelida, Borstenwürmer, Polychaeta. 2., neubearbeitete Auflage. Gustav Fischer Verlag, Jena, 648 pp.

Iken K, Konar B (2003) Natural Geography in Nearshore areas (NaGISA): The nearshore component of the Census of Marine Life. Gayana 67: 153–160. doi: 10.4067/S0717-65382003000200004

Imajima M (1966) The Syllidae (polychaetous annelids) from Japan. III. Eusyllinae. Publications of the Seto Marine Biological Laboratory 14: 85–116.

Imajima M (2003) Polychaetous Annelids from Sagami Bay and Sagami Sea collected by the Emperor Showa of Japan and deposited at the Showa Memorial Institute, National Science Museum, Tokyo (II), Orders included within the Phyllodocida, Amphinomida, Spintherida and Eunicida. National Science Museum Monographs 23: 1–221. http://ci.nii.ac.jp/naid/110004708004

Jiménez M, San Martín G, López E (1994) Redescription of *Pionosyllis neapolitana* Goodrich, 1930 and *Pionosyllis nutrix* Monro, 1936, referred to the genus *Grubeosyllis* Verrill, 1900 (Polychaeta, Syllidae, Exogoninae). Polychaete Research 16: 52–55.

Karhan SÜ, Kalkan E, Simboura N, Mutlu E, Bekbölet M (2008) On the occurrence and established populations of the alien polychaete *Polydora cornuta* Bosc, 1802 (Polychaeta: Spionidae) in the Sea of Marmara and the Bosphorus Strait (Turkey). Mediterranean Marine Science 9: 5–19. http://www.medit-mar-sc.net/files/200812/15-18480210.pdf

Kudenov, JD, Harris LH (1995) Family Syllidae Grube, 1850. In: Blake et al. (Eds) Taxonomic Atlas of the Bentic Fauna of the Santa Maria Basin and Western Santa Barbara Channel. The Annelida Part 2 - Polychaeta: Phyllodocida (Syllidae and Scale-Bearing Families), Amphinomida and Eunicida. Santa Barbara Museum of Natural History, Santa Barbara, 1–97.

Kuper M (2001) Ultrastrukturuntersuchungen der Segmentalorgane, der Spermien und der Brutpflegestrukturen innerhalb der Syllidae (Annelida: Polychaeta). Osnabrück, Germany: Universität Osnabrück. http://d-nb.info/964507099/34

Lamarck J-B de (1818) Histoire naturelle des Animaux sans Vertèbres, préséntant les caractères généraux et particuliers de ces animaux, leur distribution, leurs classes, leurs familles, leurs genres, et la citation des principales espèces qui s'y rapportent; précédée d'une Int. Déterville & Verdière, Paris, 612 pp. http://www.lamarck.cnrs.fr/ice/ice_book_detail-fr-text-lamarck-ouvrages_lamarck-38-1.html

Lanera P, Sordino P, Gambi MC (1990) Annellidi Policheti nuovi o poco conoscuiti per le coste italiane. Oebalia 16: 693–695.

Langerhans P (1879) Die Wurmfauna von Madeira. Zeitschrift für wissenschaftliche Zoologie 32: 513–592.

Langerhans P (1881) Über einige canarische Anneliden. Nova Acta der Kaiserlichen Leopold-Carolin Deutschen Akademie der Naturforscher 42: 95–124.

Langerhans P (1884) Die Wurmfauna von Madeira. IV. Zeitschrift für wissenschaftliche Zoologie 40: 247–285.

Licher F (2000) Revision der Gattung *Typosyllis* Langerhans, 1879 (Polychaeta: Syllidae). Morphologie, Taxonomie und Phylogenie. Abhandlungen der Senckenbergischen Naturforschenden Gesellschaft 551: 1–336.

Licher F, Kuper M (1998) *Typosyllis tyrrhena* (Polychaeta, Syllidae, Syllinae), a new species from the island Elba, Tyrrhenian Sea. Italian Journal of Zoology 65: 227–233. doi: 10.1080/11250009809386750

Liñero-Arana I, Díaz Díaz OF (2011) Syllidae (Annelida: Polychaeta) From The Caribbean Coast Of Venezuela. ZooKeys 117: 1–28. doi: 10.3897/zookeys.117.858

López E, San Martín G (1997) Eusyllinae, Exogoninae and Autolytinae (Syllidae, Annelida, Polychaeta) from the Chafarinas Islands (Alboran Sea, W Mediterranean. Miscellània Zoològica 20: 101–111. www.raco.cat/index.php/Mzoologica/article/viewFile/90383/145375

López E, San Martín G, Jiménez M (1996) Syllinae (Syllidae, Annelida, Polychaeta) from Chafarinas Islands (Alborán Sea, W Mediterranean). Miscellània Zoològica 19: 105–118. http://www.raco.cat/index.php/Mzoologica/article/viewFile/90438/145332

Malaquin A (1893) Recherches sur les Syllidiens. Morphologie, anatomie, reproduction, développement. Mémoires de la Société des sciences, de l'agriculture et des arts de Lille 18: 1–477.

McIntosh WC (1885) Report on the Annelida Polychaeta collected by H.M.S. Challenger during the years 1873–1876. Report on the Scientific Results of the Voyage of H.M.S. Chal-

lenger during the years 1872–76 12: 1–554. http://www.19thcenturyscience.org/HMSC/HMSC-Reports/Zool-34/htm/doc.html

Men F, Ding Z-hu, Zhao J, Wu B-L (1993) A preliminary study on small syllides from the Huanghai Sea (Yellow Sea). Journal of Oceanography of Huanghai and Bohai Seas 11: 19–36.

Milne Edwards H (1845) Observations sur le developpement des annelides. Annales des Sciences Naturelles, Paris 3: 145–182. http://biodiversitylibrary.org/page/13408583

Monro CCA (1937) A note on a collection of Polychaeta from the eastern Mediterranean with the description of a new species. Annals and Magazine of Natural History. Series 10 19: 82–86. doi: 10.1080/00222933708655241

Moore JP (1908) Some polychaetous annelids of the northern Pacific coast of North America. Proceedings of the Academy of Natural Sciences of Philadelphia 60: 321–364. http://biodiversitylibrary.org/page/24597271

Moore JP (1909) The polychaetous annelids dredged by the U.S.S. "Albatross" off the coast of southern California in 1904. I. Syllidae, Sphaerodoridae, Hesionidae and Phyllodocidae. Proceedings of the Academy of Natural Sciences of Philadelphia 1909: 321–351. http://biodiversitylibrary.org/page/26288382

Musco L, Giangrande A (2005) Mediterranean Syllidae (Annelida: Polychaeta) revisited: biogeography, diversity and species fidelity to environmental features. Marine Ecology Progress Series 304: 143–153. doi: 10.3354/meps304143

Nygren A (2004) Revision of Autolytinae (Syllidae: Polychaeta). Zootaxa 680: 1–314.

Núñez J, San Martín G (1991) Two new species of Syllidae (Polychaeta) from Tenerife (Canary Island, Spain). Bulletin of Marine Science 48: 236–241. http://www.ingentaconnect.com/content/umrsmas/bullmar/1991/00000048/00000002/art00010

Núñez J, San Martín G (1996) Anélidos poliquetos de las Islas Canarias. Familia Syllidae. I. Subfamilias Eusyllinae y Autolytinae. In: Llinás Gonzales et al. (Eds) I Congresso de Oceanografía y Recursos Marinos en el Atlántico Centro-Oriental (ICCM). Las Palmas, Gran Canaria (Spain), November 1990. Cabildo Insular de Gran Canaria, Spain, 197–217. http://mdc.ulpgc.es/cdm4/item_viewer.php?CISOROOT=/MDC&CISOPTR=112681

Núñez J, San Martín G, Carmen Brito M Del (1992) Exogoninae (Polychaeta: Syllidae) from the Canary Islands. Scientia Marina 56: 43–52. http://www.icm.csic.es/scimar/index.php/secId/6/IdArt/2584/

Olivier F, Grant C, San Martín G, Archambault P, McKindsey CW (2011) Syllidae (Annelida: Polychaeta: Phyllodocida) from the Chausey Archipelago (English Channel, France), with a description of two new species of the Exogoninae *Prosphaerosyllis*. Marine Biodiversity: Online First. doi: 10.1007/s12526-011-0092-1

Ørsted AS (1845) Fortegnelse over Dyr, samlede i Christianiafjord ved Drobak fra 21–24 Juli, 1844. Naturhistorisk Tidsskrift, Köbenhavn 1: 400–427.

Parapar J, San Martín G, Besteiro C, Urgorri V (1994) Aspectos sistemáticos y ecológicos de las Subfamilias Eusyllinae y Exogoninae (Polychaeta, Syllidae) en la Ría de Ferrol (Galicia, NO España). Boletín de la Real Sociedad Española de Historia Natural 91: 91–101.

Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith VS, Brake I, Ryrcroft S, Scott B, Johnson NF, Morris RA, Sautter G, Chavan V, Robertson T,

Remsen D, Stoev P, Parr C, Knapp S, Kress WJ, Thompson FC, Erwin T (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1–16. doi: 10.3897/zookeys.50.538

Penev L, Mietchen D, Chavan VS, Hagedorn G, Remsen DP, Smith VS, Shotton D (2011) Pensoft Data Publishing Policies and Guidelines for Biodiversity Data. Available at: http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf

Perkins TH (1981) Syllidae, principally from Florida, with descriptions of a new genus and twenty-one new species. Proceedings of the Biological Society of Washington 93: 1080–1172. http://biostor.org/reference/73932

Pierantoni U (1903) La gestazione esterna. Contributo alla biologia ed all embriologia dei Sillidi. Archivio Zoologico Italiano 1: 231–252. http://biodiversitylibrary.org/page/5747694

Por FD (1989) The legacy of Tethys: an aquatic biogeography of the Levant. Kluwer Academic Publishers, Dordrecht, 216 pp.

Ramos J, San Martín G, Sikorski A (2010) Syllidae (Polychaeta) from the Arctic and sub-Arctic regions. Journal of the Marine Biological Association of the United Kingdom 90: 1041–1050. doi: 10.1017/S0025315409991469

Riser NW (1991) An evaluation of taxonomic characters in the genus *Sphaerosyllis* (Polychaeta: Syllidae). Ophelia supplement: 209–218.

Rullier F (1972) Annélides polychètes de Nouvelle-Calédone. Expédition Française sur les Récifs Coralliens de la Nouvelle-Calédonie 6: 1–167.

Russell DE (1991) Exogoninae (Polychaeta: Syllidae) from the Belizean barrier reef with a key to species of *Sphaerosyllis*. Journal of Natural History 25: 49–74. doi: 10.1080/00222939100770061

Ruíz-Ramírez JD, Salazar-Vallejo SI (2001) Exogoninae (Polychaeta: Syllidae) del Caribe mexicano con una clave para las especies del Gran Caribe. Revista de Biología Tropical 49: 117–140. http://www.scielo.sa.cr/scielo.php?pid=S0034-77442001000100012

Saint-Joseph A de (1887) Les Annelides polychetes des cotes de Dinard, pt. 1. Annales des sciences naturelles 1: 127–270. http://biodiversitylibrary.org/page/33074463

San Martín G (1984a) Descripcion de una nueva especie y revision del genero *Sphaerosyllis* (Polychaeta: Syllidae). Cahiers de Biologie Marine 25: 375–391.

San Martín G (1984b) Estudio biogeográfico, faunistico y sistemático de los poliquetos de la familia Silidos (Syllidae: Polychaeta) en Baleares. PhD Thesis, Madrid, Spain: Universidad Complutense de Madrid.

San Martín G (1991a) *Grubeosyllis* and *Exogone* (Exogoninae, Syllidae, Polychaeta) from Cuba, the Gulf of Mexico, Florida and Puerto Rico, with a revision of *Exogone*. Bulletin of Marine Science 49: 715–740. http://www.ingentaconnect.com/content/umrsmas/bullmar/1991/00000049/00000003/art00004

San Martín G (1991b) *Sphaerosyllis* and *Parapionosyllis* (Polychaeta: Syllidae) from Cuba and Florida. Ophelia supplement 5: 231–238.

San Martín G (1994) Autolytinae (Polychaeta, Syllidae) from Cuba and north American Atlantic Ocean. Mémoires du Muséum national d'Histoire Naturelle 162: 269–277.

San Martín G (2003) Annelida, Polychaeta II: Syllidae. In: Ramos MA et al. (Eds) Fauna Ibérica, vol. 21. Museo Nacional de Ciencias Naturales. CSIC, Madrid, 554 pp.

San Martín G (2005) Exogoninae (Polychaeta: Syllidae) from Australia with the description of a new genus and twenty-two new species. Records of the Australian Museum 57: 39–152. http://publications.australianmuseum.net.au/pdf/1438_complete.pdf doi: 10.3853/j.0067-1975.57.2005.1438

San Martín G (2008) Syllinae (Polychaeta: Syllidae) from Australia. Part 1. Genera *Branchiosyllis*, *Eurysyllis*, *Karroonsyllis*, *Parasphaerosyllis*, *Plakosyllis*, *Rhopalosyllis*, *Tetrapalpia* n.gen., and *Xenosyllis*. Records of the Australian Museum 60: 119–160. http://australianmuseum.net.au/Uploads/Journals/18066/1494_complete.pdf doi: 10.3853/j.0067-1975.60.2008.1494

San Martín G, Acero MI, Contonente M, Gomez JJ (1982) Una coleccion de anelidos poliquetos de las costas mediterraneas andaluzas. Actas do II Simposio Iberico de Estudos do Benthos Marinho, Liboa 3: 171–182.

San Martín G, Alvarado R (1981) Nota sobre poliquetos de la isla de Cabrera (Baleares). Boletín de la Real Sociedad Española de Historia Natural 79: 221–234.

San Martín G, Bone D (2001) Syllidae (Polychaeta) de praderas de *Thalassia testudinum* en el Parque Nacional Morrocoy (Venezuela). Revista de Biología Tropical 49: 609–620. http://www.scielo.sa.cr/scielo.php?pid=S0034-77442001000200019

San Martín G, Gonzalez G, López E (1985) Aspectos sistematicos y ecologicos sobre algunas especies de Silidos (Polychaeta: Syllidae) de las costas gallegas. Boletin Instituto Espanol de Oceanografia 2: 27–36.

San Martín G, Hutchings PA (2006) Eusyllinae (Polychaeta: Syllidae) from Australia with the description of a new genus and fifteen new species. Records of the Australian Museum 58: 257–370. http://publications.australianmuseum.net.au/pdf/1466_complete.pdf doi: 10.3853/j.0067-1975.58.2006.1466

San Martín G, López E (2000) Three new species of *Syllis* (Syllidae: Polychaeta) from Iberian coasts. Cahiers de Biologie Marine 41: 425–433. http://www.sb-roscoff.fr/cbm/cbm.htm?execution=e1s2&_eventId=viewarticledetails&articleId=2666

San Martín G, López E, Aguado MT (2009) Revision of the genus *Pionosyllis* (Polychaeta: Syllidae: Eusyllinae), with a cladistic analysis, and the description of five new genera and two new species. Journal of the Marine Biological Association of the United Kingdom 89: 1455 doi: 10.1017/S0025315409003099

Sardá R (1984) La subfamilia Exogoninae (Polychaeta, Syllidae) de Gibraltar, con la descripción de *Pseudobrania euritmica* n. sp. Publicaciones del Departamento de Zoologia, Universidad de Barcelona, Facultad de Biologia 10: 7–13.

Simboura N (1996) Marine macrobenthic Polychaetes (Annelida, Polychaeta) of Greece: Taxonomy, Ecology, Zoogeography. PhD Thesis, Athens, Greece: University of Athens. http://phdtheses.ekt.gr/eadd/handle/10442/5952

Simboura N, Nicolaidou A (2001) The Polychaetes (Annelida, Polychaeta) of Greece: Checklist, Distribution and Ecological Characteristics. Monographs on Marine Sciences 4: 1–115.

Somaschini A, Gravina MF, Ardizzone GD (1994) Polychaete Depth Distribution in a *Posidonia oceanica* Bed (Rhizome and Matte Strata) and Neighbouring Soft and Hard Bottoms. Marine Ecology 15: 133–151. doi: 10.1111/j.1439-0485.1994.tb00049.x

Somaschini A, San Martín G (1994) Description of two new species of *Sphaerosyllis* (Polychaeta: Syllidae: Exogoninae) and first report of *Sphaerosyllis glandulata* for the Mediterranean Sea. Cahiers de Biologie Marine 35: 357–367.

Southern R (1914) Clare Island Survey. Archiannelida and Polychaeta. Proceedings of the Royal Irish Academy 31: 1–160. http://biodiversitylibrary.org/page/34773787

Surugiu V (2005) Inventory of inshore polychaetes from the Romanian coast. Mediterranean Marine Science 6: 51–74. http://www.medit-mar-sc.net/files/200812/15-1718395.pdf

Tebble N (1956) The polychaete fauna of the Gold Coast. Bulletin of the British Museum (Natural History) Zoology 3: 59–148. http://biodiversitylibrary.org/page/2243542

Tebble N (1959) On a collection of Polychaetes from the Mediterranean coast of Israel. Bulletin of the Research Council of Israel B8: 9–30.

Uebelacker JM (1984) Family Syllidae Grube, 1850. In: Uebelacker JM, Johnson PG (Eds) Taxonomic Guide to the Polychaetes of the Northern Gulf of Mexico, Volume V. Barry A. Vittor & Associates, Alabama, 30.31–30.151. http://biodiversitylibrary.org/page/3248872

Westheide W (1974a) Interstitielle Fauna von Galapagos. XI. Pisionidae, Hesionidae, Pilargidae, Syllidae (Polychaeta). Mikrofauna des Meeresbodens 44: 1–146.

Westheide W (1974b) Interstitielle Polychaeten aus brasilianischen Sandstränden. Mikrofauna des Meeresbodens 31: 1–16.

Zenetos A, Christianidis S, Pancucci MA, Simboura N, Tziavos C (1997) Oceanologic study of an open coastal area in the Ionian Sea with emphasis on its benthic fauna and some zoogeographical remarks. Oceanologica Acta 20: 437–451.

Supplement information

The Scratchpads version of this publication is available at:
http://polychaetes.marbigen.org/node/1636

Figures are available at:
http://polychaetes.marbigen.org/category/image-galleries/eastern-mediterranean-syllidae

Locations and their description are available at:
http://polychaetes.marbigen.org/content/darwincorelocation

Tables with species occurences are available at:
http://polychaetes.marbigen.org/content/species-occurences-sampling-stations
http://polychaetes.marbigen.org/content/reported-distribution-records-prosphaerosyllis-species-mediterranean

The data underpinning the analysis reported in this paper are deposited at:
GBIF (Global Biodiversity Information Facility):
http://ipt.pensoft.net/ipt/resource.do?r=easternmedsyllids

Dryad Data Repository: doi: 10.5061/dryad.4b7k408g

ZooKeys 150: 327–345 (2011)
doi: 10.3897/zookeys.150.1877
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Sphaerosyllis levantina sp. n. (Annelida) from the eastern Mediterranean, with notes on character variation in *Sphaerosyllis hystrix* Claparède, 1863

Sarah Faulwetter[1,4,†], Georgios Chatzigeorgiou[2,4,‡], Bella S. Galil[3,§], Artemis Nicolaidou[1,|], Christos Arvanitidis[4,¶]

1 *Department of Zoology-Marine Biology, Faculty of Biology, National and Kapodestrian University of Athens, Panepistimiopolis, 15784, Athens, Greece* **2** *Department of Biology, University of Crete, 71409 Heraklion, Crete, Greece* **3** *National Institute of Oceanography, Israel Oceanographic & Limnological Research, POB 8030, Haifa 31080, Israel* **4** *Institute of Marine Biology and Genetics, Hellenic Centre for Marine Research, 71003 Heraklion, Crete, Greece*

† *urn:lsid:zoobank.org:author:9BF02566-AF30-47EB-840E-DFC841B6FF84*
‡ *urn:lsid:zoobank.org:author:E3A716D4-9C20-4DD0-A231-EAA04168F17D*
§ *urn:lsid:zoobank.org:author:06EF9833-A3C5-48FA-BBA8-1881AC51E361*
| *urn:lsid:zoobank.org:author:9FA2A4EE-7E52-4111-A9B6-CDB99E7C909E*
¶ *urn:lsid:zoobank.org:author:737F149F-C30C-42EB-A690-5E693AD95427*

Corresponding author: *Sarah Faulwetter* (sarifa@hcmr.gr)

Academic editor: *C. Glasby* | Received 3 August 2011 | Accepted 18 October 2011 | Published 28 November 2011

urn:lsid:zoobank.org:pub:18BC77D8-3A1A-433F-AAC2-47534DF3FC48

Citation: Faulwetter S, Chatzigeorgiou G, Galil BS, Nicolaidou A, Arvanitidis C (2011) *Sphaerosyllis levantina* sp. n. (Annelida) from the eastern Mediterranean, with notes on character variation in *Sphaerosyllis hystrix* Claparède, 1863. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 327–345. doi: 10.3897/zookeys.150.1877

Abstract

Examination of polychaete specimens from Haifa Bay (Israel, eastern Mediterranean Sea) revealed several individuals exhibiting morphological characteristics similar to *Sphaerosyllis hystrix* Claparède, 1863. A detailed morphometrical analysis of the Israeli specimens in comparison to specimens of *S. hystrix* and *S. boeroi* Musco, Çinar & Giangrande, 2005 supported the description of the former as a new species, *S. levantina* **sp. n.** Individuals of *S. hystrix* formed a very heterogeneous group with strong character variations in the analysis and the presumed cosmopolitan distribution of the species is discussed based on literature records.

Keywords
Polychaetes, Syllidae, Exogoninae, *Sphaerosyllis,* new species, Mediterranean, Cybertaxonomy, Scratchpads

Introduction

The polychaete genus *Sphaerosyllis* Claparède, 1863 (Annelida) is one of the most species-rich genera of the syllid subfamily Exogoninae. At present, ca. 48 species are considered valid within *Sphaerosyllis* after the recent split of the group into the three genera *Sphaerosyllis*, *Prosphaerosyllis* and *Erinaceusyllis* (San Martín 2005). Up to date, 18 species of the genus have been recorded from the Mediterranean Sea (Musco and Giangrande 2005), one of them described but yet unnamed (San Martín 2003), another one in the process of description (Del Pilar-Ruso and San Martín in press). In the framework of a project focusing on the soft bottom benthos of Haifa Bay (Israel, eastern Mediterranean Sea), a number of individuals of the genus *Sphaerosyllis* were found to exhibit morphological features which did not entirely correspond to any description of known *Sphaerosyllis* species, namely falcigers with a strong serration and with a subdistal spine present in all chaetigers. A subdistal spine on the blades of at least some falcigers has been described for the type species of the genus, *S. hystrix* Claparède 1863, and for *S. boeroi* Musco, Çinar and Giangrande, 2005. Re-examination of material of *S. hystrix* revealed that some individuals –contrary to descriptions available in the literature– possess a subdistal spine not only on the blades of the falcigers in anterior but also in posterior chaetigers. Consequently, this characteristic could not be used to unambiguously distinguish the Israeli material from *S. hystrix*. In order to clarify the relationship between the three very similar species possessing falcigers with a subdistal spine, a morphometric analysis has been performed, a method allowing not only to discriminate statistically significant groupings but also to identify taxonomically important characters (Costa-Paiva and Paiva 2007).

Material and methods

Specimen collection and processing

Specimens were collected on 11 Oct. 2009 in Haifa Bay, (Israel, Eastern Mediterranean Sea) from fine to medium sands in shallow waters (10 m). Sediment samples were taken with a Van-Veen grab (KAHLSICO, model WA265/SS214) 32×35 cm, volume 20 l, penetration 20 cm. The sediment was preserved in buffered formalin 10% for 3–7 days, then sieved through a 250 µm mesh sieve and subsequently stored in 70% ethanol. Specimens were examined under an Olympus SZx12 stereomicroscope and an Olympus BX50 microscope. Illustrations in pencil were made by means of a drawing tube, subsequently scanned, imported into a graphic program (GIMP), re-drawn

and saved as a vector graphic. Three specimens selected for obtaining Scanning Electron Microscope (SEM) images were dehydrated, critical point dried (Bal-Tec CPD 030), sputter-coated with gold (Bal-Tec SCD 050) and examined under a JEOL JSM-6390LV at the Department of Biology, University of Crete. Specimens are deposited in the invertebrate collection of the Smithsonian National Museum of Natural History, Washington D.C., USA (USNM) and in the Tel Aviv University Zoological Museum, Israel (TAU).

Morphometric analyses

A total of 30 individuals belonging to three species (*S. boeroi*: 3 individuals; *S. hystrix*: 21 individuals; *S. levantina* sp. n.: 6 individuals) were analysed. Twenty-five variables were measured: I. body length, to account for size-dependencies of other characters; II. number of chaetigers; III. length of blade of dorsalmost falciger of a) anterior, b) midbody, c) posterior chaetigers; IV. length of blade of ventralmost falciger of a) anterior, b) midbody, c) posterior chaetigers; V. ratio of length of blades of dorsalmost to ventralmost falciger in a) anterior, b) midbody, c) posterior chaetigers; VI. ratio of length of blades of falcigers in anterior to posterior chaetigers for a) dorsalmost; b) ventralmost falciger; VII. Ratio of length of dorsalmost falciger to body length in in a) anterior, b) midbody, c) posterior chaetigers; VIII. Ratio of length of ventralmost falciger to body length in in a) anterior, b) midbody, c) posterior chaetigers; IX. maximum length of serration of falcigerous blades in a) anterior, b) midbody, c) posterior chaetigers (smooth, finely serrated, strongly serrated); X. presence of a subdistal spine in dorsalmost falcigerous blades of in a) anterior, b) midbody, c) posterior chaetigers.

Body length was measured excluding antennae, anal cirri and palps. Falciger blade lengths were measured from point of insertion into shaft to distal tip. Falciger blade lengths could not always be measured on the same chaetiger in all animals if blades were broken. Instead, measurements were made in predefined body regions (anterior: first 1–5 chaetigers; posterior: last 5–7 chaetigers; midbody: in between). Three individuals of *S. hystrix* were excluded from the multivariate statistical analysis due to missing values for some characters.

Summary statistics (mean, minimum, maximum, standard deviation, coefficient of variation and range of values) were calculated for each species (measurements and calculations available in online supplementary material:

http://polychaetes.marbigen.org/content/measured-values-sphaerosyllis-specimens
http://polychaetes.marbigen.org/content/summary-statistics-sphaerosyllis-hystrix
http://polychaetes.marbigen.org/content/summary-statistics-sphaerosyllis-boeroi
http://polychaetes.marbigen.org/content/summary-statistics-sphaerosyllis-levantina

To take the different data types (numerical, categorical, binary) into account, Gower's similarity coefficient (Gower 1971) was chosen to calculate a similarity matrix.

Multidimensional Scaling (MDS) was subsequently employed to display the similarities of the different individuals. To test for significance of differences between species a PERMANOVA (Permutational Multivariate Analysis of Variance) was performed (Anderson 2001). A Principal Component Analysis (PCA) was used to determine variability of characters and to identify characters for the species differentiation. To determine the importance of the characters discriminating the species, the Principal Component Scores were correlated (Spearman's correlation coefficient) with the measured character values of each individual.

Multivariate statistical analyses were performed with PRIMER V6, correlation of the Principal Component Scores were calculated with the R package (R package version 2.10; http://www.R-project.org).

Electronic publication

The description of the new taxon was prepared in a Virtual Research Environment (Scratchpads) allowing for rapid and simultaneous publication of the results in print as well as electronically in a semantically enhanced form (Blagoderov et al. 2010, Penev et al. 2010). This publication and all supplementary data (measurements, results of statistical analyses, images) can be accessed on the Polychaete Scratchpads (http://polychaetes.marbigen.org).

Results

Taxonomic results

***Sphaerosyllis levantina* sp. n.**
urn:lsid:zoobank.org:act:9CEE8F90-9596-49F6-AA22-BB79C0E816D9
http://species-id.net/wiki/Sphaerosyllis_levantina
Figures 1–4

Type material. Holotype (USNM 1160540) ALA-IL-7, Haifa Bay, 10.5 m depth. Label: "*Sphaerosyllis levantina*, Haifa Bay, coll. B. Galil 11.10.09 [Holotype]". Paratypes USNM 1160541–1160573: 33 individuals, TAU-AN 25006: 10 individuals; Haifa Bay, Israel, Eastern Mediterranean Sea, Station ALA-IL-7, coll. 11.10.2009, depth 10.5 m; Labels: "*Sphaerosyllis levantina*, Haifa Bay, coll. B. Galil 11.10.09 [Paratype X]" (where X=1–43). All material preserved in 96% Ethanol.

Comparative material examined. *S. boeroi* Musco, Çinar, and Giangrande, 2005 (Southern Evoikos Gulf, Aegean Sea, Greece: 3 specimens [Label: Tribe *Sphaerosyllis*]). *S. hystrix* (Southern Evoikos Gulf, Aegean Sea, Greece: 1 specimen [Label: Tribe *Sphaerosyllis*]; Northern Evoikos Gulf, Aegean Sea, Greece: 7 specimens [Label: DI9a 7.3.91 *Sphaerosyllis hystrix*, checked S.Martín], all deposited the in Hel-

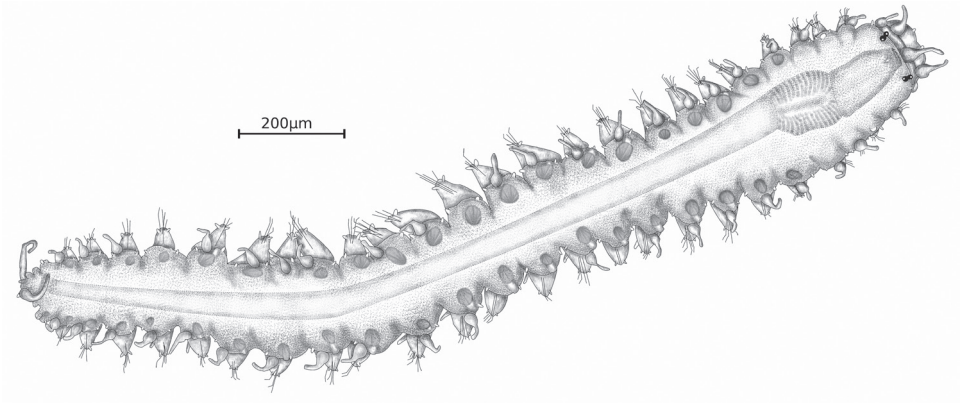


Figure 1. *Sphaerosyllis levantina* sp. n. holotype, dorsal view

lenic Centre for Marine Research, Anavyssos, Greece; Chalkida, Aegean Sea, Greece: 1 specimen [Label: 56 – *Sphaerosyllis hystrix*, κατώτερη μεσοπαραλιακή Χαλκίδας, Στενά Ευρίπου, Ξενοδοχείο Λούσι, St. 18, 25.9.97 0-0.5m, Άτομα: 1, Διδακτορικού Μίλτου] (= lower intertidal zone, Chalkida, Eviros Straight, Hotel Lousi, coll. M.S. Kitsos), Chalkida, Aegean Sea, Greece: 1 specimen [Label: 26 – *Sphaerosyllis hystrix*, κατώτερη μεσοπαραλιακή Χαλκίδας, Στενά Ευρίπου, Ξενοδοχείο Παλίρροια, St. 1a, 24.9.97 0-0.5m, Άτομα: 1, Διδακτορικού Μίλτου] (= lower intertidal zone, Chalkida, Eviros Straight, Hotel Palirroia, coll. M.S. Kitsos), Chalkida, Aegean Sea, Greece: 6 specimens [Label: 33 – S*phaerosyllis hystrix*, κατώτερη μεσοπαραλιακή Χαλκίδας, Στενά Ευρίπου, Ξενοδοχείο Παλίρροια, St. 1α, 24.9.97 0-0.5m, Άτομα: 6, Διδακτορικού Μίλτου] (= lower intertidal zone, Chalkida, Eviros Straight, Hotel Palirroia, coll. M.S. Kitsos), Chalkida, Aegean Sea, Greece: 4 specimens [Label: 80 – *Sphaerosyllis hystrix*, κατώτερη μεσοπαραλιακή Χαλκίδας, Στενά Ευρίπου, Ξενοδοχείο Παλίρροια, St. 1α, 24.9.97 0-0.5m, Άτομα: 6, Διδακτορικού Μίλτου] (= lower intertidal zone, Chalkida, Eviros Straight, Hotel Palirroia, coll. M.S. Kitsos), Thessaloniki, Aegean Sea, Greece, 1 specimen [Label: 66 – *Sphaerosyllis hystrix*, κατώτερη μεσοπαραλιακή Λιμάνι Θεσσαλονίκης, 2γ, 6.10.97 0-0.5m, Άτομα: 1, Διδακτορικού Μίλτου]) (= lower intertidal zone, Port of Thessaloniki, coll. M.S. Kitsos), all deposited the in Zoological Museum of the Aristotle University of Thessaloniki, Greece.

Type locality. Eastern Mediterranean Sea, Levantine Basin, Israel, Haifa Bay (32°54.533N, 35°04.071E).

Description. Holotype, entire animal, with 25 chaetigers, length 1.9 mm with palps but without anal cirri; width at sixth chaetiger 250 µm without parapodia, 300 µm with parapodia. Body small, slender, widest at level of proventricle (Fig. 1). Dorsal papillation on anterior chaetigers irregular, after proventricle in four longitudinal rows: two mid-dorsal rows with two papillae per segment, lateral rows with three papillae near dorsal cirri (Fig. 2a). Ventrum without visible papillation. Prostomium wider than long with 4 coalescent lensed eyes in trapezoidal arrangement. Anterior eyespots absent. Antennae pyriform with bulbous bases and elongated tips, median antenna

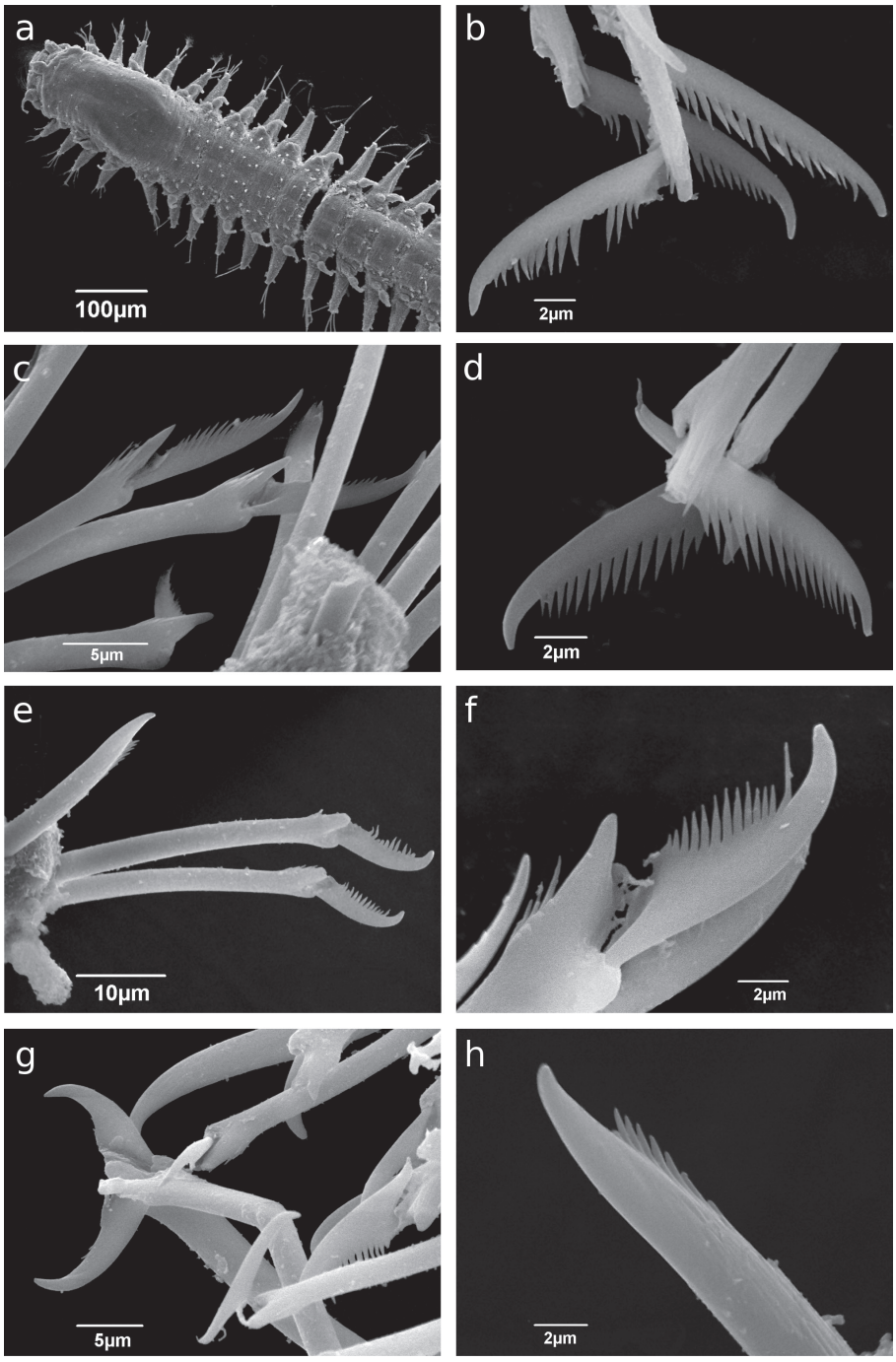


Figure 2. *Sphaerosyllis levantina* sp. n. SEM images of **a** anterior end and midbody, dorsal view **b–c** compound chaetae, anterior chaetigers **d** dorsalmost compound chaetae, anterior chaetiger **e** compound and dorsal simple chaetae, midbody **f** dorsalmost compound chaeta, posterior chaetiger **g** ventralmost compound chaetae, posterior chaetiger **h** dorsal simple chaeta

40 μm long, lateral ones 33 μm, longer than prostomium and palps together. Median antenna inserted between anterior pair of eyes, lateral ones attached on anterior margin of prostomium (Fig. 1). Palps directed ventrally, fused along their length, with a dorsal notch and few small papillae. Peristomium indistinct, dorsal fold partly covering prostomium. One pair of tentacular cirri, shaped like antennae but shorter (23 μm). Second chaetiger without dorsal cirri but with large papilla instead. Dorsal cirri similar in shape and length to tentacular cirri, anteriorly as long as parapodial lobes (23 μm), posteriorly slightly longer (28 μm). Ventral cirri conical, half as long as parapodial lobe, originating at bases of parapodia. Parapodial lobes triangular, with small papilla on each side of distal end. Parapodial glands with fibrillar material and with conical opening; from fourth chaetiger. Anterior parapodia with 4–5, rarely with 6 falcigers per fascicle; blades slender, unidentate with small subdistal spine and strong serration on 1–2 dorsalmost falcigers (Figs 2b–d, 3a). Dorso-ventral gradation in length of blades, dorsal ones maximally 14 μm, ventral ones 10 μm. Posteriorly, dorsal blades of similar length (13 μm), but stouter and more curved with robust subdistal spine and strong serration as long as subdistal spine (Figs 2e, f, 3b, c). Dorsalmost falciger posteriorly thicker than remaining ones in fascicle. Blades of ventral falcigers similar throughout body (Fig. 2g). All shafts with fine serration (Fig. 2c). Dorsal simple chaeta from chaetiger 1, subdistally serrated (Figs 2h, 4a). Ventral simple chaeta on posterior chaetigers, sigmoid, smooth (Fig. 4b). Anteriorly two aciculae per parapodium, one distally bent at right angle, acuminate tip curved upwards, the other straight and blunt (Fig. 4c); posteriorly only one acicula of the former type per parapodium. Pharynx occupying three chaetigers. Width more than ¾ of width of proventricle. Pharyngeal tooth located on anterior margin, surrounded by a crown of soft papillae. Proventricle in chaetigers 3–4 (120 μm long) with 15–17 muscle cell rows. Pygidium papillated, with two cirriform anal cirri twice as long as dorsal cirri (60 μm) (Fig. 1).

Etymology. Derived from the type locality (Levantine Basin), levantina being a neo-Latin adjective meaning "pertaining to the region where the sun raises"; feminine declination in accordance with the genus name (Syllis was a river nymph in the greek mythology and thus female).

Distribution. Israeli Coast (Levantine Basin, Eastern Mediterranean Sea).

Habitat. Fine to medium sands.

Taxonomic remarks. *S. levantina* sp. n. is similar to *S. minima* Hartmann-Schröder, 1960 in having blades of falcigers with strong serration throughout the body. However, *S. minima* has a stronger dorso-ventral gradation of the blades of falcigers (dorsal ones twice as long as ventral ones) than *S. levantina* sp. n. (dorsal ones 1.5 times longer than ventral ones) and the ventral cirrus is longer than the parapodial lobe in *S. minima*, whereas is is half as long as the parapodial lobe in *S. levantina* sp. n. *S. capensis* Day, 1953, *S. taylori* Perkins, 1981, and *S. sandrae* Álvarez and San Martín, 2009 are similar to *S. levantina* sp. n. in the shape and serration of the blades of the falcigers, but *S. capensis* has all antennae positioned in line (median one posteriorly of lateral ones in *S. levantina* sp. n.), *S. taylori* shows no dorso-ventral gradation of the falciger blade length (dorsal blade 1.5 times longer than ventral one in *S. levantina* sp. n.) and

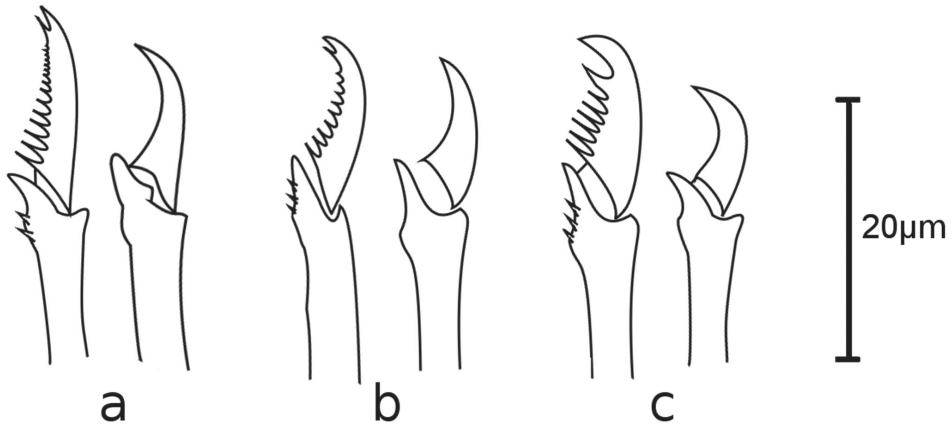


Figure 3. *Sphaerosyllis levantina* sp. n. Dorsal (left) and ventral (right) falciger of **a** anterior **b** midbody **c** posterior chaetiger

S. sandrae has smooth falcigerous blades posteriorly and parapodial glands with hyaline material (strongly serrated blades throughout the body and parapodial glands with fibrillar material in *S. levantina* sp. n.). All the above species differ from *S. levantina* sp. n. by lacking a subdistal spine on the blades of the falcigers. The only *Sphaerosyllis* species known to possess this spine are *S. hystrix* Claparède, 1863, *S. parabulbosa* San Martín and López, 2002 and *S. boeroi* Musco Çinar and Giangrande, 2005. *S. parabulbosa* clearly differs from *S. levantina* sp. n. by having minute dorsal cirri and antennae, by the presence of a subdistal spine only on blades of the posterior falcigers and by smooth blades of posterior falcigers. *S. boeroi* differs from *S. levantina* sp. n. in having much longer blades of the falcigers which show a more pronounced dorsoventral gradation (dorsal blades 2.6 times longer than ventral ones in *S. boeroi*, 1.5 times longer in *S. levantina* sp. n.) than those of *S. levantina* sp. n. (Figs 3, 5, see also tables in online supplementary material), by having a subdistal spine on blades of all falcigers (only on dorsalmost ones in *S. levantina* sp. n.) and by the dorsalmost falcigers being serrated only proximally. *S. hystrix*, according to the literature, has a subdistal spine only on the blades of the anterior dorsalmost falcigers. However, in the examined material of *S. hystrix* from the Aegean Sea 8 out of 21 specimens also possessed a subdistal spine in posterior falcigers. *S. hystrix* can nevertheless be distinguished from *S. levantina* sp. n. by having smooth or finely serrated posterior falcigers (serration less than half the length of the subdistal spine), even when the spine is present (serration almost as long as subdistal spine in *S. levantina* sp. n.) (Figs 2f, 3, 6). Furthermore, the blades of the dorsalmost falcigers show an anteroposterior gradation in length in *S. hystrix* (anteriorly 1.5 times longer than posteriorly), whereas they are of similar length throughout the body in *S. levantina* sp. n. (Figs 3, 6, see also tables in online supplementary material). Finally, *S. hystrix* has a very narrow pharynx (almost half the width of proventricle), whereas the pharynx of *S. levantina* sp. n. is wider than ¾ of the width

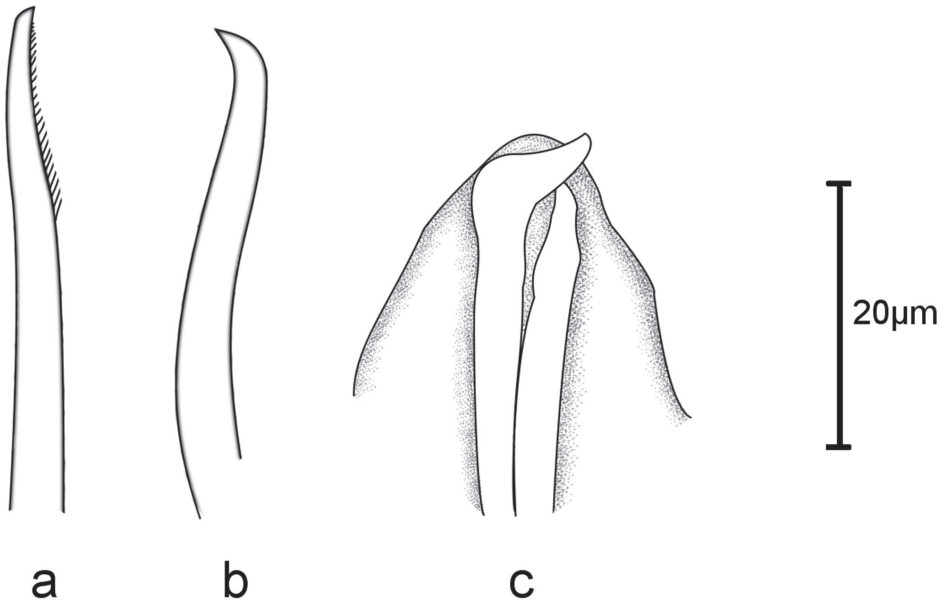


Figure 4. *Sphaerosyllis levantina* sp. n. **a** dorsal **b** ventral simple chaeta **c** aciculae, anterior chaetiger

of the proventricle. An identification key to the Mediterranean *Sphaerosyllis* species is provided at the end of this manuscript.

Ben-Eliahu (1977) discusses two different morphological forms of *S. hystrix* occurring in her samples from Israel. Based on her description and illustrations, the animal identified as *S. hystrix* sensu Westheide 1974 could potentially belong to *S. levantina* sp. n. because of the similar characters of falcigers and papillation. However, the description does not report the characteristic subdistal spine on the blades of the posterior falcigers. In addition, Westheide's (1974) description of *S. hystrix* from the Galápagos Islands differs from both Ben-Eliahu's specimen and the present material by the absence of parapodial glands (Westheide 1974), a character considered as variable and thus of no taxonomic value by Ben-Eliahu (1977) but recently accepted as a taxonomically stable character (Riser 1991).

Multivariate morphometrical analysis

The results of the Principal Component Analysis show that the first principal component (PC1) account for 77.4% of the variability, the second (PC2) for 16.4% and the remaining 3 PCs for 5.1% (eigenvector values available at http://polychaetes.marbigen.org/content/morphometric-analysis-pca-eigenvectors). The Spearman's correlation of the Principal Component scores with the measured character values of the individuals revealed that the length of the dorsalmost falcigerous blades in all

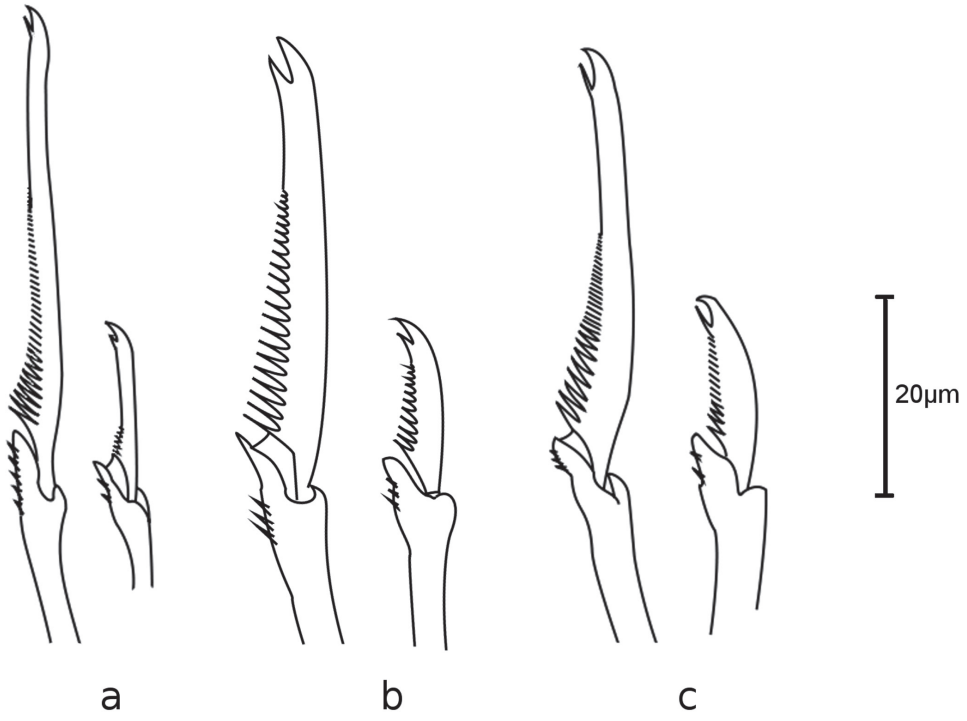


Figure 5. *Sphaerosyllis boeroi*. Dorsal (left) and ventral (right) falciger of **a** anterior **b** midbody **c** posterior chaetiger

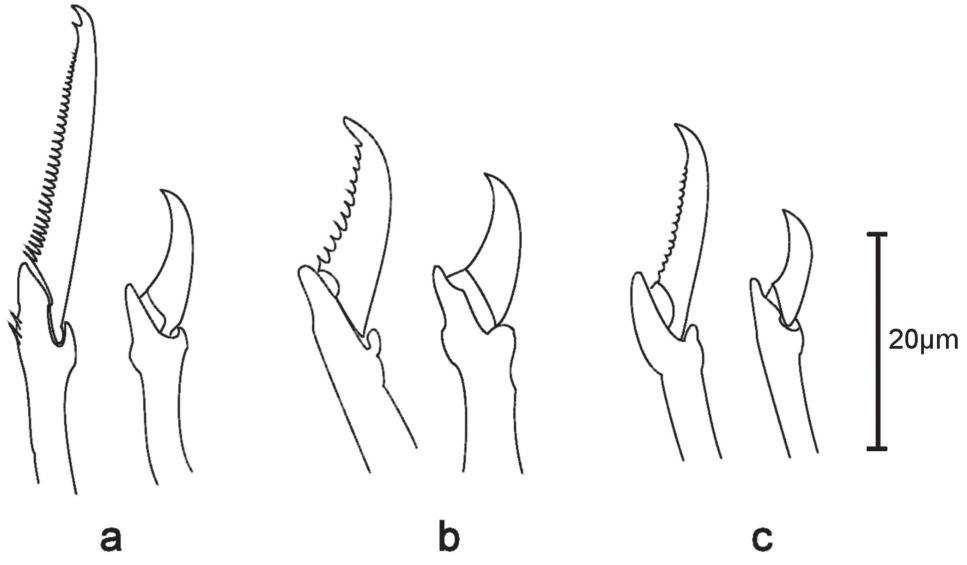


Figure 6. *Sphaerosyllis hystrix*. Dorsal (left) and ventral (right) falciger of **a** anterior **b** midbody **c** posterior chaetiger

body parts (anterior, midbody, posterior), as well as the ratio of the anterior to posterior ventralmost falcigerous blade are the most important characters discriminating between the three species (ϱ-values >0.8 / <-0.8 at $p < 0.005$) (http://polychaetes.marbigen.org/content/spearmans-correlation-principal-component-scores-vs-measurements).

The PCA plot of the first two components show a discrimination of species into three groups, with individuals of *S. levantina* sp. n. having the lowest PC1 scores, *S. boeroi* the highest scores. *S. levantina* sp. n. and *S. hystrix* show similar PC2 scores, whereas *S. boeroi* shows lower scores, and, except for one small-sized individual, forms a distinct group apart from the remaining species. Individuals of *S. levantina* sp. n. likewise form a close group, however, a couple of individuals of *S. hystrix* cannot be distinguished from this cluster (Fig. 7). The MDS diagram gives similar results, with individuals of *S. boeroi* and *S. levantina* sp. n. forming distinct groups, whereas individuals of *S. hystrix* are spread as a heterogeneous group, with some of them being plotted close to individuals of either *S. boeroi* or *S. levantina* sp. n. (Fig. 8).

The PERMANOVA analysis results in a *p*-value of 0.001 as calculated by 999 permutations, thus the null-hypothesis (no differences between the groups) cannot be sustained. Subsequent analyses of the differences between species through pairwise tests reveals significant differences between species (*S. hystrix* / *S. boeroi*: $p = 0.003$, 713 permutations; *S. hystrix* / *S. levantina* sp. nov: $p = 0.001$, 995 permutations; *S. boeroi* / *S. levantina* sp. n.: $p = 0.015$, 84 permutations).

Discussion

The genus *Sphaerosyllis* –like many of the small-sized Exogoninae genera– has a difficult and often confused taxonomy and biogeography. Among the potential causes contributing to the current confusion the following could be cited: a) lack of detail in older (before ca. 1970) species descriptions; b) difficulties of observing certain characters in fixed material (Riser 1991); c) descriptions of new species without examination of comparative material; d) ongoing discussions on the taxonomic value of characters such as the presence or absence of dorsal cirri on the second chaetiger (Fauvel 1923, San Martín 2005), presence and type of parapodial glands (Westheide 1974, Ben-Eliahu 1977, Riser 1991) and variations in chaetal structures (Riser 1991). These factors have lead to the assignment of individuals with very different character sets to the same species name and thus to wide-spread distribution records of some species. *S. hystrix* (type locality Normandy, France) is included among those species with an alleged cosmopolitan distribution, since it has been recorded from most European coasts including the Mediterranean Sea, the north-western coasts of America (Berkeley and Berkeley 1948, Hartman 1968), the Galápagos Islands (Westheide 1974), China (Men et al. 1993, Ding and Westheide 2008), Australia (Hartmann-Schröder 1984, 1985) and the Western Atlantic (Hartman and Fauchald 1971, Temperini 1981), among others. However, recent studies suggest that the North American records of *S. hystrix* and *S. pirifera* Claparède, 1868 are in fact

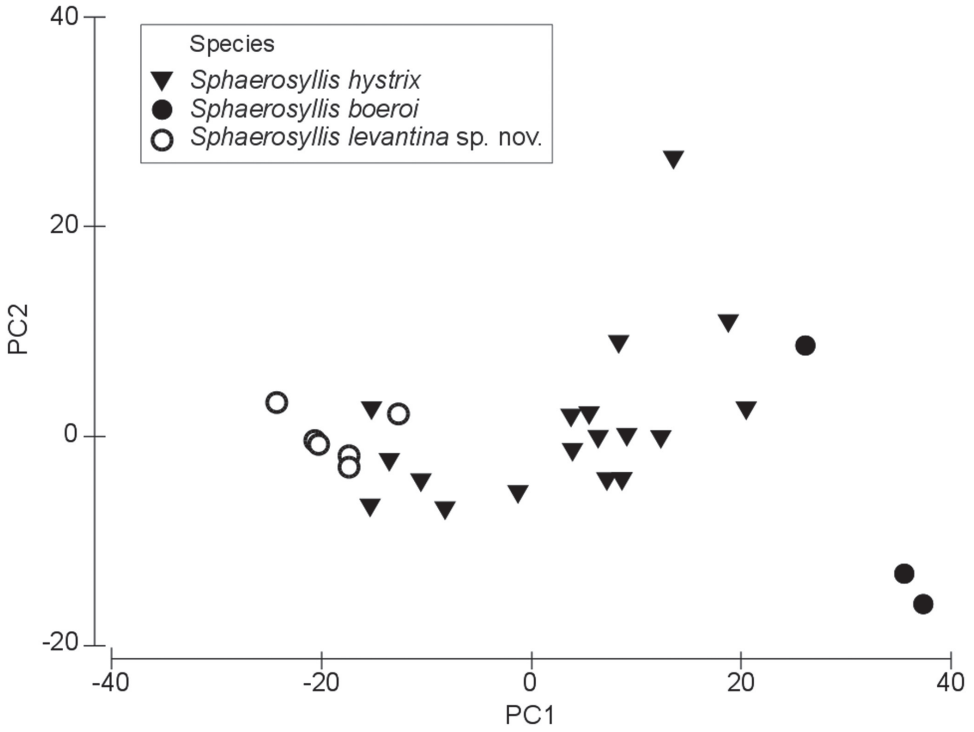


Figure 7. PCA plot.

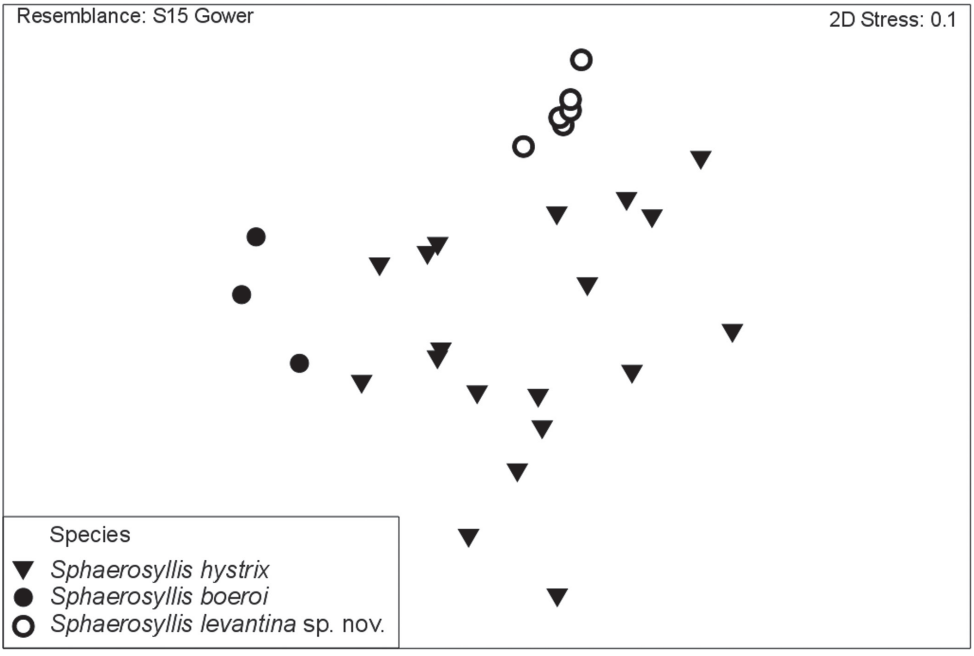


Figure 8. MDS plot.

individuals of *S. californiensis* Hartman, 1966 and that the two European species are not represented in the American Pacific fauna (Kudenov and Harris 1995). Similarly, some specimens from the Mediterranean Sea previously identified as *S. hystrix* had been re-examined and found to exhibit significant morphological differences to *S. hystrix*, leading to the establishment of a new species, *S. boeroi* (Musco et al. 2005). In the light of an ever-increasing number of molecular analyses revealing cryptic species complexes in morphologically indistinguishable polychaete species with an assumed cosmopolitan distribution (e.g. Westheide and Hass-Cordes 2001, Westheide and Schmidt 2003, Barroso et al. 2009, Bleidorn et al. 2006) it is likely that the various specimens recorded under the name *S. hystrix* may in fact form a complex of similar species, especially since many descriptions differ substantially from each other (see Ben-Eliahu 1977).

The morphometric analysis conducted in this study support the hypothesis of several morphologically very similar species co-existing in the Mediterranean. The individuals of *S. levantina* sp. n. and *S. boeroi* form distinct groups in the PCA and MDS plots, however the individuals of *S. hystrix* show a much wider spread, marginally overlapping with the other two species when only the meristic characters are taken into account. This is explained through a high character variability in the examined individuals, especially concerning the presence of a subdistal spine on the blades of the posterior falcigers and the length of the falciger blades. The presence of a subdistal spine on all dorsal falcigerous blades is invariable in *S. boeroi* and *S. levantina* sp. n., wheras individuals of *S. hystrix* with otherwise very similar chaetal structures might or might not possess such spine. Another feature that seems to be highly variable in *S. hystrix* is the length of the falciger blades in relation to body size. In fact, individuals of *S. levantina* sp. n. with short falciger blades are located at the lower end of the size spectrum of all measured blades, *S. boeroi* with almost spiniger-like blades at the higher end, whereas the blade lengths of the examined individuals of *S. hystrix* form a smooth transition between the other two species.

However, when tested by strict statistical criteria, the hypothesis of different co-existing species is significantly supported, and based on their meristic characters the species show significant differences. The results of the current study suggest that *S. hystrix* may well constitute a species complex. Given the difficult taxonomic status of the genus, similar results might be expected for other species as well, and consequently, distributions of several *Sphaerosyllis* species might be in fact questionable or unknown.

Key to the Mediterranean *Sphaerosyllis* species:

The three species *S. claparedei* Ehlers, 1864, *S. papillifera* Naville, 1933 and *S. ovigera* Langerhans, 1879 are poorly known. All have been described as having dorsal cirri on the second chaetiger, however, other species, such as *S. hystrix*, were also originally described or illustrated with dorsal cirri on the second chaetiger whereas they are in fact absent. Since the three aforementioned species are exclusively known from their

original description (or partly reproductions of these) and have never been re-described based on new material, they are tentatively included in the key below, but their identity remains questionable.

1 Dorsal cirri on chaetiger 2 present **2**
– Dorsal cirri on chaetiger 2 absent **4**
2 Papillae on dorsum absent ..... ***Sphaerosyllis claparedei*** **Ehlers, 1864**
– Papillae on dorsum present **3**
3 Parapodial glands absent ..... ***Sphaerosyllis papillifera*** **Naville, 1933**
– Parapodial glands with fibrillar material ..... ***Sphaerosyllis ovigera*** **Langerhans, 1879**
4 Parapodial glands present **5**
– Parapodial glands absent **15**
5 Parapodial glands with fibrillar material **6**
– Parapodial glands with granular material **12**
6 All antennae in line ..... ***Sphaerosyllis capensis*** **Day, 1953**
– Median antenna inserted more posteriorly than lateral ones **7**
7 Dorsal cirri shorter than parapodial lobes, at least in anterior chaetigers **8**
– Dorsal cirri longer than parapodial lobes **9**
8 Blades of falcigers strongly serrated, short (<10µm); shafts with strong spines ..... ***Sphaerosyllis thomasi*** **San Martín, 1984**
– Blades of falcigers with serration only anteriorly and dorsalmost; blades with slight dorso-ventral gradation but always longer than 10µm; shafts smooth ... ***Sphaerosyllis parabulbosa*** **San Martín and López, 2002**
9 Blades of falcigers without marked dorso-ventral gradation in length ..... ***Sphaerosyllis taylori*** **San Martín, 1984**
– Blades of dorsalmost falcigers at least 1.5 times the length of ventral ones **10**
10 Blades of posterior dorsal compound falcigers smooth to finely serrated ..... ***Sphaerosyllis hystrix*** **Claparède, 1863**
– Blades of posterior dorsal compound falcigers strongly serrated (spinules of almost same length as the subdistal spine) **11**
11 Blades of anterior dorsal compound falcigers at least twice as long as ventral ones; anteroposterior gradation of blade length; blades of both dorsal and ventral compound chaetae with a subdistal spine ..... ***Sphaerosyllis boeroi*** **Musco, Çinar and Giangrande, 2005**
– Blades of anterior dorsal compound falcigers less than twice as long as ventral ones; no anteroposterior gradation of blade length; blades of only dorsal compound chaetae with a subdistal spine ..... ***Sphaerosyllis levantina*** **sp. n.**
12 Blades of dorsalmost falcigers long (>30µm), at least twice as long as ventral ones ..... ***Sphaerosyllis magnidentata*** **Perkins, 1981**
– Blades of falcigers short (<15µm), with only slight dorso-ventral gradation ... **13**
13 Dorsal cirri clearly longer than parapodial lobes ..... ***Sphaerosyllis*** **sp. [San Martín 2003]**

– Dorsal cirri as long as or shorter than parapodial lobe **14**

14 Antennae bulbous with small tip, shorter than prostomium; dorsal simple chaetae smooth; anterior parapodia with two aciculae, one straight, one with tip bent at right angle ..... ***Sphaerosyllis*** **sp. Del-Pilar-Ruso & San Martín, in press**

– Antennae pyriform, as long as prostomium, dorsal simple chaetae serrated, all parapodia with one acicula..... ***Sphaerosyllis glandulata*** **Perkins, 1981**

15 All aciculae straight **16**

– Tip of some aciculae bent at right angle **17**

16 Dorsal cirri with conspicious papilla, giving cirri a bifid appearance..... ***Sphaerosyllis gravinae*** **Somaschini & San Martín, 1994**

– Doral cirri without papilla ..... ***Sphaerosyllis bulbosa*** **Southern, 1914**

17 All antennae in line..... ***Sphaerosyllis austriaca*** **Banse, 1959**

– Median antenna inserted more posteriorly than lateral ones..... **18**

18 Anterior parapodia with two aciculae, one straight, one with tip bent at right angle; pharyngeal glands on chaetiger 1 present ..... ***Sphaerosyllis pirifera*** **Claparède, 1868**

– All parapodia with one acicula only; pharyngeal glands on chaetiger 1 absent ..... ***Sphaerosyllis piriferopsis*** **Perkins, 1981**

Acknowlegements

The authors kindly acknowledge assistance from the following colleagues: Dr Nomiki Simboura (HCMR, Anavyssos) and Dr Miltiadis Kitsos (Aristotelian University of Thessaloniki) for loans of comparative material of *Sphaerosyllis hystrix* and *Sphaerosyllis boeroi*, Dr Guillermo San Martín (Universidad Autónoma de Madrid) for the provision of some of the literature resources, Mrs Alexandra Siakouli (University of Crete) for taking SEM pictures, Dr Vincent Smith, Simon Rycroft, Ben Scott (NHM, London) and Dr Lyubomir Penev (Pensoft Publishers, Bulgaria) for support with the electronic Scratchpads publication. The two reviewers (Dr Guillermo San Martín, Dr Luigi Musco) are thanked for their suggestions to improve the manuscript. Financial support was provided by the ViBRANT project (FP7, EU).

References

Anderson M (2001) A new method for non-parametric multivariate analysis of variance. Austral Ecology 26: 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Álvarez P, San Martín G (2009) A new species of *Sphaerosyllis* (Annelida: Polychaeta: Syllidae) from Cuba, with a list of syllids from the Guanahacabibes Biosphere Reserve (Cuba). Journal of the Marine Biological Association of the UK 89: 1427–1435. doi: 10.1017/S0025315409000654

Banse K (1959) Über die Polychaeten-Besiedlung einiger submariner Höhlen. Ergebnisse der Österreichischen Tyrrheni- Expedition 1952, Teil XII. Pubblicazioni della Stazione Zoologica di Napoli 30: 417–469.

Barroso R, Klautau M, Solé-Cava AM, Paiva PC (2009) *Eurythoe complanata* (Polychaeta: Amphinomidae), the "cosmopolitan" fireworm, consists of at least three cryptic species. Marine Biology 157: 69–80. doi: 10.1007/s00227-009-1296-9

Ben-Eliahu MN (1977) Polychaete cryptofauna from rims of similar intertidal vermetid reefs on the Mediterranean coast of Israel and in the Gulf of Elat: Exogoninae and Autolytinae (Polychaeta Errantia: Syllidae). Israel Journal of Zoology 26: 59–99.

Berkeley E, Berkeley C (1948) Canadian Pacific Fauna. Fisheries Research Board of Canada, Toronto, 100 pp.

Blagoderov V, Brake I, Georgiev T, Penev L, Roberts D, Ryrcroft S, Scott B, Agosti D, Catapano T, Smith VS (2010) Streamlining taxonomic publication: a working example with Scratchpads and ZooKeys. ZooKeys 50: 17–28. doi: 10.3897/zookeys.50.539

Bleidorn C, Kruse I, Albrecht S, Bartolomaeus T (2006) Mitochondrial sequence data expose the putative cosmopolitan polychaete *Scoloplos armiger* (Annelida, Orbiniidae) as a species complex. BMC Evolutionary Biology 6: 47. doi: 10.1186/1471-2148-6-47

Claparède É (1863) Beobachtungen über Anatomie und Entwicklungsgeschichte wirbelloser Thiere an der Küste von Normandie angestellt. Wilhelm Engelmann Verlag, Leipzig, 172pp. http://biodiversitylibrary.org/item/40310

Claparède É (1868) Les annélides chétopodes du Golfe de Naples. Mémoires de la Société de physique et d'histoire naturelle de Genève 20: 313–584. http://biodiversitylibrary.org/item/18576

Costa-Paiva EM, Paiva PC (2007) A morphometric analyis of *Eunice* Cuvier (Annelida, Polychaeta) species. Revista Brasileira de Zoologia 24: 353–358. doi: 10.1590/S0101-81752007000200013

Day JH (1953) The Polychaete fauna of South Africa. Part 2. Errant species from Cape shores and estuaries. Annals of the Natal Museum 12: 397–441.

Del-Pilar-Ruso Y, San Martín G (in press) Description of a new species of *Sphaerosyllis* Claparède, 1863 (Polychaeta: Syllidae: Exogoninae) from the Alicante coast (W Mediterranean) and first report for the Mediterranean Sea and the Iberian Peninsula of two other species of Syllidae. Mediterranean Marine Science.

Ding Z-hu, Westheide W (2008) Interstitial Exogoninae from the Chinese coast (Polychaeta, Syllidae). Senckenbergiana Biologica 88: 125–159.

Ehlers E (1864) Die Borstenwürmer (Annelida Chaetopoda) nach systematischen und anatomischen Untersuchungen dargestellt. Wilhelm Engelmann Verlag, Leipzig, 748 pp. http://biodiversitylibrary.org/item/18348

Fauvel P (1923) Polychètes errantes. Faune de France, Paris, 488 pp. http://www.faunedefrance.org/bibliotheque/docs/P.FAUVEL(FdeFr05)Polychetes-errantes.pdf

Gower J (1971) A General Coefficient of Similarity and Some of Its Properties. Biometrics 27: 857–874. http://www.jstor.org/pss/2528823

Hartman O (1966) Quantitive survey of the benthos of San Pedro Basin, southern California. Part II. Final results and conclusions. Allan Hancock Pacific Expeditions 19: 187–455.

Hartman O (1968) Atlas of the errantiate polychaetous annelids from California. Allan Hancock Foundation, Los Angeles, 828 pp.

Hartman O, Fauchald K (1971) Deep-water benthic polychaetous annelids off New England to Bermuda and other North Atlantic areas. Part II. Allan Hancock Monographs in Marine Biology 6: 1–327. http://hdl.handle.net/10088/3458

Hartmann-Schröder G (1960) Polychaeten aus dem Roten Meer. Kieler Meeresforschungen 16: 69–125.

Hartmann-Schröder G (1984) Die Polychaeten der antiborealen Südküste Australiens (zwischen Albany im Westen und Ceduna im Osten). Teil 10. Mitteilungen aus dem Hamburgischen Zoologischen Museum und Institut 81: 7–62.

Hartmann-Schröder G (1985) Die Polychaeten der antiborealen Südküste Australiens (zwischen Port Lincoln im Westen und Port Augusta im Osten). Teil 11. Mitteilungen aus dem Hamburgischen Zoologischen Museum und Institut 82: 61–99.

Kudenov JD, Harris LH (1995) Family Syllidae Grube, 1850. In: Blake et al. (Eds) Taxonomic Atlas of the Bentic Fauna of the Santa Maria Basin and Western Santa Barbara Channel. The Annelida Part 2 - Polychaeta: Phyllodocida (Syllidae and Scale-Bearing Families), Amphinomida and Eunicida. Santa Barbara Museum of Natural History, Santa Barbara, 1–97.

Langerhans P (1879) Die Wurmfauna von Madeira. Zeitschrift für wissenschaftliche Zoologie 32: 513–592.

Men F, Ding Z-hu, Zhao J, Wu B-L (1993) A preliminary study on small syllides from the Huanghai Sea (Yellow Sea). Journal of Oceanography of Huanghai and Bohai Seas 11: 19–36.

Musco L, Giangrande A (2005) Mediterranean Syllidae (Annelida: Polychaeta) revisited: biogeography, diversity and species fidelity to environmental features. Marine Ecology Progress Series 304: 143–153. doi: 10.3354/meps304143

Musco L, Çinar ME, Giangrande A (2005) A new species of *Sphaerosyllis* (Polychaeta, Syllidae, Exogoninae) from the coasts of Italy and Cyprus (eastern Mediterranean Sea). Italian Journal of Zoology 72: 161–166. doi: 10.1080/11250000509356666

Naville A (1933) Quelques formes epitoques d'annelides polychetes nouvelles ou peu connues pechees a la lumiere dans la baie de Banyuls. Annales des sciences naturelles 16: 171–208.

Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith VS, Brake I, Ryrcroft S, Scott B, Johnson NF, Morris RA, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress WJ, Thompson FC, Erwin T (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1–16. doi: 10.3897/zookeys.50.538

Perkins TH (1981) Syllidae, principally from Florida, with descriptions of a new genus and twenty-one new species. Proceedings of the Biological Society of Washington 93: 1080–1172. http://biostor.org/reference/73932

Riser NW (1991) An evaluation of taxonomic characters in the genus *Sphaerosyllis* (Polychaeta: Syllidae). Ophelia supplement: 209–218.

San Martín G (1984) Estudio biogeográfico, faunistico y sistemático de los poliquetos de la familia Silidos (Syllidae: Polychaeta) en Baleares. PhD Thesis, Madrid, Spain: Universidad Complutense de Madrid.

San Martín G (2003) Annelida, Polychaeta II: Syllidae. In: Ramos MA et al. (Eds) Fauna Ibérica, vol. 21. Museo Nacional de Ciencias Naturales. CSIC, Madrid, 554 pp.

San Martín G (2005) Exogoninae (Polychaeta: Syllidae) from Australia with the description of a new genus and twenty-two new species. Records of the Australian Museum 57: 39–152. doi: 10.3853/j.0067-1975.57.2005.1438

San Martín G, López E (2002) New species of *Autolytus* Grube, 1850, *Paraprocerastea* San Martín & Alós, 1989, and *Sphaerosyllis* Claparède 1863 (Syllidae, Polychaeta) from the Iberian Peninsula. Sarsia 87: 135–143. doi: 10.1080/003648202320205210

Somaschini A, San Martín G (1994) Description of two new species of *Sphaerosyllis* (Polychaeta: Syllidae: Exogoninae) and first report of *Sphaerosyllis glandulata* for the Mediterranean Sea. Cahiers de Biologie Marine 35: 357–367.

Southern R (1914) Clare Island Survey. Archiannelida and Polychaeta. Proceedings of the Royal Irish Academy 31: 1–160. http://biodiversitylibrary.org/page/34773787

Temperini MI (1981) "Sistemática e Distribuição dos Poliquetos Errantes da Plataforma Continental Brasileira entre as Latitudes de 23°05'S" e 30°00'S".MSc Thesis, São Paulo, Brazil: Universidade de São Paulo.

Westheide W (1974) Interstitielle Fauna von Galapagos. XI. Pisionidae, Hesionidae, Pilargidae, Syllidae (Polychaeta). Mikrofauna des Meeresbodens 44: 1–146.

Westheide W, Hass-Cordes E (2001) Molecular taxonomy: description of a cryptic *Petitia* species (Polychaeta: Syllidae) from the Island of Mahé (Seychelles, Indian Ocean) using RAPD markers and ITS2 sequences. Journal of Zoological Systematics and Evolutionary Research 39: 103–111. doi: 10.1046/j.1439-0469.2001.00166.x

Westheide W, Schmidt H (2003) Cosmopolitan versus cryptic meiofaunal polychaete species: an approach to a molecular taxonomy. Helgoland Marine Research 57: 1–6. doi: 10.1007/s10152-002-0114-2

Supplementary material

The Scratchpads version of this publication is available at:
http://polychaetes.marbigen.org/node/35

Character matrices for specimens used in the morphological analysis are available at:
http://polychaetes.marbigen.org/content/measured-values-sphaerosyllis-specimens

Summary statistics for the species are available at:
http://polychaetes.marbigen.org/content/summary-statistics-sphaerosyllis-hystrix
http://polychaetes.marbigen.org/content/summary-statistics-sphaerosyllis-boeroi
http://polychaetes.marbigen.org/content/summary-statistics-sphaerosyllis-levantina

Results of the morphometric analysis are available at:
http://polychaetes.marbigen.org/content/morphometric-analysis-pca-eigenvectors

http://polychaetes.marbigen.org/content/spearmans-correlation-principal-component-scores-vs-measurements

Illustrations and graphs of the statistical analysis are available at:
http://polychaetes.marbigen.org/category/image-galleries/sphaerosyllis

ZooKeys 150: 347–380 (2011)
doi: 10.3897/zookeys.150.1968
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Review of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae) in Japan

Marko Prous[1,†], Mikk Heidemaa[1,‡], Akihiko Shinohara[2,§], Villu Soon[1,|]

1 *Department of Zoology, Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, 51014 Tartu, Estonia* **2** *Department of Zoology, National Museum of Nature and Science, 4-1-1 Amakubo, Tsukuba-shi, Ibaraki, 305-0005 Japan*

† *urn:lsid:zoobank.org:author:5E0EFC23-7C71-42E6-A031-BB9B09AD6C30*
‡ *urn:lsid:zoobank.org:author:7B134C49-F69A-4847-9DD1-4E512C121C61*
§ *urn:lsid:zoobank.org:author:C7382A9B-948F-479B-BEE7-848DAFECD3BA*
| *urn:lsid:zoobank.org:author:73CF0534-A70B-4BAC-A91A-0FAED197E87C*

Corresponding author: *Marko Prous* (marko.prous@ut.ee)

Academic editor: *S. Blank* | Received 28 August 2011 | Accepted 8 November 2011 | Published 28 November 2011

urn:lsid:zoobank.org:pub:3BD9FFC0-EFFC-4045-A93D-C7D9CA770467

Citation: Prous M, Heidemaa M, Shinohara A, Soon V (2011) Review of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae) in Japan. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 347–380. doi: 10.3897/zookeys.150.1968

Abstract

The following eleven *Empria* species are reported from Japan: *E. candidata* (Fallén, 1808), *E. japonica* Heidemaa & Prous, 2011, *E. liturata* (Gmelin, 1790), *E. loktini* Ermolenko, 1971, *E. plana* (Jakowlew, 1891), *E. quadrimaculata* Takeuchi, 1952, *E. rubicola* Ermolenko, 1971, *E. tridens* (Konow, 1896), *E. tridentis* Lee & Ryu, 1996, *E. honshuana* Prous & Heidemaa, **sp. n.**, and *E. takeuchii* Prous & Heidemaa, **sp. n.** The lectotypes of *Poecilosoma pallipes* Matsumura, 1912, *Empria itelmena* Malaise, 1931, *Tenthredo candidata* Fallén, 1808, and *Tenthredo* (*Poecilostoma*) *hybrida* Erichson, 1851 are designated. *Empria itelmena* Malaise, 1931, **syn. n.** is synonymized with *E. plana* (Jakowlew, 1891). *Poecilosoma pallipes* Matsumura, 1912, previously assigned to *Empria*, is transferred to *Monsoma*, creating *Monsoma pallipes* (Matsumura, 1912), **comb. n.** Results of phylogenetic analyses using mitochondrial (COI) and nuclear (ITS1 and ITS2) sequences are also provided.

Keywords

Sawflies, new species, new synonymy, key, cytochrome c oxidase I, internal transcribed spacer

Introduction

With 51 valid species-level taxa (Taeger et al. 2010; Prous et al. 2011b), *Empria* Lepeletier & Serville, in Latreille et al. 1828 is one of the largest genera in the Allantinae. Nevertheless, it still remains rather poorly studied in comparison with other tenthredinid sawflies. *Empria* species are often misidentified because of the lack of easily observable diagnostic characters. Fortunately, their genitalia frequently possess clear differences even between closely related species mostly enabling their reliable identification. Though the knowledge on most of the European *Empria* species can be regarded as satisfactory (Zhelochovtsev and Zinovjev 1988; Prous et al. 2011b), very little is known about Eastern Palaearctic species. According to Takeuchi (1952a), more than seven *Empria* species had been found in Japan, but most of them remained unidentified. Until recently, only two species had been identified (Takeuchi 1952a; b; Abe and Togashi 1989), and one of them, *Empria pallipes* (Matsumura 1912), actually belongs to *Monsoma* MacGillivray, 1908 (see results). Prous et al. (2011b) reported three additional species. Here we report 11 species from Japan, two of them described as new. One male, probably representing a new *Empria* species (sp. 1) is also discussed but not yet described as new due to insufficient material.

No attempts to reconstruct the phylogeny of *Empria* have been made so far. Some preliminary results based on a limited number of species can be found in Prous et al. (2011b), which focuses on the *E. longicornis* species group. Only few intrageneric groups have been proposed, which might be monophyletic. In particular, *Empria* is sometimes divided into the subgenera *Parataxonus* MacGillivray, 1908 [now comprising *E. candidata* (Fallén, 1808) and *E. multicolour* (Norton, 1862)] and *Empria* s. str. (all other species) (Ross 1936; Zhelochovtsev and Zinovjev 1988; 1996; Yan et al. 2009). Within *Empria* s. str., the *E. hungarica* (Konow, 1895) (see Heidemaa and Viitasaari 1999) and the *E. longicornis* (Thomson, 1871) species groups (see Prous et al. 2011b) have been proposed. In addition, the *E. immersa* species group can be defined for the species possessing highly similar penis valves, which have a characteristic long apical spine (Smith 1979; Zhelochovtsev and Zinovjev 1988; Prous et al. 2011b). To examine the phylogenetic relationships within *Empria* based on DNA sequences, we here expand the dataset of Prous et al. (2011b) by including 7 more species (six outside and one inside of the *longicornis*-group). For this, we use one continuous mitochondrial region (full COI, two complete, and one incomplete tRNAs) and one nuclear region (ITS1 and ITS2 within the rRNA locus) analysed separately and in combination using Bayesian methods.

Material and methods

Pinned specimens studied are from the following institutional collections:

BMNH Natural History Museum, London, United Kingdom (G. Broad, N. Dale-Skey Papilloud, S. Ryder, N. Springate);

CSCS Central South University of Forestry and Technology, Changsha, China (M.-C. Wei);
DEI Senckenberg Deutsches Entomologisches Institut, Müncheberg, Germany (A. Taeger, S. M. Blank, A. D. Liston);
EIHU Hokkaido University, Sapporo, Japan (M. Suwa);
HNHM Hungarian Natural History Museum, Budapest, Hungary (S. Csősz, L. Zombori);
NHRS Naturhistoriska Riksmuseet, Stockholm, Sweden (H. Vårdal);
NSMT National Museum of Nature and Science, Tokyo, Japan (A. Shinohara);
SIZ I. I. Schmalhausen Institute of Zoology, National Academy of Sciences of Ukraine, Kiev, Ukraine (I. N. Pavlusenko);
TUZ Zoological Museum of the University of Tartu, Estonia (J. Luig);
UOPJ Osaka Prefecture University, Sakai, Japan (T. Hirowatari);
USNM National Museum of Natural History, Smithsonian Institution, Washington DC, USA (D. R. Smith);
UUZM Uppsala University, Museum of Evolution, Uppsala, Sweden (H. Mejlon);
YUIC Yeungnam University Insect Collections, Gyeongsan, South-Korea (J.-W. Lee);
ZISP Zoological Institute of the Russian Academy of Sciences, St. Petersburg, Russia (S. Belokobylskij, A. Zinovjev);
ZMH Zoological Museum, Helsinki, Finland (P. Malinen);
ZML Museum of Zoology and Entomology, Lund University, Lund, Sweden (R. Danielsson);
ZMUC Zoological Museum of the University, Copenhagen, Denmark (L. Vilhelmsen).

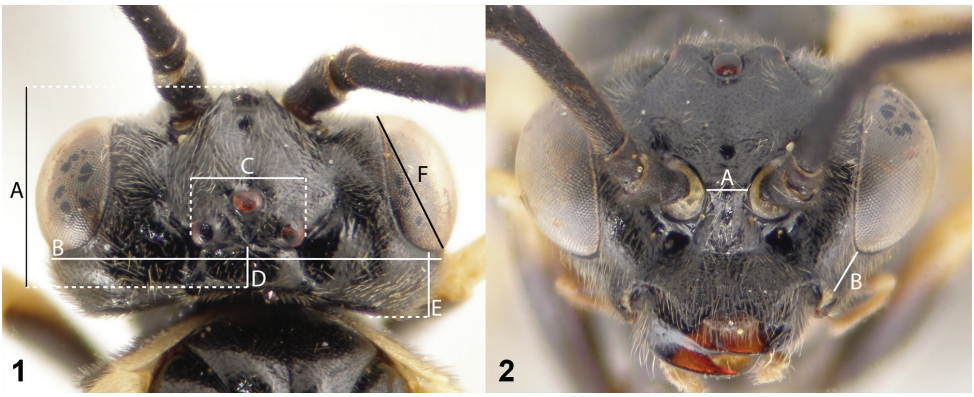
Specimens from the private collections of Erik Heibo, Guy T. Knight, and of the second author (MH) were also studied.

For morphological analyses, male penis valves, female lancets (valvula 1), and external characters of the adults were studied.

To dissect the penis valves, genital capsules were separated from the specimen and macerated in KOH or NaOH (10–15%) for 6–12 hours at room temperature, or treated with proteinase K using High Pure PCR Template Preparation Kit (Roche, Mannheim) and following manufacturer's protocol.

Imaging methods are described in Prous et al. (2011b). All images made for this study are deposited in the Morphbank database (http://www.morphbank.net/?id=592670).

Morphological terminology follows Viitasaari (2002). To differentiate between species, some distances were measured on the head capsule (Prous et al. 2011b): maximal lengths of flagellomeres, head length (Fig. 1A), head breadth behind the eyes (Fig. 1B), length between lateral margins of lateral ocelli (Fig. 1C; “breadth of postocellar area"), length of the postocellar area (Fig. 1D), head length behind the eye in dorsal view (Fig. 1E; head positioned with posterior margins of lateral ocelli



Figures 1–2. Distances measured on the head capsule. **1** *Empria quadrimaculata*, head in dorsal view, female (NSMT083) (A, head length, B, head breadth, C, breadth of the postocellar area, D, length of the postocellar area, E, minimal distance between the eye and the occipital carina = head length behind the eye, F, length of the eye) **2** *Empria quadrimaculata*, head in anterior view, female (NSMT083) (A, minimal distance between toruli, B, malar space).

and eyes aligned), length of the eye (Fig. 1F), length between toruli (antennal sockets) (Fig. 2A), maximal and minimal length of the temple (http://www.morphbank.net/?id=781392), and the length of malar space (Fig. 2B; from here on referred to as “malar space").

For molecular phylogenetic analyses, DNA sequences of the internal transcribed spacers 1 and 2 (ITS1 and ITS2), and a mitochondrial DNA (mtDNA) fragment containing tRNA-Cys, tRNA-Tyr, cytochrome c oxidase I (COI), and partial tRNA-Leu, were obtained using methods described in Prous et al. (2011b). However, because amplification of ITS2 of *Empria honshuana* sp. n. failed using the primers CAS5p8sFc and CAS28sB1d (Ji, Zhong and He 2003; Prous et al. 2011b), we used the primers AM1 (5′ TGT GAA CTG CAG GAC ACA TGA 3′) and AM2 (5′ATG CTT AAA TTT AGG GGG TAG TC 3′) (Marinucci et al. 1999; Heidemaa et al. 2004) instead. The PCR programme in this case consisted of an initial denaturing step at 95°C for 1 min, followed by 43 cycles of 20 s at 95°C, 30 s at 65–55°C (a touchdown profile was used, in which the annealing temperature decreased from 65°C to 55°C by 0.5°C every cycle), and 70 s at 68°C; the last cycle was followed by a final 7 min extension step at 68°C. For some older air-dried museum specimens, it was possible to obtain the sequences only partially. Sequences reported here have been deposited in the GenBank (NCBI) database (accession numbers JN029842–JN029898). As suggested by Chakrabarty (2010), DNA sequences from type material are here referred to as genetypes.

Boundaries of the sequenced tRNA and ITS2 genes were identified as described by Prous et al. (2011b). Phylogenetic analyses of ITS genes were performed using Bali-Phy 2.0.2 (Suchard and Redelings 2006) since this program has implementations to handle difficult-to-align sequences. In order to enhance the speed of calculation, sequences were aligned manually for detecting and fixing the conserved positions prior

to analysis with Bali-Phy. Four independent analyses were run (203 213–262 061 iterations) using the GTR + I + G[4] model. The first 10 000–60 000 iterations were discarded as "burn in" after examination of log-likelihood scores in Tracer 1.4 (available from http://beast.bio.ed.ac.uk/Tracer).

Phylogenetic analysis of the mitochondrial genes and combined analysis of the nuclear and mitochondrial genes were performed with MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) using the GTR + I + G[4] model. Mitochondrial sequences were aligned manually, and prior to phylogenetic analyses, non-coding and ambiguously aligned tRNA regions, one insertion of three base pairs in COI of *Monsoma pulveratum* (Retzius, 1783), and two to three amino acid coding codons of COI at the 3′ end (the last three codons of *E. quadrimaculata* and *E. rubicola* could not be unambiguously aligned with the last two codons of other species) were excluded. In the combined analysis we used MAP (maximum a posteriori) alignment of ITS obtained from one of the four analyses with Bali-Phy. Both mitochondrial and combined datasets were run for 5 000 000 MCMC generations, with trees and lnL's sampled at intervals of 100 generations. The first 25% of generations were discarded as "burn-in". *Monsoma pulveratum* was used to root the trees.

Data resources

The data underpinning the analyses reported in this paper are deposited in the Dryad Data Repository at doi: 10.5061/dryad.fs262s48 (Prous et al. 2011a) and at GBIF, the Global Biodiversity Information Facility, http://ipt.pensoft.net/ipt/resource.do?r=japanese_empria.

Results

Key to Japanese *Empria* and *Monsoma* (imagines)

1 Abdominal terga without pale insulated (detached) paired patches (Fig. 3); length of postocellar area more than 3.5 times diameter of lateral ocellus; first flagellomere 0.9–1 times as long as flagellomeres 2–3 combined; propleura meeting broadly in front; on hind wing cross-vein m-cu present, cell M closed; valvula 1 as in Fig. 13; Hokkaido [East Palaearctic] .... ***Monsoma pallipes***

– Abdominal terga with pale, more or less insulated paired patches (Fig. 4); length of postocellar area less than 3.0 times diameter of lateral ocellus; first flagellomere 0.4–0.7 times as long as flagellomeres 2–3 combined; propleura not meeting or meeting only narrowly in front; on hind wing cross-vein m-cu present or absent, cell M closed or open ..... ***Empria* 2**

Figures 3–6. 3 *Monsoma pallipes*, habitus in dorsal view, female (NSMT174) **4** *Empria candidata*, habitus in dorsal view, female (NSMT187) **5** *Empria candidata*, head in anterior view, female (NSMT208) **6** *Empria candidata*, head in dorsal view, female (NSMT208).

2 At least facial orbits dorsally and part of temples pale (Figs 5–6); clypeus flat without median keel; on hind wing cross-vein m-cu absent, cell M open; claws simple or with minute subbasal tooth; number of serrulae 18–21, valvula 1 as in Fig. 14; posterior margin of sternum 9 in male notched (Fig. 7), penis valve as in Fig. 25; Hokkaido [Holarctic] ..... ***E. candidata***

– Facial orbits and temples black (Figs 1–2, 9–10); clypeus with median keel (distinct mostly in anterior part of clypeus only); on hind wing cross-vein m-cu usually present, cell M usually closed; claws variable; number of serrulae 13–18(19); posterior margin of sternum 9 in male rounded (Fig. 8); penis valve different **3**

3 female ..... (female of ***E. sp.* 1** is currently unknown) **4**

– male **13**

4 Postocellar area (1.9)2.1–2.5 times wider than long (Fig.1), trochanters and trochantelli black; serrulae as in Figs 15–16; abdominal terga with 2–3 pairs of pale patches ..... ***E. quadrimaculata* group 5**

– Postocellar area 1.5–2.1 times wider than long (Figs 9–10) and / or trochanters and trochantelli pale; serrulae different (Figs 17–24); abdominal terga with 2–6 pairs of pale patches **6**

5 Abdominal terga mostly with 2 pairs of pale patches; antennae long, flagellum mostly 2.1–2.5 times longer than head breadth; in most specimens flagellomeres 1 and 2 about equally long; number of serrulae 17–19 (Fig. 15); cannot always be distinguished morphologically from *E. rubicola*; Honshu, Shikoku, Kyushu ..... ***E. quadrimaculata***

– Abdominal terga mostly with 3 pairs of pale patches; antennae short, flagellum mostly 1.9–2.2 times longer than head breadth; in most specimens flagellomere 1 longer than flgm. 2; number of serrulae 16–18 (Fig. 16); cannot always be distinguished morphologically from *E. quadrimaculata*; Hokkaido [also Sakhalin Oblast, Russia] ..... ***E. rubicola***

6 Malar space 2.2–2.5 times longer than lateral ocellus diameter and abdominal terga with 5–6 pairs of large pale patches; claws bifid; clypeus in most specimens at least distally pale; tegulae pale; serrulae as in Fig. 17; Hokkaido, Honshu (Yamagata) [East Palaearctic] ..... ***E. plana***

– Malar space 1.5–2.0 times longer than lateral ocellus diameter and abdominal terga with 2–6 pairs of small or large pale patches or malar space 1.9–2.2 times longer than lateral ocellus diameter and abdominal terga with 3 pairs of small pale patches; claws with small subbasal tooth or simple; clypeus black; tegulae black or pale; serrulae different **7**

7 Serrulae as in Figs 22–24; length of head 2.3–2.9 (2.5–3.2 in *E. tridens*) times greater than length of head behind eyes (Fig. 9); trochanters and trochantelli black or slightly pale ..... **(*E. japonica, E. loktini, E. tridens*) 11**

– Serrulae as in Figs 18–21; length of head 2.9–3.3 times greater than length of head behind eyes (Figs 1, 10) and / or trochanters and trochantelli pale **8**

8 Trochanters and trochantelli pale; tegulae completely pale **9**

– Trochanters and trochantelli black; tegulae mostly black **10**

9 Flagellum 2.2–2.4 times longer than breadth of head; abdominal terga with 3 pairs of small pale patches (Fig. 11); serrulae as in Fig. 18; Hokkaido, Honshu [East Palaearctic] ..... ***E. tridentis***

– Flagellum 1.8–2.0 times longer than breadth of head; abdominal terga with 3–4 pairs of large pale patches (Fig. 12); serrulae as in Fig. 19; Hokkaido, Honshu ..... ***E. takeuchii***

10 Basal serrulae conspicuously protruding (Fig. 20); claws simple or with minute subbasal tooth; abdominal terga with 5–6 pairs of pale patches; Hokkaido [Palaearctic] ..... ***E. liturata***

– Basal serrulae not conspicuously protruding (Fig. 21); claws with conspicuous subbasal tooth; abdominal terga with 4 pairs of pale patches; Honshu.... ..... ***E. honshuana***

11 Flagellum 2.5–2.7 times longer than breadth of head; maximal length of temple 1.40–1.55 times greater than minimal length of temple; serrulae as in Fig. 23; Hokkaido ..... ***E. japonica***

– Flagellum 1.8–2.3 times longer than breadth of head; maximal length of temple less than 1.35 times greater than minimal length of temple; serrulae as in Figs 22, 24 ..... **12**

12 Abdominal terga mostly with 5 pairs of pale patches; number of serrulae 16–18 (Fig. 22); Hokkaido [Palaearctic] ..... ***E. tridens***

– Abdominal terga mostly with 2–3 pairs of pale patches; number of serrulae 13–14(15) (Fig. 24); Hokkaido [also Sakhalin Oblast, Russia] … ***E. loktini***

13 Postocellar area (2.1)2.2–2.5 times wider than long and trochanters and trochantelli black; penis valves as in Figs 26–27 ... ***E. quadrimaculata*** **group 14**

– Postocellar area 1.7–2.1(2.2) times wider than long or trochanters and trochantelli at least partly pale; penis valves as in Figs 28–36..... **15**

14 Valviceps with small basal lobe, ventroapical part of valviceps slightly bent towards its basal part (Fig. 26); flagellum 2.9–3.3 times longer than breadth of head; in most specimens flagellomere 7 not distinctly shorter than length of eye; Honshu, Shikoku, Kyushu..... ***E. quadrimaculata***

– Valviceps with large basal lobe, ventroapical part of valviceps strongly bent towards its basal part (Fig. 27); flagellum 2.6–3.0 times longer than breadth of head; in most specimens flagellomere 7 distinctly shorter than length of eye; Hokkaido [also Sakhalin Oblast, Russia] ..... ***E. rubicola***

15 Valviceps with long apical spine (Fig. 28); malar space 1.9–2.3 times longer than lateral ocellus diameter; Hokkaido, Honshu (Yamagata) [East Palaearctic] ..... ***E. plana***

– Valviceps without long apical spine (Figs 29–36); malar space 1.3–1.8 times longer than lateral ocellus diameter **16**

16 Trochanters, trochantelli, and tegulae pale; abdominal terga mostly with 3 pairs of pale patches **17**

– Trochanters black; trochantelli black or with barely visible median pale band or patch; tegulae black or pale; abdominal terga with 2–5 pairs of pale patches ... **18**

17 Valviceps with large dorsobasally pointing spine at dorsoapical part (Fig. 29); postocellar area 1.9–2.3(2.4) times wider than long; flagellum 2.6–3.7 times longer than breadth of head; Hokkaido, Honshu [East Palaearctic] ..... ***E. tridentis***

– Valviceps with small dorsally pointing tooth at dorsoapical part (Fig. 30); postocellar area 2.0–2.7 times wider than long; flagellum 2.2–2.7 times longer than breadth of head; Hokkaido, Honshu.....***E. takeuchii***

18 Antennae short, flagellum 2.3–3.0 times longer than breadth of head.....**19**

– Antennae long, flagellum 3.2–3.8 times longer than breadth of head.....**22**

19 Valviceps with large dorsoapical spine (Figs 31–32) **20**

– Valviceps with small dorsoapical tooth (Figs 33–36) **21**

20 Dorsal margin of valviceps concave (Fig. 31); claws with minute subbasal tooth; abdominal terga with (2)3–4 pairs of pale patches; Honshu ....***E. honshuana***

– Dorsal margin of valviceps convex (Fig. 32); claws simple or with minute subbasal tooth; abdominal terga with 5 pairs of pale patches; Hokkaido [Palaearctic] ..... ***E. liturata***

21 Apical part of valvular duct extending clearly further from dorsal rim of valvura (Fig. 33); abdominal terga mostly with 2–3 pairs of pale patches; Hokkaido [also Sakhalin Oblast, Russia].....***E. loktini***

– Apical part of valvular duct reaching almost the dorsal rim of valvura or extending only slightly further from it (Fig. 34); abdominal terga mostly with 4–5 pairs of pale patches; Hokkaido [Palaearctic]..... ***E. tridens***

22 Basal lobe of valviceps short, valviceps less than 0.65 as long as valvura (Fig. 35); maximal length of temple (1.30)1.35–1.50 times greater than its minimal length; Hokkaido.....***E. japonica***

– Basal lobe of valviceps long, valviceps more than 0.8 as long as valvura (Fig. 36); maximal length of temple less than 1.35 times greater than its minimal length; Hokkaido.....***E.* sp. 1**

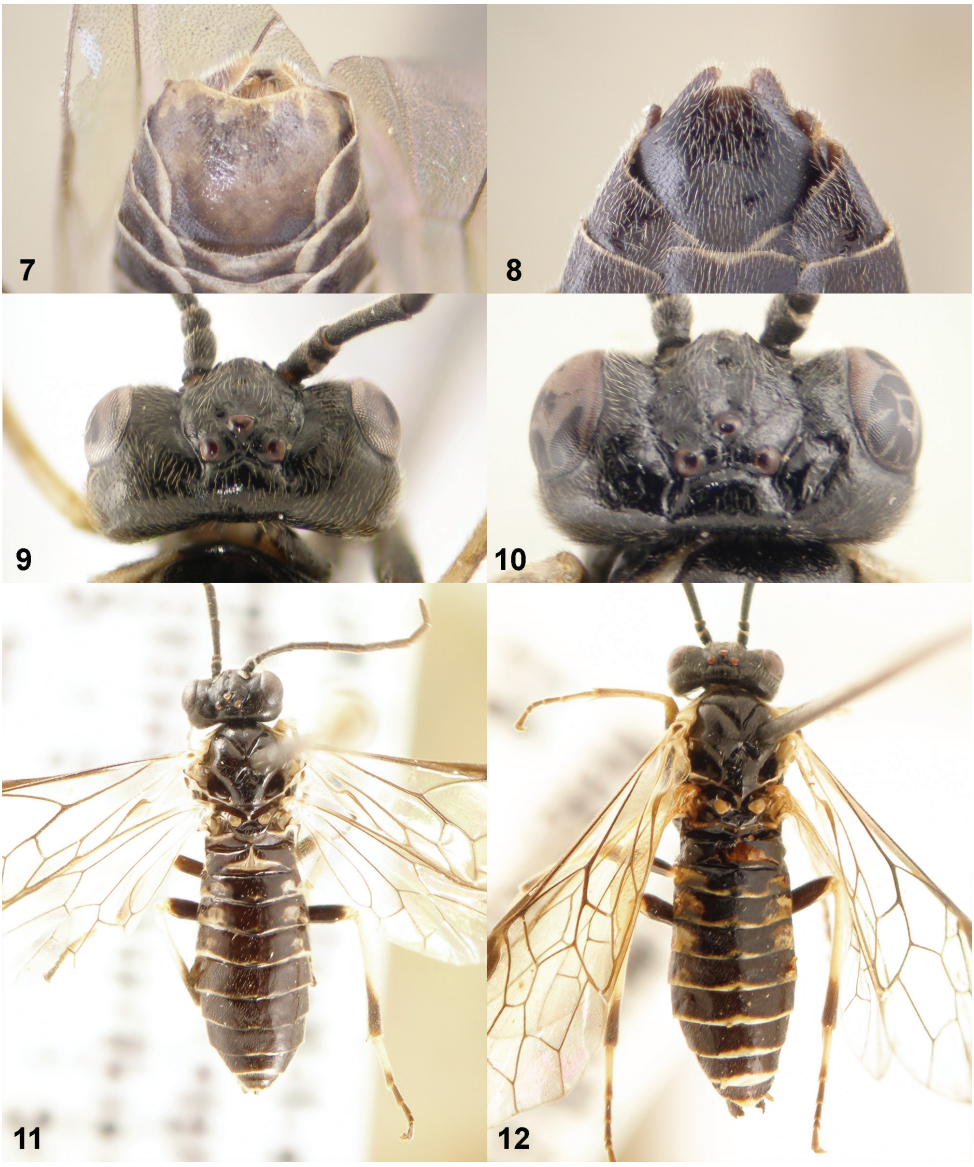
Taxonomy

***Monsoma pallipes* (Matsumura, 1912), comb. n.**
http://species-id.net/wiki/Monsoma_pallipes

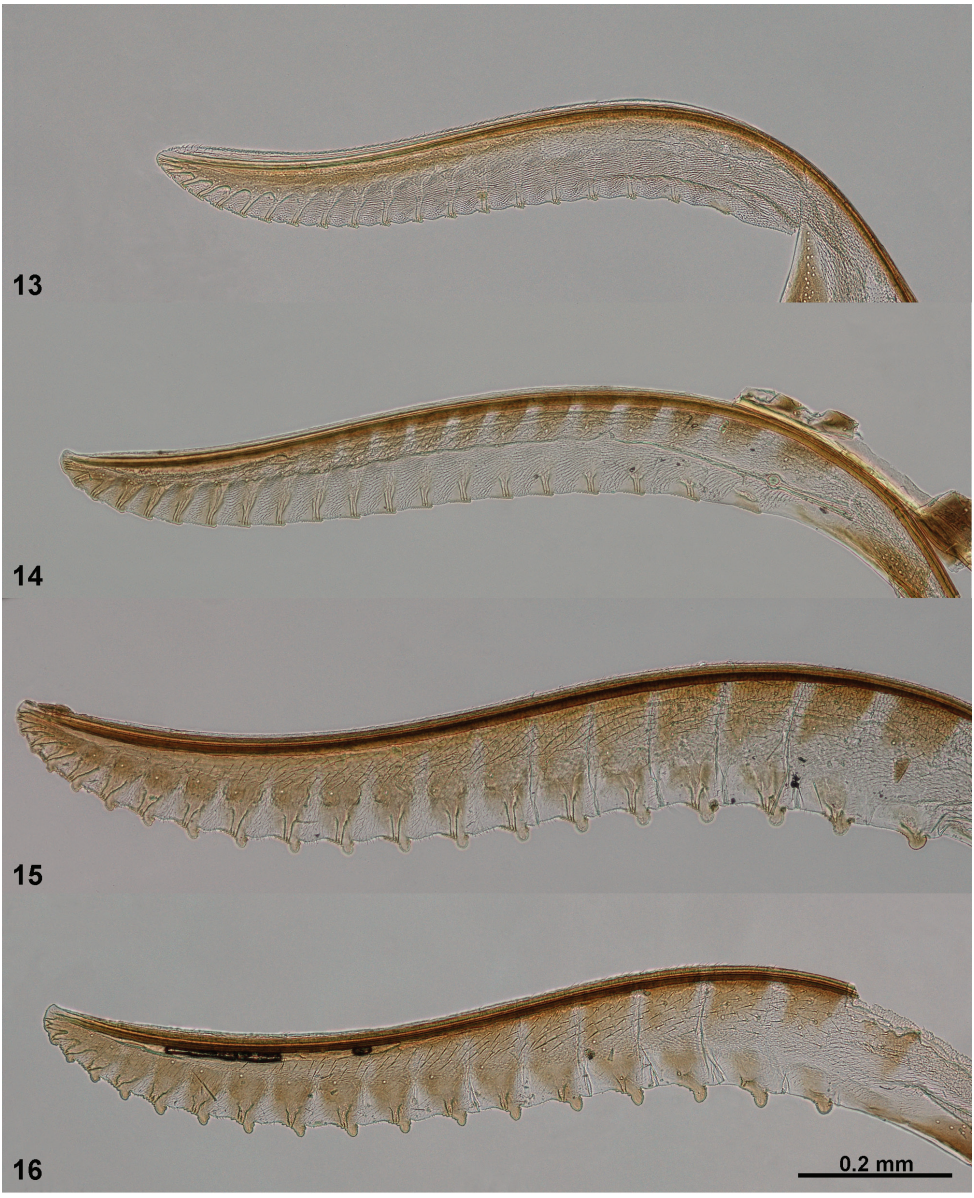
Poecilosoma pallipes Matsumura, 1912: 61–62.

Type locality. Japan, Hokkaido, Sapporo. Lectotype **(here designated)** female (Fig. 37), EIHU. Labelled: “Maruyama 5/24", “7", “Poecilosoma pallipes Mats., Type".

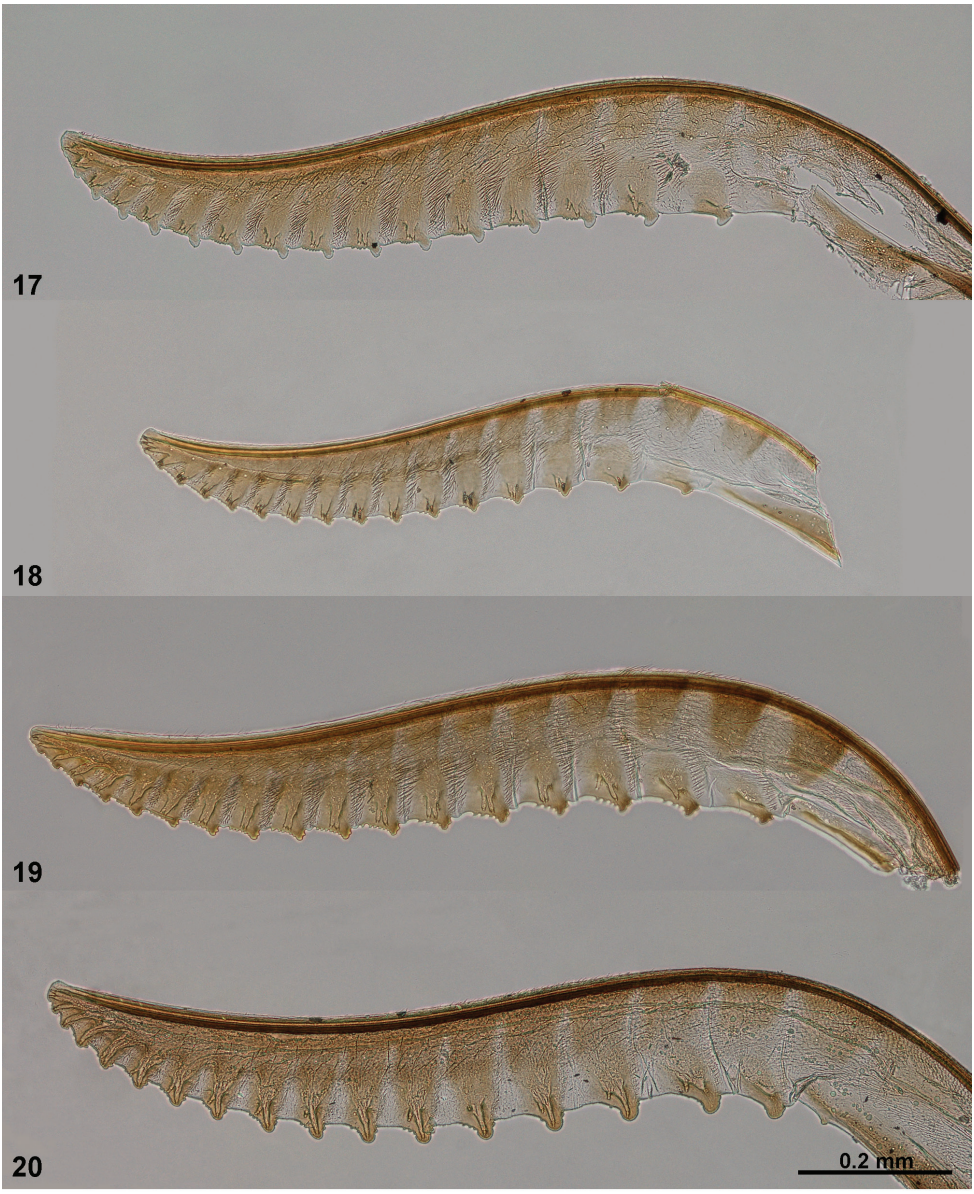
Taxonomic affinities. *Monsoma pallipes* can most easily be differentiated from the other *Monsoma* species, *M. pulveratum* (Retzius, 1783), *M. inferentium* (Norton, 1868), and *M. faustum* Zhelochovtsev, 1961, by the colouration of the head capsule: temples, genae, facial orbits, paraantennal field laterally, and area between toruli and lateral to median ocellus are pale brown in *M. pallipes*, while in the other three species the head capsule is black.



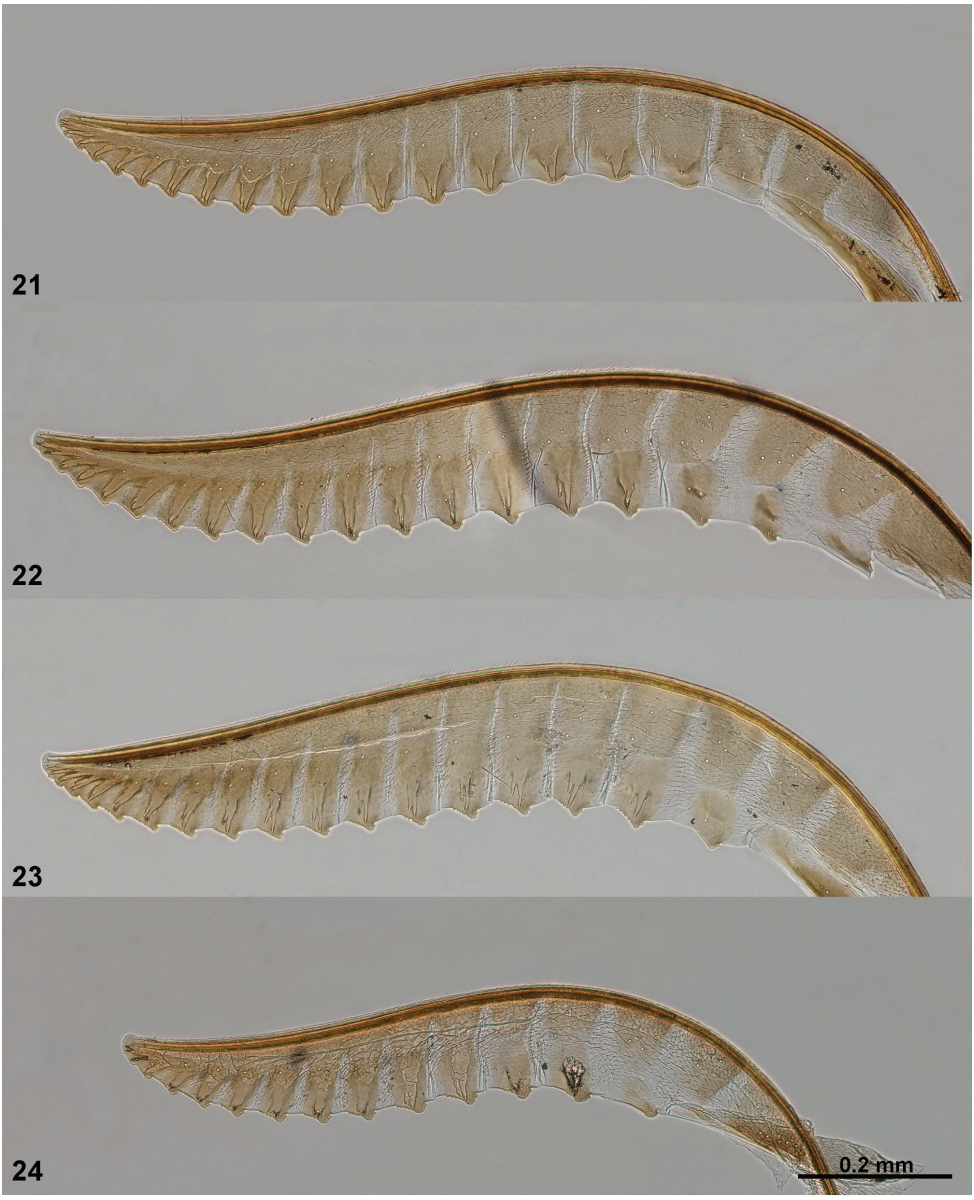
Figures 7–12. 7 *Empria candidata*, posterior tip of the abdomen in ventral view, male (TUZ282970) **8** *Empria quadrimaculata*, posterior tip of the abdomen in ventral view, male (NSMT228) **9** *Empria loktini,* head in dorsal view, female (NSMT014) **10** *Empria honshuana* sp. n., head in dorsal view, female paratype (NSMT-Hym2011-2-3-4) **11** *Empria tridentis*, habitus in dorsal view, female (NSMT051) **12** *Empria takeuchii* sp. n., habitus in dorsal view, female paratype (NSMT032).



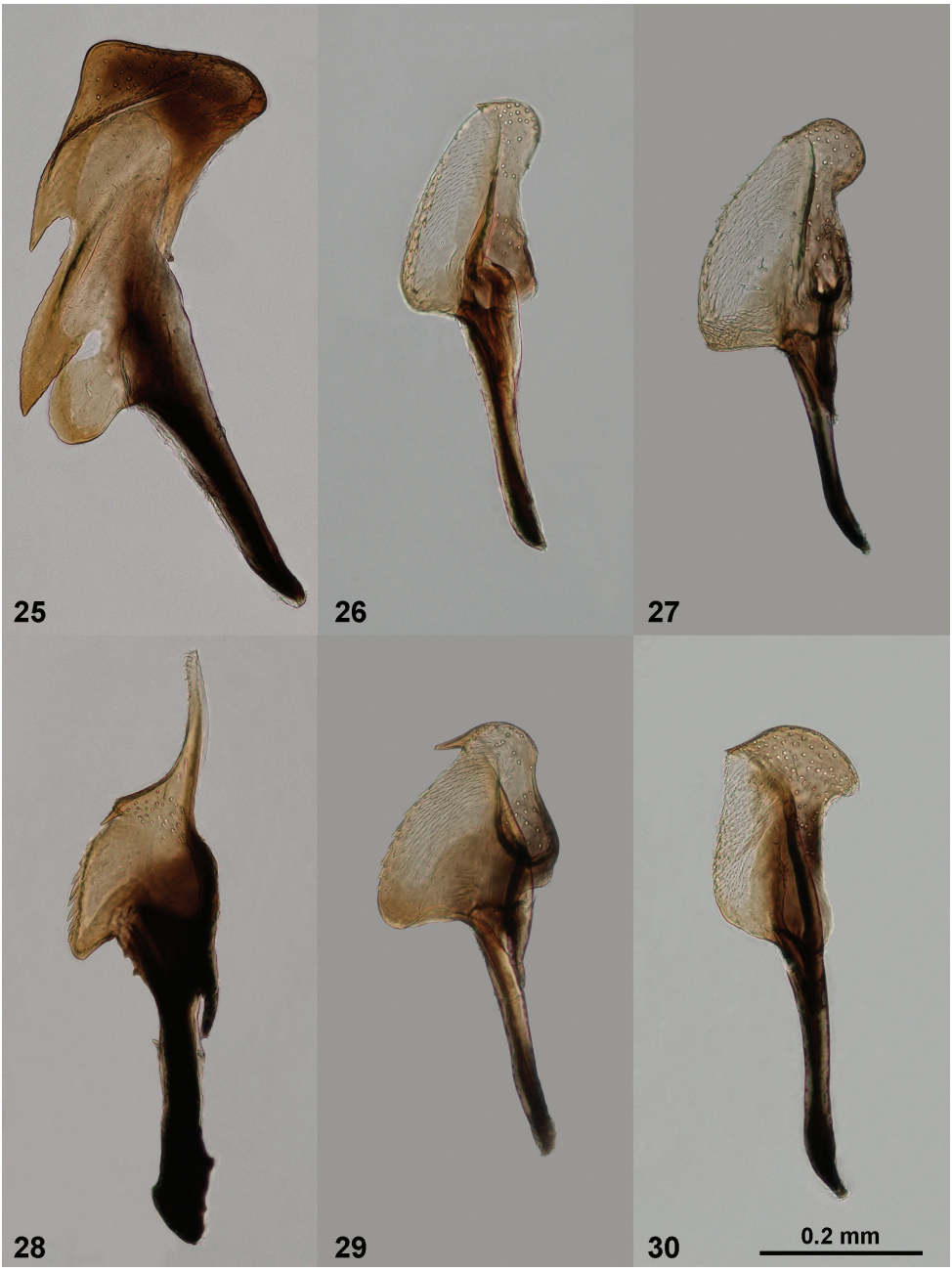
Figures 13–16. Lancets (valvulae 1) of *Monsoma* and *Empria*. **13** *Monsoma pallipes* (NSMT173) **14** *Empria candidata* (NSMT208) **15** *Empria quadrimaculata* (NSMT155) **16** *Empria rubicola* (USNM2051678_053).



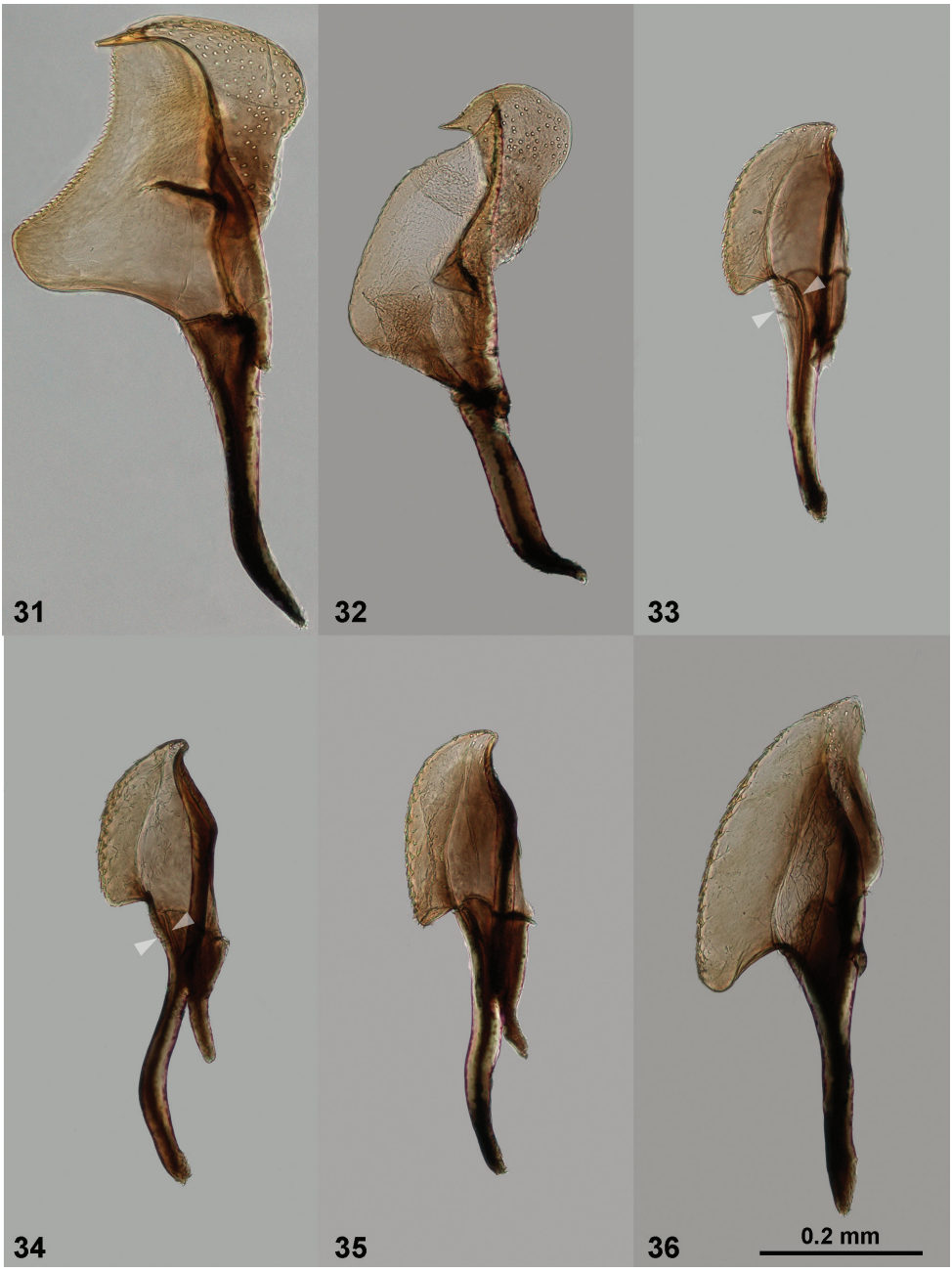
Figures 17–20. Lancets (valvulae 1) of *Empria*. **17** *Empria plana* (NSMT026) **18** *Empria tridentis* (USNM2051678_013) **19** *Empria takeuchii* sp. n., holotype (NSMT044) **20** *Empria liturata* (USNM2051678_054).



Figures 21–24. Lancets (valvulae 1) of *Empria*. **21** *Empria honshuana* sp. n., paratype (USNM2051678_016) **22** *Empria tridens* (USNM2051678_018) **23** *Empria japonica*, holotype (NSMT USNM2051678_019) **24** *Empria loktini* (TUZ615180).



Figures 25–30. Penis valves of *Empria*. **25** *Empria candidata* (NSMT036) **26** *Empria quadrimaculata* (UOPJ03) **27** *Empria rubicola* (USNM2051678_042) **28** *Empria plana* (NSMT201) **29** *Empria tridentis* (TUZ615182) **30** *Empria takeuchii* sp. n., paratype (NSMT112).



Figures 31–36. Penis valves of *Empria*. **31** *Empria honshuana* sp. n., paratype (NSMT200) **32** *Empria liturata* (USNM2051678_051) **33** *Empria loktini* (NSMT105) **34** *Empria tridens* (USNM2051678_024) **35** *Empria japonica*, paratype (NSMT009) **36** *Empria* sp. 1 (USNM2051678_040). The arrowheads illustrate the different position of valvular duct (upper right arrowhead) relative to the dorsal rim of valvura (lower left arrowhead) in *E. loktini* (Fig. 33) and other species of *longicornis*-group (Fig. 34).

Figure 37. *Monsoma pallipes,* lectotype of *Poecilosoma pallipes* Matsumura, 1912, habitus in dorsolateral view, female.

Host plants. Unknown, but could be associated with *Alnus* as for *M. pulveratum* and *M. inferentium* (Smith 1979; Pieronek 1980; Chevin 2004).

Distribution. East Palaearctic. Specimens studied are from Japan (Hokkaido) and Russia (Primorsky Krai).

Notes. Male unknown. Matsumura (1912) did not give the number of specimens he used for the original description. A female syntype bearing a red type label is hereby designated as the lectotype.

***Empria candidata* (Fallén, 1808)**
http://species-id.net/wiki/Empria_candidata

Tenthredo candidata Fallén, 1808: 105–106. **Type locality.** Sweden. Lectotype (**here designated**) female [in good condition], UUZM. Labelled: "Uppsala Univ. Zool.

Mus. Typsamlingen nr. 1940b Hymenoptera Tenthredo candidata Fallén 1808" [red, printed, partially handwritten], "♀" [pale, handwritten], "LECTOTYPUS 2008 [printed part, red label] TENTHREDO CANDIDATA FALLÉN 1808 Des. M.Heidemaa & M.Prous [handwritten part]", "*Empria* 2008 *candidata* (Fallén, 1808) ♀ M.HEIDEMAA & M.PROUS" [white, printed]. 3 paralectotype females of *Tenthredo candidata* designated ("PARALECTOTYPUS 2008 [printed part, red label] TENTHREDO CANDIDATA FALLÉN 1808 Des. M.Heidemaa & M.Prous" [handwritten part]) belong in *E. immersa* (Klug, 1818) [nr. 1940a], *E. pumila* (Konow, 1896) [nr. 1940c], and *E. fletcheri* (Cameron, 1878) [nr. 1940d] (respectively labelled by M. Heidemaa & M. Prous).

Tenthredo (Allantus) repanda Klug, 1816: 77–78.

Taxonomic affinities. The morphologically closest species is the Nearctic *E. multicolor*, from which *E. candidata* can be distinguished by the following characters: femora predominantly and most other parts of legs at least partly black (legs are almost entirely yellowish in *E. multicolor*), tarsal claws simple or with a minute inner tooth (with a long subbasal tooth in *E. multicolor*), shallowly emarginated clypeus (deeply emarginated in *E. multicolor*), and postocellar area more than 1.6 times wider than long (less than 1.5 in *E. multicolor*) (see also Smith 1979).

Host plants. *Betula* (Lorenz and Kraus 1957; Verzhutskii 1981), *B. pendula* Roth (under the name *B. verrucosa* in Verzhutskii 1966).

Distribution. Holarctic. Specimens studied are from China (Heilongjiang), Estonia, Finland, Japan (Hokkaido), Russia (Kamchatka Krai, Khabarovsk Krai, Leningrad Oblast, Primorsky Krai), South-Korea, Sweden, Switzerland, United Kingdom, USA (Maine).

Empria japonica Heidemaa & Prous, 2011

urn:lsid:zoobank.org:act:BA25596E-802D-43E3-B351-52A0BAB1B78F
http://species-id.net/wiki/Empria_japonica

Empria japonica Heidemaa & Prous in Prous et al. 2011b: 22–24. Type locality: Japan, Hokkaido, Ginsendai, Kamikawa-chô, 43°40'N, 143°01'E, 947 m, selectively cut forest. Holotype female, NSMT.

Genetype accessions in GenBank. USNM2051678_019: HM177347 (hologenetype COI), HM177397 (hologenetype ITS1), HM177299 (hologenetype ITS2); USNM2051678_009: HM177346 (paragenetype COI), HM177396 (paragenetype ITS1), HM177298 (paragenetype ITS2); USNM2051678_003: HM177345 (paragenetype COI), HM177395 (paragenetype ITS1), HM177297 (paragenetype ITS2).

Taxonomic affinities. Belongs to *E. longicornis* group (see Prous et al. 2011b). Morphologically the most similar species are *E. tridens* (Konow, 1896), *E. longicornis*, and *Empria* sp. 1, from which *E. japonica* can be distinguished by having maximal length of temple mostly more than 1.40 (in males rarely 1.30) times greater than mini-

mal length of temple (less than 1.35 in the other three species). *Empria* sp. 1 differs clearly also by its penis valve (cf. Figs 35–36).

Host plants. Unknown, but could be *Rubus idaeus* L. subsp. *melanolasius* (Dieck) Focke (see Prous et al. 2011b).

Distribution. Japan (Hokkaido).

Empria honshuana Prous & Heidemaa, sp. n.

urn:lsid:zoobank.org:act:AF95BFA0-C12F-46AB-B50B-8A8CA18B34CF
http://species-id.net/wiki/Empria_honshuana

Type-locality. Japan, Honshu, Tochigi Prefecture, Bicchuzawa, Bato, Nakagawa.

Holotype. 1 female, NSMT. Labelled: "[JAPAN: Honshu] Bicchuzawa, Bato, Nakagawa, Tochigi 13. IV. 2006 S. Ibuki", "NSMT110", "Holotypus ♀ *Empria honshuana* spec. nov. design. : M. Prous & M. Heidemaa 2011", "*Empria honshuana* sp.n. Prous & Heidemaa det. 2011".

Paratypes. "[JAPAN:Honshu] Hikagezawa Mt. Takao-san Tokyo 21. IV. 1996 A. Shinohara", 1 female, NSMT073 (NSMT); "[JAPAN: Honshu] Bicchuzawa, Bato Tochigi Pref. 9. IV. 2005 A. Shinohara" 24 males, NSMT109, NSMT115, NSMT121–137, NSMT166–170 (NSMT), 1 male TUZ615362 (TUZ); "[JAPAN: Honshu] Bicchuzawa, Bato Tochigi Pref. 23. IV. 2005 A. Shinohara" 1 male, NSMT171 (NSMT); "[JAPAN: Honshu] Bicchuzawa, Bato Tochigi Pref. 29. IV. 2005 A. Shinohara" 1 female, TUZ615361 (TUZ); "[JAPAN:Honshu] Annaigawa, nr Mt. Takao-san Tokyo 17. IV. 1994 A.&T.Shinohara" 1 female, NSMT198, 2 males, NSMT120, NSMT200 (NSMT); "[JAPAN:Honshu] Akigase-koen Saitama Pref. 14. IV. 1996 A. Ta., N. & To. Shinohara" 1 female, NSMT204 (NSMT); "[JAPAN: Honshu] Bicchuzawa, Bato Nakagawa, Tochigi 13. IV. 2006 S. Ibuki" 1 male, NSMT106 (NSMT); "[JAPAN:Honshu] Bicchuzawa Bato, Tochigi 1. V. 2010 S. Ibuki" 1 female, NSMT-Hym2011-2-3-4 (NSMT); "JAPAN: Chiba Pref. Okusa-cho, Wakaba-shi 35°36.5'N, 140°11.6E' 23 March 1997 O. S. Flint, Jr." 1 female, USNM2051678_016 (USNM); "JAPAN: Honshu Himuro-machi Utsunomyia-shi Tochiji-ken [Utsunomiya-shi Tochigi-ken], Mal. 2-15.IV.2009, Mal. trap Takeyuki Nakamura leg." 1 male, USNM2057434_04 (USNM).

Genetype accessions in GenBank. NSMT106: JN029870 (paragenetype COI), JN029890 (paragenetype ITS1), JN029854 (paragenetype ITS2); NSMT-Hym2011-2-3-4: JN029891 (paragenetype ITS1); USNM2051678_016: JN029871 (paragenetype COI), JN029892 (paragenetype ITS1).

Female. Body length. 6.0–6.9 mm.

Colour. Black; following parts unpigmented, pale: apical maxillary palpomeres; posterodorsal margin of pronotum in lateral parts; tegulae (except lateroproximal part); median band or patch of pro-, meso-, and metatrochantellus; profemur apically; protibia in anterior and partly posterior aspects; mesotibia partly in anterior and posterior aspects; metatibia basally; tarsomere 1 of hind leg basally; paired patches

on abdominal terga 2–5; at least partially posterior margins of terga (tergum 10 dorsally more widely) and sterna; and cenchri. Labrum from yellowish-brown to blackish.

Head. Head behind eyes in dorsal view subparallel sided; postocellar area trapeziform, its length equal to or longer than 2 times diameter of lateral ocellus; distinct and diverging lateral postocellar furrows going from ocelli towards occiput at least to the distance of ocellus diameter; area between frontal crests clearly exceeding the level of crests in dorsal view; postocellar area with indistinct punctures and interspaces, more or less glossy; punctures more regular on temples and postocular area, face with more irregular punctures; wrinkled interspaces more prominent on frontal area; clypeus with rough irregular punctures, more or less fused; ocellar and postocellar area convex, slightly raised; clypeus tridentate with median keel distinct mostly in anterior part of clypeus only, median tooth smaller than lateral teeth; malar space about equal to or shorter than distance between antennal sockets; frontal ridge V-shaped; pit in central part of frontal field present; median ocellus surrounded by groove, with short distinct longitudinal furrow anteriorly, and with similar but mostly less distinct furrow posteriorly. Maximal length of temple 1.2–1.4 times greater than its minimal length; flagellum 1.9–2.0 times longer than breadth of head.

Thorax. Mesoscutellum, mesoscutellar appendage, and metapostnotum more or less glossy, almost impuctate or with indistinct shallow punctures; metascutellum with irregular fine punctures; punctures on mesoscutum more evident on lateral and anterior regions of the median lobes, fading towards central regions; mesepisternal punctures variable between specimens, from rather weak with intespaces almost glossy to more distinct with sculptured, interspaces; mesepimeron with setae on posterior part; metepisternum with evenly distributed setae; metepimeron in central part without setae; distance between cenchri 1.1–1.4 times of cenchrus width; wings hyaline, venation brownish, becoming paler near junction to thorax; closed cell M in hindwing present; tarsal claws with conspicuous subbasal tooth.

Abdomen. Terga on most parts with transverse keel-like sculpticells and with short setae (about half of lateral ocellus diameter), sometimes with shallow punctures at median parts of terga 2–4; posterior parts of terga (6) 7–9 (occasionally terga 3–10) at median line with small more or less triangular pale regions; ventral margin of valvula 3 slightly bending towards apex, slightly longer than valvifer 2; serrulae of valvula 1 as in Fig. 21, number of serrulae 15–16.

Male. (Mostly the differences compared to female are given).

Body length. 4.8–5.6 mm.

Colour. Unpigmented, whitish or yellowish brown: anterolateral (seldom also posterolateral) margins of tegulae; protibia in anterior aspect, often partly also in posterior aspect; mesotibia partly in anterior aspect; outer margins of harpes; and paired patches on abdominal terga 2–(3)/4/(5).

Head. Area between frontal crests reaching or slightly exceeding the level of crests in dorsal view; malar space less than or equal to distance between antennal sockets;

length of postocellar area about 2 times of lateral ocellus diameter; maximal length of temple 1.25–1.45 times greater than its minimal length; flagellum 2.3–2.6 times longer than breadth of head.

Thorax. Distance between cenchri variable, up to 2 times width of cenchrus. Tarsal claws with minute subbasal tooth.

Abdomen. Tergum 8 with indistinct tergal hollows which form semioval or semicircular depression reaching 1/3–1/2 of tergum length and sometimes possessing indefinite central procidentia. Posterior margin of sternum 9 round; penis valve as in Fig. 31.

Taxonomic affinities. Based on the similarities in penis valves, the closest species is *E. sulcata* Wei & Nie, 1998 from China (see http://www.morphbank.net/?id=643394). While the penis valves of both species can easily be distinguished, the distinctly concave dorsal margin of valviceps of these species is a unique characteristic within *Empria*. Serrulae of the two species are clearly different (cf. Fig. 21 and http://www.morphbank.net/?id=700325). Externally the species can mainly be distinguished by colouration: in *E. sulcata* tegulae are completely pale and legs extensively yellowish, while in *E. honshuana* tegulae are at least partly and legs predominantly black.

Host plants. Unknown.

Distribution. Japan (Honshu).

Etymology. The species name refers to the type locality, Honshu, the main island of Japan.

Empria liturata (Gmelin, 1790)

http://species-id.net/wiki/Empria_liturata

Tenthredo liturata Gmelin, 1790: 2668. Type locality: Europe [type specimens probably lost (Blank et al. 2009: 13)].

Poecilosoma undulata Konow, 1885: 122. Type locality: Czech Republic, Altvater. Syntype female, DEI [examined].

See Taeger et al. (2010) for full list of synonyms.

Taxonomic affinities. The most similar species morphologically appears to be Nearctic *E. ignota* (Norton, 1867). The clearest differences between these species can be seen in the structure of penis valves (Fig. 32; http://www.morphbank.net/?id=694564).

Host plants. *Filipendula ulmaria* (L.) Maxim., *Geum rivale* L. (based on ex ovo rearings by MP in Estonia). *Fragaria vesca* has also been suggested (Enslin 1914), but this requires confirmation.

Distribution. Palaearctic. Specimens studied are from Belgium, Croatia, Czech Republic, Denmark, Estonia, France, Germany, Hungary, Italy, Japan (Hokkaido), Russia (Leningrad Oblast), Switzerland, United Kingdom.

***Empria loktini* Ermolenko, 1971**
http://species-id.net/wiki/Empria_loktini

Empria loktini Ermolenko, 1971: 22–23. Type locality: Russia, Sakhalin Oblast, Novoaleksandrovsk. Holotype female, SIZ [examined].

Taxonomic affinities. Belongs to *E. longicornis* group, morphologically the closest is *Empria basalis* Lindqvist, 1968, which can be distinguished from *E. loktini* by clearly different penis valves, lancets (see Prous et al. 2011b), and in most cases also by some external differences (in *E. loktini* metatibia is pale in basal 1/3 and the abdominal terga bear 2–3 pairs of pale patches, in *E. basalis* metatibia is mostly black and the terga have 4–5 pairs of pale patches).

Host plants. Unknown.

Distribution. East Palaearctic. Specimens studied are from Japan (Hokkaido) and Russia (Sakhalin Oblast).

***Empria plana* (Jakowlew, 1891)**
http://species-id.net/wiki/Empria_plana

Tenthredo (Poecilostoma) hybrida Erichson in: Ménétriés in: Middendorff, 1851: 60–61. Primary homonym of *Tenthredo (Tenthredo) hybrida* Eversmann, 1847. Type locality: Udskoj Ostrog [Russia, Khabarovsk Krai, Udskoe]. Lectotype **(here designated)** female, ZISP. Labelled: "Poecilostoma hybrida* Erichs. Midd. R." [pale, handwritten], "Lectotypus ♀ *Tenthredo (Poecilostoma) hybrida* Erichson, 1851 design. : M.Prous & M.Heidemaa 2011" [red, printed], "*Empria plana* (Jakovlev 1891) det. M.Prous 2008" [white, printed].

Poecilosoma plana Jakowlew, 1891: 31. Type locality: Russia, Irkutsk. Holotype female, ZISP [examined].

Empria itelmena Malaise, 1931: 23, **syn. n.** Type locality: Kamtschatka, E[lisowo] [Russia, Kamchatka Krai]. Lectotype **(here designated)** female, NHRS. Labelled: "Kamtschatka Malaise", "E", "Typus", "Lectotypus ♀ *Empria itelmena* Malaise, 1931 design. : M. Prous & M. Heidemaa 2011" [red, printed], "*Empria plana* (Jakovlev 1891) det. M.Prous 2009" [white, printed].

Empria erichsoni Liston, 1995: 241. New name for *Tenthredo (Poecilostoma) hybrida* Erichson, 1851.

Taxonomic affinities. Morphologically the closest species is *E. immersa* (Klug, 1818), from which *E. plana* can be distinguished by differences in the structure of serrulae (Fig. 17; http://www.morphbank.net/?id=694567) and penis valves (Fig. 28; http://www.morphbank.net/?id=578888). Externally, the *E. plana* specimens from mainland Asia differ clearly from *E. immersa* also by their pale clypeus (black in *E. im-*

mersa), which is, however, only partly pale or nearly black in Japanese specimens. In this regard, some disagreements concerning the taxonomic status of *E. plana* should also be noted. Some authors treat this taxon either as a geographical form, or as a subspecies of *E. immersa* (Verzhutskii 1966; Zhelochovtsev and Zinovjev 1996), but Lindqvist (1972) argues that *E. plana* (under the name *Empria hybrida* Erichson, 1851) is a separate species (followed also by Taeger et al. 2010). Because of the above mentioned differences between these two taxa, we concur with Lindqvist (1972) in treating them as distinct species. Such conclusion is supported also by current nuclear sequence data (Fig. 38).

Host plants. Possibly *Salix* sp., see Verzhutskii (1966; 1981) under the name *Empria immersa*.

Distribution. East Palaearctic. Specimens studied are from Japan (Hokkaido, Honshu), Mongolia, and Russia (Amur Oblast, Irkutsk Oblast, Kamchatka Krai, Khabarovsk Krai, Primorsky Krai).

Empria quadrimaculata Takeuchi, 1952

http://species-id.net/wiki/Empria_quadrimaculata

Empria quadrimaculata Takeuchi, 1952b: 49–50. Type locality: Japan, Kyoto, Ushio. Holotype female, UOPJ [examined].

Taxonomic affinities. The closest species are *E. zhangi* Wei & Yan, 2009 (China) and *E. rubicola* Ermolenko, 1971. *Empria zhangi* (two females and two males studied, including the holotype) can be distinguished from *E. quadrimaculata* mainly by the following two characters: 1) in female malar space clearly less than two times of the lateral ocellus diameter (about two times in *E. quadrimaculata* and *E. rubicola*), in male equal or slightly less than the ocellus diameter (clearly longer in *E. quadrimaculata* and *E. rubicola*); and 2) in female flagellum about 2.0 times longer than breadth of head (2.1–2.5 times in *E. quadrimaculata*), in male 2.4–2.5 times (2.9–3.3 times in *E. quadrimaculata*). *Empria rubicola* has shorter antennae and three pairs of pale patches (mostly two in *E. quadrimaculata*) on terga. The penis valves of *E. zhangi* and *E. quadrimaculata* are very similar (http://www.morphbank.net/?id=693502; Fig. 26), while *E. rubicola* can be distinguished from the two by relatively large basal lobe of the valviceps and by the ventroapical part clearly bent towards its basal part (Fig. 27). Valvula 1 appears indistinguishable in all three species.

Host plants. Okutani (1954) indicated *Geum japonicum* Thunb., but noted later that the specific identity of the reared *Empria* species was uncertain (Okutani 1967).

Distribution. Japan (Honshu, Shikoku, Kyushu).

***Empria rubicola* Ermolenko, 1971**
http://species-id.net/wiki/Empria_rubicola

Empria rubicola Ermolenko, 1971: 21–22. Type locality: Russia, Sakhalin Oblast, Novoaleksandrovsk. Holotype female, SIZ [examined].

Taxonomic affinities. The closest species are *E. zhangi* and *E. quadrimaculata* (see under *Empria quadrimaculata* Takeuchi, 1952 for details).

Host plants. Unknown. Holotype female and the studied paratypes (1 female, 2 males) were collected from *Rubus idaeus* L. subsp. *melanolasius* (Dieck) Focke (under the name *Rubus sachalinensis* in Ermolenko 1971), which is a common plant in Hokkaido.

Distribution. East Palaearctic. Specimens studied are from Japan (Hokkaido) and Russia (Sakhalin Oblast). Most probably this species has to be removed from the list of Chinese species (Yan et al. 2009), because *E. rubicola* has clypeus and upper half of the mesepisternum black (not yellow brown) and abdominal terga 2–4 (not 2–6) each with a pair of pale patches.

***Empria takeuchii* Prous & Heidemaa, sp. n.**
urn:lsid:zoobank.org:act:BDE02124-C81A-4705-91F4-34B40134B0C1
http://species-id.net/wiki/Empria_takeuchii

Type-locality. Japan, Honshu, Yamanashi Prefecture, Utsukushinomori, Yatsugatake Mts.

Holotype. 1 female, NSMT. Labelled: "[JAPAN:Honshu] Utsukushinomori 1500–1700m Yatsugatake Mts. Yamanashi Pref. 5–8. VI. 2000 A. Shinohara", "NSMT044", "Holotypus ♀ *Empria takeuchii* sp. n. design. : M. Prous & M. Heidemaa 2011", "*Empria takeuchii* sp.n. Prous & Heidemaa det. 2011".

Paratypes. "Shimashima Nagano Pref 16. V. 1984 A. Shinohara", 1 female, NSMT032 (NSMT); "[JAPAN:Honshu] Kamiange, Mt. Jinba Tokyo 27. IV. 2003 A. Shinohara", 1 male, NSMT037 (NSMT); "Ōmi, Ō hara [Ōhara] Kyoto Pref. 15. V. 1984 R. Inagawa", 1 female, NSMT041 (NSMT); "[Ōmi, Ōhara] Sakyo-ku, Kyoto Kyoto Pref. May, 14, 1984 T. Matsumoto leg." 1 female, NSMT211 (NSMT); "[JAPAN: Honshu] Yokotemichi, ca. 850m 35-22-39N 133-31-21E Mt. Daisen Tottori Pref. 28-29. IV. 2007 A. Shinohara", 1 male, NSMT112 (NSMT); "Takihata Kawachi-Nagano Osaka 22. IV. 1981 A. Shinohara", 1 male, NSMT213 (NSMT); "JAPAN: Ishikawa Pref., Mt. Shiritaka 637 m, May 19 1979 D. Smith & I. Togashi" 1 female, USNM2051678_047 (USNM); "JAPAN: Honshu Tamozawa, Nikkô-shi Tochigi-ken, Mal. trap 13-27.iv.2009 Takeyuki Nakamura leg.", 1 male, USNM2057434_03 (USNM).

Other material examined. "JAPAN, Hokkaido Ginsendai, Kamikawa-chô 43°40'N, 143°01'E, 947 m Selectively cut forest 6–27.vi.2008 Mal. trap, A. Ueda leg" 1 female, USNM2051678_011 (USNM); "JAPAN, Hokkaido Sekihoku-tôge, Kamikawa-chô, natural forest, 993 m 43°40'N, 143°06'E, 6–27.vi.2008 Mal. trap, A. Ueda leg." 3 males, USNM2051678_008, USNM2051678_031, USNM2051678_061 (USNM); "42°57'N,141°14'E Hakken-zan Sapporo, Hokkaidō JAPAN 16.v.2009 Takuma YOSHIDA leg." 2 males, USNM2057434_06, USNM2057434_07 (USNM).

Female. Body length. (5.1)6.4–6.9 mm.

Colour. Black; following parts more or less unpigmented, whitish or yellowish brown: labrum; apical maxillary and labial palpomeres; tegulae completely; posterodorsal margin of pronotum in lateral part rather widely, upper part of posterolateral margin of pronotum quite narrowly; pro-, meso-, and metacoxa apically; pro-, meso-, and metatrochanter partly or in most part; pro-, meso-, and metatrochantellus partly or completely; profemur in anterior, posterior, and lateral aspects; mesofemur and metafemur apically slightly; protibia in anterior and posterior aspects; mesotibia in most part; metatibia in basal 2/3; tarsomere 1 of hind leg in basal 2/3; paired patches on abdominal terga 2–4(5); posterior margins of terga and sterna; and cenchri (in one female only posterior margin).

Head. Head behind eyes in dorsal view subparallel sided; postocellar area trapeziform, its length mostly less than or equal to 2 times of lateral ocellus diameter; area between frontal crests in dorsal view reaches or slightly exceeds the level of crests; face and clypeus with somewhat irregular punctures, less shining compared to vertex and especially to postocellar area; ocellar and postocellar area at least slightly raised; clypeus tridentate, with median tooth smaller than lateral teeth; clypeus with median keel; malar space (minimal ventro-ocular distance) shorter or equal to distance between antennal sockets; frontal ridge "V"-shaped, central part of frontal field with distinct pit; maximal length of temple 1.25–1.4 times greater than its minimal length; flagellum 1.8–2.0 times longer than breadth of head.

Thorax. Anterior part of mesoscutum with more or less distinct punctures, its median and postero-lateral portions in most part with sparse indistinct punctures and glossy interspaces, or almost impunctate, glossy; mesoscutellum, mesoscutellar appendage, and metapostnotum impunctate and glossy; mesepisternum with more or less indistinct punctures, mostly glossy; mesepimeron with setae on posterior part; metepisternum with evenly distributed setae; metepimeron in central part without setae; distance between cenchri in most specimens about equal to cenchrus width, but sometimes slightly greater; wings hyaline with brownish venation; closed cell M in hindwing present; tarsal claws with conspicuous subbasal tooth.

Abdomen. Terga mostly with keel-like (sometimes mixed with scale-like) sculpticells and short setae (about half of lateral ocellus diameter); ventral margin of valvula 3 abruptly bending towards apex, about equal in length to valvifer 2; serrulae of valvula 1 as in Fig. 19, number of serrulae (15)16–17.

Male. (Mostly the differences compared to female are given).

Body length. 5.6–5.8 mm.

Colour. Unpigmented, whitish or yellowish are: meso- and metatrochanter apically; pro-, meso-, and metatrochantellus partly; mesofemur only apically, or in anterior, posterior, and lateral aspects; metafemur apically; mesotibia partly in anterior, posterior, and lateral aspects, or in most part; metatibia in basal 1/3 or in basal 1/2; outer margins of harpes; paired patches on abdominal terga 2–4(3).

Head. Area between frontal crests in dorsal view not exceeding the level of crests; length of postocellar area 1.5–2.0 times of lateral ocellus diameter; maximal length of temple 1.25–1.45 times greater than its minimal length; flagellum 2.2–2.7 times longer than breadth of head.

Abdomen. Posterior margin of sternum 9 round; tergum 8 without tergal hollows and procidentia; penis valve as in Fig. 30.

Taxonomic affinities. Morphologically, no certain closest relative can be specified. Superficially may resemble *E. rubicola* (based on males), *E. honshuana* (based on females), or *E. tridentis* (both have pale trochanters and trochantelli). Penis valve (Fig. 30) and valvula 1 (Fig. 19) clearly distinguish this species from all other known species of *Empria*. According to the molecular analyses (of ITS1 and ITS2 combined with mtDNA sequences), the closest species are those of the *E. longicornis* and *E. immersa* species groups, and *E. tridentis* (Figs 38, 40).

Host plants. Unknown.

Distribution. Japan (Hokkaido, Honshu).

Etymology. The specific name refers to Kichizo Takeuchi (1892–1968), who made great contributions to the sawfly systematics in eastern Asia.

Notes. Six additional studied specimens (1 female, 5 males) from Hokkaido were not included in the type series. The female and most of the males have a longer postocellar area (more than 2 times of the lateral ocellus diameter) compared to the specimens from Honshu (mostly less than 2 times). Serrulae of the Hokkaido female are also slightly different (cf. http://www.morphbank.net/?id=693521 and Fig. 19). No clear differences were found in the structure of penis valves between the specimens from Hokkaido and Honshu.

***Empria tridens* (Konow, 1896)**

http://species-id.net/wiki/Empria_tridens

Poecilosoma (Poecilosoma) tridens Konow, 1896: 54, 58. Type locality: Europe "Europa fere tota" [original description]. Lectotype female (designated in Prous et al. 2011b), DEI [examined].

Empria (Empria) caucasica Dovnar-Zapolskij, 1929: 38–39. Synonymy according to Conde (1940), see Prous et al. (2011b) for details.

Empria (Triempria) konowi Dovnar-Zapolskij, 1929: 39–40. Type locality: Russia, Sarepta. Lectotype female (designated in Prous et al. 2011b), ZISP [examined].

Empria (Triempria) gussakovskii Dovnar-Zapolskij, 1929: 40–41. Type locality: Russia, Kostroma District. Lectotype female (designated in Prous et al. 2011b), ZISP [examined].

Taxonomic affinities. Belongs to *E. longicornis* group. Morphologically the closest species is *E. longicornis*, from which it can be distinguished in most cases by shorter antennae and more pairs of pale patches on abdominal terga (4 large and 1 small in *E. tridens*, on terga 2–6; 3 large and 1 small in *E. longicornis*, on terga 2–5), and by its less prominent serrulae (Fig. 22; http://www.morphbank.net/?id=578850).

Host plants. *Rubus idaeus* and possibly *Rubus fruticosus* complex (Prous et al. 2011b).

Distribution. Palaearctic. Specimens studied are from Belgium, Croatia, Denmark, Estonia, Finland, France, Germany, Hungary, Japan (Hokkaido), Mongolia, Russia (Amur Oblast, Kamchatka Krai, Kostroma Oblast, Leningrad Oblast, Primorsky Krai, Sakhalin Oblast, Stavropol Krai, Volgograd Oblast), Sweden, Switzerland, Turkey, Ukraine, and United Kingdom.

Empria tridentis Lee & Ryu, 1996

http://species-id.net/wiki/Empria_tridentis

Empria tridentis Lee & Ryu, 1996: 23. Type locality: South-Korea, Goseong-gun Hyangnobong, 38.3167N 128.3E. Holotype female, YUIC [examined].

Taxonomic affinities. Morphologically, no close relatives can be identified, but in the phylogenetic analysis of the ITS and mtDNA sequences combined, the species appears as a sister of the *longicornis*-group (Fig. 40). Superficially may resemble *E. longicornis*, from which *E. tridentis* can easily be distinguished by tegulae, base of metatibia, trochanters, and trochantelli pale (all black in *E. longicornis*), and by very different structure of lancets and penis valves.

Host plants. Unknown.

Distribution. East Palaearctic. Specimens studied are from Japan (Hokkaido, Honshu), Russia (Khabarovsk Krai, Primorsky Krai), and South-Korea.

Notes. The original description of this species states that there are “a pair of large flecks on lateral portion of lst–4th tergite" (Lee and Ryu 1996), while actually no specimen studied (including the holotype) has pale patches (“large flecks") on first tergite. There is one male (NSMT018) from Honshu (Nagano) with penis valve slightly different (see http://www.morphbank.net/?id=592669) from all the other studied males, but the material is currently insufficient to decide if the specimen is aberrant or represents a separate (sibling) species.

Empria sp. 1

Taxonomic affinities. Belongs to *E. longicornis* group. Externally it is most similar to *E. japonica*, but penis valve is clearly distinct from all other known species of the *longicornis*-group (Fig. 36), being most similar to *E. alpina* Benson, 1938 (e.g. http://www.

morphbank.net/?id=577439). Can be distinguished from *E. alpina* by its colouration: in *E.* sp1 tegulae, posterior margin of pronotum, and basal 1/3 of metatibia are pale, while in *E. alpina* these are mostly black. Distinctness of this taxon is also supported by nuclear ITS sequence data (Fig. 38).

Host plants. Unknown.

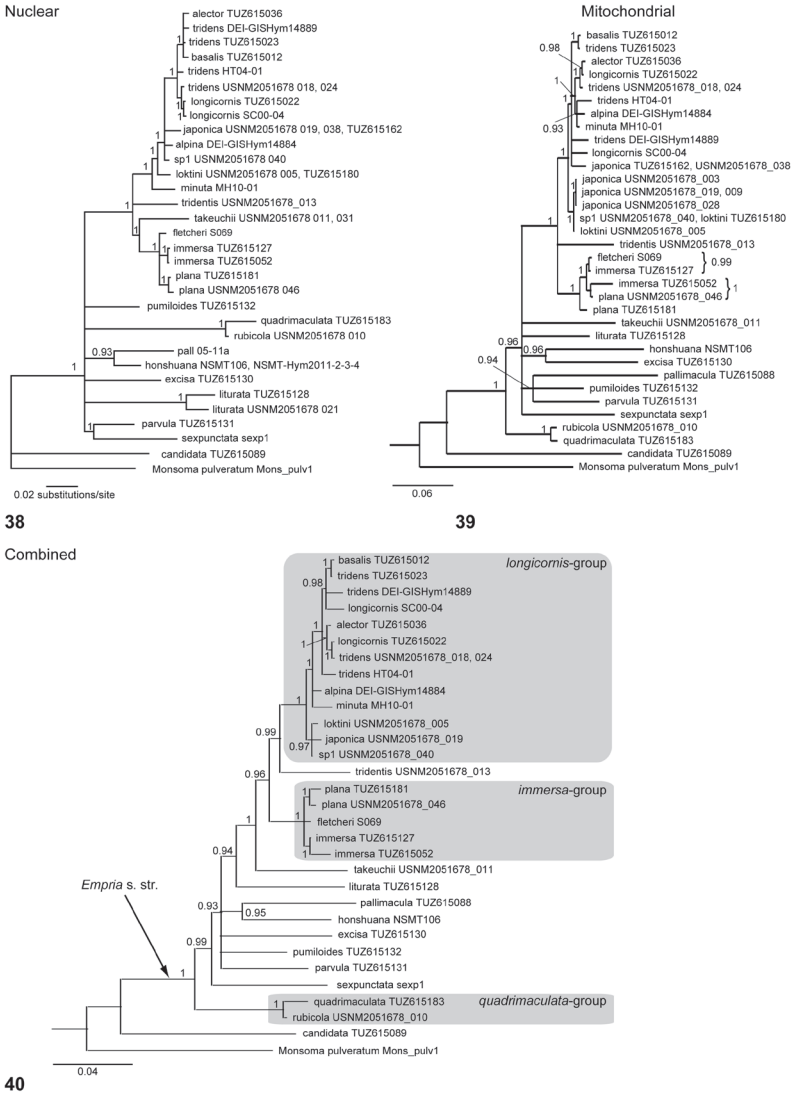
Distribution. Japan (Hokkaido).

Notes. Because taxonomy of the *longicornis*-group is quite difficult (Prous et al. 2011b) and the corresponding female remains to be found yet, additional material is needed to describe and name this presumably new species.

Molecular phylogenetic analyses

Bayesian analyses of the mitochondrial and nuclear sequences separately and in combination all resulted in somewhat different topologies (Figs 38–40), with well supported differences in some cases (especially in the *longicornis* and the *immersa*-groups). However, several clades were reconstructed in all analyses with significant statistical support (posterior probability 0.95 or more). Based on these analyses, the basal split within the genus *Empria* is between *E. candidata* and all other species (Figs 38–40), which is consistent with the division of the genus into two subgenera, *Parataxonus* MacGillivray, 1908 (*E. candidata*) and *Empria* s. str. (Zhelochovtsev and Zinovjev 1988; 1996; Yan et al. 2009). Monophyly of the *immersa*-group, the *longicornis*-group, and the *quadrimaculata*-group is well supported in all our analyses (Figs 38–40). *Empria quadrimaculata* species group is proposed here for the first time for the species sharing the same type of lancets (Figs 15–16; http://www.morphbank.net/?id=693500) and penis valves (Figs 26–27; http://www.morphbank.net/?id=693502). A clade comprising the *longicornis*-group and the *immersa*-group, *E. tridentis*, and *E. takeuchii* is well supported in the analysis of nuclear ITS and in the combined analysis of ITS and the mitochondrial sequences (Figs 38, 40). In the analysis of the mitochondrial DNA sequences, however, *E. takeuchii* is excluded from this clade, but without significant support for any other sister-group relationships within *Empria* s. str. (Fig. 39). The sister group of *E. honshuana*, revealed in the analyses of ITS and the combined sequences, is *E. pallimacula* (Figs 38, 40), but according to the mitochondrial sequences, it is *E. excisa* (Fig. 39).

Each of *E. japonica*, *E. loktini*, *E. longicornis, E. immersa*, and *E. plana* is monophyletic (as would be expected from morphology) according to the ITS sequences (Fig. 38), but not according to the mitochondrial DNA (Fig. 39). The monophyly of *Empria tridens* is supported neither by ITS nor the mitochondrial sequences (Figs 38–39; see discussion in Prous et al. 2011b). Remarkably, *Empria* sp. 1 (USNM2051678_040) has an identical mitochondrial haplotype with one specimen of *E. loktini* (TUZ615180), while morphology (cf. Figs 33 and 36, see also the key) and the nuclear ITS sequences (Fig. 38) clearly differentiate these species.



Figures 38–40. Phylogenetic analyses of the genus *Empria*. **38** Phylogeny of ITS sequences (1298–1517 bp) reconstructed using BAli-Phy (GTR + I + G[4] substitution model). Because the four independent runs of BAli-Phy produced different topologies, only clades which were found in all trees and were supported with posterior probabilities (PP) 0.9 or more are shown. Duplicate (shown behind the sequence used in the analysis) and very similar sequences (three *E. japonica*, two *E. tridentis*, and one *E. rubicola*) were removed prior to analyses to reduce computation time. **39** Phylogeny of mitochondrial sequences using MrBayes (GTR + I + G[4] model; alignment length 1642 bp). Duplicate sequences (shown behind the sequence used in the analysis) were removed prior to analyses. *Empria liturata* from Japan (USNM2051678_021) was also excluded due to incomplete sequence. **40** Combined analysis of ITS (MAP alignment from BAli-Phy analysis) and mitochondrial sequences using MrBayes (GTR + I + G[4] model). *Monsoma pulveratum* was used as an outgroup. Clades with posterior probabilities (PP) less than 0.9 were collapsed in all the trees.

Discussion

Although identification of *Empria* species using only external morphology can often be difficult, we found that females of the species reviewed here can mostly be identified without dissecting their ovipositors. Identification of the males is much less reliable without studying their genitalia because of more extensive intraspecific variation and less pronounced differences among species. The most difficult species to separate from each other on the basis of female characters are *E. quadrimaculata* and *E. rubicola*, the ovipositors of which appear nearly indistinguishable (Figs 15–16). Also the external characters applied in the present key overlap considerably between them. However, because there are consistent differences in the penis valves between the two (see Figs 26–27), they most likely represent different species.

Due to the general difficulty in identifying the *Empria* species using only external morphology, it is advisable in our opinion to leave the specimens unidentified (to avoid possible confusions in the future), especially those from the poorly studied regions (e.g. Eastern and Central Asia), as long as their identity remains problematic from external morphology and the genitalia cannot be dissected.

In addition to the 11 named *Empria* species and one presumably new but undescribed species (currently only one male is known) reported here, some additional species of the genus are likely to be found in Japan. Alpine habitats above the tree line might be inhabited by additional *Empria* species, but from there we have no samples yet.

The results of our molecular phylogenetic analyses (Figs 38–40) significantly supported the groupings within *Empria* that could be expected from morphology (*Empria* s. str., *immersa*-group, *longicornis*-group, and *quadrimaculata*-group). Although *E. pumiloides* was the only species from the *hungarica*-group in the current dataset, monophyly of this group is also supported by DNA data (unpublished results). The consistent affinity found between the *longicornis*-group, the *immersa*-group, and *E. tridentis* in all our analyses (Figs 38–40) was the only phylogenetic result not expected from morphology (though phylogenetic analyses using morphological data are still lacking). Based on the phylogenetic results presented here, we cannot draw any more definite conclusions regarding the phylogeny of *Empria*, which require, in addition to improving taxon and gene sampling, possibly also methodological advancements (e.g. using methods which take into account incomplete lineage sorting; Heled and Drummond 2010). The conflict between ITS and mitochondrial phylogenies within the *E. longicornis* and the *E. immersa* species groups (Figs 38–39; see also Prous et al. 2011b) needs further study as well (e.g. sequencing 1–3 additional nuclear markers). However, we note that incongruence between mitochondrial phylogeny with morphology and nuclear phylogeny is not uncommon among closely related species, possibly because of mitochondrial introgression (e.g. Linnen and Farrell 2007; Wahlberg et al. 2009; Near et al. 2011). Another explanation, which we cannot exclude based on current data, might be incomplete lineage sorting (for a review, see Degnan and Rosenberg 2009).

Acknowledgements

We would like to thank Sergey A. Belokobylskij and Alexey G. Zinovjev (ZISP), Olof Biström and Pekka Malinen (ZMH), Stephan M. Blank (DEI), Sándor Csősz and Lajos Zombori (HNHM), Roy Danielsson (ZML), Toshiya Hirowatari (UOPJ), Jong-Wook Lee (YUIC), Andrew Liston (DEI), Ole Martin and Lars Vilhelmsen (ZMUC), Hans Mejlon (UUZM), Inna N. Pavlusenko and Valery A. Korneyev (SIZ), Suzanne Ryder, Natalie Dale-Skey Papilloud, Gavin Broad (BMNM), David R. Smith (USNM), Masaaki Suwa (EIHU), Andreas Taeger (DEI), Hege Vårdal (NHRS), and Meicai Wei (CSCS) for loaning us material (including type specimens) from institutional collections. Stephan M. Blank clarified for us localities of some of the types. Madli Pärn and Andro Truuverk are thanked for technical assistance. Kauri Mikkola (ZMH) kindly commented on the new species names. Comments and suggestions by Stephan M. Blank, Toomas Tammaru (University of Tartu) and two anonymous reviewers helped to improve the manuscript.

The study was financially supported by the Estonian Science Foundation grant nr. 6598 to MH, the Estonian Ministry of Education and Science (target-financing project number 0180122s08) and the European Union through the European Regional Development Fund (Center of Excellence FIBIR).

References

Abe M, Togashi I (1989) Hymenoptera [Symphyta]. In: Hirashima Y (Ed) A checklist of Japanese Insects vol. 2. Entomology Laboratory, Faculty of Agriculture, Kyushu University, Fukuoka, 541–560.

Benson RB (1938) A revision of the British sawflies of the genus *Empria* Lepeletier (Hymenoptera, Symphyta). Transactions of the Society of British Entomology 5: 181–198.

Blank SM, Taeger A, Liston AD, Smith DR, Rasnitsyn AP, Shinohara A, Heidemaa M, Viitasaari M (2009) Studies toward a World Catalog of Symphyta (Hymenoptera). Zootaxa 2254: 1–96. http://www.mapress.com/zootaxa/2009/1/zt02254p096.pdf

Cameron P (1878) The fauna of Scotland, with special reference to Clydesdale and the Western District. Hymenoptera. Part I. Proceedings (& Transactions) of the Natural History Society of Glasgow 3[1875–1878](Suppl.): 1–52.

Chakrabarty P (2010) Genetypes: a concept to help integrate molecular phylogenetics and taxonomy. Zootaxa 2632: 67–68. http://www.mapress.com/zootaxa/2010/f/zt02632p068.pdf

Chevin H (2004) Biologie de *Monsoma pulveratum* (Retzius, 1783) (Hym, Symphyta, Tenthredinidae, Emphytinae). Bulletin des Naturalistes des Yvelines 31: 81–85.

Conde O (1940) Eine Revision der mir bekannten *Empria*-Arten (Hym. Tenthr.) und einige Bemerkungen zum Wesen der systematischen Forschungsarbeit. Deutsche Entomologische Zeitschrift [1940] (1–4): 162–179.

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution 24: 332–340. doi: 10.1016/j.tree.2009.01.009

Dovnar-Zapolskij DP (1929) Einige neue oder wenig bekannte Arten der Gattung *Empria* Lep. (Hymenoptera), mit einer Bestimmungstabelle der paläarktischen Arten. Russkoe Entomologicheskoe obozrenie 23: 37–47.

Enslin E (1914) Die Tenthredinoidea Mitteleuropas III. Deutsche Entomologische Zeitschrift 1914 (Beiheft): 203–309.

Erichson WF (1851) [Die neuen Arten der Hymenoptera, Diptera & Neuroptera]. In: Ménétriés E: Die Insekten (außer Parasiten). In: Middendorff AT von (1851–1853) Reise in den äussersten Norden und Osten Sibiriens während der Jahre 1843 u. 1844 auf Veranstaltung der Kaiserl. Akademie der Wissenschaften zu St. Petersburg, Vol. 2(1). Imperatorskaya Akademiya nauk, St. Petersburg, 60–69.

Ermolenko VM (1971) Novye vidy i rod pililshchikov (Hymenoptera, Tenthredinidae) s ostrova Sakhalin. Soobshhenie I. Vestnik zoologii 1971(5): 18–24.

Eversmann E (1847) Fauna hymenopterologica volgo-uralensis exhibens Hymenopterorum species quas in provinciis Volgam fluvium inter et montes Uralenses sitis observavit et nunc descripsit. Bulletin de la Société Impériale des Naturalistes de Moscou 20: 3–68.

Fallén CF (1808) Försok till uppställning och beskrifning å de i Sverige fundne Arter af Insect-Slägtet Tenthredo Linn. Kongl. Vetenskaps Academiens nya Handlingar 29: 98–124.

Gmelin JF (1790) Caroli a Linné Systema Naturae. 13. ed. Lipsiae 1: 2225–3020.

Heidemaa M, Nuorteva M, Hantula J, Saarma U (2004) *Dolerus asper* Zaddach, 1859 and *Dolerus brevicornis* Zaddach, 1859 (Hymenoptera: Tenthredinidae), with notes on their phylogeny. European Journal of Entomology 101: 637–650. http://www.eje.cz/pdfarticles/736/eje_101_4_637_Heidemaa.pdf

Heidemaa M, Viitasaari M (1999) Taxonomy of the *Empria hungarica* species-group (Hymenoptera, Tenthredinidae) in Northern Europe. Entomologica Fennica 10: 95–101.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 27: 570–580. doi: 10.1093/molbev/msp274

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755. doi: 10.1093/bioinformatics/17.8.754

Jakowlew A (1891) Diagnoses Tenthredinidarum novarum ex Rossia Europaea, Sibiria, Asia Media et confinium. Trudy Russkago Entomologicheskago Obshhestva 26[1892]: 1–62 [Separatum, preprint]

Ji Y-J, Zhang D-X, He L-J (2003) Evolutionary conservation and versatility of a new set of primers for amplifying the ribosomal internal transcribed spacer regions in insects and other invertebrates. Molecular Ecology Notes 3: 581–585. doi: 10.1046/j.1471–8286.2003.00519.x

Klug F (1816) Die Blattwespen nach ihren Gattungen und Arten zusammengestellt. Der Gesellschaft Naturforschender Freunde zu Berlin Magazin für die neuesten Entdeckungen in der gesammten Naturkunde 8[1814](1): 42–84.

Klug F (1818) Die Blattwespen nach ihren Gattungen und Arten zusammengestellt. Der Gesellschaft Naturforschender Freunde zu Berlin Magazin für die neuesten Entdeckungen in der gesammten Naturkunde 8[1814](4): 273–307.

Konow FW (1885) Bemerkungen über einige Blattwespengattungen. Wiener entomologische Zeitung 4: 117–124.

Konow FW (1895) Neue oder wenig bekannte Tenthrediniden und eine analytische Übersicht der Gattung Holcocneme Knw. Természetrajzi Füzetek 18: 50–57.

Konow FW (1896) Verschiedenes aus der Hymenopteren-Gruppe der Tenthrediniden. Wiener entomologische Zeitung 15: 41–59.

Latreille PA, Lepeletier de Saint-Fargeau A, Serville AJG, Guérin-Méneville FÉ (1828) Entomologie, ou Histoire naturelle des Crustacés, des Arachnides et des Insectes. In: Encyclopédie méthodique Histoire naturelle, 1825–1828, Vol 10(2) [ed Latreille]. Agasse, Paris, 345–833.

Lee J-W, Ryu S-M (1996) A Systematic Study on the Tenthredinidae (Hymenoptera: Symphyta) from Korea II. Ten new species of the Tenthredinidae. Entomological Research Bulletin 22: 17–34.

Lindqvist E (1968) Die *Empria*-Arten Finnlands (Hymenoptera, Symphyta). Notulae Entomologicae 48: 23–33.

Lindqvist E (1972) Zur Nomenklatur und Taxonomie einiger Blattwespen (Hymenoptera, Symphyta). Notulae Entomologicae 52: 65–77.

Linnen CR, Farrell BD (2007) Mitonuclear discordance is caused by rampant mitochondrial introgression in *Neodiprion* (Hymenoptera: Diprionidae) sawflies. Evolution 61: 1417–1438. doi: 10.1111/j.1558-5646.2007.00114.x

Liston AD (1995) A replacement name for *Empria hybrida* (Erichson, 1851) (Hymenoptera: Tenthredinidae). Entomologist's Gazette 46: 241–241.

Lorenz H, Kraus M (1957) Die Larvalsystematik der Blattwespen (Tenthredinoidea und Megalodontoidea). Abhandlungen zur Larvalsystematik der Insekten, Berlin 1: 1–339.

MacGillivray AD (1908) Emphytinae - new genera and species and synonymical notes. The Canadian Entomologist 40: 365–369. doi: 10.4039/Ent40365-10

Malaise R (1931) Entomologische Ergebnisse der schwedischen Kamtchatka-Expedition 1920–1922. (35. Tenthredinidae). Arkiv för Zoologie 23A: 1–68.

Marinucci M, Romi R, Mancini P, M., Severini C (1999) Phylogenetic relationships of seven palearctic members of the *maculipennis* complex inferred from ITS2 sequence analysis. Insect Molecular Biology 8: 469–480. doi: 10.1046/j.1365-2583.1999.00140.x

Matsumura S (1912) Thousand insects of Japan. Supplement IV. Keiseisha, Tokyo, 247 pp.

Near TJ, Bossu CM, Bradburd GS, Carlson RL, Harrington RC, Hollingsworth PR, Keck BP, Etnier Da (2011) Phylogeny and Temporal Diversification of Darters (Percidae: Etheostomatinae). Systematic Biology 60: 565–595. doi: 10.1093/sysbio/syr052

Norton E (1862) Catalogue of American species of Tenthredo, as arranged by Hartig. Proceedings of the Boston Sciety of Natural History 9[1862–1863]: 116–122.

Norton E (1867) Catalogue of the described Tenthredinidae and Uroceridae of North America. Transactions of the American Entomological Society 1: 225–280.

Norton E (1868) Catalogue of the described Tenthredinidae and Uroceridae of North America. Transactions of the American Entomological Society 2: 211–236.

Okutani T (1954) Studies on Symphyta (I). Symphyta of Sasayama with description of a new species. The Science Reports of the Hyogo University of Agriculture, Agricultural Biology 1: 75–80.

Okutani T (1967) Food-plants of Japanese Symphyta (II). Japanese Journal of Applied Entomology and Zoology 11: 90–99. doi: 10.1303/jjaez.11.90

Pieronek B (1980) On the larval *Monosoma pulverata* (Retzius) feeding on alder (Hymenoptera: Tenthredinidae-Blennocampinae). Mitteilungen aus dem Zoologischen Museum in Berlin 56: 195–199.

Prous M, Heidemaa M, Shinohara A , Soon V (2011a) Data from: Review of the sawfly genus Empria (Hymenoptera, Tenthredinidae) in Japan. Dryad Data Repository. doi: 10.5061/dryad.fs262s48; and GBIF, the Global Biodiversity Information Facility, http://ipt.pensoft.net/ipt/resource.do?r=japanese_empria.

Prous M, Heidemaa M, Soon V (2011b) *Empria longicornis* species group: taxonomic revision with notes on phylogeny and ecology (Hymenoptera, Tenthredinidae). Zootaxa 2756: 1–39. http://www.mapress.com/zootaxa/2011/f/zt02756p039.pdf

Retzius AJ (1783) Caroli De Geer (…) Genera et species insectorum e generosissimi auctoris scriptis extraxit, digessit, latine quoad partem reddidit, et terminologiam insectorum Linneanam addidit. Siegfried Lebrecht Crusium, Lipsiae, i-vi, 7–220, 1–32 pp.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574. doi: 10.1093/bioinformatics/btg180

Ross HH (1936) The sawfly genus *Empria* in North America (Hymenoptera, Tenthredinidae). The Pan-Pacific Entomologist 12: 172–178.

Smith DR (1979) Nearctic sawflies. IV. Allantinae: Adults and larvae (Hymenoptera: Tenthredinidae). Technical Bulletin, US Department of Agriculture, Washington 1595: 1–172.

Suchard MA, Redelings BD (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 22: 2047–2048. doi: 10.1093/bioinformatics/btl175

Taeger A, Blank SM, Liston AD (2010) World Catalog of Symphyta (Hymenoptera). Zootaxa 2580: 1–1064. http://www.mapress.com/zootaxa/2010/1/zt02580p1064.pdf

Takeuchi K (1952a) A generic classification of the Japanese Tenthredinidae (Hymenoptera: Symphyta). Kyoto, 90 pp.

Takeuchi K (1952b) New and unrecorded sawflies from Shikoku, Japan (II) (Hymenoptera: Symphyta). Transactions of the Shikoku Entomological Society 3: 47–54.

Thomson CG (1871) Hymenoptera Scandinaviae - Tenthredo et Sirex Lin. H. Olsson, Lundae, 342 pp.

Wei M-C, Nie H-Y (1998) Hymenoptera: Pamphiliidae, Cimbicidae, Argidae, Diprionidae, Tenthredinidae, Cephidae. In: Wu, H. (Ed.) Insects of Longwangshan Nature Reserve. China Forestry Publishing House, Beijing, 344–391.

Verzhutskii BN (1966) Pilil'shhiki Pribajkal'ja. Nauka, Moskva, 162 pp.

Verzhutskii BN (1981) Rastitel'nojadnye nasekomye v jekosistemah Vostochnoj Sibiri (pilil'shhiki i rogohvosty). Nauka, Novosibirsk, 303 pp.

Viitasaari M (2002) The suborder Symphyta of the Hymenoptera. In: Viitasaari M (Ed) Sawflies I (Hymenoptera, Symphyta) A review of the suborder, the Western Palaearctic taxa of Xyeloidea and Pamphilioidea. Tremex Press, Helsinki, 11–174.

Wahlberg N, Weingartner E, Warren AD, Nylin S (2009) Timing major conflict between mitochondrial and nuclear genes in species relationships of *Polygonia* butterflies (Nymphalidae: Nymphalini). BMC Evolutionary Biology 9: 92. doi: 10.1186/1471-2148-9-92

Yan Y-C, Wei M-C, He Y-K (2009) Two new species of Tenthredinidae (Hymenoptera, Tenthredinidae) from China. Acta Zootaxonomica Sinica 34: 248–252.

Zhelochovtsev AN (1961) Novye i maloizvestnye pilil'shhiki (Hymenoptera, Symphita) Tjan'-Shanja. Sbornik trudov zoologicheskogo muzeja MGU 8: 117–138.

Zhelochovtsev AN (=Zhelohovcev AN), [Zinovjev AG] (1988) 27. Otrjad Hymenoptera – Pereponchatokrylye Podotrjad Symphyta (Chalastogastra) – Sidjachebrjuhie. In: Zhelohovcev AN, Tobias VI, Kozlov MA (Eds) Opredelitel' nasekomyh evropejskoj chasti SSSR T III Pereponchatokrylye Shestaja chast' (Opredeliteli po faune SSSR, izdavaemye Zoologicheskim institutom AN SSSR; Vyp 158). Nauka, Leningrad, 7–234.

Zhelochovtsev AN, Zinovjev AG (1996) A list of the sawflies and horntails (Hymenoptera, Symphyta) of the fauna of Russia and adjacent territories. II. Entomologicheskoe obozrenie 75: 357–379..

ZooKeys 150: 381–395 (2011)
doi: 10.3897/zookeys.150.1566
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Cambrian archaeocyathan metazoans: revision of morphological characters and standardization of genus descriptions to establish an online identification tool

Adeline Kerner[1], Françoise Debrenne[2], Régine Vignes-Lebbe[3]

1 *CNRS (UMR 7207, centre de recherche sur la paléobiodiversité et les paléoenvironnements), Laboratoire Informatique et Systématique, MNHN Département Histoire de la Terre, Bâtiment de Géologie, CP48, 57 rue Cuvier, 75005 Paris (France)* **2** *13 rue du long foin, 91700 Sainte Geneviève des bois (France)* **3** *UPMC, université Pierre et Marie Curie, (UMR 7207, centre de recherche sur la paléobiodiversité et les paléoenvironnements), Laboratoire Informatique et Systématique, MNHN Département Histoire de la Terre, Bâtiment de Géologie, CP48, 57 rue Cuvier, 75005 Paris (France)*

Corresponding author: *Adeline Kerner* (kerner@mnhn.fr)

Academic editor: *L. Penev* | Received 14 May 2011 | Accepted 14 September 2011 | Published 28 November 2011

Citation: Kerner A, Debrenne F, Vignes-Lebbe R (2011) Cambrian archaeocyathan metazoans: revision of morphological characters and standardization of genus descriptions to establish an online identification tool. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 381–395. doi: 10.3897/zookeys.150.1566

Abstract

Archaeocyatha represent the oldest calcified sponges and the first metazoans to build bioconstructions in association with calcimicrobes. They are a key group in biology, evolutionary studies, biostratigraphy, paleoecology and paleogeography of the early Cambrian times. The establishing of a new standardized terminology for archaeocyathans description has permitted the creation of the first knowledge base in English including descriptions of all archaeocyathan genera. This base, using the XPER² software package, is an integral part of the -Archaeocyatha- a knowledge base website, freely available at url http://www.infosyslab.fr/archaeocyatha. The website is composed of common information about Archaeocyatha, general remarks about the knowledge base, the description of the 307 genera recognized with images of type-specimens of type-species for each genus, as well as additional morphological data, an interactive free access key and its user guide.

The automatic analysis and comparison of the digitized descriptions have identified some genera with highly similar morphology. These results are a great help for future taxonomic revisions and suggest a number of possible synonymies that require further study.

Keywords

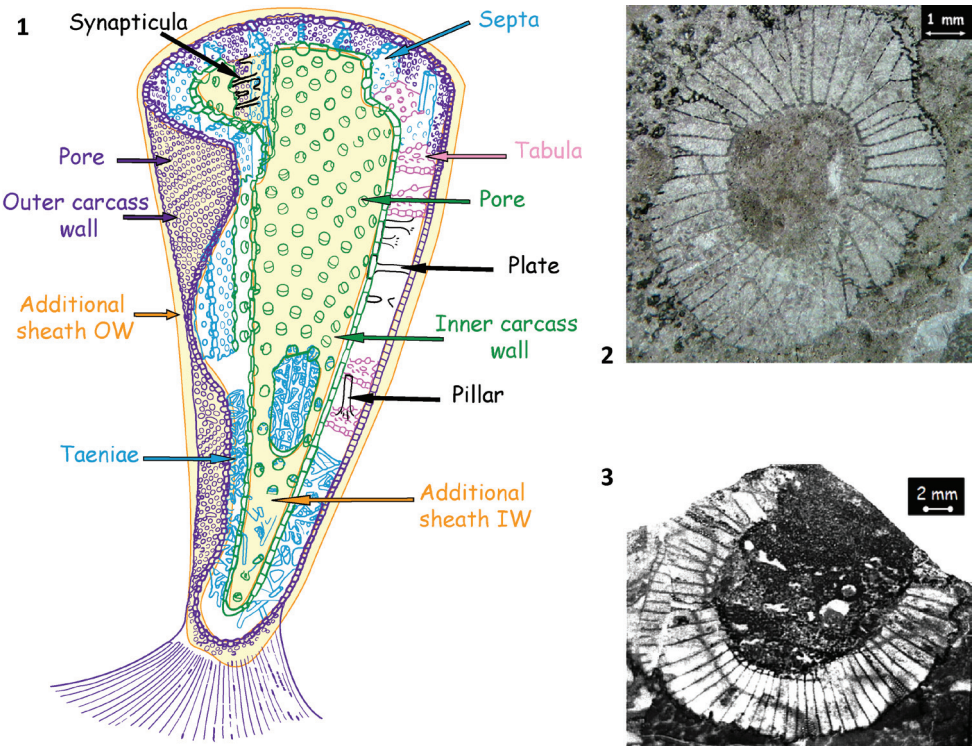
Archaeocyatha, Cambrian, standardized characters, identification key, knowledge base, XPER²

Introduction

Archaeocyathan represent the earliest reefal metazoan faunas dated at 521my, predating the Burgess Shale fauna and postdating Ediacarian faunas. They are exclusively Cambrian organisms that built the first metazoan bioconstructions as corals do today. Discovered in the middle of the XIXth century in the oldest fossiliferous rocks of Labrador, Canada, their geographical distribution is world-wide including Antarctica, Argentina, Australia, Canada, China, Germany, Greenland, France, Kazakhstan, Mongolia, Morocco, Poland, Uzbekistan, Sardinia, Serbia, South Africa, Spain, Russia and USA. Since their discovery, intensive studies have been carried out through international cooperation. Consensus about the phylogenetic relationships and biostratigraphic significance of these enigmatic organisms has been now achieved (see summary in Debrenne and Zhuravlev 1992).

As mysterious fossils without recent close-relatives, archaeocyathan represent an extinct class of the phylum Porifera, close to the Demospongiae (Debrenne and Vacelet 1984; Debrenne and Zhuravlev 1992). Their skeleton is commonly preserved as carbonate within limestone, which precludes their mechanical or chemical extraction from the surrounding matrix. Therefore, their complex, sometimes problematic morphology has to be examined through thin sections. As a consequence the orientation of the section through the skeleton, which influences the description and identification of the specimen is poorly controlled or even random. Identification of incomplete specimens is also highly problematic and the lack of specialists in the field aggravates this. As important Cambrian organisms, it is necessary for specialists and non-specialists to be able to rapidly and unambiguously identify specimens. However, easy to use identification keys are lacking despite several attempts to create such tools. Rozanov, using Vavilov's Law (Vavilov 1922), produced tables with homologous series, used as identification systems (Rozanov and Missarzhevskiy 1966). The variability in the homologous series of Archaeocyatha contains three groups of primary skeletal elements: the outer wall, the inner wall and the intervallum. Identification can be obtained using combinations of these three groups of characters but this approch, which resembles an identification key, is still complex and inapplicable to incomplete specimens.

Archaeocyathan databases have been successively developed since the 1980s. These include a database on Ajacicyathida by Debrenne and Prieur (1981), a computer aided identification with single access key, called ECAD, by M. and F. Debrenne (never circulated and stopped due to ongoing taxonomic revisions (Debrenne et al. 2002)), and a first, French version of the XPER² knowledge base (Debrenne and Kerner 2006). All these databases, with the exception of the XPER² system, used a fixed sequence of character choices that is insufficiently flexible to fit the morphological complexity of the archaeocyathans. This paper is aimed at introducing (1) a new standardized terminology for archaeocyathans description applicable in knowledge base, (2) the first knowledge base including descriptions of all archaeocyathan genera with a free access identification key, and (3) some outputs of the dataset. A brief review of archaeocyathan anatomy, systematics and importance in the Cambrian system is first given. The



Figures 1–3. 1 Stylized archaeocyathan skeleton (Debrenne, 1964 modified) **2** *Erismacoscinus* sp. in transverse section. specimen 2474 4.2Tb MNHN, Paris collection Destombes, Jbel Taissa, Morocco **3** *Coscinocyathus dianthus* Bornemann, lectotype GML An597, Canal Grande, Sardinia (Bornemann 1886).

results of this study are freely accessible on the Internet at: http://www.infosyslab.fr/archaeocyatha.

Introduction to Archaeocyatha

Anatomy and systematics of Archaeocyatha. Morphologically, the archaeocyathan skeleton is composed of two inverted porous cones, fitting into each other and interpreted as outer and inner walls delimiting the intervallum. Vertical radial elements (septa, taeniae) and/or horizontal elements (tabulae) connect the two walls (Fig. 1). The archaeocyathan cups display various architectural types: one-walled conical, single-chambered subspherical, multi-chambered conical (thalamid), chaetetid, and syringoid with solitary or modular (pseudocolonial) habits. Their skeleton is primarily made of globally polyhedral crystallites of high-magnesian calcite, probably the result of an organic matrix mediated process at a very primitive stage.

Archaeocyathan systematics is based on skeletal ontogeny determining the order of appearance of skeletal elements, their degree of complication and the stabilization of adult

features. Orders are characterized by the architecture of the cup, suborders by growth pattern models, superfamilies by the outer wall types, families by the inner wall types. Genera are differentiated by variations in walls and intervallar types, as well as distribution of pores in each element. Species are separated by different numerical coefficients (Debrenne et al. 1990; Debrenne and Zhuravlev 1992; Debrenne et al. 2002). The Class Archaeocyatha is composed of six orders and twelve suborders. The previous conventional subdivision into Regulares and Irregulares is often still used in biostratigraphy or paleoecology. These subdivisions roughly correspond to Ajacicyathida and Coscinocyathida (ex-Regulares) and Archaeocyathida and Kazachstanicyathida (ex-Irregulares).

The role of archaeocyathan in the Cambrian System. Archaeocyatha are of prime importance in biostratigraphic studies. The first stage subdivision based on archaeocyathan was established on the Siberian platform (Zhuravleva 1960). Subdivison of the Cambrian System was traditionally based on trilobite occurrences. However, the discovery of a rich archaeocyathan fauna on the Siberian Platform in horizons below the first appearance data (FAD) of trilobites, provided evidence for the establishment of a new stage, the Tommotian (Rozanov and Missarzhevskiy 1966), subdivided in 3 zones (Tab.1). Since then, archaeocyathan biozones have been used in key Cambrian areas such as Siberia, Morocco, Spain, Canada and Australia. The distribution of archaeocyathans in time, mainly early Cambrian with few relicts in middle and late Cambrian, limits their use to stages 2 to 4 of the International Stratigraphic Chart. Two parallel scales, one based on trilobites the other on archaeocyathan are established when possible for many Cambrian localities where archaeocyathan and trilobites are well studied. Under certain conditions, archaeocyathan may provide finer biozones than trilobites (Table 1).

Another interest concerns their paleoecology. Detailed studies of archaeocyathan settlements show that they were adapted to a narrow temperature range, corresponding to the intertropical zone. They were stenohaline organisms, living in the soft substrates of the intertidal zone. As passive filter-feeders, they are more adapted to habitats with reduced turbulence. Since Cambrian rocks lack usual climatic indicators such as tillites, modern phosphorites, or clays containing fossils of terrestrial vegetation, archaeocyathan with their retricted living conditions, are good indicators for ecological and environmental reconstructions (Debrenne et al. 2002; Gandin and Debrenne 2010). They are also significant in paleobiogeography. Reconstructions of land distribution are difficult for the Precambrian/Cambrian periods due to problems of paleomagnetism. Archaeocyathan reef distributions in epireic seas constrain map building. Five provinces are recognized after a first phase cluster analysis of generic distribution data: Siberia-Mongolia, Central East-Asia, Europe-Morocco, Australia-Antarctica, North-America-Koryakia. Two realms are defined by a second phase cluster analysis: Eurasia – the three first provinces – and Lauraustral – the last two provinces (Kruse and Shi 2000). Moreover, the pathways of archaeocyathan migrations inferred from the Jaccard Coefficient (Kruse and Shi 2000) confirm the early Cambrian existence of East and West Gondwana, the rifting of Laurentia from the Australian-Antarctic margin, and the drift of suspect Altay Sayan and Mongolia

Table 1. Comparison of archaeocyathan and trilobites biozones (modified after Rozanov and Sokolov 1984 and Mansy et al. 1993).

STAGES	SEBERIAN PLATFORM		ALTAI SAYAN		LAURENTIA	
	Archaeocyatha	Trilobita	Archaeocyatha	Trilobita	Archaeocyatha	Trilobita
Stage 4 ?Toyonian	*Irinaecyathus grandiperforatus*	*Anabaraspis splendens*	*Erbocyarhus heterovallum*	*Kooteniella-Edelsteinaspis*		*Plagiura / Poliella*
		Lermontova grandis	*Irinacyathus ratus*		*Tegerocyathus greenlandensis / Pycnoidocyathus pearylandicus*	
		Bergerionella ketemensis	*Adaecyathus solidus*	*Parapoliella-Oncocephalina*	*Archaeocyathus atlanticus*	
Stage 4 ?Botoman		*Bergerionaspis ornata*	*Syringocyathus aspectabilis*		*Pycnoidocyathus serratus / Tabulaconus kordae*	*Bonnia / Olenellus*
		Bergerionellus asiaticus	*Tercyathus altaicus*	*Poliellina-Laticephalus*	*Claruscoscinus fritzi / Metacyathellus caribouensis*	
		Bergerionellus gurarii				
	Porocyathus squamosus, Botomaecyathus zelenovi	*Bergerionellus micmacciformis / Erbiella*	*Clathricoscinus*		*Ethmophyllum whitneyi / Sekwicyathus nahaniensis*	
Stage 3 ?Atdabanian	*Fansycyathus lermontovae*	*Judomia*	*Arturocyathus torosus*	*Sajanaspis-Kameshkoviella*		*"Nevadella"*
			Nalivkinicyathus cyroflexus			
	Nochoroicyathus kokoulini					
	Carinacyathus pinus	*Pagetiellus anabarus*	*Thalamocyathus howelli*	*Resimopsis*		*"Fallotaspides"*
	Retecoscinus zegebarti		*Nochoroicyathus marinskii*			
		Profallotaspis jakutensis				
Stage 2 ?Tommotian	*Dokidocyathus lenaicus / Tumuliolynthus primigenius*					
	Dokidocyathus regularis					
	Nochoroicyathus sunnaginicus					
Fortunian						

terrains of the Chinese East Gondwana margin towards Siberia (Debrenne et al. 1999). These results highlight the role of archaeocyathan as key group for fundamental problems in paleobiology and geology and the important support the exhaustive compendium of archaeocyathan species and genera, along with efficient identification key, may provide for future studies.

Standardized terminology for Archaeocyatha

Proposition for an adapted terminology. A digitized knowledge base can be enhanced if the taxonomic descriptions can be compared through automatic processes. This is possible if descriptions are written using a common and standardized set of characters. However, in the literature, descriptions are often heterogeneous, using different terms or described according to a specialist's interpretation. The task to standardize the descriptions of Archaeocyatha was easier thanks to series of recent systematic revisions (Debrenne et al. 1990; Debrenne and Zhuravlev 1992; Debrenne et al. 2002). Despite this, some problems appeared due to equal states. For example, the difference between an "arcuate" structure and a "curved" structure is not immediately clear. These states may be identical, but each potential equivalent term has to be checked carefully before synonomising to a single term in the knowledge base. Reconciling traditional morphological terms is necessary in character construction for databasing. Character standardization reveals some hidden problems due to diagnoses and terminology. A classical diagnosis often follows this pattern:

Outer wall + one complex descriptive term,
inner wall + one complex descriptive term,
type of radial structure +/- other intervallar structure (tabulae…)

With such a structure, vocabulary homogenization is not adequate. Most of the terms included several concepts. The standardization step here requires the subdivision of complex descriptive terms into a list of terms with only one notion included. For example, the term "cambroid pores" contains information about the shape and the repartition of pores. Each character and state should be examined from all aspects and only basic descriptors (composed of only one notion) should be retained. This new organization of descriptors means the appearance of new terms and the disappearance of some classical terms. Moreover, in monographs, diagnoses are built only with characters that have a taxonomic interest. Some states and/or descriptors do not have any taxonomic value but are highly visual and helpful for identification e.g. descriptors 31 & 51 in the online knowledge base.

The main difference between the traditional terminology referring to Archaeocyatha and one adapted to a knowledge base concerns the description of walls: terms have been dissected into basic descriptors and grouped differently (Table 2).

We consider that a wall can be composed of one or two parts. The first one, always present, is named a carcass wall (descriptors 6, 7, 11 to 28 & 31 to 50) and the second is an additional wall (descriptor 52 to 63). A carcass wall generally has perforations (pores or canals) (descriptors 11 to 21 & 31 to 44) and may have different structures: bumps (tumuli, putulae) (descriptors 22 to 24) or external plates (spines, bracts, scales, annuli) (descriptors 25 to 28 & 45 to 50). Additional walls group together the microporous sheaths (descriptors 53 to 55, 57 to 60 & 63), sieves formed by protrusions (compound walls: incipient pore subdivision and completely subdivided pores) (descriptors 53, 55 to 58, 60, 62 & 63) and mesh (tabella, clathri, pseudoclathri) (descriptors 58 & 61). Each element is described inside these new associations. This new organization

Table 2. Classical terminology referring to archaeocyathan walls and their equivalent basic descriptors.

outer wall and inner wall													
usual terms		new standardized terminology											dependent characters
		carcass wall						additionnal sheath					
		structure well defined	perfo-rations	bumps		external plates	morphologi-cal tubes	pre-sent	micro-porous sheath	mesh	sieve		
				perforations	opening								
rudimentary		no	pores	none		none	none	no					description of types and repartition of pores
simple		yes	pores	none		none	none	no					
basic		yes	pores	none		none	none	no					
concentrical		yes	pores	none		none	none	no					
with canals		yes	canals	none		none	possible	no					description of canals
								yes	possible	none	possible		
						bracts, scales, unnuli	yes	no					
								yes	possible	none	possible		
with tumuli	simple	yes	pores	simple	downwarldy	none	none	no					description of bumps
	multiperforate			multiperforate	downwarldy								
pustular		yes	pores	simple	central	none	none	no					
with bracts scales or poretubes		yes	pores	none		yes	none	no					description of external plates
								yes	yes				
			canals	none		yes	yes	no					
								yes	possible		possible		
with annuli		yes	pores	none		yes	possible	no					
								yes	yes				
			canals	none		yes	possible	no					
								yes	yes				
compound		yes	pores	none		possible	none	yes	possible	none	possible		decriprion of additional sheath and of carcass structures
with microporous sheath		yes	pores	none		possible	none	yes	yes	none	none		
tellelar		yes	pores	none		none	none	yes	none	yes	none		
clathrate		yes	pores	none		none	none	yes	none	yes	none		
pseudoclathate		yes	pores	none		none	none	yes	none	yes	none		
tabular		link to tabulae descriptors: tabulae are extensions of inner wall, of outer wall or of both wall											

included all usual wall types apart from tabular walls that are considered to be linked to tabulae (descriptor 80). For example, a simple tabular outer wall is considered as a single character in traditional terminology, here it is decomposed into different components: outer wall is in one part (no additional sheath), (descriptor 52), this part is composed of simple pores (descriptors 11 & 12) and, moreover, in the intervallum there are tabulae (descriptor 79) stemming from the outer wall curve line (descriptor 80).

Tabulae have been subdivided into two descriptors. The first one describes their construction (descriptor 80): independent of both walls (simple, pectinate, plate and membrane tabulae) or dependent on the inner wall, the outer wall or both walls (curved, simple segmented, concentric segmented and compound segmented tabulae). The second one describes the porosity of tabulae (descriptor 81).

Modifications of the traditional terminology. Different causes can justify the modification of used terminology: a single term refers to two or several different structures, two different terms refer to different things but introduce confusion between two different structures. The first example concerns the term "spines" that was used for two different structures: 1) external plates that look like bracts and 2) skeletal elements that divided pores to form an additional sheath. In the first case the term "spines" is retained whereas the second now corresponds to "protrusions". The second example concerns sub-spherical chambered canals that may easily be confused with what we refer to as "chamber", hence our preference for the term "curved canals". However, the terminology referring to communicating canals appears difficult to understand for novices. We have chosen non porous, porous and spongiose to replace non-communicating, simple communicating and anastomosing. Finally, the difference between completely subdivided and incipient pore subdivision appears only in additional sheaths with the descriptor 62 "type of sieves". Both are considered subdivided in the description of carcass wall pores (descriptors 12 & 32).

The second instance of terms that necessitated modification concern updating the character states. The Checkbase function in XPER² shows that the *Taylorcyathus* and *Connanulofungia* descriptions are similar and that new observations of their inner wall annuli show that they are stacked differently. A new state has therefore been created inside the knowledge base to describe the annuli of *Connanulofungia*: cone in cone (descriptor 49).

Other terms illustrating complex characters become useless after their division into basic descriptors. A first case concerns tumuli and pustulae. Simple tumuli and pustulae definitions are close together: these form bumps on a carcass outer wall and have a single opening, with a difference in the direction of the opening. With basic descriptors, a bump is described with its perforations oriented to the opening direction and the terms "tumuli" and "pustulae" presence therefore become redundant and inadequate (descriptors 22 to 24). The second case is about cambroid pores and anthoid pores. Pores are defined with some basic descriptors: their type (or shape), their distribution, their arrangement (descriptors 11 to 17 & 31 to 34). Cambroid pores are simple or polygonal pores (descriptor 12) with a regular distribution on the outer wall (descriptor 15) and a random arrangement (descriptor 16). Anthoid pores are polygonal pores (descriptor 12) with an irregular distribution (descriptor 15). With such deconstruction, anthoid pores and cambroid are not considered as terms to in-

clude in the knowledge base. In the case of basic and rudimentary walls these are quite difficult to distinguish. The term "rudimentary wall" is used for imperforate walls and for a skeleton without a carcass wall well defined: Tips of intervallar structures serve as carcass, and spaces between intervallar structures as carcass pores. The term "basic wall" means a carcass wall built with tips of intervallar structures too but with additional lintels between these, forming the carcass. Imperforate walls are defined as carcass walls that are well defined without perforations. The term "rudimentary perforate wall" is decomposed into carcass not well defined (descriptors 6 & 7), carcass pores irregular (descriptor 12 & 32), irregular repartition (descriptor 15) and one row of pores per intersept (descriptor 17 & 33) and basic wall into carcass well defined (descriptors 6 & 7), carcass pores irregular (descriptor 12 & 32), irregular repartition (descriptor 15) and 2 or more than 2 rows of pores per intersept (descriptor 17 & 33). The last discarded term is pseudotaeniae, defined as "taeniae with synapticulae at each interpore node". In the knowledge base, this results in a descriptor association: vertical intervallar structures are taeniae (descriptor 65 & 66) and links are synapticulae which repartition is at each interpore node (descriptor 70 & 71).

On line service for Archaeocyatha recognition

Archaeocyathan knowledge base. This was developed with Xper², which is software, available for use under a Creative Commons by-nc-nd license. The software is dedicated to storing structured descriptive data and to provide free (matrix) access keys (http://www.infosyslab.fr/lis/?q=en/resources/software/xper2, Ung et al. 2010).
The archaeocyathan knowledge base (Kerner et al. 2011) is composed of a set of standardized descriptions, one for each genus: all the descriptions use the same set of descriptors and character states. Terminology is therefore controlled and further documented by text and images. Images each have their own copyrights. The dataset (without images) is distributed for use under a Creative Commons license (by-nc-nd, see http://creativecommons.org/licenses/by-nc-nd/3.0/).

Most of characters proposed by paleontologists to identify archaeocyathan genera were collected from relevant literature (Debrenne et al. 1990; Debrenne and Zhuravlev 1992, Debrenne et al. 2002) and confirmed by observation of about 1000 specimens in the collections of the Museum National d'Histoire Naturelle, which contain about 600 type and illustrated specimens.

The knowledge base is composed of 307 genera considered to be valid at present with a world-wide geographical coverage. Stratigraphically, the knowledge base contains all the Cambrian deposits despite the predominance of Archaeocyatha in early Cambrian deposits. Each genus is illustrated with type specimens of the type species and some additional specimens. A total of 120 descriptors are used, 85 corresponding to morphological and ontogenetic data, 8 to stratigraphic and geographic data and 27 refer to traditional classification data. To each descriptor and character state we associated a definition and images, and/or drawings.

Free access key. Incomplete specimens cannot be identified with traditional tools. Single access keys and natural keys (following the classification) are insufficiently flexible as they contain a predefined sequence of steps in the identification that rely on the presence of these distinguishing characters in the specimen. For example, if the outer walls (carcass more or less additional sheath) are not preserved, identification can only be made to suborder level. The identification service offered by Xper² is available offline or via the Internet as a free access key (Fig. 4). With free access keys (Hagedorn et al. 2010), the selection of a particular descriptor is chosen by the user at each step of the identification. The computer-aided identification tool deduces the remaining and eliminated taxa for each selected character and can display the reasons for each elimination (for various examples see Nimis and Vignes-Lebbe 2010). This type of identification system is very flexible and allows the identification of a specimen even if some of the described characters are not available, as is regularly the case for palaeontological specimens that

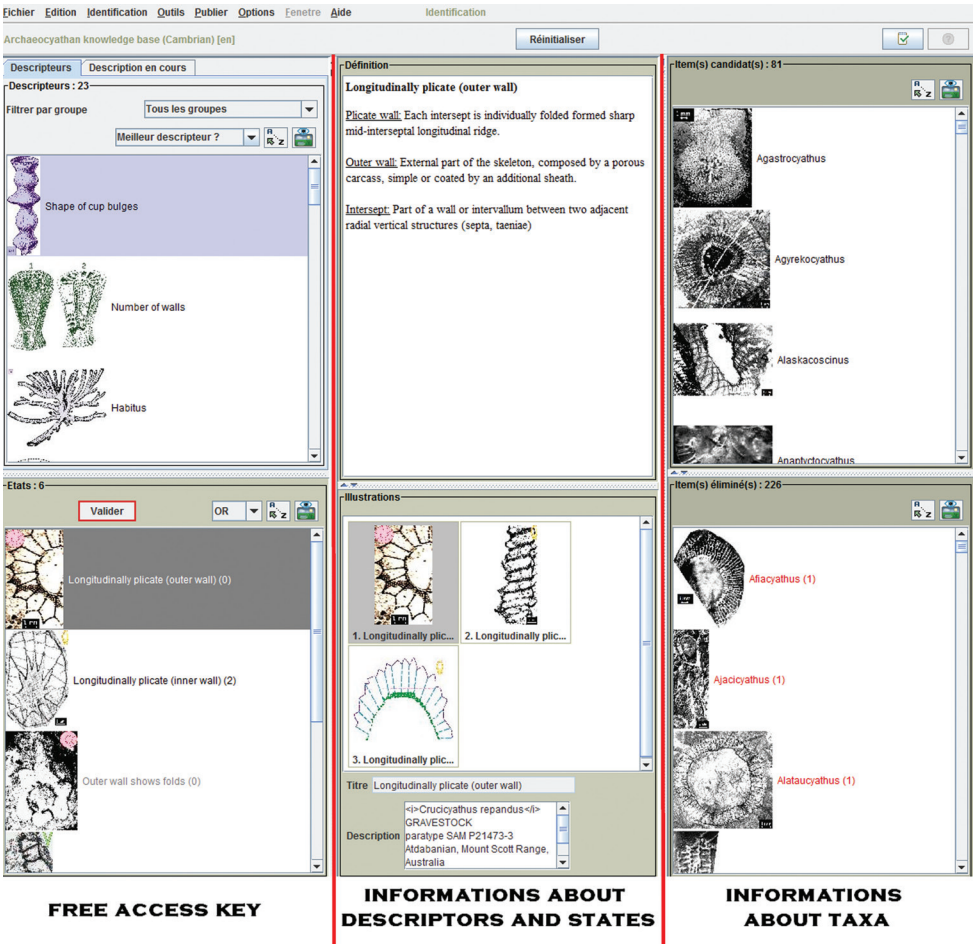


Figure 4. Screen shot of the free access key constructed in this study for archaeocyathan genera.

are often incomplete. Moreover, the user can express doubt by selecting more than one character state, or chose another descriptor. Furthermore, at any step in the identification, the user can question why a taxon should have been eliminated and can check the descriptors that are incompatible with the specimen description. Another advantage of free access keys concerns superficially similar genera that do not belong to the same Order or Suborder. For example *Coscinocyathus* and *Erismacoscinus* are two genera that look very similar in transverse section even though they are not in the same Order (Figs 2–3). *Coscinocyathus* has chambers, which is why it belongs to Capsulocyathida whereas *Erismacoscinus* does not have chambers and belongs to Ajacicyathida. The problem is that chambers are not visible in transverse section, so confusion is frequent. With free access keys, it is possible to identify these genera without using characters of the chamber.

Outputs and analysis of the knowledge base. A complete form for each genus including descriptions, pictures and information concerning their systematics can be published from the system (Fig. 5).

The Checkbase function compares all the pairs of taxonomic descriptions to see if they are distinguishable or if they overlap. If these conditions occur it means that some morphological aspects are compatible with more than one taxon. This iterative process is useful to check the consistency or misinterpretations of characters and the completeness of the knowledge base. It detected some similarities between *Graphoscyphia*, *Dictyocyathus* and *Molybdocyathus*. The three genera have an inner carcass wall with one row of simple pores per intersept and a dictyonal network. The only difference concerns the outer carcass wall which is basic or rudimentary. In recent literature, *Graphoscyphia* and *Dictyocyathus* have the same description: a basic outer carcass wall whereas *Molybdocyathus* has a rudimentary one. A fresh look at the specimens reveals that *Graphoscyphia* has a basic outer carcass wall (as originally described), *Dictyocyathus* does not have a basic outer carcass wall, but a rudimentary one and *Molybdocyathus* has a rudimentary one too (as originally described). With the change in the interpretation of the carcass wall structure of *Dictyocyathus*, the genus *Dictyocyathus* appears identical to *Molybdocyathus*. *Molybdocyathus* is now considered to be a junior synonym of *Dictyocyathus*. The automatic comparison of descriptions is displayed in a table, using different colors to highlight characters that are common or different in two or more genera. In Figure 6, we use this feature to visualize the morphological forms in the Tumulocoscinidea family. This tool has different uses. First, it can help to rapidly complete an identification when few taxa remain and differences can easily be seen. It can also be useful as a teaching tool for archaeocyathan identification. In the same way during the identification process, the information as to why a taxon is discarded (states incompatible are colored in red in the complete form of the discarded taxon) can be used as an efficient method to help the user with recognizing character states and descriptive terms.

Xper² can extract "special features" (Fig. 7), i.e. unique states present only in a single taxon. We used this feature here to check the data. It could also be used to help weight characters when creating a classical polytomous key. Some software already exists to create keys from data matrices, and in a near future we will connect our application to the webservices of the ViBRANT project (http://vbrant.eu/) to facilitate this.

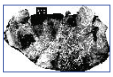
Agastrocyathus

Agastrocyathus DEBRENNE, 1964

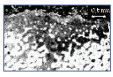
Specimen-type: *Protopharetra gregaria* holotype, DEBRENNE, 1961, pl. 2, fig. 5-6

Systematics:
Order: ARCHAEOCYATHIDA Okulitch, 1935
Suborder: ARCHAEOCYATHINA Okulitch, 1935
Superfamily: METACYATHOIDEA R. Bedford& W. R. Bedford, 1934
Family: COPLEICYATHIDAE R. Bedford& J. Bedford, 1937

Agastrocyathus
Agastrocyathus gregarius (DEBRENNE) holotype MNHN M80138, HD71 Atdabanian, Jbel Taïssa, Morocco Transverse section (DEBRENNE, A. ZHURAVLEV & KRUSE, 2002).



Agastrocyathus
Agastrocyathus gregarius (DEBRENNE) holotype MNHN M80138, HD71 Atdabanian, Jbel Taïssa, Morocco Longitudinal section (DEBRENNE, A. ZHURAVLEV & KRUSE, 2002).



Agastrocyathus
Agastrocyathus gregarius (DEBRENNE) holotype MNHN M80138, HD71 Atdabanian, Jbel Taïssa, Morocco Detail of outer wall in tangential section (DEBRENNE, A. ZHURAVLEV & KRUSE, 2002).

Outer wall

- **Structure of outer carcass wall (OW/SW) :** Carcass wall well defined
- **Carcass perforations: outer wall or single wall (OW/SW) :** Pores
- **Type of carcass pores (OW/SW) :** Subdivided
- **Size of carcass pores (OW/SW) :** Uniform size
- **Repartition of carcass pores (OW/SW) :** Regular repartition
- **Repartition types of carcass pores (OW/SW) :** Straight rows of pores
- **Relationship between rows of carcass pores and intersept (OW/SW) :** ?
- **Carcass bumps: outer wall or single wall (OW/SW) :** None
- **Carcass external plates: outer wall or single wall (OW/SW) :** None
- **Stipules :** None
- **Type of morphological tubes (OW/SW) :** None
- **Additional sheath :** Present on outer wall (or single wall)
- **Type of additional sheath (OW/SW) :** Sieve formed by protrusions
- **Repartition of bracts on additional sheath (OW/SW) :** No bract on additional sheath
- **Type of shieve (OW/SW) :** Protrusions aren't completely connected
- **Repartition of micropores (OW/SW) :** Irregular

Inner wall

- **Structure of inner carcass wall (IW) :** Carcass wall well defined
- **Stipules :** None
- **Carcass perforations: inner wall (IW) :** Pores
- **Type of carcass pores (IW) :** Simple
- **Relationship between rows of carcass pores and intersept (IW) :** 1
- **Relationship between rows of carcass pores and tabulae (IW) :** No tabula
- **Building of carcass perforations (IW) :** Without participation of vertical structures
- **Carcass external plates: inner carcass wall (IW) :** None (external plates and/or inner wall)
- **Type of morphological tubes (IW) :** None
- **Additional sheath :** Present on outer wall (or single wall)

Intervallum

- **Vertical intervallar structures :** Present
- **Type of vertical intervallar structures :** Taeniae
- **Repartition of vertical intervallar structures pores :** ?
- **Porosity of vertical intervallar structures :** ?
- **Pores of vertical intervallar structures :** Circular pores
- **Link between vertical intervallar structures :** Synapticulae
- **Repartition of synapticulae :** Random
- **Intervallar cells :** None
- **Tabulae :** none

Others informations

- **Cup shapes :** Cylindrical-conical
- **Shape of cup bulges :** None
- **Number of walls :** 2
- **Habitus :** Solitary ; Modular
- **Type of modularity :** Branching by external budding ; Branching by longitudinal subdivisions
- **Central Cavity (or internal if one wall) :** Empty ; Full
- **Thorny corolla :** None
- **Chambers :** None
- **Other aquiferous unit :** None
- **Stratigraphical extension :** Atdabanian
- **Atdabanian :** 2 ; 3 ; 4
- **Geographical repartition :** Europe ; Morocco
- **Europe :** Iberia

Systematic

- **Orders :** Archaeocyathida
- **Suborders inside the Archaeocyathida :** Archaeocyathina
- **Superfamillies inside the Archaeocyathina :** Metacyathoidea
- **Famillies inside the Archaeocyathina :** Copleicyathidae

External view

- **Cup shapes :** Cylindrical-conical
- **Shape of cup bulges :** None
- **Structure of outer carcass wall (OW/SW) :** Carcass wall well defined
- **Structure of inner carcass wall (IW) :** Carcass wall well defined
- **Habitus :** Solitary ; Modular

Figure 5. Agastrocyathus detailed sheet.

Archaeocyatha Website. The archaeocyathan knowledge base and outputs are included inside a website about Archaeocyatha. This site is composed of different information types. The first part, called Archaeocyatha brings together common information about Archaeocyatha: an introduction, covering their role in Cambrian systems, their morphology and a bibliography. The second part is about the knowledge base. This is composed of general remarks about the knowledge base and some data exports from the system: list of genera and their detailed sheets, list of descriptors, list of groups of descriptors and the base properties. The last part concerns the interactive key and its tools: user guide, matching terminologies and glossary. Matching terminologies correspond to the list of all usual terms used in archaeocyathan descriptions. From this, the user can find how a traditional term appears in the knowledge base.

The new and complete English version of the archaeocyathan knowledge base (Cambrian) can be accessed at http://www.infosyslab.fr/archaeocyatha, and be used to identify an archaeocyathan specimen to generic level (Fig. 8).

	Asterotumulus	Orbicoscinus	Retetumulus	Tumulocoscinus	UNION	INTERSECTION
Cup shapes	Cylindrical-conical	Cylindrical-conical	Cylindrical-conical	Cylindrical-conical	Cylindrical-conical	Cylindrical-conical
Shape of cup bulges	Longitudinally plicate (inner wall)	Both walls show folds	None	None	Longitudinally plicate (inner wall), Both walls show folds, None	
Type of folds affecting both walls	not applicable	Transverse folds	not applicable	not applicable	Transverse folds	
Shape of folds affecting both walls	not applicable	Deep and close	not applicable	not applicable	Deep and close	
Number of walls	2	2	2	2	2	2
Structure of outer carcass wall (OW/SW)	Carcass wall well defined	Carcass wall well defined	Carcass wall well defined	Carcass wall well defined	Carcass wall well defined	Carcass wall well defined
Structure of inner carcass wall (IW)	Carcass wall well defined	Carcass wall well defined	Carcass wall well defined	Carcass wall well defined	Carcass wall well defined	Carcass wall well defined
Habitus	Solitary; Modular	Solitary; Modular	Solitary; Modular	Solitary; Modular	Solitary, Modular	Solitary, Modular
Type of modularity	Branching by longitudinal subdivisions; Caterulate	Branching by longitudinal subdivisions; Caterulate	Branching by longitudinal subdivisions; Caterulate	Branching by longitudinal subdivisions; Caterulate	Branching by longitudinal subdivisions, Caterulate	Branching by longitudinal subdivisions, Caterulate
Type of vertical intervallar structures	Septa	Septa	Septa	Septa	Septa	Septa
Repartition of vertical intervallar structures pores	Completely porous	Completely porous	Completely porous	Aporous; Sparsely porous	Completely porous, Aporous, Sparsely porous	
Porosity of vertical intervallar structures	?	?	Finely porous	Aporous; Finely porous	Aporous, Finely porous, Coarsely porous	Finely porous
Pores of vertical intervallar structures	Circular pores	Circular pores	Circular pores	Aporous; Circular pores	Circular pores, Aporous	Circular pores

Legend
- Discrimination
- Partial discrimination
- No discrimination

6

7

Features present in this (these) and only this (these) item (items)

Special feature(s) of the item Eremitacyathus

- Carcass perforations: inner wall (IW) : Vertical tubes
- Type of plates : Denticulate
- Families inside the Loculicyathina : Eremitacyathidae

Figures 6–7. 6 Comparative table of Family Tumulocoscinidea **7** Special features of Eremitacyathus.

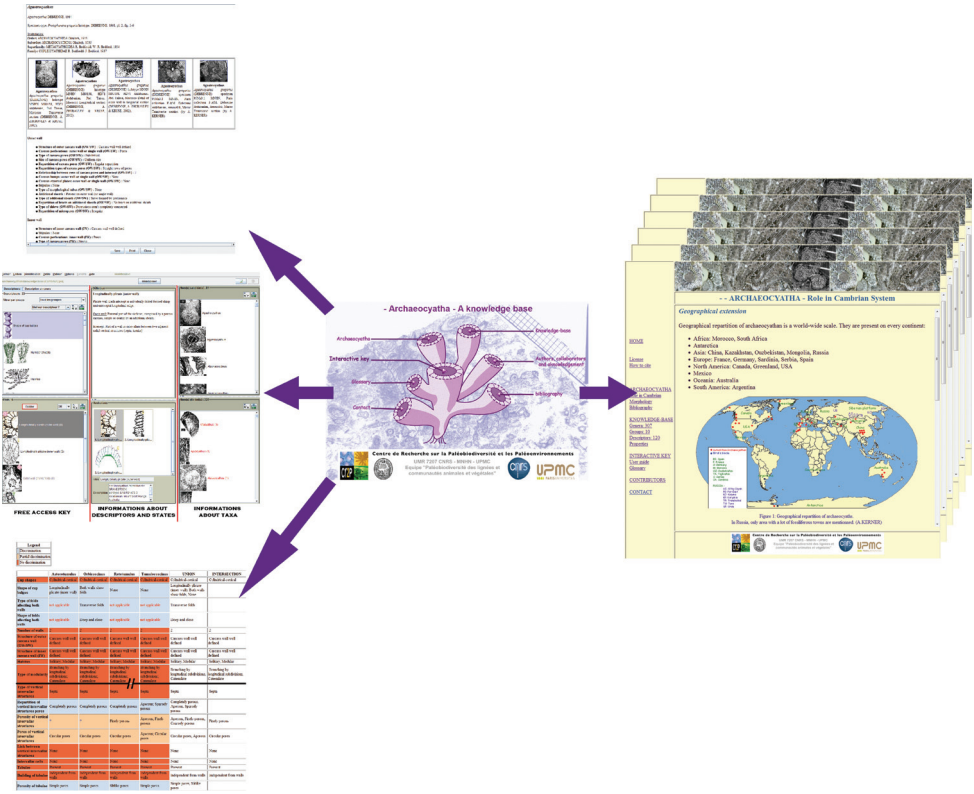


Figure 8. Archaeocyatha website structure. On the left, interactive key and some applications of it. On the right, an example of web page.

Conclusion

The identification of the Cambrian and predominantly early Cambrian metazoans referred to as Archaeocyatha, are important for a number of disciplines including biostratigraphy, paleoecology and paleogeography. Since the study of their morphological 3D-structures is complex due to different views in thin section, their identification is difficult. This problem is exacerbated by the lack of specialists in this field, with most now retired or involved in other projects. Establishing of a knowledge base for these organisms is a necessary tool and a first step to identify new field discoveries so that they can be placed in a wider context. The Xper² application for archaeocyathan genera is the first digitized content, in English, enabling identification with free access keys, and includes all currently accepted genera as well as illustrations of their nomenclatural types. A first version of the archaeocyathan knowledge base (Cambrian) is freely accessible online at URL http://www.infosyslab.fr/archaeocyatha. We hope that such an application constitutes an efficient resource for any further studies on Archaeocyatha.

The application is the first step of a general review on Archaeocyatha using the new tools for taxonomy. It will be completed and up-dated on an ongoing basis to follow and include new findings on these fossils. Content will focus on further characters analysis, both to refine the descriptions for paleontological studies, and to compute multidimensional characters. Tools will be developed to support further data analysis tool for discovering new discriminating characters. We plan to shift from a simple website (web 1.0) to a collaborative website (using Scratchpads see http://scratchpads.eu/) to open the application to the community of specialists and non specialists interested by Archaeocyatha data.

Acknowledgements

Dr Pascale Chesselet is warmly thanked for her helpful comments and style improvement. We want to thank the CNRS for funding the first author for her PhD thesis. We also acknowledge the review effort of Dr Sébastien Clausen and Dr Vince Smith.

References

Bornemann JG (1886) Die Versteinerungen des cambrischen Schichten-systems der Insel Sardinien nebst vergleichenden Untersuchungen uber analoge Vorkommnisse aus andern Landern. Nova Acta Kaiser Leopold–Carol dtsch. Acad. Naturforsch 51: 1–147.

Debrenne F, Kerner A (2006) Actualité des Archaeocyatha: état des recherches en cours. In: 21ème Réunion des Sciences de la Terre (RST), Dijon, Société Géologique de France, 29.

Debrenne F, Maidanskaya ID, Zhuravlev AY (1999) Faunal migrations of archaeocyaths and early Cambrian plate dynamics. Bulletin de la société géologique de France 170(2): 189–194.

Debrenne F, Prieur A (1981) Computerization of regular Archaeocyathan files. International symposium on conceptual methods in paleontology, Barcelona, Spain.

Debrenne F, Rozanov AI, Zhuravlev AI (1990) Regular archaeocyaths: morphology, systematic, biostratigraphy, palaeogeography, biological affinities = Archéocyathes réguliers = morphologie, systématique, biostratigraphie, paléogéographie, affinités biologiques. Ed. du Centre national de la recherche scientifique, 218 pp. + XXIII pl.

Debrenne F, Vacelet J (1984) Archaeocyatha: is the sponge model consistent with their structural organization? 4th International Symposium on Fossil Cnidaria, New York, 1983, 358–369.

Debrenne F, Zhuravlev A (1992) Irregular Archaeocyaths : morphology, ontogeny, systematics, biostratigraphy, palaeoecology = Archéocyathes irréguliers : morphologie, ontogénie, systématique, biostratigraphie, paléoécologie. CNRS Editions, 212 pp. + XXXVIII pl.

Debrenne F, Zhuravlev AI, Kruse PD (2002) Class Archaeocyatha Bornemann, 1884. Bibliography of Class Archaeocyatha. In: Hooper JNA, van Soest RWM (Eds) Systema Porifera. A Guide to the Classification of Sponges. Kluwer Academic/Plenum Publishers, New York, Springer, Vol. 2, 1553–1713.

Gandin A, Debrenne F (2010) Distribution of the archaeocyath-calcimicrobial bioconstructions on the Early Cambrian shelves. Palaeoworld 19(3/4): 222–241. doi: 10.1016/j.palwor.2010.09.010

Hagedorn G, Rambold G, Martellos S (2010) Types of identification keys. In: Nimis PL, Vignes-Lebbe R (Eds) Tools for Identifying Biodiversity: Progress and Problems, Paris (France), September 2010, Universita di Trieste: 59–64.

Kerner A, Vignes-Lebbe R, Debrenne F (2011) Computer-aided identification of the Archaeocyatha genera now available online. Carnets de Géologie/Notebooks on Geology, letter 2: 99–102.

Kruse PD, Shi GR (2000) Palaeobiogeographic affinities of Australian Cambrian faunas In: Brock GA, Engelbretsen MJ, Jago JB, Kruse PD, Laurie JR, Shergold JH, Shi GR, Sorauf JE (Ed) Palaeobiogeographic affinities of Australian Cambrian faunas. Association of Australasian Palaeontologists, Memoir 23: 1–61.

Mansy J-L, Debrenne F, Zhuravlev AYu (1993) Calcaires à Archéocyathes du Cambrien inférieur du Nord de la Colombie britannique (Canada). Implications paléogéographiques et précisions sur l'extension du continent Américano-Koryakien. Geobios 26(6): 643–683. doi: 10.1016/S0016-6995(93)80047-U

Nimis PL, Vignes-Lebbe R (2010) Tools for identifying biodiversity: progress and problems: proceedings of the international congress, Paris (France), September 2010. Universita di Trieste, 455 pp.

Rozanov A, Missarzhevskiy VV (1966) Biostratigrafiya i fauna nizhnikh gorizontov kembriya [Biostratigraphy and fauna of the lower horizons of the Cambrian]. Geologicheskiy Institut, Akademiya Nauk SSSR, Trudy 148, 126 pp.

Rozanov A, Sokolov BS (1984) Lower Cambrian stage subdivision stratigraphy. Moscow, Nauka, 184 pp.

Ung V, Dubus G, Zaragüeta-Bagils R, Vignes-Lebbe R (2010) Xper^2: introducing e-Taxonomy. Bioinformatics 26(5): 703–704. doi: 10.1093/bioinformatics/btp715

Vavilov NI (1922) The law of homologous series in variation. Journal of genetics 12(1): 47–89. doi: 10.1007/BF02983073

Zhuravleva IT (1960) Arkheotsiaty Sibirskoy Platformy [Archaeocyaths of the Siberian Platform], Akademiya Nauk SSSR, 344 pp.+ XXXIII pl.

ZooKeys 150: 397–405 (2011)
doi: 10.3897/zookeys.150.2350
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

The future of the past in the present: biodiversity informatics and geological time

Edward Baker[1], Kenneth G. Johnson[2], Jeremy R. Young[3]

1 *Department of Entomology, Natural History Museum, London, United Kingdom* **2** *Department of Palaeontology, Natural History Museum, London, United Kingdom* **3** *Department of Earth Sciences, University College London, London, United Kingdom*

Corresponding author: *Edward Baker* (edwbaker@gmail.com)

Academic editor: *V. Smith* | Received 11 November 2011 | Accepted 23 November 2011 | Published 28 November 2011

Citation: Baker E, Johnson KG, Young JR (2011) The future of the past in the present: biodiversity informatics and geological time. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 397–405. doi: 10.3897/zookeys.150.2350

Abstract

The biological and palaeontological communities have approached the problem of informatics separately, creating a divide between communities that is both technological and sociological in nature. In this paper we describe one new advance towards solving this problem - expanding the Scratchpads platform to deal with geological time. In creating this system we have attempted to make our work open to existing communities by providing a webservice of geological time data via the GBIF Vocabularies site. We have also ensured that our system can adapt to changes in the definition of geological time intervals and is capable of querying datasets independently of the format of geological age data used.

Keywords

Palaeontology, Biodiversity Informatics, Scratchpads, web services, GBIF

Introduction

Over recent years a number of projects have set out to create online communities and resources for the biological community. Similar projects have been developed by the palaeontological community to cover fossil taxa (e.g. http://www.paleodb.org) and to share information associated with geological time (for example http://www.chronos.org/).

Since the overwhelming majority of these resources are focused at workers in either the palaeontological or the neontological communities, a virtual divide is created between communities who work on the same branch of the tree of life. Especially when working with extant taxa that occur in the fossil record or when attempting to compile taxonomic information for both extinct and extant taxa within a particular group (http://corallosphere.org).

In order to address this problem we have taken the Scratchpads platform (http://scratchpads.eu; Smith et al. 2011) and developed additional functionality to allow for the recording of geological age data - a prerequisite for large-scale uptake of the Scratchpads platform by the palaeontological community. It should be noted that our solution deals only with age data and does not attempt to handle stratigraphy, although in well-studied local areas where stratigraphic terms have well-established geochronological meaning (e.g. Blue Lias around Lyme Regis) it is possible to model these names using the system we have developed.

Handling geological time – the nature of the problem

For a palaeontological database, and indeed most other types of geological data, geological age is an essential data type. For example, one might wish to record the likely age of a specimen or the age range through which a particular species is known to have lived. This sounds like a straightforward databasing problem analogous to recording the age of an historical object or geographical location data; age data or geographical location data can be converted into numerical age or geospatial coordinates on a one-off basis only needing to be revised if the original data is revised.

However, the geological timescale is not a simple known system but a constantly evolving body of knowledge. This can be illustrated by an example – consider the following statement: "The Ichthyosaur was collected from the *Arietites bucklandi* ammonite zone of the Blue Lias, at Lyme Regis (195-196Ma)." The hard data here is that the fossil was collected from the *bucklandi* zone, whilst the geological age given, 195-196Ma, is a modern estimate of the age of that zone. This interpretation has changed in the past and will change in the future as the geological timescale is refined. Changes may occur in this case either because the age of the Lower Jurassic is refined as the whole timescale is re-calibrated in the light of better radiometric data, or because the relative duration of the ammonite zones within it are refined. Indeed, whilst a modern (Gradstein et al. 2004) age estimate for the *bucklandi* zone is 195-196Ma an earlier estimate (Harland et al. 1982) was 205 to 206Ma. So even though the numerical age is needed for communication to non-expert audiences and for database queries, it is essential that only the primary data is recorded in the database and this is then dynamically used to derive the numerical age interpretations as needed, going via a separately maintained look-up table or dictionary of age definitions.

Community resources

Van Couvering and Ogg (2007) lamented the lack of an online service for geological timescale data. So, as part of the project we have added several sets of geological timescale data to the GBIF Vocabularies site (e.g. http://vocabularies.gbif.org/vocabularies/geo_chronostrat) based on the GTS 2004 (Gradstein et al 2004).

Each entry in these vocabularies has a name and either a date or date range (defined by a base and top age) as well as additional metadata where appropriate, e.g. FAD/LAD (first/last appearance datums) for nannofossil events. By providing open access to the information, we have provided a platform from which both we and others can start to build web-based timescale tools. Equally importantly since the vocabularies are stored separately from the specimen records they can readily be updated as revised timescales are developed and these revisions will then cascade to all specimen records.

The present implementation of the GBIF Vocabularies site has several issues. The first of these being a requirement for only alphanumeric characters in the name of a term (e.g. a requirement to use LowerJurassic rather Lower Jurassic). Secondly the age metadata can only be exported via the CSV export, and not the XML webservice. GBIF are currently working on improving the Vocabularies site, and we are working closely with them to ensure that the site will be capable of fulfilling our requirements.

Current scratchpad implementation

Experimental setups were created on two Scratchpads: Nannotax (http://nannotax.org) and the Indo-Pacific Ancient Ecosystems Group (IPAEG: http://ipaeg.org). Both of these examples use a predefined custom content type (GeoTime) to store information about the geological ages that can be referenced by other content types using a nodereference field. The GeoTime content type stores the name of a geological age range or date along with other essential information, including age data and event type (e.g. FAD/LAD), where applicable.

The Nannotax implementation allows for the first occurrence and last occurrence to be recorded using the data in the form it is available in (e.g. geological stage or magnetochron data). The pair of ages thus defines the total age range of the species and will allow both the age range of the species to be restated and queried in uniform formats (e.g."which species of taxa X, Y, Z occurred at time n").

The IPAEG site uses a more complex data model that acts as the foundation point for the Scratchpads 2.0 implementation in development. Like the Nannotax site it is possible to enter any number of predetermined geological ages but, in addition, it is possible to enter a custom age range or custom spot date.

In order to perform calculations with age data it is essential to access the combined range of the data entered by the user. Two useful combinations have been incorporated into the system so far: the union and intersect of the complete data set. Future work

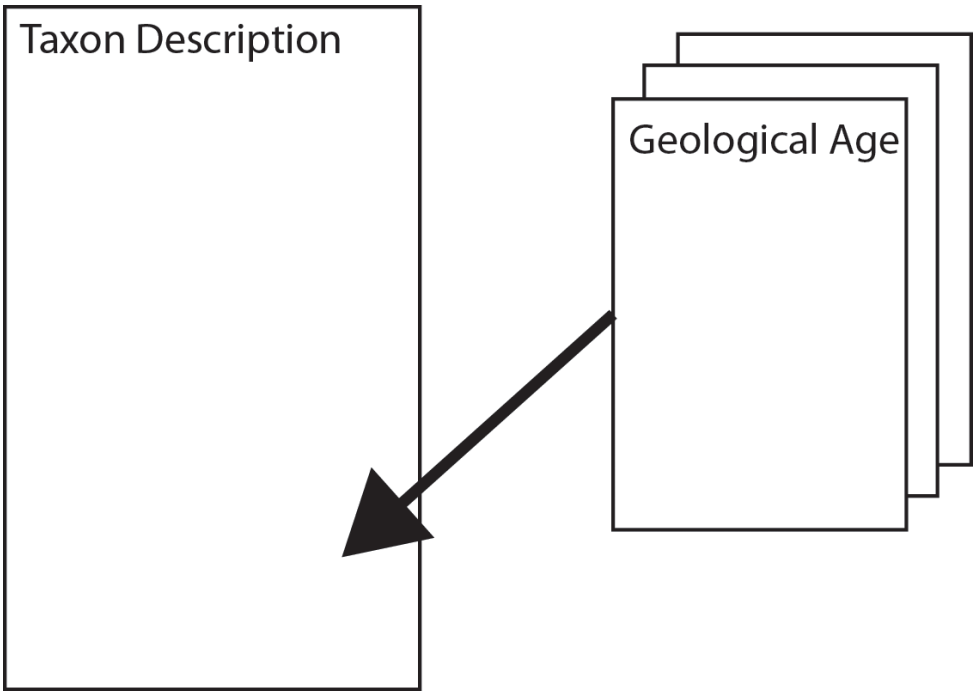


Figure 1. Data model for Nannotax.

Genus *Coccolithus* Schwarz 1894

Description: Elliptical or circular placolith coccolith showing the typical coccolthaceae rim structure (V-unit forms distal shield and lower cycle of central-area; R-unit forms proximal shield and upper cycle of central-area), central pen or spanned by disjunct bar on proximal surface.

Remarks: Circular species do not occur in Neogene.

Type species: *Coccolithus pelagicus* (Wallich, 1877) Schiller, 1930

Geological Time Data

Geological Time Periods:
Holocene
Paleocene
Biblio Reference:
Cenozoic calcareous nannofossils

Figure 2. Nannotax Screenshot.

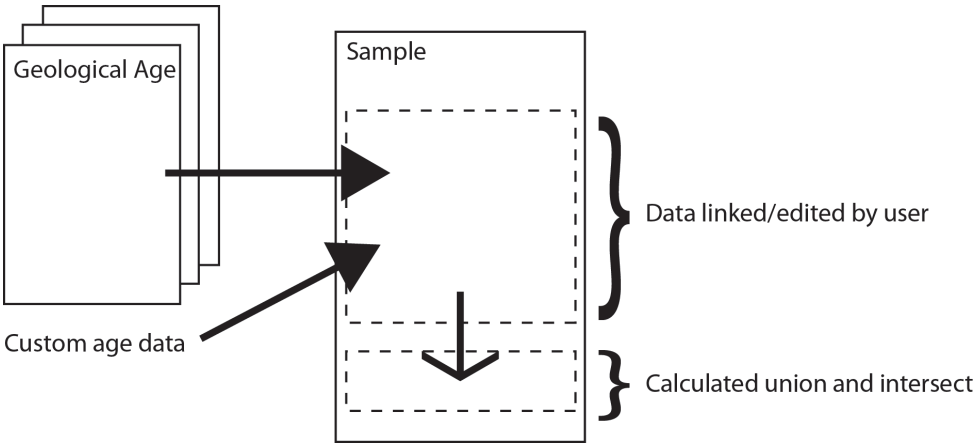


Figure 3. Data model for IPAEG.

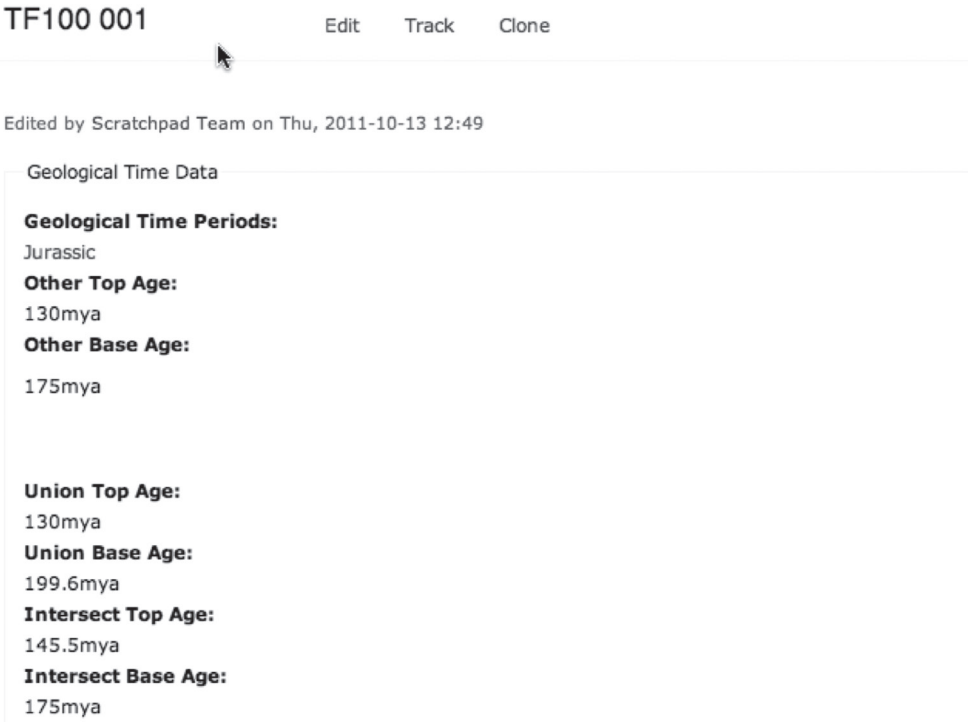


Figure 4. An example Sample record from IPAEG showing user-linked/edited age ranges (above) and calculated union and intersection dates (below).

may allow specified data to be acknowledged and referenced but excluded from the calculations.

The union gives the maximum possible time range for the species and be calculated for all GeoTime data sets. The intersect gives the overlapping range of the data sets and can only be calculated when there is a time period that is present across the data sets (Figures 5&6).

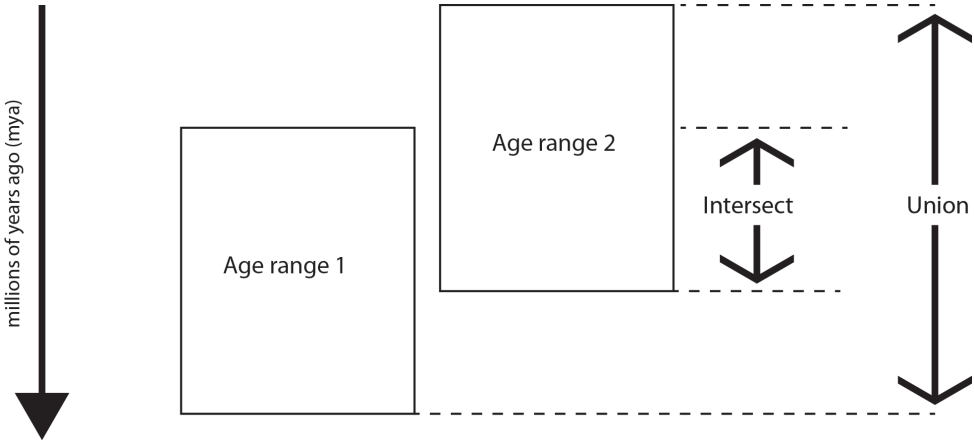


Figure 5. Calculation of the union and intersect.

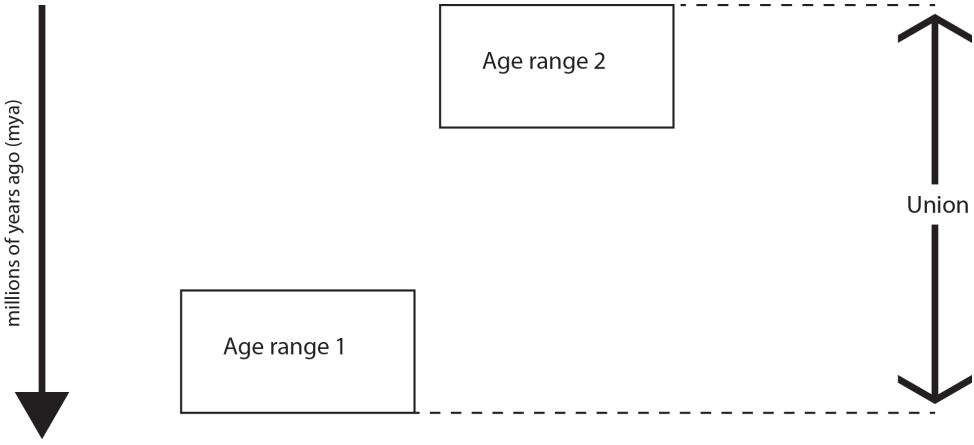


Figure 6. When there is no overlapping time periods the intersect is undefined.

Scratchpads 2.0 Implementation

The Scratchpads 2.0 implementation of the GeoTime module will allow for a variable number of age ranges (either predefined or custom) with individual references to be recorded. This is an improvement over the Scratchpads 1.0 implementation, which only allowed for one custom age range and a single reference to be given to the geological age datasets as a whole.

Issues

Given the nature of some geological age data (e.g. chronostratigraphy), it makes sense to associate these nodes with a Drupal taxonomy. In this model the Jurassic period has only one parent, the Mesozoic era, and several children, the Upper, Middle and Lower Jurassic. Attaching age metadata (e.g. top and base ages) to the taxonomy terms allows all records of a given term to be updated with a single change. The current Scratchpads implementation has a mechanism for achieving this but requires a separate content type for extending each taxonomy, plus a separate content type for ages not associated with a taxonomy. It was decided not to use this option due to the proliferation of content types required for sites dealing with multiple types of age data.

Future plans

We will create functionality to allow content to be searched using geological age data, either by union or intersect. Some example questions that could potentially be answered by this functionality are:

1. Which taxa were alive in age X?
2. Are there specimens of taxon X in age range Y?
3. Which taxa co-existed in time with taxon X?

For both questions 1 and 2 an important part of the functionality is that the age can be expressed in terms of multiple different systems - absolute age in Ma, chronostratigraphic stage or fossil zone. The query function will perform its search by converting both the recorded data, and the query parameters into absolute ages, and then converting again if necessary to display the required results. This allows any kind of primary data to be queried using the same interface and the results to be displayed in any appropriate format.

Scratchpads 2.0 will allow for data to be imported from the GBIF Vocabularies site dynamically, allowing for changes made to the metadata (e.g. base age, top age) of a geological age to be automatically propagated across the Scratchpads, making use of the GeoTime functionality.

Once a system has been created for recording geological age data the next obvious step is to create a way for these data to be displayed visually. One project that has been used to develop a relevant working example of age data is the SIMILE Timeline project (http://www.simile-widgets.org/timeline/); see http://simile.mit.edu/timeline/examples/dinosaurs/dinosaurs2.html for a geological example.

The Timeline widget has already been integrated with the Drupal views module (http://drupal.org/project/timeline) but, as yet, there is no Drupal 7 version. Migrating this code to Drupal 7 and adding support for geological age ranges (as in the above example) would allow for an aesthetically pleasing and easy-to-use visual layer to be applied to the data.

Going further

One possible use for the functionality developed here is to create a first and last occurrence database for a large number of taxa. This would become a useful resource for calibrating phylogenies (Marshall 2008) and studying changes in biotas through time (for example, Johnson et al. 2008). Additionally, range data can be used in conjunction with phylogenetic studies to help correct for incomplete sampling of the fossil record. This has been shown to alter the apparent nature of diversification in odonates (dragonflies, damselflies and extinct relatives) from an expansionist to a logistic model, with wider implications for palaeodiversity studies (Davis et al. 2011). An online repository of up-to-date range data would facilitate this type of work in future.

Although the functionality described is currently used for recording geological age data, the same functionality could be used to record and display data about other properties that can be measured in ranges, e.g. depth in sediment cores from lakes (e.g. Dalton et al. 2005).

The developed functionality could also be used in archaeological contexts by using new or modified vocabularies.

Moving beyond chronostratigraphy, it would be useful to develop processes to connect lithostratigraphic information into the scratchpad environment taking advantage of the stratigraphic lexicons published by national geological surveys (http://ngmdb.usgs.gov/Geolex) For example the formations found around Lyme Regis (e.g. Black Ven Marl, Belemnite Shales etc.). These could potentially be entered as synonyms of existing named time intervals, or added as a separate vocabulary. This method would allow for local stratigraphic data to be recorded in the Scratchpad system. An extended dataset of this nature would make it easier to integrate the Scratchpads with existing local, regional and global databases.

Acknowledgments

The authors would like to thank David Nicholson, Vladimir Blagoderov and Theresa Brown (all Natural History Museum, London) for commenting on drafts of this paper. Thanks also to David Remsen and Dag Endresen of GBIF for considering our requirements in the ongoing improvements to the GBIF Vocabularies site.

We are grateful for financial support for this project from the Gulf Coast Section of the Society of Economic Paleontologists & Mineralogists and NCB Naturalis (organised by Willem Renema) for providing financial support for software development.

This work uses infrastructure that has been developed by the EU funded ViBRANT project (Contract no. RI-261532).

References

Dalton C, Birks HJB, Brooks SJ, Cameron NG, Evershed RP, Peglar SM, Scott JA, Thompson R (2005) A multi-proxy study of lake-development in response to catchment changes during the Holocene at Lochnagar, north-east Scotland. Palaeogeography, Palaeoclimatology, Palaeoecology 221: 3–5 175–201. doi: 10.1016/j.palaeo.2005.02.007

Davis RB, Nicholson DB, Saunders ELR, Mayhew PJ (2011) Fossil gaps inferred from phylogenies alter the apparent nature of diversification in dragonflies and their relatives. BMC Evolutionary Biology 11: 252. doi: 10.1186/1471-2148-11-252

Gradstein FM, Ogg JG, Smith AG (2004) A Geologic Time Scale 2004. Cambridge University Press. doi: 10.4095/215638

Harman KT, Hyam R, Remsen DP (2009) Vocabularies - Managing Them. Proceedings of TDWG 2009. http://www.tdwg.org/proceedings/article/view/605

Johnson KG, Jackson JBC, Budd AF (2008) Caribbean reef development was independent of coral diversity over 28 million years. Science 319(5869): 1521–1523. doi: 10.1126/science.1152197

Marshall CR (2008) A simple method for bracketing absolute divergence times on molecular phylogenies using multiple fossil calibration points. American Naturalist 171(6): 726–742. doi: 10.1086/587523

Smith VS, Rycroft SD, Brake I, Scott B, Baker E, Livermore L, Blagoderov V, Roberts D (2011) Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 53–70. doi: 10.3897/zookeys.150.2193

Van Couvering JA, Ogg JG (2007) The future of the past: Geological time in the digital age. Stratigraphy. 4(2/3): 253–257.

ZooKeys 150: 407–417 (2011)
doi: 10.3897/zookeys.150.2002
www.zookeys.org

DATA PAPER

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Literature based species occurrence data of birds of northeast India

Sujit Narwade[1], Mohit Kalra[1], Rajkumar Jagdish[1], Divya Varier[1], Sagar Satpute[1], Noor Khan[1], Gautam Talukdar[2], V. B. Mathur[2], Karthikeyan Vasudevan[2], Dinesh Singh Pundir[2], Vishwas Chavan[3], Rajesh Sood[3]

1 *Bombay Natural History Society (BNHS), Shaheed Bhagatsingh Road, 400001, Mumbai, India* **2** *Wildlife Institute of India (WII), Post Box No 18, Chandrabani, 248001, Dehradun, India* **3** *Global Biodiversity Information Facility (GBIF), Universitetsparken 15, DK 2100, Copenhagen, Denmark*

Corresponding author: *Sujit Narwade* (bnhs@envis.nic.in)

Academic editor: *L. Penev* | Received 2 September 2011 | Accepted 24 November 2011 | Published 28 November 2011

Citation: Narwade S, Kalra M, Jagdish R, Varier D, Satpute S, Khan N, Talukdar G, Mathur VB, Vasudevan K, Pundir DS, Chavan V, Sood R (2011) Literature based species occurrence data of birds of northeast India. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 407–417. doi: 10.3897/zookeys.150.2002

Abstract

The northeast region of India is one of the world's most significant biodiversity hotspots. One of the richest bird areas in India, it is an important route for migratory birds and home to many endemic bird species. This paper describes a literature-based dataset of species occurrences of birds of northeast India. The occurrence records documented in the dataset are distributed across eleven provinces of India, viz.: Arunachal Pradesh, Assam, Bihar, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim, Tripura, Uttar Pradesh and West Bengal. The geospatial scope of the dataset represents 24 to 29 degree North latitude and 78 to 94 degree East longitude, and it comprises over 2400 occurrence records. These records have been collated from scholarly literature published between1915 and 2008, especially from the *Journal of the Bombay Natural History Society* (JBNHS). The temporal scale of the dataset represents bird observations recorded between 1909 and 2007. The dataset has been developed by employing MS Excel. The key elements in the database are scientific name, taxonomic classification, temporal and geospatial details including geo-coordinate precision, data collector, basis of record and primary source of the data record. The temporal and geospatial quality of more than 50% of the data records has been enhanced retrospectively. Where possible, data records are annotated with geospatial coordinate precision to the nearest minute. This dataset is being constantly updated with the addition of new data records, and quality enhancement of documented occurrences. The dataset can be used in species distribution and niche modeling studies. It is planned to expand the scope of the dataset to collate bird species occurrences across the Indian peninsula.

Keywords
India, northeast Himalaya, Assam, Arunachal Pradesh, Bihar, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim, Tripura, West Bengal, Uttar Pradesh, *Journal of the Bombay Natural History Society*, BNHS, Animalia, Chordata, Aves

Data published through GBIF: http://ibif.gov.in:8080/ipt/resource.do?r=BNHS-NEW

Taxonomic coverage

General taxonomic coverage description: The taxonomic coverage of this dataset spans Class Aves. The highest number of data records are from the family Muscicapidae (560 records), followed by Anatidae (180 records) and Accipitridae (136 records). The families with the least number of records are Hemiprocnidae, Podargidae, Indicatoridae with one data record each.

Taxonomic ranks

Kingdom: Animalia
Phylum: Chordata
Class: Aves
Order: Podicipediformes, Pelecaniformes, Ciconiiformes, Anseriformes, Falconiformes, Galliformes, Gruiformes, Charadriiformes, Columbiformes, Psittaciformes, Cuculiformes, Strigiformes, Caprimulgiformes, Apodiformes, Trogoniformes, Coraciiformes, Piciformes, Passeriformes
Family: Podicipedidae, Phalacrocoracidae, Anhingidae, Ardeidae, Ciconiidae, Threskiornithidae, Anatidae, Accipitridae, Pandionidae, Falconidae, Phasianidae, Turnicidae, Gruidae, Rallidae, Otididae, Jacanidae, Rostratulidae, Charadriidae, Scolopacidae, Recurvirostridae, Glareolidae, Laridae, Rhynchopidae, Columbidae, Psittacidae, Cuculidae, Tytonidae, Strigidae, Podargidae, Caprimulgidae, Apodidae, Hemiprocnidae, Trogonidae, Alcedinidae, Meropidae, Coraciidae, Upupidae, Bucerotidae, Capitonidae, Indicatoridae, Picidae, Eurylaimidae, Pittidae, Alaudidae, Hirundinidae, Motacillidae, Campephagidae, Pycnonotidae, Irenidae, Laniidae, Troglodytidae, Prunellidae, Muscicapidae, Aegithalidae, Paridae, Sittidae, Certhiidae, Dicaeidae, Nectariniidae, Zosteropidae, Emberizidae, Fringillidae, Estrildidae, Passeridae, Sturnidae, Oriolidae, Dicruridae, Artamidae, Corvidae.

Spatial coverage

General spatial coverage: This dataset collates species occurrences from northeast India and neighboring regions. The occurrence records collated in the dataset are *Literature-based species occurrence data of birds of North-East India 3* distributed across

eleven states of India, viz.: Arunachal Pradesh, Assam, Bihar, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim, Tripura, Uttar Pradesh and West Bengal. This region falls within the Himalayan mountain ranges, and spans an area of 234567 sq. km. The region borders with Bangladesh to the south, Bhutan to the west, Myanmar to the east and with China to the north. Minimum and maximum elevations are 2000 meters and 8000 meters above sea level respectively.

Coordinates: 24°30'0''N - 28°15'0''N Latitude; 78°22'58.8''E - 93°47'60''E Longitude.

Temporal coverage

1909–2007.

Project description

Title: This dataset is an outcome of the collaborative work carried out by three projects, viz.:, (a) Environmental Information System (ENVIS) Centre on Avian Ecology, Bombay Natural History Society, sponsored by the Ministry of Environment and Forests, Government of India; (b) Important Bird Areas Programme and Indian Bird Conservation Network (IBA-IBCN), sponsored by the Royal Society for Protection of Birds, United Kingdom; and (c) Impact of Climate Change on the Conservation of Birds, a project supported by the MacArthur Foundation.

Personnel: Sujit Narwade (Author, Content Provider, Metadata Provider), Mohit Kalra (Processor), Divya Varier (Custodian Steward/Metadata Provider), Rajkumar Jagdish (Processor), Sagar Satpute (Custodian Steward), Noor Khan (Custodian Steward), Gautam Talukdar (Publisher), V.B. Mathur (Publisher), Karthik Vasudevan (Publisher), Dinesh Singh Pundir (Publisher), Vishwas Chavan (Metadata Provider/Editor), Rajesh Sood (Metadata Provider/Programmer).

Funding: Ministry of Environment and Forests, Government of India, (Funding), the Royal Society for Protection of Birds, United Kingdom (Funding), MacArthur Foundation (Funding), Bombay Natural History Society (Host institution), and Wildlife Institute of India (publishing support).

Study area descriptions/descriptor: Northeast India is one of the most significant biodiversity hotspots of the world and among the richest bird zones in India. It is considered as the 'biological gateway' for much of India's fauna, as the Gondwana land first touched this region, during the Tertiary period. The north-eastern region is at the confluence of the Indo-Malayan, Indo-Chinese and Indian biogeographical realms. As a result, it is unique in providing a profusion of habitats that harbor diverse biota with a high degree of endemism (Chatterji et al. 2006).

Design description: The Environmental Information System (ENVIS) Centre at the Bombay Natural History Society (BNHS) is a focal point for collection, collation

and dissemination of data on avian ecology in India. The objective of this dataset is to collate avian observations documented in various research publications such as journals, magazines, newsletters, project reports, theses, books and other gray literature. However, the current version of the dataset collates occurrence records reported in research articles published in the *Journal of the Bombay Natural History Society* (JBNHS). The motivation for this approach is because of the dispersed availability of the occurrence records in published literature. Because these data records are documented in several literature sources, it is difficult to access them together, inspite of being in the public domain. Thus, for a potential user it is not possible to access and use them when he/she needs them the most. Another consideration is the quality of these data records, as they are published in peer reviewed literature and their quality is expected to be 'fit for use'. The data records were entered in a MS-Excel worksheet. The offline version of the dataset maintained by the Bombay Natural History Society is developed employing MS-Access. The key elements about which information is collated in the current version of the dataset includes: scientific name, common name, taxonomic classification, occurrence location, geo-coordinates, precision of geo-coordinates, date of observation, data collector, and primary source of the data record.

The data records were entered in the MS-Excel worksheet. The offline version of the dataset maintained by the Bombay Natural History Society is developed employing the MS-Access. The key elements about which information is collated in the current version of the dataset includes, scientific name, common name, taxonomic classification, occurrence location, geo-coordinates, precision of geo-coordinates, date of observation, data collector, and primary source of the data record.

Dataset description

Object name: Darwin Core Archive literature-based species occurrence data of birds of northeast India
Character encoding: UTF-8
Format name: Darwin Core Archive format
Format version: 1.0
Distribution: http://ibif.gov.in:8080/ipt/archive.do?r=BNHS-NEW
Publication date of data: 2011-06-30
Language: English
Licenses of use: by-nc-sa
Metadata language: English
Date of metadata creation: 2011-06-30
Hierarchy level: Dataset

Additional information

We are most thankful to the Ministry of Environment and Forests (MoEF), Government of India; the Royal Society for the Protection of Birds (RSPB), United Kingdom; and the MacArthur Foundation for financial support. We would like to express our deepest gratitude to Dr. Asad R. Rahmani, Director, BNHS and Principal Investigators of the aforementioned projects.

References

Referred for dataset

Abdulali H (1983) Occurrence of the Great Crested Grebe: *Podiceps cristatus* (Linne) at Ranchi. Bihar. J. Bombay Nat. Hist. Soc. 80 (2) 414–415.

Abdulali H (1954) More notes on Finn's Baya (*Ploceus megarhynchus*): J. Bombay Nat. Hist. Soc. 52 (2&3): 599–601.

Abdulali H (1957) Some notes on the plumages of *Centropus sinensis* (Stephens): J. Bombay Nat. Hist. Soc. 54 (1) 183–185.

Abdulali H (1961) The nesting habits of the eastern race of Finn's Baya, *Ploceus megarhynchus* Salim Ali Abdulali: J. Bombay Nat. Hist. Soc. 58 (1) 269–270.

Abdulali H (1964) Notes on Indian birds 1-*Ceyx erithacus rufidorsus* Strickland in the Sikkim Terai, Eastern Himalayas: An addition to the Indian avifauna. J. Bombay Nat. Hist. Soc. 61 (2) 439–440.

Abdulali H (1965) Behaviour of Lesser Whistling Teal [*Dendrocygna javanica* (Horsfield)] in Alipore zoo Calcutta: J. Bombay Nat. Hist. Soc. 62 (2) 300–301.

Allen D, Holt PI, Hornbuckle J (2002) Leaf-presenting as possible courtship behaviour by Pied Falconets *Microhierax melanoleucos*: J. Bombay Nat. Hist. Soc. 99 (3) 518–520.

Altevogi R, Davis TA (1979) Urbanization in nest building of Indian House Crow (*Corvus splendens* Vieillot): J. Bombay Nat. Hist. Soc. 76 (2) 283–290.

Barua M (2002) Occurrence of the Indian Skimmer *Rhynchops albicollis* Swainson in Assam: J. Bombay Nat. Hist. Soc. 99 (3) 526.

Baruah M, Chettri G, Bordoloi P (2004) Sighting of White-bellied Heron *Ardea insignis* Hume in Pobitora wildlife sanctuary: J. Bombay Nat. Hist. Soc. 101 (2) 311.

Bertram B (1967) Hill Myna *Gracula religiosa* Linnaeus breeding in artificial nests in Garo hills Assam: J. Bombay Nat. Hist. Soc. 64 (2) 369–370.

Betts FN (1952) Bird nesting on telegraph wires: J. Bombay Nat. Hist. Soc. 51 (1) 271–272.

Betts FN (1955) Notes on birds of the Subansiri area Assam: J. Bombay Nat. Hist. Soc. 53 (4) 397–414.

Burley H (1954) Occurrence of the Blacknecked Crane (*Grus nigricollis*) in Indian limits: J. Bombay Nat. Hist. Soc. 52 (2&3): 605–606.

Burnett JH (1958) Photographing a colony of Egrets (*Bubulcus ibis* and *Egretta garzetta*) in Assam: J. Bombay Nat. Hist. Soc. 55 (3) 565–566.

Chatterjee S (1995) Occurrence of Albino Lesser Whistling Teal: *Dendrocygna javanica* (Horsfield). J. Bombay Nat. Hist. Soc. 92 (3) 417–418.

Chatterjee S, Mookerjee K, Bhattacharya B, Banerjee A (1995) Occurrence of Falcated Teal *Anas falcata* Georgi in West Bengal: J. Bombay Nat. Hist. Soc. 92 (2) 262.

Chattopadhyay S (1978) Occurrence of the Thickbilled Warbler *Phragmaticola aedon rufescens* Stegmann at Baj Baj West Bengal: J. Bombay Nat. Hist. Soc. 75 (2) 491–492.

Chattopadhyay S (1980) Egg-bound death of a Purplerumped Sunbird at Baj Baj West Bengal: J. Bombay Nat. Hist. Soc. 77 (2) 333.

Chattopadhyay S(1980) Observations on parental care of a wounded chick of the Bronze-winged Jacana: *Metopidius indicus* (Latham). J. Bombay Nat. Hist. Soc. 77 (2) 325–326.

Choudhury A (1991) Purple-rumped Sunbird *Nectarinia zeylonica* (Linn.): A new record for Assam. J. Bombay Nat. Hist. Soc. 88 (1) 114.

Choudhury A (1992) Addition to the birds of Assam - Blacknecked *Grebe Podiceps nigricollis* Brehm: J. Bombay Nat. Hist. Soc. 89 (2) 245–246

Choudhury A (1993) Additions to the birds of Assam: White-tailed Sea eagle and Large Sand Plover. J. Bombay Nat. Hist. Soc. 91 (1) 139.

Choudhury A (1993) On a possible sight records of the Little Gull *Larus minutus* Pallas in Arunachal Pradesh: J. Bombay Nat. Hist. Soc. 90 (2) 290.

Choudhury A (1998) Common Myna feeding a fledgling Koel: J. Bombay Nat. Hist. Soc. 95 (1) 115.

Choudhury A (1998) The Bengal Florican *Eupodotis bengalensis* Gmelin 1789 in Dibang valley district of Arunachal Pradesh: J. Bombay Nat. Hist. Soc. 95 (2) 342.

Choudhury A (2004) Sighting of Wallcreeper *Tichodroma muraria* in Assam and Manipur: J. Bombay Nat. Hist. Soc. 101 (3) 463.

Choudhury A (2005) Great-tufted Myna *Acridotheres grandis* - An addition to the birds of Meghalaya: J. Bombay Nat. Hist. Soc. 102 (1) 117.

Choudhury A (2005) Migration of Black-eared or Large Indian Kite *Milvus migrans lineatus* (Gray) from Mongolia to North-eastern India: J. Bombay Nat. Hist. Soc. 102 (2) 229–230.

Das PK (1965) The Whitecheeked Drongo [*Dicrurus leucophaeus salangensis* Reichenow] : An Addition to the Indian avifauna. J. Bombay Nat. Hist. Soc. 62 (3) 557–558.

Datta A (2004) Sighting of the Oriental Bay-Owl *Phodilus badius saturatus* in Pakhui Wildlife sanctuary Western Arunachal Pradesh: J. Bombay Nat. Hist. Soc. 101 (1) 156.

Davis TA (1973) Mud and dung plastering in Baya nests: J. Bombay Nat. Hist. Soc. 70 (1) 57–71.

Editors (1952) The whimbrel (*Numenius phaeopus*) in Assam: J. Bombay Nat. Hist. Soc. 50 (3) 663.

Editors BNHS (1958) The Avocet (*Recurvirostra avosetta* Linn.) in Assam: J. Bombay Nat. Hist. Soc. 55 (1) 170.

Fooks HA (1966) Whistling Teal [*Dendrocygna javanica* (Horsefield)] and other memories of Alipore zoo Calcutta: J. Bombay Nat. Hist. Soc. 63 (1) 200–202.

Ganguli U (1990) Blackwinged Kite *Elanus caeruleus vociferus* (Latham) at 3650m in Sikkim: J. Bombay Nat. Hist. Soc. 87 (1) 142.

Ganguli U (1990) Brahminy Duck *Tadorna ferruginea* (Pallas) Breeding in Sikkim: J. Bombay Nat. Hist. Soc. 87 (2) 290.

Ganguli U (1990) Osprey *Pandion haliaetus* in Sikkim: J. Bombay Nat. Hist. Soc. 87 (2) 291.
Ganguli-Lachungpa U, Lucksom S (1998) Sighting of Hodgson's Frogmouth *Batrachostomus hodgsoni hodgsoni* (G.R. Gray) from Sikkim: J. Bombay Nat. Hist. Soc. 95 (3) 505.
Ganguli-Lachungpa U, Lucksom S (1998) Western Greyheaded Thrush *Turdus rubrocanus* G.R. Gray in Sikkim: J. Bombay Nat. Hist. Soc. 95 (3) 508–509.
Ganguli-Lachungpa U (1991) Occurrence of Blacknecked Grebe *Podiceps nigricollis* Brehm: Little Grebe *P. ruficollis* (Pallas) and Goosander *Mergus merganser* Linn. in West Sikkim. J. Bombay Nat. Hist. Soc. 88 (2) 280.
Ganguli-Lachungpa U (1998) Attempted breeding of the Blacknecked Crane *Grus nigricollis* Przevalski in North Sikkim: J. Bombay Nat. Hist. Soc. 95 (2) 341.
Ganguli-Lachungpa U (2002) Eurasian Eagle-owl *Bubo bubo tibetanus* Bianchi at 2,100 M in north Sikkim: J. Bombay Nat. Hist. Soc. 99 (2) 305–306.
Ganguli-Lachungpa U (2003) Common coot *Fulica atra* from Kyongnosla in East Sikkim: J. Bombay Nat. Hist. Soc. 100 (1) 121.
Ganguly JK (1986) Co-operative feeding of chicks of the Purple-rumped Sunbird (*Nectarinia zeylonica*): J. Bombay Nat. Hist. Soc. 83 (2) 447.
Gauntlett FM (1971) Durgapur barrage as a waterbird habitat: J. Bombay Nat. Hist. Soc. 68 (3) 619–632.
Gauntlett FM (1985) The birds of Durgapur and the Damodar valley: J. Bombay Nat. Hist. Soc. 82 (2) 501–539.
Gee EP (1958) The present status of the Whitewinged Wood Duck *Cairina scutulata* (S. Muller): J. Bombay Nat. Hist. Soc. 55 (3) 569–575.
George PV (1967) On the occurrence of the Fulvous-brested Woodpecker *Dendrocopos macei* (Vieillot) in Sikkim: J. Bombay Nat. Hist. Soc. 64 (3) 559–560.
Ghose D, Khan S (2005) Albino bulbul at Keibul Lamjao national park, Manipur India: J. Bombay Nat. Hist. Soc. 102 (1) 120–121.
Green MJB (1986) The birds of the Kedarnath sanctuary Chamoli district Uttar Pradesh: Status and distribution. J. Bombay Nat. Hist. Soc. 83 (3) 603 – 617.
Gupta KK (1966) Aggressive behaviour of a Spotted Owlet [*Athene brahma* (Temminck)]: J. Bombay Nat. Hist. Soc. 63 (2) 441–442.
Gupta KK (1995) A note on Baya, *Ploceus philippinus* nesting on Krishnachuda (Delonix regia) tree: J. Bombay Nat. Hist. Soc. 92 (1) 124–125.
Haribal M, Ganguli-Lachungpa (1991) Black Woodpecker *Dryocopus Sp*. In Jaldapara sanctuary West Bengal: J. Bombay Nat. Hist. Soc. 88 (1) 112.
Harington HH (1915) Notes on Indian Timaliides and their allies: Pt. 4. J. Bombay Nat. Hist. Soc. 23: 614–657.
Holmes JRS (1968) New wintering locality of the Spotted Bush Warbler *Bradypterus thoracicus* (Blyth): J. Bombay Nat. Hist. Soc. 65 (3) 779–780.
Hopper CD (1958) Occurrence of the Baikal Teal *Nettion formosum* (Georgi) in Assam: J. Bombay Nat. Hist. Soc. 55 (2) 359–360.
Hussain SA (1984) *Hypsipetes madagascariensis sinensis* (La touch): A first record for India. J. Bombay Nat. Hist. Soc. 81 (1) 195–196.

Jackson PFR (1974) Goliath Heron in the Sundarbans West Bengal: J. Bombay Nat. Hist. Soc. 71 (3) 608–609.

Javed S (1995) Hare in the diet of White-eyed Buzzard Eagle *Butastur teesa* (Franklin): J. Bombay Nat. Hist. Soc. 92 (1) 119.

Jha A (2001) Competition between Jungle Myna *Acridotheres fuscus* and Lesser Goldenbacked Woodpecker *Dinopium benghalense* for a nest hole: J. Bombay Nat. Hist. Soc. 98 (1) 115.

Jha A (2002) More evidence of Red-vented Bulbul *Pycnonotus cafer* feeding on House Gecko *Hemidactylus flaviviridis*: J. Bombay Nat. Hist. Soc. 99 (1) 118.

Jha S (1994) An Albino Myna *Acridotheres tristis* (Linnaeus): J. Bombay Nat. Hist. Soc. 91 (3) 455.

Jha S (2000) A note on the feeding of Lesser Coucal (*Centropus toulou*): J. Bombay Nat. Hist. Soc. 97 (1) 144.

Jha S (2002) Attempted feeding by a Shikra *Accipiter badius*, family Accipitridae on Buffstriped Keelback *Amphiesma stolata* family Colubridae: J. Bombay Nat. Hist. Soc. 99 (2) 298.

Kalita SK (2000) Competition for food between a Garden Lizard *Calotes versicolor* (Daudin) and a Magpie Robin *Copsychus saularis* Linn: J. Bombay Nat. Hist. Soc. 97 (3) 431

Kumar RS (2003) Ring recovery from Great Cormorants *Phalacrocorax carbo* in India: J. Bombay Nat. Hist. Soc. 100 (3) 621–624.

Kumar RS (2004) Common Starling *Sturnus vulgaris* in Arunachal Pradesh India: J. Bombay Nat. Hist. Soc. 101 (2) 320.

Lahkar B (2000) Pallas's fishing Eagle *Haliaeetus leucoryphus* (Pallas) pirates fish from an Otter *Lutra lutra* (Linn.): J. Bombay Nat. Hist. Soc. 97 (3) 425.

Lahkar K, Phukan MP (2007) Wintering range extension of White-throated Bushchat *Saxicola insignis* (Gray) in India: J. Bombay Nat. Hist. Soc. 104 (3) 348–349.

Lamba BS (1981) A queer nesting site of Bank Myna, *Acridotheres ginginianus*: J. Bombay Nat. Hist. Soc. 78 (3) 605–606.

Law SC (1952) Occurrence of the Smew [*Mergellus albellus* (Linn.)] in West Bengal: J. Bombay Nat. Hist. Soc. 51 (2) 508–509.

Lister MD (1952) Secondary song of some Indian birds: J. Bombay Nat. Hist. Soc. 51 (3) 699–706

Lister MD (1954) A contribution to the ornithology of the Darjeeling area: J. Bombay Nat. Hist. Soc. 52 (1) 20–68

Madge SC (1984) First Indian record of Chaffinch (*Fringilla coelebs*): J. Bombay Nat. Hist. Soc. 81 (3) 702–703.

Mansion P (1967) The whistling Teal [*Dendrocygna javanica* (Horsfield)] in the Calcutta environs: J. Bombay Nat. Hist. Soc. 64 (3) 558–559.

Matthews WH (1952) Breeding of *Rallina euryzonoides nigrolineata* (Gray) in the Darjeeling district: J. Bombay Nat. Hist. Soc. 51 (3) 742–743.

Meinertzhagen R(1952) A new bird for India - *Montifringilla davidiana potanini* (Sushkin): J. Bombay Nat. Hist. Soc. 51 (1) 273–274.

Mohan D, Chellam R (1990) New call record of Greenbreasted Pitta *Pitta sordida* (P.L.S. Muller) in Dehra Dun Uttar Pradesh: J. Bombay Nat. Hist. Soc. 87 (3) 453.

Mohan D, Rai ND, Singh AP (1992) Longtailed Duck or Old Squaw *Clangula hyemalis* (Linn.) in Dehra Dun Uttar Pradesh: J. Bombay Nat. Hist. Soc. 89 (2) 247.

Mukherjee AK (1969) Food-habits of water-birds of the Sundarban: 24-Parganas district West Bengal India. J. Bombay Nat. Hist. Soc. 66 (2) 345–360.

Mukherjee AK (1971) Food-Habits of water-birds of the Sundarban: 24-Parganas district West Bengal India – III. J. Bombay Nat. Hist. Soc. 68 (1) 690 – 716.

Mukherjee AK (1974) Food-habits of water-birds of the Sundarban: 24 Parganas district West Bengal India – IV. J. Bombay Nat. Hist. Soc. 71 (2) 188–200.

Mukherjee AK (1975) Food-habits of waterbirds of the Sundarban: 24 Parganas district West Bengal India – V. J. Bombay Nat. Hist. Soc. 72 (2) 422–447.

Mukherjee AK (1975) The Sundarban of India and its biota: J. Bombay Nat. Hist. Soc. 72 (1) 1–20.

Mukherjee AK (1976) Food-habits of water-birds of the Sundarban: 24 Parganas district West Bengal India – VI. J. Bombay Nat. Hist. Soc. 73 (2) 482- 486.

Mukhopadhyay A (1980) Some observations on the biology of the Openbill Stork: *Anastomus oscitans* (Boddaert) in Southern Bengal. J. Bombay Nat. Hist. Soc. 77 (1) 133–137.

Mukhrjee AK (1971) Food-habits of water-birds of the Sundarban: 24-Parganas district West Bengal India – II. J. Bombay Nat. Hist. Soc. 68 (1) 37–64.

Murthy S (1954) An intelligent Myna: J. Bombay Nat. Hist. Soc. 52 (2&3): 598.

Naoroji R, Sangha HS, Barua M (2005) The Lesser Kestrel *Falco naumanni* and Amur Falcon *Falco amurensis* in the Garo hills: Meghalaya India. J. Bombay Nat. Hist. Soc. 102 (1) 103–105.

Narayan G, Rosalind L (1991) New record of the Pied Harrier *Circus melanoleucos* (Pennant) breeding in Assam Duars, with a brief review of its distribution: J. Bombay Nat. Hist. Soc. 88 (1) 30–34.

Narayan G, & Rosalind L (1994) Wintering and time extension of Hodgson's Bush Chat *Saxicola insignis* (Gray) in India: J. Bombay Nat. Hist. Soc. 94 (3) 572–573.

Rahmani AR, Sankaran R (1990) An unusual nesting site of the Sunbird: J. Bombay Nat. Hist. Soc. 87 (1) 148–149.

Rahmani AR (1981) Large Racket-tailed Drongo and Common Babbler: J. Bombay Nat. Hist. Soc. 78 (2) 380.

Rahmani AR (1981) Narora reservoir U.P - A potential bird sanctuary: J. Bombay Nat. Hist. Soc. 78 (1) 88–92.

Rahmani AR (1991) Status of the Bengal Florican *Houbaropsis bengalensis* in India: J. Bombay Nat. Hist. Soc. 88 (3) 349–375.

Rao KR, Zoramthanga R (1978) On the phenomenon of nocturnal flights of some resident birds at Lunglei: Mizoram N.E. India. J. Bombay Nat. Hist. Soc. 75 (3) 927–928.

Rao P, Murlidharan S (1989) Unusual feeding behaviour in the Adjutant Stork *Leptoptilos dubius* (Gmelin): J. Bombay Nat. Hist. Soc. 86 (1) 97.

Rao P, Grubh RB, Muralidharan S (1989) Range extension of Eurasian Griffon Vulture *Gyps fulvus*: J. Bombay Nat. Hist. Soc. 86 (2) 240–241.

Rasool TJ (1984) Some observations on Natural Cheer Pheasant, *Catreus wallichii*: Population at Mukteshwar reserve forest. Kumaon, Nainital UP. J. Bombay Nat. Hist. Soc. 81 (2) 469–471.

Raza R (1993) Sighting of Black Bulbul *Hypsipetes madagascariensis* (P.L.S. Muller) in Gaya, Bihar: J. Bombay Nat. Hist. Soc. 90 (2) 291.

Ripley SD (1952) A collection of birds from the Naga hills: J. Bombay Nat. Hist. Soc. 50 (3) 475–514.

Ripley SD. (1980) A new species, and a new subspecies of bird from trip district Arunachal Pradesh and comments on the subspecies of *Stachyris nigriceps* Blyth: J. Bombay Nat. Hist. Soc. 77 (1) 1–5.

Ritschard M, Taschler A (2008) A recent observation of White-headed *Duck Oxyura leucocephala* at Gajaldoba barrage West Bengal India: J. Bombay Nat. Hist. Soc. 105 (1) 95.

Ritschard M, Logtmeijer P, Taschler A (2008) Two observations of Malayan Night-heron *Gorsachius melanophus* from West Bengal, India: J. Bombay Nat. Hist. Soc. 105 (1) 97–98.

Saha BC, Mukherjee AK (1978) Notes on the food of the Blackheaded Munia and the Spotted Munia in South Kamrup district Western Assam (India): J. Bombay Nat. Hist. Soc. 75 (1) 221–224.

Saha SS (1976) Occurrence of Finn's Baya (*Ploceus megarhynchus* Hume) in Dirrang district: Assam. J. Bombay Nat. Hist. Soc. 73 (3) 527–529.

Saha SS (1980) Blacknecked Crane in Bhutan and Arunachal Pradesh - A survey report for January- February 1978: J. Bombay Nat. Hist. Soc. 77 (2) 326–328.

Saha SS, George PV, Ghosal DK, Mookerjee HP, Poddar AK, Ghose RK, Das PK, Gogate VG, Biswas B (1971) Notes on some interesting birds from the salt lakes, near Calcutta: J. Bombay Nat. Hist. Soc. 68 (2) 455–457.

Sankaran R (1989) Range extension of Yellowbellied Wren-Warbler *Prinia flaviventris*: J. Bombay Nat. Hist. Soc. 86 (3) 451.

Sankaran R, Rahmani AR, Ganguli-Lachungpa U (1992) The distribution and status of the Lesser Florican *Sypheotides indica* (J.F. Miller) in the Indian subcontinent: J. Bombay Nat. Hist. Soc. 89 (2) 156–179.

Sarma P, Barua M, Menon V (1997) Orangebilled Jungle Mynah and Hodgson's Bushchat in Kaziranga National Park: J. Bombay Nat. Hist. Soc. 94 (1) 156–157.

Satheesan SM (1990) Biometrics and food of some doves of the genus *Streptopelia*: J. Bombay Nat. Hist. Soc. 87 (3) 452–453.

Satheesan SM (1993) Extension of range of the Kashmir Roller (Blue Jay) *Coracias garrulus* to Gorakhpur Uttar Pradesh: J. Bombay Nat. Hist. Soc. 90 (1) 95.

Sendall D (1952) Occurrence of the Avocet (*Recurvirostra avocetta* Linn.) in Assam: J. Bombay Nat. Hist. Soc. 50 (4) 947.

Sengupta S (1973) Significance of communal roosting in the Common Myna [*Acridotheres tristis* (Linn.)]: J. Bombay Nat. Hist. Soc. 70 (1) 204–206.

Sengupta S (1975) Further note on the pair formation of the Common Myna, *Acridotheres tristis*: J. Bombay Nat. Hist. Soc. 72 (3) 856–857.

Sengupta SN (1969) Nest protection by the Indian House Crow (*Corvus splendens* Linnaeus): J. Bombay Nat. Hist. Soc. 66 (2) 377–378.

Sharma A, Zockler C (2008) First record of Caspian Gulls *Larus cachinnans* in the Indian Sunderbans delta: J. Bombay Nat. Hist. Soc. 105 (1) 93–94.

Sharma A (2008) Record of Large congregation of Large Whistling-Duck *Dendrocygna bicolor* in the Purbasthali-ganges islets Burdwan district West Bengal: J. Bombay Nat. Hist. Soc. 105 (1) 97.

Sharma A (2008) Sighting of Indian Skimmer *Rhynchops albicollis* (Swainson) in the Purbasthali-ganges islets Burdwan district West Bengal: J. Bombay Nat. Hist. Soc. 105 (1) 92–93.

Singh KS (1998) Aerial display of Rufous Turtle Dove *Streptopelia orientalis agricola* Tickell near Nambol bazaar Manipur: J. Bombay Nat. Hist. Soc. 95 (1) 114.

Singh P (1992) Spotted Longtailed Wren-babbler *Spelaeornis troglodytoides* (Verreaux) in Arunachal Pradesh: J. Bombay Nat. Hist. Soc. 89 (3) 376.

Singh P (1995) Occurrence of Swamp Partridge *Francolinus gularis* (Temminck) in Arunachal Pradesh: J. Bombay Nat. Hist. Soc. 92 (3) 419.

Singha H, Rahmani AR, Coulter MC, Javed S (2003) Breeding behaviour of the Greater Adjutant-stork *Leptotilos dubius* in Assam, India: J. Bombay Nat. Hist. Soc. 100 (1) 9–26.

Singha H, Karim R, Rahmani AR (1999) *Menopon gallinae* infesting Greater Adjutant Stork *Leptoptilos dubius* at Nagaon Assam: J. Bombay Nat. Hist. Soc. 96 (1) 137–138.

Sivakumar S, Prakash V, (2004) Cat snake *Boiga trigonata* in diet of Jerdon's *Baza Aviceda jerdoni*: J. Bombay Nat. Hist. Soc. 101 (3) 445–446.

Taylor JN (1954) Occurrence of Bronzecapped or Falcated Teal (*Eunetta falcata*) near Calcutta: J. Bombay Nat. Hist. Soc. 52 (2&3): 607.

Yahya SA (1981) Golden Oriole (*Oriolus oriolus*) feeding a fledgling Cuckoo (*Cuculus sp.):* J. Bombay Nat. Hist. Soc. 78 (2) 379–380.

Islam MZ, Rahmani AR (2004) Important bird areas in India: Priority sites for conservation. Indian Bird Conservation Network, BNHS and BirdLife International. Pp xviii + 1133.

The following publication is referred in the metadata text

Chatterjee S, Saikia A, Dutta P, Ghosh D, Pangging G, Goswami AK (2006) Biodiversity significance of north east India: WWF-India. New Delhi. Pp 71.

Appendix I

Literature-based species occurrence data of birds of North-East India. (doi: 10.3897/zookeys.150.2002.app) File format: XLS

Explanation note: This is an Excel spreadsheet of the dataset, available through the Darwin Core Archive format at: http://ibif.gov.in:8080/ipt/resource.do?r=BNHS-NEW

Citation: Narwade S, Kalra M, Jagdish R, Varier D, Satpute S, Khan N, Talukdar G, Mathur VB, Vasudevan K, Pundir DS, Chavan V, Sood R (2011) Literature based species occurrence data of birds of North-East India. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 407–417. doi: 10.3897/zookeys.150.2002.app