ZooKeys 11: 1-8 (2009)
doi: 10.3897/zookeys.11.210
www.pensoftonline.net/zookeys

FORUM PAPER

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

Publication and dissemination of datasets in taxonomy: ZooKeys working example

Lyubomir Penev[1], Terry Erwin[2], Jeremy Miller[3,6], Vishwas Chavan[4], Tom Moritz[5], Charles Griswold[6]

1 *Central Laboratory of General Ecology, Bulgarian Academy of Sciences and Pensoft Publishers, Sofia, Bulgaria* **2** *Smithsonian Institution, Washington DC, USA* **3** *Department of Terrestrial Zoology, Nationaal Natuurhistorisch Museum Naturalis, Leiden, The Netherlands* **4** *Global Biodiversity Information Facility, Copenhagen, Denmark* **5** *1968 1/2 South Shenandoah Street, Los Angeles, California, USA; formerly: Harold Boechenstein Director, Library Services, American Museum of Natural History, New York, USA* **6** *Department of Entomology, California Academy of Sciences, San Francisco, USA*

Corresponding authors: *Lyubomir Penev* (info@pensoft.net), *Jeremy Miller* (miller@naturalis.nl), *Vishwas Chavan* (vchavan@gbif.org)

Received 27 May 2008 | Accepted 30 May 2009 | Published 1 June 2009

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

Abstract

A concept for data publication and semantic enhancements proposed by ZooKeys and applied in the milestone paper by Miller et al. (2009) is described. For the first time in systematic zoology, a unique combination of data publication and semantic enhancements is applied within the mainstream process of journal publishing, to demonstrate how: (1) All primary biodiversity data underlying a taxonomic monograph are published as a dataset under a separate DOI within a paper; (2) The occurrence dataset is separately discoverable and accessible through GBIF data portal (data.gbif.org) simultaneously with the publication; (3) The occurrence dataset is published as a KML (Keyhole Markup Language) file under a distinct DOI to provide an interactive experience in Google Earth; (4) All new taxa (42) are registered at ZooBank during the publication process (mandatory for ZooKeys); (5) All new taxa (42) are provided to Encyclopedia of Life through XML mark up on the day of publication (mandatory for ZooKeys). It is proposed to clearly distinguish between static and dynamic datasets in the way they are published, preserved and cited.

Keywords

Data publication, semantic enhancements, taxonomy

Introduction

The publication of datasets is currently being extensively discussed in scientific publishing. The need for open access to research data consequently ensuring its preservation, dissemination and reuse is recognized as a strategic goal in several documents at governmental and international levels. For instance, at the European Union (EU) level it is recognized that long-term sustainability will be achieved not only by digital data storage but even more so by increasing probability of data reuse. Such a perspective is reflected by the statement of the Council of Europe on "the importance of better access to unprocessed data and repository resources for data and material that allows fresh analysis and utilisation beyond what the originator of the data had envisaged" (2832nd COMPETITIVENESS (Internal market, Industry and Research) Council meeting, Brussels, 22 and 23 November 2007). Moreover, in 2008 the 7th Framework Program of EU opened a special call entitled "Rehabilitation of data from biodiversity-related projects funded under previous framework programmes".

Data publication has become a common practice in other branches of science (i.e. Altman and King 2007). In a recently published white paper of OECD (Green 2009), there are described working examples of a quite simple approach to data publishing. Publishing of data is not primarily a technical problem. The main question is how to publish data under the open access model and how to motivate data collectors and creators? There are also other unresolved questions related to metadata, consequent data use and reuse, conventions of citation, author's and/or institution's recognition, challenges of publishing dynamic datasets (databases) and so on.

The benefits from data publishing for authors and for society seem obvious. Since the data is online and freely available, authors gain broader recognition for their work. But more importantly, published data when well-described and validated, contribute to data discovery, access, and publishing infrastructure such as the Global Biodiversity Information Facility (GBIF) and larger data aggregations through organizations like ZooBank, Morphbank, Encyclopedia of Life (EOL). Increasingly, analyses and meta-analyses of aggregated data will unlock new potentials for academic science, applied science (i.e., conservation), and informed public discourse leading to better public policy. The indexing practices described here would establish authorship of datasets, crediting researchers and institutions for their productivity and initiative in piloting non-traditional forms of publication. Sponsors, funding agencies and society in general would benefit from the full disclosure, aggregation and reuse of the results of publicly funded scientific research.

In systematic zoology we have already excellent examples of semantic enhancements to research papers (i.e., Pyle et al. 2008; Fisher and Smith 2008; Talamas et al. 2009), however we still lack a comprehensive "full-life-cycle" model for data management, from a manuscript through the peer-reviewed publication process to data aggregation, dissemination, and stable long-term deposition. This particularly concerns primary biodiversity (mostly based on but not limited to specimen and observation-by-locality) data.

We describe here a vision for data publication and dissemination in taxonomy. The most ambitious illustration of this publishing process (Fig. 1) is the milestone paper of Miller et al. (2009) published in the present issue. Following its declared policy to develop and apply innovative ways of publishing towards a "web-based" taxonomy (Penev et al. 2008), ZooKeys, in close cooperation with the authors and GBIF, provide in this paper for the first time in systematic zoology a unique combination of data publication and semantic enhancements (Shotton et al. 2009), applied within the mainstream process of journal publishing:

1. All primary biodiversity data underlying a taxonomic monograph are published as a dataset under a separate DOI within the paper
2. The occurrence dataset is indexed to GBIF simultaneously with the publication
3. The occurrence dataset is published as a KML (Keyhole Markup Language) file under a distinct DOI to provide an interactive experience in Google Earth. The interactive map features collection occurrence data for all specimens in the monograph and links to collections of images for each species posted on Morphbank. Data can be filtered to display or hide any family, genus, or species.
4. All new taxa (42) are registered at ZooBank during the publication process (mandatory for ZooKeys)
5. All new taxa (42) are provided to Encyclopedia of Life through XML mark up on the day of publication (mandatory for ZooKeys)

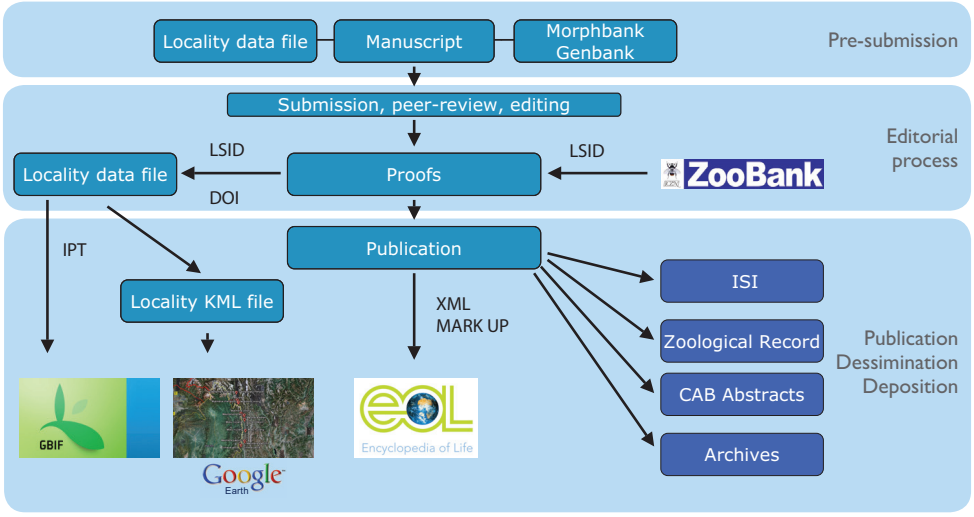

Fig. 1. The ZooKeys model for data publication and semantic enhancements workflow in taxonomy

Description of the concept

Static and dynamic datasets

A dataset is understood here as a logical file – analog or machine-readable – presenting a collection of facts (observations, descriptions or measurements) formally structured in records; each record is structured in fields. Within the domain of biological taxonomy, a dataset can be any discrete collection of data underlying a taxonomic paper – e.g., a list of all occurrence data published in the paper, data tables from which a graph or map is produced, an appendix with morphological measurements, etc. Each dataset has its own DOI within a paper.

We propose a clear distinction between fixed "*data tables*" that represent precisely the set or sets of data upon which the analyses and conclusions of a given scientific paper are based and dynamic "*databases*" that represent larger and more extensive collections of data that may or may not include the precise data table or tables that are the referent(s) for a given scientific paper. Both data tables and databases are of strong potential scientific interest and application but *publication of data tables is inextricably linked to a scientific paper and the publisher must assure consistent and secure access, in perpetuity, to referent data tables.* In other words, a newer version of a data table cannot be uploaded on the journal's website and in this way the publisher guarantees its consistency and perpetuity in time, similarly to the content of an electronic publication.

We propose the use of digital object identifiers (DOI), semantically related to the DOI of the respective paper, to be assigned to each discrete referent data table. This will insure that there is a fixed and citable entity to which subsequent reviews and revisions can refer. The referent data table can be downloaded in conjunction with a paper and critically analyzed. It can also be freely used for subsequent analyses and publications given proper citation and attribution.

Publication and citation of dynamic *databases* is supplementary and discretionary in the context of a given paper but should certainly be encouraged for the greater good of science. Sustained maintenance of databases is not an obligation for a publisher but is – by emerging scientific norms – an obligation for scientists, scientific institutions and agencies, libraries and archives. Ultimately we can envision aggregations – at various chronologic, geographic or taxonomic scales – of hundreds of databases composed of thousands of datasets.

Respecting the citation of databases, we note that conventions of citation must include careful specification of the version, date and time of accessioning. Database design should at all points support easy "stamping" of data usage.

Identification and location

We propose also to distinguish between static and dynamic datasets semantically to facilitate harvesting through scripts and other machine-generated methods. Data tables,

as the referents of a scientific paper, can be identified with the acronym "dt" (from *data table*) within its DOI, which is a semantical extension of the paper's DOI (working examples from the paper of Miller et al. (2009) in the present issue):

> If a DOI of a paper is: doi: 10.3897/zookeys.11.160
>
> then the DOI of a static dataset (*data table*) will have the form:
>
> doi: 10.3897/zookeys.11.160-app.B.dt
>
> which means Appendix B of the paper.

To distinguish between static and dynamic datasets, DOI of data, if published as a *database,* would have the form:

> doi: 10.3897/zookeys.11.160-app.B.db ("db" marks up a dynamic dataset to differ it from "dt" which means a static dataset or *data table*).

DOI of another *data table,* i.e. data underlying a graph, within the same publication would have the form:

> doi: 10.3897/zookeys.11.160-fig.2.dt

Stand-alone publication of a dataset

"Datasets", as collections of data, can be published separately in a form of a conventional publication containing textual description of key features of the dataset (e.g. "metadata") – i.e., introductory information on a taxon/taxa being subject of this dataset, principles of architecture, size, technical description, history of the dataset/database itself, hosting, relations to other datasets, ownership and copyright issues and so on. Within a journal, such a publication may be classified as "Dataset" analogously to other type of publications, i.e., "Editorial", "Research article", or "Correspondence". Formal protocols for what will constitute complete and acceptable metadata for data (data tables, datasets and databases) will need to be prescribed in subsequent analysis. Significant progress on this problem is being made in other domains (Green 2009).

Peer-review

The publication – either as a scientific paper containing referent data tables – or a stand-alone publication of a dataset – will be subject to standard peer-review processes in accordance with the editorial requirements of specific journals.

Citation

Citation of a data will vary in correspondence with the form of publication and preservation.

If published within a paper:

> <Author> <(Year)> <Legend of the dataset>. DATA TABLE. <File format> <DOI of the dataset> <Journal, volume, issue, pages> < DOI of the publication>

In the case of the working example of Miller et al. (2009), the citation of the dataset published under Appendix B has the form:

> Miller JA, Griswold CE, Yin CM (2009) Appendix B. Locality data (XLS format) for all specimens of the spider families Theridiosomatidae, Mysmenidae, Anapidae, and Symphytognathidae collected during an inventory of the Gaoligongshan, Yunnan, China, 1998-2007. DATA TABLE. File format: Microsoft Excel (1997-2003). doi: 10.3897/zookeys.11.160-app.B.dt. ZooKeys 11: 9-195. doi: 10.3897/zookeys.11.160

If published as a stand-alone:

> <Author> <(Year)> <Title of the publication> <Journal, volume, issue, pages> < DOI of the publication> DATASET. <Dataset format, version> <DOI of the dataset>. Accessed <Day of accession>

Dissemination and usage

ZooKeys was the first journal to offer mandatory ZooBank registration and to supply species descriptions of all new taxa to Encyclopedia of Life on the day of publication. Thanks to that new taxa described on the pages of the journal become quickly known to the world. What happens, however, with the species-by-occurrence records (i.e., what is normally called "primary biodiversity data")?

Description of a new species is not easy and it requires investnments of time, energy, knowledge, and resources. Similarly, the collection of specimen records requires serious effort. At ZooKeys we are convinced that each species-by-occurrence record collected on Earth has its own discrete value and deserves proper registration, publication, preservation and dissemination with proper attribution for authors, data suppliers and publishers.

Still the job is not complete merely with publication of data. The larger challenge is to accumulate a dataset in a repository, or repositories, where it will be preserved, integrated and reused. The role of such discovery and mobilisation of primary biodiversity is major reason for GBIF's existence. The recently launched Integrated Publish-

ing Toolkit (ipt.gbif.org) by GBIF allows data to be published in the form of a dataset using the DarwinCore protocol. Besides the standard DarwinCore fields describing taxonomic position and rank of each taxon, locality, collectors, ID of specimens in a collection, etc., the data table published as a downloadable Excel file as Appendix B (doi: 10.3897/zookeys.11.160-app.B.dt) in the paper of Miller et al. (2009) was completed with the following additional fields:

§ LSID (ZooBank) of each taxon
§ DOI of the dataset
§ DOI of the publication
§ Citation

The dataset was published through GBIF simultaneously with the publication of the paper, allowing its reuse within the scope of an unlimited number of cross-cutting, larger datasets compiled at larger geographic, taxonomic or chronologic scales, e.g. of all spider species recorded in China, all plant and animal species recorded in the same localities or region and so on. The LSID link of each species offer unlimited possibilities for use of each specimen record in any branch of science and nature conservation.

However, how to use the data independently, in the form "as-it-is-published"? Under the terms of the Creative Commons Attribution License anyone can download the dataset and use it, provided that the original author and source are credited. Alternatively, the Creative Commons "CC 0" license may be used – it places work in the public domain – this license relies on conventional norms of citation to insure proper attribution. The same paper however, offers one more innovative way to display and use this particular dataset. In Appendix C, the data are published in the form of KML, a file format which allows interactive mapping in Google Earth, and incorporates display on maps of all localities and species, descriptions of each specimen records, hierarchical filtering and mapping of higher taxa (genus, family), and links to species descriptive morphology in Morphbank. It remains only to include the links to EOL species pages, which will be easy to do, when EOL starts to provide LSIDs to their species pages *prior to the day of publication*. Another great feature of our time to dream for!

We believe that our proposed approach will motivate authors to publish data and to receive recognition in the form of citations, future co-authorship, and credentialing for career development. We look forward to the time when data discovery and publishing initiatives like GBIF will automatically index such datasets and incorporate them along with the correspondent descriptive metadata details.

Acknowledgments

The authors thank Pensoft's team (Teodor Georgiev, Ivailo Stoyanov and Veselin Kostadinov) and Andrea Hahn (GBIF) for preparation and publication of this special issue in unprecedented time limits.

References

Altman M, King GA (2007) Proposed Standard for the Scholarly Citation of Quantitative Data. D-lib Magazine 13 (3/4).

Green T (2009) We Need Publishing Standards for Datasets and Data Tables. *OECD Publishing White Paper*, OECD Publishing. doi: 10.1787/603233448430.

Fisher BL, Smith MA (2008) A Revision of Malagasy Species of *Anochetus* Mayr and *Odontomachus* Latreille (Hymenoptera: Formicidae). PLoS ONE 3(5): e1787. doi: 10.1371/journal.pone.0001787

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

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, Lourenco 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

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.

Shotton D, Portwin K, Klyne G, Miles A (2009) Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. PLoS Comput Biol 5(4): e1000361. doi:10.1371/journal.pcbi.1000361

Talamas EJ, Johnson NF, van Noort S, Masner L, Polaszek A (2009) Revision of world species of the genus *Oreiscelio* Kieffer (Hymenoptera, Platygastroidea, Platygastridae). ZooKeys 6: 1-68. doi: 10.3897/zookeys.6.67

2832nd COMPETITIVENESS (Internal market, Industry and Research) Council meeting, Brussels, 22 and 23 November 2007. http://ue.eu.int/Newsroom