No specimen left behind: mass digitization of natural history collections

Edited by

Vladimir Blagoderov & Vincent S. Smith

PENSOFT

Sofia–Moscow

2012

ZooKeys 209 (Special Issue)

No specimen left behind: mass digitization of natural history collections

Edited by Vladimir Blagoderov & Vincent S. Smith

First published 2012
ISBN 978-954-642-645-1 (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, July 2012

Contents

ZooKeys 209: 1–6 (2012)
doi: 10.3897/zookeys.209.3699
www.zookeys.org

EDITORIAL

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

Bringing collections out of the dark

Vincent S. Smith[1], Vladimir Blagoderov[2]

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

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

Received 14 July 2012 | Accepted 16 July 2012 | Published 20 July 2012

Citation: Blagoderov V, Smith VS (2012) Bringing collections out of the dark. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 1–6. doi: 10.3897/zookeys.209.3699

Natural history collections are an incomparable treasure and source of knowledge. Collected over centuries of field exploration, these repositories contain a sample of the world's biodiversity, and represent a monumental societal investment in research and applied environmental science (Network Integrated Biocollections Alliance 2010). Knowledge derived from the 1.5–3 billion specimens (Ariño 2010, Duckworth et al. 1993) within these collections has made vital contributions to the study of taxonomy, systematics, invasive species, biological conservation, land management, pollination and biotic responses to climate change (Chapman 2005). Despite these activities, natural history collections are significantly underutilised due to the difficulty of obtaining and analysing data within and across collections. Digitisation and mobilisation of specimen and associated data removes this impediment, but presents major technical and organisational challenges. The largest of these is how to capture specimen data fast enough to achieve digitisation of entire collections while maintaining sufficient data quality.

Until recently, episodic and incremental funding has had limited success with natural history digitisation, largely addressing local projects within single institutions or across niche research communities. New funding, coupled with more collaborative approaches to digitisation, and technical advances with scanning and imaging systems have begun to change this. The collection of eighteen articles published here examines some of these developments, providing a snapshot of current digitisation efforts and progress across these themes.

The first of these papers by Reed Beaman and Nico Cellinese (2012) looks at the transformative potential of natural history specimen digitisation, both in terms of driving new developments in technical infrastructure, as well as in new applications for the digitised products of this work. Fundamental to the increase in efficiency of these programmes is the modularisation of the digitisation process. Collections digitisation is broadly defined to include transcription into electronic format of various types of data associated with specimens, the capture of digital images of specimens, and the georeferencing of specimen collecting localities. These steps are examined by Gill Nelson and colleagues (2012), who are quite literally based at the 'hub' of National Science Foundation efforts to advance the digitisation of North American biological collections in the United States. Based on studies of major digitisation efforts across the U.S., Nelson et al. break down the clusters of digitisation activities into workflows that can be adopted by other digitisation efforts.

A fundamental step in any digitisation programme is the aggregation or federation of digital output so it can be collectively searched and discovered. The European Union funded *Open-UP* project is one such effort within Europe, and is described by Anton Güntsch and Walter Berendsohn (2012) in their paper on the mobilisation of natural history multimedia resources through the *EUROPEANA* data portal. The challenges surrounding the coordination of digitisation efforts are also looked at through a series of projects trying to address these problems, nationally or via thematic networks. In some cases these are best practice networks such as the *U.S. Virtual Herbarium* described by Mary Barkworth and Zack Murrell (2012). In other cases these projects provide a service infrastructure such as the Finnish *Digitarium* (Tegelberg et al. 2012). Even operating within the confines of a single large institution can be a challenge: different stakeholders have different priorities that can be difficult to accommodate within the budgets of single institutions. Marc Gofferjé and Jon Peter van den Oever (2012) describe a range of solutions to address these issues at NCB Naturalis. Part of the solution lies in improving the efficiency of an institutions digitisation process, as illustrated at the New York Botanic Gardens (Tulig et al. 2012) and the Royal Botanic Gardens Edinburgh (Haston et al. 2012).

Attempts to automate digitisation are confounded by the fact that different types of organisms require very different types of preservation. Plants and fungi are typically prepared as dried, flattened specimens attached to archival quality paper, with printed label data mounted on the sheet. This pre-adapts herbaria to rapid digitisation. In contrast insects, which are the most numerous organisms in collections, are typically mounted by pinning individuals on entomological pins, which are accompanied by tiny (often folded) labels beneath each specimen. The particular demands of mass digitising entomological specimens are the subject of five papers, which have methodologically converged on the scanning whole collection drawers. *GigaPan*, described by Matthew Bertone and colleagues (2012) was arguably the first of these approaches,

enabling the low cost capture of gigapixel panoramas of insect museum drawers containing many hundreds of specimens. More recently *SatScan*, developed in association with the Natural History Museum London (Blagoderov et al. 2012), and in use at the Australian National Insect Collection (Mantle et al. 2012) has enabled these panoramas to be obtained with minimal distortion. *SatScan* is accompanied by software used to select and annotate images of individual specimens. The drawer scanning approach has been incorporated as part of the U.S. *InvertNet* digitisation programme (Dietrich et al. 2012), and has resulted in a new, low cost instrument called *DScan* (Schmidt et al. 2012). A contrasting approach to accessing digital images is described by Quentin Wheeler and colleagues (2012), who are exploring the use of telemicroscopy to enable remote researchers to access and manipulate specimens beyond their physical reach. Although not strictly mass digitisation, the potential effect of this network of remote access microscopes is similar, enabling researchers to examine insect material located at major institutions over a network connection.

Even with this automation, a significant labour force is still critical for many digitisation projects. Paul Flemons and Penny Berents (2012) explore the use of volunteers to increase the rate of digitising insect collections. This has enabled the Australian Museum to capture label data and images for 16,000 specimens in just 5 months. Label data transcription is a major problem in many digitisation projects. Andrew Hill and colleagues (2012) describe their software to crowdsource label transcription through a workforce of citizen scientists. Embedding quality control techniques and design elements to keep contributors motivated, *Notes On Nature* provides a toolkit for transcription of ledgers and labels of natural history specimens. Andrea Thomer and colleagues (2012), extend this transcription work into new territory using Wiki-style templates to crowdsource data extraction from century-old field notebooks. This enables interoperability of the underlying data without losing the narrative context from which these observations are drawn. The series closes with a paper by Randall Schuh (2012), who looks at methods to integrate specimen databases into the practice of revisionary systematics, closing the loop between digitising, extracting and reusing data in taxonomic research.

In bringing together this special issue on digitisation we have sought to represent a wide selection of projects and techniques. These papers provide a snapshot of activity in what is a fast moving field that is seeing ever-increasing degrees of collaboration across disciplines and between collection-based institutions. Many of these projects deal with the unique challenges associated with major collections that have built up over several centuries, with different communities of practice and different user groups. Despite these differences, the standards for collection acquisition, preservation and documentation are broadly consistent, meaning that there is sufficient common ground to bring together the enormous amounts of data that are being exposed through these activities. We expect that in the next decade these data will become the new frontier for natural history collection management and research.

Acknowledgements

We sincerely thank the authors and reviewers of these articles who have responded, often at very short notice, to our requests for assistance. This work was supported by the Natural History Museum, London and the EU funded FP7 ViBRANT project (contract number RI-261532).

References

Ariño A (2010) Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7(2): 81–92.

Barkworth ME, Murrell ZE (2012) The US Virtual Herbarium: working with individual herbaria to build a national resource. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 55–73. doi: 10.3897/zookeys.209.3205

Beaman RS, Cellinese N (2012) Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 7–17. doi: 10.3897/zookeys.209.3313

Bertone MA, Blinn RL, Stanfield TM, Dew KJ, Seltmann KC, Deans AR (2012) Results and insights from the NCSU Insect Museum GigaPan project. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 115–132. doi: 10.3897/zookeys.209.3083

Blagoderov V, Kitching IJ, Livermore L, Simonsen TJ, Smith VS (2012) No specimen left behind: industrial scale digitization of natural history collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 133–146. doi: 10.3897/zookeys.209.3178

Chapman AD (2005) Uses of Primary Species-Occurrence Data, Version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. http://www.gbif.org/orc/?doc_id=1300

Dietrich CH, Hart J, Raila D, Ravaioli U, Sobh N, Sobh O, Taylor C (2012) InvertNet: a new paradigm for digital access to invertebrate collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 165–181. doi: 10.3897/zookeys.209.3571

Drew J (2011) The role of natural history institutions and bioinformatics in conservation biology. Conservation Biology 25(6): 1250–1252.

Duckworth WD, Genoways HH, Ros CL (1993) Preserving natural science collections: chronicle of our environmental heritage. Washington, D.C. iii+140 pp.

Flemons P, Berents P (2012) Image based Digitisation of Entomology Collections: Leveraging volunteers to increase digitization capacity. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 203–217. doi: 10.3897/zookeys.209.3146

van den Oever JP, Gofferjé M (2012) 'From Pilot to production': Large Scale Digitisation project at Naturalis Biodiversity Center. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 87–92. doi: 10.3897/zookeys.209.3609

Berendsohn WG, Güntsch A (2012) OpenUp! Creating a cross-domain pipeline for natural history data. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 47–54. doi: 10.3897/zookeys.209.3179

Haston E, Cubey R, Pullan M, Atkins H, Harris DJ (2012) Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 93–102. doi: 10.3897/zookeys.209.3121

Hill A, Guralnick R, Smith A, Sallans A, Gillespie R, Denslow M, Gross J, Murrell Z, Conyers T, Oboyski P, Ball J, Thomer A, Prys-Jones R, de la Torre J, Kociolek P, Fortson L (2012) The notes from nature tool for unlocking biodiversity records from museum records through citizen science. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 219–233. doi: 10.3897/zookeys.209.3472

Mantle BL, La Salle J, Fisher N (2012) Whole-drawer imaging for digital management and curation of a large entomological collection. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 147–163. doi: 10.3897/zookeys.209.3169

Nelson G, Paul D, Riccardi G, Mast AR (2012) Five task clusters that enable efficient and effective digitization of biological collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 19–45. doi: 10.3897/zookeys.209.3135

Network Integrated Biocollections Alliance (2010) A Strategic Plan for Establishing a Network Integrated Collections Alliance http://digbiocol.files.wordpress.com/2010/08/niba_brochure.pdf

Pyke GH, Ehrlich PR (2010) Biological collections and ecological/environmental research: a review, some observations and a look to the future. Biological Reviews 85: 247–266.

Schmidt S, Balke M, Lafogler S (2012) DScan – a high-performance digital scanning system for entomological collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 183–191. doi: 10.3897/zookeys.209.3115

Schuh RT (2012) Integrating specimen databases and revisionary systematics. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 255–267. doi: 10.3897/zookeys.209.3288

Tegelberg R, Haapala J, Mononen T, Pajari M, Saarenmaa H (2012) The development of a digitising service centre for natural history collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 75–86. doi: 10.3897/zookeys.209.3119

Thomer A, Vaidya G, Guralnick R, Bloom D, Russell L (2012) From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old

field notebooks. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 235–253. doi: 10.3897/zookeys.209.3247

Tulig M, Tarnowsky N, Bevans M, Kirchgessner A, Thiers BM (2012) Increasing the efficiency of digitization workflows for herbarium specimens. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 103–113. doi: 10.3897/zookeys.209.3125

Wheeler Q, Bourgoin T, Coddington J, Gostony T, Hamilton A, Larimer R, Polaszek A, Schauff M, Solis MA (2012) Nomenclatural benchmarking: the roles of digital typification and telemicroscopy. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 193–202. doi: 10.3897/zookeys.209.3486

ZooKeys 209: 7–17 (2012)
doi: 10.3897/zookeys.209.3313
www.zookeys.org

REVIEW ARTICLE

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

Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science

Reed S. Beaman[1], Nico Cellinese[1]

1 *Florida Museum of Natural History, University of Florida, Dickinson Hall, Museum Rd, Gainesville, Florida 32611-7800, U.S.A.*

Corresponding author: *Nico Cellinese* (ncellinese@flmnh.ufl.edu)

Academic editor: *V. Blagoderov* | Received 1 May 2012 | Accepted 9 July 2012 | Published 20 July 2012

Citation: Beaman RS, Cellinese N (2012) Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 7–17. doi: 10.3897/zookeys.209.3313

Abstract

New information technologies have enabled the scientific collections community and its stakeholders to adapt, adopt, and leverage novel approaches for a nearly 300 years old scientific discipline. Now, few can credibly question the transformational impact of technology on efforts to digitize scientific collections, as IT now reaches into almost every nook and cranny of society. Five to ten years ago this was not the case. Digitization is an activity that museums and academic institutions increasingly recognize, though many still do not embrace, as a means to boost the impact of collections to research and society through improved access. The acquisition and use of scientific collections is a global endeavor, and digitization enhances their value by improved access to core biodiversity information, increases use, relevance and potential downstream value, for example, in the management of natural resources, policy development, food security, and planetary and human health. This paper examines new opportunities to design and implement infrastructure that will support not just mass digitization efforts, but also a broad range of research on biological diversity and physical sciences in order to make scientific collections increasingly relevant to societal needs and interest.

Keywords

Scientific collections, biodiversity, digitization, specimen access, biodiversity informatics, data sharing, linked data, interoperability

Introduction

Understanding biodiversity is one of five grand challenges identified by US National Research Council Committee on Forefronts of Science at the Interface of Physical and Life Sciences (2010). Broadly defined, the study of biodiversity addresses variation among living things and systems, ranging in scale from molecules, genes, cells, individual organisms, to species through ecosystems. Specimens, and now the digital proxies for specimens, are a critical underpinning in documenting biodiversity (Berendsohn and Seltmann 2010, Berents et al. 2010, Scoble 2010, Vollmar et al. 2010). Improving infrastructure for digital specimen data comes at a time when basic biodiversity science is itself undergoing rapid change.

Investments in digitization will ultimately yield a better return if use expands and specimen data are linked across a wide array of related biotic and abiotic data. The specimen objects provide a physical basis for linking data to other biodiversity science domains. Scientific collections document the who, what, where, and when of biological diversity. Digitization, beyond making collections more accessible to researchers, provides access to downstream users such as the general public, government and non-government agencies and private enterprises.

Many researchers still fail to realize the importance of vouchering specimens to their community's practice. Whether they study molecules or ecosystems, many are content to document the organisms they work with by taxonomic name alone. Even researchers in the closely aligned field of molecular systematics have previously failed to grasp the importance of citing specimen vouchers, evidenced, for example, in the lack of voucher data cited in GenBank, other repositories, and in publications. How can we know that the sequence deposited in GenBank belongs to the taxon under which it is filed? Whether alpha taxonomy or a synthesis of large phylogenetic trees based on molecular sequences, citing vouchers remains essential to a scientific process that is repeatable and verifiable.

In order for research communities to stay abreast and benefit from opportunities of new information technology environments (e.g., cloud computing, linked data and ontologies, social and computational virtual networks), increasing multi-disciplinary collaboration between biologists and computer and information scientists and engineers is a must, as few scientists in representative domains have all the necessary skills to "do it all." Across the biological sciences, where new tools such as next generation sequencing and environmental sensors challenge network design and contribute to the now well-known data deluge (Kahn 2011, McNally et al. 2012, Michener and Jones 2012, Kolker et al. 2012), robust cyberinfrastructure that facilitates collaboration, data automation, sustainable software development, and high performance computing is a priority (Donoghue et al. 2009, Hendry 2010). Digitization of scientific collections is no exception, as two- and three-dimensional images, video, audio, and other media derived from physical specimens and observations and measurements proliferate, they add significantly to the data deluge, and to the need for long-term data storage archives and data curation. It is also essential to recognize that digitized collections perma-

nently document resources that are held in museums and herbaria, and so have a place in foundational biodiversity infrastructure.

Some of the necessary organizations are already in place, e.g., Global Biodiversity Information facility (GBIF: http://www.gbif.org), Atlas of Living Australia (ALA: http://www.ala.org.au), Virtual Biodiversity Research and Access Network for Taxonomy (ViBRANT: http://vbrant.eu), DataONE (http://dataone.org), and the US Integrated Digitized Biocollections (iDigBio: https://www.idigbio.org), which are at various stages of implementation and operation. Each, however, has limitations on scope, and the resulting infrastructure remains an innovative yet incomplete patchwork of distributed data, archival resources, tools and software. For example, GBIF has no mandate as a primary resource provider, and instead serves as an aggregator, indexer, and distributed portal; iDigBio is not funded to develop new digitization tools, and like ALA has a national mandate.

The gaps in scope present both a need and opportunity to further conceptualize and develop an international infrastructure and missing components that will fully support the broad definition of biodiversity research that coordinates and integrates with existing infrastructure, including tools developed by individuals and small teams. Coordinating biodiversity research and cyberinfrastructure requires nimble computational resources, an ability to support heterogeneous distributed data, robust and sustainable software development, and an innovative and well-trained workforce, along with the social and research infrastructure that supports them, to answer challenges that have previously been beyond the scope of traditional scientific methods and organizations.

This paper is a call to the community to define a comprehensive conceptual plan that will allow scientists across multiple disciplines to coordinate a community able to capitalize on cutting edge computational infrastructure, economies of scale, with the innovation and needs of a broad community of other scientific organizations. So far, the biocollections community has operated in an ad hoc, geographically fragmented way. As research has become increasingly collaborative, interdisciplinary, and international, new social challenges arise around how scientists work together, across disciplines, institutions, and geographic and political boundaries. Community based planning allows consideration of critical elements of sustainable infrastructure, including:

- Setting priorities and identifying use cases.
- Identifying stakeholders, collaborators, and communities of practice.
- Specifying computational infrastructure, software, and data storage requirements and dependencies.
- Practices, methods, standards, and interoperability.
- Management, organizational structure, and sustainability.
- Risk assessment.

Formal conceptual planning and development of standards is common in engineering, industrial, and biomedical sectors, but in basic biological research, a per-

ception remains that innovation and individual research are not as dependent on foundational infrastructure as in the physical sciences. As networks of biodiversity researchers grow, they have an increased need to plan effective infrastructure to support collaboration, distributed data management and access. As an example, extensive planning and design processes are documented in a NASA (2007) handbook on systems engineering, including lifecycle documentation, establishing user requirements, and management. The elements listed above and discussed below are not exhaustive, and are described in a context of how digitized collections can underwrite a larger community in the biodiversity sciences.

Priorities and use cases

A challenge of scale for this community is in the numbers. Over a billion specimens exist in thousands of collections, and most are managed independently within standalone museums, universities, and government agencies (http://nscalliance.org/wordpress/wp-content/uploads/2009/11/iwgsc-report.pdf). Digitizing an institution's collection from A-Z may be the most efficient means, but feasible only in certain circumstances, such as large-scale moves or renovations (e.g., the recent renovation of the Paris Herbarium). Funds, personnel, and time are typically limiting, so priorities must be set. Type collections, historical collections, special collections are common priorities, but identifying and increasing relevance of collections to the research community and other stakeholders is another strategy.

The aggregation of digital data through portal infrastructure such as the Global Biodiversity Information Facility (GBIF: http://www.gbif.org), VertNet (http://vertnet.org), Morphbank (http://www.morphbank.net), the Paleontology Portal (Paleoportal: http://www.paleoportal.org), among others, added to the realization that specimens are useful for much more than simple mapping of species occurrences. Digital specimen data is a proxy or surrogate of physical objects and appropriate use may be limited. However, digitized data can be used to study morphology (Corney et al. 2012), identify, classify, map and spatially model taxa (Thuiller et al. 2009, Soberón 2010). Where expertise is a limiting resource, for example in the study of hyper diverse groups (e.g., insects), cyberinfrastructure can help leverage that expertise (Moore 2011).

There is further need to establish specific use cases (or more precisely, user scenarios) whether biological, technical, or a combination of both. As applied to collections digitization or other areas of biological informatics (e.g., genomics and proteomics), research is increasingly catalyzed by improved computational infrastructure to process and store large data sets and files, index and link billions of data records, data-mine existing resources, and incorporate ontologies to support semantic reasoning. Engineering breakthroughs in optical sensors and robotics have had and will continue to have enormous potential to guide and impact digitization efforts, but the needs of the biology domain can also drive technology.

Stakeholders, collaborators, and communities of practice

Stakeholders, both primary users (e.g., curators, collection managers) and downstream users (e.g., climate researchers, resource managers, educators), are the most appropriate source of user scenarios. It is the stakeholders that build communities of practice from the ground up and define what is really needed, what is novel, and add value to current practice. Users define the need to scale infrastructure capabilities to support the science (e.g., geospatial and phylogenetic analyses). Users also compose the social networks, crowd-sourcing workforce, and ultimately provide intellectual capacity for digital markup and annotations, development of linked data applications, ontologies, automation, and workflows.

In 2010, the scientific collections community within the United States outlined a strategic plan for digitizing scientific collections, including the establishment of the Network Integrated Biocollections Alliance (NIBA, http://digbiocol.wordpress.com). The plan defined digitization to encompass a broad range of digital data capture about biological specimens, from field collection events to cataloging and accessioning metadata, images and other media derived from field and laboratory work, and set the stage for establishing priorities based upon how a specimen and its occurrence relate to research. Additionally, the physical specimens can be re-sampled, e.g., for epiphytes, parasites, mineral deposits, bio-medically active compounds, re-purposing not just data, but the specimen objects themselves, for research on many functional elements of biodiversity, including mutualism, co-evolution, lateral gene transfer, parasitology, and community ecology.

The U.S. National Science Foundation responded to elements of the NIBA plan by establishing a program for Advancing Digitization of Biological Collections (NSF-ADBC, http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503559), which funds digitization based on scientific questions or themes through extensive collaborative networks. Examples of Thematic Collection Networks (TCN) funded through this program are detailed on the iDigBio web site (https://www.idigbio.org/content/thematic-collections-networks).

Key challenges are often social and priorities may be at odds with technical needs. Solving social challenges requires different approaches and expertise not be inherently a part of existing biocollections business practices. Long adhered to curation practices may need to be revised, and interdisciplinary collaboration with social scientists and psychologists may provide useful insight, but may not necessarily be well received. For example, is it legitimate to unpin an insect to access the label data during the digitization process? As investments in digitization increase, so will the need to produce metrics of success and document outcomes. As communities of practice develop around digitization networks, social and usability considerations are essential.

Computational infrastructure

Computing, software, and data resources are clear enablers of both large-scale digitization and biodiversity research. Advance computational infrastructure, including vir-

tual and cloud infrastructure, are costly to design and deploy, so are generally viewed as resources to be adopted across all sciences. In the U.S., the nationally funded TeraGrid, and its successor XSEDE, have primarily focused on processing capability, or cycles, and benefits applications such as phylogenetic inference, image manipulation, analysis and visualization, but less so for the storage requirements of digital collections, including long-term archiving of images and other media.

Dependencies often relate to previous investments and software development in the form of libraries, services, and value added data sets. Georeferencing tools, e.g., GeoLocate (http://www.museum.tulane.edu/geolocate), are good examples of existing investment that incorporates automation, data-mining algorithms, need for gazetteer and other geospatial data, and mapping tools. Automated data capture methods, for example the use of Optical Character Recognition (OCR) may leverage commercial software and allow deployment of services or software with embedded OCR.

Practices, methods, and workflows

Digitization workflows span across human mediated processes through data and computationally intensive automation where software tools and services are the actors and intersect field collection techniques, institutional accession policy, differences in curatorial practice among domains, and involvement of the general public in crowd-sourced methods.

The workflows that represent digitization of new accessions have in many cases required, or at least highly recommended, elements of funded projects in systematics and ecology. The Moorea Biocode project (http://moorea.berkeley.edu/biocode) is an exemplar, comprehensive effort to collect data on all aspects of a biodiversity survey, including vouchers, tissues, photos and other media. Expanding on efforts such as this has potential to test capacity for digitization and physical curation. BioBlitzes are similar approaches that typically utilize a combination of expert and citizen scientists over a short period of time (a day or few).

Digitization of existing collections is an enormous undertaking. Initial digitization efforts focused on assembling very complete data records and access to researchers and the public was granted only after extensive quality control. More recently, it has been recognized that not every element of a collection record needs to be recorded in a single digitization event (Granzow-de la Cerda and Beach 2010). For example, recording of an image and "filed-under" taxon name are sufficient to start the process. Digital capture of useful information can follow at a later stage and be treated as annotations (e.g., a history of taxonomic determinations). Some aspects of data capture, like data curation, can be costly when it involves expert judgments. In entomology, for example, the initial capture of a box of specimens that may contain hundreds of individuals represents a further extension of a modular workflow. This works effectively with high-resolution sensors that allow users to scale their view appropriately.

Imaging methods have great growth potential for mass digitization efforts. Those new to digital imaging may find the array of possibilities overwhelming. Sensor resolu-

tion, pixel size, noise sensitivity, and cost are among the factors that must be weighed. Considering fitness for use means that there are no one-size-fits-all solutions; collections inherently vary in the ways that physical objects and their associated data are stored, and differ in size (from a few thousand to millions), use cases, and available budgets.

Another consideration that may ultimate affect use of a digital media objects are the formats in which they are stored, archived, and made available to researchers. Metadata, annotations, color profiles, etc. can be stored within the image, as in the case with EXIF metadata (Romero et al. 2008) or in separate databases. These presence and access to such metadata affect whether viewers can display certain media types, decode metadata, and access or provide new digital annotations. Whether the image formats are proprietary or open source, the type and level of file compression, e.g., lossless vs. lossy, are particularly important in biodiversity research applications, and especially when data are to be archived over the long-term.

Standards and interoperability

Data sharing requires that the resources be communicated in standard formats, consistent usage of vocabulary and concepts, and through protocols understood by each of the nodes of a network. In the biodiversity domain, Darwin Core (DwC, http://rs.tdwg.org/dwc), a TDWG supported standard (Wieczorek et al. 2012), is widely adopted, including by GBIF and it is used by many of GBIF's data providers in the context of the Integrated Publishing Toolkit (IPT), a recently developed tool for easy data sharing (http://code.google.com/p/gbif-providertoolkit/). In its current instance, DwC is for all intents and purposes a controlled vocabulary of terms that describe scientific collections, biodiversity observations, basic taxonomies, and localities, among others. Concepts are defined in human readable language and implementations are independent from any one format (e.g., XML, RDF, or tab-delimited). This creates flexibility to link data from the collections to virtually any other digital record in related domains. Recent harmonization efforts, for example through the Genomic Standards Consortium (http://gensc.org/gc_wiki/index.php/Main_Page), which is developing profiles for minimum information standards (MIxS), make it possible to link genomics data to scientific collections. While very preliminary, such efforts herald recognition that information needs to be exchanged across multiple domains in biology, geo-sciences, and other physical sciences.

Linked data environments are evolving quickly and increasing capacity for data discovery. A collection event may generate a number of specimens that are independently imaged and annotated; tissues may be subsampled from any specimen, its DNA extracted and sequenced. Specimens, annotations, images, tissue samples, DNA may be accessioned into collections at different institutions, and sequences deposited in GenBank. It is a challenge to track the data across different institutions, and especially across digital repositories in different domains. Linked data approaches can provide sufficient provenance to allow discovery of not just how a specimen may have been used, but if a digital annotation occurs (such as a change in identification) this can be

propagated into downstream analyses. Projects like the BiSciCol Biological Science Collections Tracker (http://biscicol.blogspot.com) aim at filling the gap in reconciling specimen data with their derivatives when these are scattered across independent digital repositories to support projects like Moorea Biocode. However, linked data approaches are successful only when data are served to the community and tracking can be achieved with the use of persistent Globally Unique Identifiers (GUIDs). As linked data efforts increase, it is becoming progressively evident that the persistence of GUIDs is both a necessity and a challenge. The responsibility of establishing a persistent GUID lies with the provider (see https://www.idigbio.org/content/idigbio-guid-statement), although other scenarios that may include large data aggregators taking on the responsibility of assigning unique identifiers are also possible. In addition, identifiers need to be associated with individual data objects, and not just data sets.

The development of formal ontologies compliments and extends efforts on controlled vocabularies and linked data. Data modeling associated with ontologies can provide a powerful approach to synthesis in semantic web environments. The biomedical community has invested heavily in initiatives such as the Open Biological and Biomedical Ontologies (OBO Foundry, http://www.obofoundry.org) and Gene Ontology (http://www.geneontology.org). One advantage inherent to biocollections data is that a long history of practice has already led to structural understanding of ontological relationships, and biological classification has served as an example in the general literature on ontologies (Heuer and Hennig 2008). While relationships between collecting events, observations, organism occurrence, and taxonomy may never be solved in a philosophical context, in a pragmatic context, the definition of terms and the use of concepts may be more precisely aligned in shared data environments by consideration of ontological relationships. As the implementation of standards and the underlying terms and concepts is a matter of practice, technology may provide partial solutions, such as in the support of mapping semantic meaning across multiple ontologies and linked data environments.

Risk assessment

While the promise of access and relevance to biological collections data are over-arching goals, digitization can also mitigate, to a very limited extent, the loss of physical collections. However, new field collections can never replace the original, especially when it comes to type specimens and historical collections, even if the localities from which they were collected still exist. Specimen acquisition, curation and preservation of specimens are an enormous long-term capital investment, and the digital capture and dissemination of data is a relatively minor cost in comparison.

Technology develops at such a rapid rate that long-term planning carries uncertainty and risk. For example, as digitization efforts begin to use cloud computing resources for data storage, they may not consider an element of vendor lock-in, i.e., that bandwidth costs may preclude them from migrating their data elsewhere. A related question is whether biodiversity data managers should even manage their own hardware resources,

which often carry hidden costs such as system administration, electric power bills, and other needs that are often not scalable. Hardware lifespan is generally in the 3–5 year range, but carefully planned software and database designs can have much longer shelf life. Optimal methods to develop, maintain, and sustain software applications and data resources are not always clear, and even innovative tools focused on highly specific tasks (e.g., in genomics, proteomics, metabolomics) are unlikely to have a sufficient user base to gain commercial viability. In limited communities of practice, therefore, other business models such as subscription services are more likely to be sustainable in such cases. Collections are generally housed in organizations (museums and academic institutions) that already have a long-term commitment to their physical collections and are managed with public, private or endowed funding. Therefore, extending that commitment to digital information follows logically, but it should not be an unfunded mandate.

The potential for failure lurks around every corner. Many risks are social as much as technical. The individuals in the biodiversity research community may not be able to communicate user scenarios that are adequately understood by technical implementers. Additionally, potential collaborators may have conflicting needs, or may not have a sufficiently innovative vision to create opportunities in a multi-disciplinary environment. There are also significant challenges to broad adoption of digitized collections data, because users outside the immediate circle of formally trained scientists may not be interested in subtleties that drive extensive discussions in the biocollections community, e.g., taxonomic concepts. Downstream users, for example, often want to know only the names of the organisms they are sampling or studying.

Conclusions

In recent years we have witnessed a renewed interest in natural history collections and with that, the leading edge of a deluge of digital biocollections data. Mass digitization approaches, driven by specific research questions, require a variety of methods tailored to the different nature of the specimens in question and requirements of the user scenarios. Rapid advances in technology allow us to implement a variety of tools and workflows that are well adapted to the needs of each collection, including specimen objects, methods of storage, available informatics and human resources. Mass digitization, no matter how achieved, offers the incredible opportunity for using biocollections to address and meet scientific grand challenges at small and large scale, within and across domains. The combination of human pressure on natural systems and new technologies for digitization creates a perfect storm of social imperatives and scientific opportunities to mobilize data and further explore under-described biodiversity still locked within museum cabinets.

The ultimate payoff for broad adoption of biocollection data resides in the synthesis of biodiversity data across domains spanning systematics, evolution, genetics, ecology, and to the physical and social sciences. If we link that knowledge only to a taxonomic name and not to a specimen, we are linking to a subjective judgment about an organism's identity and not to the physical documentation of the organism itself.

By linking experimental data to voucher specimens, experiments become more objective, repeatable, and the data gathered re-usable. Without the evidentiary documentation the investments in experimental research lose their value.

The massive amounts of digital data that we now generate are hard to manage or synthesize with lack of an appropriate infrastructure that helps tracking data provenance, metadata, and all specimen derivatives. This requires a cyberinfrastructure capable of accommodating multi institutional needs and a well-developed knowledge environment in which data can be easily synthesized and semantic reasoning applied. Two important messages arise, one social the other technical. First, in a broad, heterogeneous biodiversity research environment, we need a singular community effort to conceptualize and communicate necessary infrastructure at a larger scale than so far considered perhaps building upon the Global Biodiversity Informatics Conference (GBIC: http://links.gbif.org/supporting_biodiversity_science.pdf) initiative via GBIF. Second, approaches in heterogeneous and distributed data environments that characterize biology require at a minimum persistent GUIDs associated with every specimen and digital data object. Metadata about collective data sets is insufficient. The digitization process is only part of a large data mobilization effort for biodiversity science. It is the very first step forward in order to make data discoverable and facilitate its synthesis.

Acknowledgements

We wish to thank Vladimir Blagoderov for soliciting a special and timely issue on mass digitization of natural history collections and inviting us to contribute. We much appreciate Rod Page and Vincent Smith's constructive comments that helped us improve this manuscript. Finally, we are very grateful to the National Science Foundation (DBI 0956371) for supporting our work that fostered the ideas expressed here.

References

Berendsohn WG, Seltmann P (2010) Using geographical and taxonomic metadata to set priorities in specimen digitization. Biodiversity Informatics 7: 120–129.

Berents P, Hamer M, Chavan V (2010) Towards demand-driven publishing: approaches to the prioritization of digitization of natural history collection data. Biodiversity Informatics 7: 113–119

Granzow-de la Cerda Í, Beach JH (2010) Semi-automated workflows for acquiring specimen data from label images in herbarium collections. Taxon 59: 1830–1842.

Committee on Forefronts of Science at the Interface of Physical and Life Sciences; National Research Council (2010) Research at the Intersection of the Physical and Life Sciences. The National Academies Press. 124 pp.

Corney DPA, Clark JY, Tang HL, Wilkin P (2012) Automatic extraction of leaf characters from herbarium specimens. Taxon 61: 231–244.

Donoghue MJ, Yahara T, Conti E, Cracraft J, Crandall KA, Faith DP, Häuser C, Hendry AP, Joly C, Kogure K (2009) bioGENESIS: providing an evolutionary framework for biodiversity science. DIVERSITAS Report No. 6, 52 pp.

Hendry AP, Lohmann LG, Conti E, Cracraft J, Crandall KA, Faith DP, Häuser C, Joly CA, Kogure K, Larigauderie A, Magallón S, Moritz C, Tillier S, Zardoya R, Prieur-Richard AH, Walther BA, Yahara T, Donoghue MJ (2010) Evolutionary biology in biodiversity science, conservation and policy: A call to action. Evolution 64: 1517–1528. doi: 10.1111/j.1558-5646.2010.00947.x

Heuer P, Hennig B (2008) Classification of Living Beings. In: Munn K, Smith B. Applied Ontology: An Introduction. Ontos Verlag. Heusenstamm, Germany. 197–217.

Kahn SD (2011) On the Future of Genomic Data. Science 331: 728–729. doi: 10.1126/science.1197891

Kolker E, Stewart E, Ozdemir V (2012) Opportunities and Challenges for the Life Sciences Community. OMICS 16: 138–147. doi: 10.1089/omi.2011.0152

McNally R, Mackenzie A, Hui A, Lam DC, Tomomitsu J (2012) Understanding the 'Intensive' in 'Data Intensive Research': Data Flows in Next Generation Sequencing and Environmental Networked Sensors. International Journal of Digital Curation 7: 81–94. doi: 10.2218/ijdc.v7i1.216

Michener WK, Jones MB (2012) Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27: 85–93. doi: 10.1016/j.tree.2011.11.016

Moore W (2011) Biology needs cyberinfrastructure to facilitate specimen-level data acquisition for insects and other hyperdiverse groups. ZooKeys 147: 479–486. doi: 10.3897/zookeys.147.1944

NASA (2007) Systems Engineering Handbook, Revision 1, NASA/SP-2007-6105, NASA. http://education.ksc.nasa.gov/esmdspacegrant/Documents/NASA%20SP-2007-6105%20Rev%201%20Final%2031Dec2007.pdf

Romero NL, Gimenez Chornet VVGC, Serrano Cobos J, Selles Carot ASC, Canet Centellas F, Cabrera Mendez M (2008) Recovery of descriptive information in images from digital libraries by means of EXIF metadata. Library High Tech 26: 302–315

Scoble MJ (2010) Natural history collections digitization: rationale and value. Biodiversity Informatics 7: 77–80.

Soberón JM (2010) Niche and area of distribution modeling: a population ecology perspective. Ecography 33: 159–167. doi: 10.1111/j.1600-0587.2009.06074.x

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD – a platform for ensemble forecasting of species distributions. Ecography 32: 369–373. doi: 10.1111/j.1600-0587.2008.05742.x

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

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, de Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7: e29715

ZooKeys 209: 19–45 (2012)
doi: 10.3897/zookeys.209.3135
www.zookeys.org

RESEARCH ARTICLE

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

Five task clusters that enable efficient and effective digitization of biological collections

Gil Nelson[1], Deborah Paul[1], Gregory Riccardi[1], Austin R. Mast[2]

1 *Institute for Digital Information, Florida State University, Tallahassee, FL 32306-2100, United States* **2** *Department of Biological Science, Florida State University, Tallahassee, FL 32306-4295, United States*

Corresponding author: *Gil Nelson* (gnelson@bio.fsu.edu)

Academic editor: *V. Blagoderov* | Received 27 March 2012 | Accepted 22 June 2012 | Published 20 July 2012

Citation: Nelson G, Paul D, Riccardi G, Mast AR (2012) Five task clusters that enable efficient and effective digitization of biological collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 19–45. doi: 10.3897/zookeys.209.3135

Abstract

This paper describes and illustrates five major clusters of related tasks (herein referred to as *task clusters*) that are common to efficient and effective practices in the digitization of biological specimen data and media. Examples of these clusters come from the observation of diverse digitization processes. The staff of iDigBio (The U.S. National Science Foundation's National Resource for Advancing Digitization of Biological Collections) visited active biological and paleontological collections digitization programs for the purpose of documenting and assessing current digitization practices and tools. These observations identified five task clusters that comprise the digitization process leading up to data publication: (1) pre-digitization curation and staging, (2) specimen image capture, (3) specimen image processing, (4) electronic data capture, and (5) georeferencing locality descriptions. While not all institutions are completing each of these task clusters for each specimen, these clusters describe a composite picture of digitization of biological and paleontological specimens across the programs that were observed. We describe these clusters, three workflow patterns that dominate the implemention of these clusters, and offer a set of workflow recommendations for digitization programs.

Keywords

Biological specimen collections, paleontological specimen collections, biodiversity informatics, workflow, digitization, curation, imaging, task cluster, iDigBio, ADBC

Introduction

This paper presents an analysis and characterization of digitization practices that will help organizations produce and improve effective practices for the digitization of their biological and paleontological collections. The focus is on *digitization workflow*, the sequence of tasks that are performed in order to create digital information that characterizes individual specimens. These tasks typically include photography of specimens and labels, image processing, capture of label information as text, and locality georeferencing. The presentation of workflow characteristics in this paper provides the framework for analyzing the effectiveness and efficiency of workflows and for the development of new effective workflows. It should be noted that the workflows we observed represent a major departure from a historical practice of pulling a single specimen, creating a comprehensive database record, including researching localities, georeferences, collectors, taxon names, nomenclature, and other related details, then moving on to the next specimen (Humphrey and Clausen 1977). This slow data capture process provides an important contrast to the efficient data capture processes examined in this study. It should be further noted that the generalizations we draw here are based on our observations at a select number of institutions and may not encompass the universe of possible digitization workflows. For example, for new specimens, there is a clear trend toward collectors entering data into a database while in the field and this topic is not within the scope of this paper.

We use the term 'digitize' to represent the capture and recording of information about a specimen or collection. Specimens typically include labels, accession books, and field notes that have typed or handwritten information about the collection event (e.g. collector's name, date, locality) and the specimen itself (e.g. scientific name and identifying number). Digitization of label information includes capturing the text as characters, dividing the text into specific properties, and storing this information in a database. Digitization may also include capturing digital images and other media. References to media objects are added to the database records.

The collections community has recognized that digitization processes need to be made more efficient to meet pressing scientific and societal needs (a topic broadly reviewed by Chapman 2005a), a notion supported by such initiatives as GBIF (http://www.gbif.org), iDigBio (http://www.idigbio.org) and the Thematic Collections Networks funded by the National Science Foundation's Advancing Digitization of Biological Collections (ADBC) program (http://www.nsf.gov/pubs/2011/nsf11567/nsf11567.htm), Atlas of Living Australia (http://www.ala.org.au/), ViBRANT (http://vbrant.eu/), and VertNet (http://www.vertnet.org). However, little has been published that characterizes modern existing and effective digitization workflows for a broad range of collections (e.g. plant, insect, vertebrate, fossil, microscope slides). We believe such characterizations are an early step in the process of building a common framework for sharing efficiencies across biological and paleontological research collections.

(URLs provided for first mention only. Please see Appendix 2 for URLs of software and websites.)

Method

This study used the qualitative, *grounded theory research methodology* (Glaser and Strauss 1967, Charmaz 2006) as a general conceptual framework for guiding data collection and analysis. Grounded theory is an inductive social science research method that begins with data collection and leads to qualified conclusions (theories) about those data. The method relies on several techniques useful to our study including simultaneous data collecting and analysis, constructing categories from the data rather than from hypotheses, using a constant comparative method during data collection and analysis, advancing theoretical conclusions during the period of data collection, and sampling aimed at theory construction rather than population representativeness. In the case reported here, categorized concepts from our visits and interviews provided the basis for constructing a modular representation of digitization that we found helpful in describing and elucidating clusters of associated tasks. Data collection included a combination of onsite interviews and observations, analysis of written policies, protocols, and procedures, and the use of multiple observers.

Authors Nelson and Paul, from iDigBio, the U.S. National Science Foundation's National Resource for ADBC, made onsite visits to 28 programs in 10 museums and academic institutions for the purpose of documenting digitization workflow components and protocols and assessing productivity (Table 1). Workflows were documented photographically, through field notes, and from collected protocol documents provided by visited institutions. Staff members across administrative levels were interviewed, and workflows were carefully observed where possible, either through demonstrations or during real-time data and image capture. Those interviewed included institutional level administrators, biodiversity informatics managers, collections managers, taxonomists and systematists intimately familiar with digitization of specific organismal groups, workflow coordinators, and data entry and imaging technicians. Institutions selected for visitation varied on institution size, collection size, number of ongoing

Table 1. Summary List of Collections Visited.

Institution	Collections/Programs Visited	Collection Size ‡	Database Software	Database Platform
Yale Peabody Museum (YPM)	**Entomology** †	>1000000	KE EMu	Proprietary
	Invertebrate Zoology	3000000		
	Invertebrate Paleontology†	350000 lots		
	Vascular Plants	350000		
	Global Plants Initiative			
	Connecticut Plants Survey			
Harvard Museum of Compartive Zoology (MCZ)	**MCZ, Entomology (Lepidoptera)** †	several hundred thousand	MCZbase (Arctos)	Oracle
	MCZ, Entomology (Hymenoptera - Formicidae)	1 million pinned Formicidae		

Institution	Collections/Programs Visited	Collection Size ‡	Database Software	Database Platform
Harvard University Herbaria (HUH)	**HUH, Global Plants Initiative (GPI)** †	> 5 million	Specify 6, custom	MySQL
	HUH, California Plants			
American Museum of Natural History (AMNH)	Division of Invertebrate Zoology	> 24000000	Planetary Biodiversity Inventory (*PBI*) for Plant Bugs custom database	MySQL
American Museum of Natural History (AMNH)	Ornithology	> 1000000	KE EMu Microsoft Access	Proprietary
New York Botanical Garden (NYBG)	**Global Plants Initiative (GPI)** †	> 7000000	KE EMu	Proprietary
	Bryophytes and Lichens (LBCC) TCN†			
	Tri-trophic (TTD) TCN†			
	Barnaby Legume Monographs			
	Intermountain Flora†			
	Caribbean Project (ledgers & notebooks)			
	Amazon Project			
	Kohlmeyer Marine Fungus Collection			
University of Kansas (KU)	Biodiversity Institute, **Entomology Collection**†	> 4.8 million pinned	Specify 6	MySQL
Botanical Research Institute of Texas (BRIT)	**Apiary Project**† software demo (into ATRIUM database)	> 1000000	Apiary	MySQL
Valdosta State University Herbarium (VSC)	**Vascular Plants**†	> 60,000	Specify 6	MySQL
	Bryophytes†			
Tall Timbers Research Station (TTRS)	**Vascular Plants**†	11,000	custom database	MySQL Microsoft Access
	Lepidoptera†	1200		
	Ornithology†	4000		
	Mammalian†	1000		
Robert K. Godfrey Herbarium (FSU)	**Vascular Plants**†	> 200,000	custom database	MySQL

† indicates where observers saw the actual digitization process in action.

‡ number of specimens (unless otherwise stated).

digitization projects, organismal group(s) being digitized, and longevity with digitization activities.

Each site we visited received a questionnaire prior to our visit that examined several categories of digitization tasks that we wished to observe (see Appendix 1). We asked that they use the questionnaire as a guide to prepare for the types of questions we would be asking. The questionnaire was divided into several sections and focused on digitization workflows and tasks. Some institutions completed the questionnaire.

Task clusters

In the digitization workflows we observed, protocols for the digitization of biological and paleontological specimens were typically divided into clusters of related tasks. The order in which these task clusters were accomplished was based on a combination of staff availability, equipment, space, facilities, institutional goals, and the type of collection being digitized. Hence, though there was a general pattern to the components included within a particular task cluster, the order of accomplishment of the clusters and the tasks within each cluster varied by institution.

These five task clusters were important components of digitization, but not all were essential to meeting the digitization goals of every organization or of every specimen for every organization. These clusters are presented here in a common order of operation:

- pre-digitization curation and staging,
- specimen image capture,
- specimen image processing,
- electronic data capture, and
- georeferencing specimen data.

It should be noted that quality control and data cleaning tasks were integral to each of these task clusters (a topic reviewed by Chapman 2005b, 2005c, Morris 2005, Harpham 2006). Some institutions included a post-digitization quality control step during which data were internally compared for obvious inconsistencies or anomalies, such as discrepancies between the series of a collector's numbers and the collection dates, data incongruities between local records and duplicates at other institutions, and collection localities outside of a collector's expected geographic range (a topic reviewed by Morris 2005). This could be considered a sixth task cluster, but we chose to consider it an important part of each of the five task clusters.

Observed workflow components

Pre-digitization specimen curation and staging

Curation and staging typically constituted the first step in the digitization workflow, and often had benefits that extended beyond the immediate needs of the digitization program. This step was usually viewed as essential to efficient digitization. Collections managers also reported that it provided a stimulus for attending to needed or neglected curatorial tasks, including opportunities to do the following:

- inspect for and repair specimen damage and evaluate collection health,
- re-pin or remount specimens and replenish or replace preservatives in containers,
- treat specimens for pests,

Figure 1. Pre-digitization specimen curation and staging. Preparing barcodes and imaging labels, affixing barcodes, updating taxonomy. L to R: University of Kansas – Entomology, New York Botanical Garden and Yale Peabody Museum.

- attach a unique identifier (most often a 1- or 2-D barcode) to a specimen, container, or cabinet,
- discover important but previously unknown, lost, or dislocated holdings (e.g. those owned by other institutions or the federal government),
- update nomenclature and taxonomic interpretation,
- reorganize the contents of cabinets, cases, trays, and containers, especially when these are the units of digitization,
- vet type specimens, and
- select exemplars for digitization, when that approach is appropriate.

The last five activities in this list may require the greatest knowledge of the organismal group of any during digitization. Many institutions use students, interns, dependable volunteers, or other full- or part-time technicians to accomplish the other pre-digitization curatorial tasks on this list, including the selection of exemplars for digitizing. However, some institutions also reported success with allowing technicians to take on more responsibility for at least some of the last 5 tasks in the above list (Munstermann and Gall 2010).

In addition, as collections data become more generally available online, updating nomenclature and taxonomic interpretations and vetting type specimens can occur after the publication of data and images on the internet, providing an opportunity for off-site experts to comment on the specimens. The latter approach will avoid what can become a bottleneck in the digitization workflow caused by the limited availability of in-house taxonomic experts or well-trained technicians.

Although the application of specimen barcodes is treated here as part of pre-digitization curation, this placement in the digitization workflow is not universal. Some institutions applied barcodes at or just prior to the time of image or data capture, depending on the customized order of operations. In all cases where barcodes were used, they were applied prior to image capture to allow for the barcode value to be seen in the image, and prior to data capture to ensure that the physical specimen identifier is accurately included in the electronic data record.

Barcodes were used for two primary purposes. For individual specimens, barcodes were affixed or pinned to the single specimen or inserted into a wet container that held a single specimen. For specimen groups, such as taxon trays, wet containers, or a col-

lection of specimens from a single collecting event, barcodes were sometimes affixed to or inserted into the enclosing container. In most instances, when a container was barcoded, the number of specimens within the container was recorded, but individual specimens within a common container and not segregated by separate vials were neither barcoded nor otherwise individually identified. When individual vials containing single specimens were aggregated into larger jars, a replica of the label for the containing jar was sometimes inserted into each vial. In a few cases, the container was barcoded as were the individual specimens within that container (e.g. with Lepidoptera). In this latter case, the specimens were digitized individually, with both the individual specimen and container barcodes recorded in the database.

Linear, one-dimensional barcodes are relatively large and are used in cases where sufficient space is available, for example on vascular plant specimens, bryophyte and lichen packets, and other dry, flat specimens. A smaller version of this type of barcode, printed the size of a standard insect label, was also used in entomology collections. Space is an important constraint in barcode selection.

One-dimensional barcodes used for insect collections had two advantages. They mimicked the other labels in size, thus conserving space between specimens, and, if positioned near the bottom of the pin, were easily viewed and hand scanned without removal.

Two-dimensional barcodes were also used, especially for small specimens. They were preferred by some entomology collections because they could be included on an insect pin with the coded end clearly visible and easily scanned.

Specimen image capture

Determining what to image varied by institution and collection type. Most herbaria imaged entire specimen sheets. Close-up images of particular morphological features (e.g. fruit, flower, or leaf detail) were also sometimes captured. Certain entomological (e.g. ants, butterflies), paleontological, and ornithological collections captured several images of the same specimen with various views (e.g. dorsal, ventral, lateral, hinge, head-on, etc.).

Image acquisition and storage formats also varied by institution (a topic discussed by Morris and Macklin 2006). Many institutions used the Joint Photographic Experts Group (http://www.jpeg.org/committee.html) (jpeg or jpg) file format for distribution on the internet. Some institutions preferred camera raw formats for archiving images as these formats retain all data originally recorded when the image was made. Others preferred the well-documented and widely used Tagged Image File Format (http://partners.adobe.com/public/developer/tiff/index.html) (tiff or tif), which retains all of the original image data and most of the Exchangeable Image File Format (EXIF) data (a topic reviewed by Häuser et al. 2005b). Some manufacturers, notably Nikon and Canon, store images in a proprietary raw format that is easily read by manufacturer-produced software, but usually requires software plug-ins to be manipulated by other image editing applications (e.g. Adobe Systems Inc. Photoshop (http://www.adobe.com/products/photoshop.html) and Lightroom (http://www.

Figure 2. Specimen image capture. Fossil specimen imaging, specimen label imaging. Two very different imaging set-ups. Yale Peabody Museum, University of Kansas - Entomology.

adobe.com/products/photoshop-lightroom.html)). It should be noted that capturing and preserving high quality specimen label images offers opportunities to take advantages of future improvements in image analysis (La Salle et al. 2009), optical character recognition (Haston et al. 2012), natural language processing, handwriting analysis, and data-mining technologies.

Manufacturer-controlled raw formats are not openly documented and are subject to change without public notice. Hence, in 2004, Adobe, Inc. developed the publicly documented digital negative format (dng) as well as a freely accessible software application that converts many proprietary raw formats to digital negatives with little or no data loss (http://www.adobe.com/digitalimag/pdfs/dng_primer.pdf). A few camera manufacturers (e.g. Hasselblad, Leica, Pentax, Ricoh, Samsung) have adopted the digital negative format as the native output for some of their cameras.

From our observations, imaging requires significant specimen handling with attendant opportunities for damage. Hence, most institutions are careful in personnel selection and produce detailed written imaging protocols. However, once an imaging station is installed and properly configured, image acquisition does not appear to be technically challenging and in most institutions we observed is one of the most efficient and productive steps in the digitization process.

Large insect collections sometimes imaged only one label from a single collecting event and applied those data to all specimens associated with that event. Few entomological collections we observed imaged all specimens.

Whereas some institutions imaged only specimens or specimen labels, others included ancillary materials such as collection ledgers (Harpham 2006). Institutions that digitize ledgers typically associate specimen records with the ledger page images that contained additional information about those specimens (see discussion in Australian Museum 2011). Several institutions, especially those with mature digitization programs, expressed the desire to reference external digital objects, such as monographs, published papers, field notebooks, and gray literature to specimen images and records. It is projected that linking such material to specimen records will increasingly become an important enhancement to current specimen digitization protocols.

Imaging station components varied by institution, organism being imaged, and intended use of the resulting images. Most common was a single-lens reflex digital camera fitted with a standard or macro lens and connected to manufacturer or third-party camera control software. A typical station included:

- camera and lens, microscope (for a related discussion, see Buffington et al. 2005), or scanner (HerbScan (see JSTOR PLANTS Handbook http://www.snsb.info/SNSBInfoOpenWiki/attach/Attachments/JSTOR-Plants-Handbook.pdf) or a custom-designed replica), SatScan (Blagoderov et al. 2010), GigaPan (Bertone and Deans 2010),
- cable connecting camera to computer,
- camera control software (third party or camera manufacturer produced),
- image processing software (most common are Canon Digital Photo Professional (http://www.canon.com), Nikon Capture NX2 (http://www.nikonusa.com), Photoshop, and Lightroom), image stacking equipment and software, for example Helicon Focus (http://www.heliconsoft.com/heliconfocus.html) or Auto-Montage (http://www.syncroscopy.com/syncroscopy/automontage.asp) (for a related discussion of Auto-Montage, see Antweb (2010)),
- remote shutter release (wireless or tethered),
- copy stand and/or specimen holder,
- studio lighting, flash units, or light/diffuser box (e.g. MK Digital's Photo EBox Plus (http://www.mkdigitaldirect.com/products/lighting-systems/mk-photo-ebox-plus-1419.html)),
- scale bar,
- color standard,
- stamp to mark that a sheet, jar, tray, or folder had been imaged, and
- associated instruments (pinning blocks, forceps, latex gloves, etc.).

The most common brand of camera in use across collections was a Canon DSLR equipped with a medium-length macro lens, although Nikon DSLR cameras were also sometimes used. Megapixel ratings generally ranged from about 17 to 21.5, but were sometimes lower or higher, depending upon the expected use of the images.

It is instructive to note that generally, the larger the megapixel rating, the better the quality of the resulting images. Hence, images to be used for morphological

study were usually captured at megapixel ratings of 17 and above. Macro lenses in the range of 50–60 mm were common, but a few institutions used macro lenses in the range of 100–105 mm, which allowed for close focusing and performed well for smaller objects, such as small birds and mammals. Collections requiring macro images of very small specimens usually used a Leica microscope equipped with a Canon, Nikon, or Leica camera.

To control for image quality, some institutions located the imaging station in a darkened or minimally lit windowless room. This prevented strong extraneous light, like that from a window, from contaminating or overpowering studio lighting or producing visible shadows on the resulting images. Light control was also sometimes accomplished by draping diffuser material across studio lights. A more elegant solution utilized a diffuser box with internal lighting that can be closed prior to image capture. Preferred for this was the MK Photo-eBox Plus Digital Lighting System, originally designed for photographing jewelry, coins, and collectibles. The box is slightly larger than a standard herbarium sheet, rests on a copy stand, includes halogen, fluorescent, and LED lighting, and is equipped with an oval port on the upper surface that allows an unobstructed camera view of the specimen. Herbaria using this system usually place the color bar and scale at the top of the sheet to preserve the aspect ratio of the resulting image, thus obviating the need for image cropping and reducing the number of steps required for image processing. Although the requirement to open and close the doors of the light box seemingly slowed the imaging rate, time lost was likely recaptured from a reduction in time spent on post-imaging batch cropping and light level adjustments.

HerbScan is the imaging system used for scanning type specimens for the Global Plants Initiative (GPI) project (http://gpi.myspecies.info/). GPI specifications require that specimens be scanned at 600 ppi resolution, beyond the capacity of most DSLR cameras when used for whole sheet images of herbarium specimens. HerbScan uses a flatbed scanner (Epson Expression Model 10000XL, Graphic Arts, USB2 and Firewire interfaces) and a platform that raises the specimen sheet to the face of the inverted scanner. Scanning requires 4-6 minutes per scan for a maximum effective rate of about ten images per hour. Because the specimen sheet is pressed against the rigid glass face of the scanner, the acceptable depth of the specimen sheet is limited to about 1.5 cm, hence some specimens are too bulky for this equipment.

Keeping up with what has and has not been imaged can be daunting, especially in large collections. Many collections that we observed used the presence of a barcode or a stamp to indicate whether a particular specimen had been imaged and/or digitized. Herbaria often stamped the sheet or folder at the time of imaging to provide a visible demarcation. Some institutions also used a written or electronic tracking system to track digitization in an orderly fashion. Electronic tracking was usually accomplished within the database management system being used for data storage. For many institutions, deciding what to digitize was based on such criteria as responding to special projects, processing loan requests, emphasizing centers of interest, a desire to focus on unique or important parts of the collection, or other

priorities. In such instances, an electronic tracking system ensured that specimens were not overlooked.

Maintaining an organized tracking system for actively growing collections is especially dependent on effective protocol. Some institutions included digitization within the accessioning workflow, ensuring that all newly acquired specimens, especially those to be inserted into parts of the collection that had been previously digitized, were handled at the time of specimen acquisition.

Workflow requirements for imaging varied by institution, but generally followed a similar pattern:

- pre-imaging equipment configuration and initialization,
- procuring/organizing the next batch of specimens for imaging,
- acquiring the image, and
- moving specimens to the next station or re-inserting them into the collection.

Pre-imaging equipment configuration and initialization was generally a one-time task accomplished at the beginning of an imaging session. It involved:

- connecting or ensuring the connection of computer to camera,
- starting external studio lighting, or checking, adjusting, and testing flash units and power supplies,
- starting camera control and image acquisition software,
- starting the camera,
- setting camera aperture, shutter speed, and focus point (or loading these attributes from a previously configured settings file),
- adjusting camera height,
- changing or attaching lenses, and
- loading ancillary image management/processing software.

In some institutions, especially those where all specimens are similarly sized (e.g. herbaria), camera settings and equipment mountings were usually not changed from session to session and required only a spot check prior to commencing a new imaging session. With collections of variously sized organisms (e.g. paleontological, ornithological, Lepidopteran), camera distance to subject was frequently adjusted, lighting re-arranged, camera settings altered, and custom or specialized specimen holders repositioned. In some instances, grouping like-size specimens alleviated the need for continuous camera adjustment and increased workflow efficiency. In these situations, the potential increase in imaging error due to increased demands for technician judgment were effectively offset by a higher level of detail in written protocols, elevated attention to specialized training, and diligent monitoring during the early phases of a new technician's tenure. Institutions that imaged only labels that required only moderate resolution sometimes dispensed with much of the equipment listed above in favor of a small digital camera and less elaborate copy stand that afforded more mobility (Figure 2).

Procuring and organizing the next batch of specimens for imaging was sometimes facilitated by ensuring proximity of the specimens to the imaging station. Institutions used mobile carts or cabinets to transport specimens from the pre-digitization curation or data entry areas to a location in close proximity to the imaging station. Moving specimens from station to station rather than returning them to storage cabinets and re-retrieving them reduced the amount of time devoted to travel and handling. From our observations, workflows that began with image capture, imaged every specimen, and extracted data directly from the image rather than the physical specimen effectively eliminated the need to handle or move specimens beyond the imaging stage, facilitating re-storage immediately following imaging (Figure 6c). To ensure that specimens did not get misplaced and potentially lost within the collection, re-filing specimen drawers, trays, containers, or folders was often reserved for curators or technicians intimately familiar with collection organization. To facilitate the smooth flow of specimens, staging space was often made available at every station where physical specimen handling was required.

Image acquisition focuses on the process of camera operation for image capture. For collections with standard sized specimens (e.g. herbaria), the process involved repeating a rote procedure for each new specimen. Even for such collections, however, the technician was required to pay close attention to quality by periodically examining images to ensure that:

- lighting, exposure, and focus remained constant,
- file naming progressed according to plan,
- exposure was correct,
- focus remained sharp,
- images lacked imperfections such as blemishes or streaking,
- files were not corrupted, and
- barcodes or identifiers were in place and readable.

For wet collections, exemplar specimens were usually removed from the container before imaging. One successful technique we observed for imaging fish, reptiles, amphibians, and other organisms with a reflective epidermis submerged them in a shallow, ethanol-filled container, allowed the ripples to settle, and acquired the image through the ethanol. This method increased detail by reducing reflectance and increasing contrast. Coating fossil specimens with a thin layer of alcohol also increases contrast and provides for a sharper image (Paul Selden, personal communication, 2012).

Protocols and workflows for efficiently imaging insects—with the possible exceptions of bees, ants, and butterflies—are under development and continue to pose special challenges. In nearly every case where we observed butterflies being imaged, specimens were removed from the pinning substrate, labels were carefully removed and placed on a custom-designed holder with the labels and barcodes (or other identifier)

Figure 3. Custom specimen holder. Museum of Compartive Zoology (MCZ) Rhopalocera (Lepidoptera) Rapid Digitization Project.

clearly visible in the resulting image. One institution (Museum of Comparative Zoology) designed and constructed a custom specimen holder (Figure 3) with sufficient space to include all labels and the specimen in a single image (Morris et al. 2010). Other institutions rested the specimen on a parallel pair of taut monofilament lines and recorded two views (dorsal and ventral), each with one or more labels visible (see Häuser et al. 2005a). Some institutions combined the dorsal and ventral views side-by-side in a single composite image using image management software such as ImageMagick (http://www.imagemagick.org/).

Imaging productivity varied by collection. For herbaria, rates per imaging station ranged from as few as 10 sheets per hour using a single HerbScan, to 75–120 sheets per hour using a camera (average rate slightly less than 100 sheets per hour). Imaging rates for insects are not well documented and their derivation is sometimes confounded by the inclusion of data entry and image acquisition in a single, linear workflow that makes it difficult to segregate strictly imaging tasks from data entry. For example, the imaging step might include removing the label from the pin, taking the photo, and putting the label(s) back on the specimen pin.

Specimen image processing

Image processing involves all tasks performed on an image or group of images following image capture. Nine tasks are addressed here, reflecting common practices:

- quality control,
- barcode capture,
- file conversion,
- image cropping,
- color balance or light level adjustments,
- image stacking,
- redaction,
- file transfer, and
- optical character recognition (OCR).

Some institutions include one or more of these nine tasks (e.g. barcode capture, OCR) at other stages of the digitization process, as noted in the discussion below.

Quality control was usually effected by selecting and examining sample images at regular intervals. In some institutions, all images were visually scanned for obvious deficiencies before individual images were selected for more thorough review. Selected images were evaluated for correct focus and exposure, blemishes, scan lines, mismatches between file names and barcode values (in situations where these are expected to match), and other obvious signs of imperfections or errors. Imperfections in camera images usually related to incorrect focus or exposure. Institutions using HerbScan, especially as part of the GPI, followed a more elaborate and rigorous process (not detailed here) that included converting images to high contrast in Photoshop and running scripts that track pixilation and banding, and that expose scanner-produced flaws such as minute streaks and lines caused by wear and tear on scanner parts. The standard for GPI images, coupled with mechanical parameters of the scanners, demanded these enhanced quality control procedures (http://www.snsb.info/SNSBInfoOpenWiki/attach/Attachments/JSTOR-Plants-Handbook.pdf).

Barcode values were captured in several ways and for several purposes. Many institutions preferred specimen image file names to match corresponding specimen barcode values. Hence, the image file for a specimen with barcode value XXX123456, might be named XXX123456.tif, where XXX is replaced by the institution code. This worked well for cases in which each specimen was represented by a single image, but less effectively for cases in which a specimen might be represented by multiple images. In these latter cases, multiple image files of the same specimen often used an appended value, such as XXX123456A, XXX123456B, and so forth. Although matching the image filename to the specimen's barcode value is not a requirement, it is a common practice that helped ensure that all image files for a specific collection were uniquely named.

Based on our observations, collections that chose to use barcode values as filenames generally used one of several options. Most high-end DSLR cameras allow for cus-

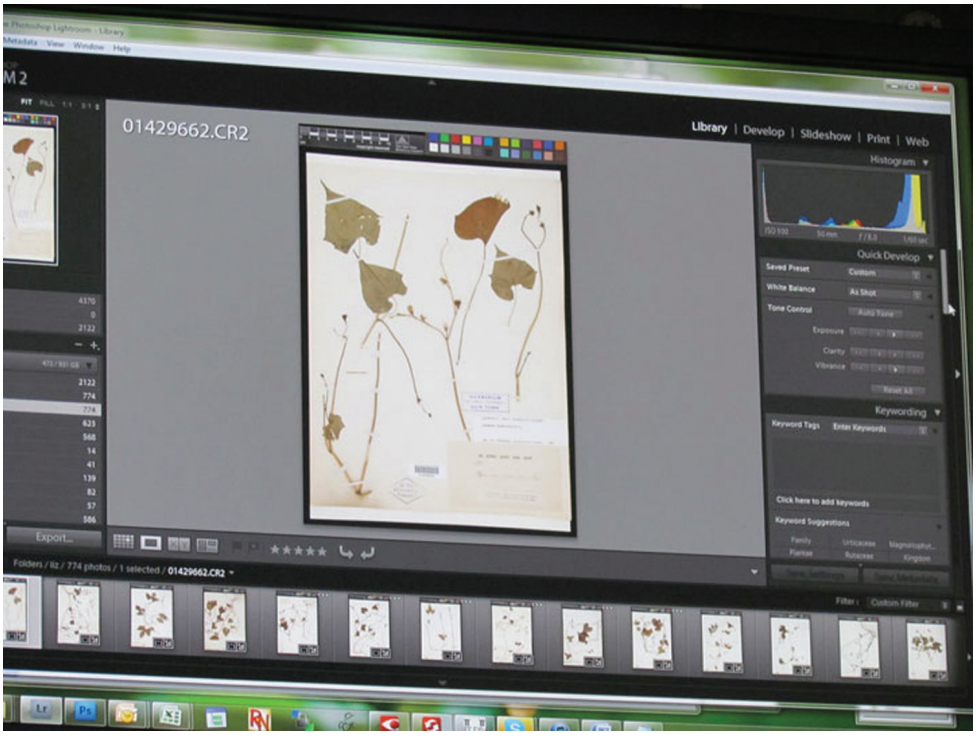


Figure 4. Specimen image processing. Using Adobe Photoshop Lightroom software to process images. New York Botanical Garden.

tomized file naming and auto-incremented file numbering, features sometimes used in herbaria. When these features were used simultaneously, the camera was configured to produce file names that matched the barcode value. This increased efficiency when specimens were arranged and imaged in sequential barcode order, but was cumbersome and inefficient when specimens were arranged in random barcode order. It also led to file naming errors when one or more specimens were unexpectedly mis-ordered. A second practice used a barcode scanner to read the barcode into the file name field or the image EXIF data as the file was imaged or saved. A third strategy used Optical Character Recognition (OCR) software to scan the image file for a barcode value and rename the file to the barcode value detected. The benefits of the latter approach included reduction of potential naming errors and greater efficiency due to reduced camera manipulation.

However, OCR software sometimes failed at detecting barcodes within images due to image quality or other issues, resulting in files not being appropriately renamed. According to our observations, barcode extraction failure rates on bryophyte packets ranged from 0.2–3%, based on tests with ABBYY Finereader Corporate edition (http://finereader.abbyy.com/corporate/) at the herbarium of Valdosta State University, where barcodes were carefully affixed in precise horizontal or vertical orientation. A fourth approach used custom-designed software to intercept the filename generated by the camera, simultaneously creating an associated record in the database for later data entry from

the image. Image filenames were unique for the collection, and the image files were usually stored in a repository and linked to database records through a software interface.

A two-part strategy we observed that addressed file naming issues used a hand-held scanner to scan the barcode value into the image EXIF via Canon Digital Professional software. Subsequent processing extracted the image's barcode value using ZXing (Zebra Crossing, http://code.google.com/p/zxing/), compared the value to the image's EXIF data, and created a database record containing the image filename and barcode value. This allowed database records to be created by software without regard to the image's filename. The key point of this process is that camera-generated filenames can be stored verbatim in a database if software is responsible for associating image files with specimen records (Morris and Macklin 2006).

Conversion involves converting camera raw images to a preferred archival or display format. In some instances, conversion is avoided by setting the camera to record images in the preferred final archive format, usually as a tagged image file (tif).

Cropping is used to trim excess image data in order to achieve an acceptable aspect ratio or to reduce unnecessary borders surrounding the specimen. Where cropping was utilized, it was accomplished in large batches that did not require monitoring once set into motion. However, cropping was not universal.

In general practice, it is considered unwise to use photo manipulation software to alter color balance, saturation, sharpness, or other image features (Cromey 2010). Doing so runs the risk of creating an image that does not faithfully represent the source specimen. Based on our observations, adjustment of light levels is an exception to this rule. Herbarium specimens, in particular, sometimes benefitted from an automatic levels adjustment. An auto levels adjustment essentially sets the white and black points in the image and spreads the available tones between these two extremes. Using an auto-levels adjustment worked best when the image contained a color bar that included true black and white reference points. This gave a better representation of the tonal values between the extremes, and usually resulted in a more lifelike image without distorting color or other attributes. Since all herbarium specimens in a specific photographic session were presumably recorded with equal illumination, consistent camera settings, and the same lens, all images made within that session benefited equally from a batched adjustment. The same was not always true for colorful subjects, such as birds or butterflies, which often responded to auto levels adjustments in a way that distorted the resulting images, often rendering them more colorful and brighter than the original.

Specimens with significant depth, such as fossils, some insects, birds, mammals, and even some herbarium sheets, make it difficult to achieve sharp focus throughout the depth of field. Institutions used one of several stacking software packages to rectify this problem. Focus stacking (http://en.wikipedia.org/wiki/Focus_stacking) involved recording several images of a stationary specimen at varying depths of field, processing them through a stacking algorithm that essentially merged the several layers into a single image while preserving properly focused pixels in each layer. The result was a sharply focused image throughout the specimen's depth. Software packages in common use included proprietary Auto-Montage (see discussion in Antweb 2010) and

Helicon Focus. No-cost software included CombineZ (http://www.hadleyweb.pwp.blueyonder.co.uk/CZP/Installation.htm). Stacking worked best with cameras that supported a live view of the specimen in conjunction with camera control software that allowed precise focus control targeted to small percentage regions of the specimen.

Electronic data capture

Electronic data capture involves extracting label data and entering those data into an electronic database. Depending on protocol, data capture can occur before, after, or simultaneous with image capture. For collections we observed in which all or nearly all specimens were to be imaged, entering data from specimen images reduced specimen handling and potential damage, eliminated multiple trips to storage locations, and allowed technicians to digitally enlarge labels for better readability. For collections that did not image specimens, or imaged only exemplars, data entry was usually the second step in the digitization sequence (Figure 6a).

Several methods were used for data capture, the most common being keystroke entry, sometimes with the support of related technologies such as OCR or voice recognition. Efficiently designed software interfaces that allowed user customization were important and increased the efficiency of data entry by eliminating duplicative or unnecessary keystrokes and arranging icons in convenient positions or in logical tab orders (see related discussion in Morris 2005). We noted that in almost all cases, the database software used in a given collection was not used out-of-the-box. Often, software was customized or custom-designed user interfaces were built by biodiversity informatics managers.

Advances in voice recognition technology are evident in computer, tablet, and smart phone applications. Nevertheless, we saw only a single use of this technology, and this only for capturing a limited set of data, but we note that some institutions are experimenting with this technology. IBM ViaVoice (now produced by Nuance Communications, Inc. (http://www.nuance.com/)), Microsoft Voice Recognition (a standard component of the Microsoft Windows® operating system), and Dragon Naturally Speaking (http://www.nuance.com/for-business/by-product/dragon/dragon-for-the-pc/dragon-professional/index.htm) are three software packages being used or tested. We note that programmers at the Botanical Research Institute of Texas (BRIT) are testing the Application Programming Interface that is packaged with the Microsoft Windows® operating system. We believe that voice recognition shows great potential for data capture and that the comparatively small cost for appropriate commercial products will be offset by greater workflow efficiencies. Most modern operating systems include built-in voice recognition capabilities of various qualities that should be tested using a high quality microphone. From our experience, the potential drawback to this technology is that substantial training to particular voices is often required for the software to perform adequately, which may limit its use where several data entry technicians are involved or when the rate of technician turnover is high. In addition,

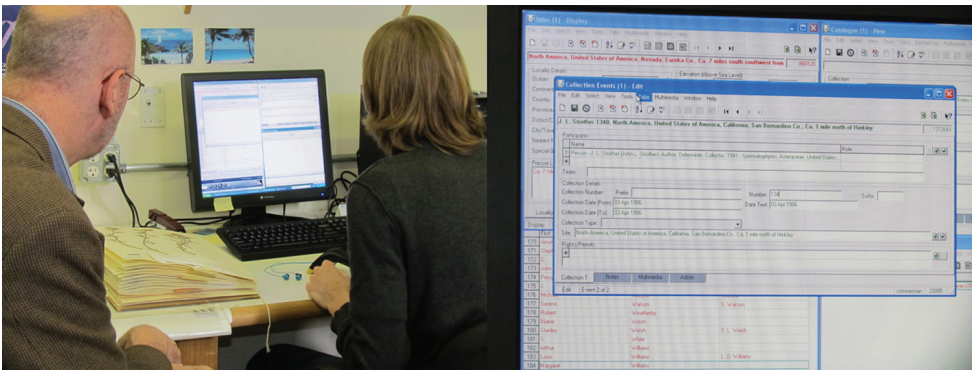


Figure 5. Electronic data capture. Entering data straight from the specimen label into the database. New York Botanical Garden.

we noted from our interviews that simultaneous data entry by several technicians in close proximity might lead to distortion and interference, or be distracting to workers.

Optical character recognition (OCR) was also being used or considered by several institutions. Two of the most effective uses we observed included the Apiary Project (http://www.apiaryproject.org/) at BRIT and the Symbiota Software Project (http://symbiota.org/tiki/tiki-index.php) at Arizona State University. Each of these interfaces simultaneously displays a specimen image, an OCR-rendered version of label data extracted from the image, and a collection of database fields into which data can be transferred. Apiary allows users to demarcate OCR regions of interest within the image and highlight OCR-generated text that can be transferred to associated data fields by mouse click. Symbiota provides for moving data to fields manually, but additionally includes functionality for searching the databases of the Consortium of North American Byrophyte Herbaria (http://symbiota.org/bryophytes/) and Consortium of North American Lichen Herbaria (http://symbiota.org/nalichens/) for previously digitized duplicates from which data can be imported.

Other institutions routinely process all images through OCR and store the OCR-generated output in text files, or import it into a field within the database for subsequent editing, data cleaning, and searching. Popular OCR software packages included Tesseract (http://code.google.com/p/tesseract-ocr/), OCRopus (http://code.google.com/p/ocropus/), and JOCR (GOCR) (http://jocr.sourceforge.net/), all of which are open source, and the proprietary ABBYY Finereader corporate version (http://www.abbyy.com/) and Adobe Acrobat Professional version (http://www.adobe.com/products/acrobatpro.html), both of which can batch process large numbers of images. There is significant interest in natural language processing (NLP), which is designed to parse OCR text into fields, as well as intelligent character recognition (ICR) or handwriting analysis, but effective systems for using these technologies to extract data from biological specimens were not observed.

In some instances data entry is accomplished by electronic import from spreadsheets or other delimited lists. Some software interfaces, e.g. Specify (http://specifysoftware.org/) (via Workbench), Brahms (http://herbaria.plants.ox.ac.uk/bol/) (via Rapid Data

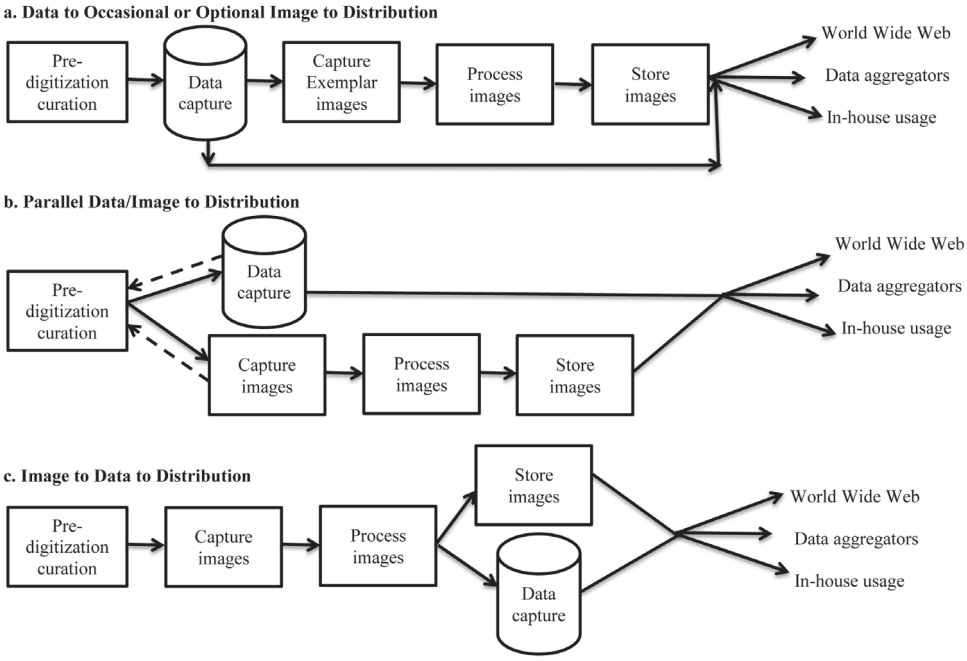


Figure 6. Dominant Digitization Workflows Observed.

Entry http://herbaria.plants.ox.ac.uk/bol/BRAHMS/Documentation), and KE EMu (http://www.kesoftware.com/) provide this capability. Issues to resolve when importing legacy or external data include data quality, mapping imported data fields to those in the preferred database, dealing with imported fields that do not have database correlates, and time required for post-import data cleanup. In many cases, importing and transforming legacy data can be efficiently managed, resulting in large dataset acquisitions for relatively small investment in time, especially when compared to keystroking.

Georeferencing

Georeferencing is the process of transforming textual descriptions of geographical data into a pair of X, Y coordinates, with an accompanying estimation of precision. Precision is usually denoted by one of several methods, including a bounding polygon, a point and its associated radius of uncertainty, or designation of the extent of the known area in which the point occurs, such as a county, park, township, range, or section (Chapman and Wieczorek 2006). Best practices suggest that each georeferenced point also include notation of the point's datum, geographical coordinate system, and georeference remarks that explain how the point, polygon, and estimate of precision were derived (Chapman and Wieczorek 2006). Coordinate pairs that do not include notation of the underlying datum upon which the point is based may include uncertainties up to about 3.5 km (Wieczorek et al. 2004).

Based on our observations, the process of georeferencing biological and paleontological specimens was typically ancillary to and discontinuous with the digitization workflow. Although digitization workflows often captured locality information from specimen or collecting event labels, these data—especially legacy data—generally did not contain geographical coordinates and most institutions chose not to georeference these data at the time of data entry. In the case of more recently collected specimens on which latitude and longitude values were included on the label, the values were typically captured at the collecting event or specimen record level at the time of data entry. It is clear from our observations that the community consensus for legacy specimens is for bulk georeferencing of unique localities as a separate step in the digitization workflow (Chapman and Wieczorek 2006).

We observed three georeferencing methodologies in use where coordinate values were not present on the specimen. Geolocate (desktop and web-based interfaces, and web services; http://www.museum.tulane.edu/geolocate/) and Biogeomancer (web-based; http://bg.berkeley.edu/latest/) are software applications designed to assist in assigning latitude/longitude coordinates to textually described localities. Both of these applications convert locality descriptions into coordinate pairs based on statements of state, county, orthogonal direction, distance, and place names of geographical features. Both also provide protocols for uploading datasets for processing and bulk georeferencing similar localities. Each returns a map of the estimated location of each described locality, including a point-radius estimate of precision. Map interfaces allow technicians to manipulate and refine the georeferenced locations of these points before recording a final determination of the point's coordinates. Technician manipulation was required for points to be reliable. Both Geolocate and Biogeomancer are free to use. The third method we observed was based on the use of standard and customized map layers in conjunction with GIS software (such as ArcMap http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=An_overview_of_ArcMap) and paper maps to pinpoint locations. For best results, all of these systems rely on a technician's knowledge of the region in which a collection is made, facility with desktop GIS or online mapping software, general understanding of maps and mapping, and ability to recognize habitat signatures on aerial photographs.

Dominant digitization workflows observed

Based on our observations, three workflows dominated digitization programs in the institutions we visited (Figure 6). The three presented here are not intended to represent a comprehensive collection of workflows. Here we call them by their characterizing patterns: *data to occasional or optional image to distribution*, *parallel data/image to distribution*, and *image to data to distribution*. All patterns begin with pre-digitization curation and terminate with distributing data directly to the World Wide Web, to data aggregators, and/or to internal users. In all three, specimen data are stored in database records that include references to associated images or other media. Images are stored

in a computer file system and are not embedded in the database. We have not measured the throughput of these patterns in a controlled experiment.

It is worth noting that the capture of specimen data from ledgers without reference to the specimens has been a dominant digitization workflow for many decades and represents the method by which the majority of existing vertebrate collections data were digitized (Humphrey and Clausen 1977). With one exception, this method was absent from the workflow patterns we observed in this study, likely due to the transition in recent years to digitizing directly from specimens.

We note that Tann and Flemons (2008) and Granzow-de la Cerda and Beach (2010) provide examples of how one might measure a data capture workflow for a given collection type. These might serve as models for setting up comparisons of workflows across or within collection types.

The *data to occasional or optional image to distribution* pattern fits those institutions in which few or no specimens are imaged. Data capture follows curation and may include decisions about which specimens to submit for imaging. Rarely, imaging of exemplars is simultaneous with data entry of those exemplars.

The *parallel data/image to distribution* pattern includes both data and image capture but treats them as independent and simultaneous rather than as sequential steps. This pattern is likely the most labor intensive of the three, especially when it requires specimen handling at two stages of the workflow, with attendant need for multiple trips to storage locations and increased opportunities for specimen damage. This pattern is made more efficient when data capture proceeds from bulk data sources (ledgers, cards), which requires specimen handling only during image aquisition.

The *image to data to distribution* pattern fits institutions that image all specimens (e.g. most herbaria) and captures data from these images. It reduces specimen handling and with it the likelihood of specimen damage, increases efficiency by eliminating the need for return trips to storage locations, and offers the capacity to incorporate Optical Character Recognition and similar technologies within the data capture workflow.

Recommendations

Based on our observations, interviews, discussions, and readings, we offer the following recommendations for establishing and improving biological and paleontogical collections digitization programs.

1. With planning, the pre-digitization curation step is an opportunity for the goals of specimen digitization and collection curation to be merged into an efficient workflow. Curation tasks that cannot be efficiently addressed in the workflow can be identified so that adequate resources can be assigned to them in the future (Sumpter 1991).
2. Biodiversity informatics managers and other digitization personnel should look for bottlenecks in digitization workflows and seek ways to make them

more efficient (Tann and Flemons 2008; Granzow-de la Cerda and Beach 2010). We recognize that much work remains for devising and disseminating strategies for evaluating and analyzing existing workflows, encouraging the application of automation, and exploring the relevance of industrial process control to workflow design.

3. There should be clear institutional policies guiding which specimens to expose to public access, including policies governing whether to redact or not redact locality data for sensitive species (Canhos et al. 2004) and ensuring that permission is obtained for privately controlled donations and collections from federal installations. We note, for example that funds from NSF's Advancing Digitization of Biological Collections are not permitted to be used in the digitization of federally owned specimens (National Science Foundation 2011).
4. Barcodes should be used only as identifiers; encoded barcode data should not incorporate taxonomic or related information that might change with time.
5. Where possible, the aspect ratio of specimen to camera should be synchronized to eliminate the need for image cropping.
6. Image processing should not include color balancing or other adjustments that result in images inaccurately reflecting actual specimens (Cromey 2010).
7. A color bar and scale should be visible in all images (Taylor 2005).
8. Protocols for periodic quality control should be established for all stages in the digitization workflow to ensure data accuracy and the production of high quality digital images (Chapman 2005a).
9. For institutions in which imaging is paramount, acquiring images of labels prior to data entry reduces specimen handling by allowing for data extraction from images rather than from specimens.
10. Attention to the digitization of gray and published literature related to specimen data is an important consideration and should be accomplished whenever possible (cf. Australian Museum 2011).
11. Georeferencing should be treated as an essential part of digitization protocols (Canhos et al. 2004, Chapman and Wieczorek 2005, Morris 2000).
12. Quality control should be integral to all steps in the digitization workflow, including post-digitization review and targeted testing should be designed to expose data inconsistencies or suspected anomalies (Morris 2005).
13. Detailed written protocols should guide every step of the digitization workflow, be uniquely designed for a given institution, and be amended regularly to reflect emerging technologies and improved efficiencies. These protocols should be electronically stored in a common folder that allows technicians to insert comments and suggestions to be reviewed and potentially adopted by biodiversity informatics managers.
14. Selection of data entry and imaging technicians should be guided by employability skill sets strongly associated with success in digitization tasks, with particular attention to potential technicians' attention to detail, orientation to increased efficiency, and commitment to high productivity.

15. Institution-wide digitization tasks should be periodically evaluated for overall progress, organizational collaboration and cooperation, and compatibility with new and emerging technology, with plans to use results of the evaluation to implement improvements (Kalms 2012).
16. Digitization workflows should be coordinated by a designated biodiversity informatics manager with IT experience, preferably from a biological sciences and collections background, to bridge the potential knowledge gap between collections managers and information technology professionals (Kalms 2012).
17. Biodiversity informatics managers should construct a frequently asked questions document that outlines common problems and offers instructions about how to address these problems, whom to contact with questions about specific categories of problems, and guidelines for which types of problems should be elevated to a higher administrative level.
18. Institutions should utilize a digitization workflow strategy that captures problems, remedies, lessons learned, and technician input for use in improving digitization protocols, and remain open to investigating possible changes in current practice (Kalms 2012).
19. Determining an appropriate storage format for archived images is an important decision that should precede image capture. Here we recommend capturing images in native camera raw and converting them from camera raw to dng or tif (a topic addressed by Häuser et al. 2005b). Alternatively, images can be natively captured and archived in tif format. Jpg format is not recommended for archived images.

Acknowledgements

The staff of iDigBio thanks all participants listed in Table 1 for efforts extended on our behalf. Every institution graciously accorded open access and provided us with a very special opportunity to see so much in such a very short time. We also thank Paul Morris and David Roberts for their helpful comments on the manuscript.

This work was made possible by the U.S. National Science Foundation's Advancing Digitization of Biological Collections Program, grant (#EF1115210).

References

Antweb (2010) Automontage imaging guidelines. http://www.antweb.org/homepage/AntWeb%20Imaging%20guidelines%20v01.pdf [accessed 7.XI.2011]

Australian Museum (2011) A Guide to Handling and Digitizing Archival Material–Registers. http://australianmuseum.net.au/Uploads/Documents/22932/Archive%20Training%20Compressed.pdf [accessed 17.XI.2011]

Bertone MA, Deans AR (2010) Utility (and shortcomings) of high resolution drawer imaging for remote curation and outreach. Presentation to the Entomological Collections Net-

work – Annual Meeting. San Diego, CA December 11, 2010. http://www.ecnweb.org/dev/files/17_Bertone_2010.pdf [accessed 2.X.2011]

Blagoderov V, Kitching I, Simonsen T, Smith VS (2010) Report on trial of SatScan tray scanner system by SmartDrive Ltd. Available from Nature Precedings http://precedings.nature.com/documents/4486/version/1/files/npre20104486-1.pdf [accessed 17.XI.2011]

Buffington ML, Burks RA, McNeil L (2005) Advanced Techniques for Imaging Parasitic Hymenoptera (Insecta). American Entomologist 51(1):50–56. http://www.entsoc.org/PDF/Pubs/Periodicals/AE/AE-2005/Spring/Buffington.pdf

Canhos VP, Souza S, Giovanni R., Canhos DAL (2004) Global Biodiversity Informatics: Setting the Scene for a "New World" of Ecological Modeling. Biodiversity Informatics 1:1–13.

Chapman AD (2005a) Uses of Primary Species-Occurrence Data, Version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. http://www.gbif.org/orc/?doc_id=1300

Chapman A (2005b) Principles and Methods of Data Cleaning–Primary Species and Species-Occurrence Data. Version 1. Copenhagen: Report for the Global Biodiversity Information Facility. http://www.gbif.org/orc/?doc_id=1262

Chapman A (2005c) Principles of Data Quality. Version 1. Copenhagen: Report for the Global Biodiversity Information Facility. http://www.gbif.org/orc/?doc_id=1229

Chapman A, Grafton O (2008) Guide to Best Practices for Generalizing Sensitive Species Occurence Data. Version 1. Copenhagen: Report for the Global Biodiversity Information Facility. http://www.gbif.org/orc/?doc_id=1233 [accessed 26.XII.2011]

Chapman AD, Wieczorek J (Eds) (2006) Biogeomancer Guide to Best Practices For Georeferencing. Copenhagen: Global Biodiversity Information Facility. Available online at http://www.gbif.org/orc/?doc_id=1288 [accessed 21.XII.2011]

Charmaz K (2006) Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. SAGE Publications, London.

Cromey DW (2010) Digital Imaging: Ethics. University of Arizona. http://swehsc.pharmacy.arizona.edu/exppath/resources/pdf/Digital_Imaging_Ethics.pdf [accessed 28.X.2011]

Glaser BG, Strauss AL (1967) The Discovery of Grounded Theory: strategies for qualitative research. Aldine Publishing Company, Chicago.

Global Biodiversity Information Facility (2008) GBIF Training Manual 1: Digitisation of Natural History Collections. http://www.infoandina.org/system/files/recursos/GBIF_TM1.pdf

Granzow-de la Cerda Ì, Beach JH (2010) Semi-automated worflows for acquiring specimen data from label images in herbarium collections. Taxon 59(6):1830-1842 [accessed 20.II.2012].

Harpham S (2006) Documentation Standards Review: Procedures for Database Upgrades. Collection Forum 21(1–2):192–202 [accessed 22.V.2012]

Haston E, Cubey R, Harris DJ (2012) Data concepts and their relevance for data capture in large scale digitisation of biological collections. International Journal of Humanities and Arts Computing 6(1–2):111–119. doi: 10.3366/ijhac.2012.0042

Häuser CL, Holstein J, Steiner A (2005a) Digital imaging of butterflies and other Lepidoptera. More or less "flat" objects? In: Häuser CL, Steiner A, Holstine J, Scoble MJ (Eds) (2005) Digital Imaging of Biological Type Specimens. A Manual of Best Practice. Results from a study of the Europoean Network for Biodiversity Information. Stuttgart, 254–261.

Häuser CL, Steiner A, Holstine J, Scoble MJ (Eds) (2005b) Digital Imaging of Biological Type Specimens. A Manual of Best Practice. Results from a study of the Europoean Network for Biodiversity Information. Stuttgart http://imsgbif.gbif.org/CMS_ORC/?doc_id=2429 [accessed 29.X.2010]

Humphrey PS, Clausen AC (1977) Automated Cataloging for Museum Collections: a model for decision and a guide to implementation. Association of Systematics Collections, Lawrence, Kansas, 79 pp.

Joint Photographic Experts Group. http://www.jpeg.org/committee.html [accessed 29.X.2010].

JSTOR Plants Handbook http://www.snsb.info/SNSBInfoOpenWiki/attach/Attachments/JSTOR-Plants-Handbook.pdf

Kalms B (2012) Digitisation: A strategic approach for natural history collections. Canberra, Australia, CSIRO. Available at http://www.ala.org.au/wp-content/uploads/2011/10/Digitisation-guide-120223.pdf [accessed 6.III.2012]

La Salle J, Wheeler Q, Jackway P, Winterton S, Hobern DL (2009) Accelerating taxonomic discovery through automated character extraction. Zootaxa 2217: 43–55. http://www.mapress.com/zootaxa/2009/f/zt02217p055.pdf [accessed 28.XI.2011]

Lichens, Bryophytes and Climate Change LBCC http://lbcc.limnology.wisc.edu/

Mares MA (2010) A Strategic Plan for Establishing a Network Integrated Biocollections Alliance [brochure]. Available March 3, 2011. http://digbiocol.files.wordpress.com/2010/08/niba_brochure.pdf. See also http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503559 [accessed 18.XI.2011]

Morris PJ (2000) A Data Model for Invertebrate Paleontological Collections Information. In: White RD, Allmon WD (Eds) (2000) Guidelines for the Management and Curation of Invertebrate Fossil Collections Including a Data Model and Standards for Computerization. National Science Foundation Workshop at the North American Paleonotological Conference, Washington DC, June 7–14, 1996. The Paleonotological Society, New Haven (USA), 105–108.

Morris PJ (2005) Relational Database Design and Implementation for Biodiversity Informatics. Phyloinformatics 7:1–66. http://systbio.org/files/phyloinformatics/7.pdf [accessed 5.III.2011]

Morris P, Eastwood R, Ford L (2010) Innovative Workflows for Efficient Data Capture in an Entomological Collection: The MCZ Rhopalocera (Lepidoptera) Rapid Data Capture Project. Presentation to the Entomology Collections Network. http://www.ecnweb.org/dev/files/12_Eastwood_2010.pdf [accessed 2.X.2011]

Morris PJ, Macklin JA (2006) Tools, Techniques, and Code for Supporting Image Databases of Natural History Collections Materials. Collection Forum 21(1-2): 203-222 [accessed 22.V.2012]

Munstermann L, Gall L (2010) Digitizing the Yale Collections—it takes a Village. Presentation to the Entomological Collections Network--Annual Meeting. San Diego, CA December 11, 2010. http://www.ecnweb.org/dev/files/gall-ecn-posted.pdf [accessed 2.X.2011]

National Science Foundation (2011) Advancing Digitization of Biological Collections (ADBC), Program Solicitation, NSF 11-567, p.4. http://www.nsf.gov/pubs/2011/nsf11567/nsf11567.pdf [accessed 1.X.2011]

Sumpter PM (1991) Curation of inverterbrate fossil collections at the Milwaukee Public Museum. Collection Forum 7(1):1–9.

Tagged Image File Format TIFF. http://partners.adobe.com/public/developer/tiff/index.html [accessed 29.X.2010]

Tann J, Flemons P (2008) Data capture of specimen labels using volunteers. Australian Museum. http://australianmuseum.net.au/Uploads/Documents/23183/Data%20Capture%20of%20specimen%20labels%20using%20volunteers%20-%20Tann%20and%20Flemons%202008.pdf [accessed 17.XI.2011]

Taylor H (2005) A photographer's viewpoint. In: Häuser CL, Steiner A, Holstine J, Scoble MJ (Eds) (2005) Digital imaging of biological type specimens. A manual of best practice. Results from a study of the Europoean Network for Biodiversity Information. Stuttgart, 126–152.

Tri-Trophic Thematic Collection Network TTD. http://sites.google.com/site/ttdtcn/

Virtual Biodiversity Research and Access Network for Taxonomy ViBRANT. http://vbrant.eu/

Wieczorek J, Guo Q, Hijmans R (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science 18(8): 745–767. http://herpnet.org/herpnet/documents/wieczorek.pdf, doi: 10.1080/13658810412331280211

ZXing (Zebra Crossing) http://code.google.com/p/zxing/ [accessed 25.I.2012]

Appendix 1

Questionnaire. (doi: 10.3897/zookeys.209.3135.app1) File format: Reach Text Format (rtf).

Explanation note: Questionnaire used for interviewing staff of visited collections.

Citation: Nelson G, Paul D, Riccardi G, Mast AR (2012) Five task clusters that enable efficient and effective digitization of biological collections. In: Vladimir Blagoderov (Ed) Mass digitization of natural history collections. ZooKeys 209: 19–45. doi: 10.3897/zookeys.209.3135.app1

Appendix 2

Web resources. (doi: 10.3897/zookeys.209.3135.app2) File format: Microsoft Word Document (docx).

Explanation note: Web resources mentioned in the paper.

Citation: Nelson G, Paul D, Riccardi G, Mast AR (2012) Five task clusters that enable efficient and effective digitization of biological collections. In: Vladimir Blagoderov (Ed) Mass digitization of natural history collections. ZooKeys 209: 19–45. doi: 10.3897/zookeys.209.3135.app2

ZooKeys 209: 47–54 (2012)
doi: 10.3897/zookeys.209.3179
www.zookeys.org

RESEARCH ARTICLE

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

OpenUp! Creating a cross-domain pipeline for natural history data

Walter G. Berendsohn[1], Anton Güntsch[1]

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

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

Academic editor: *Vladimir Blagoderov* | Received 2 April 2012 | Accepted 13 July 2012 | Published 20 July 2012

Citation: Berendsohn WG, Güntsch A (2012) OpenUp! Creating a cross-domain pipeline for natural history data. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 47–54. doi: 10.3897/zookeys.209.3179

Abstract

Multimedia data held by Natural History Museums and Universities are presently not readily accessible, even within the natural history community itself. The EU project OpenUp! is an effort to mobilise scientific biological multimedia resources and open them to a wider audience using the EUROPEANA data standards and portal. The connection between natural history and EUROPEANA is accomplished using well established BioCASe and GBIF technologies. This is complemented with a system for data quality control, data transformation and semantic enrichment. With this approach, OpenUp! will provide at least 1,1 Million multimedia objects to EUROPEANA by 2014. Its lean infrastructure is sustainable within the natural history community and will remain functional and effective in the post-project phase.

Keywords

OpenUp!, BioCASe, EUROPEANA, GBIF, Multimedia, ABCD, ESE, EDM, Biodiversity Informatics, Collections, Natural History

Introduction

The vast majority of global collections of biological organisms and images of organisms are held by institutions such as natural history museums and universities, in the realm of natural sciences. Nevertheless, nature is of course a major subject in the context of cultural history and humanities, and numerous cultural objects represent organisms

(Fig. 1). Both communities have started to digitise their objects and to publish the resulting multimedia data to make them accessible to a wider audience. The prevalent disjunction between them, however, has led to procedures, technologies, and data standards being optimized for the respective community's needs. The resulting incompatibilities prevent semantic linking and joined access.

In fact, there is a significant need for convenient joint access to the collection and multimedia holdings of different scientific communities. In the context of art history, for example, access to plant identifications provided by herbaria can be an important tool for the analysis of, e.g., ornaments in works of art. In turn, linking artwork with natural history specimens raises the general awareness of this important research tool and thus serves the museum community. And cultural background may be documented with natural history specimens; e.g. the collections during famous expeditions like those of Humboldt and Bonpland, and data on local uses recorded with the description of the collected organism.

EUROPEANA is the European portal to museums, libraries, archives, and audiovisual collections (Purday 2009). EUROPEANA has the potential to bridge the gulf between multimedia collections held by different communities by providing a common cross-domain user portal and web services based on unified metadata standards. During its first years of construction, EUROPEANA was clearly focused on cultural content,

Figure 1. Herbarium specimen Crocus vernus L. (© Botanic Garden and Botanical Museum Berlin-Dahlem, Germany) and Tapestry called Krokus by Britta Rendahl (1976) (© Upplandsmuseet, Uppsala, Sweden).

largely neglecting natural science objects. A series of biodiversity-related EU-projects such as STERNA (Sterna 2008), BHL-Europe (BHL-Europe 2009), Natural Europe (Natural Europe 2012), and OpenUp! (Berendsohn et al. 2011) widened EUROPEANA's scope to include natural history content. OpenUP! is the instrument for mobilizing and providing high volumes of biological multimedia collection objects for EUROPEANA. By end of the project (March 2014), OpenUp! will have delivered access to at least 1,1 Million objects and their corresponding data and metadata. More importantly, OpenUp! implements a sustainable pipeline from natural history collections to EUROPEANA (and potentially to other portals using the EUROPEANA standards). Recent initiatives to further digitisation of specimens (e.g. in the context of the industrial-scale e-RECOLNAT project in France, digitising all French herbarium specimens; or the NSF-funded iDigBio initiative in the US) will bring massive amounts of such objects on line. Using the OpenUp! approach, collection holders can publish their metadata and image locations, making them available to a wide audience beyond the natural history community. This pipeline scales up and will continue to function and provide access to the rapidly growing stock of multimedia content held by natural history institutions.

Of course we are fully aware of the problems of semantic mapping of metadata, especially with the taxonomic concepts represented by the name (e.g. Geoffroy and Berendsohn 2003). However, though this (as most of the retrievable information on the Internet) is not satisfying from a scientific view, we still posit that exposing natural history object information to a hugely enlarged audience (as offered by EUROPEANA) will help both the data providers as well as the users. The former will gain by the raised awareness of their holdings and by drawing attention to their cultural context, the latter will (in many cases for the first time in their life) become aware that such collections exist. And as a major side effect of mobilising the information for various networks simultaneously, researchers can choose to access the information through other interfaces that are less fuzzy in that respect (e.g. Güntsch et al. 2009).

The OpenUp! approach

OpenUp! creates an information flow from holders of collection multimedia data to the EUROPEANA data portal and services, but it avoids as much as possible the development and deployment of project-specific software modules. Rather, existing and well established protocols, standards, and software tools are used, resulting in an infrastructure that can be maintained with low maintenance costs beyond the funded project phase (Fig. 2).

OpenUp! data providers are usually connecting their existing collection management databases to the network. These databases are part of their institutional work flow so that maintenance and updating is part of the institutional setup. Connection is accomplished by equipping the local database with an installation of the BioCASe provider software package (Holetschek et al. 2009), and by mapping the local data definitions to the TDWG Biodiversity Information Standard "Access to Biological

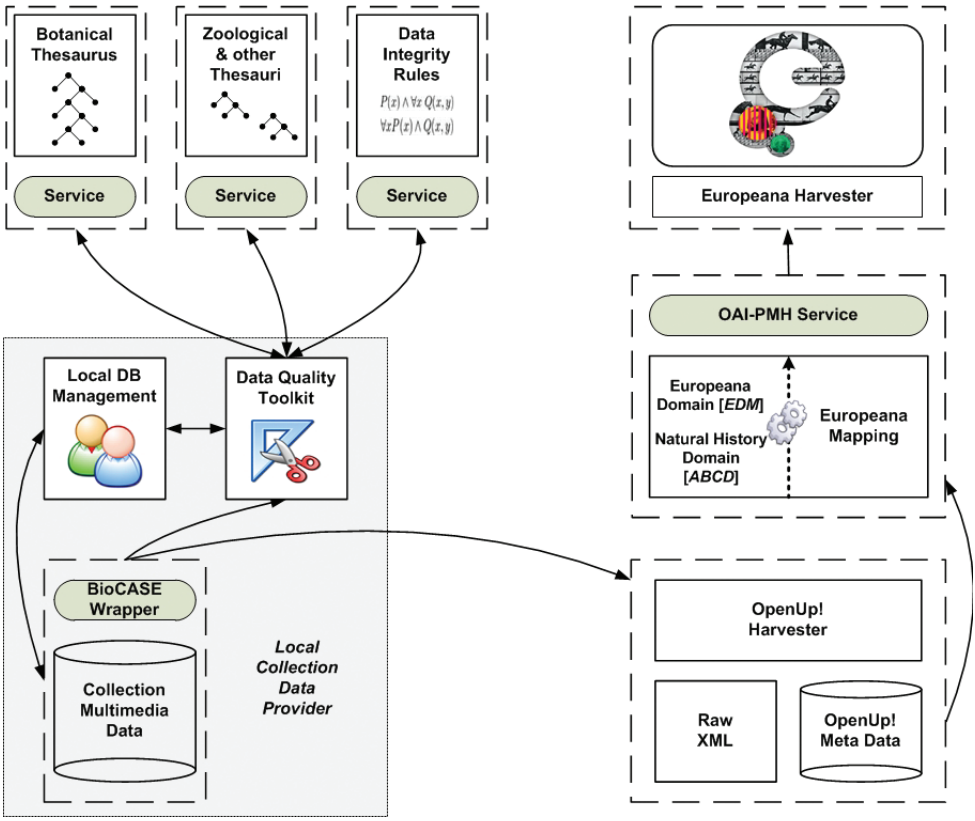


Figure 2. Information flow from a collection data provider via the central OpenUp! aggregator to the EUROPEANA harvester and portal. The collection database uses standard BioCASe/ABCD technology for connecting up to the network.

Collection Information" (ABCD, Berendsohn 2005). The software translates the local data to ABCD and allows querying the database over the Internet. The same installation is also used to provide data to the GBIF network. The only difference is that the configuration of the provider software for OpenUp! has to ensure that a minimal set of data elements required by the EUROPEANA portal is made available. The central OpenUp! aggregator notifies providers if this condition has not been met.

Harvesting of ABCD data and storage on the central aggregation server is performed using the GBIF Harvesting and Indexing Toolkit (HIT, GBIF 2011). The aggregator database stores only the textual data, including the URIs of the multimedia data. It is implemented using the same system that is used by the BHL Europe project. From there, the data from the ABCD standard used by the natural history domain are transformed into ESE (ESE 2011), which is used as a cross-domain metadata standard in EUROPEANA. The transformation is carried out using Pentaho Data Integration (aka Kettle, Pentaho 2011). The mapping between ABCD and ESE concepts is based on a thorough analysis of both standards, considering the semantics of natural history data elements used in a cross-domain context (Theeten et al. 2012).

OpenUp! metadata are periodically harvested by EUROPEANA via a single OAI-PMH access point at the aggregator database. Previews of multimedia objects for presentation and queries in the EUROPEANA portal are generated by EUROPEANA from full object URLs given in the metadata. The object itself and its presentation (e.g. using an image server or streaming software for audio files) stay with the provider, who also retains full rights of the multimedia file. The existence of the file is checked during the ABCD/ESE conversion process. Additionally, the central OpenUp! server will cyclically check the links to multimedia files and warn data providers if files become unavailable. In case of enduring problems, the links metadata will be excluded from the process.

Data Quality Control

Organising the basic information flow and data transformation process from biological multimedia collections to the EUROPEANA portal took considerable project resources. However, improving the content with regard to data quality and usability is the main item in the OpenUp! budget (which is co-funded by the European Union and the participants in the project). To support this process, some tools were implemented to support providers in the detection of data quality problems in their databases. Again, this "Data Quality Toolkit" mostly relies on existing systems and only a relatively lightweight interface layer is specific to OpenUp!

The OpenUp! Data Quality Toolkit (Fig. 3) operates directly on a given individual installation of the BioCASE provider software. It pages through a subset of ABCD records defined in its web-based user interface (OpenUp! 2012). Based on the user's choice of data quality rules to be applied, ABCD elements are then sent to an evolving set of data quality services analysing particular aspects of the data. This includes botanical and zoological name and concept checks for identifications, checks of compliance of ABCD elements to controlled vocabularies (e.g. country codes, mime types

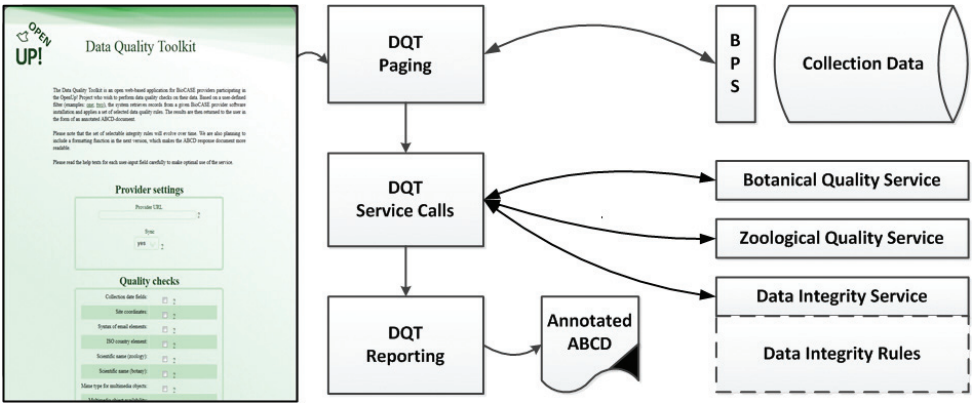


Figure 3. The OpenUp! Data Quality Toolkit

```
- <TaxonIdentified>
  - <HigherTaxa>
    - <HigherTaxon>
        <HigherTaxonName>PINACEAE</HigherTaxonName>
        <HigherTaxonRank>familia</HigherTaxonRank>
      </HigherTaxon>
    </HigherTaxa>
  - <ScientificName>
    - <FullScientificNameString>
        Abies apollinis Link
      - <!--
          <Annotations>
            <Annotation>
              <Context>OpenUP</Context>
              <ISODateTime>2012-03-26T09:21:33.873Z</ISODateTime>
              <MethodOrAgent>Kew concept reconciliation service (Botanical data quality service - concept check)</MethodOrAgent>
              <Type>Warning</Type>
              <Message>The name is a synonym.</Message>
              <Suggestion>Abies cephalonica Loudon</Suggestion>
            </Annotation>
          </Annotations>
        -->
      </FullScientificNameString>
    - <NameAtomised>
      - <Botanical>
          <GenusOrMonomial>Abies</GenusOrMonomial>
          <FirstEpithet>apollinis</FirstEpithet>
          <AuthorTeam>Link</AuthorTeam>
        </Botanical>
      </NameAtomised>
    </ScientificName>
```

Figure 4. OpenUp! Data Quality Toolkit annotation indicating that an identification is using a name which is a synonym (according to a concept reconciliation service provided by Kew Gardens).

for multimedia objects), and syntax of email-elements, dates and URLs. The toolkit then writes potential data problems as XML-encoded annotations directly into the ABCD-records they refer to and sends the compilation of all problem-records back to the user (Fig. 4). Users may also choose asynchronous access to avoid waiting periods. The tool provides suggestions to providers, which they may (or may not) take up in their OpenUp! quality enhancement task.

By decoupling the Data Quality Toolkit user interface layer from the underlying data quality services, the services themselves can be used in other contexts, and in turn, OpenUp! can integrate data quality services provided by other projects or initiatives. Collaborations have already started with the EU project BioVeL (Biodiversity Virtual e-Laboratory, BioVeL 2012) and the reBiND project funded by the German research foundation (Güntsch and Berendsohn in press).

Semantic Enrichment

The impact of the presentation of natural history specimens in a cross-domain context like EUROPEANA will partly depend on the possibilities for semantic linking with other content. Semantic linking is made possible by the metadata provided, so it can be enhanced by enriching the domain vocabularies used by the providers in the metadata. For example, in natural history databases typically the Latin scientific name is entirely sufficient (and indeed the most precise way) to denote the identification of the specimen. In contrast, content from the cultural domain will usually refer to an organism by

means of a common name. Users from that domain would not find the corresponding natural history object with their searches. Enhancing the natural history metadata by adding common names will close that gap.

In OpenUp! the botanical and zoological name services will be used to add synonym lists to the Latin names provided by the collection holders. A forthcoming OpenUp! service will be used for adding multilingual common names to the scientific names. In addition, external services will be used for adding further geographic information to the place names contained in the specimen data.

Outlook

During the first project year, OpenUp! has mobilised more than 220,000 natural history multimedia objects and made them available through EUROPEANA and GBIF, and the numbers are rapidly growing. Specimens displayed in the EUROPEANA portal demonstrate the feasibility of the principle data flows in OpenUp!. However, they also brought to light the weakness of the portal or in fact of the underlying ESE standard. Multimedia objects representing collection objects often have a strong relation to each other (e.g. several images from one specimen), which the portal does not adequately represent in its present stage. With the transition to the new metadata standard EDM (Europeana Data Model, Doerr et al. 2010) planned for 2012, nested object structures will be implemented. The millions of objects expected from the Natural History world will provide an ideal test bed for both metadata for linked objects and portal user interfaces and services providing searchable access to complex structured data.

References

Berendsohn WG (2005-) (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 26 Mar 2012]

Berendsohn WG (2011) OpenUp! Creating a cross-domain pipeline for biodiversity data. Abstracts. TDWG 2011 Annual Conference. https://mbgserv18.mobot.org/ocs/index.php/tdwg/2011/paper/view/96 [accessed 26 Mar 2012]

BHL-Europe (2009) Biodiversity Heritage Library Europe. http://www.bhl-europe.eu/ [accessed 26 Mar 2012]

BioVeL (2012) BioVeL – Biodiversity Virtual e-Laboratory. http://www.biovel.eu/ [accessed 26 Mar 2012]

Doerr M, Gradmann S, Hennicke S, Isaac A, Meghini C, Van de Sompel H (2010) The Europeana Data Model (EDM). World Library and Information Congress: 76th IFLA General Conference and Assembly. http://www.ibi.hu-berlin.de/forschung/publikationen/wissensmanagement/DoerrEtAl2010 [accessed 26 Mar 2012]

ESE (2011) Europeana: Europeana semantic Elements (ESE). http://www.europeana.eu/schemas/ese/ [accessed 26 Mar 2012]

GBIF (2011) The GBIF Harvesting and Indexing Toolkit (HIT). http://code.google.com/p/gbif-indexingtoolkit/ [accessed 26 Mar 2012]

Geoffroy M, Berendsohn WG (2003) The concept problem in taxonomy: importance, components, approaches. Schriftenreihe Vegetationsk 39: 5–14.

Güntsch A, Berendsohn WG (in press) A rescue strategy for threatened biodiversity data. In: Proceedings of PV2011 - Ensuring long-term preservation and adding value to scientific and technical data, Toulouse.

Güntsch A, Hoffmann N, Kelbert P, Berendsohn WG (2009) Effectively searching specimen and observation data with TOQE, the Thesaurus Optimized Query Expander. Biodiversity Informatics 6: 53–58. https://journals.ku.edu/index.php/jbi/article/view/1631/3472 [accessed 22 Jun 2012]

Holetschek J, Kelbert P, Müller A, Ciardelli P, Güntsch A, Berendsohn WG (2009) International Networking of Large Amounts of Primary Biodiversity Data. In: Fischer S, Maehle E, Reischuk R (Eds) INFORMATIK 2009, Im Focus das Leben. Lecture Notes in Informatics (LNI) 154: 23; 552–564.

Natural Europe (2012) Natural Europe. http://www.natural-europe.eu/ [accessed 26 Mar 2012]

OpenUp! (2012) The ABCD Data Quality Toolkit. http://services.bgbm.org/DataQualityToolkit/ [accessed 29 Mar 2012]

Pentaho (2011) Pentaho Kettle Project. http://kettle.pentaho.com/ [accessed 26 Mar 2012]

Purday J (2009) Think culture: Europeana.eu from concept to construction, The Electronic Library 27(6): 919–937. http://www.emeraldinsight.com/journals.htm?articleid=1827227 [accessed 22 Jun 2012]

Sterna (2008) About Sterna. http://www.sterna-net.eu/index.php/en/about [accessed 26 Mar 2012]

Theeten F, Roca Ristol P, Jacob B, Jõgeva J (2012) OpenUp! Guidelines for users and content providers v. 1. http://open-up.cybertaxonomy.africamuseum.be/file/722/download/739 [accessed 26 Mar 2012]

ZooKeys 209: 55–73 (2012)
doi: 10.3897/zookeys.209.3205
www.zookeys.org

RESEARCH ARTICLE

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

The US Virtual Herbarium: working with individual herbaria to build a national resource

Mary E. Barkworth[1], Zack E. Murrell[2]

1 *Intermountain Herbarium, Department of Biology, Utah State University, 5305 Old Main Hill, Logan, Utah, USA 84322-5305* **2** *Biology Department, Appalachian State University ASU Box 32027, Rankin Science Building, 527 Rivers Street, Boone, North Carolina, USA 28608*

Corresponding author: *Mary E. Barkworth* (mary.barkworth@usu.edu)

Academic editor: *V. Blagoderov* | Received 7 April 2012 | Accepted 25 June 2012 | Published 20 July 2012

Citation: Barkworth ME, Murrell ZE (2012) The US Virtual Herbarium: working with individual herbaria to build a national resource. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 55–73. doi: 10.3897/zookeys.209.3205

Abstract

The goal of the US Virtual Herbarium (USVH) project is to digitize (database, image, georeference) *all* specimens in all US herbaria, enabling them to be made available through a single portal. Herbaria house specimens of plants, fungi, and algae, so USVH will offer a rich portrait of biodiversity in the US and in the other countries represented in US herbaria. Equally importantly, working towards this goal will engage people with herbaria and the organisms they house, expanding their appreciation of both the power of biodiversity informatics and the demands that it places on data providers while developing improved communication among those working in and with herbaria. The project is not funded but has strong support among those working in herbaria. It works through regional herbarium networks, some of which existed prior to the USVH project, while others are still in gestation. It differs from most digitization projects in its emphasis on helping those involved with herbaria become part of a national enterprise, an aspect that is seen as critical to creating the resources needed to develop and sustain the project. In this paper, we present some of the lessons we have learned and the difficulties we have encountered during the first few years of the project.

Keywords

Herbaria, networks, plants, fungi, algae, digitization, online databases

Origin of the US Virtual Herbarium project

The US Virtual Herbarium project was started in 2008 at a meeting held in conjunction with the annual meeting of the Botanical Society of America. Those present were asked whether they were in favor of attempting to develop integrated access to specimen information residing in all US herbaria, creating in essence, a US Virtual Herbarium (USVH). The meeting followed 20+ years of digitizing efforts (primarily databasing) within US herbaria. It had been called because, despite these efforts, there was no evidence of a program to build a national resource that would include all herbaria. Some of those voting had been involved in digitization efforts. Others came looking for help, both financial and technical, in starting the process. At the end of the meeting, all those present endorsed the concept. Thus the project started, not in direct response to a national initiative or program but as a statement of interest by those directly involved with herbaria.

The meeting was held under the auspices of the Western Association of Agricultural Experiment Station Directors (WAAESD). Each state has an Agricultural Experiment Station (AES) and their directors work together, regionally and nationally, in areas of joint interest. Although it was AES directors in the western states who sponsored the meeting, the USVH project has always been national in scope. Formally speaking, the purpose of the meeting was to determine whether there was sufficient support to justify WAAESD sponsorship of a 5-year committee to coordinate work towards a single access point to information from all US herbaria. Given the support expressed, formation of the committee was approved.

WAAESD sponsorship provides a formal but flexible structure within which to operate. It does not provide funding; it does provide freedom in determining how best to pursue a group's objectives. It also provides a mechanism for disseminating information through the National Information Management and Support System (NIMSS). Reports and announcements posted to NIMSS are sent to AES directors in each state as well as to registered participants. Because most herbaria are not connected with AES, the sponsorship by WAAESD immediately increased awareness of herbaria.

The executive committee's first task was to develop explicit goals for the project. After considerable debate, it agreed that the overall goal of the US Virtual Herbarium project should be digitizing all specimens in all US herbaria. The result will be a major new scientific resource but the greatest benefits will result from working towards this overall goal, a process that will require helping collectors and curators record information in a manner that maximizes the value of a specimen, use the tools being developed for capturing and sharing collection information, and make use of the resulting information in their research, education, and outreach activities. It will also require increasing interaction among those who work in herbaria and educating users in diverse disciplines about the value and use of collection data. Much of the value of the project lies in ensuring that these benefits are experienced by all those involved with herbaria and in teaching students about algal, fungal, and plant diversity.

Herbarium specimens provide a particularly rich information layer to the world's biodiversity resources because they represent sessile organisms. They show the ability of

a taxon to complete its life cycle at a particular location and time and, in some instances, provide information about the prevailing growing condition (see, e.g., Woodward and Bazzaz 1988; Kouwenberg et al. 2003; Zangeri and Berenbaum 2005; Johnson 2011). Thus the value of the digital herbarium layer is clear. The optimal path (or paths) to providing it is less clear. The task of the US Virtual Herbarium project is to accelerate the process and ensure that all herbaria become involved because in that way more individuals will learn about the organisms present in herbaria, what digitization involves, and the power of biodiversity informatics. It will also result in a more dense information layer. The project does not focus on developing better ways to digitize herbaria; that is the focus of specific programs within the National Science Foundation and Institute for Museum and Library Services. Instead, the project aims to foster the collaborations needed to establish networks and enable rapid dissemination of better procedures as they become available. In this paper, we share some of the lessons we have learned in reaching the current level of digitization in the US.

Herbaria in the US

There are 729 registered herbaria in the US (Thiers et al. 2012+). They are scattered throughout the country but are more abundant in densely populated states (Fig. 1). Seventeen herbaria have a million or more specimens each; about 300 have fewer than 17,000 specimens. About 150 of the US herbaria listed in Thiers (2012+) have been transferred or closed; there are also many herbaria not listed by Thiers (2012+), most of which have fewer than 10,000 specimens. Our current estimate is that there are about 800 active herbaria and over 90 million herbarium specimens in the US.

About 78% of US herbaria are owned by an academic institution. Academic herbaria, particularly those in smaller institutions, offer excellent opportunities for involving students. Countering this potential is the fact that small herbaria often receive little or no formal support from their institution and may not be actively curated. Of the remaining herbaria, about 13% are owned by a government entity, usually federal but in some cases state, county, or municipal. About 9% are associated with botanical gardens or independent museums; among these are eight of the herbaria with a million or more specimens.

In 2009, Thiers provided Barkworth with a list of US herbaria registered with *Index herbariorum* at that time. Of these, 601 appeared to be active. "Appeared to be" because there is no guarantee that Thiers is notified when a herbarium is closed or transferred. In 2010 a survey (via paper questionnaire, with reminders by email or telephone call to some non-respondents; see Appendix 1) of all 601 herbaria resulted in 287 responses (Barkworth 2011, unpubl. data). The data revealed that many of the smaller, non-responding herbaria had been transferred or closed. Of the responding herbaria, 154 (54%) had a herbarium database and 70 (24%) were imaging their specimens. Collectively, the 287 herbaria held 50,583,000 specimens, of which 16,880,000 (33%) had been databased and 1,510,000 (3%) imaged. Most of the databasing herbaria (150/154) made specimen information available on the web through their own

web site; 39 did so through a regional website; 38 made their records available to the Global Biodiversity Information Facility. These data indicate strong commitment to digitization and data sharing among US herbaria.

In addition to there being many herbaria in the US, there are many different taxonomic opinions, particularly with respect to vascular plants. These are reflected in state and regional floras. There are resources to help interpret the resulting complexity, e.g., *Flora of North America* (FNA; Flora of North America Editorial Committee 1993-present), which is developing a single taxonomic treatment for all bryophytes and vascular plants in North America north of Mexico. These are not always accepted but Tropicos (http://www.tropicos.org/, see the list of relevant websites in Appendix 2) shows how different floristic treatments have treated a particular name. *Index fungorum* (http://www.indexfungorum.org/names/names.asp), and Algaebase (http://www.algaebase.org/) are internationally respected indices to fungal and algal names, respectively. The US Virtual Herbarium project accepts that records in different herbaria may reflect multiple taxonomic concepts, a reality that can only partially be accommodated by alternative tables of synonyms. There are undoubtedly instances where this creates problems, for instance, when interpreting the distribution of a taxon that is sometimes interpreted narrowly, sometimes broadly, but such situations are probably less common than problems caused by misidentifications.

Table 1 shows the current status of herbarium digitization in the US from a network perspective. The six existing regional networks involve about 200 herbaria, rang-

Table 1. Overview of US regional and taxonomic herbarium networks. The Southwest and Intermountain Regions share a database but have different portals. "Herbaria" indicates the number of herbaria currently providing information to the network; numbers in parentheses are for extra-regional herbaria. Records are text-based records. Geo: percentage of georeferenced records. Most data obtained from web sites or node managers, March 31, 2012

Network	URL	Taxonomic scope; Location of source herbaria	Herbaria	Records
Existing networks				
California herbaria (CA)	http://ucjeps.berkeley.edu/consortium	Vascular plants; California	20 (1)	1,454,000
Pacific Northwest Herbaria (PNW)	http://www.pnwherbaria.org	US: Alaska to Oregon + Idaho and Montana. CANADA: British Columbia, Yukon.	57	1,763,040 (174,160 images)
Southwest (SEINet) and Intermountain (IRHN)	http://swbiodiversity.org/seinet/index.php http://intermountainbiota.org/portal/index.php (Shared database; different portals)	US: Southern California east to New Mexico, north to Nevada, Idaho, and Colorado MEXICO: Baja California, Sonora; Vascular plants.	32 (2)	2,069,025 (67% Geo)
Pacific Islands (CPH)	http://www.herbarium.hawaii.edu/cph/index.html	Hawai'i and the Pacific basin [Currently 3 of 15 herbaria connected] Vascular plants.	15	60,000

Network	URL	Taxonomic scope; Location of source herbaria	Herbaria	Records
Existing networks				
Northeast (CNH)	http://neherbaria.org/CNH	US: north and east from Pennsylvania CANADA: Ontario eastward; All taxa.	58	409,883
Southeast (SERNEC)	http://www.herbarium.unc.edu/seflora/firstviewer.htm	From Eastern Texas to Virginia to the Atlantic and Gulf Coasts; All taxa.	14	140,000
Wisconsin Flora	http://www.botany.wisc.edu/wisflora/	Wisconsin; Vascular plants, lichens	8	370,000
Alabama Plant Atlas	http://www.floraofalabama.org	Alabama; Vascular plants	9	78,000
Bryophytes	http://symbiota.org/bryophytes/index.php	North America; Bryophytes.	10	922,047 (38% Geo)
Lichens	http://symbiota.org/nalichens/index.php	North America; Lichens.	16 (1)	627,756 (55% Geo)
Macrofungi	http://mycoportal.org/portal/index.php	North America; Macrofungi	5	154,526 (13% Geo)
American Myrtaceae	http://cotram.org	Myrtaceae in the Americas	4	64158 (63%)

ing from small, unlisted herbaria to the largest herbaria in the country. Some herbaria contribute to multiple portals. The number of records available is over 7,665,000. This count does not differentiate between those that are fully databased, imaged, and georeferenced and those that have minimal information, possibly only the image of a label. Progress in the different aspects is hard to assess. Only the Pacific Northwest Herbaria (PNW) portal shows the number of specimen images available and only Symbiota portals show how many records have georeference data. Many georeferenced records do not include uncertainty estimates. The California, Pacific, and Pacific Northwest networks use software developed within each region; the portal for the southeastern US uses a mixture of software; the others use Symbiota (http://symbiota.org).

Lessons learned

- Commitment, energy, time, resources, and funding are the most critical needs of the USVH project. Of these, time is usually the most scarce resource, particularly in smaller herbaria in which a single individual has to fulfill many different functions. It can, of course, be alleviated to some extent by funding but digitization will require a time commitment on the part of the person or persons responsible for a herbarium. Funding for other resources is also needed but much can be done with minimal financial support now that effective software and work flows have been developed, particularly if hardware is shared.

- The range in size of US herbaria (from less than 1000 to over 8 million) and their diverse roles is matched by the diversity of their resources and goals. Many have little or no IT support and little or no budget; others, even some smaller herbaria, have strong IT support, significant endowments, and substantial volunteer support. Goals range from research on a global level to being a reference collection for training of seasonal employees.
- Curators have diverse backgrounds. Most, particularly in mid-sized to large herbaria, are professionally trained taxonomists with memberships in professional societies such as the Botanical Society of America and the American Association of Plant taxonomists. Others have backgrounds that range from ecology to paleobotany, with their professional associations being equally diverse. This presents a challenge to developing an effective information flow among all herbaria. Regional collaborations on multiple scales are effective in addressing this challenge but require a leader with time to commit to the task.
- There is no best approach for digitizing herbaria; there are multiple effective approaches. The needs and resources of large research herbaria with multiple type specimens and collections from many countries and multiple centuries differ from those of small herbaria serving a forest district or a teaching institution. In working with those in charge of herbaria, one must recognize and respect their differing priorities and resources. Adopting theoretically suboptimal procedures for digitization may be the best procedure if the resources needed for adopting a better procedure are not available.
- Broadening participation requires minimizing barriers while maximizing benefits. Symbiota (http://symbiota.org/tiki/tiki-index.php), open source software available through SourceForge (http://sourceforge.net), accomplishes this by enabling direct data entry into the central database, providing tools for preparing labels, and integrating images of living organisms into checklists, species pages, and flash card quizzes. In August 2011, Barkworth switched the Intermountain Herbarium (258,000 specimens) to databasing directly into the regional database (SEINet/IRHN) which uses Symbiota. It was so easy to use that she persuaded two colleagues, Gordillo and Anderson, each of whom is responsible for a small herbarium (6000 and 4000 specimens, respectively), to employ it to bring their herbaria into the network. The financial cost for the two was less than $400 each, the cost of preprinted barcode labels and a barcode scanner. Data entry is being done by volunteers. Of equal importance, students introduced to the program and associated portal immediately see value in the resources provided. Once imaging equipment is available, the two herbaria will adopt procedures that exploit the advantages images offer but, in the meantime, their students are learning to record better information and their institutions can boast about contributing to a major resource.
- It does not matter whether a herbarium starts with imaging or databasing. The important thing is to start. Specimen records that consist only of text-based information can be used for generating checklists, georeferencing, and searching.

Specimen records that consist only of an image are of little value until the label information is databased but imaging can accelerate databasing and enable offsite-databasing. Establishing both of these, however, requires infrastructure development, both technical and human.

- Remote data entry and incorporation of optical character technology into the data entry process can speed up data entry but it requires access to images which, in turn, requires access to appropriate equipment. Legler (2011) has designed equipment that has been widely adopted because it is effective, easily transported, and does not take much space. The problem is that the initial cost (about $6000) is large compared to the budgets of most herbaria. Once purchased, it can be shared among neighboring herbaria, a process that also fosters the kind of social network needed to disseminate information.
- Integrating optical character recognition (OCR) technology into data entry tools will accelerate data entry for the very large number of specimens with clean, typewritten or computer generated labels but entries need to be reviewed before being accepted. Major obstacles to widespread adoption of OCR-assisted data capture are a) lack of imaging equipment and b) the need to incorporate OCR-assistance into the data entry module of the various database systems used in herbaria, a process that is underway. For interpreting hand-written or unclear labels, OCR is less effective than humans.
- Automated georeferencing tools, such as Geolocate (http://www.museum.tulane.edu/geolocate/) greatly accelerate georeferencing and can provide an estimate of uncertainty but, as with OCR data entry, the results, both for the locality and the uncertainty, need to be reviewed. At present, most programs for sharing information can only store point-radius uncertainties, not a polygon. This limits their value because plant collectors often collect along a trail. Another potential problem is that all values are calculated based on current geographic information. Even with such limitations, georeferencing is valuable. Applied to the thousands of specimens in herbaria, it enables patterns to be detected even if some of the individual locations are fuzzy. Those using the data should be aware of the inherent problems, grateful for the amount of data being provided, and willing to assist in improving its quality.
- Batch georeferencing, in which multiple specimens with the same locality information are georeferenced simultaneously, greatly accelerates georeferencing. The acceleration is greatest if records from multiple herbaria can be georeferenced simultaneously. Technological impediments to effective batch georeferencing include the absence of a mechanism for sharing specimen records among networks and the need for tools that "repatriate" the georeferencing information back to the specimen records. The human impediments include lack of knowledge as to how to georeference specimens and/or use the tools available for assisting in the task, impediments that can be overcome by workshops and online tutorials. Another impediment is the need for effective management of such collaborations.

- Enabling collectors to enter their collection information directly into a database that can both generate labels and provide data to the databases of recipient herbaria should be given high priority. Ideally, such programs should make it possible to enter information whether offline or online and for multiple taxonomic groups because individuals frequently collect more than one kind of organism. If data are entered offline, it should be possible to clean them when the connection is restored. (see, e.g., Atrium http://www.atrium-biodiversity.org/about.html). Label-making modules should also enable students to use the module while taking a class without the data being displayed so that they learn to record and store data in a manner that maximizes its utility.

  Label generating tools will not help digitize the specimens currently in herbaria but early adoption of database-driven label production combined with aggressive pursuit of funding opportunities enabled the herbaria of the University of Wyoming and the Missouri Botanical Garden (1.4 and 6.3 million specimens, respectively) to have over 50% of their collections databased by the time of the survey. The only other large US herbarium to have more than 50% of its 950,000 specimens databased is the National Fungus Collection which has 89% of its collections databased, a noteworthy accomplishment.

- Regional collaborations are the most effective method of spreading digitization. They make it easier to share imaging equipment and develop the localized resources (e.g., checklists, identification tools) that give immediate, easily recognized value to regional portals. They also make establishing personal relationships among data providers easier, relationships that subsequently become effective social networks for sharing ideas and information. Development of regional networks is also critical to building the long term, broadly based support required to create and sustain a truly national herbarium network, one that involves all herbaria.
- The map (Fig. 1) shows the major regional networks but there are many smaller digitization networks in existence, some of which were initiated with federal funding, others with state or private funding. They have been critical to bringing the digitization of US herbaria to its present status. These smaller networks generally make their records available through their own web site. One of the challenges facing the US Virtual Herbarium project is to enable such networks to share their specimen information more widely. Other challenges include establishing networks for all parts of the country and persuading herbaria with their own web site to share their records on a regular basis with a regional network.
- There is often a lag time between agreeing to establish a network and actually having a network that people can use. Herbaria with their own specimen databases need to develop scripts for exporting their data to the network database and ensuring that new and modified records are exported at regular intervals. Constructing and testing these scripts takes time. It may also be found that

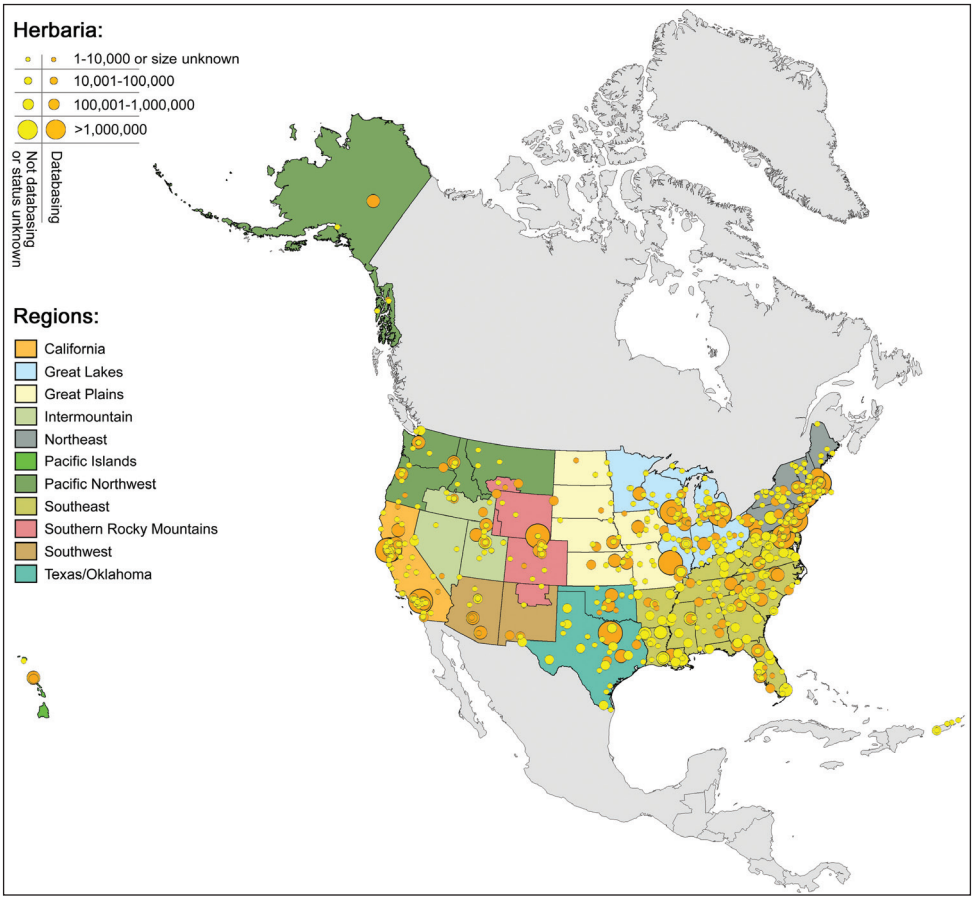


Figure 1. Regional networks and herbaria in the U.S.A. Network boundaries are guides; herbaria are free to join the network of their choice. Some herbaria contribute records to more than one network. No network has been established as yet for the Great Plains, Great Lakes, and Southern Rocky Mountain Regions. Data obtained June, 2011.

the existing data has to be cleaned up before being exported. Another source of delays can come from establishing formal memoranda of understanding. Delays are greatest if the herbaria are located in different countries or belong to a private institution. Some networks operate without formal memoranda.

- There is a need for the single, all-embracing network that is being established by iDigBio (see below). At present, herbaria with specimens from different taxonomic groups need to send their data to multiple networks (there are separate networks for bryophytes, lichens, and macrofungi). Moreover, at present regional nodes only provide access to specimens from herbaria within their region, e.g., data for specimens from the northeastern US residing in herbaria of the intermountain region are not currently made available to the northeastern network. It also means that users wishing to examine all biodiversity within a

region have to go to multiple networks to obtain the information they seek and each network. To maximize the value of a truly integrated network, however, its data must be readily accessible and easily queried not just by biodiversity informatics specialists but also by the general public and educators at all levels and in many different disciplines because it is, ultimately, these people whose support will be required to sustain the network's maintenance and development.

Interaction with iDigBio and BISON

In February 2010, an NSF-funded workshop brought together individuals with knowledge in different aspects of digitization to discuss how best to develop a national herbarium network. Several useful discussions and contacts resulted from the workshop but that fall the NSF announced its Advancing Digitization of Biological Collections (ADBC) Program. ADBC projects fall into two categories, creation of "a permanent database of digitized information from all biological collections in the U.S. (https://www.iDigBio.org/content/about-iDigBio)", the iDigBio project, and Thematic Collection Networks (TCNs) that focus on "major scientific questions" (http://www.nsf.gov/news/news_summ.jsp?cntn_id=121015). At about the same time it was announced that what is now the Biological Informatics Program of the US Geological Service had begun development of an integrated and permanent resource for biological occurrence data from the United States, Biodiversity Information Serving Our Nation (BISON). This will integrate records for the US from the Global Biodiversity Information Facility and those made available via iDigBio with multiple geographic environmental layers, thereby enabling sophisticated and complex analyses.

These two developments forced us to rethink how the US Virtual Herbarium project could best achieve its objectives, assuming they were still valuable, while complementing the work of ADBC-funded projects. The goals of the US Virtual Herbarium project are similar to those of iDigBio apart from its sole focus on herbaria, but it has a somewhat different emphasis. For iDigBio, extending participation to all collections in the US, both large and small, is a third phase, while for USVH, it is the priority. A recent analysis of the botanical capacity of the US (Kramer et al. 2010), demonstrated that the country has far fewer students entering the botanical sciences than are needed to address the major scientific questions of today. We see developing regional networks, and ultimately a national herbarium network, as one mechanism for increasing interest among such students while building an invaluable research resource. As such, it is too important to delay. We recognize that, as technology develops, new standards will be developed and new technologies become available; that is the nature of technology. The USVH organization can provide an effective conduit for rapidly sharing the benefits of such developments among all herbaria.

The BISON project should provide the access to herbarium records and tools for working with them that were part of the original vision for the US Virtual Herbarium project, at least so far as the US is concerned. It is, however, dependent on

the quantity and quality of records made available to it. The USVH project's primary focus is on helping herbaria both provide the needed records and ensure that are of the quality standards needed for use in environmental analyses. In doing so, the project will expand the number of individuals who understand the concepts involved and enable interested individuals to obtain data as it becomes available. Moreover, making information available now has resulted in the herbaria involved receiving feedback concerning some of their specimens, feedback that comes from knowledgeable individuals and will, ultimately, benefit BISON.

Future directions

Much has been learned about building a herbarium data layer in the US but the majority of herbaria are still not contributing to its development. There are some herbaria that, although digitizing their specimens, do not make the resulting resources available other than on their own network and some that have not started any part of the digitization process. In the latter cases, the problem may be that the herbarium forms a very small part of the responsibilities of the person in charge, or that the person in charge does not know how to start, or that he or she simply does not have the time. Personal contact is often a key step to bringing isolated herbaria into a network. When making such contacts, the benefits that will accrue from membership in a network need to be presented in terms that are relevant to the mission of the herbarium concerned and the person or persons running it. These benefits should, to the maximum extent possible, be immediate and direct. The greatest benefit, without question, is funding but software developments combined with the ability to share resources with and tap into the knowledge of those already in a network have substantially reduced the amount of funding required.

The benefits to medium-sized and smaller herbaria of participating in a regional herbarium include greater publicity, the ability to show how their specimens contribute to overall knowledge, and a mechanism for identifying where to focus future collecting efforts, all of which help validate their worth to institutional administrators. It provides students at academic herbaria an opportunity to participate in a regional and national informatics enterprise while improving the currency of their education. In addition, it helps build professional relationships among individuals who, because of disparate interests and obligations, might not normally connect with each other. Other benefits depend on the resources made available at the network level. These need to benefit a wide range of individuals because it is by offering such benefits that herbaria, and collections in general, earn public support. Such tools can range from quizzes about plants in a grocery store, to games where participants score points for being able to identify plants from images.

Investment in medium-sized and smaller herbaria can have major impacts on the botanical sciences in the US. These herbaria, their associated curatorial staff and users often provide the experiences that steer students towards the botanical sciences.

This is important because a disproportionate number of graduate students come from such institutions. Research intensive universities, state and federal agencies, and non-government organizations are dependent upon these "feeder institutions" to provide a flow of graduate students and professional botanists.

All larger herbaria are digitizing their collections, usually maintaining their own database and web site in addition to participating in one or more networks. If, as is the case in several large herbaria, much of their current research and collection activity lies outside the US, these activities may be most appreciated outside the US but they are essential to attainment of the US Virtual Herbarium's overall goal, digitization of all specimens in all US herbaria. Large herbaria can benefit from joining a network by becoming the "go-to" herbarium for web-related resources. They are also usually better positioned to attract funding for positions to support a regional network. In addition, contributing records to the region where they are located helps them demonstrate that they are "good neighbors" which may assist them in obtaining benefits from the jurisdiction in which they lie.

An area that still needs improvement is building the bridges needed for sharing ideas, information, and concepts between those directly responsible for herbaria and those with specialized knowledge in areas relating to digitization and use of the flood of information it is providing. There are many such areas: biodiversity informatics, information technology, computer science, geography, and education. Working with specialists in these areas will develop a richness and synergy that benefits all involved. The US Virtual Herbarium project can help extend the benefits of such interactions throughout the herbarium community. Among these benefits are increased efficiency in herbarium management which will, ultimately, free up the time of those involved for research and educational activities. Developing these interactions requires that all involved respect each other's different backgrounds, obligations, interests, and knowledge.

What of the immediate future? There are several steps that the USVH project plans to take. Regional consortia or networks are extremely beneficial in helping move multiple herbaria forward, but some parts of the country have, as yet, no effective network. One of our immediate targets is to facilitate linking all herbaria to a regional network. This can be accomplished either by expanding the region covered by an existing network, possibly with separate portals for subregions (e.g., SEINet and IRHN), or by creating new networks. Both scenarios will require acquisition of additional server space and support personnel.

Georeferencing vastly increases the value of collection records and enables searches across space which may be more relevant to some research questions than searches across taxa (Johnson et al. 2011). It is an aspect that greatly benefits from collaboration but also helps build the social infrastructure needed for effective collaboration (Constable et al. 2010). US herbaria have not, as yet, implemented collaborative georeferencing although some herbaria have georeferenced a substantial portion of their specimens. In many instances, however, this may mean only that there is a latitude and longitude associated with the record. Such limited data make it possible to obtain a picture of the overall distribution of a taxon but do not satisfy the needs of those engaged in environmental analyses (Chapman and Wirczorek 2006).

Data cleaning is another aspect that has, as yet, received surprisingly little attention from herbarium networks. The primary reason may be that the focus is on obtaining records and engaging herbaria, but there are now enough records in each network that building mechanisms for routinely identifying problems is highly desirable. These should be run at the herbarium level with cleaning at the regional level being a second line of defense. The need is for tools that check that georeference and elevation data are at least consistent with the lowest political unit used (usually county for the US, often state for other countries). The scientific name used must also be checked for accuracy because some herbaria may have recorded data in databases (or spreadsheets) without verifying that the names entered were valid. Another check, one that is probably best combined with georeferencing, is for the spelling of place names. Some will be found to be phonetic renditions (Chian for Cheyenne); others are merely misspellings.

Crowd-sourcing of data capture is already being explored in the US and elsewhere. What is not clear yet is how many volunteers can be found to take a short, online training session and then enter data for herbarium specimens online nor whether it is best to focus on identifying and capturing critical data, leaving capture of the remaining data to a later stage, or whether to try and capture all data at once. As with so many other decisions, there are pros and cons to both approaches. It is important, however, that we are transparent in reporting our accomplishments. Capturing a few fields from a million labels is not the same as capturing all label information from a million records.

Taxpayer funds, whether federal, state, or local, will not cover the cost of digitizing herbaria and maintaining herbarium networks. We must aggressively pursue other funding opportunities, including some that most of us involved with herbaria do not normally approach, such as wealthy individuals with an interest in the environment and stores that sell equipment and clothing to people who enjoy hiking. "We" in this case involves all in charge of herbaria but the approach each person takes has to reflect their abilities and interests and as well of those of the herbarium for which they are responsible. It should also complement their other responsibilities (and conform to their institution's guidelines). The US Virtual Herbarium project can help by disseminating information about successful approaches, developing templates, and seeking funds that will benefit multiple herbaria or networks.

Requests for financial support are more likely to be well received if it can be demonstrated that they will result in a product that benefits many user groups. To encourage use of the information available through existing herbarium networks, we need to work with K-12 educators to develop units that make use of network associated information while meeting state and national science standards. We must also work with state native plant societies, recognizing their value and asking their assistance in promoting use of our networks and their further development. We also need to make sure that government employees are aware of the information being made available, emphasizing its value in their work and to their constituents. And in all these interactions, we must not forget to ask what would make the resources we are developing more useful.

In addition to seeking funding from new sources, all those involved in herbaria must keep looking at work flows to see if they can be made more efficient. Sometimes simple changes, such as using preprinted barcodes to put a catalog number on a specimen rather than using a stamping machine, can save considerable time, time that can used for other purposes. Another possible change is to enable and expect those who borrow specimens to enter their information into the owner's database or into a regional database from which the owning herbarium could import the records and images. Since almost anyone borrowing specimens nowadays enters information from them into a database, this would require little additional work for the borrower but would greatly aid the loaning institution.

Sustaining the networks also requires maintaining the integrity of the data over time. The costs of doing so are non-trivial because, as Rosenthal (2011) pointed out, "digital data do not tolerate benign neglect". The specimens themselves are much more resilient in this regard. Moreover, each herbarium, even those that enter data directly into a network database, should maintain a copy of their data. This has the added advantage of ensuring that there are two copies in different locations. Another approach would be for neighboring regions to mirror each other's resources. This would increase the server space required by individual regions but in a manner that would be mutually beneficial. Eventually this task will, presumably, fall to iDigBio and BISON but, for now, herbaria and herbarium networks must adopt alternative approaches.

Conclusions

The number and distribution of herbaria in the US, together with the number of specimens they house, make them a prime resource for research in many different disciplines. Providing access to their information will enable sophisticated analyses at levels of scale, scope and accuracy that are unparalleled in the life sciences. It can also be used to introduce and encourage a fascination with plants, fungi, and algae by students at all levels in ways that incorporate inquiry. Digitizing herbaria will also enable those who work in herbaria more opportunities to study the organisms they love, and their interactions, by increasing the ease with which diverse user groups can access herbarium-based information without assistance from herbarium personnel.

The impediments to achieving the goal of the US Virtual Herbarium project, digitizing all specimens in all US herbaria, are resource-based, but they can be offset by focusing on the human factor. The project is dedicated to unlocking the vast resource represented by herbarium specimens by assisting in development of the human and knowledge infrastructure needed. It is accomplishing this task by linking people, ideas and tools into an integrated whole. Much of this involves extending the tools, knowledge, and resources developed by funded projects to more herbaria by establishing connections among people with the varied skills and interests needed, thereby building an integrated community of people working towards a common goal.

Note added in proof: Results of the 2012 herbarium survey are being posted to http://herbarium.usu.edu/SurveyResults.html. It included a question about georeferencing and asked for more details on network connections (see Appendix 3).

Acknowledgements

We thank all the herbarium curators who responded to the survey and Ben Legler for preparing Figure 1. Barkworth thanks Edward Gilbert and Curtis Dyreson for many discussions relating to all aspects of digitization and Lynette Harris, Kira Call, Thomas Phelps, and Rakelle Sanchez for their comments on an early version of the manuscript. We both thank the journal's reviewers and editors for their comments and suggestions. We hope they feel, as we do, that their efforts have resulted in a better paper. This paper is approved as journal paper #8420 of Utah Agricultural Experiment Station.

References

Flora of North America Editorial Committee (1993-present) Flora of North America north of Mexico. Oxford University Press, New York.

Chapman AD, Wieczorek J (Eds) (2006) Biogeomancer, Guide to best practices for georeferencing. Global Biodiversity Information Facility. http://www.gbif.org/orc/?doc_id=1288

Constable H, Guralnick R, Wieczorek J, Spencer C, Townsend Peterson A, The Vertnet Steering Committee (2010) Vertnet: A new model for biodiversity data sharing. PLoS Biology 8: e1000309. doi: 10.1371/journal.pbio.1000309

Johnson JP (2011) Marauding Moths. The Scientist. http://the-scientist.com/2011/10/01/marauding-moths/

Johnson KG, Brooks SJ, Fenberg PB, Glover AG, James KE, Lister AM, Michel E, Spencer M, Todd JA, Valsami-Jones E, Young JR, Stewart JR (2011) Climate change and biosphere response: Unlocking the collections vault. Bioscience 61: 147–153. doi: 10.1525/bio.2011.61.2.10

Kouwenberg LR, McElwain JC, Kürschner WW, Wagner F, Beerling DJ, Mayle FE, and Visscher H (2003) Stomatal frequency adjustment of four conifer species to historical changes in atmospheric CO_2. American Journal of Botany 90: 610–619. doi: 10.3732/ajb.90.4.610

Kramer AT, Zorn-Arnold B, Havens, K (2010) Assessing botanical capacity to address grand challenges in the United States 64 pp. [plus appendices] http://www.bgci.org/usa/bcap

Legler B (2011) Specimen Imaging Documentation, version 4.0. Consortium of Pacific Northwest Herbaria. http://www.pnwherbaria.org/documentation/specimenimaging.php

Rosenthal DSH (2011) Paying for long-term storage. Presentation at the Coalition for Networked Information Membership Meeting, December 12–13. http://www.youtube.com/watch?&gl=US&hl=en&client=mv-google&feature=m-feedu&v=_5lQxmyz3xY&nomobile=1

Thiers B (2012+) Index herbariorum: a global directory of public herbaria and associated staff. http://sciweb.nybg.org/science2/IndexHerbariorum.asp

Woodward FI, Bazzaz FA (1988) The response of stomatal density to CO_2 partial pressure. Journal of Experimental Botany 39: 1771–1781. doi: 10.1093/jxb/39.12.1771

Zangeri AR, Berenbaum MR (2005) Increase in toxicity of an invasive weed after reassociation with its coevolved herbivore. Proceedings of the National Academy of Sciences 102 (43): 15529–15532. doi: 10.1073/pnas.0507805102

Appendix 1

Survey of Digitization in US Herbaria – 2011

This shows the questions asked. It is not the original form; that had a lot more blank space. The survey was kept short out of respect for the respondent's time.

Measuring Digitization Progress

Herbarium Code: _____
Specimen total (estimate): _____
Number of specimens databased: _____
Number of specimens imaged:_____
URL for searching database: _____
URL of regional node through which data are available: _____
Other nodes through which your specimen data are available: _____

Basic information

Herbarium Name:_____
Department:_____
Address 1: _____
Address 2: _____
City: _____ Zip Code: _____
Phone: _____
PO Box:_____ Mail Stop:_____
Lat.:_____ Lon.:_____
Name of contact person: _____
Email of contact person:_____
Taxonomic focus:_____
Geographic focus:_____

Appendix 2

Web Sites

This is a listing of all web sites mentioned in the text and a brief synopsis of their significance to the paper.

Alabama Plant Atlas: Provides information about plants in Alabama, including information derived from several herbarium databases. http://www.floraofalabama.org

Algaebase: AlgaeBase is a database of information on algae that includes terrestrial, marine and freshwater organisms. http://www.algaebase.org

Apiary: Program for enabling capture of collection data in the field. http://www.apiaryproject.org

Atrium: Technology data for managing diverse biodiversity data. http://www.brit.org/explore/bioit

Consortium of California Herbaria: State herbarium network. http://ucjeps.berkeley.edu/consortium

Consortium of North American Bryophyte Herbaria: Taxonomically focused herbarium network. http://symbiota.org/bryophytes/index.php

Consortium of North American Lichen Herbaria. Taxonomically focused herbarium network. http://symbiota.org/nalichens/index.php

Consortium of Pacific Northwest Herbaria: Regionally focused herbarium network. http://www.pnwherbaria.org

Cooperative Taxonomic resource for American Myrtaceae: Taxonomically focused herbarium network.http://cotram.org/collections/index.php

Index fungorum: Synonymized list of fungal names. http://www.indexfungorum.org/names/names.asp

Institute for Museum and Library Services (IMLS): US federal agency that has funded some of the work described. http://www.imls.gov

Intermountain Region Herbarium Network: Regionally focused herbarium network. Shares database with SEINet. http://intermountainbiota.org/portal/index.php

International Plant Names Index (IPNI): List of plant names and an indication of whether or not they are valid. Only shows nomenclatural synonyms. http://www.ipni.org

Mycoportal: Taxonomically focused herbarium network. http://mycoportal.org/portal/index.php

National Information Management and Support System (NIMSS): Information systems that serves the Agricultural Experiment Stations and the Extension Service in each state. http://nimss.umd.edu

National Science Foundation (NSF): US federal agency that has funded much of the work described. http://www.nsf.gov

SERNEC: Regional network for strengthening communication and promoting data sharing among herbaria, now also serving as a regional herbarium network. http://www.sernec.org

SourceForge: Web site that provides access to open source software. http://sourceforge.net

Southwestern Environmental Information Network (SEINet): Regionally focused herbarium network. Herbaria in the Intermountain Region share data with this network. http://swbiodiversity.org/seinet/index.php

Symbiota: Open source software for promoting collaboration and data sharing among herbaria. http://symbiota.org/tiki/tiki-index.php

Tropicos: Nomenclatural resource for bryophytes and vascular plants that shows how a name has been treated in different publications. Also the specimen database of the Missouri Botanical Garden. http://www.tropicos.org

US Virtual Herbarium (USVH): Project for promoting digitization in US Herbaria. This web site is not being maintained because of funding decisions by the US government. Arrangements are being made to move it, or something similar, to another site. http://usvirtualherbarium.org

Utah State University Herbarium: Provides access to the results of the 2012 herbarium survey. http://herbarium.usu.edu

WisFlora: Provides information about plants in Wisconsin, including information derived from several herbarium databases. http://www.botany.wisc.edu/wisflora.

Appendix 3

US Herbarium Survey 2012

Presented below are the questions asked on the 2012 survey. To save space, only the questions asked about digitization are shown. For more information, see http://herbarium.usu.edu/SurveyResults.html

About how many specimens are there in your herbarium? Please provide a single number, not separate estimates for different kinds of specimens.

Databasing: Some herbaria are entering data for a few fields when imaging, then completing data entry later. For that reason, there are two questions concerning databasing.

How many specimens in your collection have been at least partially databased?

How many specimens have been fully databased (you may answer unknown)?

Imaging. The questions below distinguish between imaging specimens (biological material) and imaging labels. If you do not distinguish between the two, put an asterisk by the answer for specimens.

How many of your *specimens* have been imaged?

How many of your *labels* have been imaged?

Georeferencing. How many of your specimens have latitude and longitude information?

Access

The next questions ask about the web site(s) through which your specimen information is available. If your database cannot be searched via a web site, you have finished the survey. Thank you for taking the time to complete it. If you wish to make a comment or suggestion, please use the space the end. Hand written comments are welcome

If your records are searchable via an institutional web site, what is its URL?

If your records are searchable via one or more regional websites, what are their URLs?

If your records are searchable via one or more taxonomically focused web sites, what are their URLs?

If you provide searchable access to your records through a regional web site that lies primarily outside the US, please indicate the focus of the site(s) and its(their) URL(s).

YOUR Comments:

ZooKeys 209: 75–86 (2012)
doi: 10.3897/zookeys.209.3119
www.zookeys.org

RESEARCH ARTICLE

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

The development of a digitising service centre for natural history collections

Riitta Tegelberg[1], Jaana Haapala[1], Tero Mononen[1], Mika Pajari[1], Hannu Saarenmaa[1]

1 *Digitarium: The Digitisation Centre of the Finnish Museum of Natural History and the University of Eastern Finland, School of Computing, SIB-labs, Joensuu Science Park, Länsikatu 15 (P.O. Box 111), FIN-80101 Joensuu*

Corresponding author: *Hannu Saarenmaa* (hannu.saarenmaa@uef.fi)

Academic editor: *V. Blagoderov* | Received 23 March 2012 | Accepted 29 June 2012 | Published 20 July 2012

Citation: Tegelberg R, Haapala J, Mononen T, Pajari M, Saarenmaa H (2012) The development of a digitising service centre for natural history collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 75–86. doi: 10.3897/zookeys.209.3119

Abstract

Digitarium is a joint initiative of the Finnish Museum of Natural History and the University of Eastern Finland. It was established in 2010 as a dedicated shop for the large-scale digitisation of natural history collections. Digitarium offers service packages based on the digitisation process, including tagging, imaging, data entry, georeferencing, filtering, and validation. During the process, all specimens are imaged, and distance workers take care of the data entry from the images. The customer receives the data in Darwin Core Archive format, as well as images of the specimens and their labels. Digitarium also offers the option of publishing images through Morphbank, sharing data through GBIF, and archiving data for long-term storage. Service packages can also be designed on demand to respond to the specific needs of the customer. The paper also discusses logistics, costs, and intellectual property rights (IPR) issues related to the work that Digitarium undertakes.

Keywords

Digitisation, imaging, natural history collections, service packages, out-sourcing, mass-digitisation, automation, logistics, costs, IPR

Introduction

In Finland, the 6 largest public natural history museums contain an estimated 22 million specimens, of which 12% have been digitally catalogued (i.e., minimally digitised). In addition, private collections contain up to 8 million specimens. It has been

estimated (Pelkonen et al. 2009) that unless digitisation productivity is dramatically increased, it will take about 1,000 person years of effort to digitise these collections. Thus, in 2010, Digitarium, the Digitisation Centre of the Finnish Museum of Natural History and the University of Eastern Finland, was established in Joensuu, Finland. Digitarium aims to speed up the digitisation process through an efficient production line and knowledge management of expertise on digitisation. The main idea is to selectively outsource mass digitisation from major museums into a dedicated service centre that works in close cooperation with museum customers. In most cases this also includes return transportation of the material to the service centre.

Special features of the production process at Digitarium are imaging of all material, XML-based data management, and a distributed workflow that can employ distance workers. Automation of imaging will produce large quantities of material ready for data entry. In addition to offering the employees working on data entry the option of working from home or from a library, remote access also provides an opportunity for crowd sourcing (Howe 2006, Flemons 2011, Flemons and Berents 2012). Crowd sourcing also functions as a means of promoting free and open access to national collections.

This document briefly outlines the process of digitisation as it is being implemented at Digitarium. In addition, the paper describes the approach used to develop service packages for customers, which are formed by connecting the steps of the digitisation process in a way that the customer requires. Finally this paper visits the issues of logistics, costs, and intellectual property rights (IPR), which are important in an outsourced operation.

Process steps

The steps of the digitisation workflow process are illustrated in the functional model shown in Fig 1. The steps from Receiving to Imaging require some handling of the physical samples, whereas steps from Data Entry onwards can be distributed through the internet to the best available agent. All steps of the workflow process can be executed asynchronously, although their logical order is somewhat fixed. The process is described in more detail in Lehtonen et al. (2011).

The process and workflow described below is driven by a dedicated software workbench (Fig. 2). This tool has been written by Digitarium in Java, and it runs on Windows. The workbench manages all data in the form of XML documents, and drives the digital cameras for imaging. It can also be used for distance work, and through SSH it can remotely retrieve and write the XML documents pertinent to each step in the workflow. The produced XML data conforms by the Darwin Core and Dublin Core standards.

The metadata describing datasets (i.e., groups of Darwin Core XML documents, as well as orders by customers) are stored in XML files using the Ecological Metadata Language EML (Fegraus et al. 2005), which is a standard for describing datasets in the biodiversity science community.

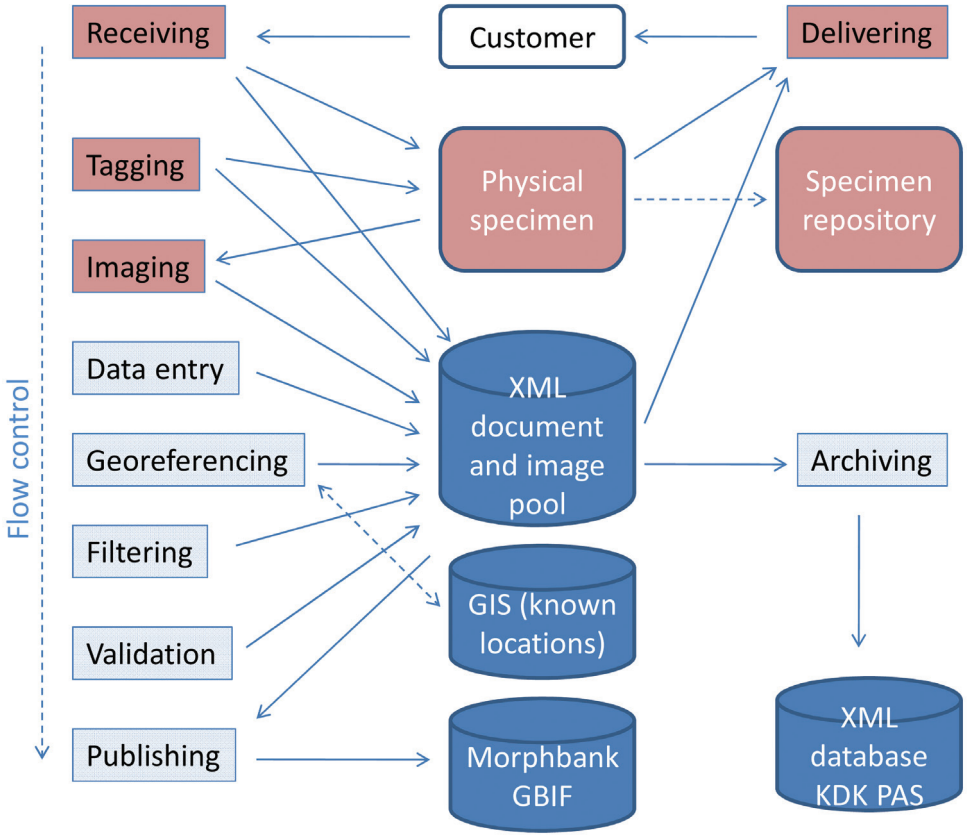


Figure 1. Functional model of the digitisation process. The steps pictured in red require handling of the physical specimens.

Receiving

Digitarium does not manage collections of its own, but is a shop for digitising materials for "customers" — that is, from museums and other institutions located elsewhere. The customer institute selects material for digitisation based on their own prioritisation. The received material and the procedures used for digitising are described in EML.

Tagging

Each sample is tagged with a label containing globally unique identifier in the form of an HTTP URI and a two-dimensional barcode (see Fig. 3 for an example). The URI can be resolvable if it is made to point to the collection database management system of the customer.

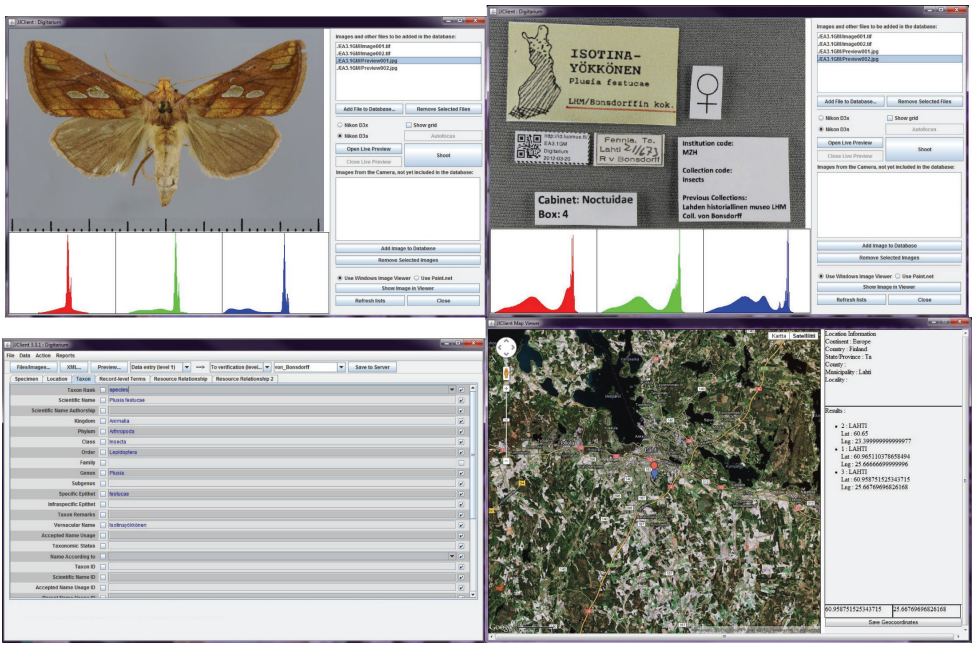


Figure 2. Selected windows of the digitisation workbench.

Imaging

The two main types of specimens that are digitised at Digitarium are plant specimens and insect specimens. A plant sheet is imaged in two pieces (see Fig. 3) with a high-end digital camera (i.e., a Nikon D3x, 24 megapixels). This way, a relatively high-quality resolution of 450 dpi over the entire sheet can be achieved at a relatively low cost. The two pieces are later joined using a panorama image stitching application based on our own algorithm, which is tuned for this kind of images. In the case of insect samples, the specimen and the labels are imaged separately with a 12 megapixel camera (i.e., a Nikon D3s using a Nikon AF-S MICRO NIKKOR 105 mm 1:2.8 objective and extension rings for the smallest objects). As the cameras are calibrated daily, no colour swatch has been included in the images. Our digitisation workbench drives all steps of image capture and annotation and all details of the imaging event and results are automatically stored in an XML document.

Delivery and optional specimen repository

After successful imaging, the specimens are returned to their institutions. Specimens can also be stored at Digitarium's repository for either short or long periods of time. This is an option for collections that are not under active study, and for excess specimens.

Figure 3. Image of a plant sheet stitched together from two parts – their boundary is barely noticeable in the middle of this sample. Notice the two-dimensional barcode, and a resolvable unique URI of the specimen details.

Data entry

The data from the specimen labels is entered manually from the images using our digitisation workbench and the vocabulary of the Darwin Core data exchange standard. In this step, we need to separate the "true and honest" reproduction of what has been written on the labels with the subsequent interpretation of that information. Any misspellings,

abbreviations, etc., are written in the "verbatim" fields of the latest Darwin Core standard vocabulary to preserve the original data. This has been somewhat problematic, as Darwin Core has not had verbatim fields available for all possible label data such as collector name, taxonomic identification, record number, label colours, etc. Thus, separating reproduction from interpretation is not yet fully supported. This led us to propose a new term, "verbatimLabel," for Darwin Core, although we have not yet implemented this new feature.

Georeferencing

Geographic coordinates are often not available from specimen labels. Our software workbench contains a function to retrieve them automatically using the web services of GeoLocate (2005). We only use the estimated latitude, longitude, and coordinate uncertainty in meters for point localities. When grid coordinates from the old Finnish national system (called "YKJ") are available, they are automatically converted into WGS-84 geographic coordinates using the point-radius method; this conversion can be well documented in Darwin Core.

Georeferencing is an optional step. It can be done by Digitarium simultaneously with data entry or verification, but it can also be left for the customer or remote expert to complete, if so agreed.

Filtering

Before publishing the data and images, the filtering of certain details such as coordinates of localities of endangered species may be necessary. For textual and numeric data, this can be done automatically based on the entered species names stored in the metadata of the dataset. Two versions of the XML file are retained: filtered and unfiltered. These details need to be masked manually from the image. Optionally, the customer may want to perform this step.

Validation

A final check of the data entry, georeferencing, and filtering is made by an experienced staff member. However, as the customer often wants to validate the digitisation result, all validation can be left to the customer.

Delivery of data

The data is delivered to the customer in the Darwin Core Archive (DwC-A) format (Wieczorek et al. 2012), which has been endorsed by Biodiversity Informatics

Standards (TDWG) (2009). Other delivery formats are available depending on the requirements of the customer. Furthermore, as the customer usually wants to have checkpoints for the work, intermediate data deliveries are often made. Delivery of the digital images has not yet taken place, as so far the customers have preferred Digitarium to host them.

Publishing

The collection data from the latest XML document version, as well as the images, are imported to Digitarium's Morphbank database service and Digitarium's GBIF IPT service. From there they are published, as agreed with the customer; if publication has not been agreed upon, the data and images remain private, and are available only to the customer and for Digitarium's internal use.

The Morphbank service, a part of the global and Nordic collaboration, is available at http://morphbank.digitarium.fi/. Morphbank is an image database tool designed particularly for natural history specimens and annotations made to them (Morphbank 2011). Morphbank provides permanent publication: after the preset publishing date has passed, the objects cannot, even in principle, be removed from the service. All Morphbank objects have stable short URIs that can be reused elsewhere.

The GBIF (2011) Integrated Publishing Toolkit (IPT) is a service for hosting biodiversity data that is intended to be shared globally. Its purposes at Digitarium are to produce the EML and DwC-A for all the datasets, and when agreed with the customer, to publish collection and specimen-level data thus promoting Digitarium's services. The IPT hosting service has also been required by several smaller museums and collections that do not have the infrastructure to connect with GBIF directly.

Archiving

All the XML documents and images will be retained indefinitely, first on Digitarium's Metacat (NCEAS 2012) service and eventually with the long-term archival service of the National Digital Library (2010). These archive functions are still under development.

Packaging of the services

The services described here are designed in cooperation with the customers to be flexible and meet the unique requirements of different clients.

Prior to each digitisation job, a formal agreement is made in terms of the details of the digitisation process, costs, and time frames. When negotiating the agreement, customers are informed of the option of customising the digitising services at Digitarium.

In the most basic case, the workflow will include steps from Receiving to Imaging. Data Entry will only include the actual information from the labels attached to the specimen, and basic interpretation that aids in later data discovery such as taxonomic group and country. Georeferencing, data filtering, and publishing may be left out, as the customer may want to perform these steps. However, the quality of the images and the technical correctness of the data entry will be verified.

In a more complete service package, descriptive data entry with full interpretation of taxonomic and locality details, georeferencing, and verification of the data will be included. Misspellings and unclear text will be retained in the "verbatim" field of Darwin Core, though. Dates and timings will be written following the ISO 8601:2004(E) standard. Country codes, institutional codes, and collection codes will be included.

In an "all-included" service package, all the steps shown in Fig. 1, as well as additional filtering, publishing, and archiving services, are included in the service. This is the most suitable method for the digitisation of an entire collection. The customer still has the opportunity to follow the process, sign off on the quality of products, and give scientific guidance. Entirely customised service packages can also be designed when needed so that resources and funding can be used to most directly answer the needs of a particular customer.

It is expected that, in the future, customers would want to monitor the progress of their digitisation jobs. For this purpose, a tracking and metadata system for the planning and scheduling of digitisation work is being prepared.

Customers are also able to participate in data entry first hand. In order to facilitate such collaboration, training on the Digitarium process can be included in the service. The aim is to produce repeatable and quality data, regardless of where the actual data entry takes place.

Finally, if a customer wants to operate these services entirely in-house, Digitarium can offer a turn-key package that includes the equipment needed to run the imaging and data entry processes. In this way, the customer may process the most delicate specimen samples in the safety of their own institution, while following the standards brought into use at Digitarium.

Logistics, costs, and intellectual property rights

Because the Digitarium service centre is located away from where the collections are housed, a few special issues must be taken into consideration. Quite rightly, transportation of the materials to the service centre is of major concern for the custodians of the collections. Not all materials can be considered for transportation (such as those stored in liquids). Materials that can be considered for transport must be carefully packed to ensure that they cannot move during transport. For botanical sheets this can be achieved, but requires some work. Insect collections are easier to package and transport; perhaps that is why most demands for Digitarium services have come from entomology collections. In a typical case, Digitarium retrieves an endowed entomological

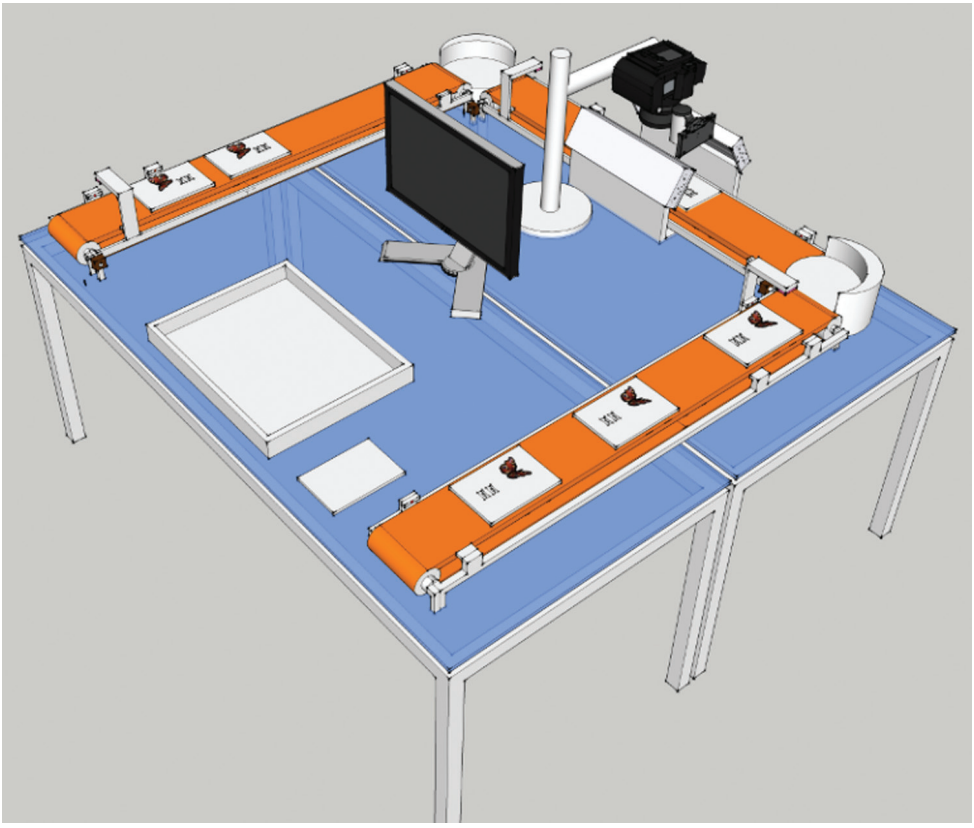


Figure 4. Conceptual design for automated imaging of entomological collections, which is currently being implemented at Digitarium.

collection, dismantles and processes it, and then delivers it to the customer institution, neatly re-packed and ordered in small units.

Receiving material also requires the extermination of possible pests that could be damaging the collection. Therefore, upon arrival, all received material is deep-frozen in a room that is in a separate building.

Processing of the material at the service centre does not necessarily take a long time, which reduces inconvenience for the customer in terms of being separated from their collection. Tagging and Imaging can in principle be done quickly, while the data entry steps can proceed at a more flexible rate (cf. Fig 1). Overall, a two- to four-week turnaround time is conceivable based on the experience of digitisation centres operating in the cultural history domain. The volume that can be processed in such time is quite variable, though, depending on type of material, and how much automation is possible.

Moving of material between organisations also requires agreement on intellectual property rights. The agreement that is made of each digitisation job transfers the copy-right to the customer when the customer has accepted the final delivery. Digitarium retains a parallel right to use the content within its own internal operations, but not

for delivery to other parties. This way, it can be ensured that no duplicate copies of the same data start circulating in global portals. In the case of images hosted by Digitarium on behalf of the customer, a rather restrictive variety of the Creative Commons license, BY-ND, is currently being applied.

For the costs of the services, only preliminary figures are available, as the process is still being formed and tested in many areas. So far, about 40,000 images have been made and 10,000 samples have been fully processed. Two-thirds of these samples have been entomological, and one-third botanical. On average, a staff member has been able to produce about 40 images or data entries per day. The cost of digitisation is currently 3.99 € per image and 5.61 € for data entry of a specimen, which makes a total of 9.60 € for a fully processed sample. These costs do not include development, administration, equipment, housing, etc. We expect the costs to reduce rapidly as the process becomes increasingly streamlined and automated.

What has been described above is still essentially a manual process. However, the separation of the different steps of the workflow offers a strong possibility for automation. In fact, Digitarium is in the process of building a conveyor belt system that moves the samples for automatic imaging (Fig. 4). We expect the costs of imaging to dramatically decrease when this system is in operation.

Conclusion

Digitarium aims to accelerate the digitisation of natural history collections, both in Finland and around the world. In order to achieve industrial-scale efficiency, we are considering the aspects of quality control, economies of scale, automation of processes, cost of labour, community resources, and workflow (cf. Speers 2009).

Progress in all these areas is being made, but a full solution has not yet been delivered. In particular, the automation of imaging and related logistics is still being crafted. The fact that all material is being imaged makes it possible to distribute data entry and subsequent steps in the process to off-site workers and to rely on crowd-sourcing for Data Entry. In these ways, processing costs can be reduced and access to remote experts can be gained for purposes such as handwriting recognition, languages, and species identification. On the other hand, digitisation technicians at Digitarium are trained to produce repeatable and qualified data from all sorts of collection material.

By offering the service packages described here, Digitarium can ensure that the wishes and needs of its customers can be met. Quality assurance not only covers the images and data, but also extends to our descriptions of the process and products. In this way, customers may choose the extent of the processing they require for a particular specimen or collection based on their own prioritisation.

The Digitarium service centre is located in Joensuu, a peripheral area of Europe, where dedicated funding sources such as the European Social Fund and the European Regional Development Fund have been available to boost the economy and build infrastructure. These funding sources are available particularly to new member coun-

tries of the EU, and offer a good opportunity for building research infrastructures such as digitisation services.

We believe that the outsourcing of digitisation to dedicated service centres with decentralised processes and well-defined service packages designed in cooperation with customers can speed the digitisation process up from the current manual practices to industrial-level efficiency (GBIF 2008, Speers 2009, Berendsohn et al. 2010).

Acknowledgments

This paper is dedicated to the memory of Larry Speers, a friend and a colleague who inspired us to believe that moving digitisation from a cottage industry into the information age is not only possible, but necessary. This work has been financed by the European Social Fund and the European Regional Development Fund.

References

Berendsohn WG, Chavan V, Macklin J (2010) Summary of Recommendations of the GBIF Task Group on the Global Strategy and Action Plan for the Digitisation of Natural History Collections. Biodiversity Informatics, 7 (2): 67–71.

Biodiversity Information Standards (2009) Darwin Core Text Guide. http://rs.tdwg.org/dwc/terms/guides/text/index.htm

Fegraus 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 (3): 158–168. doi: 10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2

Flemons PK (2011) Crowd-sourcing: perpetual valuable resource or a passing shower of dubious worth? Abstracts of the TDWG 2011 Annual Conference. https://mbgserv18.mobot.org/ocs/index.php/tdwg/2011/paper/view/118

Flemons P, Berents P (2012) Image based Digitisation of Entomology Collections: Leveraging volunteers to increase digitization capacity. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 203–217. doi: 10.3897/zookeys.209.3146

GBIF (2008) Training Manual 1: Digitisation of History Collections Data, version 1.0. Copenhagen.

GBIF (2011) The integrated publishing toolkit. http://www.gbif.org/informatics/infrastructure/publishing/

GEOLocate (2005) Georeferencing software for natural history collections. http://www.museum.tulane.edu/geolocate/

Howe J (2008) Crowdsourcing: Why the power of the crowd is driving the future of business. Three Rivers Press, New York, 304 pp.

Lehtonen J, Heiska S, Pajari M, Tegelberg R, Saarenmaa H (2011) The process of digitising natural history collection specimens at Digitarium. In: Jones MB, Gries C (Eds) Proceedings of the Environmental Information Management Conference 2011 (EIM 2011). September 28-29, 2011. Santa Barbara, CA. University of California, 87–91. doi: 10.5060/D2NC5Z4X https://eim.ecoinformatics.org/eim2011/eim-proceedings-2011

Morphbank: Biological Imaging (http://www.morphbank.net/, 31 May 2011). Florida State University, Department of Scientific Computing, Tallahassee, FL 32306-4026 USA.

National Digital Library Initiative (2010) Long-term preservation project. Final report v. 1.0. 58 p. http://www.kdk.fi/images/stories/LTP_Final_Report_v_1_1.pdf

NCEAS (2012) Metacat: Metadata and Data Management Server. http://knb.ecoinformatics.org/knb/docs/

Pelkonen V-P, Saarenmaa H, Laurenne N (Eds) (2009) Luonnontieteellisten museokokoelmien digitointi. Strategia ja toimintasuunnitelma 2010-2015. Helsingin yliopisto, Luonnontieteellinen keskusmuseo 31.12.2009.

Speers L (2009) From ink to electrons: Issues to be considered. http://www.canadensys.net/digitization

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, de Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An evolving community-developed biodiversity data standard. PLoS ONE 7(1): e29715. doi: 10.1371/journal.pone.0029715

ZooKeys 209: 87–92 (2012)
doi: 10.3897/zookeys.209.3609
www.zookeys.org

FORUM PAPER

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

'From Pilot to production': Large Scale Digitisation project at Naturalis Biodiversity Center

Jon Peter van den Oever[1], Marc Gofferjé[1]

1 *NCB Naturalis, 2333 CK, Leiden, Nederlands*

Corresponding author: *Jon Peter van den Oever* (jpvandenoever@gmail.com)

Academic editor: *V. Blagoderov* | Received 29 June 2012 | Accepted 10 July 2012 | Published 20 July 2012

Citation: van den Oever JP, Gofferjé M (2012) 'From Pilot to production': Large Scale Digitisation project at Naturalis Biodiversity Center. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 87–92. doi: 10.3897/zookeys.209.3609

Abstract

By the end of 2009 the Dutch Government awarded the establishment of NCB Naturalis with €30M funding. The amount is invested in three programs: Scientific Infrastructure for DNA Barcoding, Integration and Relocation of collections and Collection Digitisation. In this article we describe the highlights of the Digitisation Programme.

Keywords

Large scale digitisation, NCB Naturalis, Pilot, Collection, Digistreet, Programme

Introduction

Naturalis Biodiversity Center, the Netherlands Center for Biodiversity, was launched on 28 January 2010. The center is the result of the cooperation between Amsterdam University (Amsterdam Zoological Museum), Leiden University and Wageningen University and Research Centre (National Herbarium Netherlands) and the National Natural History Museum Naturalis in Leiden. The partners' collections are being brought together at Naturalis BC and will be integrated into a collection totalling over 37 million objects. In terms of collection size, Naturalis BC is one of the top five natural history museums in the world.

By the end of 2009 the Dutch Government awarded the establishment of (at that time) NCB Naturalis with €30M funding from the National Gas and oil profits (FES=funding economical structure (empowerment). This fund is responsible for

many investments in the Cultural Heritage Sector. The amount is invested in three programs: Scientific Infrastructure for DNA Barcoding, Integration and Relocation of collections and Collection Digitisation. In this article we only describe the highlights of the Digitisation Programme.

Digitisation program at Naturalis Biodiversity Center

In 2010 the preparations began to develop an overall program for the mass digitisation of the collections. The program organisation had to meet 2 main goals:

– digitise at least 7M objects of the total of 37M specimen/objects;
– develop a permanent digitisation infrastructure (to ensure the remaining objects can be processed in the near future).

The structure by which the digitisation has been developed at Naturalis is different from the classical approach. In the current economic crisis the challenge is to do more with less money. Therefore the solutions must contain new and innovative perspectives on digitisation.

When Naturalis applied for funds, the average cost of digitisation was estimated (by experience of the past) to be approximately € 5 per object. The Dutch government granted € 13 M to digitize approximately 7 million objects (average € 1.86 per object).

Therefore the following decisions had to be made:

- to digitise a large number of objects through an industrial approach.
- To collect only basic metadata associated with an object, which later can be amended.

The Prince 2 methodology is used and the projects timeframe was first set to Q4 – 2013, which was later extended to June 2015. Project governance is carried out by the Steering Committee, overseeing scientific quality of the project. The board of directors of Naturalis BC is represented in the steering committee. Program manager, project managers and project leaders are responsible for everyday work, from the project set up to hiring staff, from housing to planning of collections to operations control, from budgeting to decision preparation and execution. The entire program consists of around 80 people. Several partner institutions (Paris, London, Finland, Berlin) were visited to define best practices. A series of pilot projects were conducted before commencing large-scale digitisation projects and selecting outsourcing partners.

Several stages of the Programme implementation can be distinguished:

- Testing and selecting technologies
- Developing tools: Basic Registration Database and Central Registration System
- Conducting Pilot Projects

- Selection and Prioritization of collections for digitisation
- Choosing outsource vendors and suppliers
- Execution of projects

Approach

A tier-based approach has been developed for digitisation of the Naturalis collection (Fig. 1)

- ~2,000,000 specimens are to be digitised in-house with detailed metadata extracted ("Digistreets")
- ~5,000,000 specimens will be digitised with basic metadata acquisition through outsourced vendors
- For the rest of the collection (~30,000,000 specimens) a high-level inventory will be created.

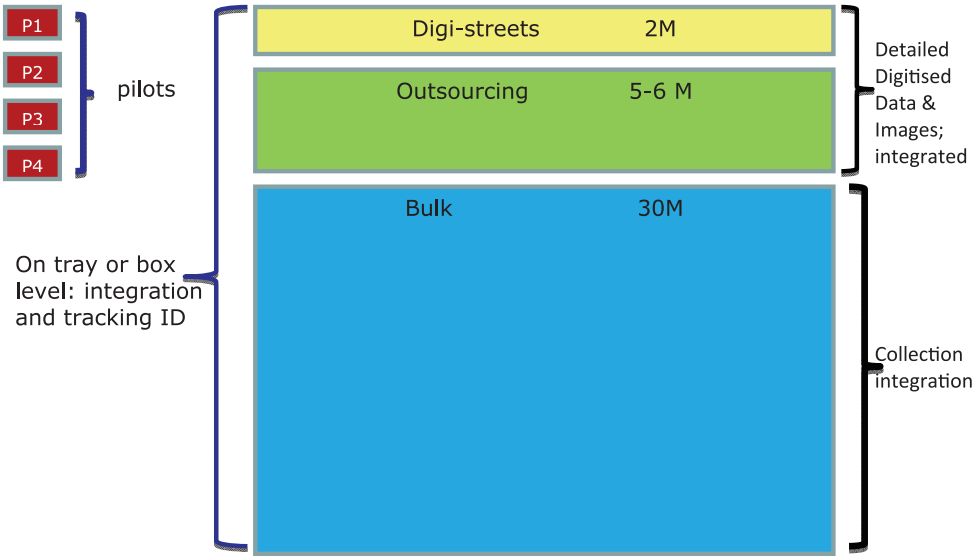


Figure 1. Structure of the digitization programme at Naturalis Biodiversity Center.

Prioritisation

When selecting parts of the collection for in-house detailed digitisation the most important factor was prioritisation by scientific or outreach value of the outcome (see below). Therefore, collections related to particular research or curatorial activities were identified. Value-for-money was a decisive criterion for outsourcing digitisation. Only collections for which industrial-scale digitisation technologies exist, which can be rela-

tively safely moved to another site, and where such service is provided for reasonable price, were selected. The most obvious example of such collections is herbaria. For collections, which are not extensively used at the present, or for which mass digitisation technologies are not yet available, or too expensive, high-level inventory will be built, describing content of every drawer or lot as detailed as practical.

One of the key strengths of the digistreets is that they must be demand driven and therefore collection independent. The Programme has developed a framework of priority setting and decision making in accordance with the institution's priorities (Fig. 2). The most treated, most important collections are key for the priority selection. This is a radical change of policy where in the past every scientist, taxonomist or biodiversity researcher had a personal history of raising funds and persuading decision makers into why their project should be prioritised. Transparency of procedure and objectified criteria of selection help to identify priority collections. Some of the indicators are:

- collaborative biodiversity projects
- European-funded and co-funded projects
- economic importance of the group
- relevance for citizen scientists and lay public
- collection conservation status

Prioritisation of projects is a multi-step process and includes (1) prerequisites: criteria mandatory for all projects, to reject unacceptable projects; (2) soft criteria: professional opinion of panel members, to create a long list of candidate projects; and (3) hard criteria: point-base factual criteria, to weigh projects and to arrange them in order of preference.

Schedule for prioritizing proposed digitizing projects within the entomology collection of NCB Naturalis

PHASE	GROUP	TASK	RESULT
Gathering project proposals	Collection managers and scientists working with the entomology collection	Call for project proposals	
		Head of entomology dept. gathers proposals and assembles them in one document	
		Document circulates within the entomology dept.	Document with all proposed projects
		Plenary discussions within entomology dept.	
Prioritizing	NCB Naturalis - By a NCB Naturalis-wide panel, through an inquiry - By the Director of Collections	Digitizing dept. checks in how far the proposed projects generally agree with NCB Naturalis policies, rejects projects that do not.	
		Panel prioritizes on general criteria	
		General manager Collections modifies priorities list following policy-based criteria	
			A prioritized list of projects
	Digitizing dept.	Feasibility within the planning period	Proposal to (partial) execution within the planning period
		Proposal to (partial) execution	
	A prioritized list is now available. It is provisionally accepted, pending decisions during the phase 'approval'. Execution of the projects can now commence.		
Execution	Production line	Execution of projects	Projects (partially) executed
Approval	Board of Directors	Review	Provisional approval
	Review Board	Review	Notification
	Scientific advisory board	Review	Comments and concurrence
	Board of Directors	Final approval	Decision taken
	Projectteam	Communicating approval	Stakeholders informed

Figure 2. Life cycle of Digistreets: stages of planning and executing of progect.

Pilots

Every digistreet is developed and derived from a pilot phase. The pilot can be defined as a proof of concept of a particular project or technology. A set of success criteria is devised and agreed upon start of every pilot. A time frame between the 3 to 6 months is needed to sharpen the requirements, the workflow, object handling and to test the technology. The Mollusc and the Entomology digistreet were the first two industrialised production projects developed. The Mollusc digistreet can be visited at the Live Science Hall at the museum. An application for iPads was developed so the visitors can be involved in transcription of scanned label information. Within 9 months 17K labels were transcribed by the members of public and checked by our taxonomists. Approximately 8K label transcriptions were useful. The data will be imported in the system. The App is enhanced and web enabled and now available for the visitors at the NCB website. The idea is that in the near future every digistreet will have an App to engage the community. After the pilot phase an evaluation report is constructed for the steering committee which made a decision on viability of a larger project. Most of the pilots were transferred into digistreets. A few pilots have not been developed to full-scale projects because the technology or process didn't meet the quality standards or requirements. An example was the 3D digitization of Bird's specimen. The quality of the images and the 3d viewing technologies were not mature enough.

Digistreets

'Digistreets' are production lines for digitisation of objects that have a lot in common from the point of view of registration, handling, and safety regulations.

Based on the overall collection characteristics nine digistreets were defined and developed:

- Wood samples
- Entomology collections
- Herbarium sheets
- Mollusc collections
- Dry mounted Vertebrates/Invertebrates;
- Alcohol/formaldehyde samples
- Microscopic slides;
- 2-D material (drawings, rare books, photographs, paintings, archives, microfiches etc.);
- Geological and paleontological collections

Each digistreet is managed as a separate project; it has a specific location, set of tools and equipment, and a more or less tailored version of the Central Registration System. Fixed targets (scope, time, quality) and a fixed budget are set for each di-

gistreet; and staff are provided for the duration of the project. Every digistreet staff member is fully aware of what they are supposed to process in what time at what cost and quality. An exception is the herbarium which combines the shipment of all the duplicate sheets at Wageningen and the development of two separate production lines: an outsourcing street in Leiden and a digistreet in Wageningen. The experience of the digistreets guidelines and requirements are being applied for the outsourcing part.

The acceptance of the goals of the digitisation project by the organisation is the key to successful projects. A collection manager is needed to instruct and manage the "streetworkers". Registrators (data entry), taxonomists and teamleaders are managed by the digistreets' projectleader. A process owner (institutionalised job role) is the leading decision maker on collections policies and priorities. He or she can oversee the individual collection requirements or the demands from the sections Collection Management, Research or Outreach. The process owner is also ultimately responsible for the safety of the staff/people or the collection objects.

Results

From the start of the programme (August 2010) until July 2012 approximately 1,000,000 objects have been internally digitised by a temporary staff of 80 people employed in digistreets. The outsourcing project, digital image bank and content management system are in a tendering phase and will be implemented in Q3/Q4 2012. Average costs per digitized object is provided in Table 1.

Table 1.

Cost of digistreets, per object	€ 2.50
Cost of outsourcing, per object	€ 0.90
Cost of infrastructure and equipment, per object	€ 0.30
Overhead (project management etc), per object	€ 0.20
Average cost per digitized object, entire programme	**€ 1.86**

Conclusions

- Mass digitisation of natural history objects is proven to be possible at reasonable costs;
- Industrial methods and concepts are a help—not a threat—to collection management and large scale object digitisation;
- By digitising the collection it is ensured that the data is available online, comparable and validated independently from location and time;
- Through the digitisation process new relations and associations can be made between taxonomies, object transcriptions, meta data, context and images;
- Data is provided using the taxonomic worldwide standards (GBIF, Darwincore) and can be accumulated, amended and used nationally and internationally.

ZooKeys 209: 93–102 (2012)
doi: 10.3897/zookeys.209.3121
www.zookeys.org

RESEARCH ARTICLE

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

Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach

Elspeth Haston[1], Robert Cubey[1], Martin Pullan[1], Hannah Atkins[1], David J Harris[1]

1 *Royal Botanic Garden Edinburgh, 20a Inverleith Row, Edinburgh, EH3 5LR, UK*

Corresponding author: *Elspeth Haston* (e.haston@rbge.org.uk)

Academic editor: *V Blagoderov* | Received 23 March 2012 | Accepted 13 July 2012 | Published 20 July 2012

Citation: Haston E, Cubey R, Pullan M, Atkins H, Harris DJ (2012) Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 93–102. doi: 10.3897/zookeys.209.3121

Abstract

Digitisation programmes in many institutes frequently involve disparate and irregular funding, diverse selection criteria and scope, with different members of staff managing and operating the processes. These factors have influenced the decision at the Royal Botanic Garden Edinburgh to develop an integrated workflow for the digitisation of herbarium specimens which is modular and scalable to enable a single overall workflow to be used for all digitisation projects. This integrated workflow is comprised of three principal elements: a specimen workflow, a data workflow and an image workflow.

The specimen workflow is strongly linked to curatorial processes which will impact on the prioritisation, selection and preparation of the specimens. The importance of including a conservation element within the digitisation workflow is highlighted. The data workflow includes the concept of three main categories of collection data: label data, curatorial data and supplementary data. It is shown that each category of data has its own properties which influence the timing of data capture within the workflow. Development of software has been carried out for the rapid capture of curatorial data, and optical character recognition (OCR) software is being used to increase the efficiency of capturing label data and supplementary data. The large number and size of the images has necessitated the inclusion of automated systems within the image workflow.

Keywords

Large-scale digitisation, curation, data entry, image capture

Introduction

The need for the digitisation of biological collections is widely recognised (eg European Commission 2011, Kroes 2011, Niggeman et al. 2011) resulting in the development of national digitisation strategies (eg Beach et al. 2010). The challenges of digitising natural history specimens have been explored (eg Beaman et al. 2007, Vollmar et al. 2010) and there have been several studies investigating data capture methods (Beaman et al. 2006, Heidorn and Wei 2008, Best et al. 2009, Lafferty and Landrum 2009, Granzow-de la Cerda and Beach 2010, Haston et al. 2012). Within this context of large scale digitisation of natural history collections, there is a need for the development of digitisation workflows to manage each of the elements of the digitisation process.

In developing workflows for the digitisation of herbarium specimens there are many factors which will influence the decisions made. Whilst it is clear that the financial costs of a digitisation programme may significantly limit the options available for equipment, software, staffing and storage, there are also other factors to consider. The funding itself may be irregular and be used for a range of diverse projects. Each institute has their own priorities and constraints and in the larger institutes there may be a range of digitisation programmes each with a different focus but which need to be integrated in some way. The recommendation of following a demand-driven digitisation model (Berendsohn and Seltmann 2010, Berendsohn et al. 2010, Berents et al. 2010) may result in an increase in the diversity of material being prioritised which will have an impact on the efficiency of the workflow. The concept of scalability is a factor which takes into account the potential increase in funding and resources. In addition, the integration of digitisation workflows into the core curation activities may play a large part in the decision-making process.

At the Royal Botanic Garden Edinburgh (RBGE), we have aimed to develop workflows which incorporate automated systems to enable us to expand and speed up the digitisation process. However, given the irregular nature of much of the funding available for digitisation, we have also based the digitisation workflows on a modular system which has the potential to be scaled up as funding becomes available. Additional modules may be added as they are developed, including a georeferencing tool (Llewellyn 2011) and additional quality control elements. A key factor in developing the workflow has been the need to continue to manage the images and data after capture. This is a very significant addition to the workload for herbarium staff and there is a requirement for this aspect of the workflow to be as efficient and simple as possible, with the aim of helping curators in the future to manage the collections.

Where possible, the digitisation workflow aims to use shared standards and formats. The adoption of standards allows easier transfer and sharing of data and is recognised as being of high importance in digitisation strategies (eg Beach et al. 2010). All data are routinely submitted to the Global Biodiversity Information Facility (GBIF), images are available on Encyclopedia of Life (EOL), a proportion of images and data are submitted to JSTOR and we are working on processes for submitting images and data to the Barcode of Life Database (BOLD) and Europeana.

Workflows and processes

The integrated digitisation workflow at RBGE has been developed over the last four years, during which time it has evolved into the present system. Over this period large digitisation projects have been undertaken wholly within this system, whilst other projects have gradually been incorporated. All digitisation is now undertaken within this integrated system and there are currently 160,000 specimens digitised and available online (www.rbge.org.uk). Whilst increasing the rate of digitisation has been a contributing factor in the development of workflows, the need for managing the data, images and processes has been the most important driver for the development of this integrated system.

There are three primary workflows within the digitisation programme (Fig. 1). The specimen workflow involves the physical movement and preparation of the specimens and folders. The data workflow focuses on the capture and management of specimen data (included within a broad "metadata" concept by Berendsohn et al. (2010)). Finally the image workflow focuses on the capture and management of images and related image management data including the equipment, operator and file location. These workflows and the interactions between them are described here.

The specimen workflow

In this context, the specimen workflow involves the physical selection and movement of specimens within the digitisation process as well as the preparation of the specimens and folders. This is closely linked with existing specimen workflows for loans, incoming specimens, destructive sampling, curation etc.

See the Specimen workflow in Figure 1.

The selection of the specimens is dependent on the outcome of the prioritisation procedure. The specimen workflow developed at RBGE predominantly focuses on large taxonomic or geographical groups to increase efficiency, and scaling up small user requests to more manageable units based on taxonomy and geography. The prioritisation of specimens within the digitisation programme has been mainly influenced by RBGE research strategy as well as external projects. This has resulted in the selection of floristic areas such as SW Asia and the Middle East, as well as focus taxonomic groups such as Sapotaceae, Zingiberaceae, Begoniaceae and Gesneriaceae. Funding from the Andrew W Mellon Foundation through the Global Plants Initiative enabled us to digitise all the type specimens, which form another significant part of the collections.

The preparation element (ie taxonomic recuration and specimen & folder preparation) of the specimen workflow is an important factor which is often under-estimated. This fundamental curatorial work includes ensuring that the specimens are correctly filed and that the filing name is legible and clearly visible, as well as ensuring that the condition of the specimens is assessed and conservation work carried out as required.

The decision to keep the herbarium open and to maintain full access to the specimens as much as possible during the digitisation programme has been necessary due to the expected duration of the digitisation work given the current funding. In practice, this has resulted in the specimen workflow for digitisation being affected by many curatorial and research activities. We have therefore aimed to integrate the workflow and other curatorial workflows currently in place. An outcome of this integration has been the modification of some curatorial practices, including loan and destructive sampling procedures.

The digitisation workstations are currently all within the herbarium area to reduce the amount of movement and to remove the need for freezing specimens on return for pest control. We have aimed to keep the number of specimens out of the cabinets at any time to a minimum whilst working with a large enough unit to be efficient.

The inclusion of an assessment of the condition of the specimens and some preservation work has reduced the rate of digitisation. However, this work is critical for the conservation of the collections and incorporating this work within the digitisation programme when the specimens are being handled is allowing us to improve the condition of the specimens. The assessment of specimen condition can also be collated and used to inform strategic decisions about the overall management of the collections.

The scope of an individual digitisation project and the arrangement of the specimens within a herbarium has a large impact on the efficiency of the specimen workflow. Whilst the most efficient workflow would generally be to work through the collections cabinet by cabinet, this can be difficult to reconcile with digitisation projects based on a particular collector or country, or with demand-driven digitisation.

The data workflow

The data workflow here includes all elements of capturing and managing data associated with the specimens, and linking these to the images and image management data. Logistically, the data associated with biological collections can be divided into three main categories for digitisation (Haston et al. 2012). Label data which are present on the specimen; curatorial data which are found on the containers holding the specimens; and supplementary data which are held separately from the collections in indexes, archives and literature. These data types can be captured using different methods at different stages of the data workflow.

Curatorial data are held separately from the specimen within the collections. At the Royal Botanic Garden Edinburgh this generally consists of two pieces of data: the filing name of the specimen and the broad geographical region from where it was collected. These represent the classification and location of the specimen within the collections, providing key information for the physical location and arrangement of specimens. Some or all of these data may not be present on the specimen itself as label data. This property means that the most efficient way to capture this data is from within the collection using information on the folders.

Label data are physically associated with the specimen and are generally visible in the corresponding digital image. This property allows these data to be captured at a later stage in the overall digitisation workflow. At RBGE, there is a small number of labels that are obscured by plant material or capsules which are not routinely captured.

Supplementary data such as field notebooks, citations in literature and online resources including Genbank, are independent from the label and curatorial data but can be used to enrich them.

See Data workflow in Figure 1.

The data workflow at the Royal Botanic Garden Edinburgh starts with the capture of curatorial data. Software written in PHP has been developed in-house to provide a simple web-based interface for rapid capture of the filing name, geographical region and barcode assigned to each specimen. The interface is designed around the fact that at the lowest level specimen storage within the herbarium is arranged into separate folders for each species within a geographical region. Within the interface users can select the species and geographical region for the folder and then add the individual specimens in each folder simply by scanning the barcode on the specimen. A specimen record is created for each barcode scanned and cross checked against any existing records in the herbarium database. After validation and error correction the new records are then batch imported into the herbarium database (*BG-BASE*™ version 6.8). A similar tool has now been developed within *BG-BASE*™.

Once the specimen has been imaged, label data can be captured during subsequent sweeps of data entry. Specimen images are processed through optical character recognition (OCR) software (ABBYY Recognition Server v. 3.0). At present the resulting text is stored unparsed as a single data string. This is then searched for recognisable tags (characters) to allow the creation of subsets of images and specimen records. These subsets are visually checked to ensure the selection process was correct and then the relevant data automatically entered. This is currently being carried out for collector and country. Finally, additional sweeps of label data entry are carried out by operators using a combination of the images and OCR text.

Within a modular system a level of data entry can be independent from imaging. This allows the ability to tailor the work being undertaken to the resources available. The use of minimal data capture methods enables the rapid creation of placeholder records that give collection managers valuable information about the number of specimens within a taxon for a particular filing region, and thus act as a catalogue of the collections. These also act as placeholder records to which images and OCR data can be attached, and which can be expanded as and when additional resources and technology become available.

The overall workflow is designed to accommodate the different requirements of the separate projects being undertaken simultaneously in the herbarium. These requirements may vary from full data entry with an image as is usually required for taxonomic or floristic work, partial data with georeferenced locality which is often all that is required for biogeographic studies, through to a basic catalogue record with minimal data for curation purposes. All these requirements can be handled within the one system. This is of particular importance due to the irregular nature of funding.

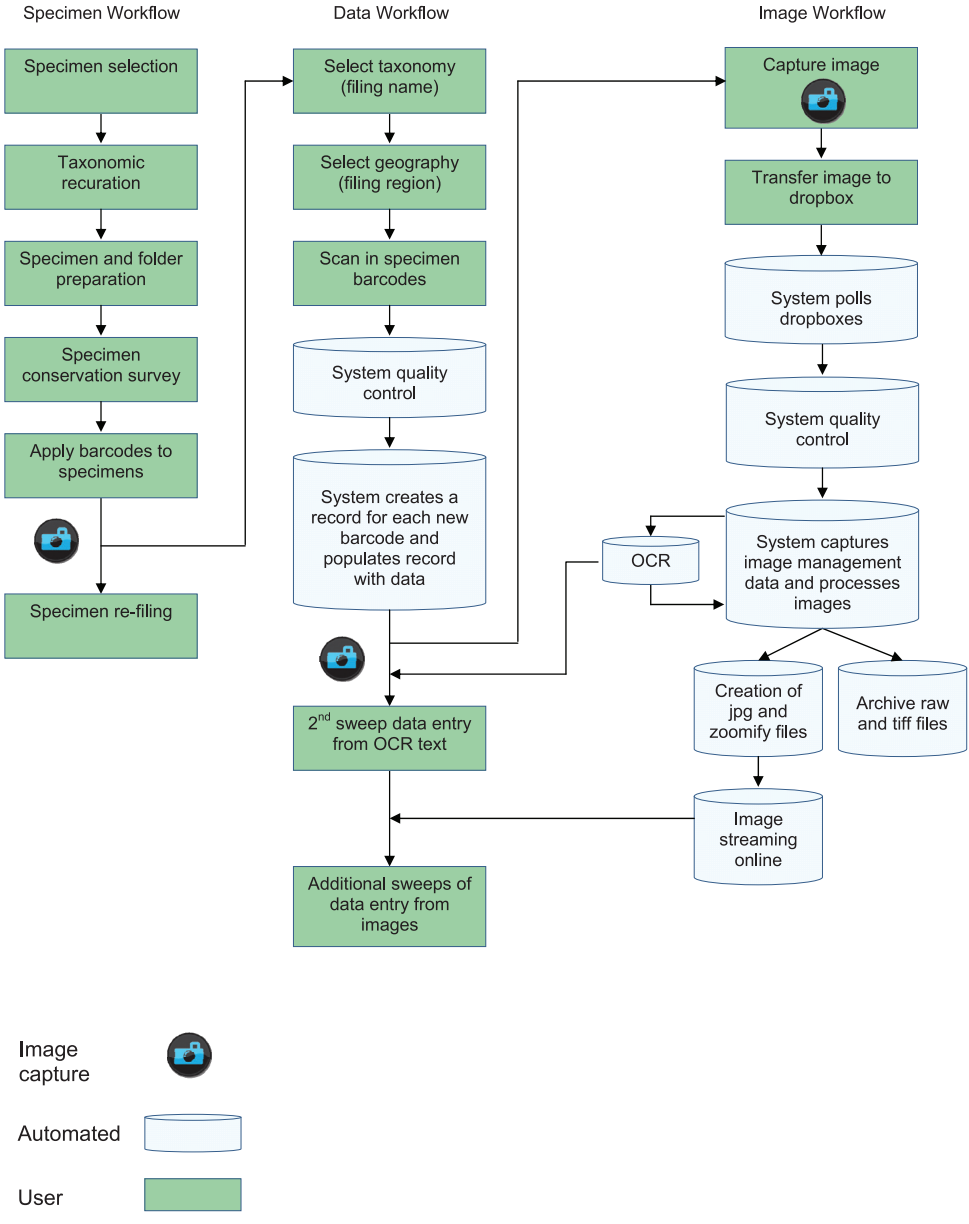


Figure 1. Diagrammatic overview of the digitisation workflows at the Royal Botanic Garden Edinburgh (RBGE)

The image workflow

Digitisation projects are resulting in large numbers of high quality images of approximately 150MB each. The scale of the digitisation programmes is too large

for completely manual processes to be used to manage these images. Image workflows being developed at the Royal Botanic Garden Edinburgh include image capture, processing, image management data recording, optical character recognition (OCR), quality control, image streaming online and archiving. This is carried out within a system based on image server software written in-house. This software has been written in Visual Basic and is designed to run as a Windows service. The software is responsible for marshalling newly scanned images into their ultimate destinations, registering the new images and all derived versions of that image in the image database, creating rescaled web viewable versions of each image, tiling the images to allow web presentation of zoomable versions of the image, submitting the images for OCR processing and recording the results of the OCR process in the image database. Multiple instances of the service can be installed in multiple servers to allow parallel processing of the new images.

See the Image workflow in Figure 1.

Image capture is carried out using two methods, but which feed into the same image workflow system. Epson Expression Model 10000XL scanners at 600dpi, result in tiff files of approximately 150–200MB each. The Leaf Aptus II-10 56 megapixel digital backs result in raw files of approximately 100MB from which tiffs of approximately 150MB are created using LeafCapture software. Preliminary quality control checks are carried out at this stage. These include manually checking the focus and cropping, as well as ensuring that the tiff file has been sucessfully created.

All images are then saved to a dropbox folder structure. The folder names comprise basic image management data including the equipment and operator's name. The image server software continuously polls the dropboxes for new files.

Additional automated quality control checks are carried out at this stage to ensure that the files are within acceptable size boundaries, that the filename fits a standard pattern, and that an electronic file with the same filename does not already exist.

As they appear, the system records associated image management data, including the equipment and operator's name, into an image database. The system then creates fully tiled image files which are stored in a zip compressed file. The tiles are extracted from the compressed file by the image server software in response to tile requests from the image viewer. The image viewer is an embedded object contained with the HTML page presented to a web browser. We currently use Zoomify image viewer software (Zoomify Enterprise ™).

A jpg file of approximately 1MB is also created and made available online. A copy of the tiff file is transferred to the ABBYY OCR workflow (ABBYY Recognition Server 3.0). Finally, the raw and tiff files are saved to archive folders, to be stored offline on tape and external hard drive storage. The locations of all these files are recorded within the image database.

This system has been developed as a modular system which can be extended as the number of cameras or scanners increase. There has been an emphasis on developing more automated systems but which can allow an element of user interaction if required, particularly within the quality control elements.

The decision to capture the images at 600dpi or equivalent was based on producing images which contain the same visible information as would be available through the standard taxonomic tool of a 10× hand lens. This results in very large images in excess of 150MB which create significant image management problems. There is currently debate about the need for such high resolution and RBGE has been involved in these discussions. We have felt the need to maintain this high level of resolution to ensure that sufficient information is retained, and in the understanding that these images can be scaled down in the future but cannot be scaled up.

Deliberately keeping both the raw and the tiff formats increases the demands on storage. In the rapidly changing environment of image file formats we aim to be inclusive and retain our ability to adapt to future developments.

Discussion

The workflows developed here have been strongly influenced by the presence of different funding streams and a diversity of digitisation projects along with the need to create a modular and integrated workflow to manage the processes, images and data. This is in contrast to an alternative digitisation approach such as that seen at the Muséum national d'histoire naturelle (MNHN) in Paris which aims to digitise all specimens within a single project, using a more unified approach.

A single unified structure may reduce the problems inherent in a modular system in which linking and maintaining links between software developed by different programmers and residing on different servers can be an issue as versions change over time. In contrast, a modular approach can potentially benefit more readily from advances in technology as modules can be added, updated or replaced as they become available.

One of the benefits of the modular system developed at RBGE has been to create an integrated but flexible management structure for specimens, data and images, which reduces the need for individual projects to create their own systems with the additional time and costs involved.

A second and highly significant benefit of an integrated system is that it helps with the curation of images and data post capture. It is essential that the on-going curation of these digital collections is considered as early as possible. The data and images need to be available and accessible but they also need to be kept up to date (with new determinations and additional data) and new file and archive formats. Having data in multiple systems, managed by different projects makes this on-going curation task almost impossible. Having them in one system makes this daunting and ever-growing task more achievable.

Acknowledgements

The authors would like to acknowledge the support of the Andrew W Mellon Foundation and the Scottish Government in funding large digitisation projects at the Royal

Botanic Garden Edinburgh. In addition, the authors would like to thank all the staff who have worked on the digitisation projects at RBGE for their dedication and willingness to explore new methods of digitisation.

References

Beach J, Blum S, Donoghue M, Ford L, Guralnick R, Mares M, Thiers B, Westneat M, Wheeler Q, Wiegmann B, the Network Integrated Biocollection Alliance (2010) A Strategic Plan for Establishing a Network Integrated Biocollections Alliance. http://digbiocol.wordpress.com/brochure

Beaman RS, Cellinese N, Heidorn PB, Guo Y, Green AM, Thiers B (2006) HERBIS: Integrating digital imaging and label data capture for herbaria [Abstract]. Botany 2006. Botanical Cyberinfrastructure: Issues, Challenges, Opportunities, and Initiatives. California State University, Chico, 28 July – 2 August 2006.

Beaman R, Macklin JA, Donoghue MJ, and Hanken J (2007) Overcoming the digitization bottleneck in natural history collections. A summary report on a workshop held 7–9 September 2006 at Harvard University. http://www.etaxonomy.org/wiki/images/b/b3/Harvard_data_capture_wkshp_rpt_2006.pdf [Accessed March, 2011]

Berendsohn WG, Chavan V, Macklin JA (2010) Recommendations of the GBIF Task Group on the Global Strategy and Action Plan for the Mobilisation of Natural History Collections Data. Biodiversity Informatics 7: 1–5.

Berendsohn WG, Seltmann P (2010) Using geographical and taxonomic metadata to set priorities in specimen digitization. Biodiversity Informatics 7: 120–129.

Berents P, Hamer M, Chavan V (2010) Towards demand-driven publishing: approaches to the prioritization of digitization of natural history collection data. Biodiversity Informatics 7: 113–119.

Best JH, Moen WE, Neill AK (2009) A framework and workflow for extraction and parsing of herbarium specimen data [Abstract]. Proceedings of the Taxonomic Databases Working Group (TDWG). http://www.tdwg.org/proceedings/article/view/567 [accessed June 2011]

European Commission (2011) Digital Agenda: encouraging digitisation of EU culture to help boost growth. http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1292&format=HTML&aged=0&language=EN&guiLanguage=en

Granzow-de la Cerda Í, Beach JH (2010) Semi-automated workflows for acquiring specimen data from label images in herbarium collections. Taxon 59(6): 1830–1842.

Haston E, Cubey R, Harris DJ (2012) Data concepts and their relevance for data capture in large scale digitisation of biological collections. International Journal of Humanities and Arts Computing 6: 111–119. doi: 10.3366/ijhac.2012.0042

Heidorn PB, Wei Q (2008) Automatic metadata extraction from museum specimen labels. In: Greenberg J, Klas W (Eds) Metadata for semantic and social applications: proceedings of the International Conference on Dublin Core and Metadata Applications, Berlin, 22–26 September 2008: Berlin, Germany. Universitätsverlag Göttingen, Göttingen, 57–68.

Kroes N (2011) The European Commission Recommendation of 27/10/2011 [No. C(2011) 7579 final] on the digitisation and online accessibility of cultural material and digital preservation.

Lafferty D, Landrum LR (2009) SALIX, a semi-automatic label information extraction system using OCR [Abstract]. Botany & Mycology 2009, Snowbird, Utah, 25–29 July 2009. http://2009.botanyconference.org/engine/search/index.php?func=detail&aid=130 [accessed June 2011]

Llewellyn C (2011) Enhancing the curation of botanical data using text analysis tools. MSc thesis, University of Edinburgh, Edinburgh.

Niggeman E, De Decker J, Lévy M (2011) The new Renaissance – Report of the Comité des Sages. http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/final_report_cds.pdf

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

ZooKeys 209: 103–113 (2012)
doi: 10.3897/zookeys.209.3125
www.zookeys.org

RESEARCH ARTICLE

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

Increasing the efficiency of digitization workflows for herbarium specimens

Melissa Tulig[1], Nicole Tarnowsky[1], Michael Bevans[1], Anthony Kirchgessner[1], Barbara M. Thiers[1]

1 *William and Lynda Steere Herbarium, The New York Botanical Garden, Bronx, New York, USA*

Corresponding author: *Melissa Tulig* (mtulig@nybg.org)

Academic editor: *V. Blagoderov* | Received 26 March 2012 | Accepted 25 June 2012 | Published 20 July 2012

Citation: Tulig M, Tarnowsky N, Bevans M, Kirchgessner A, Thiers BM (2012) Increasing the efficiency of digitization workflows for herbarium specimens. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 103–113. doi: 10.3897/zookeys.209.3125

Abstract

The New York Botanical Garden Herbarium has been databasing and imaging its estimated 7.3 million plant specimens for the past 17 years. Due to the size of the collection, we have been selectively digitizing fundable subsets of specimens, making successive passes through the herbarium with each new grant. With this strategy, the average rate for databasing complete records has been 10 specimens per hour. With 1.3 million specimens databased, this effort has taken about 130,000 hours of staff time. At this rate, to complete the herbarium and digitize the remaining 6 million specimens, another 600,000 hours would be needed. Given the current biodiversity and economic crises, there is neither the time nor money to complete the collection at this rate.

Through a combination of grants over the last few years, The New York Botanical Garden has been testing new protocols and tactics for increasing the rate of digitization through combinations of data collaboration, field book digitization, partial data entry and imaging, and optical character recognition (OCR) of specimen images. With the launch of the National Science Foundation's new Advancing Digitization of Biological Collections program, we hope to move forward with larger, more efficient digitization projects, capturing data from larger portions of the herbarium at a fraction of the cost and time.

Keywords

Herbarium specimen digitization, workflows, georeferencing, digital imaging, field books

Introduction

The specimens in the world's museums and herbaria contain a wealth of primary occurrence data that is used as the basis of many biodiversity research studies (Chapman 2005; Baird 2010; Pyke and Ehrlich 2010). Historically, herbarium specimens have only been available to researchers by visiting collections or requesting specimens on loan. Over the past 20 years, efforts have been made to make specimen data available online through the development of specimen databasing and imaging projects. While millions of specimen records are now available through institutional portals and distributed networks such as GBIF, these only represent a small fraction of the estimated 90 million herbarium specimens in the United States alone that still need to be digitized (Rabeler and Macklin 2006).

The New York Botanical Garden Herbarium (NYBG) has been digitizing its collection of an estimated 7.3 million herbarium specimens since 1995. In the first fifteen years of digitization projects, we databased 1.3 million specimens at a rate of 10 specimens an hour, leaving 6 million specimens to database. Continuing at this rate, complete digitization of the herbarium would take another 600,000 hours. Like many institutions, past digitization projects at NYBG have focused on manageable and fundable subsets of the collection ranging from 75,000–100,000 specimens that could be completed within two to three years (Vollmar et al. 2010). For example, our collection of specimens from Brazil, estimated at half a million specimens, was broken into three National Science Foundation proposals and funded over 11 years. As a result, three separate passes were made through the herbarium to locate specimens from each region of Brazil. This was an inefficient but necessary way to find the relevant specimens and complete full specimen label data entry.

With more community support for digitization of natural history collections and new programs such as the National Science Foundation's Advancing Digitization of Biological Collections (ADBC), it is necessary to develop digitization protocols and workflows that maximize the rate of specimen digitization without sacrificing the most useful information on each specimen (Granzow-de la Cerda and Beach 2010; Scoble and Bourgoin 2010). Over the course of subsequent projects, NYBG has tried several methods to develop more efficient approaches to digitization, while still providing a high level of data quality to the scientific community who use these specimens.

Digitization workflows

Strategy 1: Manual data entry

Each project started with the curation of the taxa involved to reflect currently accepted names, based on recent monographs where available such as *Flora Neotropica*, on determinations by our curators and researchers visiting the herbarium, and on data available in online resources such as TROPICOS (http://tropicos.org/) and the International

Plant Names Index (http://ipni.org/). During the curation phase, specimens related to the project were separated from all others with which they were filed. They were subsequently removed from the herbarium and brought to a cataloguer's desk for data entry.

Barcodes were applied to the specimens and data entry was keyed manually from the specimen labels. Every piece of information on the label was entered, including the complete determination history of each specimen with determiners and dates. Collection information included collector, collection team, collector number and collection date. Site information included country, province or state, and county or municipio parsed separately, as well as the precise locality in a searchable text field, and geocoordinates when on the label. Habitat and plant descriptions were included word for word in text fields. Any additional notes on the label or on the sheet in general, or notes the cataloguer needed to add about the specimen, were put in other various notes fields. Authority files were also used for all taxa, and parties involved (collector, determiner, author), as well as drop down menus and look up lists for geography. Efficiencies used during this time focused primarily on organizing the specimens by collector before starting data entry to easily copy data from one record to the next. Simple measures such as encouraging cataloguers to use key strokes rather than the mouse and organizing the windows on their screen efficiently also improved data entry rates.

Staffing for these projects consisted of information managers to oversee data entry and imaging equipment, and curatorial assistants who databased and imaged the specimens. Information managers have a background in botany or biology, preferably with an emphasis in taxonomy, and several years of experience in data entry and database management. Curatorial assistants are typically new graduates in botany or biology with some herbarium experience but usually little data entry experience.

The data entry rate in this strategy averaged 10 records per hour. This rate is meant to represent an average for employing Strategy 1. It includes data entry rates from all of our major NSF projects that used this digitization approach, spanning all groups in the herbarium, and including rates of all curatorial assistants that catalogued on these projects. Only representative specimens were imaged, typically one or two per taxon.

Strategy 2: Streamlined collection events

For the third and last leg of our Brazilian NSF projects, Species of Amazonian Brazil, we were able to leverage field book data giving us an advantage over earlier databasing projects. In the late 1970's through the 1980's, the New York Botanical Garden was involved in a massive collection program of the Amazonian region of Brazil, called Projeta Flora Amazonica. We retained the original field books from most of the major collectors on this project, representing roughly 80% of the herbarium's total Brazilian Amazon holdings.

Botanists record collection data in their field books in large blocks of specimens, collected in the same site, on the same date. Often the only data different for each collection number is the taxon and plant description. Capitalizing on this, we were able to use a template tool in our database to mass enter the majority of the collection data from

each field book rapidly, entering each collection event only once instead of repeatedly for each collection number as we came across each specimen in the herbarium. This also allowed us to georeference the site only once and apply it to all of the collection events.

In addition, we collaborated with the Instituto Nacional de Pesquisas da Amazônia (INPA) who had already catalogued most of their specimens. Because many of our specimens are duplicated there, we imported a subset of collection events from their database, adding to the pre-load of data compiled from the field books. This added data for an additional 10% of NYBG holdings for Amazonian Brazil.

Data entry then proceeded as with previous projects. The specimens were curated, separated and removed from the herbarium for data entry from the specimen labels. With this pre-load of data from the field books and imports, the only information to add was the taxon and plant description, and the completion of fully catalogued records increased to 30 records per hour. At this stage the records were made available online.

Strategy 3: Semi-automated approach

With funding from the National Science Foundation's ADBC program, our digitization strategy shifted from entering complete specimen records to entering partial records with an image for every specimen. From this point, work will be done to complete these records by several means, focusing more on automated tools to extract data from the images and by entering data from the images rather than the specimens themselves. To keep up with this new demand for images, we also upgraded our imaging protocol, as outlined below.

As with previous projects we first curate the taxa involved. This continues to be a time consuming but necessary step of the process, ensuring that the data online and in the herbarium are current. Because ADBC grants fund larger digitization projects, the usual next step of separating out project specimens has been eliminated, as we are now digitizing complete sections of the herbarium at once. This enables us to pull entire folders from the herbarium without having to separate specimens within the folders, inspect each label and make the determination as to whether or not the specimen should be included.

Using a template tool in the database, we are able to rapidly mass create partial records by barcode number range. We auto-generate the number of records based on the number of specimens we have per taxon, at a rate of 125 records per hour. This barcoding process is done in the herbarium on a cart adjacent to the cabinet in which the specimens are housed. Once they are barcoded, they are tagged and returned to the cabinets until digitization staff sweep through the cabinets and image all the specimens. While this requires us to remove specimens from the cabinets twice, we use highly-trained curatorial assistants to curate the specimens and make decisions on the current nomenclature and part-time staff or interns to image the specimens. Each staff member can then work independently, but working in teams is another approach we plan to consider.

During image processing, all images are run through optical character recognition software (OCR) to produce a text output of the specimen label. The unparsed data is

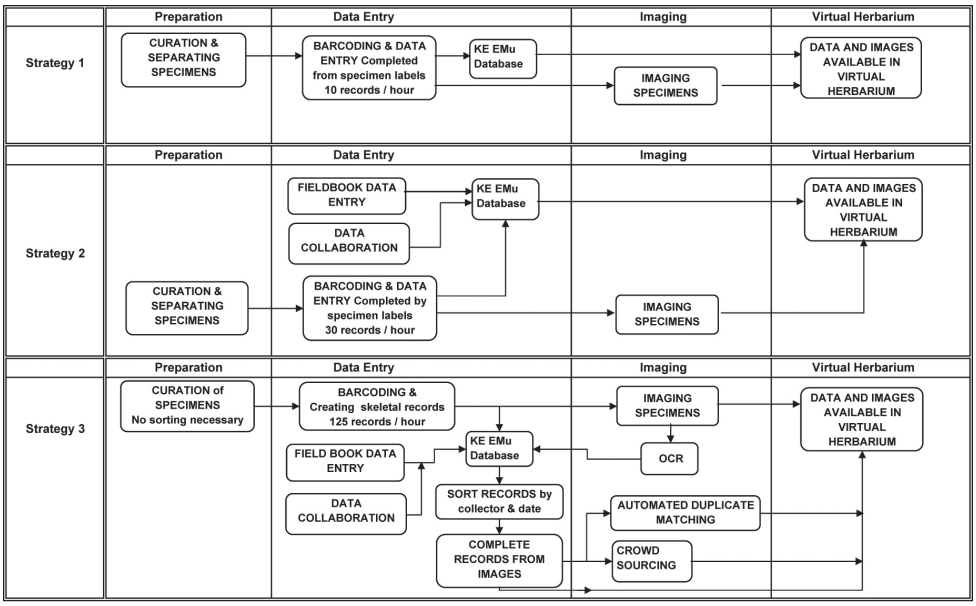


Figure 1. Digitization workflows at The New York Botanical Garden over the past 17 years.

then added to a fully-searchable text field in the specimen record. This will provide an initial way to search the records online until the records are completed and a mechanism for grouping records by collector or location during data entry. While we previously pre-sorted the physical specimens by collector before data entry, we will now attempt to pre-sort them via the OCR text. While not all labels contain typeface that will OCR, many that are partially handwritten have at least some typed "master label" information including collector name and some locality information, since a large portion of our herbarium was collected in the 20th century.

For projects where we have a large collection of field books, such as our NSF-funded Caribbean Project, we will continue to pre-load collection events records into the database. For all other records, we will parse the OCR text using automated tools in development such as Salix, the semi-automatic label information extraction system being developed at Arizonia State University (http://nhc.asu.edu/vpherbarium/canotia/SALIX3.pdf) and Apiary (http://www.apiaryproject.org/high-throughput-workflow-computer-assisted-human-parsing-biological-specimen-label-data). We will also use duplicate matching applications such as FilteredPush (Wang et al. 2009) and Specify's Scatter, Gather Reconcile (http://specifysoftware.org/content/specify-64). The end result will be records with the most pertinent data fully searchable in the database, including collector, collection number, date, current taxonomic name, and complete locality information. Since we are now taking an image of every specimen, any secondary data will still be available in the image of the label, or in the OCR text of the label that will still be available in a notes field. This includes plant description, habitat, and other notes found on the specimen.

Specimen imaging

Imaging equipment

To accommodate the image production expectations of the rapid digitization grants, several low cost imaging stations built with commercially available digital photography components were assembled. These components include: the Canon Eos 5D Mark II digital camera body, a Canon EF 50mm f/2.5 Macro lens, the Photo e-Box Plus 1419 from MK Direct, a Kaiser RS 1 copystand, and a Wasp bar code reader, and a laptop computer (Figure 2).

The Canon Eos 5d Mark II camera was selected to meet the image size requirements of 21 megapixels. With a resolution of 5616 × 3744 pixels, or 21.1 megapixels, the images can be enlarged on screen up to 78 × 52", which is roughly four times the size of the original specimen sheet. The lens used is a macro lens with a normal focal length that produces little or no edge distortion. This is optimal not only for scientific study but may also produce better results when read by optical character recognition (OCR) software.

The Kaiser copystand supports the camera so that the focal plane is 31" above the specimen. This provides for a full frame image with a quarter inch border on three sides and a one inch border on the top of the specimen. A metric scale and a Munsell color target are placed in the one inch border along the top edge of the specimen.

Specimens are illuminated by placing them inside the MK Direct Photo e-Box 1419 lightbox. 5000 Kelvin fluorescent lights provide even illumination across the entire surface of the specimen with minimal heat. Supplemental 5500 Kelvin LED lighting is used to accentuate the appearance of the surface texture of the specimen.

The imaging equipment used at The New York Botanical Garden Herbarium has now become standard equipment for the National Science Foundation's Advancing Digitization of Biological Collections project, Plants, Herbivores and Parasitoids. Identical camera work stations are being used at a number of partner institutions.

Imaging workflow

Digitizers gather the barcoded and cataloged specimens in the herbarium then transport them to the imaging station via herbarium cart. The lightbox is powered on and allowed several minutes for the lights to stabilize and the computer and camera are powered on and the camera software is started.

A specimen is placed in the lightbox. To ensure correct alignment, a template specimen sheet is affixed to the shooting surface. The digitizer aligns the specimen with the template and shuts the front panel of the lightbox. Once the specimen is placed in the lightbox the digitizer presses the shutter release button in the camera software, taking the exposure.

The camera settings are as follows: 5000 Kelvin white balance, ISO 100, 1/60th of a second shutter speed at f/9.0. To streamline image quality control and post produc-

Figure 2. NYBG imaging station consisting of a Canon Eos 5D Mark II digital camera body, a Canon EF 50mm f/2.5 Macro lens, Photo e-Box Plus 1419 from MK Direct, and Kaiser RS 1 copystand.

tion, all imaging workstations are configured identically in order to produce consistent images. The white balance of individual cameras may be manually modified to account for subtle differences in the color temperature of the lights.

The first image recorded is opened and inspected to confirm focus, exposure and color balance. Subsequent images are inspected periodically. Once each image is recorded, the digitizers rename the image files by scanning the barcodes on the speci-

mens with the barcode reader. Using a rubber stamp, photographed specimens are stamped with the word "Imaged" to avoid unnecessary reimaging in the future.

The current average imaging rate is 85 exposures per hour. This means that a full time, dedicated digitizer imaging for a full 150 hours per month, could produce well over 12,000 images per month. Each image file is approximately 25 megabytes for a total of over 300 gigabytes of data monthly.

Image quality control

Digital camera images from each imaging station are recorded in a master imaging log and the files are transferred via external hard drive to a central image quality control work station. Image quality control is performed on a single workstation with a monitor calibrated using the Xrite i1 calibrator to ensure optimal viewing. Image files are viewed and modified using Adobe Lightroom.

Image thumbnails are visually scanned en masse to confirm that the image orientation is correct and to identify any obvious defects. Periodic images are magnified to 100% magnification in order to confirm focus and that the barcode on the specimen matches the file name. Roughly every twentieth image is examined.

The image files contain technical Exchangeable Image File Format (EXIF) metadata. Additional International Press Telecommunications Council (IPTC) metadata, including Creator, Image Title, and Copyright information, is added to the image files en masse using Adobe Lightroom's Library module.

Image processing

Once quality control is assured, the camera files are enhanced for viewing using Adobe Lightroom's image editing adjustment tools. One image representative of one shooting session per camera workstation is selected and modified and the modifications are applied to all other images recorded in that session.

A more precise white balance is performed by sampling the white reference on the Munsell color target included in the image. The tonality is adjusted so that the color reference target values meet manufacturer's specification ensuring proper exposure. Sharpening is applied to enhance detail. Chromatic aberration caused by the lens is removed. For complete examples with screen shots refer to Image Editing Guidelines (http://tinyurl.com/764z7wx).

Archive

Once processed, the proprietary Canon digital camera files are converted to Adobe's DNG format and copied to an archive server. Each DNG file is approximately 25 megabytes.

Tape backups are automatically made of all new files on the server. Additionally, a complete tape backup of the entire archive takes place every six months and the tapes are stored off-site.

Access

Once saved as DNG and archived, specimen images are saved as full size, 5616 × 3744 pixel jpegs using the sRGB color space. Each jpeg is approximately 8 megabytes.

The jpegs are imported into the database where the barcode file name is matched to the corresponding catalog records and the images are made publicly available online immediately.

Optical character recognition

The New York Botanical Garden Herbarium uses ABBYY FineReader optical character recognition software to produce text files from specimen labels. An Adobe Photoshop Action (macro) is used to automatically reduce the file size of the specimen images. Each access image is cropped in half (label data is usually found on the lower half of a specimen sheet) and converted to grayscale. This reduces the file size of each specimen to less than one megabyte. The resulting grayscale jpegs are processed using ABBYY FineReader and a separate text file for each image is saved.

The temporary grayscale images and the resulting OCR text files are returned to the catalogers. Viewing the grayscale images reduces the time required to open large files, allowing the cataloger to quickly verify the OCR text which is then manually parsed into the correct database fields. In the event that the label data is not included in the cropped area, the image may be retrieved from the database and the label data can be transcribed manually. After parsing the OCR text, the grayscale images are discarded.

OCR text to database

A Powershell script is run to extract the data from each saved text file. The script opens Microsoft Excel and inserts a new row for each file, adding the barcode (which is read from the file name) and the label text. Since the barcode number is also part of the text itself, a comparison of the file name barcode and the text barcode can be made to reveal errors in either the file naming procedure or the OCR process.

Once the data are in Excel, they can be directly imported into the database to a searchable notes field. Rows in Excel can be grouped according to common textual information, such as a collector's name or an expedition title. This step allows other fields in the database to be filled when the label text is imported.

Discussion

As a result of new databasing strategies, the rate of adding specimen records to the database has gone from 10 complete records per hour to 125 partial records per hour. The resulting records have limited parsed label data initially, but are all imaged, available online immediately, and indexed by scientific name. The records will then be completed over time using the specimen image instead of the specimen itself. The result will be an index of all of our holdings for large portions of the herbarium, and eventually, for all 7.3 million specimens. It is important to note that none of these rates take into account the time put in by information management staff who oversee and train curatorial and digitization staff, import and clean database and authority files, install and troubleshoot camera equipment, process and archive images, and manage server and database upgrades.

With relatively high error rates still facing OCR and automated parsing of label data, a shift to more automated approaches has the potential to reduce the quality of information we typically provide. We feel the best first approach to complete partial records is to use database templates to mass ingest repetitive data from collector's field books for specimens deposited at NY. For some projects, we are fortunate to have the field books for the majority of the collections. This model has the potential to be useful for a wider audience in conjunction with projects like the Smithsonian's Field Book Project (http://www.mnh.si.edu/rc/fieldbooks/), which is creating an online index of these resources. Next, using duplicate matching applications such as FilteredPush and Specify's Scatter, Gather, Reconcile to search for records already fully databased by other institutions ensure that we complete the partial records with quality information.

We will then rely on automated techniques to complete the remaining partial records from the OCR text by such applications as SALIX or APIARY. It is very likely that none of these techniques will work for the completion of all labels, especially handwritten ones. Manual transcription of data will still be necessary to complete such labels. Some of this manual transcription will be done by project staff, but we also hope to enlist volunteers, especially citizen scientists with a particular interest in using these data for their own activities or research, or as a leisure activity, to help complete the records using a crowd sourcing website that we will develop for this purpose. By combining all of these approaches, we hope to rapidly catalogue the majority of the herbarium with quality information and make these records available for other institutions to download or for use in biodiversity studies.

Conclusion

The New York Botanical Garden Herbarium's cataloging and imaging procedures have evolved to the point that the limiting factor in digitization is no longer technology but manpower. As we work towards our goal of digitizing the approximately 6 million specimens remaining, we hope to continue to increase our rates and learn from new developments in the biodiversity informatics community. To supplement our efforts

The New York Botanical Garden is enlisting volunteers and citizen scientists whenever possible. While we can look forward to even greater advances in imaging technology, optical character recognition software, improved databasing and barcoding technologies, ensuring accurate data relies on well trained staff and an institutional commitment to the future growth of digital collections.

References

Baird R (2010) Leveraging the fullest potential of scientific collections through digitization. Biodiversity Informatics 7: 130–136.

Chapman AD (2005) Uses of Primary Species-Occurrence Data, version 1.0. Copenhagen: Global Biodiversity Information Facility. 106 pp. ISBN: 87-92020-01-1. Accessible at http://www2.gbif.org/Uses.pdf [March 22, 2012]

Granzow-de la Cerda I, Beach JH (2010) Acquiring specimen data from herbarium labels. Taxon 59 (6) 1830–1842.

Pyke GH, Ehrlich PR (2010) Biological collections and ecological/environmental research: a review, some observations and a look to the future. Biological Reviews 85: 247–266. doi: 10.1111/j.1469-185X.2009.00098.x

Rabeler RK, Macklin JA (2006) Herbarium networks: towards creating a 'toolkit' to advance specimen data capture. Collection Forum 21: 223–231.

Scoble MJ, T Bourgoin (2010) Natural history collections digitization: rationale and value. Biodiversity Informatics 7: 77–80.

Specify 6.4: Scatter Gather Reconcile (SGR) (2012) Specify Software Project. 18 June 2012. http://specifysoftware.org/content/specify-64

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

Wang Z, Dong H, Kelly M, Macklin JA, Morris PJ, Morris RA (2009) Filtered-Push: A Map-Reduce Platform for Collaborative Taxonomic Data Management. 2009. WRI World Congress on Computer Science and Information Engineering 3: 731–735.

ZooKeys 209: 115–132 (2012)
doi: 10.3897/zookeys.209.3083
www.zookeys.org

RESEARCH ARTICLE

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

Results and insights from the NCSU Insect Museum GigaPan project

Matthew A. Bertone[1], Robert L. Blinn[1], Tanner M. Stanfield[1], Kelly J. Dew[1], Katja C. Seltmann[2], Andrew R. Deans[1]

1 *Department of Entomology, North Carolina State University, Campus Box 7613, Raleigh, NC, 27695-7613, USA* **2** *American Museum of Natural History, Central Park West at 79th St., New York, NY 10024-5192 USA*

Corresponding author: *Matthew A. Bertone* (matthew.bertone@gmail.com)

Academic editor: *V. Blagoderov* | Received 14 March 2012 | Accepted 14 June 2012 | Published 20 July 2012

Citation: Bertone MA, Blinn RL, Stanfield TM, Dew KJ, Seltmann KC, Deans AR (2012) Results and insights from the NCSU Insect Museum GigaPan project. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 115–132. doi: 10.3897/zookeys.209.3083

Abstract

Pinned insect specimens stored in museum collections are a fragile and valuable resource for entomological research. As such, they are usually kept away from viewing by the public and hard to access by experts. Here we present a method for mass imaging insect specimens, using GigaPan technology to achieve highly explorable, many-megapixel panoramas of insect museum drawers. We discuss the advantages and limitations of the system, and describe future avenues of collections research using this technology.

Keywords

panorama, gigapixel, megapixel, specimen, collection, imaging system

Introduction

Insect specimens are integral to basic entomological research such as systematics, ecology, and applied sciences. However, most are preserved dried on pins and stored in large collections, where they remain difficult to physically access (e.g., requiring permissions and/or expensive travel). This situation leads to a massive underutilization of specimens and their associated data. While the process of physically sending (i.e., loaning) materials alleviates the need to travel to collections, it is time consuming for

collection managers and difficult for the borrower to specify which individuals are needed without knowledge of the true holdings (e.g., requesting from series of undetermined specimens). More importantly, whenever specimens are removed from their drawers they are at risk of being exposed to unfavorable conditions, including handling by untrained users, losses during transit or being misplaced, and insufficient temporary curatorial practices.

It is essential for insect collections to have a web presence and disseminate information online. Online databases of public and private collections are common practice, and usually include specimen names and taxonomic status, number of individuals of each taxon, and data from labels (such as localities, dates and other information regarding the specimen's provenance). Some collections even host images of their materials, though it is usually limited to a few photographs of exemplars or valuable specimens (e.g., types). Despite these advances, very few avenues exist to thoroughly browse the holdings of any one collection, visually, and to evaluate the extent/quality of its specimens and the degree to which they are curated.

GigaPan (www.gigapan.com) was initially developed through a collaboration between Carnegie Mellon University and the NASA Ames Intelligent Robotics Group for use on NASA's Mars Rovers (Spirit and Opportunity). It has since become a commercially available hardware and software, used to achieve many-megapixel to gigapixel (i.e., billions of pixels) images that are then represented as highly-navigable panoramas. The basic product consists of a robot that can be fitted with any digital camera (depending on camera and robot model) and mounted on typical tripod threads. Once initiated, the robot positions the camera to frame individual images across a designated area of interest and uses a robotic "finger" (or remote release) to engage the camera, which captures multiple, overlapping tiles (i.e., photos). GigaPan software is then used to stitch the resulting photos into one large panorama that has a maximum resolution roughly matching the resolution of each individual image, but across a much larger area. Further, panoramas currently can be hosted on the GigaPan website where viewers may add general comments and take snapshots of specific areas, either with annotations describing the importance of the area or questions about it. Though commonly used for capturing vast landscapes and large events, the potential of these panoramas is far reaching.

With about 1.5 million specimens, the North Carolina State University Insect Museum (http://insectmuseum.org) is the largest insect collection in North Carolina, and among the largest in the southeastern United States. The pinned collection is strong in several groups, including Hemiptera (bugs, especially Auchenorrhyncha, the holdings of which are world-renowned), Anthophila (bees, especially Megachilidae), and Pyralidae (snout moths). At a moderate size, the NCSU Insect Museum presents an important, but manageable, resource for understanding modern digitization potential of insect collections. Here we present results and insights gained from our efforts to image whole drawers using GigaPan technology. We provide details on how to achieve similar results, describe the advantages and drawbacks of the system, and discuss outcomes of the project.

Methods

Existing infrastructure

The NCSU Insect Museum has roughly 2,700 insect drawers in use, stored in 184 12- or 24-drawer metal cabinets. Drawers are U.S. National Museum (USNM) style, with the following dimensions: 45.72cm W × 45.72cm D × 7.3cm H (18"W × 18"D × 2–7/8"H; outer measurements) and 41.28cm W × 42.55cm D × 5.87cm H (16–1/4"W × 16–3/4"D × 2–5/16"H; inner measurements).

Equipment

We employed a GigaPan EPIC 100 ("silver model"), oriented horizontally on a copy stand and paired with a Canon PowerShot G11 camera. We retrofitted the GigaPan with an A/C adapter (Sargent et al. 2010) and bought a commercial A/C adapter for the Canon to alleviate the need for disposable batteries and/or charging requirements. Our lighting needs were satisfied by dual Interfit Super Cool-Lite 9 lights, each with nine 28W compact-fluorescent bulbs that produce continuous daylight spectrum (5000–5500K). Both lights were equipped with the included diffusion covers for softer lighting. Other diffused lights delivering this spectrum would be suitable. Most of the stitching was performed on an Intel i7 quad core Apple iMac (2.8 GHz, 4,096 GB RAM). The complete imaging station (without the computer) is illustrated in Figure 1.

Settings

Camera settings were based largely on those described in the GigaPan tutorials (http://gigapan.org/cms/videos) and manual (Gigapan Systems 2010), with the white balance set to daylight fluorescent (best balance for the lighting described above) and the field of view (FOV) for the camera set to 11.5º on the GigaPan unit. The FOV is dependent on the camera model, so this number is specific to the Canon PowerShot G11. The aperture was set to f/8.0 (the smallest available for the camera) to achieve the greatest depth of field (DOF; 3.5cm). The distance of the GigaPan robot plus camera was set to about 46.35 cm (18.25") from the base of the copy stand [about 43.2cm (17") above the average pinned specimen]. This height is beneficial for optimizing the DOF, quality, and size of the images at full optical zoom, while reducing curvature (see Results) and keeping the number of photos (~35 per drawer) manageable with respect to time and storage capabilities. All images were shot as large, super-fine quality JPEGs (3,648 × 2,736 pixels). The focus was locked to prevent the variable amount of time needed for the auto focus, which could result in the camera not completing the process before the robot moves to the next position. A custom timer delay of 2 sec was also added to ensure the unit was stable during photo capture. In conjunction, the "Time per Pic" on the robot was set to 4.5 sec, so movement

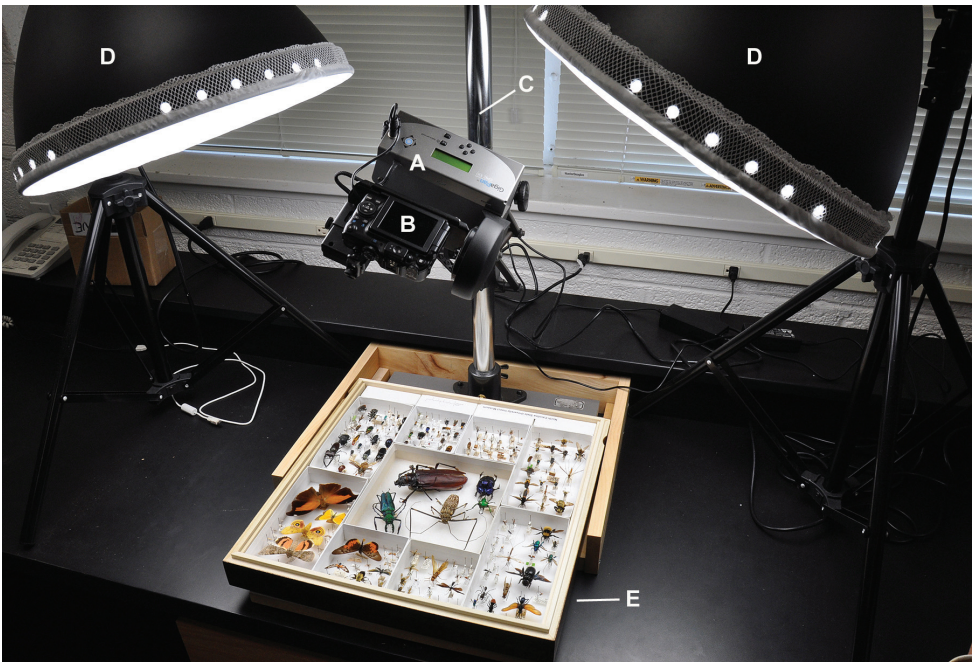


Figure 1. The complete imaging station. **A** GigaPan Epic 100 ("silver model") robot **B** Canon PowerShot G11 camera **C** copy stand **D** light with continuous compact fluorescent bulbs **E** insect drawer.

would not occur during capture. All settings were saved in one of the two custom settings slots (C1 or C2) available on the Canon G11 for recall when the camera is turned on.

Imaging workflow

Drawers were placed within the confines of a custom jig on the copy stand, with the lid removed. To prevent white space from interfering with the camera's ability to focus (an issue sometimes encountered, despite locking the focus), a Kodak Tiffen Color Separation Guide (ASIN: B00009R7G9; trimmed to fit inside a unit tray) and printed matter were placed inside empty unit trays (Fig. 2). Initially, the "New Panorama" process was begun on the GigaPan robot to define the boundaries of the drawer to be captured by the camera, and verify that the camera settings were in place and correct. After the initial setup, the Epic 100 was engaged using the "Last Panorama" function, unless the image area needed to be modified. While the robot and camera were working drawer preparation occurred for the next one in line, reducing the overall amount of time needed. After capturing all images on the camera's memory card, each completed insect drawer was given a label with the date the panorama was taken and returned to the collection. Photos for each panorama (usually n=35) were delivered manually onto a computer hard drive or external hard drive (through the computer) directly from the camera using a USB cable; using a cable bypassed the need to remove the camera and

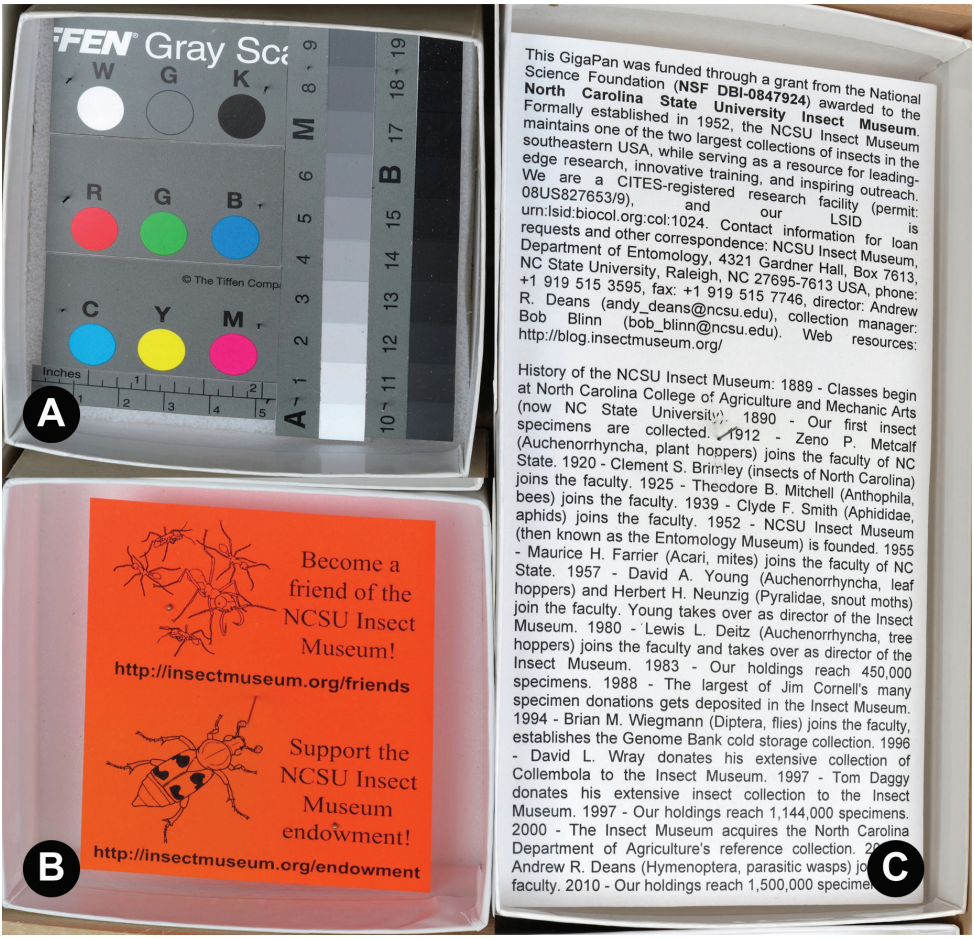


Figure 2. Color standards and white space filler. **A** Kodak Tiffen Color Separation Guide **B, C** text/picture white space filler.

memory card, potentially moving the unit from its set positions (required for using "Last Panorama" function properly). All photos were checked during/after transfer for errors, especially out of focus images, and reshot if necessary. Stitching was then initiated manually on the computer by opening the drawer images, previously transferred from the camera, in the GigaPan Stitch software (version 1.0.0804; provided by GigaPan); stitching was done either singly or as a batch of multiple drawers (10–20 at a time). Batches were possible by opening any existing .gigapan file in the stitch software and using the "New Gigapan" function (File > New Gigapan) to select the new set of photos to stitch; repeating this process resulted in multiple stitch windows open concurrently on the computer. All panoramas were checked during the preview phase of the stitching to ensure that no errors existed, most frequently misaligned tiles. If a re-stitch did not work the drawer was reshot. Finally, stitched sets were stored locally, backed up by external hard drives, and uploaded to the GigaPan website (either singly

or as batches in the same manner as described above for stitching). During uploads, each panorama was given a brief description and several keywords (usually standard words like "insect" and "museum", the order, and families present in each drawer). Throughout the entire process, custom paperwork was used to record all drawers being imaged and the status of their progression. Also, to ensure that the lights did not overheat a cool-down time of 5–10 minutes was added after shooting about 10–15 panoramas. A schematic of the entire workflow can be seen in Figure 3 and a video tutorial can be found at http://purl.oclc.org/NET/NCSU/gigapanvid.

Results

General

As of March 1, 2012, the NCSU Insect Museum had 2,124 panoramas uploaded (http://gigapan.org/profiles/ncsuinsectmuseum), or about 79% of the ~2,700 drawers. Figure 4 illustrates typical drawers, while Figure 5 shows a specialty drawer that was assembled to show insects by theme (in this case the diversity of the four largest insect orders). Final panoramas averaged about 208 megapixels in size (14,700 × 14,150 pixels).

Time to drawer completion

Average time for completing a drawer – from inserting color standards and text (not including time needed to initially create space in each drawer) through stitching and uploading – was from 12–50+ minutes. Each step required the following amount of time (single or batch; process further described in Fig. 3): drawer prep and filler placement - ~2 mins; image capture - ~4.5 mins; data transfer - ~1–3 mins (batch of 10–15); stitching images - ~3–14 mins (batch of 10–20); uploading - ~1.5+ mins (batch of 10–20). These figures were generalized over the entire life of the project, and using the latest versions of the stitch/upload software while opening multiple stitch/upload windows (described above in Methods) greatly reduced time needed to create and make public the panoramas; future, faster versions of the software should reduce these times even further. Other variables also exist that affect speed, including CPU processing power and internet connectivity (e.g., wireless vs. hard-wired connection speeds, the former usually resulting in slower uploads). Overall these figures represent a conservative estimate of 25 mins to complete each drawer.

Data storage requirements

About 150MB (typical range: 140–165MB) of storage space was required for each drawer's complete panorama data (including original photos, raw tile data, and

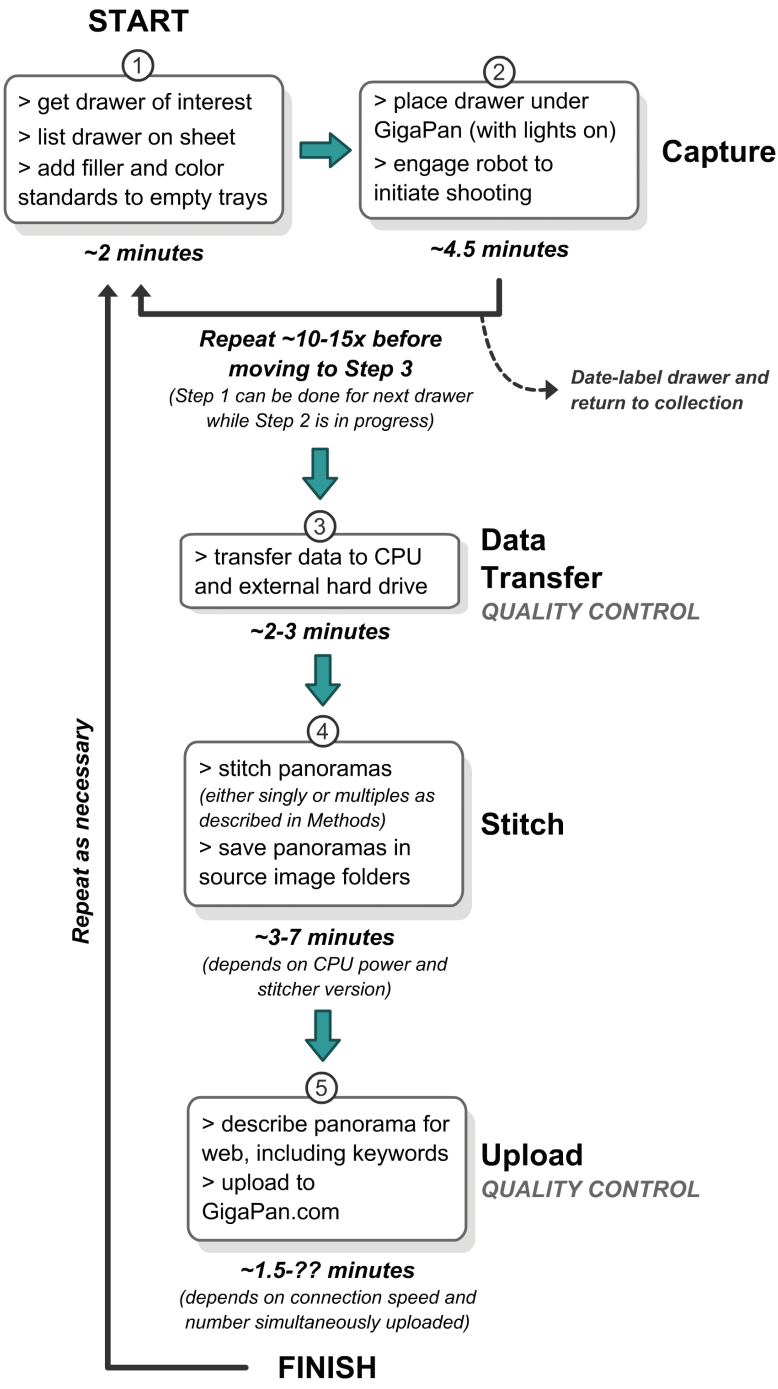


Figure 3. Schematic of project workflow. Note: times are rough estimates and prone to change depending on the efficiency of several steps.

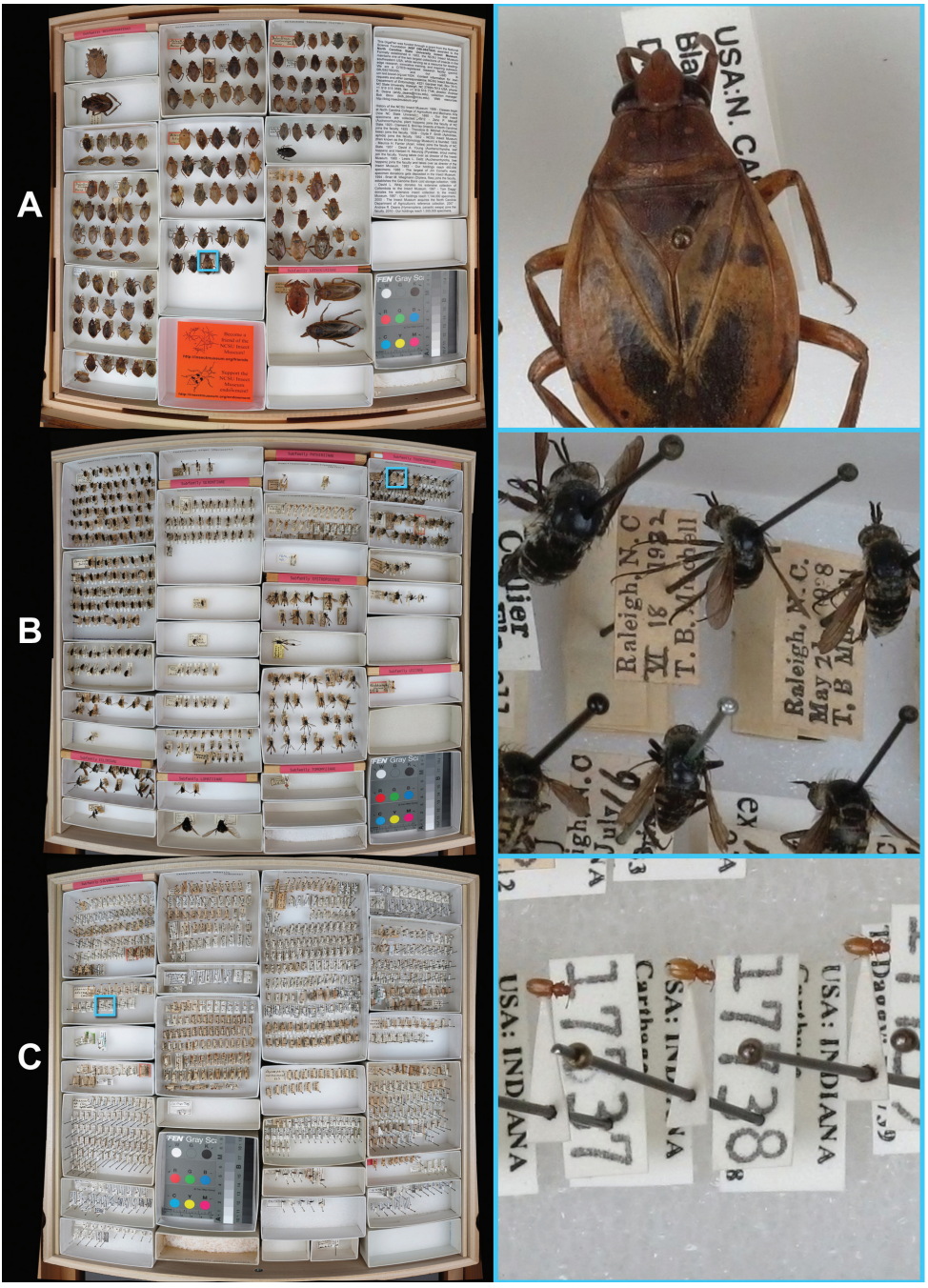


Figure 4. Examples of typical drawers, showing larger specimens, average specimens, and smaller specimens (**A, B & C**, respectively). Left – full drawer image; Right – zoomed to full resolution. **A** Belostomatidae 1 (http://gigapan.org/gigapans/96136) **B** Bombyliidae 5 (http://gigapan.org/gigapans/89195) **C** Silvanidae 2 (http://gigapan.org/gigapans/95947)

Figure 5. Example of a thematic drawer displaying the diversity of the four largest insect orders (http://gigapan.org/gigapans/49310). Clockwise from Top Left: Hymenoptera (wasps, ants & bees), Lepidoptera (moths & butterflies), Coleoptera (beetles), and Diptera (true flies). The drawer also serves as an outreach tool by containing some mistakes for people to identify and further learn the differences between the orders.

gigapan panorama file). Thus, for the entire 2,700 drawer collection, ~405 gigabytes of storage space was needed. These figures are based on JPEG images with an average size of 1.8–2.6MB each (resulting from size/resolution settings described in Methods).

Panorama quality

Panorama qualities, including resolution and distortion, were measured using a test drawer and the resulting panorama (Fig. 6). As expected, curvature/distortion (see Discussion) was found to be greatest near the edges of the drawers, i.e. furthest from the

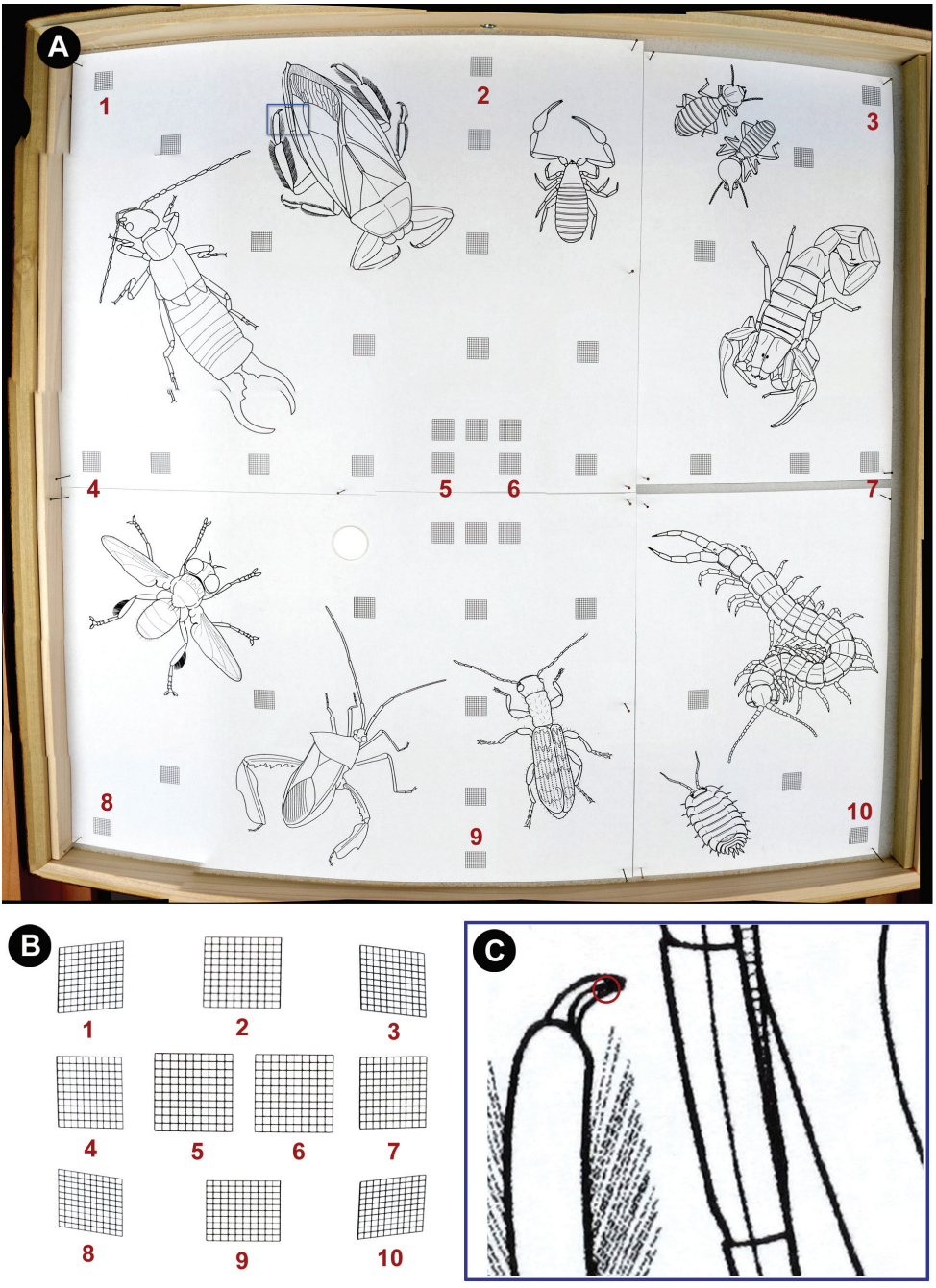


Figure 6. Panorama measures of distortion and resolution. **A** drawer with illustrations and 1cm x 1cm (1mm subunit) grids spanning the panorama **B** comparison of distortion produced across the top (1, 2, & 3), middle (4, 5, & 6), and bottom (7, 8, & 9) of drawer in A **C** smallest resolvable difference between black and white (~80μm) at 1:1 magnification (from blue rectangle in **A**).

center. Specifically, there was a 20% reduction of all lengths measured from the corners and sides of the panorama (1, 3, 4, 7, 8, & 10 in Fig. 6B), and a 20% reduction of the vertical measurements at the top and bottom positions (horizontal measurements of top and bottom appear unaffected; 2 & 9 in Fig. 6B). Further, some skewing of measurements occurred, especially at the corners of the panorama, resulting in distorted areas (see 1, 3, 8, & 10 in Fig. 6B). As for resolution, the smallest resolvable structure on a fully-zoomed panorama (discernible white space between two black spaces; Fig. 6C) measured about 80μm; thus structures smaller than this may not be discernible using the current camera optics and settings.

Online metrics

Panoramas on the NCSU Insect Museum profile at Gigapan.com (n=2,124) have been viewed a total of 326,252 times, at an average of 153.6 views and a median of 94 views. We do not have data on the percentage of unique visitors. The award-winning, specialty drawer "The Big Four" has the most views for a single panorama (24,054 as of the date above), largely resulting from widespread attention gained from GigaPan and media covering the panorama contest during the first meeting of the Fine International Conference on Gigapixel Imagery for Science (http://www.cmu.edu/news/archive/2010/September/sept30_gigapixelshow.shtml). Eighteen drawers have over 1,000 views, including both special panoramas and typical museum drawers.

Discussion

This project represents the largest and most complete effort to image and publicly-share an entire insect collection, with over 2,000 drawer panoramas available. The panoramas have been viewed many thousands of times and interactions with both experts and laypeople have occurred. While the project is not yet complete, several outcomes have materialized from the effort.

Unsolicited, remote curation has happened. Word of our insect drawer images spread quickly among insect systematists and we rapidly received communications that enhanced our holdings. In one instance, a taxonomist at a natural history museum in Ottawa, ON (837 miles north of the NCSU Insect Museum) determined a series of froghopper (Hemiptera: Cercopidae) specimens to species from an "unsorted insects" drawer (http://gigapan.org/gigapans/41421/snapshots/120403/). Along the same lines, a world bumble bee (Hymenoptera: Apidae: *Bombus*) expert provided a species name for an undetermined specimen (http://gigapan.org/gigapans/49310/snapshots/139687), and a lanternfly (Hemiptera: Fulgoridae) expert determined several specimens to species. Further, a velvet ant (Hymenoptera: Mutillidae) specialist identified several specimens (http://gigapan.org/gigapans/60116), provided new

information on the taxonomic status (synonymies) of several species, and helped resolve the identity of a wasp that had become decoupled from its pin. All interactions were communicated between coordinating members of the museum, and steps were taken to update the collection based on input from the interaction. Additionally, the project has enabled more informed donations: a world expert has contacted us to say she is using our GigaPan images to better understand our current holdings, so that she can then divide up her personal collection between natural history museums more efficiently. She wants to maximize the taxonomic coverage of her donation to our museum.

We also successfully reached out and engaged the public using these panoramas. For example, non-entomologists commented on artistic representation (http://gigapan.org/snapshots/119341/comments), made humorous comments about the insect specimens (http://gigapan.org/snapshots/117944/comments), and asked questions about insect biology (http://gigapan.org/snapshots/147239/comments). The creation and promotion of more thematic drawers, for example teaching concepts using the panoramas (as in Fig. 5) or testing knowledge using Easter eggs and treasure hunts, could easily draw more attention from the public and contribute to our mission for increased outreach, all resulting in added interest in our science.

During the project, several unanticipated outcomes occurred. One was the linking of specimen snapshots to panoramas of their locality/habitat (based on label information). Unsolicited, another member of the GigaPan community and part of the Fine Outreach for Science group, took a panorama of the cloud forest habitat in Costa Rica where one of our leafhopper specimens was collected, and linked it through a snapshot (http://gigapan.org/snapshots/127411/comments). The practical applications of these data are plentiful, including using the panorama of the habitat to estimate plant diversity related to insect specimens, or change in habitat over time. Researchers could use a GigaPan at their collecting sites in order to understand the temporal and spatial biodiversity, and further enrich the information available for the specimens taken at the site. Another potential product we had not considered, but were encouraged to contribute data for, was a 3D panorama (our example can be seen here: http://www.3d-360.com/). These are achieved by shooting two panoramas of the same drawer at slightly different angles (i.e., positioning the drawer slightly to the left or right of center to capture different perspectives). Then independent, proprietary software is used to make the panorama visible in three dimensions, either using anaglyph glasses (red/cyan) or through other methods (e.g., cross-eyed viewing, etc.). Lastly, we used GigaPan to enhance the insect collection project for the NCSU ENT 502 graduate-level course, Insect Biodiversity and Evolution, by creating panoramas of the final collections submitted by several graduate students (http://gigapan.org/gigapans?order=most_popular&page=1&per_page=10&query=ent+502). The resulting panoramas effectively archived the students' projects, either to remind them of their efforts or to guide future students making collections. We anticipate that the ease and adaptability of GigaPan will encourage even more creative applications of the technology to collection science.

Workflow improvement

Project workflow varied little after initial setup and achieving the present results. Though we did not objectively and iteratively evaluate the process along the way, several observations were made based on user experience. During drawer imaging there is down time, even when using that time to prepare the next drawer (see Fig. 3). One option for taking advantage of this time might be the incorporation of a second system, so that two drawers could be imaged in a partly overlapping time frame. Employing additional people to capture the images would not be more efficient (unless more than two systems are used at once), though having one person image the panoramas and another person stitch them after each batch reduces time. Another step that could be streamlined is data transfer, which could be done wirelessly if such technologies were incorporated (for example a wireless memory cards for the camera; http://www.eye.fi/). Additionally, upgrading the entire system to use a Digital SLR would enable options for wireless file transfer, but at a greater total cost (in addition to the cost described below in *Advantages of GigaPan*). However, the small amount of time saved may not be economically worth it. An automated batch stitch and upload could be initiated overnight to save man hours, though software for doing so is not yet available. The only drawback would be the inability to identify and correct errors in the batch process until after time has been spent stitching the panoramas (as noted in Blagoderov et al. 2010).

There is a need to formulate objective ways to evaluate the quality of the panoramas, from aesthetics like resolution, exposure and clarity, to more scientific criteria such as the potential for identifications and the amount of data that can be observed in the drawers (e.g., from labels). Furthermore, errors, such as those encountered during capture and stitching (usually involving out of focus images and misaligned tiles, respectively), were usually identified before uploading, but some subtle ones still exist in panoramas present online. To rectify the situation it will be beneficial to identify the visual clarity of the panoramas and any persisting errors; crowd sourcing the panoramas to determine these quality metrics could help to expedite the process.

General issues for mass imaging insect drawers

Imaging entire insect drawers with any system has its drawbacks. The following were identified by the authors early on, and reiterated in responses on a survey of the utility of the drawer panoramas for research (Hammond MS Thesis *in prep*).

Panoramas of pinned specimens tend to show only some angles of the insects; dorsal and some lateral aspects are usually visible, but ventral views are generally obscured. Limiting the observable amount of a specimen limits the power of these images for determining some species, especially ones where diagnostic characters are located in obscured areas. Lack of good image resolution and magnification associated with ordinary camera optics also hinders identification, especially for smaller specimens. Though higher magnification and resolution can be obtained for these panoramas, it

usually involves taking more photos of each drawer (increasing time needed for the entire project) and purchasing special lenses that are often expensive and not always available for the system being used. Another result of a single overhead panorama is that larger specimens can hide labels, further reducing the amount of information available to viewers.

Collections are consistently being updated and curated, thus many panoramas derived from such a project will become out dated at different rates and not fully represent the current state of the collection. This occurs as specimens are added to and moved around the collection, rendering the drawer images inaccurate, especially in active sections of the collection. As such, we consider these panoramas to be "snapshots" of each drawer at the time of imaging, and we provide a date on each drawer after the initial capture to hopefully aid in future evaluations of the true level of change (or stasis) for each drawer. A method for labeling the level of curation on each drawer post-panorama (e.g., number of specimens added or taken from each drawer) would help to determine which drawer images need to be updated, though such a system is not yet fully formulated and could be complicated to implement and enforce.

Advantages of GigaPan

Using GigaPan technology for drawer imaging is ideal in a number of ways. The entire system described here cost approximately $1,500 (US):

- GigaPan Epic 100 (~$450)
- Canon G11 (~$500)
- lighting (~$500)
- copy stand (~$100)
- other accessories (~$50)

Upgrading to an Epic Pro (http://gigapan.org/cms/shop/epic-pro), with a Digital SLR camera and its lenses, would increase the overall price by about $3,000. The moderate price of the system described here is financially accessible to many different collections: from small, personal collections to those with millions of specimens. The system is user-friendly, under normal circumstances after setup, initial data can be captured quickly and easily. The software is also easy to use and avenues for support are readily available through GigaPan.com. Furthermore, the ability to customize and adapt the system is highly advantageous because it does not limit the purchaser/user to particular hardware. For example, if a collection/laboratory already has an acceptable camera, it has the potential to be coupled with the system without the need to purchase a new one. Also, because the system was initially developed for work in the field, it could easily play a role in both "lab bench" research (as described here) and remote field work. Finally, the infrastructure to easily host, discuss, and annotate these immense panoramas is already present (i.e., GigaPan.com) and thus alleviates the need to

invest in ways to locally disseminate the product (e.g., buying personal servers). All of these factors contribute to increased accessibility, a critical component for widespread adoption. The formation of a vast online community of collections, and the resulting communications, could be contingent on this ease of adoption.

Limitations of GigaPan

The main difference between GigaPan and other image capturing/stitching systems is that the robot and camera are fixed and rotate around a central point. XY coordinate systems, on the other hand, pan across a fixed area and are shot in the same horizontal plane and at the same distance. Because GigaPan rotates around a point, there is always some curvature/distortion to the images (Fig. 6). The level of curvature is proportionate to the distance the unit is from the subject and the zoom (Fig. 7). Though the stitch software adjusts for these effects, measurements being made from the panoramas would not be accurate in portions of the image (see Results for distortion effects). Insects near the bottom of the drawer and their unit tray labels can be blocked by the leading edge of the unit tray, especially small trays with specimens close to the top edge. Additionally, while other drawer types (e.g., Cornell & California Academy styles) with similar dimensions should be easily accommodated using the methods described here, larger or custom drawers will need a greater distance between the insects and camera to keep the curvature to a minimum; this in turn would compromise the magnification of the images (without the use of special lenses). However, the curvature does allow for viewing vertically-oriented header labels in unit trays in the upper half of the panorama, more angled views of the insects (i.e., their sides), and specimen labels that are less hidden by the body of the insect (usually more hidden with a completely over-head camera, i.e. XY system). All of these results can actually be advantageous because they permit more information to be displayed in the panorama.

Other considerations are necessary for utilizing the system to its fullest. For an efficient workflow, an AC adapter should be integrated into the unit. The GigaPan robot normally runs on batteries that are quickly drained after several panoramas are shot. Rechargeable batteries last somewhat longer, but still need to be recharged and put back in the robot, which is time consuming; it also moves the robot, negating any saved coordinates and reducing overall efficiency. Integrating the adapter requires electrical knowledge, but can be done (Sargent et al. 2010). If the panoramas are going to be represented online an internet connection is necessary, preferably one with fast upload speed. This may be a limitation for some collections.

Annotating the panoramas on Gigapan.com is not as sophisticated as necessary for highlighting specific structures on an insect. Presently, only a rectangular snapshot can be made of an area in the panorama; more detailed description is then required to signify what the snapshot is showing. The development of better tools that could highlight specific structures would be beneficial for communicating information held within the panoramas.

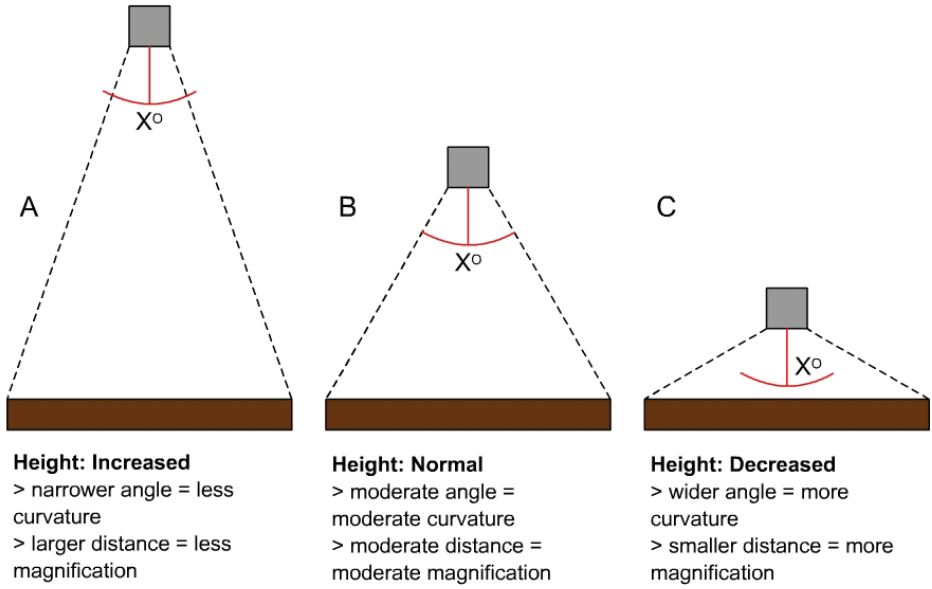


Figure 7. Illustration of the panning angle with the GigaPan robot at different heights. **A** higher than described **B** as described **C** lower than described.

Future goals

The utility of a Digital SLR equipped with a macro lens should be tested for this system. We anticipate higher quality images with better resolution of smaller specimens using better optics, though we do not entirely know how larger cameras and lenses (and their intrinsic characteristics) will affect the process. This would require both an SLR camera and a larger GigaPan robot (i.e., Epic Pro). Adding a step for post-processing images in photo editing software (e.g., Adobe Photoshop) prior to stitching, in order to enhance the sharpness, color and exposure of the panoramas, may improve final image quality.

Ongoing efforts to database the collection and apply unique specimen barcodes could be integrated into the final product. Already several drawers online have barcoded specimens (for example http://gigapan.org/gigapans/69756), though most barcodes are obscured under other labels to save space. However, modifying drawers to have the barcodes visible could allow people browsing the collection to scan the codes on their computer screen to access relevant label data or populate a list of specimens needed for loan. The system could be useful for tracking specimens that move between drawers and link them to their placement in the most current panoramas.

Many future goals involve enriching these panoramas by integrating more layers of information. We anticipate adding more keywords to each panorama to enable more powerful searches. These would include lower taxonomic ranks (subfamilies, tribes, genera, and species) and perhaps general localities. There is a great benefit to linking other information to the panoramas. For instance, a snapshot of one species (or

a series of specimens of one species) could be linked to the species' detailed images found on Morphbank (http://www.morphbank.net/), biodiversity information from GBIF (http://www.gbif.org/), genetic sequence data from Genbank (http://www.ncbi.nlm.nih.gov/genbank/), and other sources like the Encyclopedia of Life (http://eol.org/), Tree of Life project (http://tolweb.org/tree/), and many others. Additionally, if structures can be more accurately annotated (see *Limitations of GigaPan*), they could be linked to data present in various anatomy and phenotypic ontologies (e.g., OBO Foundry; http://obofoundry.org/). The possibilities are vast, but would require some added infrastructure and resourcing to achieve these results.

Other research avenues for these panoramas should be assessed. Can specimens in the image be analyzed and identified using a computer algorithm and machine learning? Can text information be extracted from the visible labels? With correction techniques, can accurate measures and morphometric analyses be performed? Could we use these panoramas to profile the state and quality of each drawer in the collection (similar to criteria described in McGinley 1993 and Favret et al. 2007)? What can the panoramas tell us about color patterns within and between species? These are a few of the uses envisioned, though they are by no means the only possibilities.

Conclusions

Overall, this project has generated excitement among entomologists and museum colleagues, which is encouraging for the future utility and adoption of this system. Many experts readily recognize the utility of drawer GigaPans, and the project has triggered several conversations about how to extend their outreach and research potential, as well as their ability to increase institutional awareness (both internally and externally). Though there are concerns about the full utility of these panoramas, especially the quality and nature of the images for identifying some insects, and their accuracy after the drawer contents go through curation, the low cost, ease of use, moderate speed, and online support make this technology a feasible system for imaging and sharing insect drawers from many settings.

Acknowledgments

We thank Matt Yoder for input during initial development of the project and Lydia Abernethy for aiding in collection management related to the project. We are indebted to Mary Jo Daines (CREATE Lab, Robotics Institute, Carnegie Mellon University), Rich Henderson (GigaPan Systems), and Randy Sargent (NASA Ames Research Center) for assistance with the GigaPan system and providing statistics on our GigaPan profile. The manuscript was improved over the original submission following advice and comments by two anonymous reviewers and the editor. This research was supported by the NSF grant DBI-0847924.

References

Blagoderov V, Kitching I, Simonsen T, Smith V (2010) Report on trial of SatScan tray scanner system by SmartDrive Ltd. Available from Nature Precedings http://hdl.handle.net/10101/npre.2010.4486.1

Favret C, Cummings KS, McGinley RJ, Heske EJ, Johnson KP, Phillips CA, Phillippe LR, Retzer ME, Taylor CA, Wetzel MJ (2007) Profiling natural history collections: A method for quantitative and comparative health assessment. Collection Forum 22(1–2): 53–65.

GigaPan Systems (2010) User Guide for Gigapan Epic and Epic 100. Available: http://gigapan.org/cms/manual/pdf/epic100-manual.pdf [accessed 22 June 2010 21:18]

Hammond BA (0000) Digital Insects: Assessing Online Presentations of Entomological Collections for Research, Education, and Outreach. A Master's Paper for the M.S. in I.S. degree. University of North Carolina, Chapel Hill, NC, USA.

McGinley RJ (1993) Where's the management in collections management? Planning for improved care, greater use, and growth of collections. In: Rose CL, Williams SL, Gisbert J (Eds) Congreso Mundial Sobre Preservación y Conservación de Colecciones de Historia Natural. Vol. 3. Temas de Actualidad, Iniciativas y Direcciones Futuras sobre Preservación y Conservación de Colecciones de Historia Natural. Dirección General de Bellas Artes y Archivos, Madrid, 309–338.

Sargent R, Denning S, Bertone M (2010) Using an external power adapter with your GigaPan. Available: http://bit.ly/GigaPanACadapter [accessed 16 June 2010, 15:42]

ZooKeys 209: 133–146 (2012)
doi: 10.3897/zookeys.209.3178
www.zookeys.org

RESEARCH ARTICLE

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

No specimen left behind: industrial scale digitization of natural history collections

Vladimir Blagoderov[1], Ian J. Kitching[1], Laurence Livermore[1], Thomas J. Simonsen[1], Vincent S. Smith[1]

1 *Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK*

Corresponding author: *Vladimir Blagoderov* (v.blagoderov@nhm.ac.uk)

Academic editor: *Lyubomir Penev* | Received 8 May 2012 | Accepted 22 June 2012 | Published 20 July 2012

Citation: Blagoderov V, Kitching IJ, Livermore L, Simonsen TJ, Smith VS (2012) No specimen left behind: industrial scale digitization of natural history collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 133–146. doi: 10.3897/zookeys.209.3178

Abstract

Traditional approaches for digitizing natural history collections, which include both imaging and metadata capture, are both labour- and time-intensive. Mass-digitization can only be completed if the resource-intensive steps, such as specimen selection and databasing of associated information, are minimized. Digitization of larger collections should employ an "industrial" approach, using the principles of automation and crowd sourcing, with minimal initial metadata collection including a mandatory persistent identifier. A new workflow for the mass-digitization of natural history museum collections based on these principles, and using SatScan® tray scanning system, is described.

Keywords

Digitization, imaging, specimen metadata, natural history collections, biodiversity informatics

Introduction

Natural history collections are of immense scientific and cultural importance. Specimens in public museums and herbaria and their associated data represent a potentially vast repository of information on biodiversity, ecosystems and natural resources for the widest range of stakeholders, from governments and NGOs to schools and private individuals. Numerous examples of the uses to which biodiversity data derived from natural history collections have been put in research on evolution and genetics, nature conservation and resource management, public health and safety, and education are widely available (summarized in

Chapman 2005, Baird 2010). The universe of natural history collection data has been estimated to be between 1.2 and 2.1 × 10^9 units (specimens, lots and collections) (Ariño 2010). To ensure efficient access, dissemination and exploitation of such an immense wealth of biodiversity relevant data, it is evident that a well-coordinated and streamlined approach to global digitization is required, in particular because it is absolutely essential for the scientific value of the generated data that the outputs (images, metadata, etc.) are linked together and also back to the original specimens via unique identifiers (uIDs).

In recent years, substantial efforts and resources have been invested into the digitization of natural history collections, with museums and herbaria routinely employing specimen level collection databases to replace older, paper-based card indexes and ledgers. In theory, this should make dissemination of specimen data through biodiversity informatics portals such as the Global Biodiversity Information Facility (GBIF; http://www.gbif.org/) very simple and straightforward. However, the truth is that natural history collections are almost as far from complete digitization as they were 20 years ago. Ariño (2010) estimated that no more than 3% of biological specimen data is web-accessible through GBIF, the largest source of biodiversity information. Consequently, there is neither a central database of collection holdings, nor a complete collection index available to users. The reason for this deficiency is partly the immense effort it would take to digitize the vast number of collections units involved (Vollmar et al. 2010). The cost of traditional digitization workflows is vast, both in financial and human terms. Our simple calculations have shown that complete databasing of the ~30 million insect specimens housed in the entomological collection of the Natural History Museum, London, would require 23 years of continuous work from the entire departmental staff to complete (65 people). Depending on the particular collections and curatorial practices used, estimates vary from US$0.50 to several dollars per specimen to capture full label data (Heidorn 2011). The cost of traditional imaging and databasing of every natural history object in all European museums was recently estimated as €73.44 per object (Poole 2010). Thus, the complete digitization of all natural history collections may cost as much as €150,000 million, and take as long as 1,500 years.

The most common solution proposed to overcome the enormous cost of digitization is prioritization based on user demand (Berents et al. 2010). Currently, most digitization projects concentrate their efforts on obtaining high quality images of selected specimens accompanied by high quality data (e.g., comprehensive and expertly interpreted label information) rather than total collections coverage. Such specimen-centric digitization efforts are thus inevitably fragmented into numerous small-scale and labour-intensive projects that usually image single specimens, one at a time.

To solve the problem of cost, as well as the inherent fragmentation in collection based biodiversity informatics, new, industrial-scale approaches to digitization are clearly needed. The larger a digitization project becomes, the lower are the transaction costs and thus the lower is the cost per specimen. Such an industrial-scale process must necessarily fulfil certain standardized criteria if it is to be of use to and adopted by a wide spectrum of natural history collections:

- As much as possible of the procedure must be automated, except when physical handling of specimens is necessary.
- The approach should, whenever possible, focus on "wall-to-wall" total digitization of entire collections, because it is faster to digitize an entire collection than to select individual specimens or drawers of particular interest.
- Complicated labour-intensive procedures must be divided into a series of separate, shorter steps, each with a distinct outcome. For example, preparation of specimens for imaging should be a separate step from the imaging itself; and unique specimen identifiers can be assigned simultaneously to all specimens in a drawer rather than individually and sequentially. Such a modularised process can then be more easily crowd-sourced among the professional and volunteer communities. Properly organized crowd-sourcing projects would be able to mobilise the efforts of thousands of enthusiasts around the world (Hill et al. 2012).
- Collection of metadata must be simplified and standardized. In most cases, digital representation of the specimen and minimal metadata (uID, specimen location in the collection) is sufficient for collection management purposes. Only minimal information should be collected when initially digitizing an entire collection, but in such a way that it can be amended and expanded upon later.

Here we describe a new method for "wall-to-wall" mass-digitization of natural history museum collections based on the SatScan® tray scanning system. The method allows for standardized scanning of museum collection trays of the highest image quality possible, followed by simplified (and easily expandable) collection of metadata.

Methods

The Natural History Museum (NHM), London, has been working with SmartDrive Limited (http://www.smartdrive.co.uk/) since 2009 on the development of one of the company's products, the SatScan® collection scanner (Fig. 1). From this collaboration, we have developed a workflow that we consider meets our needs for the industrial-scale digitization of a significant part of the NHM's collections. The system is particularly suited to the digitization of multiple, uniformly mounted or laid out specimens, such as pinned insects and smaller geological or mineralogical objects in standardized collection drawers, horizontally-stored microscope slides and herbarium sheets.

The digitization workflow envisioned for the NHM (Fig. 2) comprises three steps:

Imaging

The SatScan® collection scanner is capable of producing high-resolution images of entire collection drawers (see Table 1, Blagoderov et al. 2010, Mantle et al. 2012). The specific configuration of the system has changed somewhat from that described in

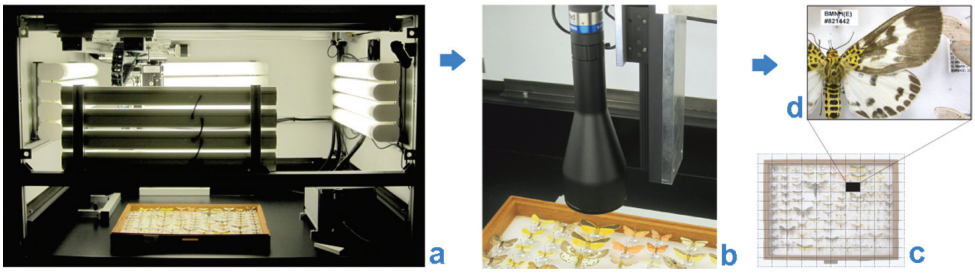


Figure 1. SatScan imaging: **a** SatScan machine **b** specimens being imaged **c** individual frames aligned **d** fragment of a stitched image; final resolution of the stitched image ~11 lines/mm.

the report, such that now a USB CMOS UEye-SE camera (model # UI-1480SE-C-HQ, 2560×1920 resolution) is used in combination with Edmund Optics telecentric TML lenses of 0.3× (#58428) and 0.16× TML (#56675). A camera with attached lens is moved in two dimensions along precision-engineered rails positioned above the object to be imaged. A combination of hardware and software provides automated capture of high resolution images of small regions of interest, which are then assembled ("stitched") into a larger panoramic image, generating the final image of the entire drawer. This method maximizes depth of field of the captured images and minimizes distortion and parallax artefacts. Analogous solutions for large-area imaging which have been developed independently include GigaPan (Bertone et al. 2012), MicroGigaPan (Longson et al. 2010) and DScan (Schmidt et al. 2012).

Metadata capture

A prototype software program, Metadata Creator, has been designed to allow fast capture of specimen data and associating these with the image of the specimen (Fig. 3). Users can mark individual specimens on the panoramic image by drawing rectangular boxes around them, selecting these areas and annotating them individually or in batches. Methods for marking the specimen, editing regions of interest and selection of multiple specimens are analogous to those used in many common graphic applications and so will be familiar, even to inexperienced users.

Specimen metadata is captured in a series of fields that are compatible with the Darwin Core 1.4.1 schema (http://rs.tdwg.org/dwc/) and which can be customized to particular user requirements. To maximize throughput, only basic metadata are collected at this stage. These will generally include a unique collection number of every specimen (see below, barcodes), collection identification (to the available curatorial level, e.g. to species/subspecies for the "Main Collection" and family/order for unsorted accessions), and, if possible, biogeographic region/country. Taxon names are looked up from an index derived from the NHM Collections Management Database. A completed project comprises a folder with an archival image of the drawer, full-reso-

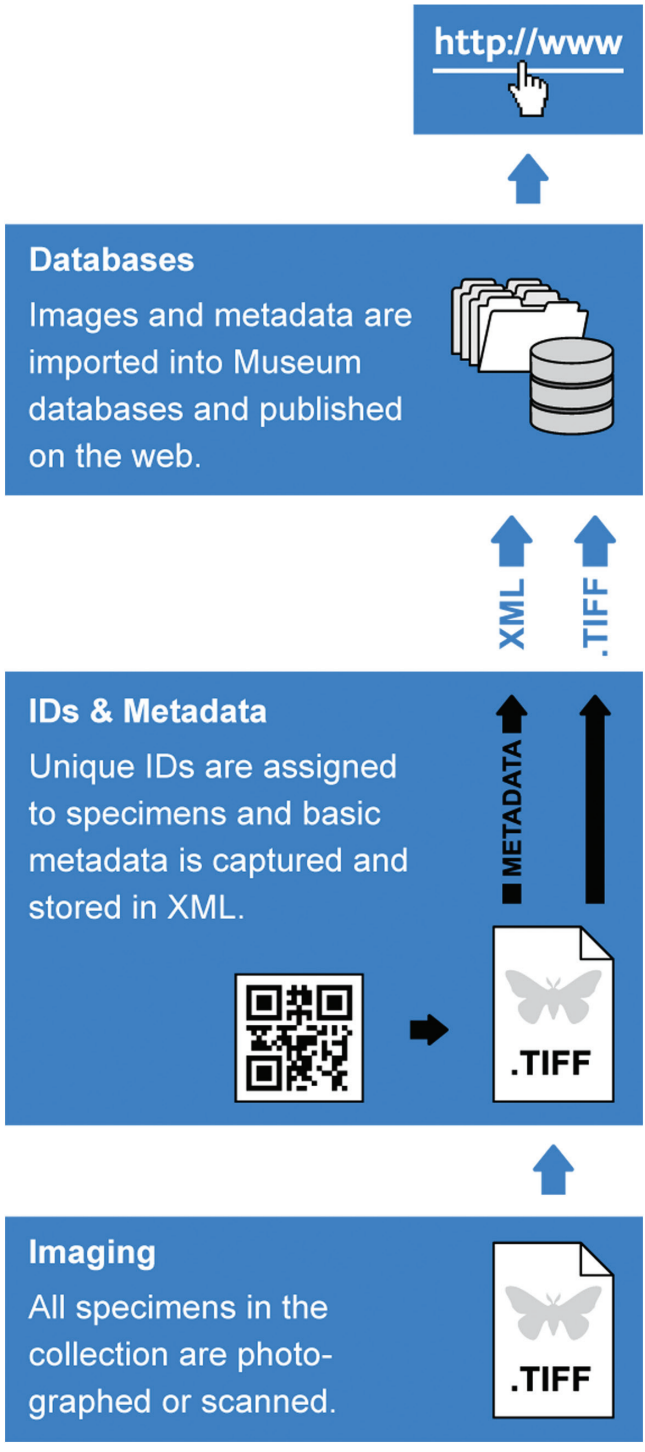


Figure 2. Image based digitization workflow consisting of four stages: Imaging, Metadata capture, Institutional databading and Publication.

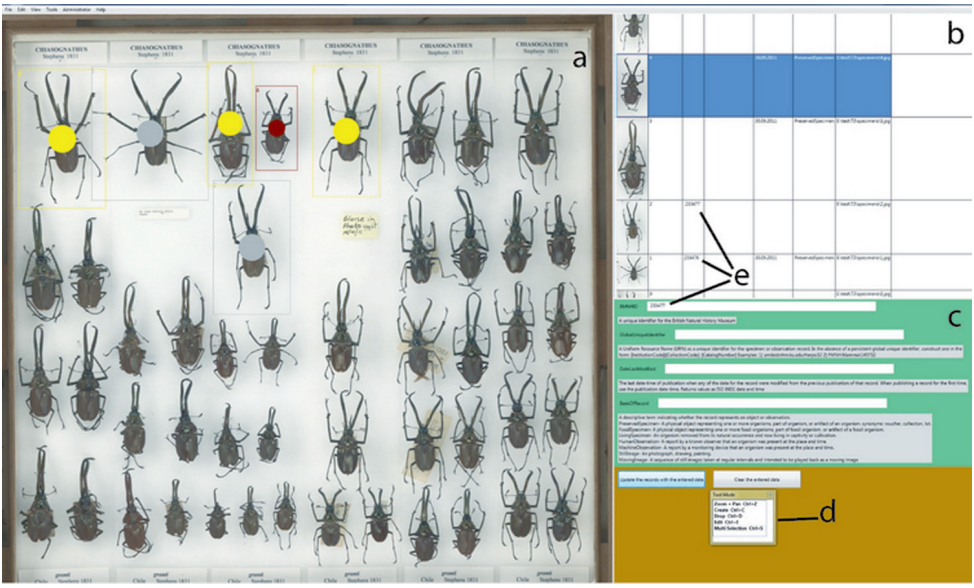


Figure 3. Metadata Creator software: **a–c** working areas **a** drawer image **b** specimen records **c** annotation fields **d** tool selector **e** unique IDs.

Table 1. Resolution and depth of field of the system as compared with a Canon EOS450D DSLR camera using a Canon MP E-65 macrolens (USAF: the smallest resolvable element on 1951 US Air Force resolution test chart; MRD: minimal resolved distance, size of the smallest visible object on image)

Objective	Sensor Resolution	Aperture	Depth of Field, mm	Resolution		
				USAF	Lines/mm	MRD, μm
SatScan 0.16× lens	1280×960	Open	5	3–4	11.3	44
		Dot	10	3–4	11.3	44
		Closed	>70	2–5	6.35	79
	2560×1920	Open	5	4–3	20.16	25
		Dot	14	4–1	16.0	31
		Closed	>70	3–2	8.89	56
SatScan 0.3× lens	1280×960	Open	2.5	4–2	17.95	28
		Dot	4.5	4–2	17.95	28
		Closed	30	3–4	11.3	44
	2560×1920	Open	1.5	5–3	40.3	12
		Dot	3	5–2	36.0	14
		Closed	35	3–5	12.7	39
Canon MP-E65 lens, 1×	4272×2848	2.8	0.5	5–6	57	8.8
		16	4	-	-	-
Canon MP-E65 lens, 5×	4272×2848	2.8	<0.3	8–1	256	2
		16	2	6–2	71.8	7

lution images of individual specimens cut-out from the drawer image, and an XML file containing annotations and links to specimen images (Appendix 1). Trials have demonstrated that 10–20 seconds per specimen is required to capture basic metadata using the Metadata Creator Software. A unique ID for the drawer is also recorded. As the NHM Collection Management System already includes a complete collections index (a brief description of the content of every drawer), no additional information is required.

Assigning uIDs

Every specimen is assigned a unique number under which it will be registered in the NHM Collections Management Database. It is a requirement of collections management procedures that a label bearing the specimen's uID is attached to the specimen. To streamline this part of the process, it is subdivided into the following steps:

1. A sequence of unique numbers is generated from the NHM Collections Management Database.
2. Labels that include both a human-readable number and a machine-readable barcode are printed.
3. The operator labels the specimens by selecting a specimen on the drawer image, pinning a label under the specimen, and scanning the barcode, thereby adding the uID into the corresponding field of Metadata Creator. Barcodes can be pinned facing up or down depending on curatorial practice; the former has the advantage of visibility on the image. In this case imaging, of course, has to take place after assigning uIDs. Images of individual specimens for which the metadata have been collected and individual numbers assigned are automatically marked on the drawer image with a grey spot, allowing easy visualization of progress.
4. When all specimens have been labelled and recorded, the XML file and corresponding specimen images are imported into the NHM Collections Management Database.

We must emphasize that Metadata Creator is a prototype software application; much more development is needed for to perfect its functionality, user interface, and integration with the Museum's information systems.

Results

A preliminary assessment of the SatScan® system was undertaken and reported upon by Blagoderov et al. (2010). Based on their findings, a series of recommendations were made for improvements and possible longer term developments to the hardware, software, imaging system and ergonomics. An updated system was delivered to the NHM in September 2011 and further trials were then conducted. This newer version of SatScan®

provides non-extrapolated resolution of the final images from 11.3 to 40.3 lines/mm and a minimum resolved distance of 79 to 12µm, depending on the lens and sensor settings employed (Table 1). The maximum depth of field has been increased slightly from 80mm to 85mm. Although focus stacking is implemented in the current version of the system, in most cases it is not necessary. For the majority of collections drawers, specimens are presented at a more-or-less uniform height and within the available depth of field; focus stacking is really only necessary for those drawers where specimens are pinned at markedly different heights or are particularly deep (e.g. fossils and mineralogical samples). The average time to scan a typical collection drawer without focus stacking is between four and six minutes, depending upon size (eight to ten minutes including logistics, Table 2). This generally translates to about two seconds per specimen. Thus, in a working day, an operator could image up to 70 drawers. These would then be stitched into the final images using an overnight batch process (see average figures in Table 2). The resulting images vary in size from 0.3 Gpx to 5 Gpx (10^9 pixels; 250 MB – 3 GB compressed TIFF files) depending on the imaging area, lens and resolution used. However, use of the highest resolution in mass digitization projects may not always be practical. We did not conduct extensive tests with the highest resolution of camera/higher magnification of lens because a 64-bit version of software is needed to handle the stitching process for files of this size, and this was not available at the time of trials.

The part of the process that involves marking of specimens and metadata capture using Metadata Creator has not been as thoroughly tested and we have yet to trial the part of the procedure that produces barcode labels and attaches these to specimens. However, preliminary results involving mock elements indicate that it will take about ten seconds per specimen. This time will be extended for those specimens that already have a human-readable uID (a "BMNH(E)" number, for example) but no barcode label, because then the former will have to be manually entered into Metadata Creator and a new barcode label printed. However, relatively few NHM insect specimens (about 1.2%) have so far been databased and assigned a uID.

The entomology collections of the NHM have about 30 million insect specimens, mostly pinned, housed in 135,000 collections drawers. Assuming that 80% of the collection is appropriate to be imaged using the SatScan® system, rough calculations based on the above figures suggest that the entire collection could be imaged and basic metadata captured in 18 person-years.

Table 2. Scanning and stitching times for different types of drawers.

Drawer type	Number of drawers in trials	Dimensions, mm	Number of frames	Average scanning time (including logistics), min	Average stitching time, min	File size, Mb
Main collection and accessions	236	500×400 or 470×450	17×14 or 16×15	8.52	12.65±1.54	488.20±30.21
Rothschild and Rhopalocera	144	560x540 or 570x 555	21×17 or 22×17	10.13	25.41±4.21	715.90±89.58

Discussion

Although images acquired through an industrial digitization process might be considered to be of limited use for taxonomic purposes, because they feature only one aspect of the specimens and may not contain necessary morphological details or label data, they could prove very useful for a variety of other purposes. Obvious collection management applications include improved collection audit and security, as well as improving accessibility of the collection. For research purposes, such acquired images could prove very valuable in morphometric analyses and phenological population studies. In addition, the public engagement aspect of industrial digitization activities should not be underestimated. Online public access to high resolution images and metadata will likely enhance public awareness of the importance of local and national collections (as well as engendering a sense of shared ownership). Moreover, high quality images will open up the possibility for fast and reliable automated or semi-automated specimen identification and thus encourage environmental "citizen-science", such as recording distributional or abundance changes of key species.

Major problems remaining with the described approach are largely concerned with the time taken to scan specimens/samples and to collect metadata. Even with a simple approach, scanning a specimen takes approximately two to four seconds followed by 10–20s for annotation and/or barcoding. Furthermore, only basic metadata are collected under the scenario described above. Indeed, in the worst case, say a drawer of unidentified mixed organisms from several phyla, only a uID will be associated with each of the specimens. It may then be argued that this will compel museums and herbaria to create essentially incomplete records with which to populate their collection databases. However, such records are comparable to stub pages in Wikipedia, empty at the moment but capable of being filled and edited in due course. Indeed, there is a case to be made for the opposite viewpoint, that there is no point collecting complete metadata if these are not going to be used for any purpose. Finally, it should be noted that the industrial digitization process described above only works relatively seamlessly for more-or-less uniformly preserved and presented specimens, such pinned insects in drawers and herbarium sheets. It is unlikely to be satisfactory for pickled specimens in jars of ethanol. These collections may have to be digitized using a different protocol.

Approximately 90% of the time required for digitization is spent on capturing metadata and labelling specimens. While the latter involves physical handling of the specimens and must be performed by experienced staff, selection of specimens in the drawer images and annotation thereof can be undertaken in a virtual environment. In many cases, the basic information to be collected can be seen in the drawer image. Implementing an open source web application that duplicates the functions of Metadata Creator and publication of drawer images using algorithms involving a pyramid of tiles (produced using Zoomify™ (http://www.zoomify.com/) or Google Maps (http://maps.google.com/), for example) will allow volunteers from around the world to participate in digitization of the collection and will decrease the time needed to process a specimen by at least 50%.

The next step in facilitating the digitization process might be to undertake "virtual curation". Here, uIDs are assigned to each specimen, records are created in the collection management database and corresponding specimen images linked to these records, but the specimens themselves are not labelled until it becomes necessary to handle the specimen physically for some other purpose (curation, loan, identification, dissection, etc.). Of course, these procedural changes would require a major cultural shift for Collections Management staff.

Revised, though still simplistic, calculations now show that the entire NHM collection of insects could be imaged in 12.88 person-years and completely digitized without crowd-sourcing in 118 person-years. Collecting basic information and attaching a barcode to a specimen would take approximately 10-30 seconds. Per-specimen cost under the current (2012) economic climate would thus be as low as £0.12. If we limit SatScan-based digitization to large and medium-size insects (up to 5 mm in length), the total time required is 58 man-years. This effort does not seem insuperable considering that the NHM insect collection is managed by 26 permanent curatorial staff, assisted by a number of people in short-term contracts and volunteers.

Despite the potential perceived drawbacks, image-based basic digitization can nevertheless mobilize hundreds of millions of biological specimens in a relatively short period of time. It is estimated that entomological specimens constitute up to 40% of all natural history specimens (Ariño 2010). Some palaeontological, zoological and mineralogical specimens, including microscopic slides, are also stored in collection drawers and trays that are amenable to simultaneous imaging. Thus, the majority of natural history specimens could potentially be digitized using industrial imaging.

The return on investment in total collection digitization will be enormous. It will open up collections to the world, facilitating their use, and help create a global collection index that can be used to set priorities for further digitization. Basic digitization of all the world's holdings of insects (800 million specimens) could be completed in less than 4000 person-years. This may sound like a huge figure, but divided among approximately 1,300 collections and potentially tens of thousands of professionals and volunteers, the work could be completed much quicker, perhaps in only a few years. "Furthermore, emerging technologies in the near future will undoubtedly decrease time and costs, while increasing data quality. Complete image-based basic digitization of insect and plant collections would produce at least 30 Pb (10^{15} bytes) of data, which constitutes ~0.0006% of the current data hosted on the Internet. At £0.2 per specimen, the cost of digitizing 2,000 million natural history specimens may appear to be an eye-wateringly high figure of £400 million. However, divided among ~4,000 natural history collections, this reduces to an average project cost of £100,000, which is equivalent to the size of a relatively modest research grant. To this the cost of imaging equipment must be added. At present, a SatScan system costs between £25,000 and £60,000, depending on the options to be implemented and the service agreement chosen, but less expensive alternative solutions are also being developed (Bertone et al. 2012, Dietrich et al. 2012, Schmidt et al. 2012).

Regardless of the technology used, mass digitization will nevertheless follow the same general approach, which includes mechanisms that enrich digital media with specimen-level metadata. This enrichment will:

1. Facilitate open dissemination of data so that it can be discovered and accessed by stakeholders, reducing both the need for physical access to collections and the number of loans;
2. Enable large-scale manipulation and integration of collection data, supporting stakeholders in their monitoring and management of information on ecosystems, biodiversity and natural resources;
3. Enhance curatorial activities, allowing the condition of loans to be tracked and reduce identification inaccuracies;
4. Protect biodiversity heritage by reducing the need to handle irreplaceable specimens;
5. Improve collections security by providing base-line images against which damage and thefts can be monitored;
6. Support disaster management, such that should the worst happen to a collection, its digital representation will continue to provide a valuable resource;
7. Raise natural history collections profiles, resulting in improved resources for further research;
8. Contribute beyond the traditional remit of museums and herbaria into new areas of interest, particularly education and public understanding of science; and
9. Support biodiversity legislation and data repatriation, which is an increasing requirement under both the 1992 Convention on Biological Diversity and the subsequent 2010 Nagoya Protocol on Access and Benefit-sharing.

Acknowledgements

The authors are very grateful to Dennis Murphy, Dave Freer, and Mike Broderick (SmartDrive Ltd.) for fruitful and ongoing collaboration; Chris Raper for help with testing the system and development of software; and two reviewers, in particular Matt Yoder, whose comments and suggestions greatly improved original version of the manuscript.

References

Ariño AH (2010) Approaches to estimating the universe of natural History collections data. Biodiversity Informatics 7: 81–92.

Baird R (2010) Leveraging the fullest potential of scientific collections through digitization. Biodiversity Informatics 7: 130–136.

Berents P, Hamer M, Chavan V (2010) Towards demand-driven publishing: approaches to the prioritization of digitization of natural history collections data. Biodiversity Informatics 7: 113–119.

Bertone MA, Blinn RL, Stanfield TM, Dew KJ, Seltmann KC, Deans AR (2012) Results and insights from the NCSU Insect Museum GigaPan project. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 115–132. doi: 10.3897/zookeys.209.3083

Blagoderov V, Kitching I, Simonsen T, Smith V (2010) Report on trial of SatScan tray scanner system by SmartDrive Ltd. Available from Nature Precedings http://hdl.handle.net/10101/npre.2010.4486.1

Chapman A (2005) Uses of Primary Species Occurrence Data, version 1.0. Global Biodiversity Information Facility, Copenhagen, 106 pp.

Dietrich CH, Hart J, Raila D, Ravaioli U, Sobh N, Sobh O, Taylor C (2012) InvertNet: a new paradigm for digital access to invertebrate collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 165–181. doi: 10.3897/zookeys.209.3571

Heidorn PB (2011) Biodiversity Informatics. Bulletin of the American Society for Information Science and Technology 37(6): 38–44. doi: 10.1002/bult.2011.1720370612

Hill A, Guralnick R, Smith A, Sallans A, Gillespie R, Denslow M, Gross J, Murrell Z, Conyers T, Oboyski P, Ball J, Thomer A, Prys-Jones R, de la Torre J, Kociolek P, Fortson L (2012) The notes from nature tool for unlocking biodiversity records from museum records through citizen science. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 219–233. doi: 10.3897/zookeys.209.3472

Longson J, Cooper G, Gibson R, Gibson M, Rawlins J, Sargent R (2010) Adapting Traditional Macro and Micro Photography for Scientific Gigapixel Imaging. Proceedings of the Fine International Conference on Gigapixel Imaging for Science, November 11–13 2010. http://repository.cmu.edu/gigapixel/1

Mantle BL, La Salle J, Fisher N (2012) Whole-drawer imaging for digital management and curation of a large entomological collection. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 147–163. doi: 10.3897/zookeys.209.3169

Poole N (2010) The Cost of Digitising Europe's Cultural Heritage. Collections Trust. http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/annexes/digiti_report.pdf

Schmidt S, Balke M, Lafogler S (2012) DScan – a high-performance digital scanning system for entomological collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 183–191. doi: 10.3897/zookeys.209.3115

Appendix 1

An example of XML output of Metadata Creator.

<?xml version="1.0" encoding="utf-8"?>
<Project xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
<Templates>
<key>BMNHID</key>
<value>A unique identifier for the British Natural History Museum</value>
<key>GlobalUniqueIdentifier</key>
<value>A Uniform Resource Name (URN) as a unique identifier for the specimen or observation record. In the absence of a persistent global unique identifier, construct one in the form: [InstitutionCode]:[CollectionCode]: [CatalogNumber] Examples: 1) urn:lsid:nhm.ku.edu:Herps:32 2) FMNH:Mammal:145732</value>
<key>DateLastModified</key>
<value>The last date-time of publication when any of the data for the record were modified from the previous publication of that record. When publishing a record for the first time, use the publication date-time. Returns values as ISO 8601 date and time</value>
<key>BasisOfRecord</key>
<value>A descriptive term indicating whether the record represents an object or observation.
</value>
</Templates>
<Specimens>
<Specimen>
<DarwinCoreData>
<key>ImageURL</key>
<value>E:\test\T3\specimens\G1_2_0000.jpg</value>
<key>BMNHID</key>
<value> </value>
<key>GlobalUniqueIdentifier</key>
<value> </value>
<key>DateLastModified</key>
<value> </value>
<key>BasisOfRecord</key>
<value>preserved specimen</value>
</DarwinCoreData>
<SpecimenIndex>0</SpecimenIndex>
<ImageDimensions>
<Left>519.635599159075</Left>
<Top>1490.562857142857</Top>

<Width>2247.9887876664334</Width>
<Height>3511.8564285714283</Height>
</ImageDimensions>
</Specimen>
<Specimen>
……
……
……..
</Specimen>
</Specimens>
</Project>

ZooKeys 209: 147–163 (2012)
doi: 10.3897/zookeys.209.3169
www.zookeys.org

RESEARCH ARTICLE

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

Whole-drawer imaging for digital management and curation of a large entomological collection

Beth Louise Mantle[1], John La Salle[1], Nicole Fisher[1]

1 *Australian National Insect Collection, CSIRO Ecosystem Sciences, GPO Box 1700, Canberra, ACT, 2601, Australia*

Corresponding author: *Beth Louise Mantle* (beth.mantle@csiro.au)

Academic editor: *V. Blagoderov* | Received 2 April 2012 | Accepted 25 June 2012 | Published 20 July 2012

Citation: Mantle BL, La Salle J, Fisher N (2012) Whole-drawer imaging for digital management and curation of a large entomological collection. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 147–163. doi: 10.3897/zookeys.209.3169

Abstract

Whole-drawer imaging is shown to be an effective tool for rapid digitisation of large insect collections. On-line, Whole-drawer images facilitate more effective collection management, virtual curation, and public engagement. The Whole-drawer imaging experience at the Australian National Insect Collection is discussed, with an explanation of workflow and examples of benefits.

Keywords

Digitisation, entomology, Whole-drawer, imaging, collections, Satscan

Introduction

"Existing taxonomic processes have served us well for centuries but are clearly inadequate for the challenge at hand. The taxonomic community must rally around a common vision……It is time to approach taxonomy as a large scale international science."

Quentin Wheeler, Peter Raven and Edward O. Wilson
Science, 2004

Libraries of printed material experienced a renaissance in the 1990s when documents were made available in a standardised, portable, digital file format, the PDF. The benefits of producing publications in both physical and digital formats were immediately clear: secure, space-efficient, resource-efficient, economical, accessible, and

so on. Arguably, the most important benefit of digitised publications is the ability to search the text within the literature, thus delivering a wealth of previously unknown and/or inaccessible data and information to users.

Natural history collections are libraries of temporal and spatial biodiversity information (Drew 2011). The data in these biological libraries are physically attached to individual specimens and, as a minimum, include information about when and where the specimen was collected, who collected it, and in the case of images what it looks like.

'Traditional' digitisation or databasing (i.e. entering label data from, or taking pictures, of individual specimens) of insect collections is inexorably slow, thus large entomology collections must seek alternative, large-scale approaches for improving delivery of biodiversity and taxonomic data to the world (Johnson 2012). Whole-drawer imaging of entomology collections is a digitisation method that is gathering momentum in a number of institutions, including the Australian National Insect Collection (ANIC) (Mantle et al. 2011), the Natural History Museum in London (BMNH) (Blagoderov et al. 2010) and the North Carolina State University (NCSU) Insect Museum (Bertone and Deans 2010). This technique produces high-quality, ultra-high resolution images of whole drawers or trays of insects for online display and extraction of specimen metadata. The resulting images of the specimen (and sometimes associated label) can be viewed, downloaded and annotated, thus providing collections and users with a remote resource for auditing, curating and accessing the collection without physically handling the specimens.

This paper will discuss the whole-drawer imaging project currently underway at the ANIC and provide an assessment against the predicted outcomes for the project. We predict that delivery of high-resolution whole-drawer images will:

1. Promote and encourage remote curation of unsorted specimens;
2. Deliver insect specimen metadata;
3. Assist with loan requests;
4. Provide a method for auditing the collection;
5. Permit morphometric analysis of at least some specimens; and
6. Encourage public engagement with biological collections.

Materials and Methods

Equipment

Imaging of collection drawers within ANIC takes place by the use of a SatScan™ prototype imaging system (Figure 1), developed by SmartDrive Ltd (http://www.smartdrive.co.uk). At the time of purchase in 2010 the complete system cost approximately AUD$80–100,000.

The SatScan system uses a combination of hardware and software that automatically captures a series of 200-400 "tile" images at precisely monitored positions. These

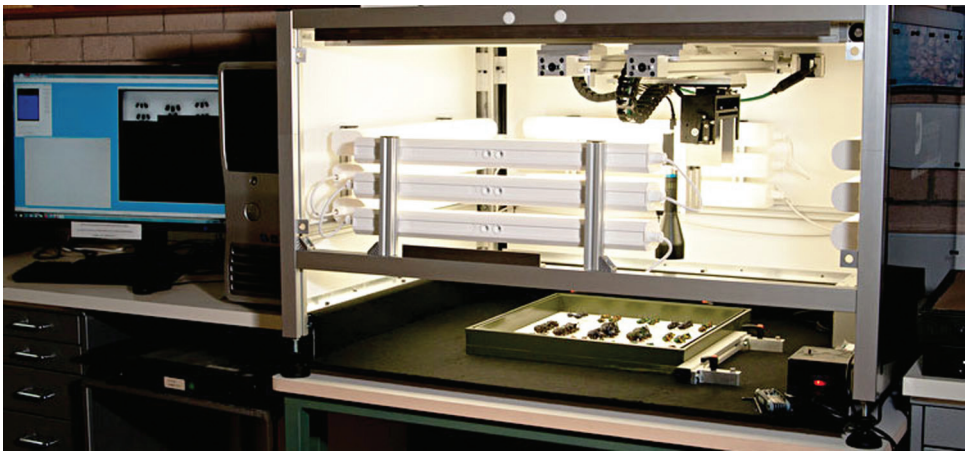


Figure 1. The SatScan imaging system used in ANIC. Shown here with the front cover removed.

tile images are then assembled ("stitched") together to form an extremely high-resolution final image of a drawer of insects.

The ANIC SatScan uses a Basler A631FC 1/2" CCD camera with Edmund Optics 0.16× telecentric lens #NT56-675 that moves in two dimensions along precision rails positioned above the drawer. In this way, the SatScan creates images with minimum distortion, no parallax artefacts and improves the overall coherence of the image. Therefore, all specimens are perfectly imaged with no occlusion from unit tray boxes and with uniform scale so that accurate measurements are valid anywhere throughout the image.

Framework surrounding the camera and lens is clad in a dark plastic material that contains twelve internal fluorescent tube lights for providing adequate light for short exposures (20–40 ms). The framework shields the drawers from surrounding ambient lighting, which could interfere with the controlled illumination inside the SatScan machine. The internal lighting is constant (not flashing) and the system operates quietly so as to not be obtrusive to the working environment.

Workflow

The SatScan captures sequential "tile" images (200 – 400 per drawer) during working hours, and then automatically "stitches" the tile images overnight to achieve a whole-drawer image. Essentially, the system captures and accurately mosaics together tile images to assemble a single, large image, covering the entire drawer area.

Given an average capture time of 5–7 minutes per drawer, a skilled operator can process up to 60 drawers of specimens each day, and up to 90 final pictures can be stitched in 12 hours (e.g. overnight). These times are typical for a trained operator and bug-free software.

Each drawer was assigned a unique identifier that also acts as a location code for the drawer within the collection. In addition, the unique identifier is the filename of the image (note – this identifier is not a GUID or LSID and is for internal ANIC use only). Hence, the image file and actual drawer can always be associated together. Figure 2 demonstrates the workflow process for digitisation of whole drawers in ANIC.

Output specifications for imaged ANIC drawers:

- Field of view: 35.5 × 27.5 mm
- Original tile images: 1280 × 960
- Final images: up to 21000 × 21000
- Resolution: ~35 px/mm
- Minimal resolved structures: 0.06–0.1 mm
- Depth of field: 10–80 mm
- File formats: 24bit BMP or LZW-compressed TIFF
- File size (15000 × 14000 px): ~780Mb (BMP), 340Mb (TIFF)
- Exposure: 1–1000 ms
- Capture time of 480 × 500 mm drawers: 5–7 min, depending on exposure
- Stitching time, 200–400 tiles: 5:30–9:30 min

Image Delivery

Whole-drawer images were uploaded to Morphbank-ALA image repository (http://morphbank.ala.org.au), where they can be viewed and navigated at a high resolution

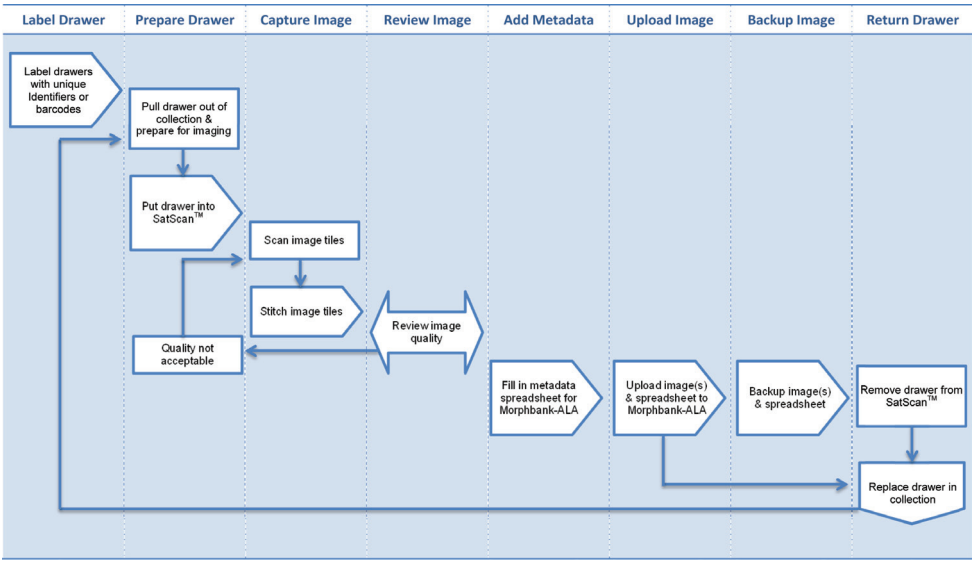


Figure 2. Workflow process in ANIC to Digitise whole drawers of insects and load images into Morphbank-ALA

(i.e. images are zoomable), edited, annotated and shared amongst the collections community, researchers and clients. Morphbank-ALA is a multi-concurrent user, web-based system, supported by all current mainstream browsers. The software is free, open-source and server-based. Images must be imported to make use of the management system, however the next version should enable referencing to externally stored images. Metadata is captured as a DarwinCore record and can be supplemented by additional user defined attributes. The system allocates stable, unique identifiers to images, which can be linked to and referenced in external publications. The system treats images as a representation of a specimen thus the subject in the image is the most important object, not the image itself. Morphbank supports assignment of taxonomic determinations and hierarchy to specimens, it supports groups and role-based security allowing for image collections to be maintained privately, within confined membership groups, and/or published to the public domain.

A typical ANIC entomology drawer measuring 480 × 500 mm produces a final image of 15000 × 14000 pixels, and file size of ~780MB (BMP) or 340MB (TIFF). Figure 3 shows an example of a TIFF drawer image displayed on the Morphbank-ALA website with the persistent URL http://morphbank.ala.org.au/?id=2075549.

At the time of publication, more than 1,500 whole drawer images were available on Morphbank-ALA. Images can be viewed by browsing the CSIRO-ANIC Group of images.

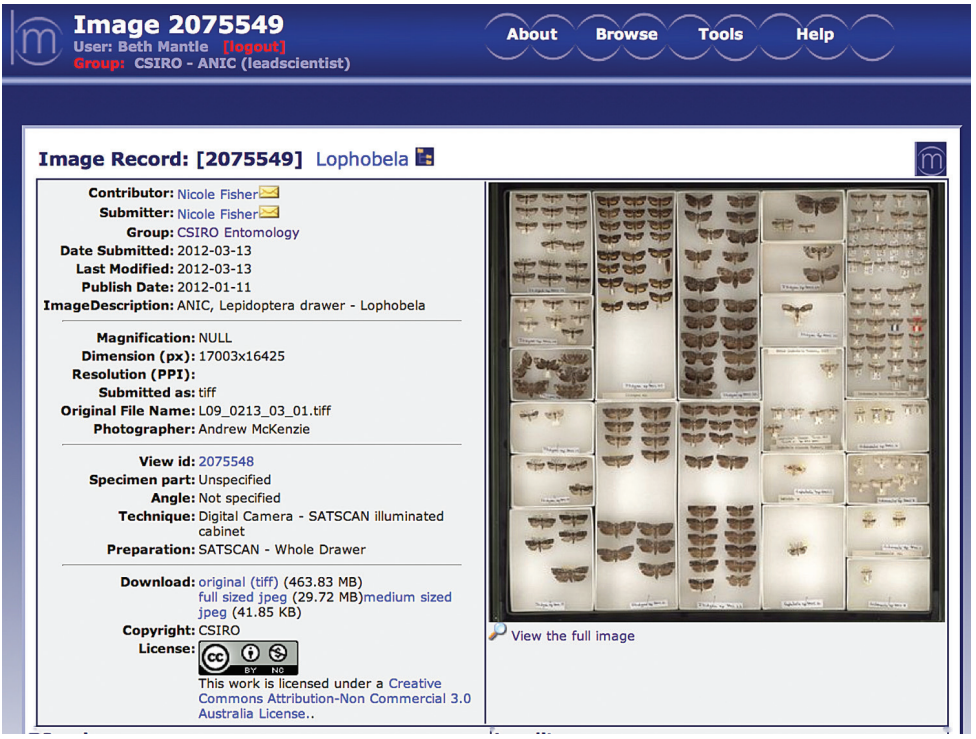


Figure 3. A whole-drawer image displayed in MorphbankALA for online for viewing, editing and download. Image properties: 17,003x16,425 pixels, 30 MB (JPEG), and 464 MB (LZW compressed TIFF).

Results and discussion

There are many challenges facing collections that plan to digitise specimen data, including: lack of funding support, loss of staff with the expertise required to accurately curate and identify specimens, and difficulty obtaining the appropriate technology and equipment (Vollmar et al. 2010). Some disciplines face greater barriers to digitisation than others. Entomological collections are particularly difficult. Insects are generally mounted on pins with very small labels attached beneath the specimen. To access the data, the specimens must be handled, the label removed from the pin and the associated data decoded and entered into a database. This is equally true for imaging individual specimens. Both forms of digitisation (data-basing, imaging) are time-consuming, and place the specimen at increased risk of damage through handling. Furthermore, entomology collections are large and contain significantly greater numbers of individual specimens than other zoological collections. The Natural History Museum in London (BNHM) boasts 28 million specimens (BMNH website), and the Smithsonian National Museum of Natural History (SNMNH) estimates holdings at more than 35 million specimens (SNMNH website).

The ANIC is the world's largest collection of Australian invertebrates and is comprised of approximately 12 million pinned, slide-mounted and fluid-preserved specimens. Based on the estimated number of specimens, and the current rate of 'traditional' digitisation at the ANIC, it will take a further 250 years to database the entire collection.

Whole-drawer imaging offers a rapid digitisation method that complements traditional databasing and has increased the rate of digitisation at the ANIC. At the time of publication, more than 1,500 collection drawers (from a current total of 22,000 drawers) have been imaged and uploaded to Morphbank-ALA . Although this project is in its early stages, the value of capturing and delivering whole-drawer images online is becoming clear.

Remote curation of unsorted specimens

Ultra-high resolution images of whole insect drawers provide enough morphological detail to facilitate identification of specimens remotely, which could contribute towards unblocking a significant "bottleneck" in the curation chain (Beaman et al. 2007). The expertise to provide accurate and reliable identifications of particular groups is often unavailable within a collection and therefore specimens cannot be appropriately identified internally. As such, entomology collections rely on visiting researchers to provide identifications and advice regarding reorganization of the collection, in this case by bringing the expertise to the specimens. However, online delivery of whole-drawer images brings the specimens to the expertise, wherever they are located, and increases the opportunity for specimens to reach a useful level of identification.

For example, an image of an unsorted drawer of Hemiptera specimens (Figure 4) was displayed to illustrate the size and quality of the images produced

Figure 4. Whole-drawer image of unsorted Hemiptera specimens with identifications provided by a remotely located expert, Dr Murray Fletcher. This drawer was subsequently re-curated according to the identifications, with specimens accessioned into the appropriate locations within the ANIC Hemiptera collection. See Appendix 1 for full list of remote identifications.

by the drawer scanner at the annual Australian Entomological Society conference in 2010. Almost immediately, several Hemiptera experts seated in the audience began calling out identifications for the specimens in the image. This exercise demonstrated the potential for remote curation of collections based on identifications of specimens in whole-drawer images.

The level of taxonomic identification using whole-drawer images varies, and is dependent on a number of factors:

1. Size of the specimens. Visual detail of diagnostic characters increases with the size of the specimens being imaged.

2. Complexity of the group. Some specimens will be unidentifiable, regardless of the quality of the image, because the group is geographically, morphologically or behaviourally complex. Non-morphological or non-visual characters, such as internal genitalia, genetics or behaviour, may be required to differentiate many species.
3. Taxonomic understanding of the group. A specimen that belongs to a group that is taxonomically poorly known and/or understood will be difficult to identify to species from an image alone. However, increased levels of curation (e.g. family level to genus level) can be achieved in almost all groups.

Images of drawers from sections of the collection that are being actively curated or revised are at risk of becoming obsolete. The imaging workflow should allow for versioning of images. Furthermore, each drawer is uniquely identified with barcodes so that changes as a result of curation or revision are captured and the drawer is flagged for re-imaging.

Insect specimen metadata

Emerging technology that can extract specimen level metadata from images of whole-drawers, specimens and specimen labels will revolutionise digitisation of entomological collections. While whole-drawer images comprised of large specimens may facilitate species identification, images of small specimens have a higher probability of revealing useful and extractable label data. This is illustrated by Figure 4, which shows an unsorted drawer containing both large and small specimens. The small specimens are hard to identify; however, as Figure 5 shows, the labels associated with smaller specimens are almost completely unobstructed from view. It is hoped that, in the future, specialised software will be capable of scanning the image, extracting and recognising the printed text associated with specimens, and automatically creating a searchable database record.

Specimens for which label data are obscured may benefit from the use of barcodes or QR codes. These codes contain the specimen metadata, are small and thus conserve space in a drawer or unit tray, and can be easily read from the specimen itself, or an image of the specimen, using a smart phone with the appropriate software. Figure 6 provides an example of a QR code attached to a large insect, with label data that can be accessed from an image:

Loan requests

Requests for loans of material from entomological collections are a resource-intensive process. When a request is received, collections staff assess whether relevant material is available (that is, a significant proportion of the material may be unsorted or unidentified), make value judgements on which material is suitable for loan (for example, damaged specimens would not be acceptable, while type specimens are often excluded from

Figure 5. Inset from previous figure (Figure 4). Label data attached to small specimens is often almost completely readable. Therefore, specimen metadata could be extracted and digitised using specialised character recognition software.

loan requests), complete the appropriate loan and permit paperwork, and securely pack and post the specimens (postage represents a significant expenditure for many large and active collections).

In some cases, the borrowed material does not match the needs of the requestor (for example, the material has been incorrectly identified, or was collected from irrelevant localities). Some loans may consist of up to tens of thousands of individual specimens, requiring days or weeks of preparation.

High-resolution whole-drawer images provide a 'virtual collection' for researchers to access and browse for specimens of interest. The images are detailed enough for potential borrowers to judge for themselves if relevant material exists, and whether they wish to request a loan. This delivers a number of savings to the lending institution:

1. Staff are not required to spend time searching the collection for relevant material.
2. If relevant, loanable material is available, the borrower can use a whole-drawer image to indicate precisely which specimens s/he wishes to borrow.
3. Large loans can be accompanied by images of the specimens, negating the need to provide detailed written lists of material on loan forms. This is also useful for tracking overdue loans or partial returns.

For example, in 2011 the ANIC received an enquiry regarding Buforaniidae grasshoppers. The ANIC holds 12 drawers of this taxon, which were imaged and provided

Figure 6. Specimen with QR Code containing label data. A smart phone with the appropriate software can read and access the label data for this specimen from the image.

on-line to the enquirer. Figure 7 shows a curated drawer arranged by species, and then by the State from which the individuals were collected. The enquirer was interested in the geographical distribution of the ANIC specimens; therefore, in this example the whole-drawer images provided all the required information. At this time, no loan was required, no further correspondence was necessary and the whole-drawer images of this group are available online for future enquiries or requests for material.

Collection auditing

Perhaps unsurprisingly, large entomology collections struggle to develop and implement practical auditing and inventorying procedures. Large numbers of individual

Figure 7. Ultra high-resolution image of Buforaniidae grasshoppers (Orthoptera) from the ANIC. Note that the specimens are arranged by species, and then by the State from which they were collected. In this example, Northern Territory specimens are pinned in the first and second columns, followed by Queensland specimens in columns three and four. The online version of this image is viewable at Morphbank-ALA.

specimens (often numbering in the millions) combined with significant gaps in taxonomic knowledge and understanding of invertebrate groups results in a challenging collection management environment. Add to this, continued annual collection growth that may contribute to backlogs of unaccessioned material.

A recent audit of the Australian Museum by the Office of the New South Wales Auditor-General (2010) highlighted three key recommendations: (1) prioritise the collections, (2) tighten inventory control and (3) plan major catch-ups on legacy material. Whole-drawer imaging provides a means for implementing all three of these recommendations.

1. Prioritise the collections.

Resourcing for collection management and development is becoming increasingly limited; therefore, it is critical that the available resources used according to a set of priorities. The Smithsonian Curation Standards and Profiling System (McGinley 1993) assigns a curation standard to individual drawers and is used to calculate a collection health index (CHI). Whole-drawer images provide a means for calculating the CHI and tracking CHI as it changes over time.

2. Tighten inventory control.

Inventory control allows risk assessment in collections. Whole-drawer images can be used to:

- Develop a map of the general locations of specimens in the collection;
- Pin-point specimens that might be considered high-risk (e.g. high monetary value in a commercial market) or high-priority (e.g. holotypes or taxa represented by a single specimen); and
- Create a visual base-line inventory to serve as a basis for future inventory control.

3. Plan major catch-ups on legacy material.

Legacy collection material or backlogs of unaccessioned specimens are at risk from neglect (such as being misplaced or damaged by pests), becoming disassociated from vital collection data (such as field note books), or not being at a curatorial level where they can be made available to experts for revisionary study or further identification. Images of drawers and boxes of legacy material makes specimens "accession-ready" by:

- Improving visibility within the collection, and
- Simplifying the accession process when resources and/or expertise become available.

Morphometric analysis of specimens.

Measurement of insect morphological characters can be done directly (on a physical specimen using callipers), or indirectly (on an image of a specimen using image analysis software). Direct measurement places specimens at increased risk of damaging through handling and the close proximity of measuring tools. Indirect measurement removes these risks but increases the risk of measurement error due to the positioning of specimens at angles other than perpendicular to the camera lens (projection distortion).

A recent pilot study was conducted in the ANIC to investigate the comparative error rate associated with direct and indirect morphometric analysis of dragonfly wings (Mantle, unpublished data). Wing length of individual dragonflies was measured using three different methods: (1) with callipers on the pinned specimen in the drawer, (2)

with callipers on wings that had been dissected from the specimen and slide-mounted, and (3) on a whole-drawer image of the dragonflies (Figure 8).

Preliminary results are encouraging and suggest that, despite variable specimen positioning, there are no significant differences between direct and indirect measures of wing length. In addition, indirect measurement on whole-drawer images was significantly faster (hours rather than days) than measurements taken from individual specimens *in situ*.

Public engagement with biological collections.

Drawers of curated insect specimens elicit wonder and delight from members of the community. Some institutions can capitalise on the community's fascination with insects through public exhibitions and educational programs. The ANIC, however, is a research-only facil-

Figure 8. Whole-drawer image of dragonfly specimens used for a pilot study investigating the error associated with direct and indirect measures of morphological characters, such as wing length.

ity that does not have front-of-house, or public displays. Delivering high-resolution whole-drawer images of some of the most visually attractive specimens from the collection may:

- Improve public engagement with the research activities of the collection;
- Increase the collection's profile within the broader community; and
- Provide a platform for delivery of virtual education and outreach services.

Furthermore, opportunities exist to collaborate with, and add value to, existing online public resources. For example, whole-drawer images illustrating various insect families could be linked to the "What Bug Is That?" interactive key (http://anic.ento.csiro.au/insectfamilies/) and to galleries of insect taxa in the Atlas of Living Australia (www.ala.org.au). Crowd-sourcing is another initiative used to actively engage the community in natural history collections by facilitating the digitisation of insect collections through online "volunteer portals" (see http://volunteer.ala.org.au/).

Conclusions

High-resolution whole-drawer imaging of the ANIC specimens has been beneficial to both the collection and its users. The project is improving curation and auditing processes by providing a mechanism for tracking specimens through space and time. Engagement with researchers has improved because the metadata available from whole-drawer images adds value to correspondence about specimens. Consequently, the imaging project will continue and it is estimated that every drawer will be available for viewing online by 2015.

Acknowledgements

The authors would like to thank the ANIC imaging volunteers: Andrew McKenzie, Emily O'Connor, Fiorella Esquivel and Frederick Michna for their assiduous efforts in preparing and imaging the insect collection drawers; Laura Johnson for the dragonfly images and measurements; Peter Brenton for assistance with delivering the images and associated metadata through Morphbank-ALA; and Murray Fletcher for identifications of Hemiptera. We are grateful to the Atlas of Living Australia for providing funding to support the purchase of equipment for this project. We also thank Matt Bertone, Stefan Schmidt and Vladimir Blagoderov for their critical reading and valuable comments in improving the contents of the paper.

References

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 (2005) ZooBank: the open-

access register for zoological taxonomy: Technical Discussion Paper. Bulletin of Zoological Nomenclature 62: 210–220.

Beaman R, Macklin JA, Donohue MJ, Hanken J (2007) Overcoming the digitization bottleneck in Natural History Collections: A summary report on a workshop held 7-9 September 2006 at Harvard University. http://www.etaxonomy.org/wiki/images/b/b3/Harvard_data_capture_wkshp_rpt_2006.pdf [Accessed 29 February 2012]

Bertone MA, Deans AD (2010) Remote curation and outreach: examples from the NCSU Insect Museum GigaPan Project. In: Proceedings of the First International Conference on Gigapixel Imaging for Science, November 11–13 2010.

Blagoderov V, Kitching I, Simonsen T, Smith S (2010) Report on trial of SatScan tray scanner system by SmartDrive Ltd. Available from Nature Precedings http://hdl.handle.net/10101/npre.2010.4486.1 [Accessed 25 February 2012]

Drew J (2011) The role of natural history institutions and bioinformatics in conservation biology. Conservation Biology 25(6): 1250–1252. doi: 10.1111/j.1523-1739.2011.01725.x

Johnson N (2012) A collaborative, integrated and electronic future for taxonomy. Invertebrate Systematics 25: 471–475. doi: 10.1071/IS11052

Mantle BL, Fisher N, La Salle R J (2011) Whole drawer imaging for curation and management of the ANIC. TDWG 2011 Conference, New Orleans, Louisiana, USA. http://www.tdwg.org/fileadmin/2011conference/slides/Mantle_SatScanANIC.pdf [Accessed 15 March 2012]

McGinley RJ (1993) Where's the management in collections management? Planning for improved care, greater use, and growth of collections. International Symposium and First World Congress on the Preservation and Conservation of Natural History Collections Vol. 3: 309–338.

New South Wales Auditor-General's Report (2010) Knowing the Collections: Australian Museum Performance Audit. http://australianmuseum.net.au/document/Knowing-the-Collections-audit-report [Accessed 15 March 2012]

Vollmar A, Macklin JA, Ford L (2010) Natural history specimen digitization: challenges and concerns. Biodiversity Informatics 7: 93–110.

The British Natural History Museum Entomology Department webpage: http://www.nhm.ac.uk/research-curation/departments/entomology/ [Accessed 2 March 2012]

The Smithsonian National Museum of Natural History Department of Entomology wepage: http://entomology.si.edu/ [Accessed 2 March 2012]

The Australian National Insect Collection webpage: http://www.csiro.au/en/Organisation-Structure/National-Facilities/Australian-National-Insect-Collection/ANIC-Profile.aspx [Accessed 2 March 2012]

Appendix 1

List of identifications provided by Dr Murray Fletcher based on high-resolution Figure 4. Identifications presented in the following order:

- Unit trays 1–5 in the upper row, left to right;
- Unit trays 6–10 in the lower row, left to right;
- Rows 1–n from top to bottom in each unit tray, left to right along each row; and
- Individual specimens separated by a comma.

Box 1	
Rows 1–3	Large Cercopidae from Malaysia
Row 4.	2 as above, ?, possibly Neuroptera, poss. Neuroptera, ?
Row 5.	?, ?, ?, *Amarusa australis* (Jacobi) (Cercopidae: Aphrophorinae), ?, ?, ?, ?
Row 6.	Flatidae, ?, Membracidae, Heteroptera, Heteroptera, Heteroptera
Row 7–10.	all Heteroptera
Box 2	
	All large Malaysian Tessaratomidae
Box 3	
Row 1.	2 × Tessaratomidae
Row 2.	Reduviidae, Reduviidae, Pentatomidae, Pentatomidae
Row 3.	Pentatomidae, Pentatomidae
Row 4.	Alydidae, 2 × *Agonoscelis rutila* (Pentatomidae)
Box 4	
Row 1.	*Mutusca brevicornis* (Alydidae) according to the label
Row 2.	3 *M. brevicornis*, + 1 scutellerid
Row 3–6.	all large Tessaratomidae
Box 5	
	All large exotic Tessaratomidae
Box 6	
	All large exotic Tessaratomidae, lower ones are nymphs
Box 7	
Row 1.	Scutelleridae, Heteroptera, Heteroptera, Heteroptera, ?, ?, ?
Row 2.	?, ?, ?, ?, ?, *Thanatodictya* sp (Dictyopharidae)
Row 3.	Flatidae (possibly *Colgar* sp.)
Row 4.	Alydidae, Heteroptera, Scutelleridae
Row 5.	Achilidae: Plectoderini, ?
Row 6.	?, ?, ?
Row 7–9	lots of little things. Last one in Row 9 might be *Dascalina* or *Massila* (Flatidae)

Box 8	
Row 1.	?, ?, ?
Row 2.	?, ?, Heteroptera
Row 3.	Scutelleridae
Row 4.	Ledrini (not Australian)
Row 5.	3 × Tessaratomidae nymphs
Box 9	
Row 1.	?, ?, ?, ?, ?
Row 2.	4 × Pentatomidae, ?
Row 3.	5 × Pentatomidae, ?, ?
Row 4.	?, Membracidae, Heteroptera, Heteroptera, Auchenorrhyncha
Row 5.	4 × Pentatomidae, Iassini, ?
Row 6.	?, ?, ?, Pentatomidae
Row 7.	Pentatomidae, Pentatomidae, ?, ?, ?, ?
Row 8.	?, ?, ?, ?, Heteroptera, Pentatomidae, Pentatomidae
Row 9.	3 × Pentatomidae
Box 10	
Rows 1–2.	Tessaratomidae
Row 3.	2 × Pentatomidae
Row 4.	2 × Pentatomidae
Row 5.	2 × Heteroptera
Row 6.	Heteroptera, ?, ?, ?, ?, ?, Pentatomidae
Row 7.	Scutelleridae, Pentatomidae, ?, ?
Row 8.	?

ZooKeys 209: 165–181 (2012)
doi: 10.3897/zookeys.209.3571
www.zookeys.org

RESEARCH ARTICLE

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

InvertNet: a new paradigm for digital access to invertebrate collections

Chris Dietrich[1], John Hart[2], David Raila[2], Umberto Ravaioli[3,4], Nahil Sobh[4], Omar Sobh[1], Chris Taylor[1]

1 *Illinois Natural History Survey, Prairie Research Institute* **2** *Department of Computer Science* **3** *Department of Electrical and Computer Engineering* **4** *Beckman Institute for Advanced Science and Technology, University of Illinois, Champaign, IL 61820 USA*

Corresponding author: *Chris Dietrich* (dietrich@inhs.uiuc.edu)

Academic editor: *V. Blagoderov* | Received 22 June 2012 | Accepted 26 June 2012 | Published 20 July 2012

Citation: Dietrich CH, Hart J, Raila D, Ravaioli U, Sobh N, Sobh O, Taylor C (2012) InvertNet: a new paradigm for digital access to invertebrate collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 165–181. doi: 10.3897/zookeys.209.3571

Abstract

InvertNet, one of the three Thematic Collection Networks (TCNs) funded in the first round of the U.S. National Science Foundation's Advancing Digitization of Biological Collections (ADBC) program, is tasked with providing digital access to ~60 million specimens housed in 22 arthropod (primarily insect) collections at institutions distributed throughout the upper midwestern USA. The traditional workflow for insect collection digitization involves manually keying information from specimen labels into a database and attaching a unique identifier label to each specimen. This remains the dominant paradigm, despite some recent attempts to automate various steps in the process using more advanced technologies. InvertNet aims to develop improved semi-automated, high-throughput workflows for digitizing and providing access to invertebrate collections that balance the need for speed and cost-effectiveness with long-term preservation of specimens and accuracy of data capture. The proposed workflows build on recent methods for digitizing and providing access to high-quality images of multiple specimens (e.g., entire drawers of pinned insects) simultaneously. Limitations of previous approaches are discussed and possible solutions are proposed that incorporate advanced imaging and 3-D reconstruction technologies. InvertNet couples efficient digitization workflows with a highly robust network infrastructure capable of managing massive amounts of image data and related metadata and delivering high-quality images, including interactive 3-D reconstructions in real time via the Internet.

Keywords

Collection digitization, collection database, image processing

Background and introduction

Invertebrate collections present one of the greatest challenges to automated specimen digitization. Not only do they represent the majority of known species and comprise the largest numbers of available specimens, but they also present a number of logistical problems that have, so far, frustrated attempts to develop automated digitization and data capture workflows.

Insect collections, which constitute the largest extant collections of invertebrate specimens, are particularly challenging. In most, a majority of the prepared specimens are pinned. Dry, pinned insect specimens, when properly housed and protected from direct sunlight, high humidity and pests (e.g. dermestid beetles), may last indefinitely. Many European museum collections include pinned specimens collected centuries ago that remain intact and useful for comparative morphological study. However, even recently collected pinned insect specimens are often extremely fragile and easily damaged through handling. Moreover, to conserve space, many curators have packed specimens very densely into unit trays and drawers, such that adjacent specimens are nearly touching each other or even overlapping. Thus extreme care must be taken in moving specimens because legs, wings or antennae may easily be broken off if specimens are brushed against one another.

Specimen data obtainable from pinned insect specimens consist of the information (morphological and otherwise) embodied in the specimens themselves, and data (metadata) printed on one or more small data labels attached to the pin below the specimen. Specimen labels include information such as the collection locality, date, and name of collector, and the determined scientific name. These may be difficult to read and interpret because of their small size, the use of non-standard abbreviations, illegible hand-writing, and/or because they may be partly or completely obscured from above by other labels and/or by the specimen itself.

The traditional approach to digitization of insect collections (reviewed by Johnson 2007, 2009) has focused almost entirely on label data capture, retrospective georeferencing, and the assignment of unique identifiers to individual specimens. The usual workflow involves manually keying in data from specimen labels and attaching a unique identifier label (machine-readable barcode and/or human readable number) to each specimen. This approach is problematic for several reasons. It is time-consuming--one reason why so many existing specimens still need to be digitized. It is expensive, with per-specimen costs estimated at US$1 or more in some recently completed or ongoing projects (Vollmar et al. 2010, Heidorn 2011 and unpublished data). It is error-prone, with typographical or other mistakes often introduced during the process of label data interpretation and transcription. It also entails substantial risk of specimen breakage due to handling, particularly if the work is being performed (as it often is) by poorly paid student technicians with little collection management experience.

Thus, the major challenges for InvertNet and similar projects are to bring the per-specimen cost of digitization down without sacrificing accuracy of data capture or risking damage to irreplaceable specimens. Indeed, the NSF ADBC program, the

source of funding for InvertNet, mandates that the average cost per specimen for digitization, program-wide, including both imaging and label data capture, be kept at or below US$0.10. ADBC aims to digitize 1 billion specimens in 10 years for a total budget of US$100 million.

Despite the problems noted above, one aspect of pinned insect collections that may prove advantageous to automated mass-digitization methods is that the specimens are usually mounted and arranged in a consistent orientation and multiple specimens of the same taxon are usually grouped together, side-by-side, within the same collection storage unit. Thus, high-resolution digital imaging methods can be used to capture images of large numbers of specimens simultaneously, thereby drastically reducing the per-specimen cost of obtaining specimen images. Other recent projects have already used this approach to acquire images of collections of pinned specimens very quickly and cheaply (Bertone and Deans 2010 and this volume, Blagoderov et al. 2010 and this volume). Immediate access to the images may then be provided via the Internet, which, in turn, may facilitate at least partial acquisition of specimen metadata (i.e., label data) by the broader community of potential users.

Some problems remain to be addressed, however. These include the need to acquire specimen-level label data and to assign unique identifiers that allow individual specimens to be tracked. Top-down images of whole drawers of pinned insects allow users to view some specimen label data, but labels are often at least partly obscured by the specimens. In cases where series of specimens from the same collection lot are placed together, it may be possible to assemble all the label data by examining different specimens in the series because different parts of the labels of different specimens may be visible. 3-D reconstructions that allow virtual tilting of drawers or specimens may reveal parts of labels obscured in a strictly top-down view. Unfortunately, even 3-D reconstructions will not allow labels placed beneath the top label on the pin to be viewed if the labels are pushed together. Use of even more advanced technologies such as micro CT scanning may eventually allow data to be captured from labels that are completely obscured by specimens or other labels, but at present, such data are accessible only through physical manipulation of specimens and labels. In such cases, the added value of gleaning this additional information needs to be balanced against the risk to the specimen posed by physical handling. Fortunately, for most specimens, a large proportion of the crucial occurrence data are printed on the top label and, because most insect specimens are small, these labels may be read without physically manipulating the specimens themselves. Examination of gigapan images (see gigapan.org) of whole drawers of pinned insects from the North Carolina State University insect collection indicates that more than 75% of the drawers and ca. 90% of the specimens imaged have text on the top label visible; this label usually comprises at least the locality name and, in most cases, also the date of collection and name of collector. Because the arrangement of pinned specimens in the NCSU collection is typical for insect collections in general (at least in the USA), large amounts of species occurrence data should be obtainable directly from high quality images of entire drawers. We estimate that 3D reconstructions that allow virtual tilting of images with

similar resolution will increase the amount of label data exposed by at least 50%, i.e., by exposing more of the top label when it is partly concealed by the specimen and by exposing labels attached farther down on the pin.

Specimen tracking is another problem that may be difficult to overcome with mass specimen digitization approaches. Recently it has become standard practice for curators to attach separate barcode or other unique identifier (UID) labels to individual specimens as part of the specimen data capture/digitization workflow (Johnson 2009). In our view, the risk of specimen damage posed by attaching such labels may outweigh the need to uniquely identify each individual specimen, especially if the specimen is being handled *only* for the purpose of attaching the barcode label. A better approach might be to attach UID labels to specimens only when the specimens need to be handled for another purpose, e.g., when being transferred into a shipping container during loan processing, or when being sorted and identified by a taxonomist or curator. Because the only value of attaching a physical UID label to the individual specimen is to facilitate tracking of the specimen after it has been moved from its original location in the collection, we recommend that curators not add UID labels to specimens until they need to be moved for other reasons. Prior to being moved, individual specimens in digitized drawers and unit trays may be digitally mapped based on their physical locations. A specimen record may then be created in the collection database and include a unique identifier and information on its location, in addition to data from the specimen labels. The unique identifier, thus assigned, will remain a virtual UID until the specimen needs to be moved, at which point a physical label may be printed and attached to the specimen. Alternatives to ink-on-paper UID labels, such as passive Radio Frequency Identifier (RFID) tags (which may be pinhead sized and have recently become quite affordable) should also be explored. Because RFID tags (unlike barcodes) do not need to be visible in order to be detected and scanned, they offer the added advantage of further reducing the need for physical manipulation of specimens. They also offer the possibility of developing Augmented Reality (AR) systems capable of physically mapping the locations of specimens in three-dimensional space (e.g., within a drawer, cabinet or collection range) using radio telemetry.

Recent advances in high-throughput insect specimen imaging

Most recent collection digitization initiatives that include an imaging component have focused on capturing images of individual specimens (e.g., Lampe et al. 2005, Enriquez 2011, Ball et al. 2011, Eades et al. 2012, Harman et al. 2011, Haüser et al. 2005, Kjar et al. 2012). While this approach may have the potential advantage of producing very high-quality images of individual specimens, it also requires physical manipulation of the specimens, which entails risk of specimen damage and has a high per-specimen cost. Most digitization initiatives that have adopted this approach have focused only on high value collection holdings (e.g., type specimens). A cost/benefit analysis of this approach needs to be undertaken, since the risk of damaging such specimens during the

digitization process must be weighed against the benefits gained by providing access to the digital images (e.g., how often is a particular research need addressed by access to the image alone, rather than to the specimen itself?). If such high quality images of individual specimens are being captured for other purposes (e.g., for publication in a taxonomic paper), they should be archived and associated with the collection database record for that particular specimen.

Recent advances in digital gigapixel imaging allow images of entire drawers of pinned insects to be captured. Multiple neighboring images can be "stitched" together into a single "panoramic" image. This stitching operation is enabled by recent advances in computer vision, and relies on finding matching features in the overlapping regions shared by neighboring images. By capturing multiple high resolution images and stitching them together into a single panorama, drawers containing thousands of specimens may be digitized very rapidly and the quality of the final images may be very high. This method was used successfully at North Carolina State University (NCSU) in a recent NSF-funded project (Bertone and Deans 2010) and suggests a promising pathway toward more efficient methods for mass imaging and digitization of pinned insect (and other) collections.

Using the GigaPan robot (gigapan.org) combined with a consumer-grade digital camera, the NCSU team was able to capture images of their entire collection, comprising >2700 drawers within just a few person/months and make these high quality images available to the public via the GigaPan website. The web interface allows users to view images of entire drawers and zoom in onto individual specimens, such that label data (when not obscured from above by large specimens) and details of the morphology of the specimens may be seen. Hand-entering data for each of the approximately 2 million specimens in the NCSU collection, using traditional methods, would have required many person-years of effort. The GigaPan project provided rapid access to the entire collection.

One problem with the NCSU/GigaPan digitization methodology is that it provides only limited access to specimen label data (capture of label data was not one of the stated goals of the project). Only the label data not obscured by the specimens may be extracted from the GigaPan images and the data are neither available as text, nor have they been parsed into the standard Darwin Core database fields (http://rs.tdwg.org/dwc/) to facilitate automated searching of particular data elements. Another problem is the distortion introduced into the stitched gigapixel images caused by the fixed position of the robot-mounted camera over the center of the drawer. During image capture, the robot tilts the camera from front to back and side to side, such that the edges of the drawer are photographed at an angle while the center of the drawer is photographed with the lens pointing directly downward. The resulting stitched images show a pronounced fish-eye effect (barrel distortion) with the sides of the drawer bowed outward. Stitching software (e.g., Hugin open-source stitcher; http://hugin.sourceforge.net/) exists that includes tools to correct for this distortion to some extent, but it is difficult to remove all distortion from the stitched image if the original images from which the stitched image is constructed are themselves highly distorted.

The SatScan system implemented at the Natural History Museum, London (Blagoderov et al. 2010), uses an alternative technology that overcomes the distortion problem. In this system, the camera does not tilt but moves horizontally, capturing images all from the same angle but at different X/Y positions over the drawer. Images produced by this system have similar levels of resolution to those obtained in the NCSU/GigaPan project, but the drawer images produced by SatScan are free of distortion, even toward the edges of the drawer (http://sciaroidea.info/node/44309).

Still the problem of capturing label data persists. Although labels attached to insect specimens are usually very small, most insects are also small, so, for a large proportion of pinned specimens in collections, label data are at least partially visible from above. As any insect taxonomist knows, it is usually possible to see more (sometimes all) of the label(s) simply by tilting the drawer or otherwise viewing the specimens from an angle. This can be seen in many of the NCSU GigaPans (http://www.gigapan.org/profiles/ncsuinsectmuseum), where the labels of specimens toward the edges of the drawers are more exposed than those near the center, simply as an artifact of the GigaPan image capture protocol. An improved system that maximizes visibility of the labels, in situ, would simply need to capture images of the drawer from multiple perspectives, including different horizontal positions over the drawer (*à la* SatScan) as well as different angles (*à la* GigaPan). Technologies for combining such images to create 3-D reconstructions can then be used to allow virtual tilting, maximizing the user's ability to read the data on labels partly obscured by the specimens or by other labels. This is the approach we envision using for InvertNet.

The InvertNet approach

Our efforts to implement robust, rapid and cost-effective solutions for mass digitization of invertebrate collections focus on four main areas: 1) use of improved image capture hardware; 2) application of improved image processing and visualization methods; 3) development of user-friendly, semi-automated workflows; and 4) establishment of robust cyberinfrastructure for data ingest, storage and delivery.

Improved image capture hardware

The primary goals of an ideal capture system include:

1. The system should be as automated as possible to minimize operator activity and therefore human error;
2. It should capture an array of high resolution images from multiple viewpoints, to support zooming in to reveal specimen detail, viewing otherwise occluded portions of pin labels, and 3-D reconstruction;
3. It should be inexpensive to purchase, operate, and maintain;

4. It should be flexible to adapt to operator and scientific feedback from operations when deployed.
5. It should be upgradable once deployed, to take advantage of improvements in imaging technologies (sensors, processing, etc.) as they become available.

We have investigated three options for capturing such images. We first investigated combining multiple GigaPan-style panoramas from different viewpoints, such as from four corners of the specimen drawer, and using prost-processing to create composite images. However, this can increase time, effort, and the probability of human error if the drawer and/or camera must be re-positioned manually during processing of a single tray.

A more reliable option used a robotic camera positioning system based on a modified Computer Numerical Control (CNC) machine (similar to a plotter, except with the pen replaced by a camera) to position camera/lens precisely and repeatably in an x-y grid to complete the panorama. Such robotic systems are capable of moving tools (including cameras) rapidly and precisely in three dimensions offer great advantages in terms of adaptability by programming various capture "recipes" based on tray geometry, specimen layout, and specimen scale and density within a tray.

In order to minimize distortion between neighboring images and reduce stitching artifacts, we use a telecentric lens that captures an orthographic (not perspective) projection of the image on the sensor. The telecentric lens shoots the same image area regardless of how far away it is, and one cannot enlarge or reduce the area being photographed by moving the camera closer or farther away. This is beneficial for measurements, image processing and stitching, but precludes the use of neighboring image overlap processing for multiple view (3-D) and occluded label processing.

To accommodate these multiple viewpoints, we extended the CNC camera positioning machine with a computer controlled pan-tilt mechanism that provides the ability to capture grids of overlapping images at various positions, and also at various oblique angles in order to simultaneously support accurate panorama generation, 3-D reconstruction, and occluded label capture.

CNC systems were developed for machining dense materials with industrial power tool heads. They are large and heavy, often hundreds of pounds. Furthermore the physical size of the moving parts of such machines complicates lighting, as large machine parts move through the path of lighting sources during capture, altering lighting conditions and casting shadows which can affect feature matching algorithms such as panoramic image stitching. Because of their industrial development for machining, CNC machines are large and not able to be easily disassembled, massive – hundreds of pounds, require high power, and are not easy to move, ship, and locate in a laboratory setting.

We are currently testing a more lightweight prototype that is based on the Delta Robot (http://en.wikipedia.org/wiki/Delta_robot), which resembles a three-legged spider that suspends the camera over the specimen drawer, with much less hardware to interfere with fixed lighting systems. These robotic systems are very fast and accurate, and are used in "pick and place" factory lines for purposes such as picking items and aligning them for packaging. Such a machine is inexpensive to build and can be pro-

grammed to accomplish very rapid, precise and complex movements (for example see: https://www.youtube.com/watch?v=foTE0Mau5a8). We are currently working with a 3-arm design with additional pan/tilt motors that allow the camera to be rotated in addition to precisely placed in x-y-z position over the drawer of pinned specimens. The machine is far less massive when compared to a CNC style system - tens of pounds, and is easily disassembled and reassembled without machinists tools and expertise. This facilitates shipping, lab positioning, movement, and physical requirements of the system.

Stitching software

Software capable of combining multiple images into a single panorama is now widely available. The GigaPan software system, used successfully in the NCSU digitization project, is one example. One current disadvantage of the GigaPan software is that it requires that final, stitched images be posted to the GigaPan.org website in order to be viewed and manipulated via the Internet. Open-source stitchers (e.g., Hugin, OpenCV) and Zoomable User Interfaces (ZUIs) such as Zoomify, required to view and manipulate the image are now also available and provide greater flexibility for the development of customized interfaces and workflows (see below).

Stiching algorithms rely on feature detection and matching across the raw images, which can be computationally demanding for large numbers of images. Two of our team, Hart and Raila, are participants in the Illinois-Intel Parallelism Center (I2PC) which is focused on new multicore parallel computing architectures, techniques, and tools. In collaboration with I2PC we are exploring parallel implementations of stitching codes. Results to-date have shown order of magnitude performance increase (from 500 seconds to 40 seconds) on modern commodity desktop computer systems, and we believe that the next generation of processors should accelerate the performance to levels that should not impede the workflow of digitization when run on commodity systems, but the stitching codes are also able to be run on large scale super-computing systems within the server-side of the InvertNet infrastructure if needed.

3-D Reconstructions

In addition to providing a means for creating distortion-free 2-D gigapixel images of entire specimen drawers, by using advanced hardware to vary the viewpoint and direction of image capture, we enable two new and exciting capabilities. From different vantage points, we can better see beneath the specimens to better capture the data from the labels pinned below them, and images from multiple view directions can be used to reconstruct 3-D models of the specimens themselves, potentially facilitating capture of more morphological data than is possible using 2-D, top-down images. We have tested multi-view stereo (MVS) reconstructions on specimen capture images and reconstructed 3-D models from them. MVS takes a pair of photographs from two different

viewpoints and "rectifies" them, distorting them so corresponding points in each image have the same "y" coordinate. It can then search along horizontal lines for these matching points and uses the disparity in their alignment to estimate their distance from the viewer. Such estimates can be error prone and require further smoothing. Our current MVS reconstruction is based on a state-of-the-art algorithm developed by Disney Research Zurich for reconstruction of human faces for feature film production (Beeler et al. 2010). However, the smoothing designed for facial geometry does not work well on the insect specimens tested so far and we are researching new methods that work better on the dark, sparse and fine features from high-resolution invertebrate images.

Digitization of other kinds of specimen storage units

Invertebrate collections consist not only of pinned specimens stored dry in drawers and unit trays, but also include fluid (usually ethanol) preserved specimens in vials or jars, and specimens mounted on microscope slides. The methods described for capturing images of whole drawers may be extended to these other storage types. Images of multiple slides or jars may be captured simultaneously and then segmented to facilitate data capture for individual units. This is the approach taken by another project at the Illinois Natural History Survey and University of Minnesota, recently funded by NSF (Tinerella 2010). Slide mounted specimens are perhaps the easiest to digitize: they may be treated as two dimensional objects and, because they are of standard size, individual slides may be imaged in groups placed in fixed positions on a tray and then segmented using a simple pixel map of the tray. Once digitized, the specimens and labels are clearly visible on the image and the image may then be used as a surrogate for the physical slide during subsequent label data capture. Following this approach, InvertNet is capturing images of 20 slides at a time by arranging them in fixed positions on a clear plastic template placed on the bed of a consumer-grade flatbed scanner. Images captured in this way are of sufficient quality to reveal label text and the general condition of the specimens but, in most cases, not good enough to reveal details of specimen morphology sufficient for species identification or morphological study. A variety of automated systems are available commercially for digitizing collections of microscope slide-mounted specimens, combining robotic slide loaders with high quality microscopes or scanners (Rojo et al. 2006) but, to our knowledge, none have yet been applied to large-scale digitization of slides in natural history collections.

Fluid-preserved specimens in vials present a greater challenge. Multiple specimens are often stored in the same vial and the orientations of specimens and labels vary among vials. Views of vial contents are distorted by the refractive properties of the glass and fluid and the labels may obscure the specimens, or vice versa, to greater or lesser extent. Complete digitization of ethanol-preserved specimens now requires laborious removal of the specimens from the vials so that they may be spread apart and imaged. We are experimenting with methods for capturing images of multiple vials simultaneously. At present, the relatively low-cost proposed approach for InvertNet uses a flatbed scanner to cap-

ture images of multiple vials simultaneously using customized vial racks with clear sides. Racks containing vials are oriented so that as much as possible of the label(s) in each are in view, the racks are then placed on their sides on the scanner bed and scanned (Fig. 1). The racks are then flipped over (180 degrees vertically) to capture a second image of the opposite sides of the vials. This approach allows entire collections of vials to be digitized quickly because handling is minimal. It also reduces distortion of labels and specimens because placing the vials on their sides causes these objects to float down and rest against the glass. The main disadvantage of this approach may be the failure to expose/capture all label data if multiple labels are included in a vial and/or labels are oriented in such a way that the text cannot be seen. Also, in most cases, images of the specimens themselves will not be of high enough quality to facilitate species identification or morphological study. In some cases, single specimens from lots of larger invertebrate species (e.g., crustaceans) may be removed from jars and imaged next to jars and labels. More advanced 3-D imaging technologies may eventually provide the means to capture and segment undistorted images of fluid-preserved specimens and labels in situ, although vials containing numerous individual specimens and/or labels will continue to present difficulties.

Figure 1. A set of three-dram vials scanned using a color flatbed scanner showing the front (left) and back (right) of the same set of vials. Note that the position of empty spacer vials (e.g., sixth from top in middle column) is the same, but inverted, in the two images because the vial racks are flipped vertically between scans. This relatively quick and inexpensive procedure exposes at least some label data for subsequent capture and reveals the general condition of specimens.

Invertnet mass digitization workflow

Combining the hardware and software technologies described above, InvertNet will implement a semi-automated workflow that is user-friendly, requires minimal training for the end user, and meets the goals of reducing the per-specimen cost of invertebrate collection digitization while minimizing risk of damage to the specimens. Design of the InvertNet workflow and user interfaces is underway and is addressing several important points.

1. **Ease of use.** Given the anticipated heavy use of the system by non-skilled workers (e.g., students) the capture hardware operation and data input workflows will be required to minimize errors, verify correct inputs, and support corrective measures.
2. **High performance capture and input.** The overall workflow should not be impeded by capture hardware, client-side processing, or data transfer. The operator should be able to work in a sustained manner.
3. **Fault resilience.** The workflow should not be impeded by transient network conditions between the worksite and the InvertNet website, which can manifest as network delay, connection failures, and off-line operations.
4. **Security.** Data, raw and processed, should not be lost in the capture and upload process and should be transferred from the capture site into secure storage as quickly as network connectivity permits.
5. **Flexibility.** The workflows and hardware should be adjustable to site-specific preferences such as batch processing, variations due to collection attributes, and in general be flexible.
6. **Maintainability.** The systems in participating sites will run identical software releases, be remotely supported and upgradable, and consistent across sites in hardware and software versions.

To support these goals we are implementing the following generic workflow:

1. **Capture Workstation Preparation.** Stage drawers to be digitized, power up capture station, preform calibration operation.
2. **Capture Operations.** Operator selects among capture recipes, inserts prepared drawer, initiates capture. When capture is complete, operator reviews real-time processed images for completeness and accuracy.
3. **InvertNet Login.** Operator logs into digitization software/portal within InvertNet "Digital Collections" space.
4. **InvertNet Input.** Operator creates capture record for each tray processed above, providing appropriate metadata into system, with automation support to avoid entering redundant data.

Cyber-infrastructure

Providing access to large digital collections of invertebrate specimens will require a robust, Internet-based, information technology (IT) infrastructure to store and provide access to the data and images via the Internet, and ensure that access to the data is maintained over the long term. To do this, InvertNet has implemented a cloud based infrastructure based on the open source cloud project OpenStack (http://openstack.org). This allows the InvertNet website and databases to be mirrored across web servers at multiple locations, which yields faster response times for users of the website and allows for rapid and complete disaster recovery.

Website and content management system

The InvertNet web site (Fig. 2) is built on a robust cyberinfrastructure platform called HUBzero. HUBzero was developed with NSF support and designed specifically to support the kinds of large-scale, massively collaborative scientific research platforms that the ADBC program aims to build. HUBzero was originally designed to support a large community of nanotechnology researchers, but has since been adopted by a wide variety of other communities of researchers. The main advantage of HUBzero over other open-source content management systems is that it integrates a traditional CMS (Joomla; (http://www.joomla.org/) with powerful and highly customizable tools for data sharing, data analysis, data archiving. This gives InvertNet the ability to customize both back-end and front-end components of our cyberinfrastructure to meet our users' needs for ingesting, processing, and visualizing digitized biological collections that include both traditional occurrence data and high-resolution graphics.

For example, to provide redundancy and preservation of contributed digital collections, we integrated HUBzero with an extensible cloud storage infrastructure (http://openstack.org), which allows us easily to scale up storage as the number of contributed collections increases, as well as spread storage resources over multiple redundant sites, improving security.

To facilitate ingest and management of large collections of specimen images and data, we integrated HUBzero with the Medici multimedia content management system (http://medici.ncsa.illinois.edu/). Medici is a flexible, extensible semantic system designed to support any data format and multiple research domains and contains three major extension points: preprocessing, processing and previewing. When new data are added to the system, whether directly via the web application or desktop client, or through web services, preprocessing is automatically off-loaded to extraction services in charge of extracting appropriate data and metadata. The extraction services attempt to extract information and run preprocessing steps based on the type of data. For example, in the case of images, a preprocessing step creates previews of the image and automatically extracts metadata from the image and assigns a persistent, globally unique identification (GUID). Medici allows users to manage and aggregate collec-

Figure 2. Current HUBzero-based InvertNet homepage showing top menu bar with content areas accessible to registered users.

tions comprising distributed sub-collections, track internal processing of resources and the creation of derived resources, provide GUIDs for resources suitable for citation and export metadata in globally understood standard formats. It also enables users to use desktop analysis tools via a remotely hosted web service in concert with the knowledge-space (i.e., digitized collections) without having to deal with download, installation, licensing, etc. Medici's web interfaces (Fig. 3) are highly customizable, which enables us to create custom forms for capturing various kinds of metadata for different collection objects (e.g., whole drawers of pinned specimens). By making the clients and preprocessing steps independent and using Resource Description Framework (RDF) as a common domain-neutral data representation, the system can grow and adapt to different user communities and research domains, HUBzero also supports the development and integration of data processing (e.g., image analysis) and analytical tools that will allow users to manipulate and analyze data directly within the InvertNet platform.

Coupling the Medici content repository system with HUBzero will enable InvertNet to act as a collaborative social platform that can scale effectively and allow for submission of image collections. It will incorporate the digitization workflows, image post-processing, databases, environments for community building and collaboration, analytical tools, developer tools, and tools for education and outreach. To our knowledge, no other platform or website/application combines all of these capabilities and features to date.

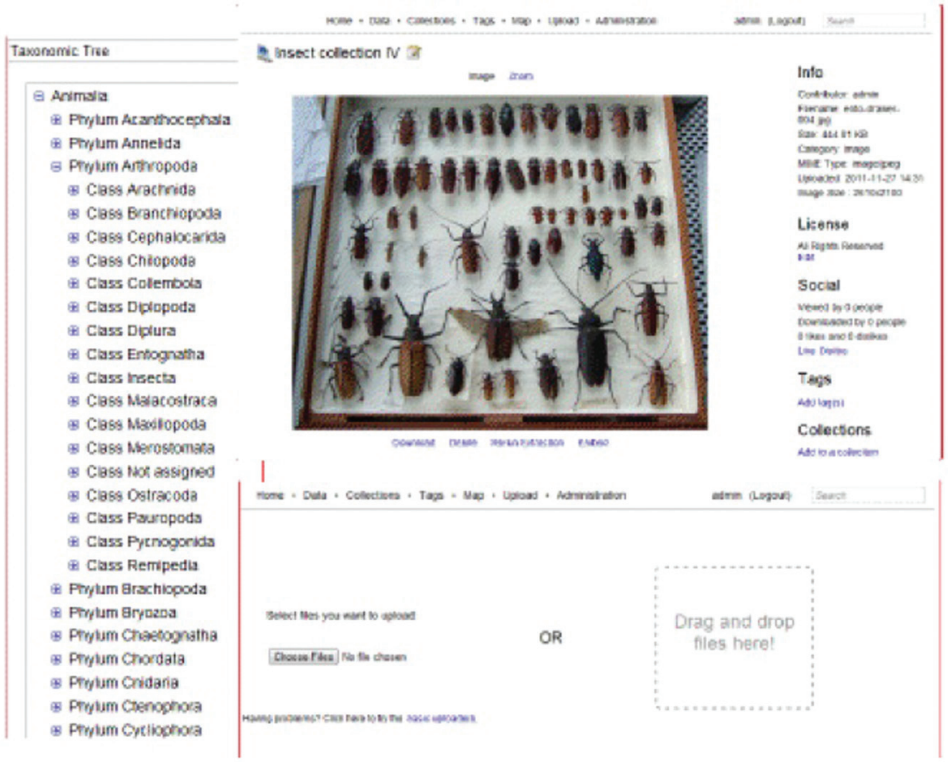


Figure 3. Current version of InvertNet's Medici multimedia semantic content management system interface, accessible from InvertNet digital collections tab on homepage, showing taxonomic tree, drag and drop file upload space, and zoomable user interface for viewing gigapixel images.

A few words on specimen-level label data capture

As already demonstrated by the NCSU GigaPan project, use of advanced imaging techniques can provide rapid access to large numbers of invertebrate specimen images and the variety of potential uses of such images in research and education have only begun to be explored. Nevertheless, current biological collection database standards require capture of data at the specimen level. Images of entire storage units (e.g., drawers) may be segmented using image analysis software with the images of individual specimens placed in separate database records (Fig. 4). Because most insect specimens are small, labels pinned beneath them are often visible and, if the image quality is sufficient, the text of such labels may be read and interpreted. Thus, at least partial specimen occurrence and taxonomic data may be obtained directly from the images of many specimens. Even more advanced image capture and reconstruction techniques than those produced by the GigaPan or SatScan systems, including those being incorporated into the InvertNet digitization workflows, should provide even greater access to specimen-level label data, given the capability these techniques provide for viewing specimens from multiple perspectives. However, attempts to

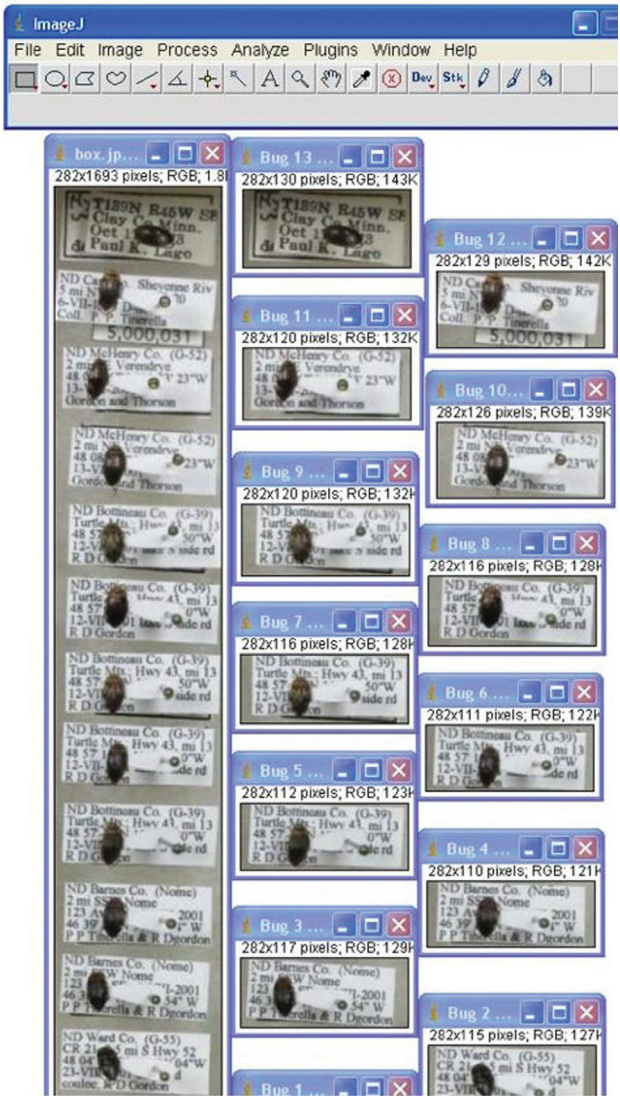


Figure 4. Image of multiple pinned insect specimens in unit tray (left) and same specimens segmented into separate files (right) using customized ImageJ image processing protocol.

further automate the process of reading and interpreting specimen labels have, so far, had mixed success. The performance of available optical character recognition (OCR) software tested so far on insect specimen labels is generally poor. In most cases, more time must be spent detecting and correcting errors than would be required simply to enter the data into the appropriate fields of a database by hand. At present, the crowd-sourcing/citizen science approach to label data capture (Hill et al. 2012, this volume) appears to be the most promising avenue for entering such data as text into a relational, standards-compliant database. We anticipate that, by com-

bining our advanced imaging protocols with a crowd-sourcing approach to label data capture, InvertNet will be able to deliver specimen-level occurrence and taxonomic data for a high percentage of the specimens present in the insect collections being digitized, all without the need for handling individual specimens. Ultimately, we envision InvertNet providing a digitization toolkit and research platform available to the entire natural history museum community.

Acknowledgments

We thank the editors for inviting us to contribute to this special volume and Paul Tinerella (U. Minnesota) for initial discussions on digitization procedures. We also thank the following InvertNet collaborators for encouragement and support: A. Cognato (Michigan State U.), G. Courtney and J. VanDyk (Iowa State U.), J. Holland (Purdue U.), L. Gruber (Milwaukee Public Museum),R. Holzenthal (U. Minnesota), P. Johnson (South Dakota State U.), J. Klompen and M. Daly (Ohio State U.), J. Rawlins, R. Davidson and J. Fetzner (Carnegie Museum, Pittsburgh), D. Rider (North Dakota State U.), A. Short (U. Kansas), R. Sites (U. Missouri), D. Young (U. Wisconsin-Madison), J. Zaspel (U. Wisconsin-Oshkosh), G. Zolnerowich (Kansas State U.), J. McPherson (Southern Illinois U.), K. McCravy (Western Illinois U.), M. Goodrich (Eastern Illinois U.), E. Mockford (Illinois State U.), A. DeLorme (Valley City State U.), Neil Cumberlidge (Northern Michigan U.),and J. Vaughan (U. North Dakota). An anonymous referee made several helpful suggestions that greatly improved the manuscript. InvertNet is funded by the U.S. National Science Foundation's Advancing Digitization of Biological Collections (grant# EF11-15112). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Ball J, Gross J, Gryzmala T, Nishida G, Oboyski P, Gillespie R, Roderick G, Will K (2011) Calbug: a case study of digitization challenges for Entomology collections. Entomological Collections Network, Reno, Nov 2011. http://www.ecnweb.org/dev/files/talks_ecn_calbug_kwill2.ppt

Beeler T, Bickel B, Beardsley P, Sumner P, Gross M (2010) High-Quality Single-Shot Capture of Facial Geometry. Proceedings of ACM SIGGRAPH, ACM Transactions on Graphics 29(3): 40.1–40–9. http://graphics.ethz.ch/publications/papers/paperBee10.php

Bertone MA, Deans AR (2010) Remote curation and outreach: examples from the NCSU insect museum GigaPan project. Proceedings of the Fine International Conference on Gigapixel Imaging for Science, November 11–13 2010. http://www4.ncsu.edu/~ardeans/BertoneDeansFAFS.pdf

Blagoderov V, Kitching I, Simonsen T, Smith VS (2010) Report on trial of SatScan tray scanner system by SmartDrive Ltd. Nature Precedings. hdl:10101/npre.2010.4486.1 http://precedings.nature.com/documents/4486/version/1

Eades DC, Otte D, Cigliano MM, Braun H (2012) Orthoptera Species File Online. http://orthoptera.speciesfile.org/HomePage.aspx

Harman KJ, Fitton M, Honey MR, Martin G (2011) The Linnaean Society's insect collection: increasing access through digitization. http://www.linnean.org/fileadmin/images/Collections/LS_Entomology_poster_FINAL_PRESS.pdf

Haüser CL, Steiner A, Holstein J, Scoble MJ (Eds) (2005) Digital imaging of biological type specimens: a manual of best practice. European Network for Biodiversity Information, Stuttgart, 309 pp. http://www.gbif.org/orc/?doc_id=2429

Heidorn PB (2011) Biodiversity Informatics. Bulletin of the American Society for Information Science and Technology 37: 38–44. doi: 10.1002/bult.2011.1720370612

Hill A, Guralnick R, Smith A, Sallans A, Gillespie R, Denslow M, Gross J, Murrell Z, Conyers T, Oboyski P, Ball J, Thomer A, Prys-Jones R, de la Torre J, Kociolek P, Fortson L (2012) The notes from nature tool for unlocking biodiversity records from museum records through citizen science. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 219–233. doi: 10.3897/zookeys.209.3472

Johnson NF (2007) Biodiversity informatics. Annual Review of Entomology 52: 421–438. doi: 10.1146/annurev.ento.52.110405.091259

Johnson NF (2009) Insect biodiversity informatics. In: Foottit RG, Adler PH (Eds) Insect Biodiversity. Wiley-Blackwell, Oxford, UK, 433–443. doi: 10.1002/9781444308211.ch18

Kjar D, Patel M, Klopfer M, Kweskin M, Schultz T (2012) Smithsonian formicid type database. http://ripley.si.edu/ent/nmnhtypedb/public/namelisttemplates/longoutput-namelist.cfm?publicconsumption=1&typeid=663

Rojo MG, Garcia GB, Mateos CP, Garcia JG, Vicente MC (2006) Critical comparison of 31 commercially available digital slide systems in pathology. International Journal of Surgical Pathology 14: 285–305. doi: 10.1177/1066896906292274

Tinerella P (2010) Automation of natural history collections and bioinformatics: rapid optical data acquisition and automated computerization at INHS. National Science Foundation Award Abstract #1132188. http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=1132188&WT.z_pims_id=5448

Vollmar A, Macklin JA, Ford L (2010) Natural history specimen digitization: challenges and concerns. Biodiversity Informatics 7: 93-112. https://journals.ku.edu/index.php/jbi/article/view/3992

ZooKeys 209: 183–191 (2012)
doi: 10.3897/zookeys.209.3115
www.zookeys.org

RESEARCH ARTICLE

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

DScan – a high-performance digital scanning system for entomological collections

Stefan Schmidt[1], Michael Balke[1], Stefan Lafogler[2]

1 *Zoologische Staatssammlung, Münchhausenstr. 21, 81247 Germany* **2** *Technisches Büro München, Thierschstr. 20, 80538 München, Germany*

Corresponding author: *Stefan Schmidt* (hymenoptera@zsm.mwn.de)

Academic editor: *V. Blagoderov* | Received 22 March 2012 | Accepted 29 May 2012 | Published 20 July 2012

Citation: Schmidt S, Balke M, Lafogler S (2012) DScan – a high-performance digital scanning system for entomological collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 183–191. doi: 10.3897/zookeys.209.3115

Abstract

Here we describe a high-performance imaging system for creating high-resolution images of whole insect drawers. All components of the system are industrial standard and can be adapted to meet the specific needs of entomological collections. A controlling unit allows the setting of imaging area (drawer size), step distance between individual images, number of images, image resolution, and shooting sequence order through a set of parameters. The system is highly configurable and can be used with a wide range of different optical hardware and image processing software.

Keywords

Entomology, insect collection, insect drawer, CNC technology

Introduction

Natural history collections are nature's treasure houses. About 80 million objects are deposited in German natural history collections alone, including about 65 million insects (Brake and Lampe 2004). The Zoologische Staatssammlung in Munich, Germany (ZSM) holds about 25 million zoological objects. About 90% of the collection are insects, including 10 million Lepidoptera, 3-4 million Coleoptera, and about three million Hymenoptera, stored in about 100,000 standard sized drawers (51 × 42 cm).

The material deposited in natural history collections like the ZSM is principally held and intended to support research purposes. Natural history collections are indis-

pensable scientific resources that play a central role in biodiversity research (Wheeler et al. 2012). However, the level of documentation of entomological collections is very low, and even basic data about specimens or metadata about collections are often completely missing (Brake and Lampe 2004). Moreover, digitisation of natural history specimens is labour-intensive and usually proceeds at a very slow pace. This is partly due to a regrettable lack of personnel, a situation that is not going to change in the foreseeable future. Technical solutions have the potential to aid our digitisation efforts by reducing the need for extensive human resources. However, these solutions need to be developed. Our aim is to use innovative approaches to develop new methods for the rapid digitisation of entomological collection drawers, and the subsequent extraction of relevant metadata from drawer images.

DScan is a prototype scanning machine and the foundation of a digitisation system that allows fast and efficient digitisation of entomological drawers. Our primary aim is the optimisation of this system for on-demand-digitisation requirements. Because the contents of and arrangement of specimens within drawers will change if they are part of an active research collection, re-scanning of drawers needs to be as fast and as easy as possible.

The resulting images allow inspection of insect specimens at high resolution without the need to access the collection itself physically. The level of detail can be adjusted as required, for instance in relation to the size of the insect specimens, and, is in most cases sufficient for specialists to recognize the taxon at genus or even species level.

Mechanics of the drawer scanning system

DScan is made of a sturdy, industrial standard aluminium frame (LWH = 1080 × 1080 × 1500 mm) with linear units as used by Computer Numerical Control (CNC) positioning machines (Fig. 1, and YouTube video under youtu.be/zyT7l-CZego). Servo drives and precision ball screw spindles allow a minimum step distance of 0.02 mm at a maximum speed of 100 mm/s. Effective travel ranges are 600 × 600 mm horizontally (x- and y-axis) and 200 mm vertically (z-axis). The system is operated by a PC-controlled console (netbook) with ProNC software (DNC Software Ltd, www.pronc.com). The left and right sides and back of the scanner are covered by white panels. The front is closed by a curtain with a reflective inner surface that is closed during scan operations.

Optics

Choice and selection of the optical components of the drawer scanning system are largely unconstrained by the mechanics of the positioning components of the DScan mechanism. A wide range of different camera systems can be adapted to work with the DScan, provided that the camera has remote control capability because the shutter release needs to be triggered by the control unit. Currently the best option includes a digital single-lens reflex (DSLR) camera, although the recent introduction of mir-

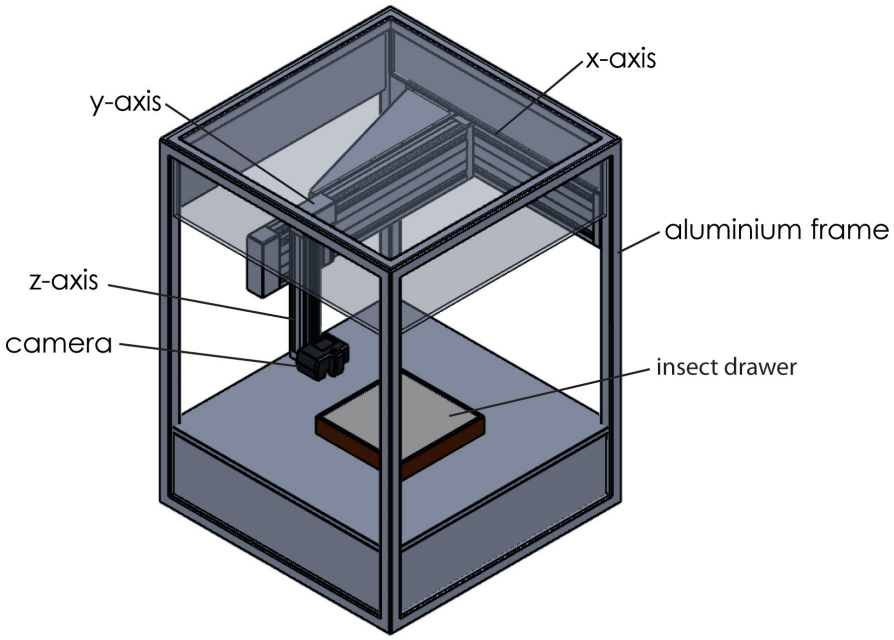


Figure 1. Schematic drawing of the DScan system. Flashes (not shown) are placed inside the scanner.

rorless system cameras with interchangeable lenses and comparatively large sensors will increase the range of suitable optical equipment. In addition, mirrorless cameras employing electronic shutter mechanisms avoid the wear that can be significant when using a DSLR for creating large numbers of images.

Our current system comprises a Nikon D300 DSLR camera equipped with a 12 megapixel APSC sensor, attached to a Voigtländer 90 mm Apo-Lanthar macro lens. As light sources we use two studio flash lights that are placed inside the scanner compartment. The white inner surfaces of the top and side panels, and the white front curtain are highly reflective and allow for maximum lighting efficiency. The flashes are directed toward the side and top panels to achieve an even and non-reflective illumination of specimens. The indirect lighting reduces the risk of blown-out highlights caused by reflections from insects with smooth surfaces and exceedingly high contrast. Typically, these effects are caused by direct, punctual lighting when photographing insects with strongly sculptured and shining, in particular metallic, surfaces.

Scanning process

The system takes images of a single drawer in a sequential order as determined by the controlling unit. The shooting order is customizable and can be configured by modifying parameters of the control program. The best results are achieved when each photograph overlaps its neighbours by about 30-40%, which enables the stitching software to gener-

ate smooth transitions between images. The number of images per drawer can be adjusted by changing the distance between drawer and camera (z-axis). A lower z-distance results in a larger number of images and a larger size and higher resolution of the final megapixel image. A full scan of a standard sized drawer (51 × 42 cm) at a distance of 60 cm takes about 2.5 minutes and produces a set of 56 images (see the DScan in action on YouTube, youtu.be/zyT7l-CZego). At a distance of 52 cm, which is the minimum distance of the macro lens we are using without close-up lens, a scan comprises 99 images and the scanning process takes about 4.2 minutes.

Image processing

To obtain the final high-resolution image, the captured images from each drawer need to be assembled or "stitched" using dedicated stitching or panorama software. For this purpose, we use AutoPano Giga (Kolor, www.kolor.com). Images are captured in RAW format, developed using Capture NX2 (www.capturenx.com) and saved as 8- or 16-bit TIFF images. Alternatively, images can be captured in JPEG (Figs 2a; 3a, d, g) or TIFF format (Figs 2b; 3b, e, h) and directly assembled without the need of image development. Using JPG format reduces the post processing effort but produces images of slightly inferior quality compared to RAW (cf. Fig. 3a, d, g vs 3c, f, i, for high resolution versions of Figures 2 and 3 see media.zsm-entomology.de/suppl/zookeys_mass_digitisation_volume/Fig_2.png and media.zsm-entomology.de/suppl/zookeys_mass_digitisation_volume/Fig_3.png).

With 56 images per drawer, the final stitched image has a size of ca. 300 megapixels, whereas images that are assembled from 99 photographs result in pictures of about 500 megapixels. The resulting images are far too large for display in a web browser and need to be made "zoomable" by tiling and creating a low resolution version of the original image. This can be achieved by dedicated software such as Zoomify (Zoomify, Inc., www.zoomify.com) or Krpano (krpano GmbH, krpano.com). During this process, tiles are created at different resolutions, allowing zoom-and-pan viewing of the drawer image so that if parts of a drawer image are enlarged, the corresponding tiles are loaded, thus avoiding the need to load the full high-resolution image before it can be viewed. Sample images are available online at zsm-entomology.de and show insect drawers containing Coleoptera (zsm-entomology.de/wiki/Drawer_Digitization_Project_-_Coleoptera), Hymenoptera (zsm-entomology.de/wiki/Drawer_Digitization_Project_-_Hymenoptera), and Lepidoptera (zsm-entomology.de/wiki/Drawer_Digitization_Project_-_Lepidoptera).

Performance

With an optimised workflow in place, from capturing individual images of a single drawer to the final megapixel image, processing of about 100 insect drawers per day

Figure 2. Partial drawer images taken at the same position using three different file formats: captured as JPEG (**a**), captured as TIFF (**b**), and TIFF converted from RAW (**c**). A high resolution version of the image is available under media.zsm-entomology.de/suppl/zookeys_mass_digitisation_volume/Fig_2.png

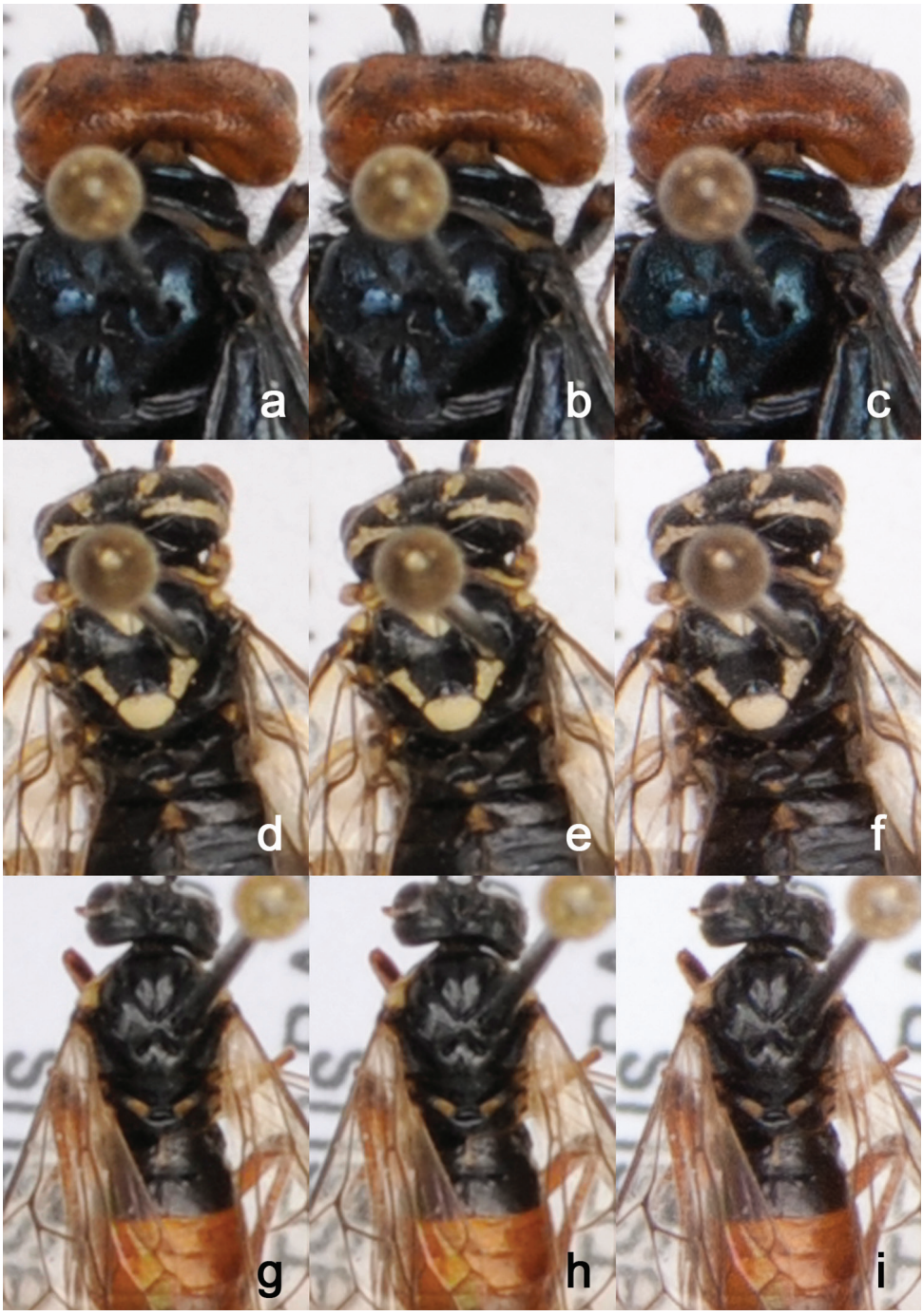


Figure 3. Enlargements of drawer images from Fig. 1 to show quality differences between image file formats. Each of the three specimens was captured in JPEG (**a, d, g**), TIFF (**b, e, h**), and RAW (**c, f, i**). A high resolution version of the image is available under media.zsm-entomology.de/suppl/zookeys_mass_digitisation_volume/Fig_3.png

seems technically possible. This assumes a scan rate of 20 drawers per hour plus 10 hours for processing the images in batch mode, which can be done overnight. Image processing (developing, stitching, and image adjustment) depends largely on the computer hardware used. Using a workstation equipped with two Intel Xeon processors with 12 MB cache and a speed of 2.26 Ghz each, 24 MB of RAM, and a high-end graphic card, the stitching process of 56 individual TIFF images using the software AutoPano Giga takes about 4.5 minutes. The subsequent generation of multi-resolution images in jpeg format using Krpano takes an additional 1.5 minutes, resulting in a total of about 6 minutes for the computational part of the scanning process, starting from a set of individual images to high-resolution, zoomable images that are ready for dissemination on the internet.

Costs

The costs for the CNC system itself without optics, computer hardware, and software amount to about USD $25,000. The camera system, i.e. a digital SLR with macro lens and studio flash lights, comes to about $2,000–$3,500, adding up to about $30,000 in total for the system including software (AutoPano Giga, krpano) but without computer hardware for image processing and storage. However, as mentioned before, available hardware can be used and a range of suitable cameras can be fitted to the system, requiring only minor modifications to the controlling unit and cable connectors. Standard computer hardware can be used for the stitching of images although the processing will take longer than with a dedicated workstation.

Further developments and prospects

The DScan system aims to achieve rapid digitisation of entomological collections. The high-resolution images of insect drawers themselves contain a wealth of information. However, additional processing is required to extract that information from the images and make computable metadata about the drawer content available and searchable. Currently, we generate basic metadata associated with each drawer manually, including taxon information and geographic coverage. Several ways to extract metadata from drawer images are currently being explored and evaluated:

- Counting the number of specimens in a drawer and at the same time assigning numbers to each specimen using image analysis software like ImageJ (http://rsbweb.nih.gov/ij/) (Fig. 4). The number and position of each specimen can be exported and used, for instance, for image analysis purposes. For example, the position (x-, y-coordinates) of specimens can be used to automatically crop an image around the position of a specimen. This would allow to create individual images of specimens in a drawer.

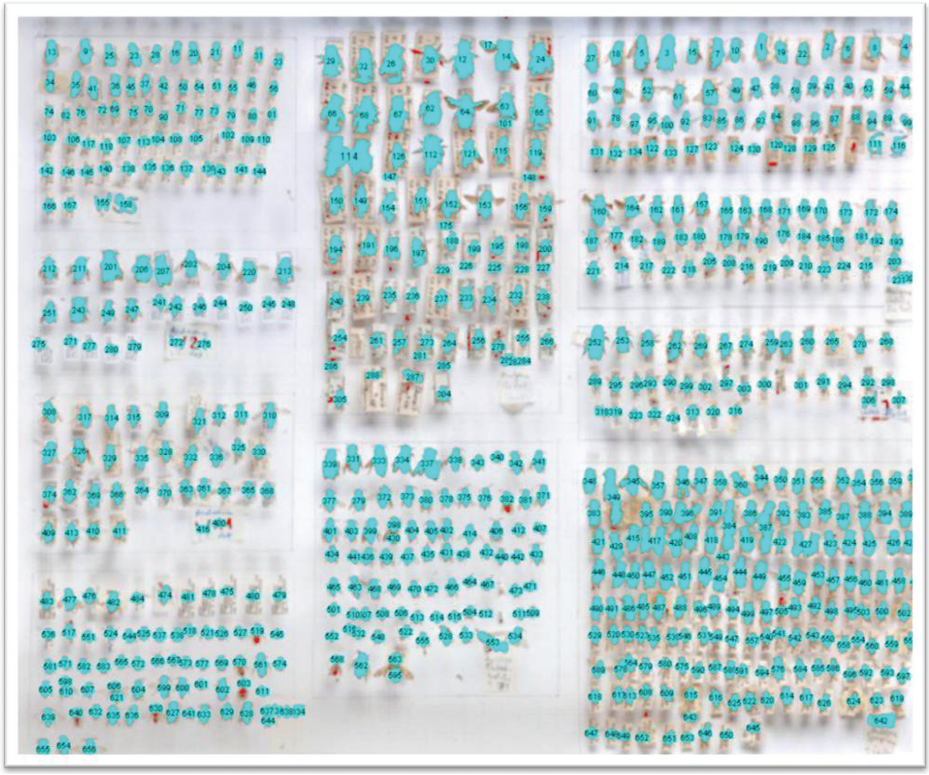


Figure 4. Automatic numbering of specimens using ImageJ. For details see text.

- Adding a clickable "hot spot" to individual specimens in a drawer that shows, when selected, specimen- or species-related information, for example, type data. The trigger can also be used to open a text box, open an external web site, or submit a database query.
- There could be an option for users to interactively mark certain specimens, for example taxonomists who would like to borrow certain specimens for closer examination. Additionally, users may be given the opportunity to add information to specimens, e.g., identifications.
- Metadata could be extracted using Optical Character Recognition software.
- Specimens with a Quick Response (QR) code label that is visible from above can be tracked in a collection.
- Cutting out individual specimens from the drawer image and associating metadata with them will bridge the gap between digitisation at drawer and at specimen level.
- Extended depth-of-field photography to avoid parts of the image (e.g. bottom labels) being out of focus. This is particularly important when depth-of-field becomes very short at close object distance.

The above list includes only some ideas that seem worthy of exploration in the future. More applications and analysis methods will surely emerge once the system is used routinely, provided that funding opportunities permit further development.

Acknowledgements

We particularly thank Katja Neven (ZSM Entomology) for patiently testing the D-Scan, despite frequent modifications of the system, and for assisting in producing the sample drawer images. The DScan system was developed as part of the joint project "Kompetenzzentren innovativer Datenmobilisierung" (Competence Centres for innovative Data Mobilization, grant 01 LI 1001 B), funded by the Federal Ministry of Education and Research (BMBF). The project is part of the German Global Biodiversity Information Facility (GBIF-D) through the Evertebrata II node.

References

Brake I, Lampe K-H (2004) ZEFOD - Zentralregister biologischer Forschungssammlungen in Deutschland. Ergebnisse des Teilprojektes über zoologische Sammlungen an Museen und Universitäten. Mitteilungen der deutschen Gesellschaft für allgemeine und angewandte Entomologie 14: 475–478.

Wheeler QD, Knapp S, Stevenson DW, Stevenson J, Blum SD, Boom BM, Borisy GG, Buizer JL, De Carvalho MR, Cibrian A, Donoghue MJ, Doyle V, Gerson EM, Graham CH, Graves P, Graves SJ, Guralnick RP, Hamilton AL, Hanken J, Law W, Lipscomb DL, Lovejoy TE, Miller H, Miller JS, Naeem S, Novacek MJ, Page LM, Platnick NI, Porter-Morgan H, Raven PH, Solis MA, Valdecasas AG, Van Der Leeuw S, Vasco A, Vermeulen N, Vogel J, Walls RL, Wilson EO, Woolley JB (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. Systematics and Biodiversity 10(1): 1–20. doi: 10.1080/14772000.2012.665095

ZooKeys 209: 193–202 (2012)
doi: 10.3897/zookeys.209.3486
www.zookeys.org

RESEARCH ARTICLE

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

Nomenclatural benchmarking: the roles of digital typification and telemicroscopy

Quentin Wheeler[1], Thierry Bourgoin[2], Jonathan Coddington[6], Timothy Gostony[1], Andrew Hamilton[1], Roy Larimer[5], Andrew Polaszek[3], Michael Schauff[4], M. Alma Solis[4]

1 *International Institute for Species Exploration, Arizona State University, Tempe, AZ 85287 USA* **2** *Laboratoire d'Entomologie, Museum National d'Histoire Naturelle, Rue Buffon, Paris, France* **3** *Department of Life Sciences, The Natural History Museum, London SW7 5BD, U.K.* **4** *United States Department of Agriculture, Systematic Entomology Laboratory, Beltsville, MD 20705 USA* **5** *Visionary Digital, Palmyra, VA 22963 USA* **6** *National Museum of Natural History, Smithsonian Institution, Washington, DC 20530 USA*

Corresponding author: *Quentin Wheeler* (quentin.wheeler@asu.edu)

Academic editor: *Vladimir Blagoderov* | Received 8 June 2012 | Accepted 13 July 2012 | Published 20 July 2012

Citation: Wheeler Q, Bourgoin T, Coddington J, Gostony T, Hamilton A, Larimer R, Polaszek A, Schauff M, Solis MA (2012) Nomenclatural benchmarking: the roles of digital typification and telemicroscopy. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 193–202. doi: 10.3897/zookeys.209.3486

Abstract

Nomenclatural benchmarking is the periodic realignment of species names with species theories and is necessary for the accurate and uniform use of Linnaean binominals in the face of changing species limits. Gaining access to types, often for little more than a cursory examination by an expert, is a major bottleneck in the advance and availability of biodiversity informatics. For the nearly two million described species it has been estimated that five to six million name-bearing type specimens exist, including those for synonymized binominals. Recognizing that examination of types in person will remain necessary in special cases, we propose a four-part strategy for opening access to types that relies heavily on digitization and that would eliminate much of the bottleneck: (1) modify codes of nomenclature to create registries of nomenclatural acts, such as the proposed ZooBank, that include a requirement for digital representations (e-types) for all newly described species to avoid adding to backlog; (2) an "r" strategy that would engineer and deploy a network of automated instruments capable of rapidly creating 3-D images of type specimens not requiring participation of taxon experts; (3) a "K" strategy using remotely operable microscopes to engage taxon experts in targeting and annotating informative characters of types to supplement and extend information content of rapidly acquired e-types, a process that can be done on an as-needed basis as in the normal course of revisionary taxonomy; and (4) creation of a global e-type archive associated with the commissions on nomenclature and species registries providing one-stop-shopping for e-types. We describe a first generation implementation of the "K" strategy that adapts current technology to create a

network of Remotely Operable Benchmarkers Of Types (ROBOT) specifically engineered to handle the largest backlog of types, pinned insect specimens. The three initial instruments will be in the Smithsonian Institution(Washington, DC), Natural History Museum (London), and Museum National d'Histoire Naturelle (Paris), networking the three largest insect collections in the world with entomologists worldwide. These three instruments make possible remote examination, manipulation, and photography of types for more than 600,000 species. This is a cybertaxonomy demonstration project that we anticipate will lead to similar instruments for a wide range of museum specimens and objects as well as revolutionary changes in collaborative taxonomy and formal and public taxonomic education.

Keywords
Types, typification, digital imaging, biodiversity informatics, taxonomy, nomenclature, natural history museums

Introduction

Our ability to explore, sustain, and utilize biodiversity depends on accurate species identifications, predictive phylogenetic classifications, and reliable scientific names. Biodiversity informatics relies on scientific names and the field continues to expand uses of binominals in information management and analysis (Patterson et al. 2006, 2010).

Species-level binominals are objectively applied due to the practice of typification in which a single specimen is designated to function as a representative or standard for the name (Blackwelder 1967, ICZN 1999, McNeill et al. 2006) Nomenclatural benchmarking is the periodic alignment of species names with changing theories of the limits of species and involves the reexamination of type specimens. Although the Code aims to promote stability in nomenclature Eugene Gaffney (1979) observed that taxonomic stability *is* ignorance. New data, specimens, and analyses inevitably change and improve our understanding of species. These changes variously require coining new names, redefining concepts attached to existing names, or resurrecting names from synonymy. Unless binominals keep pace with the growth of knowledge and changing concepts of species, their information content and reliability as tools of communication and data management decline over time.

The process of nomenclatural benchmarking is the examination of type specimens of all available species-group names (i.e., all species-group names meeting the requirements of the prevailing Code) to ascertain which currently accepted taxonomic species the specimen bearing the name falls within. Whichever species the type specimen falls within, there follows the name attached to it. Difficulties in accessing types to inform nomenclatural decisions is slowing progress in taxonomy and threatening the integrity of biodiversity databases. Digital representations of types or e-types are clearly a major part of the solution. Where detailed images of types exist many nomenclatural decisions can be made rapidly and efficiently. Botanists have led the way in the systematic digitization of types with impressively effective results from projects of individual herbaria to coordinated community projects (e.g., Global Plants Initiative, see www.

botanischestaatssammlung.de/projects/GPI.html). Zoologists are making progress, including specialized imaging techniques for unique specimen challenges (e.g., Berquist et al. 2012), but have major challenges ahead.

Here we address four issues that we regard as major challenges for nomenclatural benchmarking. First, there is the matter of a massive backlog. It has been estimated that the nearly two million currently recognized species (Chapman 2009) are accompanied, including names in synonymy, by perhaps five to six million name-bearing types. There is no tally of the number of type specimens that have been digitized to date, but it is at most a fraction of the backlog. Second, there is the issue of adding to the backlog through the description of new species. There is no formal requirement or expectation that types of the 18,000 or so species described each year be digitized. Third, there is a need for access to type specimens by experts in cases where existing digital images (e-types) fail to reveal characters in sufficient detail for definitive decisions regarding status. And, finally, there is a global need for a portal for access to all e-types.

We propose a strategy for addressing these challenges, including (I) modifications of the Codes to assure no further accumulation of backlogs of non-digitized types, (II) an "r" strategy that relies on automation to rapidly create reasonably informative e-types without the need for expert involvement; (III) a "K" strategy that engages experts to expand and refine such first approximation e-types; and (IV) the creation of a global archive of e-types. In addition, we describe a first generation "K" strategy instrument accessible via the Internet as part of an international network of remotely operable digital microscopes that make insect types accessible to taxon experts and that we anticipate will be launched in December, 2012.

I: Digitize types for new species at time of description

We could avoid adding to an already massive backlog of un-digitized types by adopting a few simple practices. First, we believe that the Codes should be modified to mandate registration of all nomenclatural acts, including descriptions of new species (Polaszek et al. 2005). As a further requisite, e-types should be a mandatory part of the registration of new species. While the minimum requirement would be one or more images, authors should be urged to include both a habitus representation of the type, preferably from multiple angles, as well as additional annotated detailed images of diagnostic anatomical details. Successful implementation will require standards for images as well as for data and metadata capture and dissemination, but such standards are already in wide use in biological informatics and should pose no serious difficulty.

Major museums that accession large numbers of types each year should establish e-typification centers to meet their in-house needs and to serve as a regional digital typification center. E-types could be created at a nominal fee for taxonomists working outside such institutions or offered at no charge for authors willing to permanently deposit the type with the museum.

II: Rapid ("r" Strategy) e-typification

To deal with a backlog of millions of type specimens we propose the development and engineering of automated e-typification instruments capable of rapidly capturing as much visual information from the specimen as possible without the need for expert intervention. It is easy to imagine such automated instruments that rotate the specimen, orbit a digital camera, or employ a battery of digital cameras to rapidly create rotatable and scalable 2D and 3D images of types. This would capture most, but not all, characters and provide a reasonably good first approximation of an e-type. Automation will result in low personnel costs. Deployed in numbers, such instruments could quickly eliminate the backlog. Following this initial digitization of the backlog these instruments could be permanently installed at the e-typification centers discussed above.

III: Comprehensive ("K" Strategy) e-typfication

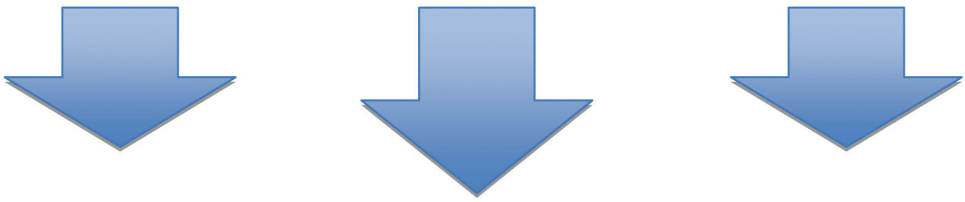
One reason that the "r" strategy is rapid is that it imposes a one-size-fits-all approach to creating reasonably good 3D composites of type specimens. While resulting e-types will enable many nomenclatural decisions, in other cases the images will be found wanting in detail, illumination, angle, or some other respect. In certain cases, such as where a dissection is necessary to reveal a character, a physical visit to the museum or shipment of a specimen is unavoidable. In other cases it may be that simply connecting an expert with a type specimen via telemicroscopy is enough. This "K" strategy takes advantage of expert knowledge to supplement existing images with those that target diagnostic characters. This is a symbiotic relationship, with the expert gaining precious access to a type and the museum profiting from expert knowledge, because the images captured from telemicroscopy will become part of the composite e-type.

Benefits of telemicroscopy are obvious. They can save a great deal of time and money compared to visits by experts to museums, they can virtually repatriate types to scientists in countries of origin allowing a level of interaction not possible with archived images, they can dramatically decrease wear and tear on specimens, and they further democratize taxonomy by leveling the field for amateurs and scientists at small institutions who will have equal access to types.

IV: A global e-type archive

A comprehensive, distributed, open-access global e-type archive is urgently needed. In fact, next to completing a catalog with the status of all available species names, such an archive ranks among the greatest needs for advancing biodiversity exploration and informatics. A global e-type archive would provide one-stop access to images of the

I: Triage **Avoiding additional backlog**	**II: "r"** **Rapid e-Typification**	**III: "K"** **Comprehensive e-Typification**
ICZN should mandate that all newly described species of animals are registered in ZooBank and that the registration process include digitization (e-typification) to meet a minimum standard (perhaps dorsal plus lateral and ventral). Collections that house types should be equipped to digitize. At least one museum in each country should be designated a typification center, offering service at little or no cost (perhaps in exchange for deposition of types). This will avoid any additions to the backlog of non-digitized types.	In order to deal with a massive backlog of insect types that are not yet digitized in any form, we propose development of an automated 3D imaging instrument capable of rapidly creating as close to a full 3D image of a type specimen as possible. Dozens of rapidly acquired stills would be sutured into a rotatable, zoom-able representation of the type specimen allowing a view of almost every angle of the specimen. The emphasis is on reasonably good and rapid documentation, not on high quality capture of any single morphological structure.	The comprehensive strategy takes a different approach, connecting via cyber space taxon experts with type specimens on a need to know basis. As types must be examined to resolve nomenclatural issues, the expert is allowed to manipulate and photograph a specimen remotely so that s/he can capture key characters in detail. Over time, as an archive of images grows, the need to access specimens will decrease and images can also be ingested into existing 3D composite images. Our ROBOT instrument is the first realization of "K".



GeTA

Global e-Type Archive

The goal should be established to create a comprehensive archive of digital images of all type specimens. The above strategy is our recommended vision for creating and populating the insect digitized (e) types, all fed into an archive that provides open access to both initial 3D representations of types and, over time, accumulated detailed images as well.

Figure 1. Three-part strategy to (a) avoid further growth of backlog by digitizing all new species, (b) rapidly create 3D e-types for all existing species, and (c) open access to types for experts to facilitate their nomenclatural decision-making while simultaneously expanding and enhancing comprehensiveness of digital images of informative characters of type specimens. All images should be available through an open access public "Global e-Type Archive," whether managed by ZooBank or a community-level organization.

type specimens for any species and would be complementary to, and possibly accessible through, portals such as ZooBank, the Encyclopedia of Life, and the Biodiversity Heritage Library. It could also be easily hyperlinked in electronic taxonomic journals and monographs.

Implementing "K" strategy for insect type specimens

ROBOT(E)

The idea of sharing specialized research instruments through Web access is not new (Hadida-Hassan et al. 1999) and, in our case, can be expanded to include specialized research resources such as specimens in collections. Histologists and pathologists have used telemicroscopy for decades and pioneered many innovative applications including robotic controls, archival images, multiple simultaneous viewing, interdisciplinary telecommunication, team consultation, and expert teleconsultation (e.g., Bellina and Missoni 2009, Leong and McGee 2001, Mea et al. 1999, Kayser 2002, Pantanowitz 2010) with application by extension to taxonomy.

Networking three leading insect collections in Washington, DC (Smithsonian Institution, National Museum of Natural History, Department of Entomology), London (Natural History Museum, Department of Entomology), and Paris (Museum National d'Histoire Naturelle, Laboratoire d'Entomologie) we set out to demonstrate that telemicroscopy could be used to implement our "K" strategy. With just these three nodes in a network of remotely operable microscopes in a network scheduled to go "live" in December, 2012 we will open potential access to a large fraction of insect type

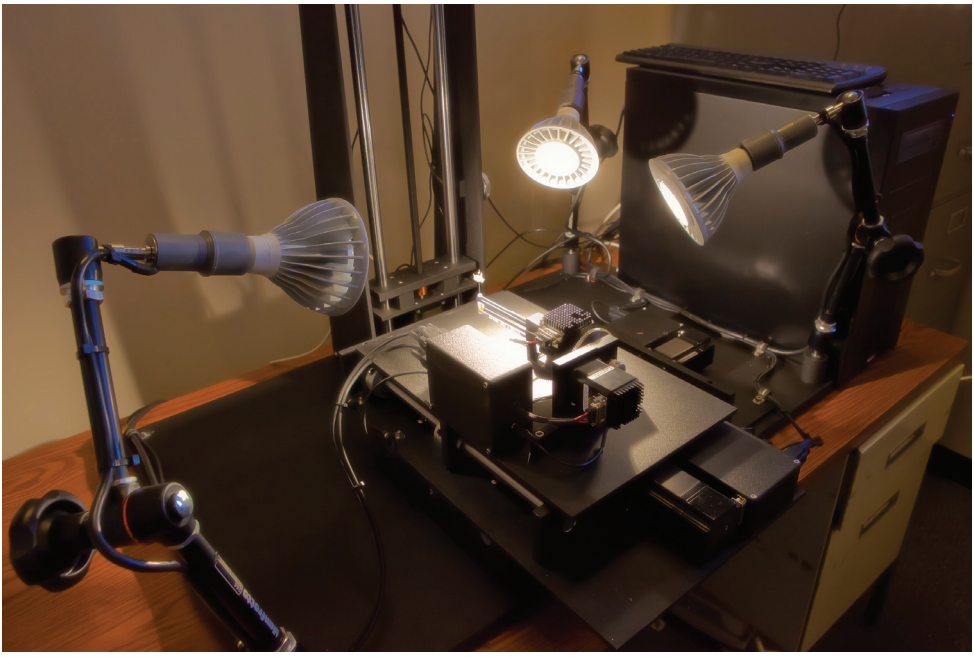


Figure 2. Two ROBOT(E) remotely operable digital imaging systems designed to allow taxonomists to examine, manipulate, and digitally photograph type specimens through a Web connection. Three such instruments are being deployed to major insect collections in Washington, London, and Paris. A prototype instrument remains with the IISE for testing and development purposes. PHOTO: Courtesy of Erik Holsinger, Arizona State University.

specimens. These three collections, the largest on earth, contain more than 600,000 insect type specimens and more than 100,000,000 specimens possibly representing as many as 80% or more of known insect species.

Our selection of insects for a demonstration project was a relatively easy one for several reasons. First, insects account for more than one million described (Foottit and Adler 2009, Zhang 2011) species and an estimated two to three million type specimens. Second, many types are preserved as dry, pin-mounted specimens, making the engineering challenge of handling them manageable. Third, many types fall within a reasonable size range, again easing the challenge of handling most of them with a single device. Finally, many insects are of great agricultural, medical, and ecological interest and their taxonomy is undergoing rapid change requiring frequent access to types.

We have named our system ROBOT (Remotely Operable Benchmarker Of Types), with the first iteration (E) specially designed to handle pinned entomological types. Our goal was to make the system as simple and reliable as possible and to minimize costs by using as much off-the-shelf technology as feasible. The heart of ROBOT(E) is a digital Canon 7D camera that gave us several critically important capabilities beyond capturing images including auto-focus and through-the-sensor high resolution viewing. For the *z* axis we used the Visionary Digital BK P-51 CamLift that has a very precise linear actuator that can be moved in increments as small as 6.0 microns. The *x* and *y* axes use precise micro-step motors to move plates that were custom manufactured by a machine shop. Heavy studio-style lamp holders were modified to secure daylight temperature (ca. 5000 K) LED lamps that would operate on 120 or 240 v current. For the specimen holder, we designed an arm linked to two additional micro-step motors so that the specimen may be spun 360 degrees and "rolled" 180 degrees to reveal the ventral surfaces of specimens. The pin is secured by a tight bundle of fine acrylic cable into which it is inserted.

We designed and wrote the ROBOT(E) software to be simple and intuitive. Several "windows" may be seen or hidden and resized or positioned to meet user preferences. Simple mouse, arrow key, and button choices operate the system's five motors. Autofocus may be alternated with fine manual focusing. Autofocus is disabled when the specimen is rolled, and an algorithm keeps the specimen in approximate focus. Images are stored in a temporary folder from which they may be downloaded to any target folder. In addition, users may create bookmarks that remember *x*, *y*, and *z* coordinates so that specific views may quickly be recovered.

This first generation of ROBOT is intended to prove the usefulness of telemicroscopy in the study of types and has limitations. Future generations could easily be modified to handle a range of museum specimens or objects with little modification. Once the systems are fully tested in museum settings, we plan to add a number of additional features, including an automated image stacking montage function and improved control over illumination. Options will likely include a choice of spot or diffuse light. By combining ROBOT with an advanced video communication software package, colleagues can examine a type or rare specimen simultaneously, a specimen intercepted at a port of entry could be identified in consultation with an expert, or an expert could

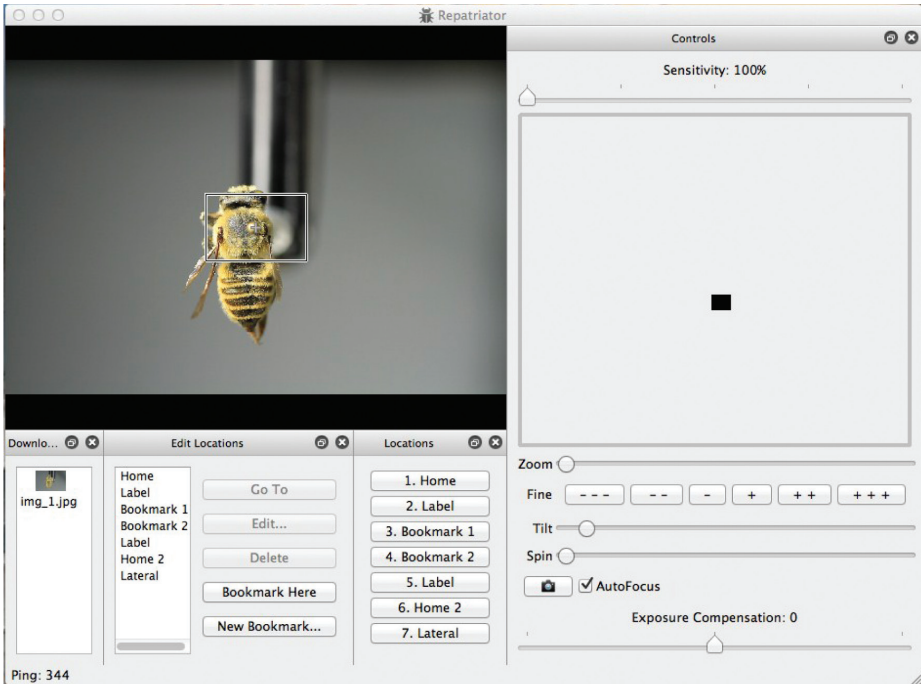


Figure 3. Screen capture of ROBOT(E) system in use. User is able to orient specimen on multiple axes by actuating micro-step motors that position on *x*, *y*, and *z* scales as well as spinning around axis of pin or tilting specimen to examine lateral or ventral perspectives.

use the specimen for advanced teaching. We hope that this project serves to encourage additional uses for remote microscopy and paves the way to open access to types.

Conclusions

Our implementation of a network of remotely operable digital microscopes serves as a demonstration that high value specimens can be accessed, examined and imaged from virtually anywhere. It is merely one step in the modernization of museum specimen access. This is not a general solution to type accessibility or a substitute for creating e-types. We propose a broader strategy of which this direct connection of expert and type is merely one component. Our other recommendations include a global archive of type images, e-typification at time of original description and registration, and engineering automated instruments to rapidly create 3D images of all types. We also foresee modifications to improve our telemicroscopes in terms of their functionality, ability to handle a wide range of specimens and objects, and coupling with automated systems that alleviate much of the need for human involvement in specimen access.

Acknowledgements

Construction of a prototype of ROBOT(E) was made possible by a grant from the Virginia M. Ullman Foundation and matching contributions from Visionary Digital, Inc. Construction of the three final instruments was funded by the International Institute for Species Exploration, Arizona State University, and Visionary Digital, Inc. Andrew Hamilton's contributions to this work were supported by the National Science Foundation under grant number SES-09083935.

References

Agosti A, Alonso-Zarazaga M, Beccaloni G, de Place Bjørn P, Bouchet P, Brothers DJ, the Earl of Cranbrook, Evenhuis N, Godfray HCJ, Johnson NF, Krell F-T, Lipscomb D, Lyal CHC, Mace GM, Mawatari S, Miller S, Minelli A, Morris S, Ng PKL, Patterson DJ, Pyle RL, Robinson N, Rogo L, Thompson FC, van Tol J, Wheeler QD, Wilson EQ (2005) A universal register for animal names. Nature 437: 477. doi: 10.1038/437477a

Bellina L, Missoni E (2009) Mobile cell-phones (M-phones) in telemicroscopy: increasing connectivity of isolated laboratories. Diagnostic Pathology 4: 19. doi: 10.1186/1746-1596-4-19

Berquist RM, Gledhill KM, Peterson MW, Doan AH, Baxter GT, Yopak KE, Kang N, Walker HJ, Hastings PA, Frank LR (2012) The digital fish library: Using MRI to digitize, database, and document the morphological diversity of fish. PLoS ONE 7: e34499. doi: 10.1371/journal.pone.0034499

Blackwelder RE (1967) Taxonomy: A Text and Reference Book. Wiley, New York. 714 pp.

Chapman AD (2009) Numbers of Living Species in Australia and the World. Australian Biological Resources Study, Canberra. 80 pp.

Foottit RG, Adler PH (Eds) (2009) Insect Biodiversity: Science and Society. Wiley-Blackwell, Chichester, 632 pp.

Gaffney ES (1979) An introduction to the logic of phylogeny reconstruction. In: Cracraft J, Eldredge N (Eds) Phylogenetic Analysis and Paleontology. Columbia University Press, New York, 79–111.

Hadida-Hassan M, Young SJ, Peltier ST, Wong M, Lamont S, Ellisman MH (1999) Web-based telemicroscopy. Journal of Structural Biology 125: 235–245. doi: 10.1006/jsbi.1999.4095

ICZN (1999) International Code of Zoological Nomenclature. 4th ed. International Trust for Zoological Nomenclature, London, 306 pp.

Kayser K (2002) Interdisciplinary telecommunication and expert teleconsultation in diagnostic pathology: present status and future prospects. Journal of Telemedicine and Telecare 8: 325–330. doi: 10.1258/135763302320939202

Leong FJW-M, McGee JO'D (2001) Automated complete slide digitization: a medium for simultaneous viewing by multiple pathologists. Journal of Pathology 195: 508–514. doi: 10.1002/path.972

McNeill J, Barrie FR, Burdet HM, Demoulin V, Hawksworth DL, Marhold K, Nicolson DH, Prado J, Silva PC, Skog JE, Wiersema JH, Turland NJ (Eds) (2006) International Code of

Botanical Nomenclature (Vienna Code). Regnum Vegetabile 146. A. R. G. Gantner Verlag KG., Ruggell, 568 pp.

Mea VD, Cataldi P, Pertoldi B, Beltrami CA (1999) Dynamic robotic telepathology: a preliminary evaluation on frozen sections, histology and cytology. Journal of Telemedicine and Telecare 5: 55–56. doi: 10.1258/1357633991932559

Pantanowitz L (2010) Digital images and the future of digital pathology. Journal of Pathological Informatics 1: 15. doi: 10.4103/2153-3539.68332

Patterson DJ, Remsen D, Marino WA, Norton C (2006) Taxonomic indexing — extending the role of taxonomy. Systematic Biology 55: 367–373. doi: 10.1080/10635150500541680

Patterson DJ, Cooper J, Kirk P, Pyle R, 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

Zhang Z-Q (2011) Phylum Arthropoda von Siebold, 1848 In: Zhang Z-Q (Ed) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. Zootaxa 3148: 99–103.

ZooKeys 209: 203–217 (2012)
doi: 10.3897/zookeys.209.3146
www.zookeys.org

RESEARCH ARTICLE

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

Image based Digitisation of Entomology Collections: Leveraging volunteers to increase digitization capacity

Paul Flemons[1], Penny Berents[1]

1 *Australian Museum, 6 College Street, Sydney 2010*

Corresponding author: *Paul Flemons* (paul.flemons@austmus.gov.au)

Academic editor: *V. Blagoderov* | Received 30 March 2012 | Accepted 26 June 2012 | Published 20 July 2012

Citation: Flemons P, Berents P (2012) Image based Digitisation of Entomology Collections: Leveraging volunteers to increase digitization capacity. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 203–217. doi: 10.3897/zookeys.209.3146

Abstract

In 2010, the Australian Museum commenced a project to explore and develop ways for engaging volunteers to increase the rate of digitising natural history collections. The focus was on methods for image-based digitising of dry pinned entomology collections. With support from the Atlas of Living Australia, the Australian Museum developed a team of volunteers, training materials and processes and procedures.

Project officers were employed to coordinate the volunteer workforce. Digitising workstations were established with the aim of minimising cost whilst maximising productivity and ease of use. Database management and curation of material before digitisation, were two areas that required considerably more effort than anticipated.

Productivity of the workstations varied depending on the species group being digitised. Fragile groups took longer, and because digitising rates vary among the volunteers, the average hourly rate for digitising pinned entomological specimens (cicadas, leafhoppers, moths, beetles, flies) varied between 15 to 20 per workstation per hour, which compares with a direct data entry rate of 18 per hour from previous trials.

Four specimen workstations operated four days a week, five hours a day, by a team of over 40 volunteers. Over 5 months, 16,000 specimens and their labels were imaged and entered as short records into the museum's collection management database.

Keywords

Digitising, image, volunteers, Australian Museum, collections

Introduction

The Australian Museum (AM) has natural science collections dating from 1806. The collections hold more than 18 million specimens of animals, fossils, rocks and minerals. Digitisation (in the form of databasing the text from specimen labels) of the collections commenced in the 1970s and in 2012 approximately 40% of the collections have a text record in the Museum's collection database. To digitise the remainder of the collections in this way, would at comparable rates, take another 50 years at least.

Funding for digitising of collections needs to be allocated as efficiently and wisely as possible to maximise the return for the investment. However it is unlikely that funding available for digitising is ever going to equal funding required for fully digitising our collections. This represents the digitising impediment.

In response to a lack of adequate resources for digitising, the Australian Museum (AM) has been exploring opportunities for engaging volunteers in image-based specimen digitisation since 2007. Initial work (Tann and Flemons 2008) demonstrated that utilising volunteers for imaging specimens and their labels was feasible and compared favourably with traditional text-only data entry techniques.

In 2010 the Australian Museum obtained funding from the Atlas of Living Australia (ALA) to develop a volunteer-based digitisation program (DigiVol). The aim of this project was to explore and develop methods and technologies for engaging volunteers to assist in the rapid digitisation and registration of museum specimens. The project focused on the entomology collection, in part because it is a big collection that is largely not digitised, yet it lends itself to a methodical volunteer-based digitising process.

It was considered essential to establish a clear project scope for setting boundaries within which to develop processes and procedures. This was particularly important for the imaging process as the choice of imaging resolution would have an enormous impact on downstream use and storage of images. Computer storage costs, network bandwidth and display capabilities often lag behind the capacity for capturing high resolution images. However, the time consuming handling of specimens suggested maximising image resolution. Staying focused on the goal at hand simplified this dilemma. The primary goal in this case was to obtain good quality label images that could be easily read; the secondary goal being to capture an image of the specimen at the same time. With this in mind we established the following criteria for the project:

- Maximum 5MB file size
- Create an image of clearly readable text on labels, this being the priority
- Produce a clear, focused image of specimens at maximum resolution allowed by inclusion of labels in the same image – these specimens which will range in size from large cicada's and moths (measured in many cm's) to small beetles and flies (measured in the few mm's) so detail that can be captured will vary.
- Attach a registration number
- Use relatively low cost imaging and computing equipment.

- Create a partial record of metadata, including the species name and registration number.
- Develop simple standardised processes that could be easily replicated and implemented on multiple workstations by volunteers
- Develop a process that would be comparable in speed to direct data entry by volunteers
- Ensure specimen safety with minimal breakages
- Maintain a harmonious working relationship with collection staff

This paper outlines the methodology of this project (for more detail on the materials and methods see [AM digitisation final report]), reports on the outputs, and discusses the issues encountered and lessons learned.

Methods

Database for storage of image metadata

An important component of digitising infrastructure was the database in which image metadata and short record information was initially entered into by the volunteers. This database was separate from the corporate collection management database for a number of reasons:

- Data Security - the corporate database had strict permissions on access for purposes of maintaining data integrity. In this project, volunteer staff, did not have data entry access
- Direct data entry into the corporate database can be slow and not as efficient or effective as using a lightweight MS Access database for data entry and validation followed by bulk importing the records into the corporate database

We chose MS Access as the platform for this database because:

- There was an existing software licence for MS Access
- The database support officer for this project has existing expertise in setting up, managing and programming in MS Access.

Where possible information stored in the database was made available as pick lists so that data entry required as little typing as possible, reducing input error and making the process faster.

At the time of image capture volunteers enter information through the MS Access database data entry form (Figure 1).

The data captured by the volunteers includes data necessary for creating a "short record". It contained the bare minimum of detail about a specimen to enable the

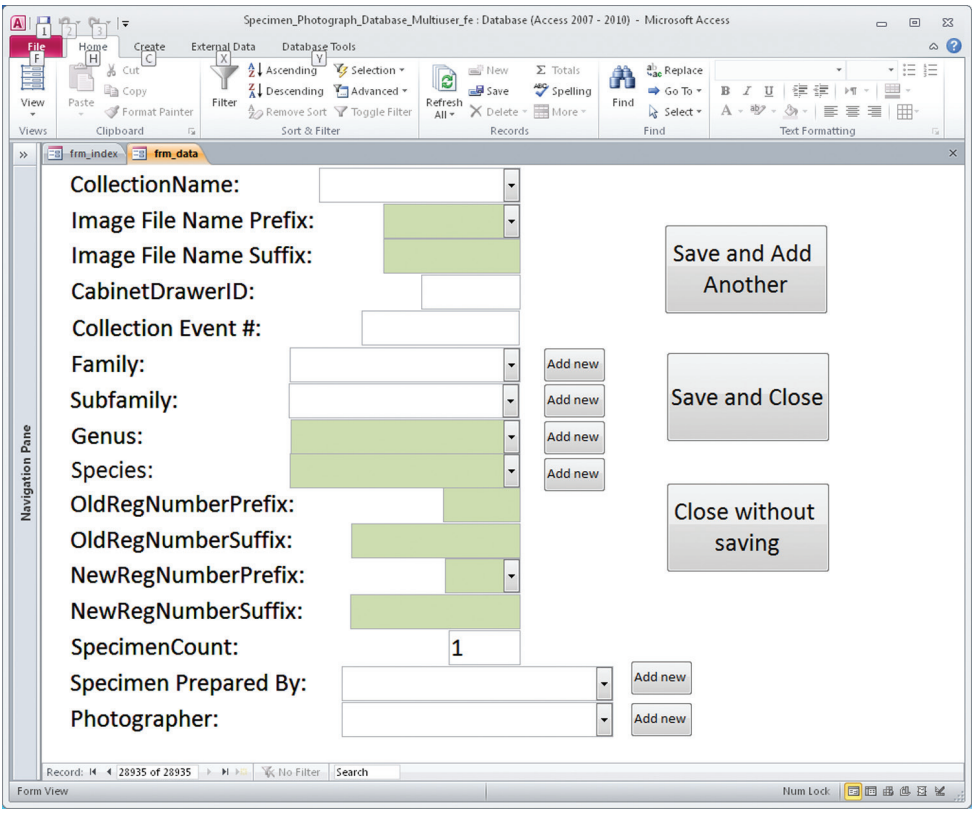


Figure 1. Database for entry of image metadata.

creation of a valid collection database record in the museum's collection management database, EMu. This short record which consisted of catalogue number, species (or a higher taxon level) name and the images themselves were imported into EMu. The rest of the label data, once captured could then be appended to this short record at a later date. The short record in the meantime is available for audit purposes, and for some collection and data management activities, as the specimen label data can clearly be seen on the image even though it is not text searchable.

Digitisation Laboratory

The Australian Museum provided a large room in which the Digitising Laboratory was established. This space was important in establishing the sense of belonging for the volunteers as it was a dedicated space for the project. The room was fitted out with power and network outlets and secure access. There was enough space for four specimen label imaging workstations, one register imaging station, a microscope camera workstation and three transcription workstations.

Specimen workstations

Each of the four specimen imaging workstations had the same equipment. Workstations were used for imaging individual specimens and their labels.

Workstation equipment:

- Two desks
- A desktop computer capable of manipulating large images
- Camera, lens and light source (flash)
- Copy stand for vertical photography
- Tools for specimen handling such as tweezers
- For a full listing of equipment see [AM digitisation final report]

Figure 2. A digitising workstation.

Process/workflow

Selecting and providing specimens for digitisation involved the Museum collection staff selecting appropriate curated drawers of specimens in preparation for imaging. Selection criteria included ensuring that the taxonomy and names for specimens in drawers were as up-to-date as possible, and unambiguous. Specimens also needed to be mounted, relatively robust (this is not essential but inclusion of less robust taxa generally led to more breakages and so required more collection staff time in resolving breakages), to avoid damage when being handled, and accessible within each drawer, for example, not cramped where labels and specimens would be damaged.

It was a large task to ensure the supply of specimens for digitising. With four workstations operating four days per week, the rate at which specimens can be digitised put a considerable burden on collection staff. Curated drawers needed to be allocated for up to a month in advance to ensure that volunteers would not run out of specimens to digitise.

Summary of the steps in handling and imaging of specimens

(for full details see [Specimen Training guide])

- Curate specimens

Before moving drawers of specimens to the imaging laboratory, collection staff ensured they were curated to a specified level for the project. Type specimens were removed, as they were considered too precious to be handled by those without appropriate training. Each specimen was checked to ensure it was labelled adequately with its taxonomic name, and that drawers were not overcrowded. This workload had significant resource implications for collection staff (see Discussion and Conclusion Table 1)

- Retrieve specimens

Drawers were removed from the collection by the digitisation officer in the order that they were numbered and transported to the Digitisation Lab. Drawers were transported on a trolley from the collection to the Digitisation Lab.

- Prepare specimens

At each workstation there were two volunteers: one volunteer handled the specimen (the specimen handler), the other volunteer photographed the specimen (the digitiser).

For more details of the process see [Specimen Training guide].

- Image specimens

Label information was entered into the database, and the specimen and its labels were imaged.

- Deal with damaged specimens/labels

Damaged specimens (broken parts are collected and included with the specimen), damaged labels, and specimens without labels were placed in a 'hospital' drawer to be returned to collection staff for assessment and repairs. Place holders were used to identify where the specimens were to be returned to. The collection manager was notified when the hospital drawer was full.

- Return specimen drawer to the collection

After a specimen had been imaged it was replaced in its drawer.

Once imaging of all specimens in a drawer is finished, a drawer could be returned to the collection.

Figure 3. An example of a specimen and label image, in this case a hawk moth.

- Review image and entered data

After each drawer had been imaged the information entered in the database was reviewed to ensure consistency and identify any obvious image or data capture problems.

- Monitoring of the process

The following information is recorded to monitor the project's outputs, staffing and volunteers:

Number of specimens digitised per day by volunteer
Number of drawers digitised per day
Number of volunteer hours per day
Number of damaged specimens
Number of specimens damaged beyond repair
Number of collection staff hours per day

Volunteer recruitment, supervision and management

Recruitment, coordination and supervision of volunteers

The development and management of a team of well trained, productive volunteers dedicated to the digitisation of museum collections, whether large or small, required the same basic approach.

A volunteer coordinator was essential. In practice, this role and its responsibilities could have been spread across one or more people or positions. Ideally, however, we felt that a single position, which in the case of this project was shared between two individuals, was likely to produce the best outcome for the museum. The large size of the volunteer team and the high throughput of the project required a dedicated co-ordinator resource to ensure that the workload of existing staff was not impacted greatly.

The volunteer coordinators were responsible for recruiting, training, coordinating and supervising volunteers. First and foremost they needed to have excellent people management skills and a good understanding of the technical processes involved in digitisation. The extent to which they need to be technically proficient was dependent on the availability of other sources of technical expertise. Coordinators were trained in specimen handling, the extent that they assisted the collection staff in developing a video demonstrating how volunteers should handle specimens.

The major steps in the creation of the volunteer team were as follows:

- Recruitment

An expression of interest email was sent out to Australian Museum Members. Potential volunteers were asked to identify their preferred days, which could be a Saturday, and their availability, to volunteer for one day a week, or one day a fortnight.

- Rostering

Potential volunteers were prioritised on their day preferences according to their response time to the expression of interest.

- Induction

New volunteers were given an introduction to the working area and other volunteers and a tour of the public exhibition within the museum.

- Training

Volunteers attended a one day training session with short videos about handling and imaging specimens. The videos were accompanied by training manuals, and followed by hands-on practice with experienced volunteers.

- Review

Digitiser volunteers undertook a six-week introductory period. At the end of this period each volunteer would complete a self-assessment review of their practice and the project.

- Ongoing Support

Digitiser volunteers received ongoing practise support from peers as well as the digitisation officer.

Results

Each specimen digitised resulted in a "short" record in EMu with taxon and registration number linking with existing taxonomic information in EMu, and included an image of the specimen with associated labels.

The following graphs show various statistics over the period of the project.

As the number of workstation hours per month varied (Fig. 4) so did the number of specimens digitised per month (Fig. 5). The drop off over December was due to the Christmas holiday period.

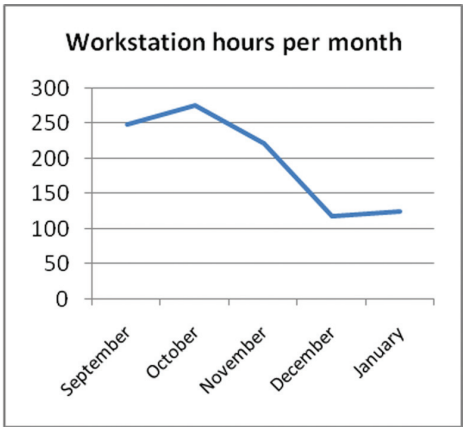


Figure 4. Workstation hours per month.

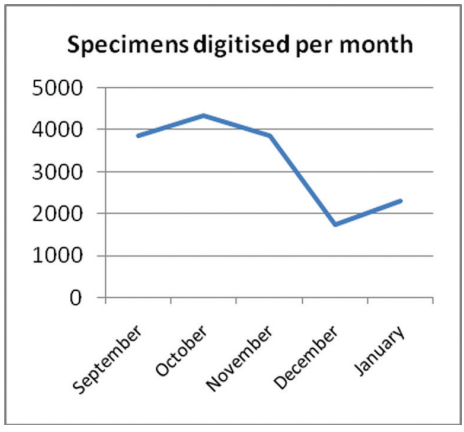


Figure 5. Specimens digitised per month.

The number of specimens digitised per workstation hour (Fig. 6) was reasonably consistent throughout the project. Compared to the relatively large changes in daily productivity (see Fig. 8) when averaged over a month productivity was reasonably constant.

The number of damaged specimens per month (Fig. 7) was related to the number of specimens being processed and the fragility of the group being worked on. The highest rate of damage was with the Sphingidae moths. All damaged specimens need to be dealt with by collection staff, so an increase in the numbers damaged, meant an increase in collection staff time to remedy.

There is great variability in the number of specimens digitised per workstation hour (Fig. 8). This was not simply a factor of the number of workstation hours per day. Other factors, such as volunteer competency and diligence, and the taxonomic group being worked on, also influenced the digitising rate. The variation in the workstation hours (Fig. 9) is due to variable volunteer attendance which was more likely to be affected by external factors than is the case with paid staff, a factor that affects productivity.

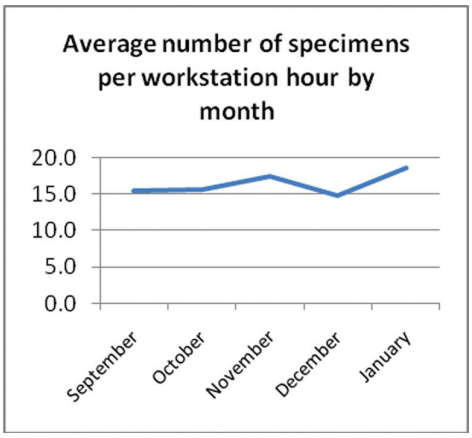


Figure 6. Average number of specimens per workstation hour by month.

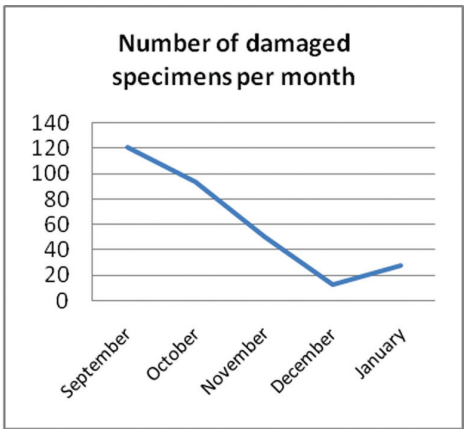


Figure 7. Number of damaged specimens per month.

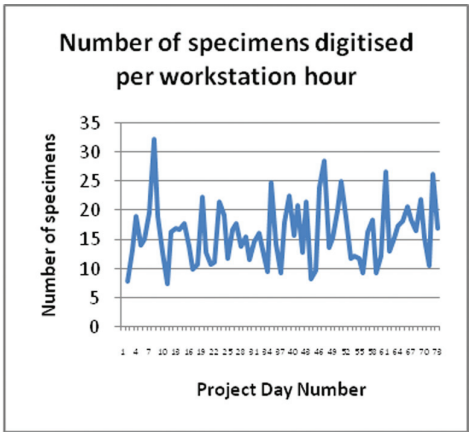


Figure 8. Number of specimens digitised per workstation hour.

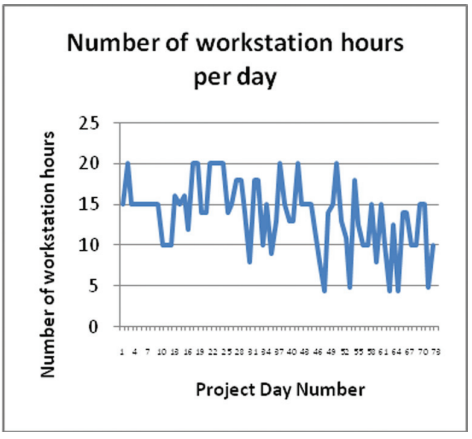


Figure 9. Number of workstation hours per day.

Discussion and conclusions

Why use image based digitisation?

There was a time when databasing (entry of text only data) of collection holdings was the preferred way of digitising a collection, for collection management and data access. This is now being challenged by the digitising of collections where specimens are imaged and their associated label data entered as complementary data.

The advent of this approach has come from the realisation that having an image of the specimen and its associated labels has strong collection data management benefits including:

- A readily accessible digital voucher of each specimen and its labels for verification and reference as a digital loan
- Reduced need for specimen handling
- A virtual specimen in the event of collection loss or damage, e.g. fire, flood, earthquake, or for when the specimen is on loan
- Remote access to original label data for review by researchers
- A capacity for using handwriting to help identify a collector in the absence of a collector name
- A limited potential for species identification from an image
- Enabling the use of 'non-experts' in data entry with the benefit of knowing data quality and enabling dubious data to be checked without having to physically visit a specimen in the collection.

Best practice

The processes and procedures detailed in this paper constitute best practice for the predefined goal of image based digitising of individual specimens and the associated labels. We have produced documentation and videos that detail the handling of specimens and registers and a handbook for volunteers involved in the digitizing project [Rapid Digitisation Project Resources].

The two most important components of this best practice are:

- the dedicated role of the digitising officer who recruits, trains and coordinates the volunteers, liaises with collection staff and implements the technical processes
- the curation of material prior to digitisation by collection staff which makes the digitising process as effective as possible in terms of consistent identification of digitised specimens, ease of handling and selection of appropriate specimens (including removal of types).

Curation of specimens by drawer in preparation for digitising

A factor that should not be glossed over is the potential resource impact of preparing drawers of specimens for digitising. This curation involves removing types (where it has been decided that types are not to be imaged by volunteers, as is the case in the AM project), ensuring specimens within the drawers are labelled adequately with taxonomic names and that the specimens are not overcrowded and thus difficult for volunteers to handle. These are tasks which must be carried out by collection staff or, where appropriate, experienced volunteers.

Institutions need to ensure that adequate resources and lead time are made available to allow collection staff to curate drawers well ahead of scheduled digitising for those

drawers. In addition, any application for funding of digitising projects needs to factor in resources required for the curation of material to be digitised.

It is difficult to estimate how much time is required specifically for the curation of specimens for digitising (Table 1) because curation is a normal part of collection management. What is clear though is that a dedicated rapid digitising project shifts the priorities of collection staff onto curation of specimens that may otherwise not have been in their workplans. Unless effectively resourced this can lead to conflict over work priorities in the collection. All insect collections contain a mixture of groups which range from well identified (usually because of associated input from skilled researchers) to completely unidentified or incorrectly identified. There may be well-identified groups that are not suitable for digitisation because of their physical state or there may be groups which would be ideal to digitise which are not well identified. The presence of well-identified collection material represents an investment that should not be taken for granted.

Table 1. This table gives some idea of the difficulty in estimating time required for curation. These are figures provided by David Britton, Collection Manager at AM.

Group	No. of drawers	No. of specimens	Curation time estimate
Notodontidae (moths)	26	~ 1000	7 days, included some identification (4 drawers)
Sphingidae (hawk moths)	50	1916	8–15 days, included some identification (6 drawers)
Cicadas	82	4386	25 days
Scarabaeidae (beetles)	5	2204	15 days including identification
Noctuidae (moths)	10	809	4 days
Leafhoppers etc	41	3385	20 days

Technical support for creating and maintaining databases and photographic equipment

The entry of metadata for each image captured is an important consideration as it has implications for resourcing, productivity and the ease in which data can be incorporated into the institutions collection management system, in the case of AM that being KE-EMu.

Databasing at the time of image capture could be carried out in a number of ways including: direct entry into a spreadsheet such as Excel, entry into a purpose built database such as MS Access for later/subsequent import/uploading or direct entry into the corporate collection database. We chose the MS Access option because it optimises data entry speed and accuracy (through the use of picklists, default data values and automated field population), and doesn't carry security overheads (volunteers accessing the corporate database has unacceptable data security issues).

Technical support is required to establish and maintain the database and the various entry forms required.

Options for Capturing Full Records

The complete label information could be transcribed and entered into a spreadsheet or database at the time of image capture. However, this approach wasn't adopted because it was felt that separating out the imaging and transcription steps has benefits for the process in terms of specialisation which is likely to result in improvements in speed, efficiency and accuracy.

Imaging the labels allows scope for unlocking and outsourcing the transcription of the complete label data to create a full record.

Two options that utilise volunteers are as follows:

- Internal Volunteers – by setting up separate computer workstations with either spreadsheets or data entry database forms, the label information could be transcribed by the volunteers in the DigiVol laboratory.
- Crowdsourcing with Online Volunteers – the approach chosen was to establish an online volunteer transcription site [Biodiversity Volunteer Portal] where the complete label information can be transcribed into defined database fields. This data can then be validated and imported back into the Emu collection database to create the full database record.

Funding Options

Our investigation of funding options came to the following conclusion on likely sources of funding for digitising projects:

It is far easier to get funds for buying equipment and building infrastructure than it is for 'bums on seats'.

With this in mind, some institutions may find it worthwhile to seek funds for equipment purchase and then allocate some existing internal resources to set up the equipment and coordinate the volunteers in a manner that is amenable to their available staff resources.

Short term projects of one or two years may be funded through trusts, particularly those related directly to the institutions activities, e.g. The Australian Museum Foundation. Such short term projects should focus clearly on delivering a specific content such as a charismatic or high profile collection in its entirety.

In the absence of either of these sources of funding it is dependent on the institution itself to determine its priorities in terms of digitisation and focus what resources it can in pursuing those priorities.

Low cost digitising options

Where institutions are unable to implement best practice because of resourcing constraints the processes and procedures outlined above can be scaled to suit available resources. For example, a single workstation could be established at minimal cost and a small team of volunteers (two to ten) trained and coordinated by an existing staff member if the workstation was located in close proximity to the staff member.

- Equipment selection

The cost of setting up a workstation is somewhat flexible in that many institutions will already have the necessary equipment for specimen handling and curation and also the necessary furniture. This can considerably reduce the costs of setting up a workstation, reducing it to just the cost of imaging equipment and computer software and hardware.

- Computer

A fast but standard specification was chosen to get the best balance between price and performance. Two screens were used: a larger screen for viewing the images (as image capture is controlled through the computer) and another screen for operating the database for data input.

- Copy stands

Good quality copy stands are essential as they provide stability for the camera and a sound platform upon which the specimens can be imaged. Kaiser makes excellent stands.

- Cameras

Any number of cameras could have been chosen and would have been suitable to the task. We chose a Canon 550D as we felt it delivered good results was sufficient to do the job and represented very good value for money. We felt there was no need for a more expensive camera nor a higher resolution camera because we wanted to keep the image size to a manageable 5MB jpeg.

- Storage

When capturing many thousands of images at 5Mb size per image, the impact on storage is significant. Images are stored on the Museums network as part of its image storage infrastructure. Funding for future image capture has been factored into the Museums overall IT planning.

Acknowledgements

The authors would like to acknowledge the assistance of Rhiannon Stephens and Leone Prater with developing the digitising project, Michael Elliott for his excellence in data-

base and imaging support, David Britton for his support and assistance with the AM Entomology collection, John Gollan and John Tann for their assistance in developing the ideas underpinning this project; the generous and dedicated volunteers involved in the DigiVol project and Isobel Kindley for assistance in establishing the Volunteer Program. The DigiVol project was made possible by the Atlas of Living Australia.

References

AM digitisation final report. http://www.ala.org.au/wp-content/uploads/2011/10/Australian-Museum-digitisation-project-final-report.pdf

Biodiversity Volunteer Portal. http://volunteer.ala.org.au/

Rapid Digitisation Project Resources. http://www.australianmuseum.net.au/Rapid-Digitisation-Project

Specimen training guide: A Guide to the Handling and Digitising of Specimens http://www.australianmuseum.net.au/document/Specimen-Training-Compressed

Tann J, Flemons PKJ (2008) Data capture of specimen labels using volunteers. Internal Report, Australian Museum, 17 pp. http://australianmuseum.net.au/Uploads/Documents/23183/Data%20Capture%20of%20specimen%20labels%20using%20volunteers%20-%20Tann%20and%20Flemons%202008.pdf

ZooKeys 209: 219–233 (2012)
doi: 10.3897/zookeys.209.3472
www.zookeys.org

RESEARCH ARTICLE

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

The notes from nature tool for unlocking biodiversity records from museum records through citizen science

Andrew Hill[1], Robert Guralnick[2], Arfon Smith[3], Andrew Sallans[4], Rosemary Gillespie[5], Michael Denslow[6], Joyce Gross[5], Zack Murrell[6], Tim Conyers[7], Peter Oboyski[5], Joan Ball[5], Andrea Thomer[8], Robert Prys-Jones[9], Javier de la Torre[1], Patrick Kociolek[2], Lucy Fortson[3]

1 *Vizzuality, New York, New York, USA* **2** *University of Colorado, Boulder, Colorado, USA* **3** *Adler Planetarium, Chicago, Illinois, USA* **4** *University of Virginia, Charlottesville, VA, USA* **5** *University of California Berkeley, Berkeley, California, USA* **6** *Appalachian State University, Boone, North Carolina, USA* **7** *Department of Zoology, Natural History Museum, Cromwell Road, London SW7 5BD, UK* **8** *University of Illinois, Urbana-Champaign, Champaign, Illinois, USA* **9** *Bird Group, Natural History Museum at Tring, Akeman Street, Tring, Herts HP23 6AP, UK*

Corresponding author: *Andrew Hill* (andrew@vizzuality.com)

Academic editor: *V. Blagoderov* | Received 6 June 2012 | Accepted 16 July 2012 | Published 20 July 2012

Citation: Hill A, Guralnick R, Smith A, Sallans A, Gillespie R, Denslow M, Gross J, Murrell Z, Conyers T, Oboyski P, Ball J, Thomer A, Prys-Jones R, de la Torre J, Kociolek P, Fortson L (2012) The notes from nature tool for unlocking biodiversity records from museum records through citizen science. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 219–233. doi: 10.3897/zookeys.209.3472

Abstract

Legacy data from natural history collections contain invaluable and irreplaceable information about biodiversity in the recent past, providing a baseline for detecting change and forecasting the future of biodiversity on a human-dominated planet. However, these data are often not available in formats that facilitate use and synthesis. New approaches are needed to enhance the rates of digitization and data quality improvement. Notes from Nature provides one such novel approach by asking citizen scientists to help with transcription tasks. The initial web-based prototype of Notes from Nature is soon widely available and was developed collaboratively by biodiversity scientists, natural history collections staff, and experts in citizen science project development, programming and visualization. This project brings together digital images representing different types of biodiversity records including ledgers , herbarium sheets and pinned insects from multiple projects and natural history collections. Experts in developing web-based citizen science applications then designed and built a platform for transcribing textual data and metadata from these images. The end product is a fully open source web transcription tool built using the latest web technologies. The platform keeps volunteers engaged by initially explaining the scientific importance of the work via a

short orientation, and then providing transcription "missions" of well defined scope, along with dynamic feedback, interactivity and rewards. Transcribed records, along with record-level and process metadata, are provided back to the institutions. While the tool is being developed with new users in mind, it can serve a broad range of needs from novice to trained museum specialist. Notes from Nature has the potential to speed the rate of biodiversity data being made available to a broad community of users.

Keywords
Natural History Museums, Biodiversity, Open Source, Museum Collections, Citizen Science, Digitization, Transcription

Introduction

Natural history collections represent irreplaceable legacy information about our biosphere. In an era dominated by planetary-scale anthropogenic change (Walther et al. 2002, Parmesan and Yohe 2003) and unprecedented biodiversity loss (Jenkins 2003, Loreau et al. 2006, Wake and Vredenburg 2008), both historical and recent biocollections and their associated data represent valuable benchmarks for analyzing the biological impacts of environmental change and determining its causal factors (Moritz et al. 2008, Rainbow 2009, Pyke and Ehrlich 2010, Erb et al. 2011). The knowledge derived from specimens has been a critical component in studies of invasive species (Giovanelli et al. 2008, Rödder and Lötters 2009); biological conservation (Pawar et al. 2007); land management (Ochoa-Ochoa et al. 2009); pollination (Biesmeijer et al. 2006); species distributional (Lyons and Willig 2002, Peterson 2003, Moritz et al. 2008, Peterson and Martínez-Meyer 2009) and phenological (Nufio et al. 2010) responses to climatic change; spread of pathogenic organisms (Moffett et al. 2009, Soto-Azat et al. 2010); species discovery (Bebber et al. 2010); and forecasting future changes (Graham et al. 2004).

It is estimated that the number of specimens in natural history collections could range anywhere from 1 billion for just arthropods (Nishida 2003) to 2 billion records for all collections (Ariño 2010). Whatever the final number, the current representation of digitized records is much less. The Global Biodiversity Information Facility (GBIF) maintains the largest single portal to digital species occurrence records -- currently provisions about 400 million records, many of which are from citizen observation networks and not natural history collections. Further, the taxonomic representation in GBIF is skewed to those taxonomic communities and regions of the world where support for digitization has been strongest. While the current digital available representation of vertebrates in Western Europe and North America may be quite good, for groups such as insects in regions such as the tropics, our data remain particularly limited (Guralnick and Hill 2009). Biocollections contain abundant historical records (Boakes et al. 2010) that help fill the gaps from early time-periods, often pre-dating massive human-caused changes to landscapes. Furthermore, these collections often contain important biological records that can help further the study of biodiversity today (Pyke and Ehrlich 2010).

Despite the well-documented value of biocollections for science and society, the ability of researchers and policy makers to utilize this resource is hampered because many specimen data remain sequestered within institutions in non-digital formats. Digitization, transcription, description, and mobilization of specimen data (including label data, images, field notes, illustrations, and gene sequences) improves data discovery, interoperability, and enhancement (Edwards et al. 2000, Canhos et al. 2004, Soberón and Peterson 2004, Guralnick and Hill 2009), but these activities are not automatic, and present technical and organizational challenges (Pennisi 2005, Berendsohn and Seltmann 2010). Many institutions lack the financial, technological, or staffing resources needed to complete the many tasks required to deliver well-described digital data to data consumers (Vollmar et al. 2010). Even those institutions fortunate enough to have the needed resources and capacity may still want to utilize new methods that engage the public, serve educational missions, and potentially deliver more error free data while also scaling down total digitization costs.

Specimen digitization (i.e. digitally capturing each component of the specimen label and at times the specimen) is a multi-step process, and one of the most expensive and time-consuming of those steps is transcribing the labels into textual formats essential for further description and querying. This is particularly challenging when labels are hand-written, rendering other techniques such as optical character recognition (OCR) mostly useless. While OCR can prove valuable with printed or typed labels, and will undoubtedly play an important role in the future, the technology is still prone to errors that need to be corrected and validated. There is, however, a potentially transformational solution to this problem: working with citizen science volunteers across the world to help with transcription tasks.

Citizen science, where volunteer researchers are asked to help create or process scientific data, is becoming popular on the web (Zooniverse, https://www.zooniverse.org/; Folding@home, http://folding.stanford.edu/) and in web-enabled field collection (eBird, http://ebird.org/; iNaturalist, http://inaturalist.org/). Biological specimen transcription is a task well suited for citizen science, and a small number of projects have already been developed. Herbaria@home (http://herbariaunited.org/atHome/) for example, provides a portal to the herbarium sheets from primarily the United Kingdom and Irish herbaria. The work done by Herbaria@home has helped unlock over 100,000 specimens, making them digitally available for further science research. A more recently launched project, Atlas of Living Australia (ALA) Biodiversity Volunteer Portal (http://volunteer.ala.org.au/), has a broader scope, digitizing records and field notes from Australia's biodiversity collection. The ALA site builds missions and encourages users to earn badges for their efforts. The Volunteer Portal has brought in around 200 volunteers who have completed nearly 20,000 transcription tasks.

Here we describe for the first time a prototype citizen science application for transcribing cross-institutional, taxonomically diverse, natural history ledgers and labels called *Notes from Nature* (http://www.notesfromnature.org/; Figure 1). In describing this tool and how it was designed, we hope to also provide insights into data management and quality assurance methods, volunteer engagement practices, and education and reward mechanisms in online citizen science project development. We frame our

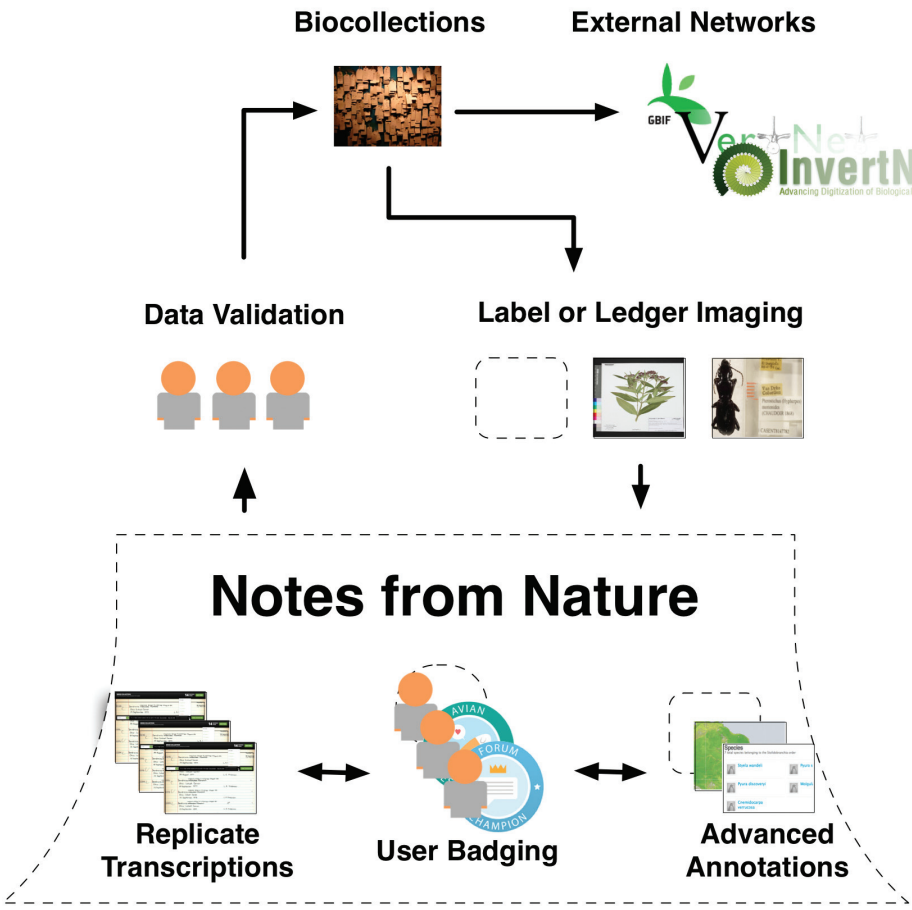


Figure 1. Organization of the Notes from Nature platform.

development process using knowledge and tools gained from other Zooniverse projects, which has pioneered web-based citizen science in other disciplines, while discussing unique aspects of working with natural history specimen based image sources. In particular, we discuss topics important to the development and management of citizen science applications, such as methods to provide user feedback, communication and rewards to volunteers, and testing accuracy compared to more traditional transcription practices.

Methods and results

Data resources for initial phase of notes from nature

Notes from Nature is currently in a prototype phase and was developed in a collaboration between institutions and consortium including: Natural History Museum London bird collection (NHMUK; http://www.nhm.ac.uk/research-curation/depart-

ments/zoology/bird-group/index.html), the Southeast Regional Network of Expertise and Collections (SERNEC; http://www.sernec.org/) organization, Calbug (http://calbug.berkeley.edu/), and the University of Colorado Museum (http://cumuseum.colorado.edu/Research/Zoology/). The NHMUK contributes an iconic group of organisms with a long history of enthusiasts and volunteer communities – birds. SERNEC is a collaboration of Southeastern United States herbaria to bring collections "online" in part through digitization efforts of herbarium sheets. Calbug is a collaboration involving multiple entomological collections in California and coordinated by the University of California Berkeley's Essig Museum of Entomology (EMEC); one goal is to provide a model for the digitization of diverse and digitally underrepresented arthropod specimens. The University of Colorado Museum of Natural History (UCMNH) is providing a unique validation dataset discussed in more detail below.

The input data and images from these three groups fall into three different categories. The NHMUK data consist of images of hand-written ledger pages that contain each component of a record organized in rows and columns (Figure 2a). SERNEC provides images of plant specimens with associated labels: in this case, specimens are flat, and are therefore particularly amenable to photographing, and suffer minimal image loss or distortion in the third dimension (Figure 2b). The Calbug digitization processes are particularly challenging because individual specimens are mounted, along with labels, on pins (Figure 2c). Each specimen is carefully removed and photographed alongside each associated label. The three projects have independent, and for SERNEC and Calbug, ongoing imaging initiatives that are driving content for Notes from Nature.

We have collected an additional 100 images, representing ledger pages of bird specimens containing over 1000 records from UCMNH, to be used as reference standards. The full set of these records has already been databased once, creating an objective standard of quality for comparison. These images were then re-transcribed by trained museum staff in Fall of 2011 using current best practices in order to calculate rate and current cost. The transcription of these records will then also be duplicated by Notes from Nature volunteers. Local "staff" and citizen science retranscriptions will then be compared to the original datasets in order to generate statistics regarding accuracy, speed, and required training of the volunteer community to create data on the Notes from Nature platform. We will make such statistics publicly available on the Notes from Nature blog. We note that this initial comparison, although useful, may not

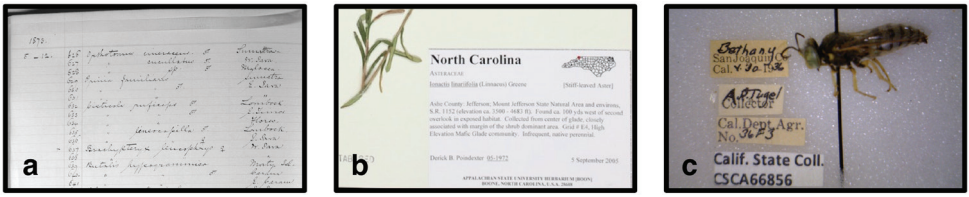


Figure 2. Example biocollections source images showing (**a**) The Natural History Museum, London bird specimen ledger; (**b**) The Southeast Regional Network of Expertise and Collections herbarium sheet label; **(c**) Calbug specimen and label image.

generalize to other types of material (e.g. herbarium sheets, specimen labels). However, such initial statistics are of high value given only anecdotal information by which to judge cost efficiency and quality. Further such tests can only help provide assessment of the cost and quality effectiveness of the citizen science approach.

Notes from nature platform design overview

Notes from Nature is being developed with personnel and programming support from The Citizen Science Alliance (CSA; http://www.citizensciencealliance.org/), which develops and maintains a roster of projects called the Zooniverse (http://www.zooniverse.org/), and Vizzuality (http://www.vizzuality.com/), a CSA parter that specializes in biodiversity visualization. A core team of CSA developers, designers and educators is funded by a grant from the Alfred P. Sloan Foundation that promotes the development of new citizen science projects at the Zooniverse. Zooniverse projects are growing in diversity but each project builds upon a set of technologies that aid common features across projects such as transcription data collection and user communication (https://github.com/zooniverse).

The front end of the platform is built on a stack of the latest web-technologies using JavaScript and HTML5. The transcription tool, for example, uses a mix of HTML5 Canvas and JavaScript to give the user a simple mechanism for capturing each record's location and content. The system is designed to have different user-interfaces tailored to the image layout and information displayed. For example, the transcription tool layout for row-and-column based ledger page images (Figure 3) will differ from the layout for mounted plant specimen and label images. The tool is open-source and code is available online at https://github.com/Vizzuality/BioTrans.

The design of Notes from Nature takes it cues from other successful Zooniverse projects. Any person with Internet access can create a Zooniverse account and join the project (or any other project in the Zooniverse). Prior to performing any transcription, a new user is led through a short series of tutorials. These demonstrate the process of accurate transcription, but more importantly explain how and why the data are important to scientists. In previous Zooniverse projects, orientation tutorials have proven especially valuable for imparting the urgency and value of the work which in turn provides initial motivation for involvement (Raddick et al. 2010).

Notes from Nature organizes the raw data – digital images – in three different ways: by projects, by collections, and by missions. "Projects" are large, unified, datasets provided by partner museums or consortiums or museums. SERNEC and Calbug are two distinct examples of projects. "Collections" are the organizing subunits within projects. For example, Calbug is a collaboration across eight different institutions, and each institution that has records in Notes from Nature will be referred to as a "collection". The three projects are shown on different pages of the Notes from Nature site so that volunteer transcribers can learn about the projects and collections that interest

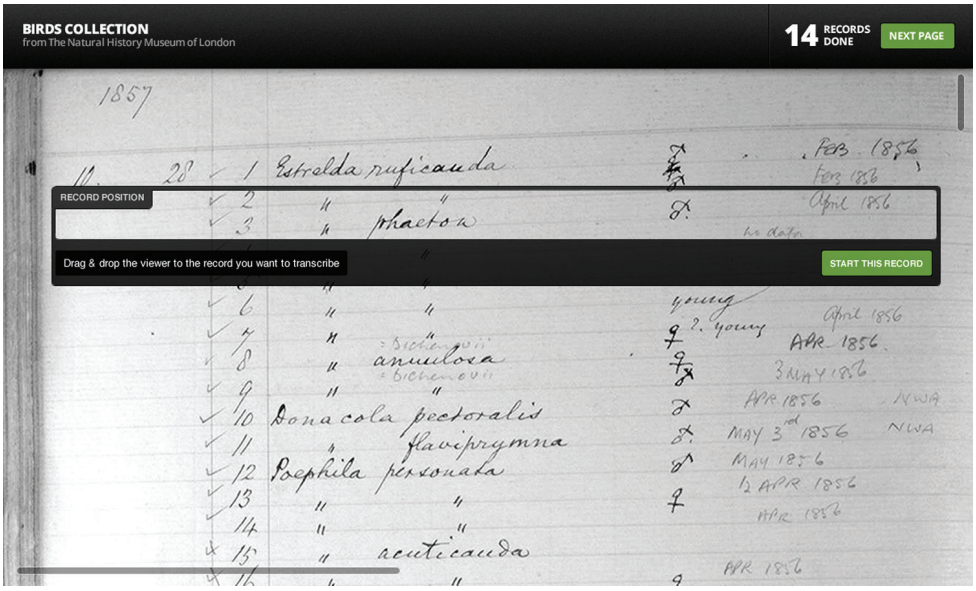


Figure 3. The Notes from Nature transcription tool for NHMUK museum ledgers. The tool gives users basic methods to navigate through a page of collections records while transcribing each major component of the record, viewing help dialogs, or skipping difficult to transcribe record entries. For help dialogs, we provide more than one example for each record element. The record outline is a movable window and, during transcription, the image and the tool location on that image is also captured as metadata, so that data managers can return quickly return to the source material for any record.

them them most. While the real world organization of projects and partners can be complex, the simplification is intended to help users find relevant information about the specimens they are transcribing. Finally, the Notes from Nature team is developing "missions" that thread narratives across or within projects and collections. Missions are meant to engage the users, especially those with special interests in a particular organism or group of organism (e.g. beetles) or regions (e.g. west African tropics). Each missions has a clear end-point, where every record in the mission is transcribed or determined to be too challenging for transcription and the mission is considered complete.

During the transcription process on Notes from Nature, the user examines and transcribes records or ledger pages one at a time. The work a user performs is recorded, and elements of that work will be displayed as part of their personal profile page; a user's personal data may include what collections they have worked, how many missions in which they have taken part, or on what missions they are currently working. As discussed below in more detail, transcribers are also rewarded for completing certain kinds of tasks, acquiring badges for different kinds of activities such as completing a certain number of records in a particular taxonomic group or geographic area, finding new and unusual records such as previously unrepresented species of organisms.

Transcription and storage of results using notes from nature

The transcription tool is the workhorse of Notes from Nature, capturing both text inputs from the user along with its own position and the page on which it is being used. Volunteers move the tool to overlap a single specimen record among the many on a ledger sheet, and then transcribe and categorize the components of each record, such as collector, geographic, temporal, and taxonomic fields. In all cases, a record of the image or page of the scanned material, the record's identification in a collection or project, and the location of the transcription on the digital image are stored in a MongoDB back end hosted by the Citizen Science Alliance.

The accuracy of transcriptions generated in Notes from Nature is evaluated by collecting at least three replicate transcriptions for every record (Figure 4). The level of convergence by volunteers is used to evaluate confidence in the output (Lintott et al. 2008). The accuracy for each field within a record (such as date of

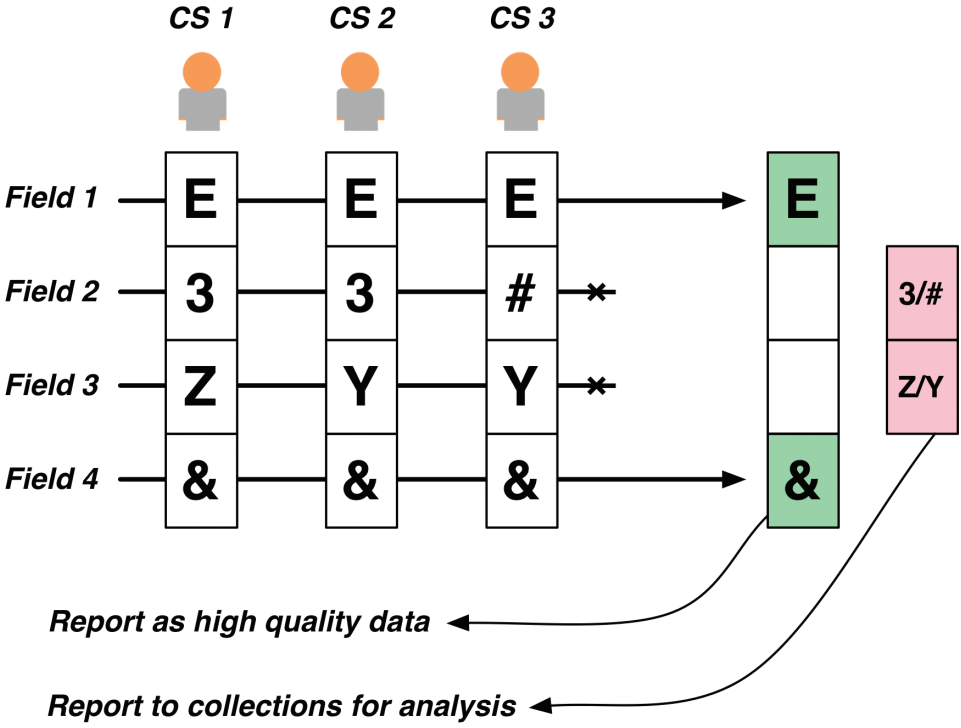


Figure 4. The simplified transcription replication and validation step. Following three independent transcriptions of a record, data is reconciled and returned to the original data provider. Records sent back to the provider can be fully complete, partially complete, of fully incomplete. Fully complete records are those where all three citizen scientist volunteers (CS) agree on every field of the record. Partial records include only those fields where CS agree. Fully incomplete records indicate that volunteers were largely unable to transcribe the record consistently. Data collected that does not become part of the final record is still made available for further review by the data provider.

collection or species name) can be measured independently, allowing trained staff to then revisit problematic records and work to resolve discrepancies outside of the Notes from Nature platform.

The full record collected at transcription, including all multiple replications, are returned to the original data providers as both "raw" outputs and summaries that can provide quick views of progress (number of records transcribed on a day, total hours spent, etc). Notes from Nature will assure that the core fields, and other parts of records that are valuable to collect but might be idiosyncratic to a collection, meet community standards (Wieczorek et al. 2012). We will ask all users to transcribe records verbatim. The task of the citizen scientist is not to correct the original data, but instead to make it digitally available. In later versions of Notes from Nature, we plan to include interfaces for advanced users to suggest corrections to the original record. Part of this future work will be cleaning records to conform to the controlled vocabularies in standards such as Darwin Core.

For the Notes from Nature initial prototype, the goal is to assure that the essential fields of each partner institution are captured verbatim, with metadata about collection and replication. Core members of the Zooniverse and Vizzuality teams will be working with the project leads to ensure the data is captured effectively and returned to the home institutions in formats most useful for further integration back into databases. As per collaboration agreements, all data collected from this project will be made freely available online in usable formats (e.g. Darwin Core records) by the collaborating projects (NHMUK, SERNEC, Calbug) or their member institutions.

Volunteer engagement and incentives

The methods for engaging volunteers in the Notes from Nature project can be categorized in three ways: communication, transcription feedback and narratives, and incentives.

> Communication: Notes from Nature, like most projects on Zooniverse, encourages users to interact with both scientists and other volunteers in a purpose-built discussion platform (https://github.com/Zooniverse/Talk) and via live-virtual discussion. The live discussion interfaces serve as an excellent medium for comments and questions and also become a focal point of communication to and from the researchers that are interested in seeing this data inform future science and conservation. Like other CSA projects, Notes from Nature will have a blog for communicating and archiving major news, discoveries, and milestones to the community. The blog will also become a tool for outreach, seeking new volunteers from existing clubs and communities.
>
> Transcription feedback and narratives: Notes from Nature will provide immediate information about how a user's actions are expanding the library of information for scientific research. Records transcribed can be shown as part

of a "collective map" illustrating how new records streaming in from all Notes from Nature volunteers are closing gaps in our knowledge. Similarly, users will be given data-driven narratives such as collector histories, where we will create maps showing where collectors have travelled, telling small stories about the scientific work and contribution of the people who helped create the biological collections. Users will also get feedback about the taxa they are transcribing utilizing taxon resolvers and displaying content such as images or narratives from EOL and Wikipedia in the Notes from Nature interface.

Incentives: Users will receive badges that are marks of accomplishment that can be kept on the Notes from Nature site and shared with others broadly via other social media sites. Distributing digital badges to represent new skills or achievements and thus promote learning and further engagement is a trend emerging in education fields (Goligoski 2012); however, rigorous studies demonstrating whether or not badges enhance citizen science motivation and learning have yet to be performed. Examples of badges in Notes from Nature may include "World Explorer" for those who complete transcriptions in a large number of countries, or "Bird Expert" for those who transcribe the top number of bird records.

Conclusion

The development of web-based citizen science endeavors stems from a long tradition of utilizing volunteers with a strong interest in the scientific subject matter (Cohn 2008). Such volunteer work has typically taken place locally at museums or other institutions, but the rise of the World Wide Web has provided a new, global platform for unpaid citizen efforts (Cravens 2000). Citizen science projects have taken many forms, the most well known among the biology community being outdoors-based reporting of species geographic distribution (e.g. iNaturalist, eBird; Sullivan et al. 2009) and phenology (e.g. Project Budburst; Meymaris et al. 2008). These projects are facilitated by the Internet, but have their roots in citizen volunteer efforts that, in cases like the Christmas Backyard Bird Count, stretch back more than a century.

A new category of citizen science leverages the Internet to disperse, transform, and reassemble information at unprecedented rates. These citizen science projects focus less on the creation of new scientific records, and more on the interpretation or enhancement of existing data sources and grow from a legacy of online volunteer transcription and proofreading started over a decade ago (See Distributed Proofreaders, http://www.pgdp.net/). Transcription of natural history collections records is a particularly strong fit for this new form of web-enabled citizen science, given the scope of the challenge, the scientific need for these data, and the inherently interesting subject matter. Other projects attempting similar outcomes are underway, including the Atlas of Living Australia Biodiversity Volunteer Portal and Herbaria@

home, but each of these vary from Notes from Nature in scope and the tools deployed. However, with existing projects in place and future projects being considered, a key question is whether the approach will capture the imagination of enough people to remain a reasonable, cost-effective and long-term solution to the challenge of transcribing as many as a billion objects.

Citizen Science on the web is in its infancy, and our knowledge about what works and why is still developing. The methods and product we are developing for Notes from Nature are helping to expand and build upon that knowledge. In particular, working within the Zooniverse offers experience with a legacy of technological tools, such as live-chat and reusable back-ends, a consistency across citizen science projects, and a strong focus on understanding and replicating successes while avoiding pitfalls. As importantly, the Zooniverse has generated a critical mass of volunteers and has established itself as a key member in the community creating citizen science projects. While initial citizen science applications in the Zooniverse focused on classifying and annotating anomalies across many astronomy images (e.g. Planet Hunters, http://www.planethunters.org), the roster of applications continues to grow. Old Weather (http://www.oldweather.org), for example, utilizes a simple transcription mechanism to collate temperature and other weather variables to determine past ocean climates. The project initially focused efforts on Royal Navy ship logs of the 20th century, but has since expanded to new sources of historic ship logs. The project, collaboratively developed by archivists, climate scientists, and citizen science experts has already transcribed over a million pages of such logs through engaging over 25,000 active volunteers since its start in 2010.

Notes from Nature is in many respects "experimental," and is still in its prototype phase. Many different enhancements will be tested, such as badges. Rewarding users is a complex topic in citizen science, as many considerations need to be made about how it could affect the quality and accuracy of data being collected. In Notes from Nature, the primary role of badges is to bring attention to particular work or achievements that can be made by volunteers in topics or datasets of interest. Ultimately, this will build into a Zooniverse-wide badge system, allowing users can collect badges from multiple domains of citizen science work. Badges will be an ongoing development in Notes from Nature, and the tool itself is expected to go through further iteration and refinement long after its initial full public release in August 2012.

The current focus of Notes from Nature is on accurate transcription of data exactly as it is recorded in the non-digital version. The first release will offer no opportunities for interpretation or annotation. We will continue to improve the transcription tool built for each of the data sources and add new interfaces for users, including tools for improving the quality of data and fitness for use. Examples to be developed in the near future include performing taxonomic and geographic "referencing". Taxonomic referencing would allow users to use services to check if names on labels are still valid, and if not, locate and provide an interpreted valid name (Thomer et al. 2012). Geographic referencing would provide means to convert textual locality descriptions into latitude, longitude, uncertainty triplets (Hill et al. 2009).

After Notes from Nature demonstrates that it works and is of wide interest, we hope grow our network of biocollections collaborators. We do so recognizing there is also a set of responsibilities to the community, including: 1) developing a reasonable and clear process for new biocollections to participate; 2) assuring that Notes From Nature does not overwhelm the community of citizen scientists with seemingly insurmountable tasks; 3) recognizing room for growth in this domain such that Notes From Nature can help address the needs of many citizen science transcription efforts. This challenge has been faced previously in Old Weather, where it is apparent that a much greater need for ledger transcription exists than was first thought. Our design architecture anticipates such growth, with Projects and Collections, built to facilitate local control of material coming from individual and partnering biocollections, and Missions, which target interests of citizen scientists and cut across any one project or collection.

Through Notes from Nature, we hope to team with citizen scientists to further widen the pipeline of digital biodiversity data for research. Both the application, and the new digitization it facilitates, may prove transformative for biological collections, citizen science and biodiversity science respectively. For biological collections and citizen scientists, we hope to bring new attention to those collections and the institutions that house them by connecting volunteers around the world to stories those data can tell. For biodiversity sciences, Notes from Nature will help unlock historical records that can help create and refine biodiversity baselines essential for documenting biodiversity change now and into the future.

References

Ariño AH (2010) Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7: 82–92. https://journals.ku.edu/index.php/jbi/article/viewArticle/3991

Bebber DP, Carine MA, Wood JRI, Wortley AH, Harris DJ, Prance GT, Davidse G, Paige J, Pennington TD, Robson NKB, Scotland RW (2010) Herbaria are a major frontier for species discovery. Proceedings of the National Academy of Sciences 107: 22169–22171. doi: 10.1073/pnas.1011841108

Berendsohn WG, Seltmann P (2010) Using geographical and taxonomic metadata to set priorities in specimen digitization. Biodiversity Informatics 7(2): 120–129. https://journals.ku.edu/index.php/jbi/article/viewArticle/3988

Biesmeijer J, Roberts S, Reemer M, Ohlemüller R, Edwards M, Peeters T, Schaffers A, Potts S, Kleukers R, Thomas C, Settele J, Kunin WE (2006) Parallel declines in pollinators and insect-pollinated plants in Britain and the Netherlands. Science 313: 351–354. doi: 10.1126/science.1127863

Boakes EH, McGowan PJK, Fuller RA, Chang-qing D, Clark NE, O'Connor K, Mace GM (2010) Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. PLoS Biol 8(6): e1000385. doi: 10.1371/journal.pbio.1000385

Canhos VP, Souza S, Giovanni R, Canhos DAL (2004) Global Biodiversity Informatics: setting the scene for a "new world" of ecological forecasting. Biodiversity Informatics 1: 1–13. https://journals.ku.edu/index.php/jbi/article/viewArticle/3

Cohn JP (2008) Citizen science: Can volunteers do real research? BioScience 58(3):192–197. doi: 10.1641/B580303

Cravens J (2000) Virtual volunteering: Online volunteers providing assistance to human service agencies. Journal of Technology in Human Services 17: 119–136. doi: 10.1300/J017v17n02_02

Edwards JL, Lane MA, Nielsen ES (2000) Interoperability of biodiversity databases: biodiversity information on every desktop. Science 289: 2312–2314. doi: 10.1126/science.289.5488.2312

Erb LP, Ray C, Guralnick R (2011) On the generality of a climate-mediated shift in the distribution of the American pika (Ochotona princeps). Ecology 92: 1730–1735. doi: 10.1890/11-0175.1

Giovanelli JGR, Haddad CFB, Alexandrino J (2008) Predicting the potential distribution of the alien invasive American bullfrog (Lithobates catesbeianus) in Brazil. Biological Invasions 10: 585–590. doi: 10.1007/s10530-007-9154-5

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19(9): 497–503. doi: 10.1016/j.tree.2004.07.006

Goligoski E (2012) Motivating the Learner: Mozilla's Open Badges Program. Access to Knowledge: A Course Journal 4(1) https://www.stanford.edu/group/opensource/cgi-bin/showcase/ojs/index.php?journal=AccessToKnowledge&page=article&op=viewArticle&path%5B%5D=217

Guralnick R, Hill A (2009) Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. Bioinformatics 25(4): 421–428. doi: 10.1093/bioinformatics/btn659

Hill AW, Guralnick RP, Flemons P, Beaman R, Wieczorek J, Ranipeta A, Chavan V, Remsen D (2009) Location, Location, Location: Utilizing pipelines and services to more effectively georeference the world's biodiversity data. BMC Bioinformatics. 10 (Suppl 14): S3. doi: 10.1186/1471-2105-10-S14-S3

Jenkins M (2003) Prospects for biodiversity. Science 302(5648): 1175–1177. doi: 10.1126/science.1088666

Lintott CJ, Schawinski K, Slosar A, Land K, Bamford S, Thomas D, Raddick MJ, Nichol RC, Szalay A, Andreescu D, Murray P, Vandenberg J (2008) Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society 389: 1179–1189. doi: 10.1111/j.1365-2966.2008.13689.x

Loreau M, Oteng-Yeboah A, Arroyo M, Babin D, Barbault R, Donoghue M, Gadgil M, Häuser C, Heip C, Larigauderie A, Ma K, Mace G, Mooney HA, Perrings C, Raven P, Sarukhan J, Schei P, Scholes RJ, Watson RT (2006) Diversity without representation. Nature 442: 245–246. doi: 10.1038/442245a

Lyons SK, Willig MR (2002) Species richness, latitude, and scale-sensitivity. Ecology 83(1): 47–58. doi: 10.1890/0012-9658(2002)083[0047:SRLASS]2.0.CO;2

Meymaris K, Henderson S, Alaback P, Havens K (2008) Project BudBurst: Citizen Science for All Seasons. AGU Fall Meeting Abstracts 1: 614.

Moffett A, Strutz S, Guda N, González C, Ferro MC, Sánchez-Cordero V, Sarkar S (2009) A global public database of disease vector and reservoir distributions. PLoS Neglected Tropical Diseases 3: e378. doi: 10.1371/journal.pntd.0000378

Moritz C, Patton JL, Conroy CJ, Parra JL, White GC, Beissinger SR (2008) Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. Science 322(5899): 261–264. doi: 10.1126/science.1163428

Nishida GM (2003) Museums and display collections. In: Resh V (Ed) Encyclopedia of insects. Academic Press, 768–775.

Nufio CR, McGuire CR, Bowers MD, Guralnick RP (2010) Grasshopper community response to climatic change: variation along an elevational gradient. PLoS ONE 5(9): e12977. doi: 10.1371/journal.pone.0012977

Ochoa-Ochoa L, Urbina-Cardona JN, Vázquez LB, Flores-Villela O, Bezaury-Creel J (2009) The effects of governmental protected areas and social initiatives for land protection on the conservation of Mexican amphibians. PLoS ONE 4(9): e6878. doi: 10.1371/journal.pone.0006878

Parmesan C, Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. Nature 421: 37–42. doi: 10.1038/nature01286

Pawar S, Koo MS, Kelley C, Ahmed MF, Chaudhuri S, Sarkar S (2007) Conservation assessment and prioritization of areas in Northeast India: priorities for amphibians and reptiles. Biological Conservation 136: 346–361. doi: 10.1016/j.biocon.2006.12.012

Pennisi E (2005) How did cooperative behavior evolve? Science 309(5731): 93. doi: 10.1126/science.309.5731.93

Peterson AT (2003) Predicting the geography of species' invasions via ecological niche modeling. Quarterly Review of Biology 78(4): 419–433. doi: 10.1086/378926

Peterson AD, Martínez-Meyer E (2009) Pervasive poleward shifts among North American bird species. Biodiversity 9: 14–16.

Pyke GH, Ehrlich PR (2010) Biological collections and ecological/environmental research: a review, some observations and a look to the future. Biological Reviews 85(2): 247–266. doi: 10.1111/j.1469-185X.2009.00098.x

Raddick MJ, Bracey G, Gay PL, Lintott CJ, Murray P, Schawinski K, Szalay AS, Vandenberg J (2010) Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. Astronomy Education Review 9(1): 010103. doi: 10.3847/AER2009036

Rainbow PS (2009) Marine biological collections in the 21st century. Zoologica Scripta 38(Suppl S1): 33–40. doi: 10.1111/j.1463-6409.2007.00313.x

Rödder D, Lötters S (2009) Niche shift versus niche conservatism? Climatic characteristics of the native and invasive ranges of the Mediterranean house gecko (Hemidactylus turcicus). Global Ecology and Biogeography 8(6): 674–687. doi: 10.1111/j.1466-8238.2009.00477.x

Soberón J, Peterson T (2004) Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359: 689–698. doi: 10.1098/rstb.2003.1439

Soto-Azat C, Clarke BT, Poynton JC, Cunningham AA (2010) Widespread historical presence of Batrachochytrium dendrobatidis in African pipid frogs. Diversity and Distributions 16(1): 126-131. doi: 10.1111/j.1472-4642.2009.00618.x

Sullivan BL, Wood CL, Iliff MJ, Bonney RE, Fink D, Kelling S (2009) eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation 142(10): 2282–2292. doi: 10.1016/j.biocon.2009.05.006

Thomer A, Vaidya G, Guralnick R, Bloom D, Russell L (2012) From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 235–253. doi: 10.3897/zookeys.209.3247

Vollmar A, Macklin JA, Ford L (2010) Natural history specimen digitization: challenges and concerns. Biodiversity Informatics 7: 93–112. https://journals.ku.edu/index.php/jbi/article/viewArticle/3992

Wake DB, Vredenburg VT (2008) Are we in the midst of the sixth mass extinction? A view from the world of amphibians. Proceedings of the National Academy of Sciences 105 (Suppl 1): 11466. doi: 10.1073/pnas.0801921105

Walther GR, Post E, Convey P, Menzel A, Parmesan C, Beebee TJC, Fromentin JM, Hoegh-Guldberg O, Bairlein F (2002) Ecological responses to recent climate change. Nature 416: 389–395. doi: 10.1038/416389a

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An evolving community-developed biodiversity data standard. PLoS ONE 7(1): e29715. doi: 10.1371/journal.pone.0029715

ZooKeys 209: 235–253 (2012)
doi: 10.3897/zookeys.209.3247
www.zookeys.org

RESEARCH ARTICLE

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

From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks

Andrea Thomer[1], Gaurav Vaidya[2], Robert Guralnick[2], David Bloom[3], Laura Russell[4]

1 *University of Illinois, Urbana-Champaign, Graduate School of Library and Information Science, 501 E. Daniel Street, Champaign, Illinois, 61820, USA* **2** *University of Colorado, Boulder; University of Colorado Museum of Natural History, Henderson Building, Boulder, Colorado, 80309, USA* **3** *University of California, Berkeley, Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building, Berkeley, California, 94705, USA* **4** *University of Kansas, KU Biodiversity Institute, 1345 Jayhawk Blvd., Room 606, Lawrence, Kansas, 66045, USA*

Corresponding author: *David Bloom* (dabblepop@gmail.com)

Academic editor: *Vladimir Blagoderov* | Received 18 April 2011 | Accepted 12 July 2012 | Published 20 July 2012

Citation: Thomer A, Vaidya G, Guralnick R, Bloom D, Russell L (2012) From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 235–253. doi: 10.3897/zookeys.209.3247

Abstract

Part diary, part scientific record, biological field notebooks often contain details necessary to understanding the location and environmental conditions existent during collecting events. Despite their clear value for (and recent use in) global change studies, the text-mining outputs from field notebooks have been idiosyncratic to specific research projects, and impossible to discover or re-use. Best practices and workflows for digitization, transcription, extraction, and integration with other sources are nascent or non-existent. In this paper, we demonstrate a workflow to generate structured outputs while also maintaining links to the original texts. The first step in this workflow was to place already digitized and transcribed field notebooks from the University of Colorado Museum of Natural History founder, Junius Henderson, on Wikisource, an open text transcription platform. Next, we created Wikisource templates to document places, dates, and taxa to facilitate annotation and wiki-linking. We then requested help from the public, through social media tools, to take advantage of volunteer efforts and energy. After three notebooks were fully annotated, content was converted into XML and annotations were extracted and cross-walked into Darwin Core compliant record sets. Finally, these recordsets were vetted, to provide valid taxon names,

via a process we call "taxonomic referencing." The result is identification and mobilization of 1,068 observations from three of Henderson's thirteen notebooks and a publishable Darwin Core record set for use in other analyses. Although challenges remain, this work demonstrates a feasible approach to unlock observations from field notebooks that enhances their discovery and interoperability without losing the narrative context from which those observations are drawn.

Keywords

Field notes, notebooks, crowd sourcing, digitization, biodiversity, transcription, text-mining, Darwin Core, Junius Henderson, annotation, taxonomic referencing, natural history, Wikisource, Colorado, species occurrence records

"Compose your notes as if you were writing a letter to someone a century in the future."
Perrine and Patton (2011)

Introduction

Our species has analyzed and documented the natural world for millennia, in media as diverse as Paleolithic cave paintings, handwritten field notes, and structured databases of sequences sampled from the environment. While structured data facilitate long-term ecological monitoring, the "first-person precision" (Grinnell 1912) of an idiosyncratic, unatomizable narrative about nature — be it a drawing on a cave wall or a handwritten page in a field journal — gives these data context that does not readily fit into a spreadsheet, and which may form the nucleus of an important new insight or discovery. Field notes in particular sit at the crossroads of these qualitative and quantitative methods; in them, structured and unstructured data are necessarily intertwined (Kramer 2011).

The observations contained in field notebooks take on particular importance given the current biodiversity crisis (Jenkins 2003, Heywood and Watson 1995, Loreau et al. 2006, Wake and Vredenburg 2008) — a crisis which threatens the fabric of ecosystems on which our own species depends (e.g. Millennium Ecosystem Assessment 2005, Worm et al. 2006). Legacy occurrence records extracted from field notebooks provide essential baselines of past community biotic state for resurvey efforts such as the Grinnell Resurvey Project (Moritz et al. 2008, Tingley et al. 2009) and the Alexander Grasshopper Project (Nufio et al. 2010).

The growing use of such records for global change biology creates new challenges and opportunities for their digitization, transcription, representation, and integration with other sources of historical data. All these challenges ultimately depend on pulling structured data from unstructured text, while somehow maintaining a link to the original texts. Solving these challenges is key to realizing their value in research and policy-making.

Here we present a case study that makes occurrence records in field notebooks available by utilizing something of a rarity in this arena: a fully scanned and tran-

scribed set of field notebooks, penned by University of Colorado Museum of Natural History founder Junius Henderson (http://en.wikisource.org/wiki/Field_Notes_of_Junius_Henderson). We provide a pragmatic approach for utilizing free, relatively easy-to-use technologies to annotate these notes, and discuss some of the remaining gaps in our toolkits and cyberinfrastructure. We also present a workflow for extracting occurrence records from field notebooks that requires minimal resources (beyond the authors' time), fosters community involvement, and abstracts the necessary information while maintaining links to its original text, thereby preserving the context that only "first-person precision" can provide. The primary challenges we address are how to: 1) publish these field notes in a way that supports annotation of species occurrence records; 2) extract these records efficiently; 3) convert these records to the most interoperable format; and, 4) store these records and maintain their link to the original field notes.

Background

Remsen et al. (2012) identified conversion of unstructured text into structured data as a key challenge in biodiversity informatics, and showed a working methodology for creating a Darwin Core archive from a conventional floristic checklist. We follow the path laid by those authors, but focus on mining observations from field notebooks. Field notebooks are often "hidden" in archives of institutions, and unlike formally published sources, typically lack a centralized access point (Sheffield et al. 2011), a standardized mark-up language, and any sort of reliable or scalable method of mining content from the notes. Sheffield and Nakasone (2011) from the Smithsonian's *Field Book Project* present an excellent high-level view of how existing metadata standards could be used to semantically link collections and field notes. This collections-level schema, however, does not address the need to annotate and extract data from documents. Furthermore, though work has been done linking digital collections to Wikipedia articles (e.g., Lally and Dunford 2007), and though the National Archives have recently partnered with Wikisource to upload their materials for transcription (http://transcribe.archives.gov/), neither of these projects have attempted to annotate or extract data from the materials.

In light of this lack of prior work, and given the observational nature of the notes, we decided that these observations would be best published as Darwin Core records. Though there are other standards used in the digital humanities to mark up scholarly texts (e.g. the Text Encoding Initiative's standard, http://www.tei-c.org/), none of these are tailored for the encoding of biodiversity data. Darwin Core, on the other hand, is a commonly used metadata schema for describing and exchanging a range of biodiversity data, from museum specimen records to field observations (Wieczorek et al. 2012). In particular, the Global Biodiversity Information Facility (GBIF) uses it for storage, transfer and presentation of biodiversity data.

The study corpus: Junius Henderson's field notes

Junius Henderson was appointed the first curator of the University of Colorado Museum of Natural History (CU Museum) in 1902. He kept handwritten field notebooks describing his expeditions across the Southern Rocky Mountains and elsewhere over a 26-year period. Henderson completed 13 notebooks and 1,672 pages of entries, augmented by other materials such as photographs and a locality ledger. Henderson's notes are arranged as entries (Figure 1), which usually contain some kind of header denoting date and place. All entries are separated by a blank space, so even if header text is not strictly standardized, the beginning and end of each entry is quite clear. Although Henderson did keep a locality ledger, he did not directly or systematically reference specimens to field note entries. Thus, if there are direct links between collected specimens and field notes, they have yet to be discovered.

Henderson's notebooks are a chronicle of the American West in transition and paint a vivid picture of a changing landscape as cities expand, wild places retreat, and horse-and-buggies give way to cars. His journal entries describe everything from mollusks in freshwater and marine systems, to the geology of the Rocky Mountains, to the more mundane aspects of fieldwork (e.g., "Train again so late as to afford ample opportunity for philosophic meditation upon the motives which inspire railroad people to advertise time which they do not expect to make except under rare circumstances,") (Henderson 1907).

From February 2000–02, former CU Museum Director and Curator Peter Robinson transcribed all thirteen volumes of Henderson's notes into Word documents — a herculean task given Henderson's handwriting. In 2006, the National Snow and Ice Data Center (NSIDC) scanned Henderson's thirteen notebooks for a large glaciology project. Through a lengthy series of events, documented more fully in a series of blog posts (http://bit.ly/jhfnblog), the scans and transcriptions, separated from each other for several years, were reunited once we began work on this project.

The existence of both scanned images and typed transcriptions made Henderson's notes an excellent test case for annotation and automated occurrence extraction; transcriptions could be tagged and annotated via a markup schema, and checked against scanned images of the original pages to ensure accuracy. As of this writing, only the first three notebooks have been annotated.

Methods

We documented this project using a blog as an open notebook and a means to communicate our goals, ideas, and progress. Those goals were: (a) to make Henderson's notes easily discoverable, publicly accessible, freely reusable and sustainably preserved and, and (b) to extract taxonomic occurrences from these notes.

A platform for field notebook access and annotation: Wikisource

We quickly realized we needed a way to support the annotation of species occurrences on an open platform so that anyone interested could help with the task. We decided on the Wikipedia-related project Wikisource (http://wikisource.org) for the following reasons:

Ease of use. The process of uploading scanned pages is simple. PDFs are uploaded to the Wikimedia Commons and pulled into Wikisource. Once in Wikisource, hyperlinked index pages can be created and transcribed text can be matched with the scanned image of each field book page (Figure 1). The wiki markup language is similarly easy to learn and use. The language is the same as that used in Wikipedia, which means skills developed in Wikipedia can be brought to Wikisource easily.

Completely open access. Everything on Wikisource can be edited by anyone, giving us a way to crowdsource annotation to citizen scientists and archivists. All Wikisource pages have a built-in means of tracking edits that ensure that all changes made to the transcriptions are documented and reversible.

An existing community of developers. Wikisource uses the same software as Wikipedia (a PHP application named "MediaWiki"), which is under active development by a core team of developers. Sharing the same software and licensing terms means that content can be shared between the two projects freely. Additionally, pages designed to be incorporated into other pages (known as *templates* in Wikispeak; see http://en.wikipedia.org/wiki/

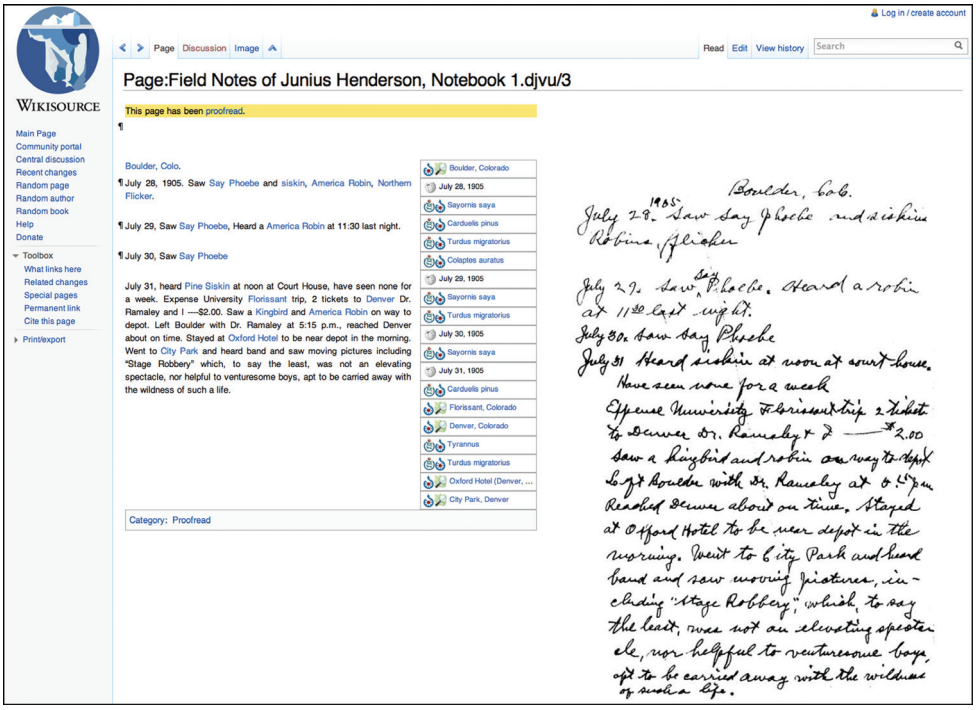


Figure 1. Web browser view of a scanned page of Henderson's journal displayed side-by-side with transcriptions and annotations using the MediaWiki *Proofread Page* extension.

Template:Cleanup for an example) can be moved from one project to another easily, speeding development. The Wikipedia community also carries out software development for Wikisource-specific features; our project relied on the *Proofread Page* extension to provide side-by-side views of transcriptions and their corresponding scanned images (Figure 1).

An existing community of users, transcribers, and proofreaders. There is an active Wikisource community improving Wikisource's content and to transcribing newly uploaded texts (see http://en.wikisource.org/wiki/Wikisource:Community_collaboration). We hoped to draw some of these community members into our project.

Uploading content

The ideal upload to Wikisource is a Portable Document Format (PDF) or DjVu multipage image file containing the entire scanned document along with its OCRed text (sometimes referred to as a "searchable PDF"). Such files retain their text in Wikisource, making transcription easy. In our case, we uploaded handwritten scans as-is and inserted the transcriptions manually. PDF or DjVu files are uploaded to the Wikimedia Commons using the Upload Wizard (http://bit.ly/wcupload) and reused in Wikisource. One important note: both the Wikimedia Commons and Wikisource only allow the upload of materials in the public domain or published under liberal open source licenses (such as the Creative Commons Attribution or Creative Commons Attribution-ShareAlike licenses). Materials that have only been made available for non-commercial use may not be uploaded to the Wikimedia Commons. This means that data from the Biodiversity Heritage Library, which uses a Creative Commons Non-Commercial Share-Alike license, could not be uploaded to Wikisource. For a thorough discussion of the effect of these licenses on biodiversity science, see Hagedorn et al. (2011).

While uploading images to the Commons is simple, reusing them in Wikisource can be tricky (a guide to this process — updated by us — is available on Wikisource: http://bit.ly/wsindexhelp). After setting up the Index page (Figure 2) and copying the transcriptions into Wikisource manually, we were ready to begin annotation.

Creating annotation templates

In Wikisource, annotations are best made through the use of templates. Templates are a feature of the MediaWiki software that allows one wiki page to be inserted into another. While usually used to embed common design elements across Wikipedia (such as the *Unbalanced* template, used to warn readers that an article might be unbalanced: http://en.wikipedia.org/wiki/Template:Unbalanced), they can also provide complex functionality, such as creating a standardized citation format (see http://en.wikipedia.org/wiki/Template:Cite_journal) or calculating ages from birthdates. We developed our own templates to not only tag the elements of an occurrence record but also create links to other web resources.

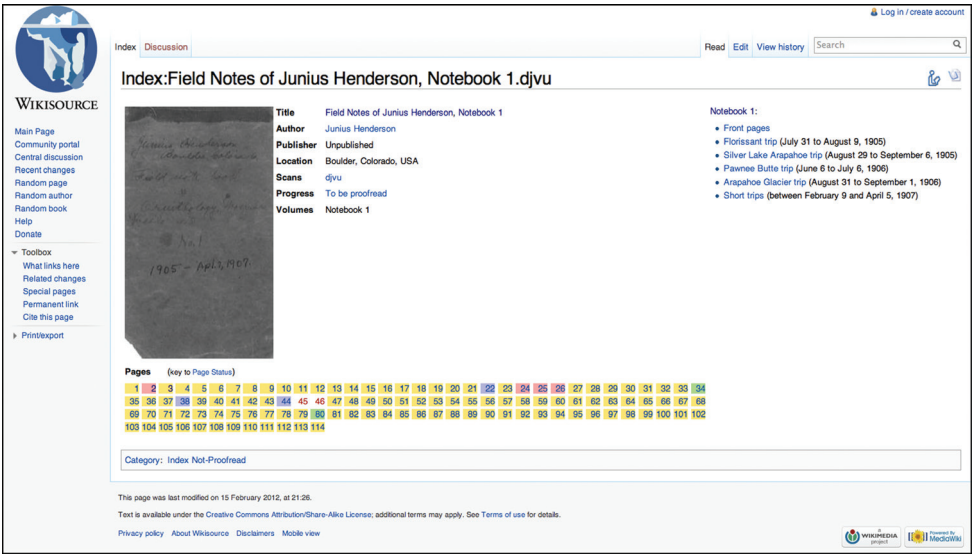


Figure 2. Index page for Notebook #1. Each Index page corresponds to a multipage file. The Index page displays volume metadata and links to sections of the notebook, while also providing links out to each notebook page and color-coding to determine which pages have been already transcribed and proofed.

The elements of an occurrence record

A species occurrence record should contain the following basic elements in order to be fit-for-use in biodiversity science: 1) the species' name, and 2) the place and 3) time in which it was observed. Also important, but slightly less crucial, is additional information describing the observation event: the name of the person making the observation, any equipment used, the sampling method, and so on.

Thus, because our goal was the extraction of occurrence records, we created annotation templates for *taxa*, *locations*, and *dates*. A triplet of all three annotations would, in theory, be attributable to an observation event and could be pulled from the annotated text as an occurrence record. The templates link these elements to Wikipedia pages, and provide a means to show annotations separately from the text itself.

The first sentence of Henderson's first field book contains a simple example of the type of text we hoped to annotate with Wiki markup (Figure 3):

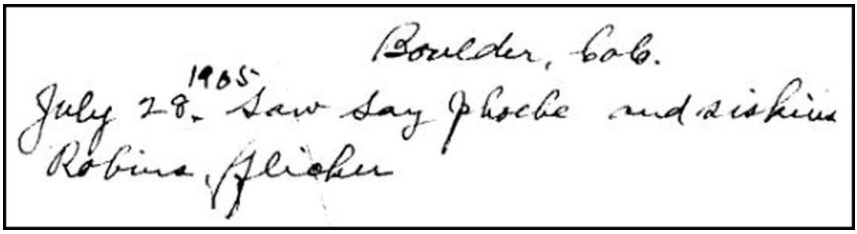


Figure 3. Henderson's first sentence. "Boulder, Colo. July 28, 1905. Saw Say [sic] Phoebe and siskins, [American] Robins, [Northern] Flicker."

This single sentence contains six annotatable terms: a *location* (Boulder, Colo), a *date* (July 28, 1905), and four *taxa* (Say['s] Phoebe, Pine Siskin, American Robin, Northern Flicker). Each template attempts to link the annotated element to associated pages in the Wikimedia Commons and Wikipedia. Thus, templates include the verbatim text from Henderson and an interpretation of that element's formal name (as determined by the annotator) that resolves to other Wiki-resources. The general syntax of these templates is:

{{*element*|f*ormal name of this element*|*element as written by Henderson*}}

For example, the first taxon annotation in the text reads:

{{*taxon*|*Sayornis saya*|*Say Phoebe*}}

While the process of creating these annotations is relatively simple, we soon discovered that each requires substantial decision making on the part of the annotator, leaving ample room for variation.

In the case of the "Siskin" above, annotators could make several interpretations. An experienced birder may reason that based on Henderson's location at that time, he is referring to a Pine Siskin (*Carduelis pinus)* and create the following annotation:

{{taxon|Carduelis pinus|siskins}}

But it's just as likely that a less experienced annotator would create the following less specific, though technically correct, annotation:

{{taxon|Siskin|siskins}}

This latter annotation links to a Wikipedia disambiguation page listing 18 different bird species, a kind of British aircraft, and a Canadian junior ice hockey team (http://en.wikipedia.org/wiki/Siskin).

We allowed our annotators complete flexibility in interpreting vernacular names as they saw fit while editing notebook pages (Figure 4); this meant that we had to review and resolve taxonomic annotations to a best valid taxon name, just as a lab supervisor would need to check a volunteer's work in a museum. In future work, we will take steps to prescribe best practices based on what we learned in this pilot project.

The full process of determining a valid scientific name from Henderson's verbatim description is *taxonomic referencing*, analogous to georeferencing for localities. As with georeferencing, there is uncertainty in the process of linking legacy observations to current valid names; the level of uncertainty depends on who did the referencing and when. We discuss our approach to taxonomic referencing below.

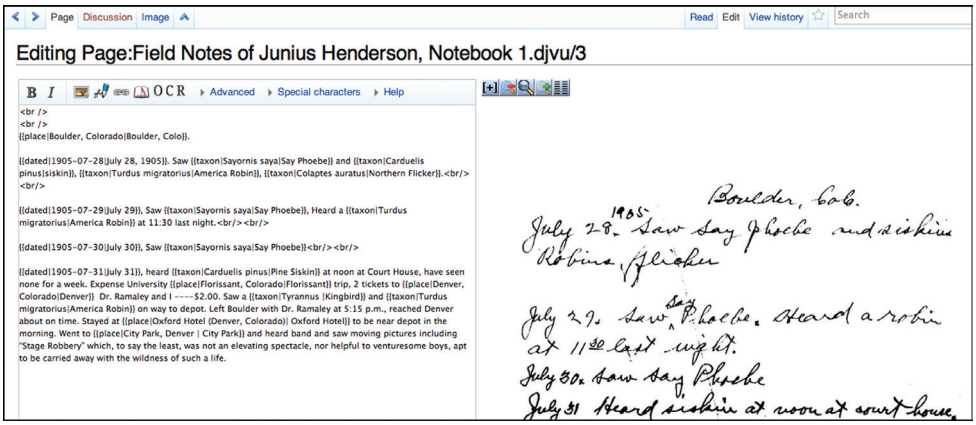


Figure 4. Editing a notebook page on Wikisource. This screenshot shows side-by-side transcription and wiki markup syntax.

Data extraction: Seeking efficiency and accuracy

The annotated text from Henderson's first three notebooks was downloaded using the MediaWiki API (http://bit.ly/mediawikiapi). Individual annotations were then identified using regular expressions. We have described this process in detail in supplementary file 1: "Methods Supplement_Henderson.pdf." The Perl module and scripts used for this process are available at https://github.com/gaurav/henderson.

In summary, the steps were to:

1) Retrieve the number of pages in the file; 2) Extract the wiki markup from each individual page; 3) Write the wiki markup to a single XML file, which was divided into individual pages; 4) Concatenate this page-by-page file into one single text file to account for entries split across pages (Figure 5); 5) Divide the file into entries rather than pages; and 6) walk through the file, keeping track of the last location and date annotation encountered. Each taxon in an entry, coupled with the entry date and the preceding location, was tagged as an occurrence. Each triplet of elements that made up the occurrence was written to a CSV file, along with some text from the entry itself, the page number in the notebook, and a permanent link to the version of the Wikisource page containing the entry at the time the XML file was downloaded.

Converting records into interoperable formats

After pulling occurrences into a CSV, we cross-walked this data into several fields selected from the Darwin Core Standard and added whatever supplementary information we could (e.g. by extrapolating higher taxonomy; see Appendix 1). Content in most fields depended on the four variables extracted from our dataset (taxon, date, location, page number), though some content was fixed (e.g., recordedBy always read

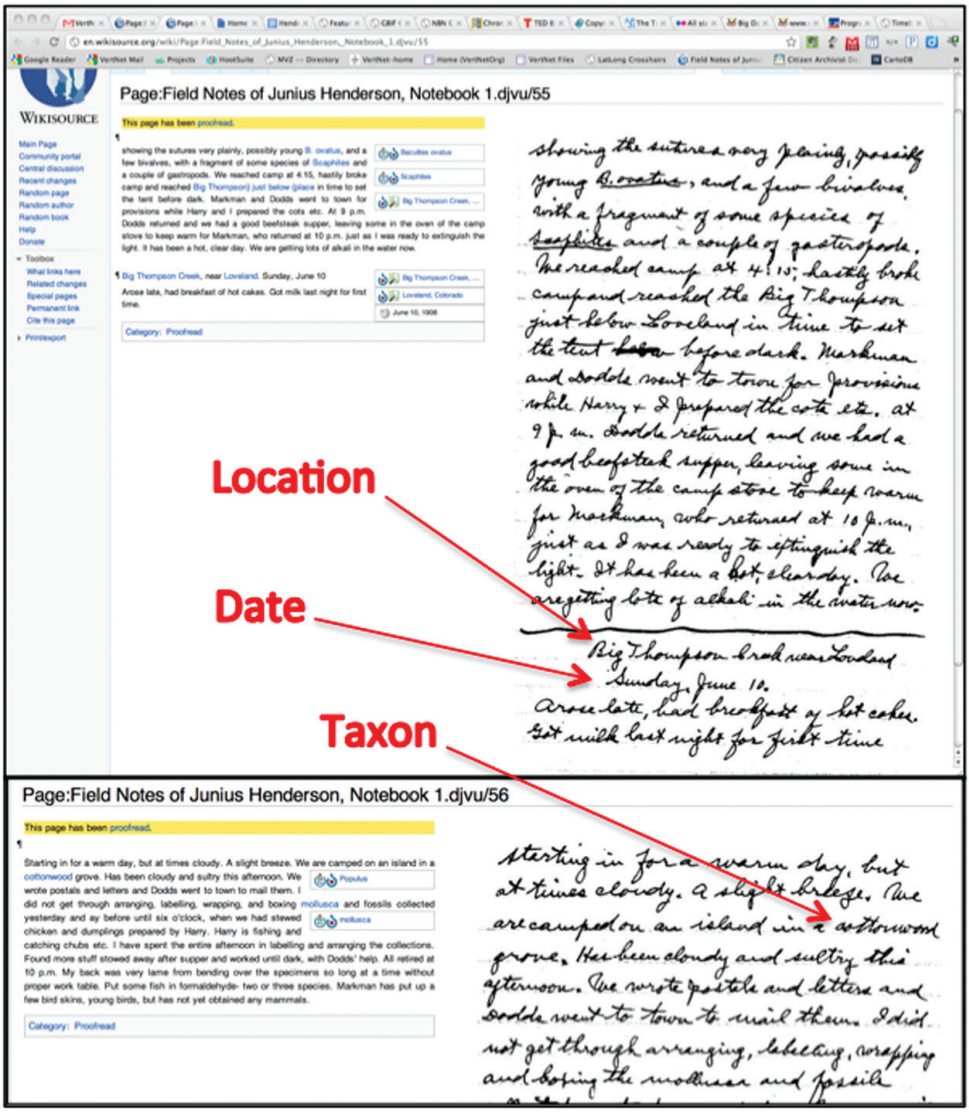


Figure 5. An example of how a location (Big Thompson Creek near Loveland), a date (Sunday, June 10, 1906), and a taxon (Cottonwood, genus Populus) are grouped from across multiple pages.

"Junius Henderson"), and other content required manual determination or validation before being entered.

Proofing the Darwin Core record set

The process of extracting taxon-location-date triplets is imperfect and requires vetting by proofreaders to ensure accuracy of the automated process, which does not consider

contextual data. For example, our automated extraction scripts would incorrectly assume the following passage refers to a presence, not an absence: "Am perplexed by the entire absence of robins on this trip" (http://bit.ly/jhfn1-43). In future work, we plan to alter our templates to give annotators the ability to record whether an observation marks a presence or absence of a taxon.

As mentioned above, taxonomic names need special vetting, too. Henderson freely mixed vernacular and scientific names in his notes, and annotators consequently did as well. We performed taxonomic referencing using Google Refine, Encyclopedia of Life (EOL), and Integrated Taxonomic Information System (ITIS) name resolvers, following instructions from an *iPhylo* blog post by Rod Page (http://iphylo.blogspot.com/2012/02/using-google-refine-and-taxonomic.html). First, we loaded our CSV files from each field notebook into Google Refine. We then reconciled names assigned by annotators against the ITIS Freebase namespace (integrated within Google Refine) and the EOL service (developed by Page), and accepted the best judgments (as determined by probability scores). Those *best names* from each service were placed into two separate columns for further expert validation. The rows that produced consistent results from both EOL and ITIS name services were considered correct after a quick check for accuracy. One of the authors (Vaidya) checked each record in which EOL and ITIS suggested different best names and either chose the EOL name, the ITIS name, both, or neither. In many cases, one service provided a clear best fit at the right taxonomic depth compared to the other. In cases where both provided poor results, we did not choose a name. On those records where ITIS was found to be the best fit, we used the ITIS Taxonomic Serial Number to populate the vernacularName and the higher taxonomy fields. We also recorded the taxonomic resolution service used (EOL, ITIS, or EOL & ITIS) in the identificationRemarks field of the Darwin Core file we produced.

We also checked for annotation errors directly on Wikisource. One of the authors (Guralnick) went through each page of Notebook 1 on Wikisource to check for any obvious problems, such as poor formatting, mislabeling, or missed annotations (e.g., dates, locations, or taxa that could have been annotated but were not). He also checked all three notebooks for annotations that noted absences or that otherwise were not obviously observations.

Data archiving and maintaining links to the original notes

All generated Darwin Core occurrence records include a URL to the page in Wikisource from which they are drawn in the Source field, i.e., they will take you to the version of the page that was live at the time at which the original XML file was created, not the latest version of the file. Additionally, each record is assigned an automatically generated catalog number as the record is extracted from the notebook.

Data resources

The data presented in this paper are available for download in a Darwin Core Archive via VertNet, http://ipt.vertnet.org:8080/ipt/resource.do?r=hendersonnotebooks1-3. The archive includes taxon occurrences extracted from the field notes of Junius Henderson as he traveled through Colorado and the western United States.

Results

After advertising our project via the blog, Twitter, and emails to relevant listservs, a total of three notebooks were transcribed and annotated, largely by volunteers (Table 1): 352 pages of notes and 222 entries in all. As of March 27, 2012, 10 registered Wikisource users and 11 anonymous users helped annotate these notebooks. All three notebooks were annotated within four to six weeks each. Again, only three of Henderson's thirteen notebooks were uploaded for the purposes of this pilot project; we hope to upload and annotate the remaining notebooks soon.

A total of 1,087 taxon annotations were created across all three books, with each entry having between zero and 33 taxon annotations. Taxonomic resolution led to 560 records that were identified as valid by both EOL and ITIS taxonomic name resolvers. Expert validation led to 195 records as judged to be matched better by EOL than ITIS, and 83 records wherein the ITIS match was preferable to EOL's. A total of 238 records could not be validated by either EOL or ITIS.

In Notebook 1, only two of 634 annotations were poorly formatted, caused by missing brackets. Only one date was transcribed incorrectly: "Apl 5/07" was annotated incorrectly as "April 7, 1907" (http://bit.ly/enws3614593). Also in Notebook 1, ten places and taxa could have been annotated but were not, and in all cases these were very broad taxonomic groups (e.g., Crustacea). A total of eleven taxon annotations across all three notebooks were manually identified as not denoting presence, and removed from the

Table 1. Summary information on each notebook.

	Notebook 1	Notebook 2	Notebook 3
URL	http://bit.ly/jhfn1-indexpg	http://bit.ly/jhfn2-indexpg	http://bit.ly/jhfn3-indexpg
Number of annotations	632	703	1007
Taxon annotations	349 (201 unique)	224 (125 unique)	514 (248 unique)
Place annotations	219 (115 unique)	419 (154 unique)	401 (139 unique)
Date annotations	64 (63 unique)	60 (59 unique)	92 (90 unique)
Dates in range	July 1905 to April 1907	May 1907 to October 1908	January 1909 to September 1909
Time spent annotating	6 weeks	4 weeks	6 weeks

final dataset. Overall, the error rates and false positives were very low. After eliminating records of absence and some incorrect annotations, 1,068 valid observations remained; these were exported to a final Darwin Core Archive is included in the supplemental materials of this paper (see supplemental file 2: "dwca-hendersonnotebooks1-3.zip").

Discussion

Wikisource as a medium for open provisioning and annotation of field notebooks

Our work is part of a larger set of efforts to transcribe, and ultimately mine, the extensive library of historical biodiversity literature (Gwinn and Rinaldo 2009). The choice to use Wikisource for provisioning and annotation of field notes well served our needs, but we recognize the tremendous efforts made by developers to build their own platforms for notebook and journal transcription projects, especially *From The Page* (http://beta.fromthepage.com/), which is being used to transcribe the field notes of renowned herpetologist Lawrence Klauber, of the San Diego Zoo (http://bit.ly/fromthepage-lmk). The primary benefit that *From The Page* offers over Wikisource is that of customization. In the Klauber interface, for instance, developers were able to add a sidebar listing of Klauber's "slang": the common names he used to refer to animals in lieu of their scientific names. This could potentially be a great help to volunteer annotators, but is not currently supported by the Wikisource interface.

Wikisource is a relatively new part of the Wikimedia world, and continues to grow to accommodate new uses, as our project demonstrates. The annotation mechanisms we developed were new to Wikisource and pushed the bounds of accepted community practice, especially the relatively obtrusive "link-out boxes" that are placed inline with the text. While there have been some community discussions about the best way of visualizing annotations on Wikisource (e.g., http://bit.ly/N7woun), there has been no major opposition to our templates as yet. We also created community resources to encourage the use of our templates by other notebook annotation projects in the future (see http://en.wikisource.org/wiki/Wikisource:WikiProject_Field_Notes), but, as of this writing, we remain the only field notebook project on Wikisource.

We were able to speedily annotate three notebooks because our crowdsourcing approach worked as well as, or better, than expected, albeit in unexpected ways. Though we attempted to motivate volunteer efforts by promising acknowledgement in this paper and offering a free coffee mug featuring one of Henderson's field photos in exchange for service, such incentives were ineffective. Instead, two hard-working, anonymous users, known only by IP addresses, completed the majority of annotations. This may indicate that there are motivating factors beyond reward and acknowledgement that spur people to volunteer for these projects.

It is an open question whether using Wikisource fostered or limited participation. There is a learning curve when using Wikimedia products — not just one of learning a new technology, but also of learning the social mores of the existing wiki-community.

Potential volunteers and digitization project managers alike may be put off by both barriers to entry, relatively low though they are. On the technology side, we found the Wikisource GUI to be simple and effective, but not always intuitive. For example, despite good help guides, it took some members of our team (who shall remain unnamed) over a month to discover forward and back arrows that allow navigation between sequential notebook pages without returning to the Index. On the social side, posting to the "talk" pages to discuss new policies or initiatives requires learning new ways of communicating with, and integrating into, an online community, which takes time and emotional energy. We wonder if annotator anonymity reflects a desire to avoid entanglement in this community, and simply do a task that is enjoyable.

Challenges storing and extracting and converting records into interoperable formats

Though Wikisource *can* function as a repository of sorts, it is unclear whether the Wikimedia Foundation *wishes* for it to function as the primary home for digital manifestations of primary source documents. Because there is little easily found documentation describing its long-term digital preservation plans or strategies, we hesitate to call Wikisource a repository. The Wikimedia Foundation may wish to be more deliberate and less opaque in communicating these strategies, especially if it wishes to encourage continued annotation work. Clear digital preservation policies could better assure Wikipedians of their contributions' relative permanence – whether document uploads, transcriptions or annotations.

We also faced challenges when attempting to capture our workflow in the same structured format as the occurrence records we were extracting: that is, we had more data than we could "fit" into Darwin Core fields. Our solution was to create two sets of files: one composed of simple Darwin Core terms (see supplemental file 2: "dwca-hendersonnotebooks1-3.zip"), and another with a richer set of provenance data showing the process of taxonomic referencing and data processing (see supplemental file 3: "HendersonDwCfull.csv"). This allowed us to present a simple, interoperable dataset while still preserving a record of the densely idiosyncratic process unique to our project and workflow for the purposes of this paper. However, proliferating slightly different versions of this recordset could ultimately cause more confusion than clarity.

Darwin Core's limited expressivity became especially evident when performing taxonomic referencing; the lack of best practices and vocabularies for describing this multistep process is a notable gap in biodiversity informatics workflows. We particularly note the lack of a VerbatimName term in Darwin Core. Introducing VerbatimName would provide the means to capture the original string as expressed in an occurrence record or field notebook as a starting point to tracking that taxonomic referencing process. Just as VerbatimLocality and GeoreferencingMethod are recorded for future rein-

terpretation, new terms such as VerbatimIdentification and TaxonResolutionMethod could provide the means to capture essential processing steps as well.

The problems we faced using name resolution services were typical of attempts to automatically extract and parse taxonomic names, thus underscoring the need to better support taxonomic referencing workflows. Though both ITIS and EOL name resolution services returned a substantial number of matches to our names, human validation showed that these resolvers often performed mysteriously, sometimes providing well-resolved binomials when only a genus was entered, or resolving vernacular names in unexpected ways. EOL, for instance, consistently mapped "mouse" to *Amphipyra tragopoginis*, the Mouse Moth. Homonyms across different kingdoms further complicated matters, such as *Crucibulum*, which may be a genus of gastropod or of fungi.

Challenges with data storage and lasting linkages to sources

Field notebook data and specimen records are often recorded in the field, at the same time, but need to be reconnected after the fact. It is unclear which of Henderson's observations resulted in collecting events, but re-associating data from these different sources will help enrich local knowledge of biodiversity. A next step will be comparing and contrasting University of Colorado Museum of Natural History zoological specimen catalogs with field notebook observation datasets, both now represented in Darwin Core files. One simple approach is to search on date, and compile taxonomic matches between notebook observations and specimen records. Also of great value will be georeferencing field notebook records to further simplify direct comparisons with other contemporaneous species occurrence records.

We close by noting a final and perhaps most vexing challenge: keeping field note annotations on Wikisource synchronized with the extracted occurrence records. During the occurrence extraction process, we assigned catalog numbers to each occurrence. However, we do not presently have a workflow to then annotate Wikisource with these numbers. Because Wikisource is a necessarily live platform, there is a possibility that additional occurrences will be found and annotated after our initial extraction. Our script, as it is written, would re-catalog these occurrences from the top of the page to the bottom; in short, our catalog numbers are neither stable, nor permanent nor globally unique. This will be hugely problematic if our workflow is implemented in other projects with longer time horizons. In the future, we either need to find a way to annotate occurrences in Wikisource with unique identifiers, or edit our script and cataloging process to remember what we have or have not counted as an occurrence. Although excellent versioning in Wikisource and inclusion of some content from the notebooks in the final CSV files may allow checks for old and new entries, the more stable and reliable solution is to amend the script to automatically annotate references to taxa in Wikisource with such identifiers.

Acknowledgements

Many thanks to Allaina Wallace and Ruth Duerr for their crucial contribution of the original scans of Henderson field notebooks. Thanks especially to Ben Brumfield for thoughtful discussions and support through his excellent blog *Collaborative Manuscript Transcription* (http://manuscripttranscription.blogspot.com/). We also appreciate the support and encouragement of Rusty Russell and Carolyn Sheffield of the Smithsonian Field Book Project (http://www.mnh.si.edu/rc/fieldbooks/). Many thanks to Paul Flemons and other commenters on our So You Think You Can Digitize blog (http://soyouthinkyoucandigitize.wordpress.com); their feedback has been instrumental in moving us forward. Elizabeth Merritt and the Center for the Future of Museums helped us reach a much broader audience via the CFM blog. Although we don't know the names of our mystery annotators, we want to especially thank them for doing a tremendous amount of painstaking work with high accuracy. We only know a pseudonym for one presumed annotator, "R. U. Testing Me." Nic Weber reviewed parts of the manuscript and provided useful feedback. Finally, we appreciate the opportunity to be included in this special issue and want to thank the reviewers for their excellent comments, and Vladimir Blagoderov for his patience and support.

References

Grinnell J (1912) An Afternoon's Field Notes. The Condor 14(3): 104–107. Retrieved from http://www.jstor.org/stable/1362226

Gwinn N, Rinaldo C (2009) The Biodiversity Heritage Library: Sharing biodiversity literature with the world. IFLA Journal 35(1): 25–34.

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

Heywood VH, Watson RT (1995) Global Biodiversity Assessment. Cambridge University Press, Cambridge.

Henderson J (1907) Notebook 1. Unpublished; available online at http://commons.wikimedia.org/w/index.php?title=File:Field_Notes_of_Junius_Henderson,_Notebook_1.djvu&oldid=68218270 and http://en.wikisource.org/w/index.php?title=Field_Notes_of_Junius_Henderson/Notebook_1&oldid=3721123

Jenkins M (2003) Prospects in biodiversity. Science 302: 1175–1177.

Kramer KL (2011) The spoken and the unspoken. In: Canfield MR (Ed) Field Notes on Science & Nature. Harvard University Press, Cambridge, Massachusetts.

Lally AM, Dunford C (2007) Using Wikipedia to Extend Digital Collections. D-Lib Magazine, 13(5/6). Retrieved from http://www.dlib.org/dlib/may07/lally/05lally.html

Loreau M, Oteng-Yeboah A, Arroyo MTK, Babin D, Barbault R, Donoghue M, Gadgil M, Häuser C, Heip C, Larigauderie A, Ma K, Mace G, Mooney HA, Perrings C, Raven P,

Sarukhan J, Schei P, Scholes RJ, Watson RT (2006) Diversity without representation. Nature 442: 245–246.

Millennium Ecosystem Assessment (2005) Ecosystems and Human Well-being: Synthesis. Island Press, Washington, DC. Retrieved from http://www.maweb.org/documents/document.356.aspx.pdf

Moritz C, Patton JL, Conroy CJ, Parra JL, White GC, Beissinger SR (2008) Impact of a Century of Climate Change on Small-Mammal Communities in Yosemite National Park, USA. Science 322: 261–264. doi: 10.1126/science.1163428

Nufio CR, McGuire CR, Bowers MD, Guralnick RP (2010) Grasshopper community response to climatic change: Variation along an elevational gradient. PLoS ONE 5: e12977. doi: 10.1371/journal.pone.0012977

Perrine JD, Patton JL (2011) Letters to the Future. In Canfield MR (Ed) Field Notes on Science & Nature. Harvard University Press, Cambridge, Massachusetts.

Remsen D, Knapp S, Georgiev T, Stoev P, Penev L (2012) From text to structured data: Converting a wordprocessed floristic checklist into Darwin Core Archive format. PhytoKeys 9: 1–13. doi: 10.3897/phytokeys.9.2770

Sheffield C, Nakasone S, Ferrante R, Peters T, Russell R, Van Camp A (2011) Merging Metadata: Building on Existing Standards to Create a Field Book Registry. Libreas: Library Ideas 7: 66–74. Retrieved from http://www.libreas.eu/ausgabe18/texte/08sheffield.htm

Sheffield C, Nakasone S (2011) Together under one roof: Combining collection and item level description through multiple metadata schemas. Proceedings of the American Society for Information Science and Technology 48(1): 1–4. doi: 10.1002/meet.2011.14504801312

Tingley MW, Monahan WB, Beissinger SR, Moritz C (2009) Birds track their Grinnellian niche through a century of climate change. Proceedings of the National Academy of Sciences 106: 19637–19643. doi: 10.1073/pnas.0901562106

Wake D, Vredenburg VT (2008) Are we in the midst of the sixth mass extinction? A view from the world of amphibians. Proceedings of the National Academy of Sciences USA 105(1): 11466–11473. doi: 10.1073/pnas.0801921105

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. doi: 10.1371/journal.pone.0029715

Worm B, Barbier EB, Beaumont N, Duffy JE, Folke C, Halpern B, 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.

Appendix 1

Darwin Core categories and field names used in this project. The authors generated the non-Darwin Core Terms and associated fields.

Darwin Core Class	**Terms included in Darwin Core file**
Record-level Terms	dcterms:modified, basisOfRecord, institutionCode, collectionCode, source
Occurrence	catalogNumber, recordedBy
Event	eventDate, year, month, day, verbatimDate, fieldNotes
Location	country, countryCode, stateProvince, locality, verbatimLocality
Identification	identifiedBy, identificationRemarks,
Taxon	taxonID, scientificName, kingdom, phylum, class, order, family, genus, species, vernacularName, taxonStatus, taxonRemarks
Non-Darwin Core Terms	– *ScrapedName* records the scientificName for the organism observed as entered by Henderson and transcribed by us. – *AnnotatorName* records the corrected ScrapedName as recorded by the annotators. The annotators had the option of leaving this field blank, in which case we use the ScrapedName as the AnnotatorName. – Both ScrapedName and AnnotatorName were fed through a taxonomic resolution process (see Methods, section "Proofing the Darwin Core record set"). Three taxonomic resolvers were used for some of the records: the Global Names Index (GNI), the Encyclopedia of Life (EOL) and the Integrated Taxonomic Information System (ITIS). The resulting identifiers and best-matched scientificNames are provided for all three services; additionally, our ITIS service returned vernacular names, which are also recorded. The *Source of correct name* field indicates whether EOL, ITIS or Both services were returned the correct name. – *canonicalScientificName* is the scientificName with the authorship information deleted. – *AnnotatorLocality*: Annotators were asked to provide a corrected, modern place name for the verbatimName; these are recorded here. – Higher taxonomy (kingdom, phylum/division, etc.) were only extracted from ITIS for records where the ITIS name was correct. The *taxonID* field contains the ITIS Taxonomic Serial Number (TSN) used to look up the higher taxonomy; the *scientificName from TSN* field contains the scientific name that ITIS associates with that TSN.

Appendix 2

Data extraction methodology. (doi: 10.3897/zookeys.209.3247.app2) File format: PDF.

Explanation note: This supplement contains a detailed description of the steps we carried out to extract transcriptions and annotations, from Wikisource via the MediaWiki API. The Perl scripts we used to carry out these steps are available online at https://github.com/gaurav/henderson.

Citation: Thomer A, Vaidya G, Guralnick R, Bloom D, Russell L (2012) From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 235–253. doi: 10.3897/zookeys.209.3247.app2

Appendix 3

Text file containing all occurrence records. (doi: 10.3897/zookeys.209.3247.app3) File format: CSV.

Explanation note: A complete set of occurrence records extracted from Henderson's notebooks 1-3.

Citation: Thomer A, Vaidya G, Guralnick R, Bloom D, Russell L (2012) From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 235–253. doi: 10.3897/zookeys.209.3247.app3

ZooKeys 209: 255–267 (2012)
doi: 10.3897/zookeys.209.3288
www.zookeys.org

RESEARCH ARTICLE

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

Integrating specimen databases and revisionary systematics

Randall T. Schuh[1]

1 *Division of Invertebrate Zoology, American Museum of Natural History, New York, New York 10024 USA*

Corresponding author: *Randall T. Schuh* (schuh@amnh.org)

Academic editor: *V. Blagoderov* | Received 25 April 2012 | Accepted 22 June 2012 | Published 20 July 2012

Citation: Schuh RT (2012) Integrating specimen databases and revisionary systematics. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 255–267. doi: 10.3897/zookeys.209.3288

Abstract

Arguments are presented for the merit of integrating specimen databases into the practice of revisionary systematics. Work flows, data connections, data outputs, and data standardization are enumerated as critical aspects of such integration. Background information is provided on the use of "barcodes" as unique specimen identifiers and on methods for efficient data capture. Examples are provided on how to achieve efficient workflows and data standardization, as well as data outputs and data integration.

Keywords

specimen databases, workflows, revisionary systematics

Introduction

Meier and Dikow (2004) argued that biodiversity data should come from revisionary studies–rather than from uncritical digitizing of museum specimen data, because such revisions 1) provide the most accurate identifications, 2) provide the most complete taxonomic coverage, 3) and they satisfy these points in a cost-effective way. Nonetheless, revisions are what might be viewed as the traditional approach to creating a database of specimens for a taxon. In the following pages I will provide a rationale and a roadmap for satisfying both the acquisition of high-value biodiversity data while at the same time creating a structured database of that same information during the revisionary process.

The creation of specimen databases–a subset of a field that has frequently been referred to as biodiversity informatics (Johnson 2007)–has reached a point in its maturity that has brought down per-specimen digitization costs and increased accessibility of available tools to a much broader range of systematists than was the case 15 years ago. Movement into the Internet Age, the more widespread use of digital technologies such as barcodes, and the increasing sophistication and availability of database technology are all contributing factors.

One manifestation of the maturity of biodiversity informatics can be seen in the United States National Science Foundation (NSF) program Advancing Digitization of Biological Collections (ADBC 2011), a ten-year initiative designed to promote and fund the digitization of biological collections. The core digitization activities are in Thematic Collection Networks (TCN), funded projects that bring together a group of collections focusing on a common research or investigative theme. The TCNs are coordinated through a "national resource" or HUB (Home Uniting Biocollections). Through the activities of the HUB we should anticipate seeing the dissemination of more tools and improved access to relevant technology and the methods by which data can be integrated across collections and which would also be of use to revisionary systematists.

Most of the tools applied in specimen data capture—such as databases and barcodes–were initially developed for use in industry. Their application in the realm of biological collections was originally in collection management, rather than as an adjunct to the preparation of scientific publications such as taxonomic revisions. Even though the technology is available, the full integration of biodiversity databases into revisionary studies is far from a fully realized objective. The reasons may include the foreign nature of the technology to older investigators, the lack of direct access to the tools, the lack of technical expertise for implementation of the technology, and simple reluctance to alter traditional approaches to the preparation of revisions.

In the following pages I will argue for the adoption of database tools as an integral part of the revisionary process. This is not just an argument for the adoption of modern technology. Experience suggests that the benefits accrued will more than justify the costs incurred, both in terms of money spent to acquire the necessary equipment and software as well as time spent learning to incorporate "databasing" into one's day-to-day taxonomic labors.

I have already written about aspects of this subject in two prior papers which focused on the methods for the solution of large-scale taxonomic problems (Cassis et al. 2007) and the use of Web-based data capture as a model for multi-national systematic research projects (Schuh et al. 2010). The lessons learned, and approaches outlined, in those papers derived largely from experience gained in the conduct of an NSF-funded Planetary Biodiversity Inventory (PBI) project (http://research.amnh.org/pbi/) for the study of the plant-bug subfamilies Orthotylinae and Phylinae (Insecta: Heteroptera: Miridae). As was the case in those works, this paper is based largely on approaches developed during the PBI project. The present paper will not attempt to resolve the intertwined issues of 1) whether databases should be collection based, with research data gathered from across a spectrum of such information repositories, 2) whether databases

should be project based and integrate data across taxonomic lines or research themes, or 3) whether both types of databases can and should co-exist. Rather, I will focus on workflow, data connections, data outputs, and data standardization, issues that are central to enhancing the revisionary taxonomic process.

Database choice

The arguments to be made in this paper assume that one has access to a specimen database with certain "basic" features. These include the capability to efficiently capture all relevant and necessary data in a highly structured format, the capability to organize those data in ways useful to the reviser, and the capacity to output data for direct use in revisions as well as for the production of maps and other visual aids. A number of such database products exist, some free of charge, and most capable of performing the necessary functions. They exist as stand-alone products, as institutional tools functioning on a local area network, or as Internet-based tools. Because information on these databases is not the primary intent of this article, and because the logic of choice is beyond the scope of this article, I will not dwell further on the issue database choice. As sources of further information the reader might wish to consult Schuh et al. (2010) and the abstracts in Session 1 from the 2011 meeting of the Entomological Collections Network (http://www.ecnweb.org/dev/AnnualMeeting/Program).

Unique specimen identifiers (USIs)

The use of *barcodes* to uniquely identify individual specimens goes back at least to the work of Daniel Janzen and the InBio collections in Costa Rica (Janzen 1992). In the intervening 20 years, code technology has advanced, such that many applications now use matrix codes (Fig. 1, right) which can store much more information in a smaller format than is the case with linear barcodes (Fig. 1, left). Whatever technology you choose, the use of unique specimen identifiers (USIs) provides the capacity to track individual specimens with exactitude, and to directly associate a variety of information sources with them.

Machine readability, although not an essential component of a USI, is a valuable aspect of barcode and matrix code labels. At $250 or less, the cost of code readers is now about one-tenth what it was in 1994 (Thompson 1994), making them a truly affordable databasing asset. The most convincing argument for the use of machine reading is that the readers do not make mistakes, whereas human transcription is prone to error. Once their use becomes part of your work routine, barcode readers significantly enhance the speed and accuracy with which USI data can be entered into the database, either when doing original data entry or when retrieving specimen data. Some have worried that barcode reading technology will change over time, and that encoded labels will therefore become obsolete. In anticipation of this potential reality, all such labels should include the alphanumeric representation of the code as well as the code itself (Fig. 1).

Figure 1. Linear barcode label (left), matrix code label (right).

Production of barcode labels can be contracted out to specialized suppliers or can be done in house. Because of the widespread use of the technology, appropriate tools for their preparation and printing are readily available. Nonetheless, a distinct difference between the commercial application of these technologies and their use in biological collections is that the latter group of users expects the labels to be permanent, suitable for alcohol and dry storage, and for the printed matter to be of high resolution, whereas none of those criteria is important in industrial applications such as package delivery and airline baggage identification. Although most any printer can be used to print barcodes, specialized software is required to produce individual labels with sequential numbering (e.g., BarTender 2012). Many database applications expect coded information to be in a certain format. Thus, when preparing barcodes, it is important to verify that the format of the code, such as the institutional acronym/collection code and numerical string that follows, are in a format accepted by your database.

Curators of biological collections have long applied catalog numbers to specimens, although such practice has been much less common with insect collections than with those of recent vertebrates, fossils, and plants, for example. Although these "catalog" numbers were often not unique within institutions, let alone across institutions, they did offer a way to uniquely associate specimens with log-books of data, accession information, field notebooks, and other written resources. Most barcode implementations come much closer to globally-unique identification than was the case with traditional catalog numbers, through the use of codes that combine an institution code + a collection code + plus a catalog number. This approach complies the with Darwin Core standard promoted by the Taxonomic Database Working Group (2012), with the caveat that a single code is sometimes applied to a group of specimens, often referred to as a *lot*, in which case the unique identifier applies to more than one specimen.

The use of barcodes has resulted in the frequent attachment of multiple codes to individual specimens, often in addition to traditional catalog numbers. Several factors are at play, including the use of barcodes as the modern equivalent of catalog numbers as well as to identify specimens used in independent research projects. Sometimes these two uses are included in a single label, sometimes on separate labels. Recent Internet-based discussions suggest that prevailing opinion regards the attachment of multiple labels as acceptable, often unavoidable, and that the all of the codes should remain on the specimens in perpetuity. Some or all of these codes may be globally unique.

Verbatim vs Transformed Data: A choice mediated by the use of USIs

A recent symposium organized for the 2011 meeting of the Entomological Collections Network (Reno, Nevada; http://www.ecnweb.org/dev/AnnualMeeting/Program), included a more or less equal number of presentations arguing for 1) the verbatim capture of all label data in a single text field with subsequent transformation into a more highly structured format, or for 2) transformation of label data into a publication-ready format as an integral part of the data-capture process. Schuh et al. (2010) made the argument for the latter approach, but to my knowledge there are few 1) published arguments concerning the merits and demerits of these alternative approaches or 2) quantitative studies analyzing the efficiency of the alternative approaches.

Verbatim data capture allows for data acquisition with minimum training of the data-entry personnel. The only real requirement would seem to be the ability to read the labels and convert them into a text string. Those data must then be transformed into a structured format and written to the database tables by the use of some software algorithm or other automated data-parsing approach. Finally, the accuracy of the transcription must to be checked, an additional step, and one that will require greater expertise in interpretation of label data than did the initial data entry.

Transforming data as part of the data-capture process, so that the data are in the exact form used by the database requires additional training of personnel over what is needed for verbatim data capture. Nonetheless, because the data are structured during the process of data capture, these data are ready for straightforward review for accuracy, at which point they can be considered "publication ready" and the additional training effort will be available for all subsequent data capture.

Even though errors may be made under either approach, the use of USIs allows for subsequent investigators to return to individual specimens with substantial confidence concerning the correspondence of original and transcribed data. It is my view, and that of many of my colleagues, that the capture of transformed data is more efficient because it is a one step process that allows for immediate use of the data. Data captured en masse from collections will not be available until they have undergone algorithmic transformation and been approved for upload, thus potentially presenting a time lag that will hinder the progress of the reviser or other data user.

Data-capture Work Flow in Revisionary Studies

Label generation: Capture field data to the database and generate all labels from it

Many specimens used in revisionary studies, possibly most particularly in entomology, come from the dedicated fieldwork of the reviser. Thus, the opportunity to use appropriate technology in conjunction with fieldwork would seem to be a straightforward choice. This would include the capture of latitude/longitude and altitude data in the field through the use of a GPS (global positioning system) device in the form used by

geographical information systems software and the recording of field data in exactly the format to be used in the specimen database. Thus, the choice should be degrees and decimal parts thereof for lat/long data and meters for altitude. Locality and collection-event data can be directly captured in digital form in the field, or recorded to an archival field notebook and captured in digital form at the earliest subsequent opportunity. GPS data can be downloaded directly, an approach that precludes mistakes during transcription of numbers, one of the most common errors made in the capture of field data.

The argument for using a database to capture/store field data and to produce specimen labels is bolstered by the many examples of specimens in collections where multiple collectors on the same field trip produced their own labels. Although such labels contain similar information, they are frequently not identical and thus may end up in a database as representing distinct localities. The drawbacks are one or more of the following: 1) what was actually a single locality will likely end up being georeferenced multiple times, or if lat/long data were captured in the field, those data may still not be identical on the labels; 2) one or more renderings of the collection locality may contain errors; 3) the locality may be easily interpreted in one rendering but difficult to interpret in another; and 4) some of the labels may be substandard from a curatorial point of view. Using the database from the outset, including for the generation of labels, facilitates data standardization and the uniform presentation of data in all of its subsequent uses. It also greatly facilitates the retrospective capture of data for specimens whose localities are already in the database. This last point has economic implications, because even though the personnel time available to enter all specimens collected at a given locality may not be available at the time the specimens are mounted and labeled, the cost of entering just the locality/collection event data at the time of the fieldwork will never be an issue.

Specimen data: Enter specimen data early in the revisionary process

Although it has been said many times, and therefore may seem trite, the use of a database can save many key strokes. Once the data have been entered and checked for accuracy for a given locality, they can be re-used in the generation of labels, for preparation of reports of "specimens examined", and for many other purposes. If for any reason an error is found, it can be corrected and all subsequent and varied uses of those data will be accurate and uniform. The capture early on in the revisionary process of as much specimen data as possible allows for the structuring and examination of those data in ways that are otherwise difficult and cumbersome. What is paramount is that the data are captured once but useable in many ways without the need for re-keyboarding. Nonetheless, it is probably fair to say that in the traditional preparation of a revision, the last thing to be done was to capture specimen data, whether using a word-processing file, spreadsheet, or relational database. The use of a specimen database facilitates the capture of specimen data much closer to the beginning of the revisionary process, so that all relevant observations on specimens can be managed through the medium

of the database and available over the entire course of the revisionary process. In addition to locality data, such observations might include host data, habitat descriptions, museum depository information, dissections, images, measurement data, and DNA sequence files, to name just some of the possibilities.

Capturing specimen data: Organize specimens before capturing data

With some forethought and advance preparation the process of retrospective specimen data capture can be made more efficient and also facilitate other aspects of the revisionary process. Collective experience of participants on the Planetary Biodiversity Inventory project, and other colleagues, recommends the following sequence of events for dealing with specimens from any given institution (Fig. 2):

1. Sort specimens by provisional species criteria (morpho species, etc.)
2. Sort specimens by locality
3. Sort specimens by sex
4. Affix sequential unique specimen identifiers (barcodes, matrix codes)
5. Enter data in database

This workflow is efficient because it allows for series of specimens of the same species, sex, and locality to bear USI codes in sequential order and for data for all of those

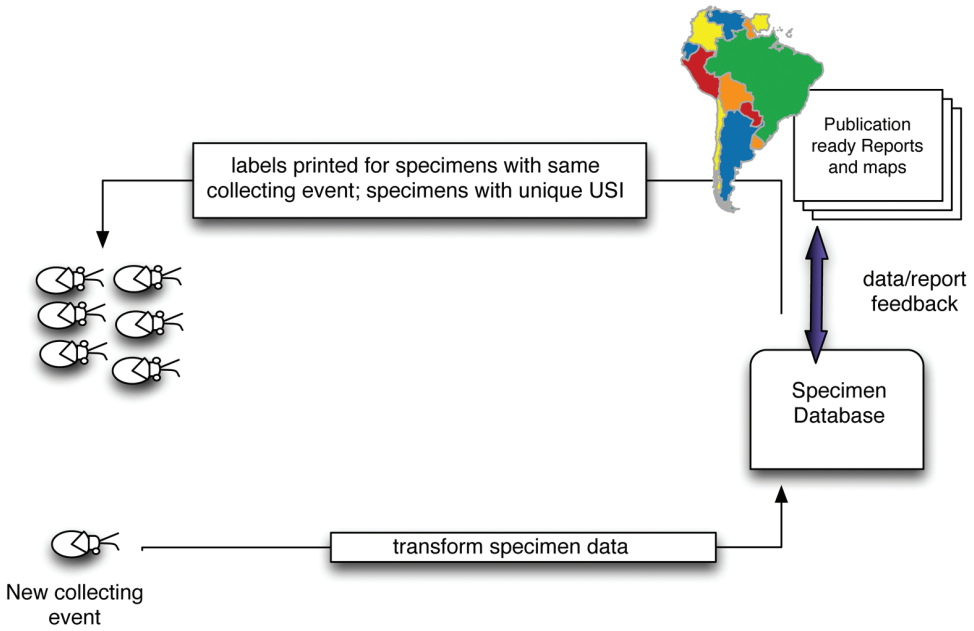


Figure 2. Diagram of specimen data connections and work flows.

specimens to be captured as a single action. Of course, this approach is most important in those cases where there are multiple examples of a species from a single collecting event.

Although sexing specimens may not be necessary or possible for all taxa, in many groups the standard description is based on one sex, or the other. Sorting by sex before specimen data are entered facilitates comparisons, adds a logical aspect to the organization of the material in collections, and helps to produce sequential USIs, which saves space in presenting data on specimens examined. If during the course of preparing a revision specimens are found to have been initially misidentified, the records for those specimens can be readily retrieved via the barcode and the identifications in the database can be corrected.

Data Connections

Georeferencing and mapping: Using the database as an analytic tool

Georeferencing–the addition of latitude/longitude data to individual specimen records–permits the mapping of specimen distributions in space. Such mapping should be part of the revisionary process, rather than taking place near the end, as has traditionally been the case. As a matter of standard practice, lat/long data should be available on all specimen labels being produced as a result of fieldwork in this day and time. And, as mentioned above, data from modern fieldwork should desirably be captured to a database for the preparation of all labels, such that no manual georeferencing will be required. Under this approach, georeferencing is intimately related to the issue of workflow, because the earlier in the revisionary process the specimen data can be mapped, the more useful they will be. Nonetheless, lat/long data will have to be determined for legacy material.

Georeferencing was at one time a time-consuming and tedious process. It is now much easier, due to the ready availability of automated tools such as GeoLocate (2010), unrestricted access to quality gazetteers for much of the world (Fuzzy Gazetteer 2003, Geonames: http://www.geonames.org/, GNIS 2011), and the universal accessibility of Google Earth (2012) and Google Maps (2012), among other sources. Thus, there is a strong argument for georeferencing of specimen data in close coordination with initial capture of those data. Such an approach will allow for the visualization of distributions early in the revisionary process. This will provide a feedback loop concerning the accuracy of the georeferencing itself, the interpretation of distributional patterns, and the on-the-spot investigation of suspect identifications as recognized by the visualization of distributional outliers.

Even if your database application does not have integrated mapping tools, the simple ability to export lat-long data will permit the easy visualization of those data and the creation of maps (fig. 3). Some of the tools freely available are the Simple Mapper (Shorthouse 2010), Google Earth, and the Global Mapper of Discover Life (2012). All allow for lat-long data in decimal format to be pasted into the application for production of maps useful for publication or for the preparation of presentations.

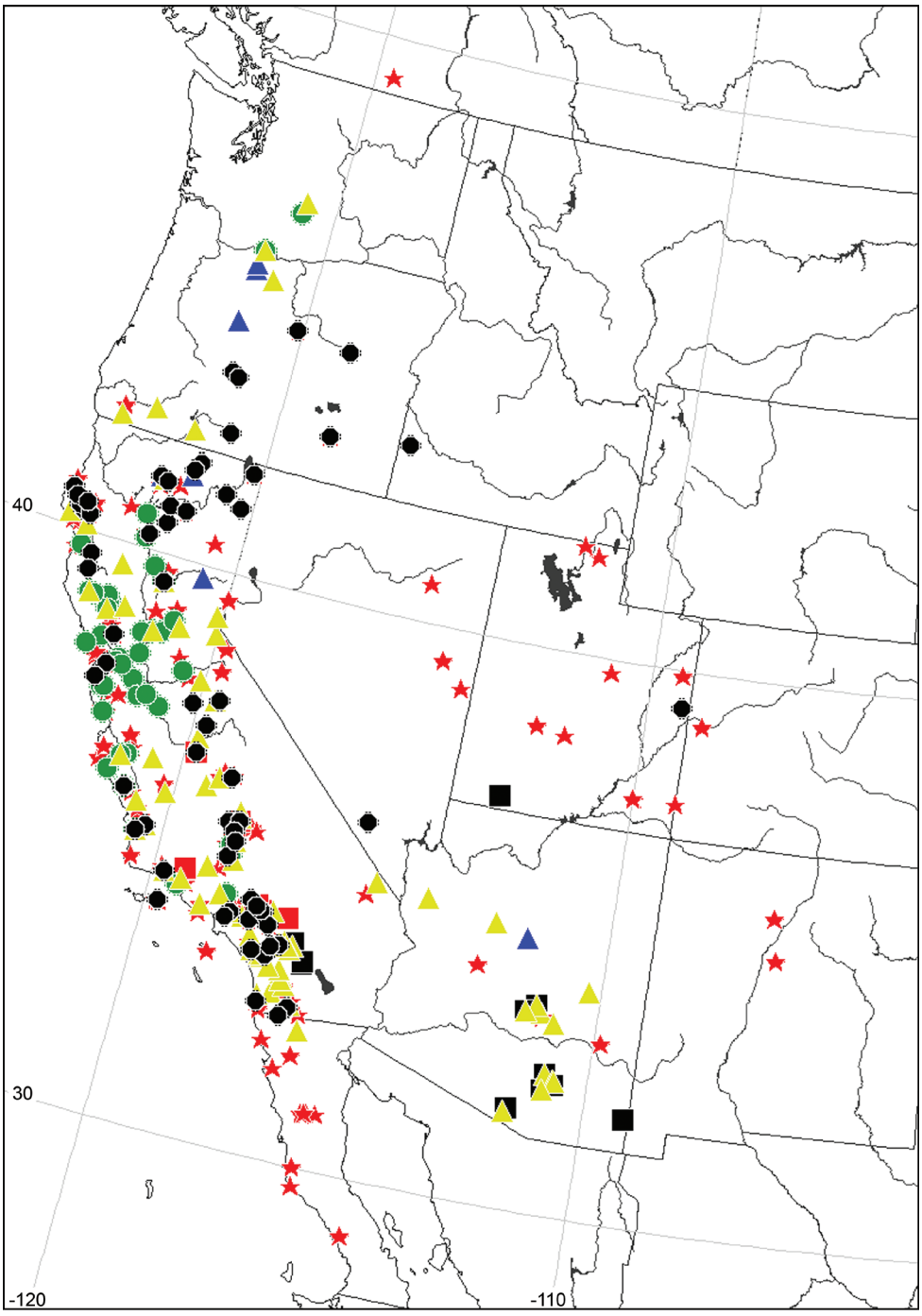


Figure 3. Map of species distributions in western North America created using the Simple Mapper.

Measurements, images, etc.: Integrating other data sources

As is the case with georeferencing early in the study of specimens, the use of USIs as labels for images, measurement data, and DNA sequences allows these data sources to become an integral part of the data record for the specimens under study, and for tracking those data in an unequivocal manner.

Data outputs: Organizing data through the power of report writing

Reports of specimens examined

Once specimen data have been captured, checked for accuracy, and georeferenced, the real power of the database for revisionary studies comes from the ability to generate reports. Possibly most valuable is the preparation of reports of specimens examined, a core component of traditional revisions (Fig. 4). The reports can be written, revised, and rewritten in a matter of seconds or minutes, and preclude retyping and reformatting of data; the same can be said for the preparation of maps. Other types of reports, such as species by locality, hosts by species, and range of collection dates–among many other possibilities–are also easily produced and complement the contents of many revisions.

Beckocoris inventarium

Holotype: **USA: California:** ***Los Angeles Co.:*** Largo Vista Rd 3.1 mi S of Rt 18, SE of Llano, 34.45251°N 117.7651°W, 1275 m, 17 May 2004, Schuh, Cassis, Schwartz, Weirauch, Wyniger, Forero, *Tetradymia stenolepis* E. Greene (Asteraceae), det. A. Sanders UCR 140645 Field ID H10, 1;m (AMNH_PBI 00297367) (AMNH).

Paratypes: **USA: California:** ***Los Angeles Co.:*** Largo Vista Rd 3.1 mi S of Rt 18, SE of Llano, 34.45251°N 117.7651°W, 1275 m, 17 May 2004, Schuh, Cassis, Schwartz, Weirauch, Wyniger, Forero, *Tetradymia stenolepis* E. Greene (Asteraceae), det. A. Sanders UCR 140645 Field ID H10, 2;f (AMNH_PBI 00297369, AMNH_PBI 00297384), 3;m (AMNH_PBI 00297364-AMNH_PBI 00297366), 19;f (AMNH_PBI 00297370-AMNH_PBI 00297383, AMNH_PBI 00297385-AMNH_PBI 00297389) (AMNH), 1;m (AMNH_PBI 00297368), 1;f (AMNH_PBI 00297390) (CNC). ***San Bernardino Co.:*** Phelan, Rt 138 at Phelan Road, 34.42531°N 117.6174°W, 1310 m, 16 May 2004, Schuh, Cassis, Schwartz, Weirauch, Wyniger, Forero, *Tetradymia stenolepis* E. Greene (Asteraceae), det. A. Sanders UCR 140645 Field ID H10, 2;m (AMNH_PBI 00297392, AMNH_PBI 00297398), 6;m (AMNH_PBI 00297391, AMNH_PBI 00297393-AMNH_PBI 00297397), 11;f (AMNH_PBI 00297400-AMNH_PBI 00297409, AMNH_PBI 00297411) (AMNH), 1;m (AMNH_PBI 00297399), 1;f (AMNH_PBI 00297410) (USNM).

Other Specimens Examined: **USA: California:** ***San Bernardino Co.:*** Apple Valley, 34.53139°N 117.28278°W, 830 m, 15 May 1955, W. R. M. Mason, 2;f (AMNH_PBI 00381924, AMNH_PBI 00381925) (CNC). Victorville, 34.53611°N 117.29028°W, 09 May 1955, W. R. M. Mason, 2;f (AMNH_PBI 00381926, AMNH_PBI 00381927) (CNC).

Figure 4. Report of specimens examined, including unique specimen identifiers.

The power of database query languages facilitates the preparation of counts of total specimens examined, specimens examined by museum, specimens dissected, and other summary information that helps to clarify the sources and uses of data.

Species pages: Integrating all data sources in electronic form

Species pages have become the Internet equivalent of species treatments in traditional print publications. The Encyclopedia of Life (EOL 2012) is centered around this approach and promotes the goal of creating a page for every known species. "Web aggregators" such as Discover Life (2012) produce species pages through highly automated means, providing images, keys, and maps for a very large number of taxa. The research efforts of my colleagues and myself resulted in the creation of the Heteroptera Species Pages (2012; http://research.amnh.org/pbi/heteropteraspeciespage/) which assembles available data from a specimen database and creates pages on the Web in real time.

Descriptive databases: Adding the descriptive component

More has probably been written on the use of descriptive databases in revisionary systematics than has been the case for specimen databases. These products allow for the creation of character descriptions, natural language descriptions, interactive keys, and phylogenetic matrices. The most longstanding version of such a database is DELTA (Dallwitz 2010); a more recent entrant is Lucid Builder (Lucidcentral.org 2012), which has the advantage of employing the TDWG SDD (Structure of Descriptive Data) protocol which allows for the interchange of data with other platforms. One example of moving the descriptive database concept to the Internet is that of Norman Platnick and his NSF-funded team working on the spider family Oonopidae (http://research.amnh.org/oonopidae/index.php). Descriptive databases and specimen databases are a logical complement to one another. The former require a controlled set of character descriptions in order to function effectively, a time-consuming activity, but one that can pay off handsomely in groups with many species to be described and where ongoing identification of specimens—such as in groups of insects of great economic importance—is a major issue. The latter require the capture of specimen label data, but allow for extensive and continued reuse of those data once acquired.

In my own work, I have created matrices in the program Winclada (Nixon 1999) and used the facilities of the program to output descriptions that can be utilized in publication with minimal editing (e.g., Schuh and Pedraza 2010). As is the case with descriptive databases such as DELTA and Lucid Builder, or with programs such as mx (http://mx.phenomix.org/index.php/Main_Page), the matrix that is used to prepare descriptions and keys will often not be identical to a matrix

well suited to phylogenetic analysis. Nonetheless, the gap between these two uses is oftentimes small, and minimal modification will allow for both matrices to be derived from essentially a single effort.

Conclusions

In summary, the affordable technology for capture, manipulation, and sharing of specimen data awaits revisers to avail themselves of the opportunity to harness the power of these tools (see Johnson 2007). Experience suggests that seamless integration of revisionary research and database technology will not necessarily take place overnight, but once the logic of using a database as part of revisionary studies is in place, the database will take on the status of a research tool, not just as a way to capture structured specimen data. The time spent on specimen data capture will be quickly repaid through the ability to use those standardized data at every step of the revisionary process, beginning with the standardization of labels by creating the database record of all relevant data at the time of field work, continuing with the creation of maps and reports during the process, and concluding with use of the identical data in the published product. These benefits accrue not only to the individual investigator, but more particularly to research teams where multiple investigators are involved in the preparation of revisions and other specimen-based research products.

Acknowledgments

The approaches described in this paper were derived from experience gained during participation in, and administration of, a USA National Science Foundation-funded Planetary Biodiversity Inventory award for the study of the true bug family Miridae. Many PBI project colleagues helped shape my views on specimen databases and their place in revisionary studies and collection management. Among others, these include co-principal investigator Gerry Cassis, Michael Schwartz, Sheridan Hewson-Smith, Christiane Weirauch, Denise Wyniger, Fedor Konstantinov, and Dimitri Forero. I am also indebted to the late James S. Asche and his colleagues at the University of Kansas, Lawrence, for generous discussion of their own experiences with specimen databasing and to John S. Ascher (AMNH) for his contributions to the logic of specimen data capture. I thank: Katja Seltmann (AMNH) and Michael Schwartz for discussion and encouragement and for their critical comments on earlier versions of the MS which helped to form the final product; Dimitri Forero, Ruth Salas, and two anonymous reviewers for comments on the manuscript; and Katja for the conception and creation of figure 2. Finally, I thank Nina Gregorev (AMNH) for her expert programming contributions and many suggestions on how to produce the truly user-friendly database implementation used in the PBI and TCN projects. This work was funded by NSF awards DEB 0316495 (PBI) and EF1115080 (ADBC-TCN) to the American Museum of Natural History.

References

Advancing Digitization of Biological Collections (ADBC). National Science Foundation Program Solicitation (2011) http://www.nsf.gov/pubs/2011/nsf11567/nsf11567.htm

BarTender (2012) http://www.seagullscientific.com/aspx/bar-code-label-software.aspx

Cassis G,Wall MA and SchuhRT (2007) Insect biodiversity and industrializing the taxonomic process: A case study with the Miridae (Heteroptera). In: Hodkinson T, Parnell J (Eds) Towards the Tree of Life: taxonomy and systematics of large and species rich clades. CRC Press, Boca Raton, 193–212.

Dallwitz MJ (2010) Overview of the DELTA system. http://delta-intkey.com/www/overview.htm

DiscoverLife (2012) http://www.discoverlife.org

EOL (2012) Encyclopedia of Life. http://www.eol.org

Fuzzy Gazetteer (2003) http://isodp.hof-university.de/fuzzyg/query/

Geolocate (2010) A Platform for Georeferencing Natural History Collections Data. http://www.museum.tulane.edu/geolocate/

Global Mapper | Discover Life (2012) http://www.discoverlife.org/mp/20m?act=make_map

GNIS (2011) Geographic Names Information System. http://nhd.usgs.gov/gnis.html

Google Earth (2012) http://www.google.com/earth/index.html

Google Maps (2012) http://maps.google.com/

Heteroptera Species Pages (2012) http://research.amnh.org/pbi/heteropteraspeciespage/

Janzen DH (1992) Information on the barcode system that INBio uses in Costa Rica. Insect Collection News 7: 24

Johnson NF (2007) Biodiversity informatics. Annual Review of Entomology 52: 421–438. doi: 10.1146/annurev.ento.52.110405.091259

Lucidcentral.org (2012) http://www.lucidcentral.org/en-us/software/lucid3.aspx

Meier R, Dikow T (2004) Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. Conservation Biology 18: 478–488. doi: 10.1111/j.1523-1739.2004.00233.x

Nixon KC (1999) Winclada Program and documentation. http://www.cladistics.com/about_winc.htm

Schuh RT, Pedraza P (2010) *Wallabicoris*, new genus (Hemiptera: Miridae: Phylinae: Phylini) from Australia, with the description of 37 new species and an analysis of host associations. Bulletin of the American Museum of Natural History 338, 117pp. http://hdl.handle.net/2246/6066

Schuh RT, Hewson-Smith S, Ascher JS (2010) Specimen databases: A case study in entomology using Web-based software. American Entomologist 56: 206–216.

Shorthouse DP (2010) SimpleMappr, an online tool to produce publication-quality point maps. http://www.simplemappr.net

Taxonomic Database Working Group (2012) http://wiki.tdwg.org/twiki/bin/view/DarwinCore/DarwinCoreDraftStandard

Thompson FC (1994) Bar codes and specimen data management. Insect Collection News 9: 2–4.