Biodiversity Informatics, 7, 2010, pp.113 – 119.

TOWARDS DEMAND-DRIVEN PUBLISHING: APPROACHES TO THE PRIORITIZATION OF DIGITIZATION OF NATURAL HISTORY COLLECTION DATA

PENNY BERENTS (1), MICHELLE HAMER (2), AND VISHWAS CHAVAN (3)*
(1)Australian Museum, 6 College Street, Sydney, NSW 2010, Australia. Email: penny.berents@austmus.gov.au
(2) South African National Biodiversity Institute, 2 Cussonia Ave, Brummeria, Pretoria, South Africa and School of Biological & Conservation Sciences, University of KwaZulu-Natal. Email: m.hamer@sanbi.org.za
(3)Global Biodiversity Information Facility Secretariat, Universitetsparken 15, DK 2100, Copenhagen, Denmark. Email: vchavan@gbif.org
**Corresponding author*

Abstract – Natural history collections represent a vast repository of biodiversity data of international significance. There is an imperative to capture the data through digitization projects in order to expose the data to new and established users of biodiversity data. On the basis of a review of the current state of digitization of natural history collections, a demand-driven approach is advocated through the use of metadata to promote and increase access to natural history collection data.

Key Words. – Natural History Collections, publishing, demand-driven digitization, prioritization, metadata, biodiversity data

Natural history collection data is a critical component of biodiversity data which has widespread application in biodiversity research, natural resource management and biosecurity (Chapman, 2005; Tann, et al., 2008; Pyke & Ehrlich, 2010). The Global Biodiversity Information Facility (GBIF) Global Strategy and Action Plan for Mobilization of Natural History Collections Data (GSAP-NHC) Task Group was charged with examining priorities for the digitization of natural history collection data (GBIF, 2008a). The drivers for specimen level digitization are many and varied, and often intrinsically linked (Vollmer, et al., 2010), therefore it is not realistic for the Task Group to dictate priorities for specimen level digitization. The proportion of a collection which is digitized varies greatly from one collection to another with some collections fully digitized and others having made little progress. The opportunities and priorities for digitization vary from one institution and one country to another. There is no single answer to setting priorities for specimen level digitization and therefore, we argue that the use of metadata to describe a collection be a key component to achieve demand-driven data digitization and publishing.

We are further of the opinion that:

1. Metadata must be used to expose data to users and to expand the user base,
2. Metadata and specimen level digitization should be considered as a part of the digitization process and be prioritized together,
3. Metadata creation can be considered on a scale from local to global.

We therefore recommend that the creation of metadata records for a collection is essential. This, however, does not replace the need for digitization itself but should be considered as a part of any digitization project. Metadata creation will lead to digitization by increasing the exposure of the data.

DEMAND-DRIVEN PRIORITIZATION OF COLLECTION DIGITIZATION

The majority of digitization activities and initiatives are opportunistic in nature (Vollmer, et al., 2010) but with such an opportunistic approach, digitization of the world's natural history collections will not be achieved in the foreseeable future. Further, as stated by Scoble & Bourgoin (2010), and Berendsohn & Seltmann (2010), with current resource allocation, and socio-political and scientific priorities, it may not be possible to achieve digitization of all the specimens housed in the world's natural history collections, which comprise more than 3 billion specimens. The management of collections is greatly enhanced by digitization but despite this strong internal driver, institutions have been unable to generate adequate resources to achieve this goal. It is therefore necessary to mobilize external resources and expose the collections and data to a wider user group who will drive and resource digitization priorities (National Science and Technology Council, Committee on Science, Interagency Working Group on Scientific Collections, 2009). 'Demand-driven digitization' will address the immediate needs of stakeholder communities and increase the potential for attracting financial and human resources and improved infrastructure.

We therefore recommend that natural history collections adopt the approach of 'demand-driven' digitization of collections to address the requirements of stakeholder communities. We strongly advocate that the collection management and curatorial community develop an institutional or collection specific demand-driven prioritization plan for collection digitization on the basis of content needs of the major stakeholders and the user community. The integration of such institutional or collection specific plans would then help to develop a national or thematic collection digitization strategy and action plan. We further recommend that GBIF works with major natural history collection stakeholder communities to develop best practice guidelines for developing (a) institution or collection specific, demand-driven plans for collection digitization, and (b) national or thematic collection digitization strategy and action plans. Chris Frazier, et al. (2008) has written a very useful guide on initiating a collection digitization project, as part of the GBIF Training Manual on digitization of natural history collection data (GBIF, 2008b). In our opinion, this will further help federal science funding agencies and private donors to cooperate in developing comprehensive funding strategies which will result in key scientific, ecological, and social issues being addressed.

METRICS FOR PRIORITIZING THE DIGITIZATION

We reviewed a variety of factors which influence decisions regarding digitization of natural history collections. These factors include (1) type specimens, (2) collections associated with projects, (3) historical significance, (4) taxonomic priorities, (5) ecosystem relevance, and (6) species of special concern. We suggest that the following criteria should be considered when setting priorities for digitization:

1. **Type specimens**: Type specimens are important not only for taxonomists, but also as a reference for accurate identification and naming of biological specimens, which is fundamental to all other biological and biology-related fields (agriculture, medicine, conservation, environmental management, etc.). Type specimen records should allow links to the Catalogue of Life, the Encyclopedia of Life and other databases of spatial and temporal information from collections to provide ready access to information such as the type locality and the date of collection. Type data must indicate the status of the type specimen, such as whether it refers to the type specimen of a currently valid species name. Holotype (primary type) specimens should be a top priority, but other types such as paratypes should also be included in the databases. Ideally, each type specimen (at least the holotype) should also be photographed or scanned so that access to information associated with type material can be made globally accessible in perpetuity. A global account of type specimens and the collections in which these specimens are housed will also allow some assessment of the value of different collections. The prioritization of type specimens in collections has three main benefits:
   a. profiles the jewels in the museum and herbarium collections and thus enhances their profile (prestige effect),
   b. creates an achievable target and gets institutions comfortable with what is involved in digitization and thus they may continue with these activities, and
   c. provides another register of all known species that can be compared to Catalogue of Life and other names received by GBIF. GBIF can then link species to their type specimen

locations, thus directing researchers to these essential resources.

2. **Digitization of collections associated with projects**: e.g. museum exhibitions, biodiversity hotspots, biodiversity surveys, conservation questions such as important bird areas, or biological or other research projects. We strongly recommend that digitization of specimens associated with projects should be an integral part of a project and thus be achieved prior to the close of projects. In the case of completed projects an independent assessment needs to be carried out to decide the order of preference for digitization of specimens associated with such projects.
3. **Historical significance**: The unique value of many natural history collections lies in the historical nature of the data. What constitutes an "historical" dataset may be debatable and be regionally variable depending on the time since the greatest change. For example, in developing countries, pre-1980 may be considered historical, while in industrialized countries pre-1900 may be of more value. Specimens to be considered for these datasets should be identified to species level, and have locality data provided as accurately as possible, with some indication of uncertainty such as an uncertainty radius, and include at least the year of collection.
4. **Taxonomic priorities**: Digitization may be focused on taxonomic priorities as a result of taxonomists bringing resources for digitization driven by their need for specimens and data. Taxonomists contribute the greatest amount of material and data to collections as specimens and specimen data are critical for taxonomists' work. It is the taxonomic cadre that provides the value of the specimens and collections to society. Ideally taxonomic priorities should be linked to global or national initiatives, involving several taxonomists and / or institutions working on a single taxon. For example, in South Africa, the South African Butterfly Conservation Assessment has driven the need for specimen data capture from all public institutions and private collectors and has resulted in more than 400,000 butterfly records being captured.
5. **Ecosystem relevance**: This could involve prioritizing the collections which have resulted from biodiversity surveys or research projects on ecosystems or habitats that provide critical services (e.g. freshwater ecosystems, wetlands, coral reefs, forests, rangelands etc).
6. **Species of special concern**:
   a. *Invasive alien species*: invasive species are considered to be one of the greatest threats to biodiversity, and specifically to ecosystem functioning and resilience to change. Information on the diversity and distribution of alien invasive species has global as well as regional and local relevance, and global data sets will be of value to a large number of biologists, environmental managers and conservationists. Having large, global datasets of alien invasive species will also highlight the value of digitization of specimens or observations and the value of GBIF activities to a wide range of stakeholders.
   b. *Species of direct relevance to people*:
      i. *Harvested species (e.g. crops,* medicinal plants, line fish*)*
      ii. *Pests*
      iii. *Diseases or disease vectors (of humans, livestock or crops).*

      The rationale for the selection of the species needs to be explicit and data must include species level identification, accurate locality data as well as the date (minimum of the year) of collection or observation.
   c. *Threatened, endangered, endemic species*: data on these species are critical for natural resource management, conservation planning, and decisions about land use.

The factors discussed in this section are considered useful for determining priorities which will generate demand-driven digitization. Data capture initiatives based on these priorities will provide data of direct use to a wide range of stakeholders, perhaps expanding the traditional users of collection data, and thus increasing the value of GBIF's initiatives (and therefore the possibilities for funding). Understanding current and past distributions of species of direct relevance to human survival is critical not only for human well-being, but also to illustrate the value of collections and taxonomy and of the biodiversity sciences in general.

Having recognized that metadata is essential for the demand-driven digitization of natural history collection data, in the remainder of this paper we discuss various aspects dealing with metadata authoring and publishing.

METADATA, A PRIORITY: WHY?

There are several advantages to authoring and publishing enriched metadata documents, including:

1. Increased discovery and visibility
2. Stimulation of demand-driven digitization
3. Increased usage and user base
4. Comprehensive tracking of the progress of national to global scale digitization
5. Early detection of collection risk assessment – identify collections at risk
6. Improved capacity management – technical, infrastructure, human resources and finance
7. Improved estimation of the scale of biological collections

METADATA: CHALLENGES OR CONSTRAINTS?

- What is metadata?

Discussions with curators and collection managers reveal that there is very poor understanding of what metadata is and its significance for improved discovery and visibility of the collections. This calls for increased awareness and outreach amongst the natural history collections community about the importance of metadata and how it can contribute towards the sustainability and increased use of collections.

- Metadata scale.

One of the very critical decisions that influence the usefulness of a metadata document is the scope of the collection which the document describes. Collections are arranged and organized on multiple bases, such as taxa, projects, collector, ecosystem etc. With such complexity, decisions about whether to describe collections on the basis of taxa, size, or any other criterion will determine the usefulness of the metadata document. Further, inclusion of both digital and non-digital specimens in the same document adds to the challenge.

METADATA: CRITERIA FOR DETERMINING SCOPE OF METADATA DOCUMENTS

Authoring a metadata document that describes a collection adequately, resulting in sufficient exposure, visibility, renewed interest and increased support for digitization is a challenge. Therefore, determining the scope of the metadata document is essential. Some of the frequently asked questions are, (a) whether a single metadata document could be good enough to describe the collection, (b) how lengthy or detailed a metadata document should be, and (c) to what level of granularity / depth it should collate the details.

We believe that answers to these questions largely depend on the answers to the following questions (1) how big is the collection? (2) what human resource capacity is available to do the metadata authoring? (3) how you would like to project the collection and (4) who is the target audience?

The following criteria should be used to determine the scope of the metadata document, and for deciding whether single or multiple metadata documents will best describe the collection (Table 1).

TABLE 1. CRITERIA FOR DETERMINING SCOPE OF THE METADATA DOCUMENTS.

Aspects	Scale	Issues to Consider
Taxon	Family or lower taxa	Size of the collection Age of collection Level of curation / digitization Complexity and diversity of the level of metadata record
Geographic scope	Country or Ocean / Seas Province / State Biogeographic regions	Political boundaries v/s bioregions Ease of management, and organization of collection
Projects	Individual project	Complexity in collection management (which is

	Expeditions or Cruise	often taxon based, and projects / expeditions often cut across taxa) A collection from an expedition or cruise may be deposited across multiple institutions / countries Temporal scale
Collector	One record per collector or group of collectors	Homogeneity vs heterogeneity Complexity in collection management (which is often taxon based, and projects / expeditions which encompass many taxa) A collection from an expedition or cruise may be deposited across multiple institutions / countries Temporal scale
Size of the collection	Multiple metadata documents will describe extensive collections better	<1000 specimens – single metadata document >1000 specimens – multiple metadata documents

As a general principle large collections will require multiple metadata documents. Furthermore, large collections may consider a taxon or region specific approach for collating metadata documents, whereas in small collections the author may employ a project or collector specific approach. However, the decision for adopting a specific criterion or combination of criteria is influenced by multiple factors.

WHAT METADATA ELEMENTS ARE ESSENTIAL?

On the basis of our assessment of the central question as to what will describe the collections best, we suggest that details about the following elements must constitute the core component of the metadata document:

1. List of taxa – preferably to the level of family, but in the case of insects or invertebrates this could be to a higher taxon level (e.g. Class or Order) (low granularity) and where possible, also to a lower taxon level (Family, Subfamily, Tribe) (higher granularity).
2. List of regions – preferably include biogeographical regions as it would enhance the use of metadata.
3. Temporal scale – granularity depends on the size of the collection and temporal range of collection events (e.g. from 1990-2000).
4. An estimate of the size of the collection - i.e. specify by order of magnitude of 100s, 1000s, or 10000s) (e.g. approx 1000-2000 specimens).
5. State of accession or curation - e.g. state if the collection is sorted and pinned or not sorted yet, and whether the collection is accessioned into a catalogue book.
6. State of digitization – metadata, extent of digitization (e.g. %), detail of data captured (e.g., taxonomic details only, or locality data, collection data, imaging of each specimen or % of specimens).
7. Type status – How many type specimens v/s non-type specimens.
8. Persistent Identifier (i.e. a unique number or code that unanimously identifies the record) for collection, curator and metadata record itself. Interlinking between these Persistent Identifiers is crucial for easy and efficient discovery.
9. Special significance – e.g. historical or social (productivity and public health), economic or environmental significance of collection.
10. Collection risk assessment: level and description of the potential risk to the collection and reasons for such a risk.

METADATA: IMPLEMENTATION APPROACH

We recommend the following step-wise approach for constructing a metadata document:

1. Work at curation or collection manager level
2. Adopt a hierarchical approach:
   a. Taxa (higher to lower taxonomic levels)
   b. Bioregions (larger to smaller regions)
3. If tackling digitization from a low base (where few details are available) a few metadata records should be created to describe the collection on a large scale to achieve data exposure. The next priority would be digitization of the top priority elements of the collection. As digitization proceeds finer level metadata records should be created :

e.g. (fictional example) Metadata record 1: For the entire Australian Museum collection – we have a faunal collection > 16 million specimens from Australia and the Indo-Pacific from 1800's to present in various stages of curation and digitization.

Metadata Record 2: Mollusc collection (~50,000 specimens) from Indo-Pacific from 1850-2000.

Metadata Record 3: Create a metadata record for each of the 10 major families in the mollusc collection.

METADATA: EXEMPLAR USE CASES

John studies the impact of Climate Change on Amphibians in Madagascar. He searches on the GBIF data portal and other amphibian specific portals, which results in 2000 data records. A search on GBIF Data Portal leads to 20 metadata records with 20000 specimens of which 18000 specimens are not digital. A scan through 20 metadata records reveals that there are an additional 4000 Madagascan specimens in eight museums not digitized. John approaches the curators, and in the following month he has an additional 2500 records for analysis.

CONCLUSIONS AND FUTURE WORK

A demand-driven approach is considered to be the most successful approach to the daunting task of digitizing the data of the world's natural history collections. However, we currently lack best practice guidelines on how to develop demand-driven strategies and action plans for digitization of natural history collections data. In the near future GBIF, together with professional societies, needs to develop such guidelines. The use of metadata will expose the data to stakeholders and increase the resources available for data capture. A hierarchical and prioritized approach is recommended for the creation of metadata records. Institutions and GBIF should develop guidelines and plans for the digitization of collections including the use of metadata.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGEMENTS

Penny Berents is grateful to the Australian Museum for support to work on this project. Michelle Hamer is grateful to the South African National Biodiversity Institute. Vishwas Chavan is grateful to the Global Biodiversity Information Facility.

REFERENCES

Berendsohn, W. G., and P. Seltmann. 2010. Using geographical and taxonomic metadata to set priorities in specimen digitization. Biodiversity Informatics 7: 120-129.

Chapman, A.D. 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. (Accessed September 20, 2010).

Frazier, C. K., Wall, J., and Grant, S. 2008. Initiating a natural history collections digitization project, version 1.0. Copenhagen: Global Biodiversity Information Facility. 75 pp. ISBN: 87-92020-05-4 (PDF: http://www.gbif.org/) Accessible at http://www2.gbif.org/Digitization.pdf. (Accessed September 20, 2010).

GBIF 2008a. Terms of Reference for "Task Group on a Global Strategy and Action Plan for the Mobilisation of Natural History Data"[1]. Accessible at http://tinyurl.com/gsaptg. (Accessed September 20, 2010).

GBIF, 2008b. GBIF Training Manual 1: Digitization of Natural History Collections Data, version 1.0. Copenhagen: Global Biodiversity Information Facility.

[1] http://tinyurl.com/gsaptg

ISBN 87-92020-07-0. Accessible at http://www.gbif.org. (Accessed September 20, 2010).

National Science and Technology Council, Committee on Science, Interagency Working Group on Scientific Collections, 2009. Scientific Collections: Mission-Critical Infrastructure of Federal Science Agencies. Office of Science and Technology Policy, Washington, DC. Accessible at http://www.whitehouse.gov/sites/default/files/sci-collections-report-2009-rev2.pdf. (Accessed September 20, 2010).

Pyke, G.H. and Ehrlich, P.R. 2010. Biological collections and ecological/environmental research: a review, some observations and a look to the future. Biogical Reviews., 85: 247 – 266.

Scoble, M. J., and T. Bourgoin. 2010. Natural history collections digitization: rationale and value. Biodiversity Informatics 7: 77-80.

Tann, J., Kelly, L., and Flemons, P. 2008. Atlas of Living Australia – User Needs Analysis. (User needs analysis report | Atlas of Living Australia). Published electronically at: http://www.ala.org.au/documents/user-needs-analysis-report.html. (Accessed September 20, 2010).

Vollmar, A., Macklin, J. A., and Ford, L.S. 2010. Natural history specimen digitization: challenges and concerns, J. Biodiversity Informatics 7: 93-112.