Biodiversity Informatics, 7, 2010, pp. 77 – 80.

NATURAL HISTORY COLLECTIONS DIGITIZATION: RATIONALE AND VALUE

MALCOLM J. SCOBLE
Department of Entomology, Natural History Museum, Cromwell Road, London SW7 5BD, UK, m.scoble@nhm.ac.uk

Abstract. – The value of digitizing Natural History Collections is well attested, although their rate of digitization should be increased. The task group of the Global Strategy and Action Plan for the Digitization of Natural History Collections agreed that the only way of achieving a significant increase is by capturing metadata to encourage digitization at the specimen level. Encouraging a metadata solution appears to be the best way of mobilizing the community responsible for caring for and providing access to such data. Moreover, a user-driven approach is likely to offer the best means of prioritizing what should be digitized.

Key words. – Natural History Collections; digitization, museums, herbariums, metadata, GBIF.

Questions about the natural world may be addressed by natural science collections, even if they comprise a small component of a more extensive source of data. Although there may be as many as 2-3 billion specimens in natural science collections across the world (Duckworth, Genoways and Rose 1993; Ariño 2010), the amount of data is not large compared with the vast and increasing number of digital observations produced and used by monitoring and other projects. Although data in collections are complementary to these other data, they offer, when digitized, an exceptional resource. Within collections lies the most extensive dataset that exists of the planet's biodiversity – a dataset that is relevant to research and decision-making. Unlike observational data, which are restricted to relatively few of the estimated 2 million described species, collections hold a recoverable record of what species exist, or have existed over the past three hundred years (to a time even before Linnaeus), and where they occur or occurred. This record is usually biased: so far, collections are anything but a representative sample of life on earth, although they provide, at least, minimum estimates. Indeed, our entire knowledge of most species is based on just one or very few specimens housed in natural science collections. Nevertheless, for most species, collections provide us with the best public record available. And, unlike observational data, the physical presence of specimens allows us to examine them many times using new techniques. Examples include the extraction and study of molecular data, although many museum curators restrict destructive use of specimens. The recent push for stable isotopes, DNA and fatty acid analysis presents a good example.

Collections are physical databases of the natural world. The specimens they house contain a wealth of taxonomic, spatial and temporal data, albeit with much variation in detail, quality and coverage. The problem is that most of this information is trapped in various museums, herbariums and private holdings. While it is accessible to *bona fide* researchers who have the means to visit collections, few outside the discipline of taxonomy have made much use of it. Recently, modern methods have given us the capability to capture digital information from collections and expose it globally through computerized networks via a web interface - either as data associated directly with individual specimens or as metadata describing collections.

Natural science collections have been used mostly for taxonomy (or systematics), the discipline associated with inventorying the planet's biodiversity and describing its evolutionary relationships. These will remain

primary tasks, although doubtless one that will occur increasingly online in a collaborative virtual environment. With the means of digitization at our disposal, however, we might not only help improve the species inventory but also realize a far wider potential of biological collections. Such data could help researchers address questions on natural resource inventories, the effect of environmental change on biodiversity, on how to gain a better understanding of species distribution over time and why changes in species' distributions have occurred. Stated more broadly, specimen information in collections still has the great potential to be used in research, resource management, education and sustainability science. Digitization priorities should, therefore, be set with the wider user community in mind. Yet for the most part they have not, although there are notable exceptions where much taxonomy-related information is available (e.g. in FishBase, VertNet, GBIF).

While digitizing material in collections has a wider value, the effort expended will have to be proved cost-effective against other demands on those who digitize. This question is relevant at any time, but particularly so in a straightened economic environment. Yet, there are well documented and compelling case studies of the use of information from natural history collections, particularly when integrated with data from research programs especially in the fields of biogeography, ecology and evolution (Graham et al. 2004). Clearly, both volume and quality of data are critical factors, which means that if digitization is to be achieved on a more comprehensive scale, a shift in the working patterns and current aims of curators and others managing collections will be required. In particular, effort will need to be prioritized, focused and sustained.

THE VALUE OF DIGITIZATION

Locked up in collections, is information potentially relevant to broad questions that require information about species diversity and species distribution and their change through time. Much information is available already from the Global Biodiversity Data Portal (http://data.gbif.org), but much more resides undigitized in collections. There is more to do in the mobilization of primary species data in collections in terms of encouraging further digitization and prioritizing what should be digitized. But collection managers should take heart from encouraging noises being made about the value of natural history collections to address a variety of questions.

One of the most comprehensive accounts of the value of digitizing specimens in collections was written by Chapman (2005), in a paper on the uses of species-occurrence data. This was preceded by a shorter review by Graham *et al.* (2004). By providing a series of examples, Chapman examined the uses both of specimen and observation data to a very wide range of fields from taxonomy (the traditional use), through biogeography, species diversity and invasive species, to education, and art and history. The value of specimen data in collections for addressing these questions varies considerably, with those from modern surveys having a greater variety of uses, primarily because of the higher quality of information associated with more recent specimens. It is crucial that appropriate metadata are collected to enable the widest use of this information. Chapman also addressed the criticism that museum specimen data were outdated and unreliable, explaining that while some undoubtedly are, many are not only usable but can be improved by rendering them accessible across digital networks. A compelling case for the use of natural history collection data in modeling, alone or, particularly, combined with other types of data, was made by Graham et al. (2004). These authors noted, for example, studies demonstrating the value of such data in predicting future distribution of invasive species.

SCALING UP

The response of curators to digitizing natural science collections has been rather more technology driven than strategically planned. But many curators and managers have adopted new technology to digitize specimen data opportunistically, often without any obvious purpose. Where digitization has been more purposeful, it has been developed largely for the close community of taxonomists rather than the wider group of stakeholders and global users. While there are exceptions to this statement, large-scale digitization is more likely to succeed if taxonomists are familiar with the major

applications and if they forge partnerships with users and align their digital outputs with producers of other kinds of digital data.

If a functional global infrastructure for biological collections is to be achieved, a responsiveness is needed that contributes to and helps those working in domains other than just in natural history collections. Routine digitization can certainly contribute towards building capacity, even if in a limited way. For example, curators now frequently digitize label data and make images of specimens that can be sent to borrowers *in lieu* of the actual specimens, or they may lend the specimens while keeping the digital data as security. Activities of this kind can form valuable contributions to mobilizing the digitization of natural history collections, but it is a relatively slow way of building a critical mass of digital objects.

Automated or semi-automated digitization is probably the only way in which digitization at the specimen level will be achieved at anything like the scale needed if data are to be useful for more quantitative studies. Some kinds of specimens are intrinsically easier to subject to such an approach. Herbarium sheets are probably the best example, having the advantage of being mounted flat with associated data written on labels attached to the sheet in the same plane. There are many examples of herbarium sheets being digitized as major scanning programs, in the Botanischer Garten und Botanisches Museum (Berlin), Kew Gardens (London), and many other herbariums. Specimens on microscope slides share with herbarium sheets similar characteristics. By contrast, digitization of label data on dried insect specimens poses an immense challenge given that labels are attached, often as a series, on the same pin as the specimen and underneath it. This arrangement renders it impossible to photograph drawers of specimens together with the associated specimen label data. Solutions, which appear promising, are being sought for specimens of all kinds housed in drawers (Blagoderov et al., 2010).

Digitization of natural science collections (images, information or digital surrogates) has often been undertaken because it is worthy and increasingly expected rather than overtly targeted to specific uses. Although further digitization of collections (whether at specimen or metadata level) will almost certainly lead to uses as yet unanticipated, particularly when a critical mass of data becomes available, a more strategic approach to digitization will be developed better by partnering with users (particularly ecologists). Creating partnerships is more beneficial than simply anticipating user needs. Progress might be made by mobilizing the user community to fund digitization for specific purposes and to use offers of funding to prioritize, whether it be for a specific project or user community (e.g. specific taxa across many collections, or all collections from one area – for example for data repatriation).

THE METADATA APPROACH

A metadata approach seems the most expeditious solution to the challenge of digitizing collections. Berendsohn and Seltmann (2010) state that capturing metadata is the only realistic solution to providing the scale of digitization across all kinds of collections that will mobilize the data gathering in a timely way. Capturing specimen-level data from all collections is simply not possible in the short or even medium term with the kind of resources available to the collections community. This statement was confirmed by a survey of 228 respondents to the task group of the Global Strategy and Action Plan for the Digitization of Natural History Collections. It is certainly not meant to suggest that the prioritized capture of specimen-level data should not be undertaken, for there are certain kinds of collections that are eminently capable of being digitized at scale (herbarium sheets being particularly tractable as already noted) or that have been digitized on a large scale already (e.g. the US vertebrate collections). But providing metadata is a means of providing the community with an understanding of what is potentially available at the specimen level across the universe of biological collections.

Such metadata might include the number of specimens an organization holds for a particular taxon and the country or countries of origin. This approach is demanding enough in its own right, but if collections-rich organizations are resolved to digitize their holdings, this method would provide a realistic and cost-effective initial solution to the problem of scaling up the digitization of collections. An important proviso is that the

providers should aim for high quality metadata (see e.g. Global Biodiversity Information Facility 2008). The main conclusion of the task group was that metadata records describing a collection are an essential and achievable prerequisite for meeting user demands and best practices, but that this process could and should lead to prioritized specimen-level digitization.

A metadata approach to the digitization of collections will in itself help achieve a number of ends. Notably, it will facilitate the finding and using of data, identify gaps in coverage of our samples of biodiversity, help expose errors and other shortcomings of data quality, improve quality and peer-review, provide data on density of sampling to assess their suitability for analytical work, accelerate data capture, and improve the management and enhancement of collections.

Achieving even this metadata solution is an immense task and GBIF has key roles to play as a facilitator for the international community of collection holders and as a data broker. It will require resolve and persistence to keep the international collections community engaged actively in digitization, a process that has hardly started in terms of what needs to be achieved to form a critical mass of information. It will also require leadership in acting as a forum for the debate about how to prioritize what to digitize. Production of metadata about collections, however, will help inform the public and scientists alike.

ACKNOWLEDGEMENTS

We thank three referees for thoughtful comments.

REFERENCES

Ariño, A. 2010. Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7: 81-92.

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

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

Chapman, A. D. 2005. Uses of Primary Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Duckworth, W.D., Genoways, H.H. & Rose, C.L. 1993. Preserving natural science collections: chronicle of our environmental heritage. Washington, D.C.: iii+140 pp.

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

Graham, C. H., Ferrier, S., Huettmann, F., Moritz, C. and Peterson A.T. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19: 497-503.