Global Strategy and Action Plan for the Digitisation of Natural History Collections

GBIF and Specimen Information: the rationale

Mobilising the biodiversity information intrinsic to the specimen holdings of natural history museums and herbaria of the world was one of the core aims of establishing the Global Biodiversity Information Facility[1] and has been an integral part of GBIF's DIGIT work programme ever since.[2]

GBIF has now made accessible on-line more than 150 million primary biodiversity data[3] records, about 40% of which are specimen data. In contrast to observation data, specimen data comprise a much wider temporal range, with many collection events dating back 200 years or more. As a consequence, most specimen data are not initially available in digital form. However, they are, for example, of prime interest in documenting climate change events of the last 2 centuries at a wide range of scales. The approximately 60 Million specimen records now accessible through the GBIF infrastructure are thought to represent a large proportion of the data available in digital form, especially if historical specimens are considered. Effectively, the "low hanging fruit" has been picked. An action plan is needed to implement a further mobilisation of specimen data. This is especially urgent for historical data, derived from specimens assembled from the biodiversity rich countries of the world and preserved in institutions often situated in temperate regions.

There are several billions of natural history specimens. However desirable, digitization of all specimens across the globe is a noble but impracticable goal. Digitization will need to be carefully prioritized to have the maximum impact in the shortest time frame.

Impediments and Need for Global Action

There are three main obstacles to increasing the rate of digitisation and the impact of specimen data. First, digitisation is a costly and labour-intensive process. Second, although innovative ideas abound, there is a marked lack of coordination, coherence and encouragement for the ongoing digitisation efforts in collection institutions. Third, there is no mechanism to globally request information about relevant holdings of collection institutions nor to answer such requests, and the purpose of specimens digitisation is thus not widely appreciated by the wider user community.

This situation calls for guidance and for a general strategy to make critical specimen information universally available. GBIF constituted a Task Group of experts in the field[4],

---

[1] Final Report OECD Megascience Forum Working Group on Biological Informatics. OECD, Paris, Jan. 1999.

[2] Of course, digitisation of specimens has not been GBIF's only activity, observation records, the taxonomic backbone, and organising the community were other important tasks beside the enormous effort spent on rolling out the IT infrastructure. Within its new work plan GBIF plans to reach out to further communities and by initiating several GSAPs for mobilising different types of primary biodiversity data.

[3] Together with observation data (e.g. from floristic and faunistic mapping projects, nature watchers, bird ringers etc.), which are similarly centred around a specific organism found at a specific time in a specific place, specimen data are now known as primary biodiversity data, as opposed to secondary data such as species descriptions and taxonomic hierarchies, which represent a syntheses or hypothesis based on primary data.

[4] The GBIF Task Group on the Strategy and Action Plan for the Digitisation of Natural History Collections includes: Dr. Arturo Ariño, University of Navarra, Pamplona, Spain; Roger Baird, Canadian Museum of Nature, Ottawa, Canada; Dr. Walter Berendsohn, Botanic Garden and Botanical Museum Berlin-Dahlem, Germany (Chair);Dr. Penny Berents, Australian Museum, Sydney, Australia; Dr. Thierry Bourgoin, Museum National d'Histoire Naturelle, Paris, France; Dr. Michelle Hamer, University of Kwazulu-Natal, South Africa; Dr. Tsuyoshi Hosoya, The National Museum of Nature and Science, Tsukuba, Japan; Dr.

with the aim of developing a draft action plan that is relevant at the global level and informs regional and national action plans for the digitisation of natural history collections. The draft plan will be developed in consultation with stakeholder communities in and beyond the GBIF sphere, such as GBIF Nodes, the Clearing House Mechanism focal points, organisations of collection-holding institutions, and societies in the field, as well as with the OECD Global Science Forum's activities in the field. The consultation process itself will be mediated by the GBIF Secretariat. Implementation of the plan relies on the capacity of the institutions holding collections to participate in a globally coordinated approach. It should be noted, however, that the global action plan is not intended to replace any current ongoing digitisation activities, but to complement and help prioritise them.

The strategy

The development of the plan will be guided by a single basic strategic principle: **user demand will be the driver of the detailed digitisation of individual specimens.** Accordingly, priorities for digitisation should be set either according to the demand from ongoing or projected research, or in accordance with socio-political demands (Conventions etc.). Funding of digitisation activities should be linked directly to these priorities, i.e., the costs are either to be incorporated into research proposals, or covered by (international) organisations, foundations, or governments.

To make this possible, collection-holding institutions will need to co-ordinate their efforts on a national and regional, if not global level, and agree to implement the necessary mechanisms.

Request to the GBIF Governing Board

This document is the result of a preparatory meeting of the Task Group in Copenhagen, in October 2008. A preliminary list of activities has been elaborated, a timeline drafted, and resource requirements are proposed (see annex). Within the time period of a mandate running until February 2010, the Task Group will produce a report to the Governing Board as well as various documents supporting the development of regional and national activities.

GBIF participants are asked to:

- comment on the strategic approach outlined
- actively support participation the ensuing consultation process
- support the research activities outlined in the annex, either by means of supplementary funds or by in-kind staff support, and
- support the coordination of collection institutions at national and regional levels.

November 23, 2008

Sven Kullander, Swedish Museum of Natural History, Stockholm, Sweden; Dr. James Macklin, Harvard University Herbaria, Cambridge, USA; Dr. M. Sanjappa, Botanical Survey of India, Kolkatta, India; Ángela Suárez-Mayorga, Alexander von Humboldt Biodiversity Research Institute, Colombia; Dr. Malcolm Scoble, Natural History Museum, London, UK. Liaison with GBIF Secretariat: Vishwas Chavan, (GBIF Secretariat, Copenhagen, Senior Program Officer for DIGIT).

Annex: Preliminary workplan for task group coordinated activities

1. Rationale for digitising natural history collections

Unify the available information using a Wiki-based communication process. Related questions include:

- Who are the direct users of data / metadata? (Report underpinned by a list of publications and reports based on specimen data, showing range of applications.)
- Who are the end-result beneficiaries?
- What previously unknown inherent values of collections have been uncovered by digitisation?
- What risks of specimens loss exist that can be (partially) mitigated through digital preservation methods?

2. Estimating the size of the universe of collection data

Unify the available information using a Wiki-based communication process and to initiate further research. Related questions include:

- How to estimate the number of undigitised collections and specimens? Can these figures be inferred soundly from existing data or known models?
- How many collection institutions are known to have digitization activities, and to what extent?
- How many specimens have actually been digitized, and what fraction of accessible/available/existing data in known collections or taxa do they represent?
- Are there discoverable patterns in the set of digitised data that allow for gap analysis and resource discovery?

3. Identify barriers to the digitisation of specimen information

Establish a liaison with the CollectionsWeb Research Coordination Network (www.collectionsweb.org) to identify infrastructural, human, social, legal, and political barriers.

4. Categorise possible priorities for digitisation and define mechanisms to serve them

Two categories can be distinguished that relate to the current capabilities of access to non-digitised specimens in the collections:

- **Metadata-enabled priorities.** This includes all digitisation demands that today cannot readily be served (or even assessed) by collections, such as the request to digitise all material from a certain geographic area.
- **Prioritised taxa** (e.g. invasives, cultivated, rare or endangered, of legislative relevance, ...). These requests can readily be serviced by collections, since most of them are ordered by taxonomic criteria.

In both cases, there is no mechanism to request globally information about relevant holdings of collection institutions nor to answer such requests.

A second set of priorities result from properties of the specimens or collections themselves, e.g.

- **Collection-related priorities** (collections which are endangered due to lack of curation or inadequate management; exceptionally valuable parts of the collection which cannot be made available other than in digital form, ...)
- **Quality-related priorities** (specimen records from digitised taxonomic literature, highly curated subcollections, exsiccata, etc.)

Both categories need a coordinated approach by collection institutions.

5. Develop a metadata-based approach to digitisation activities

Collection data can be considered at three levels: (i) detailed specimen data, (ii) metadata about specimens in a collection (e.g. “400 specimens of Coleoptera from Brazil”), and (iii) collection-level metadata (institutional, personnel etc.). The third level is not within the scope of this group, but the Biodiversity Collections Index will be necessary as a framework to associate the specimen metadata with the site where the specimen is housed.

The group will look closely at specimen metadata as a means of resource discovery (a “finding aid”), (i) from the collection’s perspective (what can be provided at reasonable cost?), (ii) from the researchers point of view (by looking at studies that have used or are using specimen data), and (iii) on a political level (e.g. country, mainly by liaison with focal points etc.). The costs related to the mobilisation of metadata will be investigated, taking existing sources into consideration. This will depend on an effective scheme to categorise (sub-)collections, because costs will be strongly influenced by factors such as taxonomic scope, management regime, and physical location.

6. State of the art of digitising detailed specimen data

Building on the substantial effort that GBIF has already made on developing and documenting 'best practice 'guidelines on digitization of collections (methods, quality, data cleaning, geo-referencing etc), the following questions need to be addressed with respect to mobilising the information contained in the specimens already present in collections:

- Which categories of specimens are needed to describe adequately digitisation processes? (E.g. flat sheets, pinned specimens, fluid-preserved, collection lots, ecological samples, specimen data from literature, specimens with data in ledgers, ...). This is one of the important factors determining digitisation costs.
- What are the concrete processes to complete the digitisation of each category of specimens? Create flowcharts to document the process and to link to available sources of information for specific parts of it. Related questions include the role of imaging in specific taxonomic groups, the role of other digital surrogates (x-rays, spectra), as well as the minimal requirements in terms of metadata items (finding aids).
- How can we rationalise digitisation, i. e. what mechanisms for 'industrialisation' of the process can be harnessed to reduce costs and time? What targets in the workflow lend themselves to rationalisation, what is the potential of automated procedures (e.g. image feature recognition, robotics, ...), can the general public be involved in data capture?
- Can a schema be developed to categorise the costs incurred by collections when digitising specimens on demand? The cost of digitising detailed specimen data depends on a number of factors, e.g. local labour costs, the selection of data items, the desired accuracy, the number of items to be digitised, the time span available to execute the task, etc. For a demand-driven digitisation process, collections need to specify costs in a transparent way, preferably according to a defined schema, which is designed around a “as-simple-as-possible” philosophy.

7. Changes required in the treatment of new accessions

How should collections adapt to the new information environment? What are the minimal requirements of data capture at the point of accessioning specimens? Models for effective information flows starting in the field? Can collections influence individuals (collectors) to partake in the process? How can processes be defined for today's collection, to avoid the backlog of specimen data tomorrow?

Resource requirements:

Apart from coordinative tasks, tackling the questions under 2, 5 and 6 above in the short time-period available to the task group will only be possible if additional funding is provided for research projects.[5]

1. Research Projects	€ 105,000.00
2. GSAP-NHC TG meeting(s)	€ 020,000.00
3. GSAP-NHC intern for 15 months	€ 020,000.00
4. In-country investment for National Action Plans	varies

Timeline for Task Group activities

24 October 2008: Task Group agreed on document for GB circulation

03 November 2008: Presentation of initial general strategy to GB15

26 Nov. 2008: Request for comment (RfC) circulated to heads of delegation

19 Dec 2008 Deadline for comments on first RfC documents

January 2008: Task group analyses outcome of first RfC and agrees on supplementary documents to be circulated for second RfC

January 2008: Research projects outlined and funding sources identified

May 2009: Circulation of discussion document to wider audience in second RfC

July 2009: Analysis of response to 2nd RfC

September 2009: Draft GSAP-NHC report circulated for discussion at GB16

October 2009: Presentation of draft GSAP-NHC report to GB16

February 2010: Final GSAP-NHC Report and group adjourns

[5] In the proposed Workplan for 2009/2010, the GBIF Secretariat has committed 40.000 Euro towards the execution of this workplan, 20.000 of which are earmarked for the intern at the Secretariat. The other 20,000 can help to initiate the research activities. In addition, there are 15.000 Euro available from the Freie Universität Berlin, BGBM, to continue research in the area of metadata-mediated access to collection information. We are looking for supplementary funds to support the research activities and the 2nd and last meeting of the Task Group.