Accelerating the discovery of biocollections data

Final task group report: November 2016

Leonard Krishtalka
Eduardo Dalcin
Shari Ellis
Jean Cossi Ganglo
Tsuyoshi Hosoya
Masanori Nakae
Ian Owens
Deborah Paul
Marc Pignal
Barbara Thiers

Suggested citation

Krishtalka L, Dalcin E, Ellis S, Ganglo JC, Hosoya T, Nakae M, Owens I, Paul D, Pignal M & Thiers B (2016) Accelerating the discovery of biocollections data. Copenhagen: GBIF Secretariat. Available online at: http://www.gbif.org/resource/83022.

Persistent URI

http://www.gbif.org/resource/83022

Task group coordination, GBIF Secretariat

Siro Masinde (smasinde@gbif.org), Programme Officer for Content Mobilization

Cover Image

'Diversity of orders' insect reference collection, Zoological Museum, Natural History Museum of Denmark, Copenhagen University by Kyle Copas. Photo 2016 licensed under CC BY 4.0.

Table of contents

Executive summary

Biocollections and their associated data document the life on our planet, past and present. They are fundamental for understanding, advancing and applying biodiversity science to the discovery of knowledge and advancing human well being. Nevertheless, the estimated 2.5–3 billion specimens of plants and animals in worldwide museums, herbaria and like institutions remain largely underutilized, as only ~10% of their associated data has been digitized for deployment by the educational, scientific and policy communities.

Biocollections institutions therefore face a challenging dilemma: how to prioritize and fund the digitization of massive amounts of data associated with billions of voucher specimens of animals, plants, fungi and other organisms that document the planet's biodiversity and undergird natural and human ecosystems.

To examine this issue, and as part of a broader global strategy for mobilizing primary biodiversity data, GBIF convened a task force (2015–2016) to help accelerate the discovery, digitization and access to biocollections data.

The task force's main operating principle was the value-chain of Data →Knowledge →Applications, i.e., when biocollections data is mobilized, analysed and converted by research into knowledge, the data will be highly valued and funded by diverse constituencies because the knowledge will inform solutions to current and future critical challenges of human, economic and environmental well-being.

The task force's main objectives were to:

1. Document best practices from ongoing content mobilization initiatives for small, large and different kinds of collections
2. Document successful business models for mobilizing resources for digitization
3. Consult with ongoing capacity-building and content-mobilization initiatives
4. Provide guidance in the development of training and outreach materials to help institutions and interested stakeholders to implement a metadata approach for content mobilization
5. Provide guidance on establishing priorities for digitizing biocollections in order to serve institutional, national and global needs and achieve the greatest economies of scale
6. Make recommendations for achieving long-term sustainability of mobilizing and providing access to biocollections data

The task force conducted a global survey of biocollections to determine and demonstrate the digital readiness of the world's biocollections and their institutions, as well as the realized benefits and impediments of digitization to the collection/institution. More than 800 responses from 2000 collections in 72 countries were received, with 76% at publicly funded institutions—40% at universities and 36% at non-university institutions. Key findings are encouraging. Digitization of biocollections data is an ongoing, valued enterprise in most of the world's museums and herbaria.

- 86% (615 respondents) are currently digitizing or have completed digitizing at least some or all of their collections. Among collection types, 13 of 15 are more than 50% digitized. Only 1% are not digitizing their collections and have no plans to do so.
- The major realized benefits of digitization are: increased use, exposure and knowledge of the institution's collections; more effective and efficient management and preservation of data and associated physical specimens; enhanced data quality; staff acquisition of new informatics skills.
- Major obstacles to digitization are: lack of funding, time, credit and/or expertise for digitization; not an institutional priority; data has errors; effort exceeds perceived benefit.

- The top three criteria for determining digitization priorities are research (53%), funding/grant opportunities (51%) and select taxa (42%).

Several institutions, organizations and projects, e.g. GBIF, iDigBio, VertNet, SPNHC, ALA, TDWG, Canadensys and so on, document community best practices for different aspects of data mobilization that address and remove perceived barriers to digitization. Among them are: setting smart digitization priorities, schedules and workflows; curating and cleaning data; adopting appropriate institutional policies and data-licensing practices to facilitate data dissemination and reuse; and data management and archiving.

Most digitization projects—more than 80% according to the survey—are government-funded and most often from one or two sources. Accordingly, institutions and digitization projects should diversify their funding sources to minimize the impact of potential funding cuts by governments and maximize investments in digitization. Some institutions and projects have achieved such diversification through partnerships with information technology companies. e.g. NHM London. Others have developed crowdsourcing programmes to engage citizens in transcribing specimen data labels, e.g. WeDigBio (international), Les Herbonautes at the Muséum national d'Histoire naturelle (France) and Naturalis' crowdsourcing projects, *Glashelder* (Dutch) and LiveScience (international). Governments, the private sector, foundations and citizen science programs are best engaged if the digitization priorities are demand-driven by the data imperatives of human, economic and environmental well-being that biocollections can inform (see Recommendations). At the country level, institutions should present and highlight the value of their collection-data digitization to governments and other entities as fulfilling the value-chain of Data →Knowledge →Applications.

Capturing and publishing collection metadata is a critical first step in exposing non-digitized collections and their value to global discovery and access. Metadata also provides a framework for institutions and biocollections to develop a comprehensive understanding of their holdings, a consequent, prioritized digitization plan and potential business-use cases to recruit research and resource partners.

Biocollections are therefore encouraged to adopt a tiered strategy for worldwide collections-data capture (and imaging where appropriate), i.e., a staged approach to digitization. Such an approach can start with less expensive but rapid steps to capture and share metadata about a collection-holding institution and an overview of its collections, then progress to more expensive and time consuming steps to capture and share specimen-level data and images at a finer granularity.

In the value-chain framework of Data →Knowledge →Applications, the task force's major recommendation is that institutions should set "demand-driven" digitization priorities to fulfil the first link in that value chain, i.e., data-to-knowledge. Specifically, this entails mobilizing and deploying the best biodiversity data to enable the best science for understanding and sustaining human systems and Earth's biological systems—and to do so in time to make a difference. Once demonstrated that collection data is essential for smart, science-driven solutions, data digitization—the first link in the value chain—will be valued and supported by entities that "demand" the second link, knowledge-to-applications.

It is clear that setting digitization priorities involves serving competing institutional, local, regional, national and global imperatives: individual research interests; institutional mandates; science agendas; and various environmental concerns (e.g. endangered species, invasives, disease vectors/hosts, pollinators, pests). Moreover, each imperative has its particular calculus of taxonomic groups, geographic areas, time periods and ecosystems/habitats. Overlying these permutations are the missions of different stakeholders and funders: intergovernmental bodies (e.g. IPBES, CBD), government agencies, NGOs, private foundations and corporations.

As such, the task force recommends that in a resource-limited world, a digitization strategy of maximum efficacy will require all parties to collaborate on setting demand-driven, overarching priorities that, simultaneously:

- Target the most urgent social, environmental, economic and biodiversity science imperatives of our time
- Are underpinned by sophisticated gap analyses
- Include the greatest commonality among competing imperatives and interests
- Tackle what is most pragmatic, first
- Promise the most immediate impacts

Such a strategic and collaborative approach can evolve the current cottage industry of biocollections digitization into an enterprise that is industrial strength, globally effective and efficient, and funded by consortia of entities that value the result: governmental science, health, natural resources and agricultural agencies; intergovernmental agencies; and NGOs, corporations and private foundations with missions in these and other sectors.

To enact this strategy, the task force recommends that GBIF and its partners convene a series of high level discussions among these constituents to fund and implement these five, long-term strategic priorities for mobilizing the remaining 90% of the world's biocollections data and bringing them into currency for science and society.

Background

The discovery and access to primary biodiversity data is critical for informed decision-making to achieve sustainable use of biotic resources and to address many of the world's key challenges, such as the impacts of climate change, invasive species, zoonotic disease outbreaks and food security. It is estimated that natural history collection institutions collectively house 2.5 to 3 billion specimens that document more than 300 years of the biological exploration of the Earth. Biocollections are the single largest source of information on biological diversity outside nature itself (Scoble 2010, Buerki & Baker 2015, Holmes et al. 2016). Physical specimens and their associated data constitute a vast biodiversity library of spatial and temporal occurrence of species, populations, individuals, and their morphology and genetic traits. From the study of fossil specimens tens of millions of years old to specimens of modern organisms, extant and historical collections enable reconstruction of past and present biodiversity and forecasting of future environmental states.

As such, these biocollections underpin much of our knowledge and ongoing research in biodiversity science, including irreplaceable documentation of evolutionary and ecological patterns and processes among fossil and recent organisms, as well as current threats to and conservation of biodiversity (Holmes et al. 2016). Despite this foundational importance, the world's biocollections remain largely untapped, as only about 10% have been digitized. Furthermore, only a small percentage of the digitized and imaged collections have been optimally mobilized to make them discoverable, accessible, interoperable and reusable.

In 2009, a survey carried out by GBIF, on the challenges and concerns related to digitizing natural history specimens, found that lack of funding was the overwhelming barrier, followed in descending order by lack of time and staff, lack of institutional support, infrastructural/technological constraints and challenges due to curation practices (Vollmar et al. 2010). The survey also found an uneven digitization landscape that led to a patchy accumulation of data at varying qualities and based on different priorities, ultimately influencing the fitness-for-use of the data.

In 2010, a GBIF Task Group on the 'Global Strategy and Action Plan for the mobilization of Natural History Collections data (GSAP-NHC) recommended the capture of essential metadata as a first step toward making non-digitized collections discoverable and accessible (Berendsohn et al. 2010). This metadata approach captures data on the different kinds of collections at various scales, from single units to large groupings.

Based on recommendations of the GSAP-NHC and our own deliberations, we summarize the value of capturing and sharing metadata for non-digitized collections as follows:

1. It is a rapid method of evaluating and assessing collections.
2. Sharing metadata makes collections discoverable and accessible.
3. It is a stepping stone towards making collections deliver immediate value and transforming them into a global resource.
4. It enables quick reporting on gaps such as taxonomic and geographic coverage, curation and physical state, and high-value series, among others.
5. It is a framework for helping prioritize a collection's digitization projects taxonomically, geographically or temporally.
6. It provides institutions the knowledge of their holdings required to build the business case for digitization.
7. It is a rapid and concise way of communicating and advertising an institution's collection holdings and potential, which can be key in attracting partnerships and the necessary funding for digitization.

Since the GSAP-NHC task group report was published in 2010, much progress has occurred, some of it in response to the recommendations of the task group. Following are some of the noteworthy developments.

- Mobilization of natural history collections has remained an important component of the GBIF strategy and the GBIF Secretariat's work programme. The same has been promoted through GBIF nodes and the entire GBIF community.
- A GBIF metadata profile incorporating elements to describe non-digitized collections was developed but it has not adequately served the intended purpose
- To encourage scholarly credit for metadata publishing, the concept of "Data Papers" was implemented and is now well established.
- To improve citation of and credit for data publishers, GBIF has implemented tools such as the digital object identifier (DOI), making all data downloads traceable.
- To facilitate online curation of specimens, GBIF plans to implement annotation tools and feedback mechanisms for data publishers.
- To recognize and give due credit to those involved in the enormous work of curating specimens and managing the associated digital data, a new joint RDA / TDWG working group on metadata standards for attribution of physical and digital collections stewardship (https://rd-alliance.org/group/metadata-standards-attribution-physical-and-digital-collections-stewardship/case-statement) is currently in place and expects to complete its work by the end of 2017.
- The number of specimen-based data records in GBIF.org has increased tremendously and currently stands at about 125 million. This has been accelerated by increased digitization in both small and large institutions worldwide as well as improved infrastructure for increased collaboration, sharing and management of NHC data. Examples among large projects include
    - The mass digitization effort at the Natural History Museum, London (Blagoderov et al. 2012)
    - The US iDigBio consortium (https://www.idigbio.org) funded through the Advancing the Digitization of Biological Collections (ADBC) programme of the US National Science Foundation
    - The Canadensys consortium (http://www.canadensys.net) in Canada
    - The Atlas of Living Australia (http://www.ala.org.au)
    - The Consortium of European Taxonomic Facilities (CETAF) (http://cetaf.org)
    - SYNTHESYS (http://www.synthesys.info)
- To accelerate data capture and citizen involvement, a number of GBIF nodes and other GBIF collaborators are implementing crowdsourcing programmes especially in transcribing specimen data labels, e.g. WeDigBio (international: https://www.wedigbio.org), Naturalis' Dutch crowdsourcing project, *Glashelder*, which uses *VeleHanden* application (https://velehanden.nl/projecten/bekijk/details/project/nat_nbc) as well as the more international counterpart, LiveScience (http://www.naturalis.nl/en/museum/livescience/crowd-sourcing), and Les Herbonautes (http://lesherbonautes.mnhn.fr) at the Muséum national d'Histoire naturelle (France).
- Much progress has been made in developing hardware, methods and tools to industrialize the digitization of NHCs. Examples include the Digistreet conveyer belt system employed by Naturalis Biodiversity Centre (Netherlands) for imaging herbarium specimens (Heerlien et al. 2015), and the Inselect tool—a modular, easy-to-use, cross-platform suite of open-source software tools that supports the semi-

automated processing of specimen images generated by natural history digitization programmes (Hudson et al. 2015).

- Commercial companies that can provide large scale, high quality digitization and imaging of biological collections as well as collection management software at reasonable cost have been outsourced for some of the mass digitization projects carried out at some of the very large museums. For example, Naturalis (Netherlands), Smithsonian Institution, and Muséum national d'Histoire naturelle (France) have used Picturae (https://picturae.com) to carry out digitization. Several natural history collections including Smithsonian Institution, New York Botanical Garden and Natural History Museum, London, use the Emu collection management software by Axiell (http://alm.axiell.com).
- An online metadata resource for biodiversity collections, the institutions that contain them and associated staff members, namely, the Global Registry of Biodiversity Repositories (GRBio: http://grbio.org), was set up (Schindel et al. 2016).

Importantly, what has also emerged is a tiered strategy for worldwide collections digitization (plus imaging where appropriate) and model concepts, such as Linked Open Data (LOD) (Berner-Lee 2009) and the *Digitization Maturity Model* in the ALA's Guide to Digitization (Kalm 2012). In this strategy, the tiers are:

a) *Level One*—Metadata I: Sharing and publishing Institution/Organization/Collection-level information. The who and where of a collection and, broadly, its content and history. Also, registering collections with GRBio supports the adoption of collections metadata standards and globally unique identifiers. Analogous to the Linked Open Data level one "on the web", this level is not costly or time-consuming, but is invaluable as it is key to a collection's *discoverability*.

b) *Level Two* - Metadata II (spindex, *sensu* Mason 2016). Producing and publishing species-level (or perhaps cabinet-level) collection inventories. Such collection inventories provide excellent data for tracking collection health, space requirements and gaps in taxonomic, geographic and/or temporal coverage and follow-on strategic planning for collection growth, conservation and digitization. For example, recently the Academy of Natural Sciences of Drexel (ANSP) provided a Species Index, or #spindex for short (Mason et al. 2016) as a basis for a strategic specimen digitization programme. Producing and sharing these in a standard, machine readable and non-proprietary format, analogous to LOD levels two and three, ensure global access to data for planning and funding initiatives.

c) *Level Three* – Specimen Data I (skeletal data). Capturing and publishing at least skeletal-level data (with or without locality georeferencing) and perhaps images for each specimen or lot, preceded by a strategically chosen level of pre-digitization curation (Nelson et al. 2012). Where possible, all shared data is mapped to currently accepted data standards, while advancing standards development as needed. Imperfect or erroneous data are then exposed to machine algorithms and the expertise in the worldwide community for effective, efficient correction and improvement.

d) *Level Four* – Specimen Data II (richer data). Locality data is georeferenced where possible and the specimen record is enriched with other data, such as: field notes, grey literature, note cards, etc. Field notebooks may be captured at either Level Three or Level Four—whichever is most appropriate for the particular collection digitization project.

e) *Level Five* – Specimen Data III (born digital data). The data associated with all new collections is "born digital" and incorporated into the existing, georeferenced database, with collection management documents, e.g. specimen labels, generated from the database. Digital records may also have links to GenBank accessions,

BCoL IDs, etc., and must include *globally unique identifiers*, which makes Linked Open Data a reality, a “web of data” (Berners-Lee 2009).

Ultimately, for global digitization, collections worldwide should collaborate (see Recommendations) to achieve critical mass, least redundancy and economies of scale in their digitization, and to meet demand-driven global imperatives that depend on collection data for solutions.

The GBIF Task Force

Objectives

As part of a broader global strategy for mobilizing primary biodiversity data, GBIF convened a task force to help accelerate the discovery and access to biocollections, especially those yet to be digitized. The task force commenced its work in March 2015 and runs up to the end of 2016. The objectives of the task force are to:

1. Document best practices from ongoing content mobilization initiatives, taking into account their applicability to large and small collections as well as different kinds of collections (e.g. wet, dry, mounted, pinned).
2. Document successful business models for mobilizing resources for digitization.
3. Consult with ongoing initiatives that target capacity building and content mobilization (e.g. SEPDD (http://www.sud-expert-plantes.ird.fr/sepDD), the GBIF BID programme (http://www.gbif.org/bid), ALA (http://www.ala.org.au), iDigBio (https://www.idigbio.org), specialist groups and the GBIF community in order to bring together the different stakeholders and catalyse activities around metadata capture.
4. Provide guidance in the development of training and outreach materials to help institutions and interested stakeholders to implement a metadata approach for content mobilization.
5. Provide guidance on setting priorities for digitizing biocollections to serve institutional, national and global needs and achieve the greatest economies of scale.

Task Force membership

The task force comprised eight members with diverse international experience and expertise along with three *ex officio* members. It consulted widely with stakeholders including experts, institutions, initiatives and projects as well as potential funders.

- Leonard Krishtalka (chair), Director, Biodiversity Institute, University of Kansas, USA
- Barbara Thiers, Director of the Herbarium and Vice President for Science Administration, New York Botanical Garden
- Deborah Paul, Digitization and Workforce Training Specialist, iDigBio–HUB, Florida State University, USA
- Eduardo Dalcin, Biodiversity Informatics Expert, Rio de Janeiro Botanic Gardens
- Ian Owens, Director of Science, Natural History Museum, London
- Jean Ganglo, Professor of Forestry, University of Abomey-Calavi, Lomé, Benin
- Marc Pignal, Musuém national d’Histoire naturelle, Paris
- Masanori Nakae, Curator, National Museum of Nature and Science, Tsukuba, Japan
- Shari Ellis, consultant to the task force and iDigBio External Evaluator
- Tsuyoshi Hosoya, Division Head, Fungi and Algae Research, National Museum of Nature and Science, Tsukuba, Japan
- Siro Masinde, Programme Officer for Content Mobilization, GBIF Secretariat

Operating vision

The task force adopted the Data-Knowledge-Application value chain framework as its operating vision. It is imperative to demonstrate how digitizing and sharing biocollections data contributes to this value chain, especially if all stakeholders are to be persuaded to contribute significant resources to accelerate the mobilization of biocollections data. The natural history collections community has made good progress in recent years in the first link in this value chain, mobilizing data through digitization and publishing to GBIF and other data portals, and converting that data to published knowledge through research, mostly in academic institutions.

However, the knowledge-to-application link in the value chain remains weak, despite the number of compelling, documented use cases. Two possible reasons among others are: (1) the number of institutions devoted to this link are much fewer, less networked and more poorly funded than institutions devoted to the data-to-knowledge link; and (2) the results of data-to-knowledge, published in professional journals are highly inaccessible to the governmental and NGO communities devoted to the knowledge-to-application component. Clearly, it is incumbent on the biocollections community, working with other communities across the value chain, to demonstrate the knowledge-application link if it is to rally the business community, major donors, funding agencies and governments to support and accelerate the digitization of primary specimen data.

Meetings and outreach activities

Task force members have held 16 monthly group meetings. Other meetings and consultations have been held in between as and when needed. Task force members have also participated at other conferences and workshops and continue to consult widely with stakeholders on pertinent issues, such as setting strategic institutional, national and global digitization priorities, as well as making compelling cases that can attract buy-in from potential funders in the public and private sectors. Table 1 summarizes the main meetings and outreach activities in which the members have participated.

Table 1. Meetings and outreach activities by GBIF task force members

Event and attendee's initials	Venue	Date
16 group meetings	Virtual	April 2015 - present
TF face-to-face meeting (LK, BT, DP, MN, ED, IO, SE, SM)	Washington DC	3 Nov 2015
GRBio outreach (BT)	Washington DC	27-28 April 2015
SPNHC (BT, DP, SE)	Gainesville, Florida	17-23 May 2015
Biodiversity Collections Network - BCoN (LK)	Field Mus., Chicago	1-2 Sep 2015
TDWG (LK, DP, JG, SM)	Nairobi	28 Sep-1 Oct 2015
iDigBio Summit V (LK, BT, DP, MN, ED, IO, SE, SM)	Washington DC	4-6 Nov 2015
Entomological Coll. Network - ECN presentation (DP)	Minneapolis	Dec 2015
Amer. Inst. Biol. Sc.-AIBS & BCoN, capacity bldg (BT)	Washington DC	Dec 2015
SPNHC TF symposium (LK, SM, DP, BT, MN, IO)	Berlin	20-24 June 2016
Entomological Coll. Network – ECN presentation (DP)	Orlando, Florida	23-24 Sep 2016

A global survey of natural history collections

Purpose of survey

In late 2015, the task force carried out a global survey on natural history collections. The purpose of the survey was to enable the task force to determine and demonstrate: (1) The digital readiness of the world's biocollections and their institutions; (2) the benefits to the collection/institution that digitization engenders; and (3) the impediments to collection data digitization.

Survey methodology

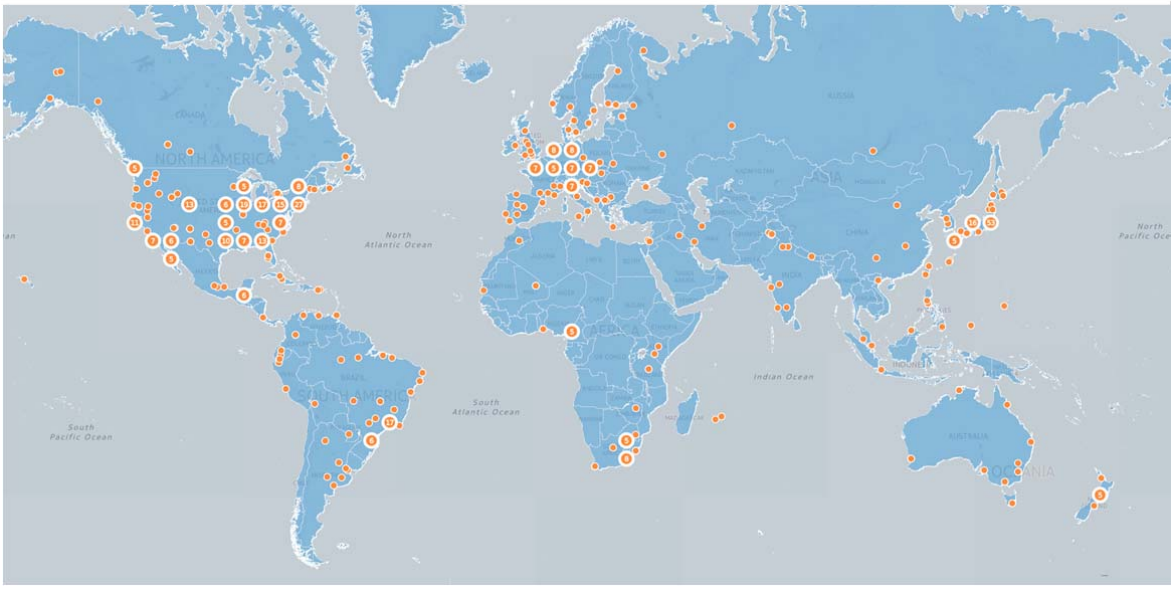
The survey questionnaire was prepared through meetings, consultations and research, and mainly administered online through Qualtrics software (http://www.qualtrics.com). A Microsoft Word version was however also made available for respondents that had difficulty in filling out the online survey. We distributed the survey to individuals affiliated with collections worldwide through a variety of channels including listservs and member lists such as Index Herbariorum and GBIF nodes, as well as personal contacts with institutions. We also announced the survey at relevant meetings, workshops and conferences attended by potential target respondents. The survey was translated into Japanese and distributed throughout Japan via the community of "Science Museum Network (S-Net)". Those contacted were requested to help with the further distribution of the survey in order to help maximize the global reach. Key distribution channels were:

- GBIF node managers
- GRBio
- Herbaria-L
- iDigBio
- Index Herbariorum
- MUSEUM-L
- NHColl
- SERNEC
- S-Net (http://science-net.kahaku.go.jp/)
- Taxacom
- TDWG
- Conferences including TDWG (Nairobi, 2015), iDigBio V Summit (Washington DC, 2015), Entomological Collections Network Conference (Minneapolis, 2015).

Summary of survey results

We present a summary of the survey results here, but a more detailed report (Annex 2) from a first analysis was distributed to the survey respondents that expressed interest in being contacted directly. More than 800 individuals completed at least a portion of the survey, of which 617 were complete enough to be counted. Note that the number of respondents who answered each question varied either because in the case of the online version, some questions were automatically skipped based on prior responses, or the respondents elected to not answer the question. Respondents represented almost 2000 collections distributed over 72 countries. Map 1 shows the locations of the survey respondents based on the registered IP addresses from where the online survey was completed. The distribution and density of respondents—and by extension the locations of natural history collections—mirrors that for the global distribution of the Internet infrastructure (http://internetcensus2012.bitbucket.org/images/worldmap_16to9) as well as the GBIF species occurrence map (http://www.gbif.org/occurrence).

Map 1. Location of survey respondents by IP address, CartoDB by Kevin Love (iDigBio)



This is the most comprehensive survey of NHCs that has been carried out in the past six years, as it reached all continents and achieved a large number of responses. A similar, 2009 survey (Vollmar et al. 2010), but of more limited scope elicited 201 responses, mostly from North America (62%) and Europe (22%), with none from Africa. Of the respondents in the current survey, 76% were at publicly funded institutions—40% at universities and 36% at non-university institutions. Almost all (92%) of respondents were primarily curators or collection managers with 10% as head of research and collections.

Key findings from the task force survey

- Of the usable 617 responses, 86% are currently digitizing or have completed digitizing some or all of their collections.
- 13 out of 15 collection types report more than 50% data capture.
- Very few respondents (1% or 5 individuals) reported they are not digitizing and have no plans to do so.
- Major obstacles to digitization were: funding, time (lack of), size of task, not an institutional priority, data has errors, limited expertise in databasing or processing specimen data, no credit (tenure, reappointment) for digitization effort, effort exceeds perceived payoff.
- Major priorities for collection digitization are research (53%), funding/grant opportunities (51%) and select taxa (42%).
- The major realized benefits of digitization are: increased use, exposure and knowledge of the institution's collections; more effective and efficient management and preservation of data and associated physical specimens; enhanced data quality; staff acquisition of new informatics skills.

Responses by taxonomic collections included vascular plants (20%), bryophytes (10%), fungi (10%), algae (9%), arthropods (8%), mammalogy (6%), ornithology (5%), herpetology (5%) and ichthyology (5%), with the remaining representing malacology, marine invertebrates, terrestrial invertebrates and fossil invertebrates, vertebrates and non-vascular plants.

It is clear that the value of databasing and publishing collection data is now embedded in the community—most collections are digitizing or trying to do so. To that point, 86% (615 respondents) indicated that they are *currently databasing some of or have completed*

databasing their collections. The percentage varies across collection type, as does the mean portion of the collection that has been digitized. But, on average, 13 of 15 taxonomic collection types are more than 50% databased. Also, across organismal groups, the average of all collection types digitized is more than 50%, except arthropods at 38% and invertebrate fossils at 47%.

Barriers to digitization were raised by 587 respondents, most citing either funding or time, which is similar to the findings reported by ITHAKA for digitization of special collections (Maron and Pickle 2013). The top 10 barriers to digitization, for our survey respondents (n=587), were:

1. Funding and other resources (80%)
2. Personnel time/effort (80%)
3. Task is overwhelming (40%)
4. Not an institutional priority (35%)
5. Collection data has errors (30%)
6. Limited digitization expertise among personnel (25%)
7. Insufficient information on digitization process (15%)
8. No benefit to job advancement, tenure (14%)
9. Not a priority of the individual in charge of the collection (12%)
10. Effort exceeds payoff (10%)

Virtually all of the cited barriers to digitization after the first two (funding; effort) fall into four categories:

1. Size of task is overwhelming.
2. Digitization is not institutional priority because of perceived mismatch between effort required and benefit achieved.
3. Sentiment that data cannot be digitized because of the number of errors
4. Personnel require greater experience and expertise in digitization workflows and applications

The TF determined *sharing collection metadata* can help overcome the major barrier of insufficient resources, as it makes an institution's collections discoverable to funding by various stakeholders and potential funders. For example, the British Library (BL) currently has more than 30 digitization efforts; all funded by various foundations and other entities with a vested interest in particular materials held by the BL (from a presentation by the British Library, at the Cisco Pitstop (Jackson 2016)).

How are collections deciding what to digitize?

With 519 respondents, the top three variables driving digitization priorities include: research (52%), funding/grant opportunities (51%) and taxonomic focus (42%). Other criteria include partnership in a larger community effort (24%) and a geographic focus (23%). About 18% cited opportunistic digitization and about 5% health and human services.

Who is doing the digitization?

Collections rely on a variety of personnel to perform digitization tasks, usually staff, students, volunteers and *rarely*, third-party organizations (2–10 %). Most often, digitization is being completed by paid personnel (90% respondents), as opposed to paid staff (53%) or students (59%; paid or unpaid) "frequently" doing the digitization.

How are collections funding digitization?

Of the 523 respondents to this question, 87% cited receipt of some funding, of which 69% reported external sources and 61% regular institutional sources. Slightly more than half (53%) cited only one source of funding (internal, external, ad hoc), whereas 36% received

funds from two sources and 30% from three. Most of the external funding (80%) is from government agencies, with 72% receiving funds from just one kind of external entity (i.e., industry, foundations, government, other).

Benefits of digitization

The top 8 perceived benefits of digitization (516 respondents) are: increased use of collections, increased exposure, better knowledge of holdings, better management of data, digital data preservation, enhanced data quality, new skills for staff and better management of physical specimens. Thirty percent reported new communities using the data, 35% saw increased publicity and reduced physical handling of the collection and 40% cited increased use of their collection data in research and publications, as well as increased public awareness of the importance of collections.

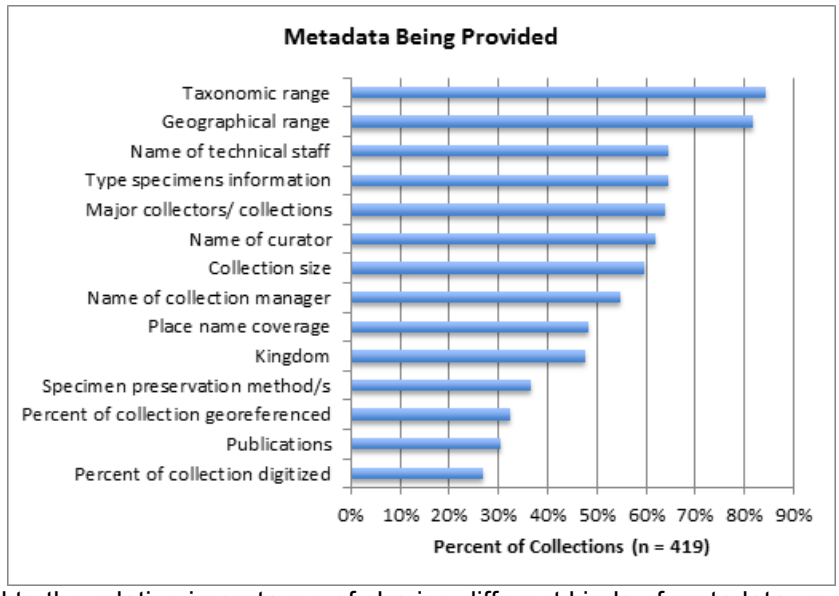
Institutional commitment

Of the 516 respondents, 72% confirmed their institution's commitment to continued digitization. Of 422 respondents, 50% reported that their institutions were providing staff training and seeking continued funding. More than 60% indicated plans for long term archival storage and more than 70% for long-term data curation.

Metadata

What collection metadata do we need as a community? The survey requested information on the kinds of metadata collections are currently providing and consider important (see Figure 1 for results).

Figure 1. Metadata being provided



With regard to the relative importance of sharing different kinds of metadata, more than 50% cited taxonomic and spatial data as critical, with 45% indicating type-specimen data as important. Metadata values ranked as critical by 20%–30% of respondents included: percent collection digitized, notable publications, percent georeferenced, place name coverage, name of collection manager, collection size and name of curator. Respondents (506) report most metadata being shared via GBIF (43%), Index Herbariorum (36%) and institutional websites (21%).

Long-term plans

Encouragingly, more than 80% of 513 respondents indicated their institution/organization intends to digitize their entire collection(s), and more than 30% that they plan to prioritize digitization in response to research needs/requests. About 12% will focus on strategically

important or unique collections, such as type specimens or endemic species. Others (about 12 %) cited new collections.

Selected use cases: Collection-based data informing solutions

Discovery and access to primary biodiversity data are indispensable in ensuring informed decision-making on human, environmental and economic well-being (Kremen et al. 2008; Gaiji et al. 2013; Peterson et al. 2015). The principle is that Digitally Accessible Knowledge (DAK), e.g. primary biodiversity data that is digital, published and therefore accessible worldwide, can be integrated into the broader global storehouse of biodiversity information (Sousa-Baena et al. 2014) for fulfilling the knowledge-to-application link of the value chain. The Society for the Preservation of Natural History Collections (SPNHC) outlines the following use case themes that illustrate how DAK of collection-based biodiversity informed critical solutions for science and society. We illustrate each theme with selected examples.

Public health: Zoonotic diseases and environmental contaminants

Perhaps the most compelling examples of the importance of natural history collections are in the area of public health and safety (Suarez & Tsutui 2004). In several important cases, collections have been used to track the history of infectious diseases, identify their sources or reservoirs and pinpoint geographic areas for potential interdiction.

Zika, dengue and chikungunya viruses

Zika, dengue and chikungunya are mosquito-borne diseases that have recently re-emerged in epidemic proportions across wide geographical areas hitherto not known to harbour them (Kraemer et al. 2015; Bogoch et al. 2016; Cao-Lormeau and Musso 2014). These diseases are spread by two mosquito species, *Aedes aegypti* and *Aedes albopictus.* Essential for health planning, is mapping the global distribution of these vectors using all known sources of data, including specimens in natural history collections that indicate the geographical determinants of their ranges. When Kraemer et al. (2015) mapped the global distribution of these mosquitoes; it showed that they are more widespread than previously known, thus predicting an increased risk of new infections in areas hitherto thought to be infection-free zones.

Hantavirus

In the 1990s, a Hantavirus (*Bunyaviridae*) caused a pulmonary infection that was fatal for most people who contracted it. Public health officials could not identify the vector of this disease until evidence of the virus was found in museum tissue collections of deer mice from the American Southwest (Yates 2002).

Ebola

The 2014 outbreak of the Ebola virus in West Africa was the largest of this deadly disease and the first in this region. A public health priority is the ability to identify the potential spread of this zoonotic disease and geographic areas at greatest future risk. Natural history collections are central to this task. Three different species of bats are suspected to play an important role in the life cycle of Ebola and similar viruses. Using data from natural history collections, researchers have determined the geographic range of these three species. Niche modelling of these data enabled researchers to pinpoint the geographic areas and communities at highest risk for future outbreaks across Central and West Africa. These data will help prioritize surveillance for Ebola outbreaks and improve the diagnostic capacity in these at-risk regions, which have a combined human population of 22 million people (Piggott et al. 2014).

Anthrax

In 2001, anthrax (*Bacillus anthracis*) was sent through the mail to media outlets and government officials in the U.S., resulting in five deaths. Researchers from the Centers for

Disease Control and Prevention (CDC) compared isolates from the 2001 anthrax attack in the United States with stored museum specimens to differentiate and identify the strain used in these attacks (Hoffmaster et al. 2002).

Lassa fever

Lassa fever (LF) is a zoonotic disease caused by Lassa virus (LASV), a member of the *Arenaviridae* family. Introduction of the virus into humans occurs through direct or indirect contact with excreta of the natural reservoir, the rodent *Mastomys natalensis*, although precise modes of transmission are not well characterized (Fichet-Calvet and Rogers 2009, Peterson et al. 2014). Peterson et al. (2014) applied ecological niche modelling, based on museum collection data, of the rodent's geographic occurrence to 107 data records of LF from an initial dataset of 111 records collected by Fichet-Calvet and Rogers (2009) in seven West African countries: Nigeria, Benin, Côte-d'Ivoire, Burkina Faso, Sierra Leone, Guinea and Liberia, all LF prevalence areas. Their results indicate that West Africa and particularly the southern humid forest habitats of the sub region as well as the drier Sahel zone of the continent are suitable for LASV transmission and therefore are at risk for LF. These areas are the priority for a health surveillance infrastructure so that health officials and decision-makers can detect and stem the spread of the disease. The study also revealed that additional data is needed from field and museum collections to add geographic and ecological resolution to the risk assessments in West Africa, including Benin, Togo and Ghana.

Environmental contaminants—Mercury, DDT, Atrazine

Environmental contamination affects human health as well as ecosystem health. Museum specimens have been used to track ambient mercury levels over time (Berg et al. 1966). In a classic study from the 1960s, eggs from museum collections were critical in establishing the link between the chlorinated hydrocarbons in DDT to the sharp decline in bird species. Specifically, measurement of bird eggs collected over a century demonstrated a marked decrease in shell thickness that coincided with the widespread use of DDT (Radcliffe 1967, Hickey and Anderson 1968). More recently, museum collections were used to demonstrate that sexual abnormalities in frogs increased after the widespread adoption and use of atrazine as an herbicide (Hayes et al. 2002).

Food security

Agricultural diseases

Phytophthora infestans, the cause of the potato blight that triggered famine in Ireland in the 1840s, continues to cause damage to potato fields around the world, resulting in huge losses annually to the world's third largest food crop. Yoshida et al. (2013) compared the genomes of herbarium specimens of *P. infestans* with that of modern strains and determined that outbreaks in the 19th century were caused by a single lineage, which is not the direct ancestor of the strains that have come to dominate more recent global populations.

Bioterrorism

Collections can be used to determine whether or not emerging pests have spread naturally, accidentally or deliberately by comparing specimens of pests across temporal and geographic ranges. Bioterrorism through deliberate introduction of agricultural pests has been identified as a threat by the US National Research Council, which has stressed the need for "reference specimens and other taxonomic information for rapid and accurate identification" of newly discovered pests (NRC 2003).

Invasive species

Cheatgrass

Invasive species in the U.S. cause environmental damage and loss totalling more than $130 billion per year. Cheatgrass, *Bromes tectorum,* one of the most damaging invasive species, crowds out wheat plantations and fodder crops. The grass has limited nutritional content and the long, sharply pointed fruit can penetrate the skin of livestock, causing injury or infection. Cheatgrass is native to Europe and Central Asia, where it does not exhibit the invasive tendencies that it does in North America. Comparison of genetic information between plants from the species' native range and those from historical and modern herbarium collections in North America (Novak & Mack, 2001), revealed that cheatgrass was introduced multiple times to North America from different areas of its native habitat. When these different strains came in contact with one another in North America, they interbred, creating novel strains with invasive qualities.

Argentine ant and green alga

Museum specimens have also been used to elucidate the invasion history of the Argentine ant (*Linepithema humile*) (Suarez et al. 2001) and the green alga *Codium fragile* (Provan et al. 2007).

Climate change impacts

Food stocks

Climate change is expected to cause dramatic shifts in the distribution of species, with serious implications for natural ecosystems, crop plants, their pollinators and food supply. Studies using museum collections demonstrate distributional shifts and extinctions in butterflies (Parmesan, 1996) and changes in nesting times in tree swallows (Dunn & Winkler, 1999). More recently, Jones et al. (2014) used GBIF's collection-based species occurrence data to model predicted changes in distribution of aquatic food species around Great Britain. The results project decreases in species diversity and catch weight, which will reduce the profitability of the fishing industry and threaten its decline. Accordingly, the authors recommend changes in the British fishing industry that may help to offset a future drop in revenue.

Bees

Over the past 40 years, drier weather has limited the growth of some populations of alpine plants in the Rocky Mountains of North America, making it more difficult for bees to obtain nectar. The paucity of nectar-producing flowers as well as the warmer temperatures favour bees with shorter tongues that can access a broader range of flowers. Miller-Strotman et al. (2015) documented this phenomenon by measuring the tongues of specimens of bees in museum collections—bees now inhabiting this region indeed had shorter tongues. If this trend continues, it could lead to the extinction of plants whose flowers can only be pollinated by bees with long tongues.

Baobab tree

The baobab, with more than 300 product uses and ensuing commercial value in the EU and US, is one of the most important trees to be conserved and domesticated in Africa, given the impact of this industry on African economies and livelihoods. Sanchez et al. (2011) using available DAK (480 records) on the African baobab tree (*Adansonia digitata*) and niche modelling, found that under IPCC scenarios of climate change only a percentage of the present distribution of the species in Africa will remain viable in the future.

Their results informed useful strategies for baobab conservation—*in-situ* in protected areas, *ex-situ* in seed banks and sustainable use of the species. The existence of only 480 records for African countries, of which less than 100 belong to West Africa, indicates that field and collection-data on baobab occurrences in West and East African countries need to be

digitized and published to the research community to increase the resolution of baobab distribution models and conservation strategies under different scenarios of climate change.

African palms

Palms (Arecaceae) are a multi-use resource for many African economies and communities but especially in West Africa. For example, they are the source of income through the palm oil and wine trades, local palm alcohol production and sales, and multiple uses of palm branches and leaves. Blach-Overgaard et al. (2010, 2015) applied ecological niche modelling on 1920 occurrence records of 29 palm species obtained mainly from herbarium specimens, to assess the degree to which African continental-scale palm species distributions are controlled by climate, non-climatic environmental factors such as habitat and human impact, or non-environmental spatial constraints such as biotic interactions and/or dispersal limitations. They found that, at the continental scale, climate, especially water-related factors, constitutes the only strong environmental control of palm species distributions in Africa. Furthermore, due to the strong response of palm distributions to climate in combination with the importance of non-environmental spatial constraints, African palms will be sensitive to future climate change in that their ability to track suitable climatic conditions will be spatially constrained. As with the baobab, the collection-based modelling studies of the African palms can inform *in-situ*, *ex-situ* and other species conservation measures.

Extinction lessons from deep time

Among the many studies based on paleontological and recent collections and their associated data are analyses published in *Nature* (Barnosky et al, 2011) and *Science* (Ceballos et al, 2015) indicating that the Earth's sixth mass extinction may well be underway at the present time, given the documented species losses over the past few centuries and millennia. Current extinction rates were shown to be higher than expected compared to those documented in the fossil record.

In another study (Barnosky 2008), analysis of megafaunal collection data and climate records revealed that an increase in human biomass and impacts, along with climate change were the fingerprints on the Quaternary megafaunal extinction and its subsequent ecological threshold event. Humans have since become the dominant ecological species which, with higher rates of climate change, will induce extinctions across taxa of all body sizes and possibly a near-future biomass crash that will have a severe impact on humans and their domesticates.

Habitat and species loss

The greatest threat to biodiversity and its contribution to ecosystem function is habitat loss (Millennium Ecosystem Assessment, 2005). Studies involving museum collections have successfully documented such shrinking habitats and the effects on their biodiversity. The loss of prairie habitat has led to the decline of its small mammals (Pergams and Nyberg 2001).

Based on a review of thousands of herbarium specimens of lichens, Lendemer and Allen (2014) identified a previously unknown biodiversity hotspot in the Mid-Atlantic Coastal Plain of North America. Projections expect this region, already under threat from encroaching development, will be completely inundated due to climate-induced sea level rise within the next century. Development pressures have seriously reduced the availability of corridors through which species in the areas affected by sea level rise could migrate to higher ground.

Endangered and threatened species

Since 2010, the Brazilian National Centre for Flora Conservation (CNCFlora) is responsible, at the national level, for assessing the conservation status of the Brazilian flora and

developing recovery plans for species threatened with extinction. CNCFlora is the Red List Authority for plants in Brazil and adopts the standards and procedures recommended by the International Union for the Conservation of Nature (IUCN). So far, CNCFlora has assessed the extinction risk of 5,165 species of the Brazilian flora (11.2% of the national flora). For this assessment, the CNCFlora has built up a database of species occurrences from two main sources: Rio de Janeiro Botanical Garden Virtual Herbarium and speciesLink.

The case of one species, *Abatia microphylla* (Family Salicaceae) reveals a risk assessment in which every herbarium occurrence record is critical. Initially, the evidence of only four valid records for two different localities caused this species to be listed as "Critically Endangered". A further query of GBIF's records for Brazil identified two new records of this species from one new locality, effectively an "Extension of Occurrence" of 273,8Km^2, and a revised listing of "Endangered".

The CNCFlora continues working on the assessment of all species of Brazilian flora—ca. 40,000 species—based on the best knowledge available. Whereas 80% of the Brazilian herbaria records are digitized, only ~20% are reliably georeferenced, effectively relegating 12% of the 5,165 assessed species to IUCN's "Data Deficient" category. Sousa-Baena et al. (2013) however demonstrated that about 40-54% of the 934 Data Deficient angiosperm species that were listed at the time had considerable digitally accessible knowledge available. The problem was knowledge deficiency because the available data remained unanalysed and dormant for conservation decision-making.

Plants as indicators of minerals (metallophytes)

Miners and scientists have long known that certain plant species can be a signal for ore-bearing rocks (Brooks et al. 1985; Ernest 2006). For example, *Lychnis alpina*, a small pink-flowering plant in Scandinavia, and *Haumaniastrum katangense*, a white-flowered shrub in central Africa, are both associated with copper. Haggerty (2015) discovered that *Pandanus candelabrum* is closely associated with kimberlite pipes that are mined for diamonds in Liberia. Using herbarium specimens to map the distribution of *Pandanus candelabrum* can help in diamond prospecting since this species is restricted to the diamond-bearing kimberlite dykes and is not found even in the alluvium covering the adjacent dikes.

Other major documented use-cases

Following are other major studies and compendia that demonstrate the use of collections and associated data for important research and findings across a broad suite of subjects. The links to the publications are given after each summary.

GBIF-mediated data (2012-2015) use cases

The GBIF Secretariat has systematically reviewed and compiled peer-reviewed research using and applying GBIF-mediated data since 2008, and since 2012, these have been published in an annual *Science Review*. Each issue below documents additional use cases, mostly addressing the data to knowledge value chain. It should be noted that as of 2016, the name reflects the year of publication of the Science Review rather than the papers summarized therein. As such, the 2016 Science Review chiefly summarizes articles published in 2015.

- 2016 GBIF Science Review: *http://www.gbif.org/resource/82873*
- 2014 GBIF Science Review: http://www.gbif.org/resource/82191
- 2013 GBIF Science Review: http://www.gbif.org/resource/80915
- 2012 GBIF Science Review: http://www.gbif.org/resource/80847

Chapman report (2005) use cases

Chapman (2005) in a GBIF report on the uses of primary species occurrence data discusses a wide range of use cases of natural history collections data.

http://www.gbif.org/resource/80545

UK Natural Science Collections Association, (NatSCA 2005) use cases

In 2005, the Natural Science Collections Association, UK, published a report entitled, “A matter of life and death: Natural science collections: why keep them and why fund them?”, in which they emphasized the importance of NHCs and the uses they serve.

http://www.natsca.org/sites/default/files/publications-full/A-Matter-Of-Life-And-Death.pdf

US Interagency Working Group on Scientific Collections (IWGSC 2009) use cases

The IWGSC report published in 2009 documents a number of use cases for the US federal scientific collections, both biological and non-biological. The US federal collections are viewed as part of the global scientific infrastructure and enterprise. The use cases illustrate the impact of scientific collections in the following areas: economy and trade, environmental change over time, environmental quality, invasive species, scientific treasures, food and agriculture, public health and safety, national security and unanticipated uses.

https://usfsc.nal.usda.gov/sites/usfsc.nal.usda.gov/files/IWGSC_GreenReport_FINAL_2009.pdf

Virginia Tech - Biological collections as a resource for technical innovation

Virginia Tech recently launched a three-year project (2015-2017) funded by the US National Science Foundation that addresses the unanticipated uses of biological collections - biocollections inspiring engineering innovation. The project aims at giving scientists and engineers from diverse backgrounds specific suggestions as to how natural history collections could be leveraged for engineering innovation. Secondly it will also provide policy makers and the general public a well-justified outline of the innovative and economic potential of natural history collections as well as estimates for the effort that would be required to realize this potential.

https://www.nsf.gov/awardsearch/showAward?AWD_ID=1521072

Outcomes for this project will be archived at http://bist.centers.vt.edu.

Data Gap Analysis: Setting priorities for digitization

For all of biodiversity science—and knowledge in general—data gap analysis (DGA) enables us to "know what we don't know" and then to prioritize the filling of those gaps according to strategic imperatives (Arturo et al. 2015). With regard to biocollections institutions, worldwide they are faced with the challenging dilemma of how to prioritize the digitization of massive amounts of data associated with millions of voucher specimens of animals, plants, fungi and other organisms that document the planet's biodiversity. Setting digitization priorities, informed by data gap analyses, is essential to having the best biodiversity data enable the best science for understanding and advancing social, economic and environmental well-being—and to do so in time to make a difference. Indeed, setting such gap-based priorities is incumbent on biocollections if they are to speed the flow of data-to-knowledge-to-application in the value chain.

The Task Force recognizes that setting digitization priorities involves serving competing institutional, local, regional, national and global imperatives. These include, but are not limited to: individual investigator research interests; institutional mandates; science agendas; and various pressing environmental concerns (e.g. endangered species, invasive species, zoonotic diseases, pollinators, pests). Moreover, each imperative has its particular calculus of taxonomic groups, geographic areas, time periods and ecosystems/habitats. Overlying these permutations are the missions of different stakeholders and funders: intergovernmental bodies (e.g. IPBES, CBD), government agencies, NGOs, private foundations and corporations. In a resource-limited world, a digitization strategy of maximum efficacy will require all parties to collaborate on setting overarching priorities that, simultaneously: (1) target the most urgent environmental imperatives of our time; (2) are underpinned by sophisticated data gap analyses of those imperatives; (3) include the greatest commonality among competing interests; (4) tackle what is most pragmatic; and (5) promise the most immediate impacts (see Recommendations).

To that end, the Task Force convened a symposium on "Setting Global and Local Digitization Priorities" at the SPNHC conference in Berlin, June 20-25, 2016 (http://www.spnhc2016.berlin/page40.html). The five presentations in the symposium centred on setting digitization priorities in a variety of situations from the global, regional, national and institutional level, and across different collection sizes and themes to satisfy a variety of competing needs. A summary of the results from the NHC survey (section 2 in this report) including how respondents have set digitization priorities, was also presented.

The task force also convened a side meeting to plan the review of the Natural Collections Description (NCD) standard and metadata needs for NHCs.

Next Steps

The Task Force recommends the following steps as a follow up to this report:

- Draft a priority-setting framework for individual biocollections institutions.
- Based on the draft framework, convene a series of meetings with stakeholders to develop strategic frameworks for helping biocollections deliberate and set their digitization priorities.
- In partnership with the RDA/TDWG joint working group on metadata standards for the sciences, evaluate the application of NCD standard
- Develop roadmap documents to assist institutions in mobilizing biocollections metadata
- Help form a closer-working cooperative network of global biocollection entities and societies to achieve a critical mass for planning, policy impact and generating resources.
- Hone and tailor biocollections use-cases for specific communities (researchers, corporations, foundations, policy makers, educators, etc.) to demonstrate the benefit of published, vouchered biodiversity data for science, society, governments and the private sector across a series of thematic imperatives.
- Convene major summits of government, corporate and foundation institutions to develop a funding mechanism to complete the strategic, priority-based digitization of biocollections data worldwide.

Recommendations

Setting priorities for digitization and data gap analysis

1. The community, stakeholders and individual biocollections should establish collaborative, integrated priorities for digitization of biocollections data based on the value framework of data-knowledge-application. Within this framework, priorities should be demand-driven by global, national, regional and local concerns and required research that simultaneously: (a) target the most urgent social, environmental, economic and biodiversity science imperatives of our time; (b) are underpinned by sophisticated gap analyses; (c) include the greatest commonality among competing imperatives and interests; (d) tackle what is most pragmatic, first; and (e) promise the most immediate impacts.

   Such a demand-driven approach addresses both links in the value chain of data-to-knowledge-to-application. Fulfilling the first link—data-knowledge—provides the raw, vouchered data that catalyses research results that inform solutions. Fulfilling the second link—knowledge-to-application—recruits demand-driven investments for such solutions from governments, foundations and the corporate sector. Mathematical algorithms may be a useful tool for calculating and modelling such priorities (see Butts et al. 2010).

   Extensive gap analyses of existing digitized data will identify the critical taxonomic and geographic data gaps and data enhancements (e.g. georeferencing) that need to be filled to address the strategic priorities. The GBIF DGA report by Arturo et al. (2016) provides such DGA methodologies. Critical geographic gaps can be inferred from GBIF's overall occurrences plot densities - http://www.gbif.org/occurrence. Community data aggregators such as GBIF and iDigBio should work together on providing robust DGAs.

2. The community, led by GBIF and international and national partners, should convene a series of high-level summits among biocollections, governments, corporations and foundations to develop, fund and implement the five, long-term strategies (Recommendation 1) for mobilization the remaining 90% of the world's biocollections data and bringing them into currency for science and society. A component of this investment to be explored is an international fund for accelerating digitization of biodiversity data in regions with the largest data gaps.

Digitization and best practices

3. Biocollections should employ the proven best practices and community standards to digitize their collections, with examples and guidance from GBIF, iDigBio, SPNHC, ALA, TDWG, etc. One of the best practices is adoption of a tiered strategy for worldwide collections digitization (plus imaging where appropriate) and model concepts, such as Linked Open Data (LOD) (Berner-Lee 2009) and the *Digitization Maturity Model* in the ALA's Guide to Digitization (Kalm 2012). Specifically, the five tiers in this strategy are:
   a) *Level One*—Metadata I: Make the collection globally discoverable by publishing the Institution/Organization/Collection-level information, i.e., the who and where of a collection and, broadly, its content and history and by registering collections with GRBio.
   b) *Level Two* - Metadata II: Produce and publish species-level or cabinet-level collection inventories as a basis for strategic planning of collection growth, conservation and follow-on digitization.
   c) *Level Three* – Specimen Data I: Capture and publish skeletal-level data (with or without locality georeferencing), mapping data to currently accepted

standards, and exposing imperfect or erroneous data to machine algorithms and global expertise for correction and improvement.

d) *Level Four* – Specimen Data II: Georeference locality data and enrich specimen records with other data, such as: field notes, grey literature, note cards, etc.

e) *Level Five* – Specimen Data III: Data associated with all new collections is "born digital", i.e., immediately captured and incorporated into the existing, georeferenced database. Links to GenBank accessions, BCoL IDs, etc., must include *globally unique identifiers*.

Global collaboration, where possible, can help meet Recommendation 1 (see above) to address global imperatives and to achieve critical mass, least redundancy and economies of scale.

4. Biocollections should work with community partners to remove their perceived barriers to their digitization efforts. The three major barriers cited by respondents in the TF survey have been successfully eased or removed by numerous institutions and biocollections. For example, setting strategic priorities and schedules for data capture will ease the perception that the "task is overwhelming." Productive collection-based research and funding from government and private entities can reverse the perception that the efforts/resources expended exceed the benefits of digitization. And digitization of collections is the fastest, most efficient method of identifying and correcting collection data errors, often en mass, so that both the physical specimens and their data can be readily deployed and trusted for the very purposes those collections were intended to serve. Many biocollections-focused organizations, e.g. iDigBio, BCoN, NSCA, SYNTHESYS3, SPNHC, TDWG, etc. can provide the expertise, training and tools to advance efficient and effective collection data capture and publication to the worldwide community.

Metadata

5. Natural history collections should publish their rich metadata for both digitized and non-digitized specimens in order to make them discoverable and advertise their value to science and society.
6. To accomplish this, the community should develop robust yet user-friendly APIs for metadata and evolve metadata standards, including resolving the semantic problem with "metadata" as it means different things to different people. Current metadata standards could be adapted and emended, for example, Ecological Metadata Language (EML) with an extended profile for NHC, and mapping data from Natural Collections Description (NCD) to EML.
7. GBIF's dataset metadata should be presented in a structured way so that it can be searchable by using various parameters
8. The community, led by GBIF, SPNHC, TDWG, GRBio, iDigBio and other entities, should develop an educational campaign on the importance of metadata, how it advertises the collection to the world of users and its value, how to capture and report critical metadata fields, etc.
9. Following the lead of the IWGSC, government agencies should improve and provide access to the documentation of the contents of their scientific collections.

Partnership and collaboration

10. Current biocollections-centred organizations should strongly consider integration into a federated union with greater critical mass, impact and effectiveness nationally and internationally. For example, organizations that focus mainly on physical specimens and their data (e.g. SPNHC), data standards (TDWG), collections

metadata (GRBio), policy and funding (NSCA), data portals (GBIF, VertNet, BISON) and regional/national efforts (iDigBio, SiBBr (http://www.sibbr.gov.br), NBN (https://nbn.org.uk), ALA, etc.) have extensively overlapping missions and interests that would be less dispersed and more efficiently fulfilled through federated cooperation. One model in this domain is the World Federation of Culture Collections.

11. Following the example of the IWGSC (2009), agencies should collaborate much more closely in setting and implementing the policies, procedures and protocols for managing their scientific collections.

References

Arturo A, Chavan V & Otegui J (2016) Best Practice Guide for Data Gap Analysis for Biodiversity Stakeholders. Copenhagen: GBIF Secretariat. http://www.gbif.org/resource/82566

Barnosky, AD (2008) Megafauna biomass tradeoff as a driver of Quaternary and future extinctions. *Proceedings of the National Academy of Sciences of the United States of America* 105, Supplement 1: 11543-11548. http://dx.doi.org/10.1073/pnas.0801918105

Barnosky AD, Matzke N, Tomiya S, Wogan GOU, Swartz B, Quental TB, Marshall C, McGuire J, Lindsey EL, Maguire K, Mersey B & Ferrer E (2011) Has the Earth's sixth mass extinction already arrived? *Nature* 471:51-57. http://dx.doi.org/10.1038/nature09678

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. http://dx.doi.org/10.17161/bi.v7i2.3989.

Berg W, Johnels A, Sjostrand B & Westermark T (1966) Mercury contamination in feathers of Swedish birds from the past 100 years. *Oikos* 17: 71-83.

Berner-Lee T (2009) Linked Data. https://www.w3.org/DesignIssues/LinkedData.html

Blach-Overgaard, A., Balslev, H., Dransfield, J., Normand, S. & Svenning, J. C. 2015. Global-change vulnerability of a key plant resource, the African palms. Sci. Rep. 5, 12611; doi: 10.1038/srep12611

Blach-Overgaard, A., Svenning, J. C., Dransfield, J., Greve, M. & Balslev, H. 2010. Determinants of palm species distributions across Africa: the relative roles of climate, non-climatic environmental factors, and spatial constraints. Ecography 33, 380–391; DOI: 10.1111/j.1600-0587.2010.06273.x

Blagoderov, V., Kitching, I.J., Livermore, L., Simonsen, T.J., and Smith, V.S. 2012. No specimen left behind: industrial scale digitization of natural history collections. ZooKeys 209:133-146. doi:10.3897/zookeys.209.3178

Bogoch, I.I., Brady, O.J., Kraemer, M.U., German, M., Creatore, M.I., Kulkarni, M.A. et al. 2016. Anticipating the international spread of Zika virus from Brazil. Lancet 387:335–336. http://dx.doi.org/10.1016/S0140-6736(16)00080-5

Brooks, R.R., Malaisse, F., and Empain, A. 1985. The heavy metal tolerant flora of southcentral Africa: A multidisciplinary approach. A.A. Balkema, Rotterdam and Boston. Available at: http://trove.nla.gov.au/work/18240023

Buerki, S. and Baker, W. 2016. Collections-based research in the genomic era. Biological Journal of the Linnean Society 117:5-10. http://dx.doi.org/10.1111/bij.12721

Butts, S.H., Bazeley, J.A., and Briggs, D.E.G. 2010. A curatorial assessment for stratigraphic collections to determine suitability for incorporation into a systematic collection. Collection Forum, 24(1-2):46-51. Available at: www.spnhc.org/media/assets/cofo-24.pdf

Cao-Lormeau, V.M., and Musso, D. 2014. Emerging arboviruses in the Pacific. Lancet 384: 1571–1572. http://dx.doi.org/10.1016/S0140-6736(14)61977-2

Ceballos, G., Ehrlich, P.R., Barnosky, A.D., García, A., Pringle, R.M. and Palmer, T.M., 2015. Accelerated modern human–induced species losses: Entering the sixth mass extinction. Science advances, 1(5), p.e1400253. DOI: 10.1126/sciadv.1400253

Chapman, A.D. 2005. Uses of Primary Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. 100 pp. Available online at: http://www.gbif.org/resource/80545

Dunn, P. O. and Winkler, D.W. 1999. Climate change has affected the breeding date of tree swallows throughout North America. Proceedings of the Royal Society of London B 266. 2487–2490.

Ernest, W.H.O. 2006. Evolution of metal tolerance in higher plants. For. Snow Landsc. Res. 80, 3: 251–274. http://www.wsl.ch/publikationen/pdf/7764.pdf

Fichet-Calvet, E., Rogers, D.J. 2009. Risk maps of Lassa fever in West Africa. PLoS Neglected Tropical Diseases 3: e388. doi: 10.1371/journal.pntd.0000388

Gaiji, S., Chavan, V., Arino, A.H., Otegui, J., Hobern, D., Sood, R., and Robles, E. 2013. Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. Biodiversity Informatics, [S.l.], v. 8, n. 2, July 2013. doi:http://dx.doi.org/10.17161/bi.v8i2.4124

Haggerty, S. 2015. Discovery of a kimberlite pipe and recognition of a botanical diagnostic indicator in NW Liberia. Economic Geology 110: 851-856. DOI: 10.2113/econgeo.110.4.851

Hayes, T., Haston, M., Tsui, M., Hoang, A., Haeffele, C., and Vonk, A. 2002. Herbicides: Feminization of male frogs in the wild. Nature 419:895–896.

Heerlien, M., van Leusen, J., Schnoerr, S., de Jong-Kole, S., Raes, N., and van Hulsen, K. 2015. The natural history production line: An industrial approach to the digitization of scientific collections. ACM Journal of Computing and Cultural Heritage 8(1), Article 3, 11 pages. DOI: http://dx.doi.org/10.1145/2644822

Hickey, J. J. and Anderson, D.W. 1968. Chlorinated hydrocarbons and eggshell changes in raptorial and fish-eating birds. Science 162:271–273

Hoffmaster, A. R., Fitzgerald, C.C., Ribot, E., Mayer, L.W., and Popovic, T. 2002. Molecular subtyping of *Bacillus anthracis* and the 2001 bioterrorism-associated anthrax outbreak, United States. Emerging Infectious Diseases 8:1111–1116.

Holmes, M.W., Hammond,T.T., Wogan, G.O., Walsh, R.E., LaBarbera, K., Wommack, E.A., Martin, F.M., Crawford, J.C., Mack. K.L, Bloch, L.M., and Nachman, M.W. 2016. Natural history collections as windows on evolutionary processes.Molecular Ecology 25(4):864-8. doi: 10.1111/mec.13529.

Hudson, L.N., Blagoderov, V., Heaton, A., Holtzhausen, P., Livermore, L., Price, B.W., et al. 2015. Inselect: Automating the Digitization of Natural History Collections. PLoS ONE 10(11): e0143402. doi:10.1371/journal.pone.0143402

IWGSC (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.

Jackson, J. 2016. Digital Collections: the Cisco Pitstop | Digital Museum https://blog.nhm.ac.uk/2016/02/26/digital-collections-the-cisco-pitstop/ 26 February 2016.

Jones MC, Dye SR, Pinnegar JK, Warren R & Cheung WWL (2014) Using scenarios to project the changing profitability of fisheries under climate change. *Fish and Fisheries* 16: 603-622. http://dx.doi.org/10.1111/faf.12081

Kalms B (2012) Digitisation: A strategic approach for natural history collections. Atlas of Living Australia, CSIRO Ecosystem Sciences, Canberra, ACT, Australia. http://www.ala.org.au/wp-content/uploads/2011/10/Digitisation-guide-120326.pdf

Kraemer MUG, Sinka ME, Duda KA, Mylne AQN, Shearer FM, Barker CM, Moore CG, Carvalho RG, Coelho GE, Van Bortel W, Hendrickx G, Schaffner F, Elyazar IRF, Teng HJ, Brady OJ, Messina JP, Pigott DM, Scott TW, Smith DL, Wint GRW, Golding N, Hay

SI (2015) The global distribution ofthe arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. *eLife* 4: e08347. http://dx.doi.org/10.7554/eLife.08347

Kremen C, Cameron A, Moilanen A, Phillips SJ, Thomas CD, Beentje H, Dransfield J, Fisher BL, Glaw F, Good TC, Harper GJ, Hijmans RJ, Lees DC, Louis E Jr, Nussbaum RA, Raxworthy CJ, Razafimpahanana A, Schatz GE, Vences M, Vieites DR, Wright PC, Zjhra ML (2008) Aligning Conservation Priorities Across Taxa in Madagascar with High-Resolution Planning Tools. *Science* 320: 222-226. http://dx.doi.org/10.1126/science.1155193

Lendemer JC & Allen JL (2014) Lichen biodiversity under threat from sea-level rise in the Atlantic Coastal Plain. *Bioscience* 64: 923-931. http://dx.doi.org/10.1093/biosci/biu136

Maron NL & Pickle S (2013) Appraising our Digital Investment: Sustainability of Digitized Special Collections in ARL Libraries. Ithaka S+R. http://dx.doi.org/10.18665/sr.22363

Mason SC Jr., Betancourt I, Gelhaus JK (2016) A digital species index of the entomology collection at the Academy of Natural Sciences of Drexel University. Presented at Entomological Collections Network Meeeting September 23, 2016. Orlando, Florida, USA

Millennium Ecosystem Assessment (2005) *Ecosystems and Human Well-Being: Biodiversity Synthesis*. Washington, DC: World Resources Institute.

Miller-Struttmann, N. E., Geib, J., Franklin, J.D & Galen C (2015) Functional mismatch in a bumble bee pollination mutualism under climate change. *Science* 349: 1541-1544. http://dx.doi.org/10.1126/science.aab0868

Nelson G, Paul D, Riccardi G, Mast A (2012) Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys* 209: 19-45. http://dx.doi.org/10.3897/zookeys.209.3135

Novak SJ & Mack RN (2001) Tracing plant introduction and spread: Genetic evidence from *Bromus tectorum* (Cheatgrass). *BioScience* 51: 114-122. http://dx.doi.org/10.1641/0006-3568(2001)051[0114:TPIASG]2.0.CO;2

National Research Council (2002) *Countering Agricultural Bioterrorism*. Washington (DC): National Academies Press.

Parmesan C (1996) Climate and species' range. *Nature* 382:765-766. http://dx.doi.org/10.1038/382765a0

Pergams ORW & Nyberg D (2001) Museum collections of mammals corroborate the exceptional decline of prairie habitat in the Chicago region. *Journal of Mammology* 82: 984-992. http://dx.doi.org/10.1644/1545-1542(2001)082<0984:MCOMCT>2.0.CO;2

Peterson AT, Moses LM, Bausch DG (2014) Mapping Transmission Risk of Lassa Fever in West Africa: The Importance of Quality Control, Sampling Bias, and Error Weighting. *PLoS ONE* 9(8): e100711. http://dx.doi.org/10.1371/journal.pone.0100711

Peterson AT, Soberón J & Krishtalka L (2015) A global perspective on decadal challenges and priorities in biodiversity informatics. *BMC Ecology* 15: 15. http://dx.doi.org/10.1186/s12898-015-0046-8

Piggott DM, Golding N, Adrian M et al. (2014) Mapping the zoonotic niche of Ebola virus disease in Africa. *eLife* 3: e04395. http://dx.doi.org/10.7554/eLife.04395

Provan J, Booth D, Todd NP, Beatty GE & Maggs CA (2007) Tracking biological invasions in space and time: elucidating the invasive history of the green alga *Codium fragile* using old DNA. *Diversity and Distributions* 14(2): 343-354. http://dx.doi.org/10.1111/j.1472-4642.2007.00420.x

Ratcliffe DA (1967) Decrease in eggshell weight in certain birds of prey. *Nature* 215: 208-210. http://dx.doi.org/10.1038/215208a0

Sanchez AC, Osborne PE & Haq N (2011) Climate change and the African baobab (*Adansonia digitata* L.): the need for better conservation strategies. *African Journal of Ecology* 49: 234-245. http://dx.doi.org/10.1111/j.1365-2028.2011.01257.x

Schindel D, Miller S, Trizna M, Graham E & Crane A (2016) The Global Registry of Biodiversity Repositories: A Call for Community Curation. *Biodiversity Data Journal* 4: e10293. http://dx.doi.org/10.3897/BDJ.4.e10293

Scoble M (2010) Rationale and Value of Natural History Collections Digitisation. *Biodiversity Informatics* 7: 2. http://dx.doi.org/10.17161/bi.v7i2.3994.

Sousa-Baena MS, Couto Garcia L & Peterson AT (2013) Knowledge behind conservation status decisions: Data basis for "Data Deficient" Brazilian plant species. *Biological Conservation* 173: 80-89. http://dx.doi.org/10.1016/j.biocon.2013.06.034

Sousa-Baena MS, Couto Garcia L & Peterson AT (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions* 20(4): 369-381. http://dx.doi.org/10.1111/ddi.12136

Suarez AV & Tsutui ND (2004) The value of museum collections for research and society. *Bioscience* 54(1): 66-74. http://dx.doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

Suarez AV, Holway DA & Case TJ (2001) Patterns of spread in biological invasions dominated by long-distance jump dispersal: Insights from Argentine ants. *Proceedings of the National Academy of Sciences* 98(3): 1095-1100. http://dx.doi.org/10.1073/pnas.98.3.1095

Vollmar A, Macklin JA, Ford L (2010) Natural History Specimen Digitization: challenges and concerns. *Biodiversity Informatics* 7: 93-112. http://dx.doi.org/10.17161/bi.v7i2.3992

Yates TL (2002) The ecology and evolutionary history of an emergent disease: Hantavirus pulmonary syndrome. *BioScience* 52 (11): 989-998. http://dx.doi.org/10.1641/0006-3568(2002)052[0989:TEAEHO]2.0.CO;2

Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, Lanz C, Martin FN, Kamoun S, Krause J, Thines M, Weigel D, Burbano HA (2013) The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *eLife* 2013(2): e00731 http://dx.doi.org/10.7554/eLife.00731

Annex I: Acronyms and abbreviations

ADBC	Advancing the Digitization of Biological Collections
ALA	Atlas of Living Australia
API	Application programming interface
BCOL	Barcode of Life
BCoN	Biodiversity Collections Network
BID	Biodiversity Information for Development
BISON	Biodiversity Information Serving Our Nation, USA
Canadensys	Network of Canadian biological collections
CBD	Convention on Biological Diversity
CETAF	Consortium of European Taxonomic Facilities
GRBio	Global Registry of Biodiversity Repositories
iDigBio	Integrated Digitized Biocollections, USA
IPBES	The Intergovernmental science-policy Platform on Biodiversity and Ecosystem Services
IWGSC	Interagency Working Group on Scientific Collections
MUSEUM-L	A general purpose, cross-disciplinary electronic discussion list for museum professionals, students and all others interested in museum related issues, under the auspices of the International Council of Museums
NBN	National Biodiversity Network, UK
NCD	Natural Collections Description
NGO	Non-Governmental Organization
NHC	Natural History Collections
Nicol-L	Natural History Collections List of Society for the Preservation of Natural History Collections
NSCA	Natural Science Collections Alliance, USA
RDA	Research Data Alliance
SERNEC	SouthEast Regional Network of Expertise and Collections, USA
SiBBr	Sistema de Informação sobre a Biodiversidade Brasileira (Information System on Brazilian Biodiversity)
SPNHC	Society for the Preservation of Natural History Collections
SYNTHESYS	An EU-funded project creating an integrated European infrastructure for natural history collections
Taxacom	Biological Systematics Discussion List
TDWG	Biodiversity Information Standards also known as Taxonomic Databases Working Group
TF	The GBIF Task Force
VertNet	Database of vertebrate biodiversity data from natural history collections
WeDigBio	Worldwide Engagement for Digitizing Biocollections

Annex II: Summary of results from a first analysis of NHC survey

GBIF Task Force Report on Survey Subsection: A Global Survey of Natural History Collections. http://www.gbif.org/newsroom/news/accelerating-discovery-of-biocollections-data

Introduction

As part of a broader global strategy for mobilizing primary biodiversity data, GBIF convened a Task Force (TF) to help accelerate the discovery and access to both digitized and non-digitized collections[1]. In the initial meetings the TF realized that it was important to get data on the current state of NHCs in order to inform future discussions and recommendations. Some of the pertinent questions that we sought to understand included the following:

- What methods and models are proving most successful for worldwide collections digitization?
- What metadata are key to sharing so that other stakeholders can discover collections?
- Who is doing the digitization?
- What are the obstacles most often encountered?
- What variables are driving what gets digitized?

With no available data, the GBIF Task Force on Accelerating the Discovery of Biocollections Data decided to carry out a global survey as the best way to find out the state of affairs of NHCs worldwide.

Purpose of survey

In the late 2015, the task force designed and administered a global survey on natural history collections. The purpose of the survey was to enable the task force to determine and demonstrate: (1) The digital readiness of the world's biocollections and their institutions; (2) the benefits to the collection/institution that digitization engenders; and (3) the impediments to collection data digitization.

Survey methodology

The survey questionnaire was prepared by the TF through meetings, consultations and research, and mainly administered online through Qualtrics software. An MS Word version was made available for respondents that had difficulty in filling out the online survey. We distributed the survey to individuals affiliated with collections worldwide through a variety of channels including listservs and member lists including Index Herbariorum, and GBIF nodes, as well as personal contacts with institutions. We also announced the survey at relevant meetings, workshops and conferences attended by potential target respondents. The survey was translated into Japanese and distributed throughout Japan via the community of "Science Museum Network (S-Net)". Those contacted were further requested to help with the further distribution of the survey in order to help maximize the global reach. Table 1 summarizes the key distribution channels.

[1] http://www.gbif.org/newsroom/news/accelerating-discovery-of-biocollections-data

Table 1 Main distribution channels for the global NHC survey

Taxacom
Herbaria-L
iDigBio
GBIF Node Managers
SERNEC
NHColl
GRBio
Index Herbariorum
TDWG
MUSEUM-L
S-Net (http://science-net.kahaku.go.jp/)
Conferences, e.g. TDWG (Nairobi), iDigBio V Summit (Washington DC), ECN, etc.

Summary of survey results.

We present a summary of the survey results here but a more detailed report with further analyses, interpretation, and discussions will be made available at a later date as part of the final report of the GBIF Task Force. Over 800 responses representing nearly 2000 collections distributed over 72 countries were received (Map 1). Deleting very incomplete responses resulted in 617 usable responses. The survey had a wide geographic reach across all the major continents but with much higher response rates from Western Europe, North America in particular the USA and Japan because the survey was translated into Japanese. Note that the number of respondents that answered each question varied either because either in the case of the online version, they were not shown the question based on prior responses, or the respondents elected to not answer the question. We present a summary of the key results.

Map 1 shows the locations of the survey respondents based on the registered IP addresses from where the online survey was completed.

Map 1. Survey respondents by IP address, by Kevin Love, iDigBio, using CartoDB

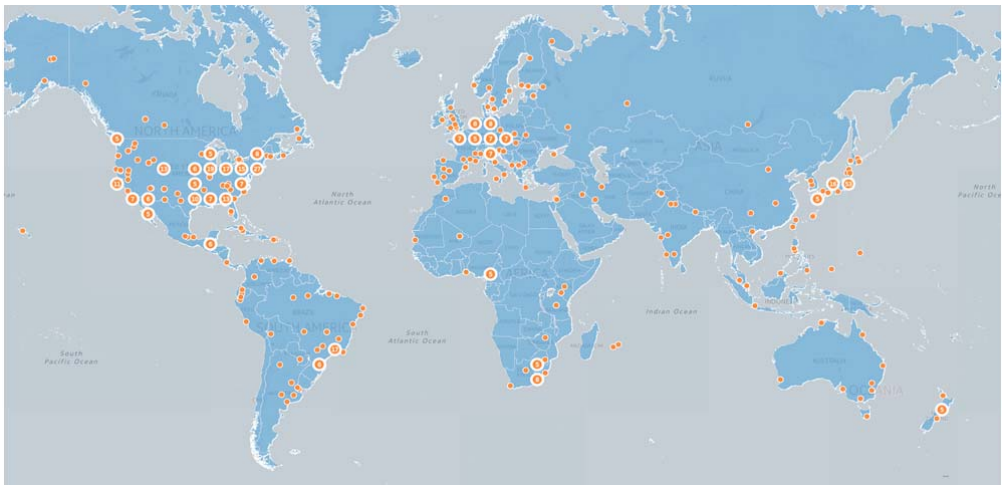


Table 2. Summary of Key data from the global NHCs survey

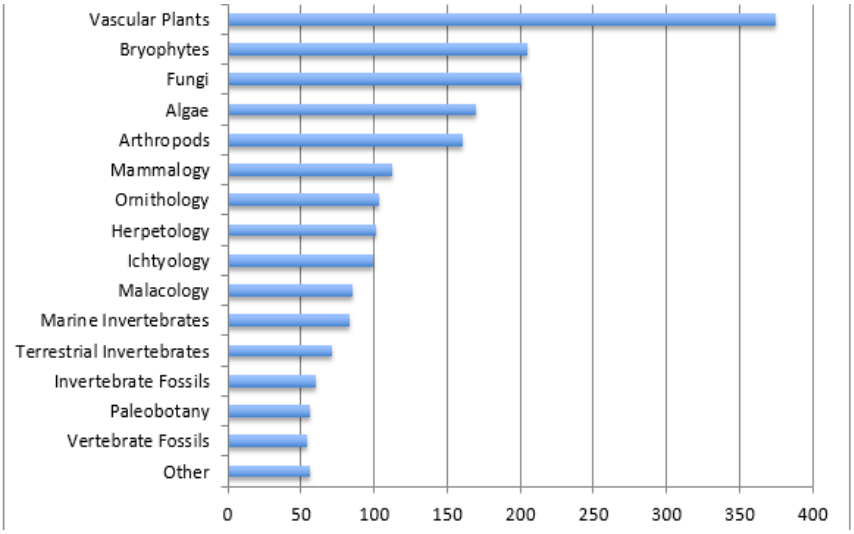
86% (615 respondents) indicate they are currently digitizing or have completed digitizing at least some or all of their collections.
76% of respondents were individuals based at publicly funded institutions, with 40% universities and 36% non-university institutions. Nearly all (92%) of respondents were primarily curators or collection managers with 10% as head of research and collections.
13 out of 15 collection types, when averaged, report their collections over 50% electronically databased.
Very few respondents (1% or 5 individuals) reported they are not digitizing and have no plans to do so.
The top 10 obstacles to digitization are: funding, time (lack of), size of task, not institutional priority, data has errors, limited expertise, lack of digitization process knowledge, no credit (tenure, reappointment) for this work, not priority for those in administration, not a good payoff for needed effort
The top three priorities when deciding what to digitize are research (53%) and funding / grant opportunities (51%), followed by taxonomic priorities (42%).
Benefits of digitization - top eight responses: increased use of collections, increased exposure, better knowledge of holdings, better management of data, digital preservation, enhanced data quality, new skills for staff, better management of physical specimens

Details

Collections responding

Collection type response rate varied (see Figure 1). Vascular plant collections represent nearly 20% of total respondents followed by Bryophytes (10%), Fungi (10%), Algae (9%), Arthropods (8%), Mammalogy (6%), Ornithology (5%), Herpetology (5%), and Ichthyology (5%) with the remaining representing Malacology, Marine Invertebrates, Terrestrial Invertebrates, Invertebrate Fossils, Paleobotany, Vertebrate Fossils, and other.

Figure 1. Collection types represented by our survey respondents (number of collections n=1992)



Staff responding to survey

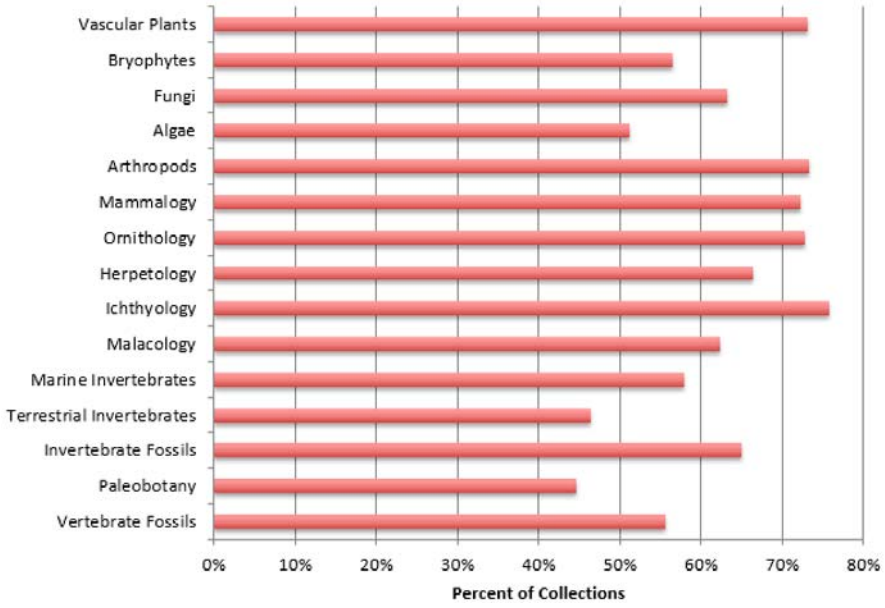
For n=615 respondents, curators made up 57%, collection managers 35%, faculty 34%, head of research and collections 13%, information manager 10%, director/CEO 8%, and other 4%.

Digitization and databasing trends

Overall, our data seems to indicate that most collections are digitizing at least some part of their collections or trying to do so. Of the 615 respondents who answered the question, 86% indicated that they are currently or have completed databasing their collections. The percentage varies across collection type, as does the mean portion of the collection that has been completed.

If we average answers for “percent of collection databased” and group by collection type, we are encouraged to note that 13 out of 15 collection types are over 50% databased, collectively (Figure 2). In other words, if one vascular plant collection said it was 25% digitized, and another vascular plant collection said it was 75% digitized, then between the two collections - they are 50% digitized.

Figure 2. If digitizing all or part of a collection, what percent is databased?

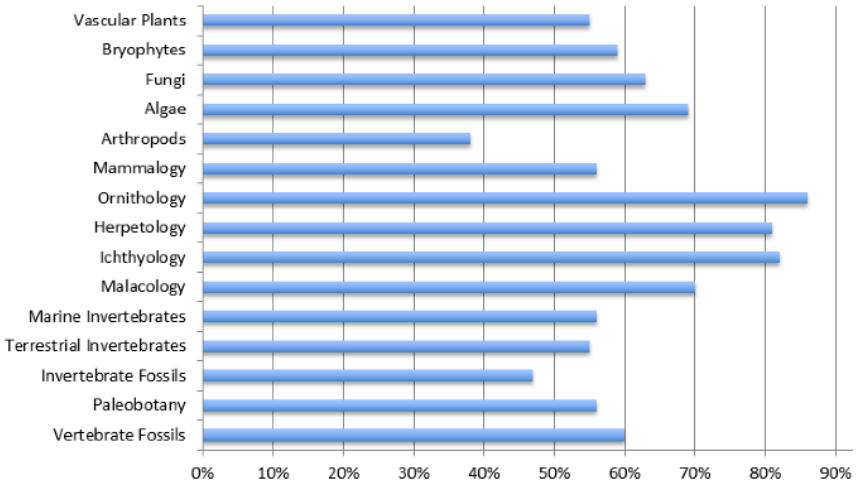


When looking at these averages of the percent of collection databased, for each organismal group, it’s notable that the averages of all collection types is over 50% for all groups except arthropods which averages 38% and invertebrate fossils which averages 47% (Figure 3).

Published collections

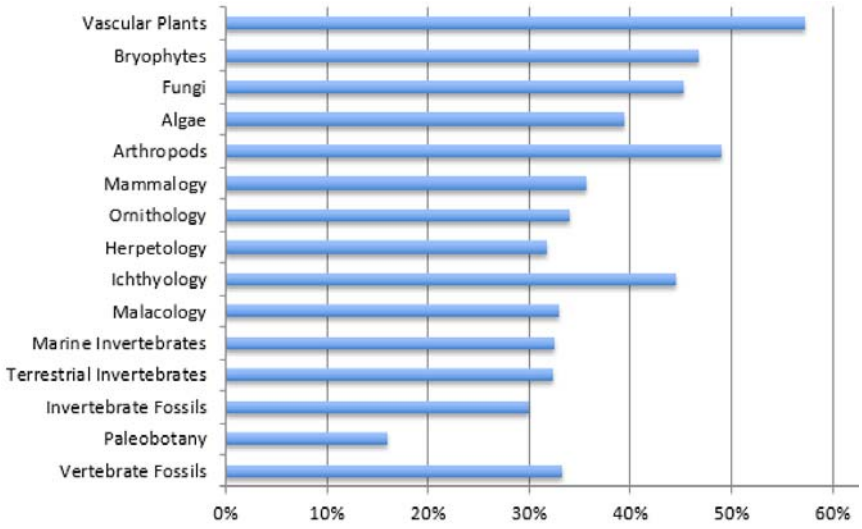
The mean percentage of collections that have been published is higher than the mean percentage of those databased, but that is because the published statistic reflects those that have already been databased. It is based on a smaller number of collections. To be exact, the number of those who reported publishing their collection data is one-third smaller than those who reported databasing their collections.

Figure 3. Averaging the percent of collection databased, across each collection type



We can illustrate with vascular plants. We have 375 vascular plant collections in the data set (or, more precisely, 375 people provided information about vascular plant collections). Of these 274 report databasing their collections. On average, 55% of the collections are databased. This is averaged across all collections that provided an actual percentage. A smaller number of individuals, 215, reported publishing their collection data. On average, 65% of the collection data is published (Figure 4).

Figure 4. Percent of collections published (partial or complete)



Barriers to digitization

If many are digitizing, or trying to do so, *what barriers get in the way of digitizing*? If we know what these barriers are, how can we use this information going forward? (See recommendations in interim and final report). Although only 1% of respondents (n=5 individuals) indicated that there are no plans to digitize their collections, far more people (n=587) offered to share what they experience as obstacles to digitization — most notably lack of funding/resources and lack of time. See similar insightful data from ITHAKA for digitization of special collections (Maron & Pickle 2013).

Table 3. The top 10 barriers to digitization, for our survey respondents (n=587)

1. Funding / resources not available (80%)
2. Lack of time among personnel (80%)
3. Size of task is overwhelming (40%)
4. Not an institutional priority (35%)
5. Collection data has errors (30%)
6. Limited expertise among personnel (25%)
7. Insufficient information on digitization process (15%)
8. No benefit to reappointment, tenure (14%)
9. Not a priority of the individual in charge (12%)
10. Not a good effort / payoff ratio (10%)

The reasons beyond the first two, can be grouped into 3 categories as suggested here, in order to address them (Table 4). (See recommendations in upcoming interim and final report).

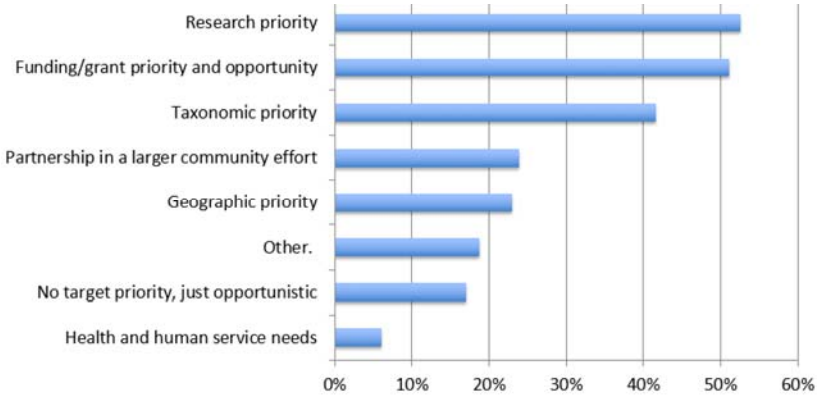
Table 4. Obstacles to digitization - 3 groups

- size of task is overwhelming
- not an institutional priority, no benefit, not a priority, not good effort / payoff ratio, not priority, lack of perceived need, deemed not valuable
- data has errors, limited expertise, lacking information on the digitization process

How are collections deciding what to digitize?

With 519 respondents, the top three variables driving the decisions for what to digitize include: research priority (52%), funding/grant opportunities (51%), and taxonomic priority (42%). Other reasons selected were: partnership in a larger community effort (24%), and geographic priority (23%). About 18 percent reported their digitization is opportunistic with no target priority, and about 5% said health and human service needs drive their digitization decisions. (Figure 5)

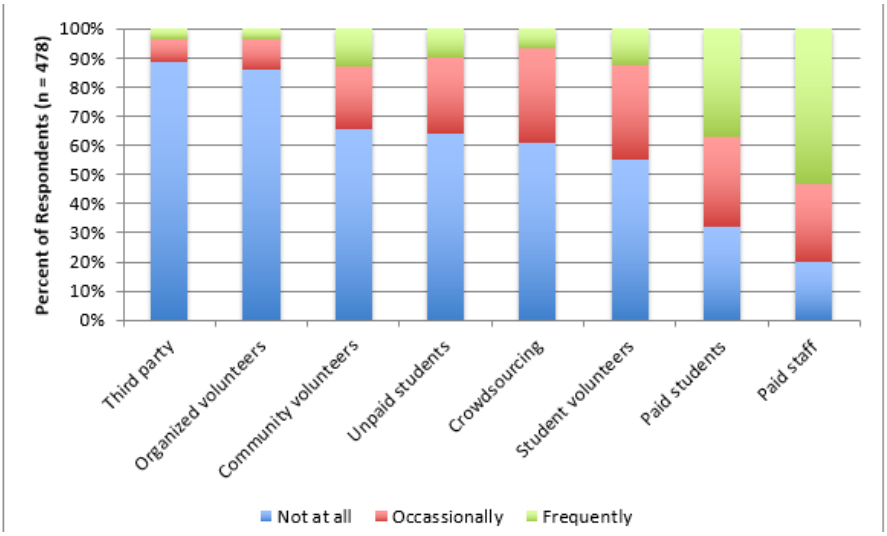
Figure 5. How collections are prioritizing digitization choices



Who is doing the digitization?

Collections rely on a variety of personnel to perform digitization tasks usually including staff, students, volunteers, and *rarely*, third party organizations (2 to 10%). Most often, digitization is being completed by paid personnel, with 90% of respondents reporting that paid staff or students "frequently" perform the tasks. While 53% of respondents reported paid staff "frequently" do the digitization, 59% reported that students (paid, unpaid, or volunteers) "frequently" digitize their collections. (Figure 6)

Figure 6. Who is doing the digitization work?



Funding

How are collections paying for this work? Of the 523 respondents who answered the question, 87% indicated that they had received at least some funding to support digitization. Of these, 69% reported that they received external funds with 61% receiving regular institutional funding. Slightly more than half (53%) reported receiving only one of these types of funding (i.e., external, institutional), while 36% received funds from two types of sources and 30% from three. Most of the external funding (80%) is from government sources with nearly three-quarters of respondents (72%) receiving funds from just one type of external source (i.e., industry, other, foundations, government).

Benefits of digitization

The top 8 responses (from 516 respondents) given for *benefits of digitization* include: increased use of collections, increased exposure, better knowledge of holdings, better management of data, digital preservation, enhanced data quality, new skills for staff, and better management of physical specimens. Thirty percent report new communities using the data, 35% see increased publicity and reduced physical handling of the collection, and 40% share awareness of an increased use of their collection data in research and publications and an increased public awareness of the importance of collections.

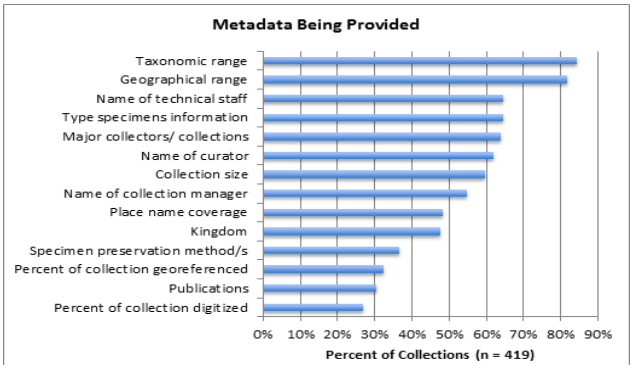
Institutional Commitment

Of the 516 individuals responding to this question, 72% indicated their institutions were committed to sustainability (or responded positively to a later question that asked about ways their institutions were sustaining digitization). Of 422 respondents, 50% are providing staff training and seeking continued funding. Over 60% have long term archival storage plans in place, and over 70% have plans in place for long term curation of the data.

Metadata

Metadata is the information about your collection/s (e.g. taxonomic, geospatial and occurrence coverage, collection contacts, etc.). As a community, what metadata do we need to effectively move forward with strategic digitization and data mobilization? To address this, we first need to know what metadata collections currently provide and consider important. 419 respondents answered the question, see Figure 7.

Figure 7 Metadata provided by collections



We also asked respondents to rank the importance of sharing each type of metadata. Over 50% marked taxonomic and geographic range of the data as critical to share and 45% marked Type Specimen data as important. Metadata values ranked as critical by 20% to 30% of respondents included: percent collection digitized, notable publications, percent georeferenced, place name coverage, name of collection manager, collection size, and name of curator. 506 Respondents report most metadata is being shared via GBIF (43%), Index Herbariorum (36%), and Institutional Websites (21%). (See recommendations in interim and final report for providing and accessing metadata). Note that reason number one in Table 4, a lack of funding and / or resources, provides a perfect reason to seek digitization opportunities *by sharing metadata* to increase discoverability / visibility.

Long term plans

Encouragingly, over 80% of 513 respondents indicate their institution / organization intends to digitize their entire collection/s. Over 30% of these same respondents share they plan to focus on digitizing specific areas of their collections in response to research needs / requests. About 12% indicate they are planning to digitize the parts of their collections that they find strategically important or unique such as type specimens or endemic species. Others (about 12 %) share they plan only to digitize new additions to the collections.

Next Steps

Please be on the lookout for our Interim Report for our recommendations and more details about our findings. Some of this work was presented in detail at The Society for the Preservation of Natural History Collections (SPNHC) 2016 Conference in Berlin in June and Botany 2016 in August – and looking for community input and feedback. We plan to release our final report near the end 2016. If you have questions or you'd like to discuss any of these points further, please do contact us through our Task Force Chair: Leonard Krishtalka (krishtalka@ku.edu) or our GBIF Programme Officer for Content Mobilisation, Siro Masinde (smasinde@gbif.org).

Survey subsection authors

Shari Ellis, iDigBio Project Evaluator, Deborah L Paul, iDigBio Digitization and Workforce Training Specialist

Reference

Maron, N. L., & Pickle, S. (2013). Appraising our Digital Investment: Sustainability of Digitized Special Collections in ARL Libraries. Retrieved from http://sr.ithaka.org?p=22363. Complete ITHAKA Survey Results here: http://www.sr.ithaka.org/wp-content/uploads/2015/08/digitized-special-collections-report-slides-15feb13.pdf

Annex 2 was emailed to survey respondents on 16 August 2016