Biodiversity Informatics, 7, 2010, pp. 93 – 112

NATURAL HISTORY SPECIMEN DIGITIZATION: CHALLENGES AND CONCERNS

ANA VOLLMAR (FORMERLY 1), JAMES A. MACKLIN (1), LINDA S. FORD (2)
(1) Harvard University Herbaria, 22 Divinity Avenue, Cambridge, MA 01238
(2) Harvard University Museum of Comparative Zoology, Cambridge, MA 01238
Correspondence e-mail: jmacklin@oeb.harvard.edu

Abstract. – A survey on the challenges and concerns involved with digitizing natural history specimens was circulated to curators, collections managers, and administrators in the natural history community in the Spring of 2009, with over 200 responses received. The overwhelming barrier to digitizing collections was a lack of funding or issues directly related to funding, leaving institutions mostly responsible for providing the necessary support. The uneven digitization landscape leads to a patchy accumulation of records at varying qualities, and based on different priorities, ultimately influencing the data's fitness for use. The survey results also indicated that although the kind of specimens found in collections and their storage can be quite variable, there are many similar challenges across disciplines when digitizing including imaging, automated text scanning and parsing, geo-referencing, etc. Thus, better communication between domains could foster knowledge on digitization leading to efficiencies that could be disseminated through documentation of best practices and training.

Key words. – Natural history collections, survey, collections, specimens, specimen data, metadata, digitization, GBIF, biodiversity research.

INTRODUCTION

Natural history collections are recognized as keepers of the primary information for the flora and fauna for both the present-day and historical record of the planet. A projected three billion or more specimens are estimated to be held in the biological and paleobiological collections of the world (Butler et al. 1998; Lane, 1999). These specimens have associated core data that is recognized as fundamental to discipline-specific research, as well as broader global issues such as invasive species, ecological/conservation issues, climate change, and emerging diseases (Araújo et al., 2005; Loarie et al., 2008; Peterson and Vieglais, 2001; Pinto, 2010; Saurez and Tsutsui, 2004; Shaffer et al., 1998; Winker, 2004).

Currently, even with demand expanding rapidly beyond traditional taxonomic/systematic research to include other academic researchers, NGOs, resource managers, governmental agencies, and citizen scientists, most of the world's holdings has not been digitized and/or made available on-line. Of the potential three billion specimens, only a small fraction have been digitized, which is evidenced by the approximately 50 million specimen records currently available through the GBIF portal (GBIF, 2010). In order to begin prioritizing data capture at a national or global level it is imperative to understand the current state of affairs of digitization initiatives at institutions housing collections. There have been other recent surveys taken on the state of natural history collections but none have specifically focused on digitization challenges (National Science and Technology Council, 2009; Synthesis, 2010). A workshop on identifying and addressing digitization bottlenecks was held at Harvard University in 2006, which brought together experts in a diverse array of natural history collections, biodiversity informaticians, and various stakeholders. The conclusions and recommendations of this report helped to focus on both the need for the survey and the questions involved (Beaman et al., 2007).

The purpose of this survey was to assess what types of digitization work were ongoing in collections, to get a sense of how resources were handled and distributed for digitization projects, and to identify the nature of impediments to digitization work and how they have been addressed in diverse collections. The data are intended to provide practical information to

curators, collection managers, and administrators at all levels about the challenges presented by digitization, and the ways in which they have been addressed, or not, in diverse collections.

METHODS

From May 20 to June 20, 2009, the survey "Natural History Specimen Digitization: Challenges and Concerns" was circulated throughout the global natural history community to curators, collections managers, and administrators. The survey was initiated by the Global Biodiversity Information Facility (GBIF) Global Strategy and Action Plan for digitization of Natural History Collections (GSAP-NHC) Task Group, in collaboration with the National Science Foundation Research Coordination Network, CollectionsWeb (www.collectionsweb.org), and the Society for the Preservation of Natural History Collections (www.spnhc.org).

In this paper, the findings from the survey will be discussed in detail, beginning with an overview of the respondents. The second section turns to broad trends emerging from open-ended questions in the survey regarding barriers to digitization, along with specific ways in which those barriers have been addressed. It discusses issues associated with having collections data available on-line, including how various collections have addressed the difficulty of ensuring feedback about their on-line data. The third section examines how funding has been distributed in collections to support digitization projects and associated staff, and presents a comprehensive list of funding sources survey respondents actively use to support digitization in their collections (see Appendix 1). Section four tackles the technological logistics of digitization, examining data entry and imaging. It also outlines the specifics of a range of equipment employed in various digitization projects and associated costs. Finally, drawing on survey and complementary interview material, the paper discusses specific digitization issues faced by different disciplines. Interviews were conducted from March to May, 2009 with staff at Harvard University Museum of Comparative Zoology (MCZ), Harvard University Herbarium (HUH), and Yale University, the Natural History Museum (NHM), UK, University of Navarra, Spain, University of Kwazulu-Natal, South Africa, and the Canadian Museum of Nature (CMN). The raw survey data is available in three files (GSAP_QUESTIONS.csv; GSAP_VAR.csv; GSAP_DB_ANON.csv) linked to on the GSAP-NHC home page on GBIF's website (http://www.gbif.org/informatics/primary-data/task-groups/gsap-nhc/)

SECTION 1: OVERVIEW OF SURVEY RESPONDENTS

In total, 201 respondents completed the survey. Reflecting a heavy bias in the geographic distribution of the survey, 62% were from North America, 22% from Europe, 9% from South and Central America, 3% from Asia, 3% from the Australasian region, 1% from the Middle East, and none from Africa. More than half (62%) of respondents were from institutions/museums affiliated with a university, 23% were from national or government-affiliated museums/institutions, and 8% were from free-standing museums. Other respondents had private collections (2%) or were affiliated with non-profit, non-governmental organizations that maintained collections (2%). Collections represented ranged greatly in size, from just 17 specimens up to 21 million specimens. The median collection size was 200,000 specimens.

When asked to describe their position within the collections where they worked, nearly half of respondents selected more than one title to characterize the work they did, suggesting that respondents often played many roles within their collections and did diverse kinds of work. Respondents who identified themselves as curators or assistant curators made up 47%, 39% were collections managers or collections assistants, 28% identified as researchers, and 18% were data managers, database technicians, or biodiversity informaticians. Another 12% had more administrative positions as directors or assistant directors of museums/institutions, and as program directors.

As with the diversity of roles respondents played in their collections, 43% of respondents worked with collections in multiple disciplines; 55% of respondents worked with botany collections; 26% with entomology collections; 18% with invertebrate zoology collections; 14-16% with each of herpetology, ichthyology, mammalogy, mycology, and ornithology; and 10%

with each of vertebrate and invertebrate paleontology. Other disciplines represented by survey respondents include geology (7%), mineralogy, anthropology, and archaeology (5% each), and ethnobotany, paleobotany, historical scientific instruments, and scientific photograph archives (less than 1% each). Since almost half of the respondents worked either in multiple collections or collections encompassing multiple disciplines, drawing correlations between particular disciplines and digitization barriers, practices, or the like will not always be straightforward.

SECTION 2: BARRIERS TO DIGITIZATION

Overall 94% of respondents reported digitization and/or imaging ongoing in their collections in the past two years, but the results also highlight impediments. When asked whether or not digitization work was *currently* ongoing in their collections, only 79% of respondents answered affirmatively, the remainder citing largely similar reasons for why digitization work was at a standstill. Respondents (n=44) ranked the reasons why digitization was not ongoing in a collection; funding was the primary reason, followed in descending order of importance by time, staff, lack of institutional support, infrastructure/technology, and curation practices. When respondents were sorted into categories based on job responsibilities, this ranking remained largely stable. Similarly, the least important reasons among respondents also were consistent, and included issues of data sharing, lack of collecting permits, sensitive species information, and indigenous rights (see Table 1).

Table 1. What are the primary reasons digitization work is not ongoing in your collection? Ranking is on a scale of 1 (most important) to 5 (least important).

All respondents (n=44)	Answering for institution (n=7)	Answering for specific collection (n=28)	Directors (n=7)	Curators/collection managers (n=39)
Funding (1.3)	Staff (1.2)	Funding (1.1)	Funding (1.0)	Funding (1.3)
Time (1.7)	Funding (1.3)	Time (1.5)	Staff (1.2)	Staff (1.6)
Staff (1.7)	Lack of institutional support (1.8)	Staff (1.9)	Time (1.3)	Time (1.6)
Lack of institutional support (2.1)	Time (2.2)	Lack of institutional support (2.5)	Lack of institutional support (1.7)	Infrastructure/ technology (2.3)
Infrastructure/ technology (2.4)	Curation practices (2.4)	Infrastructure/ technology (2.5)	Curation practices (3.0)	Project complete (2.3)
Data sharing (3.9)	Data sharing (3.6)	Data sharing (4.2)	Project complete (4.0)	Data sharing (3.7)
Collecting permits (4.2)	Collecting permits (3.6)	Collecting permits (4.2)	Storage practices (4.3)	Sensitive species data (4.0)
Sensitive species data (4.2)	Sensitive species data (3.6)	Sensitive species data (4.4)	Collecting practices (4.3)	Collecting permits (4.1)
Indigenous rights (4.4)	Indigenous rights (4.8)	Indigenous rights (4.5)	Indigenous rights (5.0)	Indigenous rights (4.5)

Among 171 respondents from collections in which digitization was either ongoing or had occurred within the past two years, the major impediments to digitization identified were similar to the reasons named by other respondents as to why digitization was not ongoing. Funding, time, and staff were consistently the top three challenges faced, followed by issues of data entry, data quality, and georeferencing. These latter three in particular suggest a need for accessible guidelines and suggestions about how to structure digitization projects. Collecting practices, sensitive species information, collecting permits, and indigenous rights were the least important barriers to digitizing collections (see Table 2).

Their status as "least important" suggest these integral issues for collections in general take a backseat to more immediate logistical constraints and concerns of digitization. In an open-ended question about the impediments to digitization work, respondents elaborated on their initial rankings, giving more detailed descriptions of these identified barriers facing their digitization projects.

FUNDING

Other survey respondents pointed to budgetary issues and the myriad of problems related to the lack of funding. One respondent from Israel described this succinctly: "[a lack of funding] did not allow good professional programmers. The process was long and not successful [so] we returned to Excel and Access data entry. [Further, the] budget enables student work only [and] they

Table 2. If digitization is ongoing in your collections or has happened within the past two years, in your experience have any of the following issues impeded digitization work in your collection? Ranking is on a scale of 1 (most important) to 5 (least important).

All respondents (n=171)	Answering for institution (n=21)	Answering for specific collection (n=104)	Directors (n=33)	Curators/ collection managers (n=132)	Data Managers (n=32)
Funding (1.5)	Funding (1.5)	Time (1.6)	Time (1.4)	Funding (1.5)	Funding (1.5)
Time (1.6)	Staff (1.7)	Funding (1.6)	Funding (1.6)	Time (1.5)	Time (1.7)
Staff (1.8)	Time (1.9)	Staff (1.8)	Staff (1.7)	Staff (1.7)	Staff (1.8)
Data entry (2.5)	Georeferencing (2.2)	Data quality (2.6)	Data entry (2.0)	Data entry (2.6)	Infrastructure/ technology (2.3)
Data quality (2.6)	Data quality (2.4)	Data entry (2.6)	Georeferencing (2.2)	Infrastructure/ technology (2.6)	Georeferencing (2.5)
Collecting practices (3.8)	Storage practices (3.3)	Collecting practices (3.9)	Curation practices (3.8)	Collecting practices (4.0)	Collecting practices (3.7)
Sensitive species data (3.9)	Indigenous rights (3.8)	Sensitive species data (4.0)	Collecting practices (3.9)	Sensitive species data (4.0)	Storage practices (3.9)
Collecting permits (4.2)	Collecting permits (3.8)	Collecting permits (4.3)	Collecting permits (4.1)	Collecting permits (4.3)	Collecting permits (4.1)
Indigenous rights (4.4)	Collecting practices (4.1)	Indigenous rights (4.5)	Indigenous rights (4.1)	Indigenous rights (4.4)	Indigenous rights (4.3)

tend to leave after a while. Teaching them each time the job is very time consuming. They need closer supervising and [the] data requires more attention for quality." Respondents also identified obtaining funding sources as barriers to digitization, writing that "digitisation [is] seen as impossible to apply [for] funding" (Denmark) and that "it is quite hard to find funds that pay just for the digitization of natural history collections" (Netherlands). In an effort to get around funding crunches and incorporate digitization into daily curatorial tasks, one respondent described a system in which "data entry and imaging are now project-based, so that herbarium data and loan requests are answered simultaneously with the advancement of digitization. In other words, digitization has become an opportunistic activity that piggybacks on more pressing day-to-day herbarium service to clients" (Canada). Funding will be further addressed in Section 3.

STAFF

Closely related to funding, if not inseparable, is the importance of staff to digitization efforts, particularly well-trained staff. As respondents emphasized repeatedly in their responses to the survey, people are key to successful digitization projects. A respondent from Canada wrote, "Staff is the greatest limitation for digitizing the collection. There is a steep learning curve for accuracy and speed of data entry." Other respondents alluded to social difficulties, writing that impediments included "attitudes (digitization being seen as scientifically unproductive)" (Denmark), and "cultural entrenchment" (USA). Getting staff (at all position levels) on board with digitization projects was seen as a key element, and this was accomplished by "training and helping the head curators and staff understand why certain methods are in place…and moving everyone (including the 'old guard') forward" (USA). Respondents also mentioned changes such as "hiring a curator concerned about digitization" (USA) and making a "change in The Management view to [the] value of digitization" (Malaysia) as important in increasing success.

Still others cited the importance of "workflow" or "staff organization and work planning" as significant to ensuring efficient digitization processes. For instance, thoughtful placement of disciplinary experts and non-experts throughout the process can be exceedingly helpful, as non-experts who know little about a given discipline can frequently do the first level of data entry, marking sites of doubt and uncertainty for experts to look over at a later time.

In many contexts, data entry staffing can prove a huge bottleneck in the digitization process but, even this fundamental step, may be conditional. Michelle Hamer (pers. comm. 2009) of the South Africa National Biodiversity Institute and the South Africa GBIF node made note of the low volume of applications for digitization funding coming to SAGBIF. She identified this particular bottleneck for digitization not as a lack of funding or available staff, but rather that receiving funding would necessitate fulfilling the terms of a grant, and this was rendered difficult in her region given significant barriers in terms of a lack of resources even for basic curation needs.

To get a digitization project off the ground, involved staff must be well trained in the use and possibilities of the data capture client and/or whatever other technologies are being used in the digitization process. Training requires time and money, both of which are limiting factors along with the question of who, in a given institution, is able and available to train staff. As indicated in interviews and by survey responses, this person is often the data manager or biodiversity informatician, which can divert effort from other necessary informatics work. Creating local tutorials that explain, for example, the purpose of each field in a database and "data dictionaries," which address common uncertainties encountered during data entry, can be very helpful in dispersing the training load and reducing the number of input errors.

Finally, databases in which staff must log in to modify or create new entries allow collection managers to track how long certain tasks take and how efficiently particular people are working. Digitization of specimens, thus, not only expands possibilities for managing collections, but also the possibilities for managing the people who work in collections. Because digitization of collections expands the access to collections, there are often increases in the number of visitors and more requests for information and loans. This can become an issue when a collection is understaffed, and highlights the importance of good planning in any digitization project; departments and

institutions must be aware and plan for the changes in the collections' use once databases are available to a broader forum.

SPACE AND CURATION

Respondents reporting on their institutions at large emphasized issues of physical space and money. On the difficulties presented by physical spaces, a respondent from Austria cited as a hurdle the "size of collections [and the] complex structure of the very heterogeneously historically grown collections dating back to the beginning of the 19th century." While many older institutions face this same challenge and are engaged in renovating buildings and internal spaces, the spatial organization of any collection impacts not only ease of use but also the process of digitizing collection data. Collections are frequently organized by mixtures of taxonomic and geographic hierarchies, while at the same time also shaped by the constraints of particular physical spaces, by the ways in which people like to use collections, and by the storage practices demanded by the nature of the specimens themselves. Ideally, the spatial and curation needs of a collection can to some extent be assisted during the digitization process. This can happen in a variety of ways, with parts of collections being reorganized and updated or, as practiced in an ornithology collection in the USA, collections are closely monitored for pests as part of digitizing specimens from their labels.

Curation of collections involves navigating the space in which a collection physically resides and, frequently, the present-day management is dependent on its historical curation. Current work is always constrained by past decisions about how to gather and record information, and the spatial decisions of how to store and organize specimens. In collections that have had many curators with different practices, specialties, and goals, the levels of curation throughout a collection can be diverse, making digitization a challenge. Curators might employ different numbering systems or organizational practices, and almost always have different research specialties that in turn shape where resources are focused. Collections with a greater continuity of curation throughout their history, with fewer staff changes (at all levels), or with curators and collection managers who are in agreement about the goals and management of a collection, can be far more effectively and efficiently digitized.

DATA ENTRY AND DATA QUALITY

Digitization (as distinct from imaging) frequently begins with recording specimen data. This basic data entry involves interactions between the person doing the data entry and with both the specimens and the data-capture client. Often catalogues and ledgers are targeted first for data entry because the data are more accessible, with a format essentially that of a basic spreadsheet. As one respondent (USA) noted, the "absence of bulk data sources such as ledgers limit[s] data capture to the handling of individual collections objects," which is significantly more time consuming. Bulk data sources, however, are not without problems. In older collections or collections that have had many curators, there are frequently multiple catalogues with overlapping or duplicate numbers for different parts of a single collection. Specimens can be entered into the database using prefixes or suffixes attached to catalogue numbers in order to differentiate objects with the same number, allowing each specimen to be renumbered uniquely. In addition, if data are entered directly from ledgers and catalogues rather than from the specimens themselves, once the collection has been databased, it is often necessary to go back into the collection and see which specimens listed in the ledger/catalogue are actually present in the collection. Even if a specimen is missing, however, its data should be recorded since the specimen may be recovered and the data, itself, has scientific merit although diminished without the voucher. For some collection types (e.g., botanical, entomological), there may only be specimen information on the objects themselves, or information may be located in multiple places. Finally, decisions must be made about whether or not to capture data from all possible places for individual specimens, weighing the benefits of more complete data capture against constraints of time, resources, and efficiency.

A usable and efficient interface for data entry that limits keystrokes and streamlines the process is to customize the appropriate fields required for data entry. Frequent dialogue between the people doing the data capture and those doing the programming is key to crafting data-entry clients that are smooth functioning and tailored to the

needs of individual disciplines and digitization projects. However, not all collections are able to work with programmers or have access to more complex clients and databases, and in such cases two survey respondents cited using Microsoft Excel and/or Access as clients that "increased efficiency of data entry" (Israel) and proved "easey [sic] for everyone – even non-educated people – to use and understand. Later other programs can be in use" (Denmark).

The quality of the specimen data as recorded on original sources, such as specimen labels, ledgers, card catalogues, and field notes, can present a variety of stumbling blocks to digitization projects. Transferring data from specimens, ledgers, and catalogues often necessitates deciphering poor handwriting, translating labels in foreign languages, making inferences about dates (e.g., day or month first), and decoding place names (e.g., historic locations no longer exist, place names change). In short, data that are not precise at the point of capture can be tricky to fit into more determined database structures. However, uncertainties in collection data must be digitized as they are originally recorded. For instance, updating names of places or taxonomic determinations without including the original designations can result in data being tied anachronistically to places and things, or to a loss of geopolitical information. Coming up with ways of recording uncertainties in the original data, and subsequent inferences or interpretations made by people doing data entry, are key for maintaining high quality data. Tagging uncertainties and problems for later review by qualified staff is a common practice in collections with staff entirely dedicated to data entry, but who may not be familiar with the discipline itself.

Streamlining workflow for data entry and imaging, solving data problems, and reducing the number of steps in handling and databasing specimens was another important issue for respondents. One respondent from Australia "[made] label generation for specimens a product of databasing, not an additional task prior to databasing." A tactic employed by a respondent from the USA was to "database 1 specimen per species as a 'first pass' to obtain a complete taxon checklist for the collection before returning to database the remaining…90% of specimens." Another respondent (USA) reported doing a "pre-digitization critical assessment of specimens, and elimination of low/no-data specimens." Others "standardized and documented the digitization protocol" (Spain) and "wrote [a] protocol manual before data entry began" (USA). Backlog, which can present difficulties in data capture, was addressed by a respondent in the Netherlands by "divid[ing] backlog digitisation of collection labels in two phases, i.e. initial image plus basic data[,] and second other data and field notes."

TECHNOLOGY

Funding can significantly limit options for using various technologies for digitization, and outdated technology and equipment can significantly slow if not halt projects. Addressing technology was an important issue for a respondent in Nicaragua, who "changed the old computer. The significant barrier is the low network we have." Other issues included the "need to create workflow to handle issues during data capture" (USA) and software being diverse (difficult to integrate data entered in different systems/projects) and/or underdeveloped. For collections with access to funds, improvements to the technologies used were key to addressing digitization bottlenecks, including the implementation of a system of unique identifiers like barcodes, purchasing digital cameras and scanners, using voice recognition software, using optical character recognition (OCR) software to enter data, joining multiuser collection management systems, using international standards like DarwinCore, and improving cataloguing software.

COLLECTIONS DATA ON-LINE: BENEFITS, CHALLENGES, AND FEEDBACK MECHANISMS

In total, 60% of all survey respondents reported that at least part of their collections data were available on-line (see Table 3).

Table 3. What are the reasons your collections data are not available online? Responses were open-ended.

	Respondents from university institutions (n=42)	Respondents from government institutions (n=19)
Staffing issues	21%	not reported
Lack of IT resources	26%	58%
Digitization and data cleaning incomplete	43%	32%

Out of the respondents with at least some on-line data, 55% were affiliated with government institutions, 62% were affiliated with university institutions, and 80% were from free-standing institutions reported that some of their collections' data were available on-line. Forty-six respondents elaborated on why their data were not yet available on-line, and the reasons largely fell into two categories. First, respondents expressed the plan to put all of their data on-line at once, and thus needed to digitize more data and check its quality before making it available (13 respondents from various countries including, in alphabetical order, Brazil, Canada, Malaysia, Netherlands, New Zealand, Spain, Sweden, and USA). Second, respondents cited a lack of reliable Internet service, web servers, and website/software support, or that the necessary IT infrastructure did not exist at their institutions (14 respondents from Canada, India, Jamaica, Malaysia, Netherlands, Spain, USA, and Uruguay).

Various reasons were cited for why collection data were not available on-line. For respondents affiliated with government institutions, the reasons included the lack of web servers, support, and IT resources, and digitization was not yet complete. For respondents from university-affiliated institutions, data were not on-line because digitization and data clean-up was not yet complete, there was not enough staff to tackle projects, and web servers and IT support were not available. Also mentioned were the lack of time and/or money, both of which might be translated into general issues of staffing and equipment. Only 3 respondents from free-standing institutions answered the open-ended question, mentioning issues of accuracy of data, institutional policies, and lack of a suitable website.

In reflecting on the benefits to their institutions of having their collections data available on-line, respondents cited an increase in use of collections and requests for data (31 out of 70 respondents, 44%), and a general heightened visibility of collections. Data availability to the general public and to remote researchers were also considered assets. Increasingly, preliminary research could be conducted on-line, and so respondents reported that questions and loan requests were more specific, resulting in less physical handling of the collection, thereby extending the longevity of the collection. Respondents also noted the benefit of data correction, and receiving feedback on errors and/or misidentifications from outside users. Additionally, one data manager from Austria mentioned the opportunity for "virtual repatriation of material to countries of origin."

Some respondents, answering the survey with respect to their institutions as a whole, reported that they had not experienced any problems since making their collection information available on-line, however, another reported: "the only downside is a LOT more work coming in: editing the data, answering questions, checking data entry, revisiting determinations, etc" (USA). Many respondents reported experiencing an increase in workload due to heightened visibility of collections, and did not have adequate staff to handle queries. Finding the funding, staff expertise, and support to keep databases up and running was also noted as a challenge.

Many curators and collections managers mentioned errors and data quality as primary issues, citing difficulties in checking data prior to posting on-line, and expressing concern at the public accessing potentially flawed data. Others (16 out of 57) cited technological problems as their top issues, ranging from a lack of IT support and training for staff, to minimal server functionality, a poor institutional network system (Colombia), and limited storage facilities for high-resolution digital images (Austria). Exposure of sensitive species data was also an issue, which

respondents reported on occasion was accidentally revealed and, at other times, was requested by researchers, necessitating an evaluation of the research. At the same time, revealing data to politically charged entities was also a concern. One respondent from Australia described their policy: "Registration is required to access specimen details; this sometimes requires arbitration to fairly address and assign access for some applicants (especially from the community or mining industries)."

One of the most significant challenges presented by data sharing is how to navigate/ensure returns on projects that share information for free. Curators and managers of collections put a great deal of time, energy, and resources into digitizing data and making it available to larger and, significantly, more remote audiences. Collection managers frequently give voice to the idea that collections are "alive" because researchers and specialists physically work with the collections, redetermining taxonomic identifications, rearranging parts of the collection, etc. While increasing remote access and on-line use of collections clearly has many benefits, at the same time, it presents difficulties in ensuring that feedback about collections is received, in particular that the work researchers do with collections data on-line make it back to the collection itself. Another challenge expressed was keeping track of who uses the data and attribution for its use. As one respondent (USA) wrote, "[I] saw a paper published referencing only data from "Ornis" [an ornithology resource pooling specimen data from many institutions] but not specifying which collections actually contributed data. We had a dozen relevant specimens to that study but were not able to determine if they were used." Another respondent (Sweden) also mentioned, "People may think what is on-line is all we have, although only a small percentage is databased."

Respondents reflecting on how they received feedback about their individual collections broadly cited two mechanisms: voluntary feedback links/forms, and restricting access by log in. More specific iterations of these tactics included conspicuously located contact information, user surveys, collection agreement pages that users must navigate through in order to access the data, requests for reprints of publications drawing on data, and required membership in order to access data. Others reported offering limited or basic specimen data, thus requiring users to contact collection managers for further detail. Nineteen out of 57 respondents (33%) noted that there was no mechanism in place to solicit feedback from on-line users.

SECTION 3: FUNDING

The primary impediment to digitization was reported as the lack of funding which is intimately linked to the people, who do digitization, in the form of salaries, and to the technology and infrastructure of digitization through the necessary purchase of equipment. In this section, we examine where different kinds of institutions look for funding, how respondents prioritized spending those funds, and how staff work time was utilized for digitization projects. We also include a list of the specific funding sources cited by respondents as providing significant resources for digitization projects.

In reporting sources of funding used for digitization projects within the last two years (since 2007), 69% (136) of respondents received internal institutional funding, 54% (107) received public funding, and 30% (59) received private funding. No official funding was received by 4% of the respondents; 3% used either personal income or pursued digitization projects in free time and 1% drew on volunteer efforts. On average, respondents received 53% of their funding from internal institutional sources, 49% from public sources, and 23% from private funding sources. Respondents' answers did not add up to 100% in this question since monies were received from multiple sources, and so the percentages of each funding category do not add up to accordingly. While the limitations of these proportions as exact measures must be recognized, they do give a general sense of where collections are finding support for digitization projects.

When applying for funding, 72% of all respondents reported that they explicitly requested funds for digitization projects, equipment, or people. Among respondents identified as directors of institutions/programs, 86% requested funds specifically for digitization projects, as did 88% of data managers, 78% of researchers and faculty, and 69% of curator/collections managers. Among respondents requesting support for collections

digitization projects, staff were ranked as their top need, followed by money, and then equipment. It should be noted that "money" in this context can not practically be distinguished from either staff or equipment needs.

When there is no funding for digitization projects, 48% of respondents reported reallocating resources from other jobs or projects to support digitization work. Other sources of resources used to cover the costs of digitization projects included annual budgets, internal and departmental funding, salaries of staff, funds available to hire students, endowments, and volunteers. Many respondents described including digitization as a routine part of collection maintenance activities, or even collecting expeditions. Finally, three very dedicated respondents reported using their own personal funds! For a list of funding sources respondents actively used within the past two years to support digitization projects in their collections, please see Appendix A.

In assessing how they would prioritize managing their collections in general if given more money, respondents identified specimen curation as their top priority (average ranking of 1.7, with 1 being most important). Second most important were research (2.2) and collection storage/equipment (2.2). Education was a distant priority at 3.2. Additional priorities mentioned included digitization (30 respondents), specimen acquisition (5 respondents), increased staff (4 respondents), and space (4 respondents).

When asked to reflect on how they would use additional funding for digitization, respondents commonly voiced the need to "simply digitize collection data," and this was often paired with an increase in staff as key to enabling the success of data capture projects. Seventy-nine out of 180 respondents said they would direct funds toward hiring more staff to work on digitization and data entry and, as per one respondent (USA), "Staff hours – time is of the essence." Respondents (40) also said they would allocate funds for improved equipment and technology, purchase more equipment, better software, and more digital storage space, and one (USA) wanted to "develop methods for the curation of digital media associated with specimens (e.g., field notes, digital images, radiographs, etc.)."

Despite the broad consensus among respondents that staff were essential, only 55% of respondents indicated that staff currently performing digitization work had such tasks specified in their job descriptions. Of staff who performed digitization work that was not specified in their job descriptions, 72% reported doing so as an extension of their daily tasks. Thus, although staff are widely recognized as perhaps the most important aspect of digitization efforts, for many this work is not formally recognized as a part of their job.

SECTION 4: TECHNOLOGY AND DATA ENTRY

In this section, we look at the ways in which collections are stored which affects access for digitization, the sources from which data are being entered, how long data entry takes from the various sources, and what kinds of collections are not using unique identifiers. We briefly look at georeferencing and imaging, and then close with an overview of various technologies and their costs being used for digitization and imaging as reported by survey respondents.

On average, of the collections reported, 56% of specimens were stored as dried and pressed on sheets, 33% of specimens were pinned, 26% were fluid specimens in vials or bottles, 16% were dried and in packets, 13% were skeletons and bones, 11% were skins and hides, 10% were slides, 7% were dried in vials or bottles, 7% were fluid in tanks, 6% were taxidermy mounts, 5% were tissues, and 4% were cleared and stained (wet skeletal preparation). Most respondents (74%) reported that their collections utilized two or more different ways of recording information about

Table 4. If digitization work has been ongoing in your collection within the past two years, from what sources are/were you entering data?

	Respondents entering data from each source (n=185)
Specimen labels	90%
Catalogues	41%
Ledgers	28%
Literature	24%
Cards	21%

Table 5. On average, how long does it take you to enter data from each of the following sources? How would you characterize the number of data fields you are entering per specimen?

	Low (1-10 data fields)		Medium (11-20 data fields)		High (20+ data fields)	
	Respondents entering data from each source	Time to enter data	Respondents entering data from each source	Time to enter data	Respondents entering data from each source	Time to enter data
Specimen labels (n=161)	15%	5-9 minutes	59%	5-9 minutes	27%	5-9 minutes
Catalogues (n=71)	15%	5-9 minutes	57%	5-9 minutes	28%	5-9 minutes
Ledgers (n=49)	16%	5-9 minutes	55%	5-9 minutes	29%	5-9 minutes
Literature (n=44)	16%	10-14 minutes	55%	10-14 minutes	30%	20+ minutes
Cards (n=36)	8%	1-4 minutes	66%	5-9 minutes	26%	5-9 minutes

specimens in their collection. It should be noted that respondents' answers did not have to add up to 100% in this question (e.g., one specimen with multiple preparations), and so the percentages of each storage/prep type do not add up accordingly. While the limitations of these proportions as exact measures must be recognized, they do give a general sense of the physical nature of the specimens in collections of respondents surveyed. It should be noted that this does not reflect a global proportioning of storage or prep types.

The majority of respondents (95%) reported that they used a database to record data about specimens in their collection, and information was recorded from various sources (see Table 4).

For the average time and number of data fields entered from each of these sources, see Table 5.

Additionally, 63% of all respondents reported georeferencing specimen data, although only 50% of those were doing so according to best practice guidelines or standards. Apart from data entry, 64% of respondents surveyed reported that they were doing some imaging of specimens, which for most took anywhere from less than 5 minutes to 10 minutes (see Table 6).

Most (91.5%) of all respondents said they assigned unique identifiers to specimens in their collection. Of those specimens associated with unique identifiers, on average 70% were associated with catalogue numbers and 22% with barcodes. In the final section, we will touch on some of the challenges faced by various disciplines in assigning unique identifiers.

DIGITIZATION AND IMAGING TECHNOLOGY

When reflecting on the primary factors they considered when purchasing digitization equipment/technology, on a scale from 1 (most

Table 6. If you are imaging specimens, how long does it take to image a specimen?

	Respondents reporting the length of time to image specimens
Less than 5 minutes (n=37)	30%
6-10 minutes (n=38)	30%
11-15 minutes (n=16)	13%
16-20 minutes (n=9)	7%
20+ minutes (n=15)	12%

Table 7. What equipment are you using for digitization projects in your collections and, if known, approximately how much did each piece of equipment cost? If you are imaging, what technology are you using to image and, if known, approximately how much did it cost?

Scanners	Price	Software	Price
HerbScan	--	Geo Locate	--
INDUS Book Scanner	$25,000	ImageMagick	--
Epson Expression 10000XL	$3,000	Nikon Capture NX	--
Microtek Scanmaker 9800XL	$2,000	Phase Capture One	--
Nikon Super Coolscan 5000	$1,200	Robogeo GPS	--
Digital Cameras	**Price**	SinarCapture Shop	--
Fuji S2 Pro	--	Leica Automontage	$3,000
Nikon Coolpix 995	--	Adobe Design Suite/ Photoshop	$400
Nikon D90	--	**Other Equipment**	**Price**
Nikon D100 plus 100mm macro lens	--	Barcode printer	--
Sony DH-5	--	Barcode scanner	--
Tethered Sony A900	--	Camera stand	--
Digital photomicroscope	--	Lighting equipment	--
X-ray imaging	--	Sound recording devices	--
Jenoptik Eyelike M22 camera back and Schneider APO-Digitar 90mm/f 4.5 lens with TTI camera stand/ lighting	$60,000	File server space (3 terabytes)	$3,600
Syncroscopy AutoMontage	$60,000	Apple Mac Pro Computer, 2x2.8Ghz	$5,312
Sinar Evolution 75H Multi Shots Digital Back System 33 mp	$39,794	Datamax printer for archival quality specimen labels	$4,000
Leica MZ16A Stereomicroscope with Leica Motor Focus and a Leica DFC 420	$30,000	Microchip labels and microchip scanner	€3,000
TTI-Repro-Graphic Workstation 3040/Digiflex 67ei/ Sinar 75H	$28,117	External hard drives for backup	$1,000
Leica MZ10	$14,000	Portable terabyte storage drives for archiving specimen scans	AU$600
Large format camera with BetterLight digital back	$10,000	Beseler copystand	$500
Coloreal Ebox	$3,000		
Canon Rebel XTi	$1,000		
USB microscope camera	€70		

important) to 3 (least important), respondents' top concern was quality (1.4), followed by suitability to project (1.5), compatibility with existing/future technology (1.66), ease of use (1.7), cost (1.8), and durability (1.9). Other factors reported included speed and efficiency (4% of respondents), availability of support/staff expertise (2% of respondents), and compliance with funding sources' requirements (2% of respondents). The proximity of importance of these factors suggests that none of them takes full precedence, and that they are significantly dependent on the needs of any given digitization project.

In Table 7, the technologies used by respondents in both digitization/data entry and imaging projects are summarized. All costs were reported by respondents in US Dollars unless otherwise specified, and have not been confirmed by any further external research. The equipment used reflects a wide variety of uses and also resource availability, with both $400 digital cameras and $60,000 imaging set-ups included in the survey results.

SECTION 5: DISCIPLINARY CONCERNS

In this section, particularities of collections in various disciplines are addressed. Because many survey respondents worked in multiple collections and answered the survey with respect to more than one scientific discipline, survey data are less useful for addressing discipline-specific trends. To supplement the survey data, interviews with a variety of staff at the Harvard University Museum of Comparative Zoology, Harvard University Herbarium, and the Yale University Peabody Museum were conducted to help form the foundation of this section. This section is not intended to establish the definitive characterizations of specific disciplines, since some observations from the interviews may be more reflective of challenges faced by individual collections and institutions. Nevertheless, many of the issues noted cross institutional boundaries and give a general sense of the problems facing specific types of collections.

BOTANY/MYCOLOGY

Storage types: Dried and pressed on sheets, dried in packets, dried fruit and seed collections ethnographic artifacts, petrified wood, fluid specimens, fossil specimens.

A total of 119 respondents indicated that they worked with either botanical or mycological collections, or both. Many of these respondents also worked with other disciplines, so observations based strictly on discipline are somewhat general. On average, these respondents reported that 77% of their specimens were dried and pressed on sheets, 18% were dried and in packets, 6% in fluid-filled bottles/vials, and 3% slides. Respondents also mentioned live plant collections, wood collections, specimens on rocks, fossils, and photographs. The majority of botany/mycology respondents (96%) said they used unique identifiers in their collection, and 84% reported there was digitization work ongoing.

Barriers to the digitization of herbarium specimens begin with the number of specimens: material frequently comes in faster than can be databased, leading to a significant backlog that is compounded in larger collections by the volume of specimens already contained within a particular collection. Because herbarium sheets are relatively large pieces of paper, this means there is often abundant data to capture, making the process of digitization time consuming, and the possibility likely of encountering information that does not fit particular fields in a given data capture client. Accordingly, 94% of respondents were entering data from specimen labels, 25% from catalogues, 21% from ledgers/accession records, and 21% from literature. Most respondents (62%) were georeferencing specimen data, and of those, 48% were doing so according to best practice guidelines/standards. A majority of the respondents (59%) were imaging specimens, although the relative importance/abundance of imaging is unknown. See Table 8 for data entry statistics.

Table 8. Data entry statistics for botany/mycology collections.

On average, how long does data entry take per specimen from each data source? (n=96)		On average, how would you characterize the number of data fields per specimen you enter?		Approximately how long does it take to image a specimen?	
Data source	Time	Data fields	Percent of respondents (n=100)	Time	Percent of respondents (n=66)
Specimen labels	6.1 minutes	Low 1-10 fields	12%	5 minutes or less	41%
Catalogues	7.5 minutes	Medium 11-20 fields	63%	6-10 minutes	32%
Ledgers/ accession records	6.2 minutes	High 20+ fields	25%	11-15 minutes	9%
Cards	6.6 minutes			16-20 minutes	6%
Literature	9.8 minutes			20+ minutes	3%

Additional barriers—none of which are unique to botanical specimens—include illegible handwriting, poor documentation, synonymy, space, and multiple sheets of a single specimen. Ethnobotanical collections—particularly historic ones—present a unique set of challenges as ethnobotanical collecting methods and associated documentation have been largely unstandardized, and such collections encompass a wide range of storage types (herbarium sheets, raw materials, artifacts, fruits and seeds). The ethical dimension of ethnobotanical collections involves questions of repatriation and indigenous rights that are faced by anthropological and archaeological collections.

ENTOMOLOGY

Storage types: Pinned insects, Riker mounts, papered specimens (dried and folded in envelopes), fluid specimens, frozen tissues.

A total of 55 respondents indicated that they worked with entomological collections. Some of these respondents also worked with other disciplines, so observations based strictly on disciplines are somewhat generalized and exact proportions of collection storage/prep types could not be calculated. In general, entomological specimens are mounted on pins and organized in drawers, which may not always be organized in lots (i.e., many individuals of the same species from the same collecting event) and may contain multiple species since groups of insects are frequently clustered together in smaller boxes called unit trays. Information about specimens can thus be associated with a specimen itself, with a unit tray, or with a drawer as a whole; parsing these different levels of information into a database can be challenging.

Entomological collections rarely have ledgers or card files; instead, nearly all data are physically associated with the specimen itself, usually on small pieces of paper pinned below the specimen. Specimen data are often recorded in nonstandard shorthand that abbreviates the location and date of collection, and species identification for each specimen. In some cases, this information is recorded in a purposefully cryptic manner to hide collecting locations, and so deciphering these labels can be quite difficult, often requiring specialized knowledge accumulated over many years of involvement in the field. Getting the necessary species data off the specimen and into a database is, thus, a time-consuming process; pins have to be removed, and specimens handled. For the same reason, adding items like barcodes to individual specimens is physically challenging. Thus, it is not unsurprising that of the survey's 18 respondents who said they did not assign unique identifiers, 14 were from entomological collections (housed in university-affiliated institutions). Eight respondents reported that

digitization work was *not* ongoing in their collection, with the top reasons cited as lack of funding, time, institutional support, staff, and infrastructure and technology. Eight respondents also did not have data from their collections available on-line, largely because data had not yet been digitized, and web servers/Internet service were unreliable or absent.

Many of the respondents in entomological collections (69%) reported that there was some kind of digitization work ongoing in their collection. An ongoing Lepidoptera imaging project at the Museum of Comparative Zoology, Harvard University uses barcodes that are double sided so that if they are not readable from above, the insect can be removed, turned over, and the barcode read without any further manipulation of the specimen. Additionally, the barcodes used are readable even if punctured with a pin. Because the specimen data is in shorthand and data entry personnel generally will be unable to interpret the meaning of what is written on each label, the project plans to utilize crowd-sourcing (i.e., upload images of specimen labels to a wiki-style website) entomologists all over the world—particularly amateurs—can then submit interpretations of labels and specimen information to the project.

MAMMALOGY

Storage types: study skins, cased skins (preparations without cuts to the abdomen), skeletons, fluid specimens, taxidermy mounts, histological slides, embryos, frozen tissues, observation data and measurements of each mammal.

A total of 38 respondents worked with both mammalogy and ornithology collections, and 33 with only mammalogy collections. Some respondents also worked with other disciplines (especially herpetology and ichthyology). Because of the nearly complete overlap between respondents working with mammalogy and ornithology collections, discipline-specific observations from the survey are problematic.

With mammalogy collections, cataloguing and databasing specimen information is not the most time consuming part of incorporating mammals into the collection; rather, the limiting step is the time it takes to prepare mammals. Specimen preparation can take a relatively long time for large mammals, so databasing is only a brief moment in a much longer process.

An abundance of diverse kinds of data about each specimen must be entered into a database, presenting challenges in crafting appropriate data entry interfaces. From the collecting end of the process, this can be streamlined by clear communication between collections managers and collectors in the field about what kinds of information are necessary to gather; one individual gives field-kits with data checklists to collectors so as to make data entry and cataloguing a smoother process.

ORNITHOLOGY

Storage types: skins, skeletons, fluid specimens, egg and nest lots, taxidermy mounts, frozen tissues, histological slides.

A total of 38 respondents worked with both ornithology and mammalogy collections, and 36 with only ornithology collections. Problems with discipline-specific observations are as noted previously.

One element of digitizing ornithology collections that requires attention is the need to associate different parts of particular specimens that are stored scattered throughout collections (e.g., skeletons with skins, or photographs of birds prior to collection and the resulting skins). Fully digitizing information about particular specimens necessitates physical sleuthing and cross-referencing within collections. Accommodating the different kinds of data that must be entered depending on the kind of specimen can also be a challenge, such as efforts to database eggs and nests, which incorporate observation data. A nest and its eggs are considered a "lot," and the information recorded includes the number of eggs per nest, the location of the nest (e.g., ground or tree and, if tree, then height), how the species was identified, if the adult was seen while collecting, and how long the eggs were incubated prior to collection. In addition, the ability for databases to incorporate auditory data of bird sounds adds another layer of complexity and possibility to digitizing ornithology collections.

ICHTHYOLOGY AND HERPETOLOGY

Storage types: fluid specimens (including in metal tanks), skeletons, cleared and stained specimens, frozen tissues, histological slides, taxidermy mounts. Specific to herpetology are skins/hides, and turtle shells.

A total of 44 respondents worked with both herpetology and ichthyology collections, 35 with only herpetology collections and 32 with only ichthyology. Respondents also worked with other disciplines (especially ornithology and mammalogy). The majority of herpetology/ichthyology collections (85%) had digitization ongoing. While many specimens are individuals, both ichthyology and herpetology collections contain lots, multiple individuals (ranging from two to thousands) gathered together in a single collecting event and catalogued as a single specimen. Each individual organism does not have a unique identifier, unless the specimen is examined for research purposes, in which case ideally each is identified individually. Digitizing lots presents challenges in terms of the number of subdivisions of the organism, and the levels of part enumeration supported by various databases.

Most of the respondents (68%) in herpetology/ichthyology collections were imaging specimens. Significant challenges for both ichthyology and herpetology collections are the issues faced in imaging fluid specimens, which can be time consuming. Some of the respondents (19%) reported that it took them less than 5 minutes to image a specimen, 41% said it took 6-10 minutes, 22% took 11-15 minutes, 4% took 16-20 minutes, and 11% took more than 20 minutes. For example, some specimens must be removed from their jars and entirely submerged in fluid to create a single focal plane. Large and oddly shaped specimens like snakes present still further difficulties.

The majority of the respondents (71%) working with herpetology/ichthyology collections reported georeferencing specimen data. Of those, who were georeferencing, however, only 53% were doing so according to best practice guidelines/standards. Georeferencing presents a number of issues for collections with localities tied to water. Historically, fresh water collections were often not identified with latitude and longitude but rather descriptions of particular places, while oceanographic collections are identified with longitude and latitude, but precision was less important. Precision becomes of utmost importance in cases like collections tied to rivers. Rivers can be thousands of miles long so, without further identifying information, it can be impossible to discern where along a river a specimen was collected. Different styles by collectors of describing localities create challenges for digitization efforts attempting to enrich specimen data by assigning precise locations to specimens.

Georeferencing is further complicated because the error radius frequently employed to designate uncertainties in location demarcates a circular area. Many of the geographic water features referenced in fish collections are not perfectly round and are near land. If the error radius is applied without consideration to the water boundaries, land will be included within the error radius. Collections at rivers frequently occur at bridges and other sites of crossing; thus, a more useful designation of uncertainty would only extend upstream and downstream rather than in a circle radiating from the assumed point of collection. Specimens collected in coastal locations can likewise be confusing since an error radius might encompass both fresh and salt water habitats, and land. Georeferencing tools that are linked to data about terrain and the presence or absence of water are more useful for ichthyology specimens.

INVERTEBRATE ZOOLOGY

Storage types: Fluid specimens, dried sponges and corals, shells, specimen slides, and frozen tissues.

A total of 21 respondents worked with invertebrate zoology collections. These respondents also were heavily involved with other collections, particularly ornithology and herpetology (18 respondents) and mammalogy and ichthyology (16 respondents). Responses by discipline, then, are difficult to gauge in this case for reasons discussed earlier.

One of the challenges in digitizing invertebrate collections is numbers: there are so many species and specimens that it is hard to database them all. Invertebrates are also often collected in large quantities and grouped in lots.

Lots in invertebrate collections can contain thousands of organisms, which presents difficulties when loaning such specimens, as typically each organism in a lot is counted before being loaned. As a result, many lots have only approximations of the number of organisms they encompass. In databases, multiple levels of enumeration in lots (i.e., if they are subdivided) can be hard to accommodate and track.

VERTEBRATE AND INVERTEBRATE PALEONTOLOGY

Storage types: Fossil specimens, microfossils on slides, slabs and oversized specimens on carts. Specific to VP are trackways and fossil skeletons.

A total of 29 respondents worked with vertebrate and invertebrate paleontology collections. These respondents were also involved in other collections, particularly herpetology (16), mammalogy (15), and ornithology (14). Again, responses by discipline are difficult to gauge. Locality data are of utmost importance in digitizing paleontology specimens; without locality data, a specimen is nearly worthless because location is as important as a specimen's taxonomic identification. However, digitizing paleontological specimens requires the inclusion not only of information about collection locality, but also geologic information like unit, age, series (upper, middle, lower), formation, and beds/members/zones. In particular, invertebrate zoology stratigraphic collections—which give data about specimens through time and in various locations—are difficult to database because of their more temporal orientation.

Specific to vertebrate paleontology is the challenge of describing what part of a specimen a given object might be. Because vertebrate paleontology specimens are composed of diverse and complex objects, over time layers of narrative attempting to describe objects can become an incomprehensible description. Attaining consistency in description, which is key to describing and identifying specimens, is thus critical.

SECTION 6: CONCLUSIONS

This paper serves to highlight the many challenges and concerns relevant to digitizing natural history collections. The detailed findings provide both a status quo for how specimens are being digitized presently, and a window into the issues that need to be resolved in order to break down barriers and make the digitization process more efficient. The survey strongly suggests that the greatest barrier to digitization is the cost of doing the work, from hiring staff to purchasing technology. The cost barrier is not a simple one to overcome as many respondents noted that there are very few funding sources available to tap. The burden of digitization, thus, generally falls to the resources at hand and institutional and/or collection priorities, making the process slow and very uneven across the collection landscape.

This survey also highlighted that although each kind of collection has some important domain-specific issues, there are many challenges that are common to all. Thus, communication of knowledge about digitization among disciplines could have great benefit toward overcoming common challenges, such as imaging, automated text scanning and parsing, georeferencing, etc. Indeed, the survey respondents suggested that there was a great need for more communication of knowledge on digitization through training and documentation of best practices. In addition, many respondents, most likely from smaller institutions, reported already working with multiple collections, especially in the vertebrate and paleontological disciplines, which would facilitate the spreading of information across collections. Community standards and collaborative efforts need to be further developed and embraced by those who create the software that the collection community relies on. The added benefit would be that newly created components or modules could then be linked together into workflows, which could address the specific needs of a particular collection.

The barriers to digitization ultimately lead to a patchy set of digitized records being available to potential users, which can clearly be demonstrated by searching the GBIF portal. This reality reduces the availability of specimens for research and broader purposes. This patchiness also extends to the amount and quality of the data records being captured, which has a direct impact on the data's fitness for use. Reduction of fitness can negatively impact the data's application to broader issues and the power of the conclusions on which any analyses are based.

Although the social barriers to digitization were not specifically addressed in the survey, it became quickly apparent that these issues were prevalent. In no small part because of the social nature of natural history collections, we hope that some elements of this survey and paper can serve as practical tools for collections embarking on digitization projects, such as approximating how long it will take to digitize a ledger based on the number of data fields. We also hope that the survey results illuminate the paucity at the level of implementation and will help show the need for, and serve as a guide to, funding sources for digitization projects.

Given the abundance of natural history specimens and the potentially broad importance of their associated data ranging from discipline-specific research, to public initiatives, to global issues, digitization must be undertaken as strategically as possible. A recurring theme in the survey was that digitization efforts must have the most impact in the shortest amount of time, and for a reasonable cost; this appeal, by its very design, requires coordination within and among collections and institutions. Much as collecting and specimen acquisition has always been a social endeavor, so must be the effort to render those same collections digital.

ACKNOWLEDGEMENTS

The authors would like to acknowledge funding for the Society of Preservation of Natural History (SPNHC) Best Practices Intern, Ana Vollmar (first author), from Alan Prather, PI of the National Science Foundation Research Coordinating Network:CollectionsWeb (# 0639214) and support from the Global Biodiversity Information Facility (GBIF). We would especially like to thank the following collection professionals who were interviewed, including (in alphabetical order): Arturo Ariño (Spain), Roger Baird (CMN), Adam Baldinger (MCZ), Susan Butts (Yale), Judith Chupasko (MCZ), Jessica Cundiff (MCZ), Rod Eastwood (MCZ), Brendan Haley (MCZ), Michelle Hamer (South Africa), Karsten E. Hartel (MCZ), Shusheng Hu (Yale), Eric Lazo-Wasem (Yale), Paul Morris (HUH and MCZ), Chris Norris (Yale), Malcolm Scoble (NHM), Patrick Sweeney (Yale), Jeremiah Trimble (MCZ), Gregory Watkins-Colwell (Yale), Andrew Williston (MCZ), Jonathan Woodward (MCZ), and Krzysztof Zyskowski (Yale).

REFERENCES CITED

Araújo, M.B., Pearson, R.G., Thuiller, W. and Erhard, M. 2005. Validation of species-climate impact models under climate change. Global Change Biology, 11, 1504-1513.

Beaman, R., Macklin, J.A., Donoghue, M.J., and Hanken, J. 2007. Overcoming the Digitization Bottleneck in Natural History Collections: A summary report on a workshop held 7-9 September 2006 at Harvard University. http://www.etaxonomy.org/wiki/images/b/b3/Harvard_data_capture_wkshp_rpt_2006.pdf [Accessed September 11, 2010].

Butler, D., Gee, H. and Macilwain, C. 1998. Museum research comes off list of endangered species. Nature: Briefings, 394: 115-117. [DOI:10.1038/28009].

GBIF. 2010. Global Biodiversity Information Repository Portal. http://data.gbif.org/welcome.htm [Accessed September 11, 2010].

Lane, M.A. 1999. Weaving a Web of Wealth: Biological Informatics for Industry, Science and Health. Australian Academy of Science, Canberra, Australia. 40 pp.

Loarie, S.R., Carter, B.E., Hayhoe, K., McMahon, S., Moe, R., Knight, C.A., and Ackerly, D.D. 2008 Climate Change and the Future of California's Endemic Flora. PLoS ONE 3(6): e2502. [doi:10.1371/journal.pone.0002502].

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, 2009. http://www.whitehouse.gov/sites/default/files/sci-collections-report-2009-rev2.pdf [Accessed September 11, 2010].

Peterson, A.T. and Vieglais, D.A. 2001. Predicting Species Invasions Using Ecological Niche Modeling: New Approaches from Bioinformatics Attack a Pressing Problem. BioScience 51(5): 363-371 [doi:10.1641/0006-3568(2001)051[0363:PSIUEN]2.0.CO;2].

Pinto, C.M., Baxter, B.D., Hanson, J.D., Méndez-Harclerode, F.M., Suchecki, J.R., Grijalva, M.J., Fulhorst, C.F., and Bradley, R.D. 2010. Using museum collections to detect pathogens [letter]. Emerg. Infect. Dis. [serial on the Internet]. http://www.cdc.gov/EID/content/16/2/356.htm [Accessed September 11, 2010], [doi:10.3201/eid1602.090998].

Shaffer, H. B., Fisher, R. N. and Davidson, C. 1998. The role of natural history collections in documenting species declines. Trends in Ecology and Evolution 13: 27-30.

Suarez, A.V., and Tsutsui, N. D. 2004. The value of museum collections for research and society. BioScience 54: 66-74.

Synthesis. 2010. Synthesis of Systematic Resources. http://www.synthesys.info/II_na.htm [Accessed September 11, 2010].

Winker, K. 2004. Natural History Museums in a Postbiodiversity Era. BioScience 54(5): 455-459.

APPENDIX A

Funding sources used within the past two years to support digitization projects in the collections of the survey respondents.

- Accessprojektet, Sweden
- African Plant Initiative/Mellon Foundation
- Amherst College, USA
- ARTstor
- Atlantic Canada Conservation Data Center
- Australia Virtual Herbarium Project
- Australian Government
- Barcelona University, Spain
- Batson Endowment for the A. C. Moore Herbarium, USA
- Biology Department, Texas A&M
- BioMAP Project
- Boeing, USA
- Canadian Foundation for Innovation
- Canadian Museums Association
- Census of Marine Life
- Concejo Distrito Capital, Colombia
- Concejo Nacional de Ciencia y Tecnología de Guatemala (CONCYT), Guatemala
- Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
- Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil
- Darwin Initiative, UK
- Department of Conservation, Terrestrial and Freshwater Biodiversity Information System (TFBIS) fund, New Zealand
- Dirección General de Investigación (DIGI), Universidad de San Carlos, Guatemala
- Dutch Scientific Foundation
- Earthwatch Institute
- Environment Canada
- Environmental Foundation of Jamaica, Virtual Herbarium Project
- European Distributed Institute of Taxonomy (EDIT)
- European Union Funds
- Finnish Ministry of Education
- Fisheries and Oceans Canada
- Friends of the University of Alberta Museums, Canada
- Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP), Brazil
- Fundação para a Ciência e Tecnologia, Portugal
- Global Environment Facility (GEF)-Andes Project
- Global Biodiversity Information Facility (GBIF)
- Government of Jamaica
- Government of Newfoundland and Labrador, Canada
- Government of Spain
- Government of the Netherlands
- Häagen-Dazs
- Hanes Trust
- Harvard University, USA
- Hearst Scholarship Foundation, California State University, USA
- Institute of Museum and Library Services (IMLS), USA

- Inter-American Biodiversity Information Network (IABIN)
- Israel Academy of Sciences
- Kansas State University, USA
- Latin American Plant Initiative (LAPI)/ Mellon Foundation
- Louisiana Board of Regents, USA
- Max Planck Gesellschaft, Germany
- Mellon Foundation
- Ministerie van Onderwijs, Cultuur en Wetenschap, Netherlands
- Ministerio de Ciencia e Innovación, Spain
- Ministerio de Educación y Ciencia, Spain
- Ministry of Science and Education, Spain
- Mondriaan Foundation, Netherlands
- Museum Assistance Program, Canada
- Museum of Comparative Zoology, Harvard University, USA
- NaGISA program and GoMA program (Alfred P. Sloan foundation)
- National Cancer Institute, USA
- National History Museum, Stockholm, Sweden
- National Science Foundation-Biological Research Collections (NSF-BRC) grant, USA
- National Science Foundation, USA
- Netherlands Organization of Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO)
- New Brunswick Wildlife Trust Fund, Canada
- New Mexico State Department of Fish and Game, USA
- Norwegian Development Agency (NORAD)
- Norwegian Ministry of Foreign Affairs
- Overseas Territories Environment Programme, UK
- Penn State, USA
- Plant Health Australia
- Pollinators Thematic Network (PTN) initiative, IABIN
- Red Nacional de Información Académica, Colombia
- RENATA, Colombian Ministry of Education
- Riksbankens jubileumsfond Formas, Sweden
- Secretaría de Estado de Medio Ambiente y Recursos Naturales (SEMARENA), Dominican Republic
- Servizio Civile Nazionale, Italy
- Smithsonian Trust Funds, USA
- Southeast Regional Network of Expertise and Collections (SERNEC), USA
- Svenska Artprojektet, Sweden
- Swedish Species Information Center, ArtDatabanken
- Swedish University of Agricultural Sciences
- The Swedish Taxonomy Initiative
- UNESCO-l'Oréal for Women in Science Fellowship
- United States Bureau of Land Management
- United States Department of Agriculture (USDA)
- USDA Current Research Information System (CRIS) project funds, USA
- United States Fish and Wildlife Service
- United States Forest Service
- United States National Park Service
- Universidad de León, Spain
- Universidad de San Carlos de Guatemala
- University of Florida, USA
- University of Iowa, USA
- University of Malay, Malaysia
- University of New Brunswick Department of Biology, Canada
- University of Wisconsin Natural History Museums Council, USA
- van Eeden, A. Mennege, H. de Vries, van Leersum Foundations, Sweden
- Young Canada Works in Heritage Institutions