Palaeontologia Electronica
palaeo-electronica.org

Digitization workflows for paleontology collections

Talia S. Karim, Roger Burkhalter, Úna C. Farrell, Ann Molineux,
Gil Nelson, Jessica Utrup, and Susan H. Butts

ABSTRACT

The development of digitization workflows is an essential part of any formalized large-scale digitization program. Paleontological collections literature has addressed the need for, and utility of, digitized collections for nearly four decades, but no modern, community-vetted set of digitization workflows to accomplish this goal has been widely adopted. With the advent of the U.S. National Science Foundation's (NSF) Advancing the Digitization of Biodiversity Collections (ADBC) program in 2011, iDigBio, NSF's national coordinating center for facilitating digitization, in collaboration with broad community representation from numerous institutions, launched a series of working groups to address workflow development across all major preparation types. Workflow modules have been developed for pre-digitization curation, data entry, imaging objects (catalogs, field notes and other materials not stored with specimens, labels, two- and three-dimensionally preserved specimens), image processing, and proactive digitization. Modules and the tasks they include may be implemented in any order and customized for specific configurations and institutional parameters. The workflows are made publicly available for download and customization at GitHub and via the iDigBio documentation pages. A review of platforms for electronic data publishing through online aggregators, a crucial step in any digitization program, is also provided.

Talia S. Karim. University of Colorado Museum of Natural History, Boulder, Colorado 80503, USA. talia.karim@colorado.edu
Roger Burkhalter. Sam Noble Museum, University of Oklahoma, 2401 Chautauqua Avenue, Norman, Oklahoma 73072, USA. rjb@ou.edu
Úna C. Farrell. Department of Geological Sciences, Stanford University, 450 Serra Mall, Stanford, California 94305, USA. ufarrell@stanford.edu
Ann Molineux. Non-vertebrate Paleontology Lab, University of Texas, 10100 Burnet Road, Austin, Texas 78758, USA. annm@austin.utexas.edu
Gil Nelson. Department of Biological Sciences, Florida State University, Tallahassee, Florida 32303, USA. gnelson@bio.fsu.edu
Jessica Utrup. Yale University, Peabody Museum of Natural History, Division of Invertebrate Paleontology, 170 Whitney Avenue, PO Box 208118, New Haven, Connecticut 06520, USA. jessica.utrup@yale.edu
Susan H. Butts. Yale University, Peabody Museum of Natural History, Division of Invertebrate Paleontology, 170 Whitney Avenue, PO Box 208118, New Haven, Connecticut 06520, USA. susan.butts@yale.edu

Karim, Talia S., Burkhalter, Roger, Farrell, Úna C., Molineux, Ann, Nelson, Gil, and Butts, Susan H. 2016. Digitization workflows for paleontology collections. *Palaeontologia Electronica* 19.3.4T: 1-14
palaeo-electronica.org/content/2016/1569-digitization-workflows

Keywords: workflow; digitization; imaging; data publishing; paleontology; collections

Submission: 4 May 2015 Acceptance: 2 August 2016

INTRODUCTION

The development and documentation of effective workflows is a primary component in the digitization of neontological and paleontological collections (Granzow-de la Cerda and Beach, 2010; Vollmar et al., 2010; Beaman and Cellinese, 2012; Haston et al., 2012; Nelson et al., 2012; Weiss, 2013; Wolniewicz, 2009). Nevertheless, there is no current workflow documentation outlining best or recommended practices for the digitization of paleontological collections. Until recently, discussions dealing with computerization focused mostly on electronic database design (e.g., Morris, 2000) or defining the role digitization should play in the paleontology community (MacLeod and Guralnick, 2000). In the last decade, large scale digitization projects in the United States (with iDigBio) and Europe (with the EUROPEANA Portal and more recently, Synthesis 3) have initiated a number of conversations and collaborations about how best to capture specimen data quickly, publish the data in an easily accessible manner with sufficient data quality (for review see Smith and Blagoderov, 2012), with discipline-specific workflows developed and published (e.g., herbaria, Nelson et al., 2015) as a product of these initiatives. However, digitization of paleontology collections is often not straightforward, as specimens vary in size and dimensionality, no published guide for standard specimen orientations exists, specimens often sit on top of labels obscuring data, and data is often stored in multiple formats and in multiple locations (e.g., labels with specimens; locality cards, field notes, and ledgers away from specimens). A case study by Weiss (2013) involving collaboration between a paleontology collection and library scientists recognized several additional concerns regarding digitization of paleontological collections, including preservation and access of data, metadata and thoroughness of the digital record, and the fidelity of rendering fossils in 2-D and 3-D images. We address these issues with the generic workflow modules presented herein and offer suggestions on how to turn some of these challenges into opportunities for more comprehensive digitization of a collection.

BACKGROUND

Paleontologists started discussing “electronic data processing” nearly four decades ago, foreshadowing the advent of digitization and digital data storage and display (see e.g., Glenister, 1977; Mello, 1969; Mello and Collier, 1972) and many paleontology collections have had some form of digital specimen database for decades. The Paleontological community has also long recognized the need for publishing these data and making them discoverable via aggregation sites. The PaleoPortal, initially funded in 2003 by the National Science Foundation, included one of the early attempts at aggregation of digital museum records via DiGIR protocol. It served as a single access point for searching multiple collections across the United States. In May 2012, Integrated Digitized Biocollections (iDigBio), Yale Peabody Museum, Botanical Research Institute of Texas, and the Biodiversity Institute at the University of Kansas co-sponsored the Developing Robust Object-to-Image-to-Data (DROID) workflow workshop in Gainesville, FL (iDigBio, 2012a) for the purpose of designing consensus workflows across various domains within the biodiversity collections community. Concomitant with the planning and execution of this workshop, staff from iDigBio conducted a series of on-site visits to a number of museum and academic collections to document existing digitization workflows within natural history collections (Nelson et al., 2012). The preliminary results of this survey were incorporated into the DROID workshop (Nelson, 2012). The primary outcome of the DROID workshop was a series of working groups focused on the production and distribution of preparation-specific digitization workflows for (1) flat sheets and packets, (2) pinned objects in trays and drawers, and (3) 3-D objects in spirits, jars, drawers, and boxes (iDigBio, 2012b). Nelson and Karim both participated in the original DROID workshop and the subsequent development of workflows by the 3-D objects working groups. This working group eventually split into two separate working groups to develop more specific workflows for fluid preserved specimens and fossil specimens, respectively. The iDigBio Paleontology Digitization Working Group also organized a series of workshops/symposia (Table 1) and webinars that

TABLE 1. Summary of iDigBio Paleontology Digitization Working Group workshops and symposia contributing to workflow development.

WORKSHOP/MEETING	DETAILS	LOCATION & DATE
Developing Robust Obect-to-Image-to-Data (DROID) Workshop	**Sponsors:** iDigBio, Yale Peabody Museum, Botanical Research Institute of Texas, Biodiversity Institute-University of Kansas. **Goal:** design consensus workflows across various collection types.	University of Florida, Gainesville, FL May 2012
Paleontology Digitization Workshop	**Sponsors:** iDigBio, Yale Peabody Museum, Fossil Insect Collaborative TCN. **Goal:** discuss digitization efforts across the paleontological community, gather feedback on DROID workflow development.	Yale Peabody Museum, New Haven, CT September, 2013
Paleontology Imaging Workshop	**Sponsors:** iDigBio, Jackson School of Geoscience-University of Texas. **Goal:** share and demonstrate various imaging techniques used for fossil specimens, including 3D scanning, CT scanning, and photogrammetry, and gather feedback on DROID workflow development.	University of Texas, Austin, TX April, 2014
Specify for Paleontology Collections	**Sponsors:** iDigBio, Biodiversity Institute-University of Kansas. **Goal:** provide training on how to use the Specify 6 collections database software for paleontology collections, gather feedback on how to structure "paleo context" data within the Specify 6 data model, and gather feedback for DROID workflow development.	Biodiversity Institute-University of Kansas, Lawrence, KS May, 2014
Digitization Symposium- Advancing the Digitization of Paleontology and Geoscience Collections: Projects, Programs, & Practices	**Sponsors:** iDigBio, GSA-Geoinformatics Division, Paleontological Society, Society for the Preservation of Natural History Collections, University of Colorado Museum of Natural History, Yale Peabody Museum. **Outcome:** 34 oral and 4 poster presentations sharing implementations of digitization projects across the geosciences.	Geological Society of America (GSA) Annual Meeting, Vancouver, BC October, 2014

addressed various aspects of paleontological specimen digitization with the aim of gathering ideas from the broader paleontological community about their workflows and to draw these efforts together into a comprehensive document of workflow modules. The results of these working groups were then synthesized into generalized workflow modules for digitizing various aspects of paleontological collections and they are presented herein.

DIGITIZATION WORKFLOWS

Following the lead of DROID, output from previous iDigBio-sponsored workflow working groups, and the digitization task clusters enumerated by Nelson et al. (2012), the paleontology digitization workflows working group pursued a modular approach to workflow development. The four primary modules for which workflows were developed are listed in Figure 1.

Figure 2 presents, in modified business process modeling format, one implementation of these modules in an illustrative workflow that essentially parallels the sequence listed above. The figure depicts the main workflow in the central lane with one-time and episodic tasks shown in adjacent parallel lanes. It also emphasizes (with a loop symbol) which of the modules encompass iterative processes. Note that quality control and conservation are associated with more than one module, as will likely be the case in any specific implementation.

These workflows are not introduced as *best practices*, a term that we define to mean broadly vetted, rigorously tested, and critically measured

Workflow Modules

Module 0: Pre-Digitization Curation
Module 1: Data Entry
Module 2: Imaging
  Module 2A: Imaging catalogs, field notes, and other materials not stored with specimens
  Module 2B: Imaging labels associated with specimens
  Module 2C: Imaging three-dimensionally preserved specimens
  Module 2D: Imaging two-dimensionally compressed specimens
  Module 2E: Image processing
Module 3: Proactive Digitization

FIGURE 1. List of workflow modules developed by the iDigBio DROID and Paleontology working groups.

protocols and procedures in use with substantial consistency across a community of users. Instead, the workflows presented here constitute a set of recommended practices drawn from numerous published and unpublished sources and known practices of institutional leaders in the digitization of neontological and paleontological specimens; they are reflective of common, successful, and implementable practices.

The format of each module is standardized into four columns and *n* rows (Table 2). Column 1 includes the Task ID of each task to be accomplished in the workflow. Tasks are numbered sequentially. Column 2 contains the name of the task, sometimes expanded into a short statement that describes the activity to be accomplished. Column 3 elaborates on the tasks; often expressing variations in strategies for executing the task, details of the task, and further explanations of what the task requires and/or is intended to accomplish. Column 4 itemizes the resources needed to accomplish the task, mentions relevant institutional protocols and manuals, and frequently includes literature citations, links to related websites, and resources such as software and equipment.

The modules address all aspects of digitization and provide a framework within which an undigitized collection could incorporate curatorial work and conservation. They also provide alternative approaches for those currently digitizing but uncertain how to handle a particular type of media, and provide other solutions to digitizing problems experienced in their present protocol. The detailed documentation for these modules is available at iDigBio's website (iDigBio, 2016).

The decision to use a modular approach in the development of these workflows stems from their generic nature. Given the breadth in database management systems, imaging equipment and configurations, institutional infrastructure, variations in physical facilities, goals for digitization programs, and other differences between institutions, a modular approach facilitates institutional adaptations and encourages the creation of workflows customized to the institution (Haston et al., 2012). Editable and PDF versions of these workflows are also available for download and customization at: github.com/iDigBioWorkflows/PaleontologyDigitizationWorkflows.

For simplicity in accounting for all tasks in a particular module, we have used a linear rather than iterative representation, although iterative sub-processes are occasionally referenced within a particular task's explanations and comments section (see third column of Table 2). For instance, in Module 2D–Imaging Two Dimensional Compressed Fossils, steps T4-T11 are listed as being iterative in the explanations and comments section for step T4. This set of steps, which includes tasks such as lens setup, specimen cleaning and positioning, and adjusting lighting, can be completed in a continuous series for each specimen and then repeated for multiple specimens, or can be adjusted as needed for a particular project. Although this approach risks obscuring embedded iterative sequences within a particular module, it reflects our goal of providing a comprehensive menu of digitization tasks while leaving specific implementations and order of execution at the discretion of users.

Some processes referenced in the workflows (e.g., focus stacking) are dependent on wide variations in specimen size and composition, making a universally relevant accounting of workflow steps difficult. We have highlighted these processes within the workflows, usually providing generic guidelines for implementation accompanied by relevant references. Recommendations of specific software is beyond the scope of these workflows, but we do reference common software used in paleontological collections digitization.

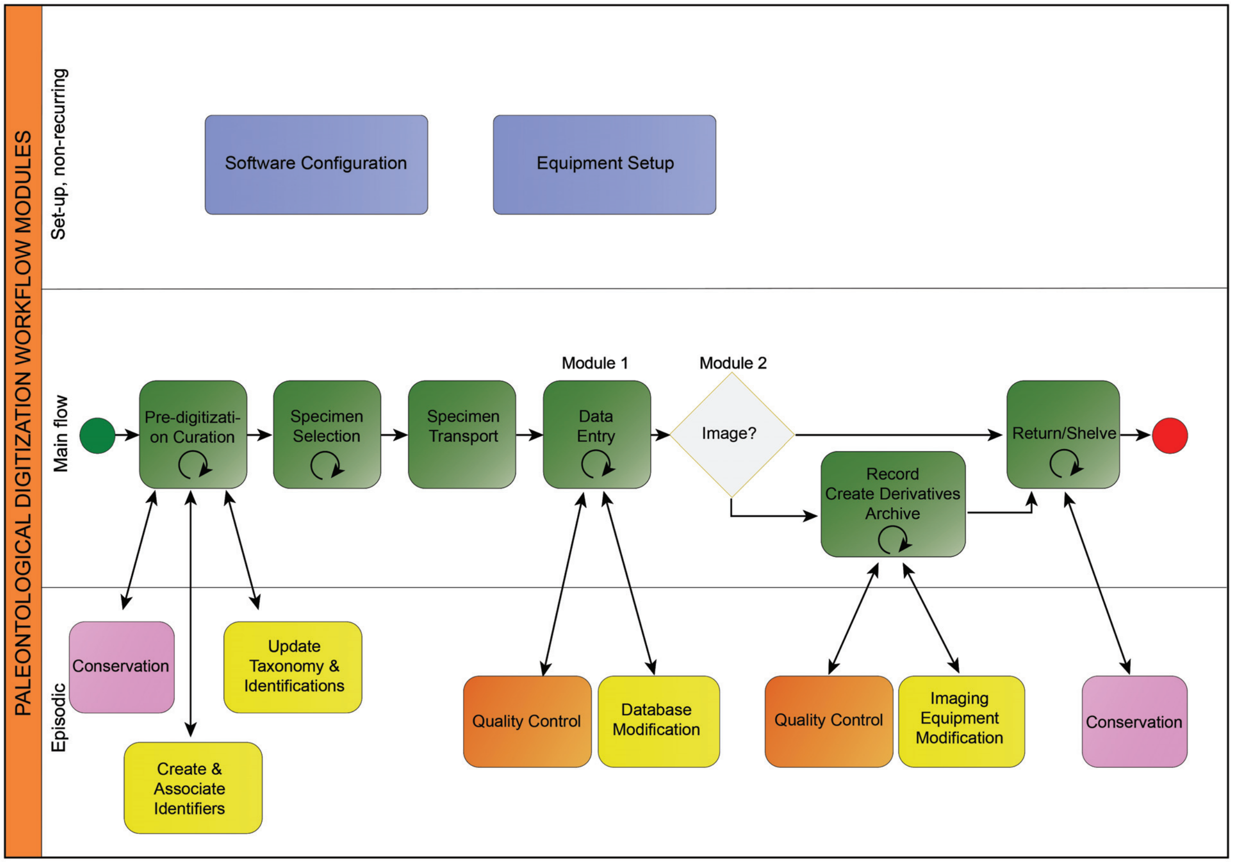

FIGURE 2. Example implementation of the workflow modules described herein.

Collections Digitization Policy/Strategic Plan

Tasks within several of the workflow modules suggest the existence of an institutional strategic plan or a collection-level policy or framework for guidance in making curatorial and digitization-related decisions. Detailed, well thought out, and written digitization policies are important in planning a digitization program (Scoble, 2010). They are commonplace in many biodiversity collections (New York Botanical Garden, 2004) and recommended herein as an essential component for ensuring consistency of practice across staff and time. Their development should precede the launch of a digitization program, and their revision should be ongoing to reflect emerging technologies, new discoveries in techniques and processes, and continual refinement of workflows. Topics to be covered within these documents include: digitization rationale, digitization goals, data and metadata standards (Häuser et al., 2005), image standards, technology replacement and data management plans, project planning and management, a digital asset management plan, and standards for data publication. The digital asset plan template created by the National Museum of Natural History is useful in developing such a plan: www.idigbio.org/wiki/images/2/20/NMNH_Digital_Asset_Plan_Template.pdf.

Decision making at the institutional level can also be useful in finding solutions to long-term data storage and data publishing. The University of Colorado Museum of Natural History has worked with the campus Research Computing Facility to acquire large amounts of storage for images and other digital data being generated by a wide variety of digitization projects across the museum. Several museum staff members have continued to work with Research Computing to develop data management tools desired by the museum as a whole. This relationship between Research Computing and the museum was made possible by leveraging support from museum leadership and cross disciplinary recognition at the collections-level of the need for long-term data storage. Similarly, Yale University has developed a collaborative framework for supporting the lifecycle management and use of Yale's digital assets. Collections from multiple units within the university can now be searched

TABLE 2. Example of modular workflow format from Module 0: Pre-Digitization Curation and Setup.

Module 0: Pre-Digitization Curation and Setup

Task ID	Task Name	Explanations and Comments	Resources
T1	Prioritize specimens, collection objects, ledgers, field notes, and catalogs to digitize.	Varies by institution. Should follow institutional digitization policies and guidelines.	Institutional policy, project guidelines, active research criteria, etc.
T2	Note damage to object to be digitized that needs immediate attention.	Route to conservation workflow as necessary, based on institutional policy or curatorial practices.	Institutionally specific curation guidelines.
T3	Update specimen taxonomy (and related authority files) as necessary.	This may happen prior to the digitization of any taxonomic group. Some institutions update specimens with expert determinations prior to digitization. Others record determinations from the label in anticipation of community involvement in helping correct determinations.	According to institutional protocol and procedures and/or project requirements.

from a single platform that combines digitized information from the art galleries, libraries, academic departments, and the Yale Peabody Museum. This platform can be explored at: discover.odai.yale.edu/ydc/.

Workflow Modules

Pre-digitization curation and set up. The pre-digitization curation and set up module (Module 0) constitutes a group of tasks to be accomplished prior to digitization (Figure 2). This module addresses numerous curatorial tasks that might otherwise be put off until time is available, resulting in substantial opportunity for collection improvement. Topics covered include: assessing curatorial and digitization priorities; selecting and configuring a database management system; updating controlled taxonomies within a database; making and updating taxonomic determinations; defining data entry standards; determining the types of objects to digitize (e.g., specimens, labels, field notes, ledgers, catalogs, etc.); assessing and noting specimen/collection object damage, and repairing if time allows and damage is minor, or routing to conservation if necessary; determining image file naming conventions; assigning/updating local catalog numbers; selecting and implementing a strategy for creating, assigning, and associating globally unique identifiers (GUID) to specimens/collection objects.

We emphasize the importance of assigning catalog numbers at this stage before digitization has started as this piece of information is required for most digital records (e.g., specimen records in a database) to be created. These locally unique catalog numbers also ensure that no two specimens within a single collection or institution bear the same identifier, and the uniqueness of these identifiers is usually enforced by database conventions. For instance, many database programs require that entries into the catalog number field be unique, thus the software will not allow you to create multiple specimen records with the same catalog number. Using this built-in safety net works well for ensuring that duplicate catalog numbers are not assigned within a single collection.

GUIDs on the other hand, are universally unique and ensure that no two specimens across the universe of collections bear the same identifier. As records of specimens are aggregated with those from other institutions, the use of GUIDs will ensure that all references to a specimen can be resolved to a single physical object stored in a specific natural history collection. The critical point here is that the GUIDs should be assigned to the specimen records by their initial creators. Retroactively physically affixing a GUID to every specimen in a collection is not feasible, especially for large collections. A more realistic option is for a GUID to be assigned to each electronic specimen record by the institution creating the record, and in fact many collections management software packages already do just this (e.g., Specify [Specify Software Project, University of Kansas, Lawrence; specifyx.specifysoftware.org/] and EMu [Axiell Group, Sweden; emu.kesoftware.com/] will automatically assign a GUID to each specimen record). The critical point here is that the GUIDs should be assigned to the specimen records by their initial creators and not secondarily by subsequent data aggregators so as to avoid records possessing multiple GUIDs. The GUID can then be tracked back to the physical specimen via the institutionally unique catalog number, which is also a part of the specimen record.

Many different types of GUIDs exist (see e.g., Page, 2009; Richards et al., 2011; Guralnick et al., 2014), but there is no clear consensus on which type should be used broadly for natural history collections (see Guralnick et al., 2015). One possible option for a resolvable identifier for fossil specimens would be to use an International Geo Sample Number (IGSN) (www.geosamples.org/igsnabout) in place of an automatically generated GUID. In the case of IGSNs, the sample is registered with a unique nine-digit alphanumeric code that serves as the GUID. A key aim of the IGSN system is to provide a robust way to link other data, such as published analytical results, back to the original physical sample or specimen. The Non-vertebrate Paleontology Lab at the University of Texas at Austin (UT-NPL) has started using IGSNs and has adopted them into their digitization workflow as a way to link specimens back to other types of data.

In practice, the actual assignment of GUIDs is most likely to occur as part of Module 1: Data Entry, and not as part of Module 0. However, we recommend that the strategy for their creation and assignment should happen as part of Module 0. Guralnick et al. (2015) discuss several ways to deal with the application of GUIDs to new and legacy data and also advocate that whatever type of GUID is used, that it should be resolvable. In practice, however, many paleontology collections are opting for the blanket assignment of GUIDs to specimens by their collections management software (e.g., University of Kansas, University of Colorado, and University of Texas using Specify; Yale Peabody Museum and Smithsonian National Museum of Natural History using EMu). This option, in which the databasing software automatically generates and associates a GUID (in the form of a universally unique identifier (UUID)) with each specimen record in the database, is becoming widespread because it offers an easy and quick solution to creating an identifier that can then be mapped to the Darwin Core occurrenceID field (Darwin Core Task Group, 2015) when data are published. Whether auto-generated UUIDs or resolvable identifiers such as IGSNs or URIs (e.g., Page, 2008) become the standard has yet to be determined.

Data entry. The data entry module (Module 1) deals with the process of entering specimen, locality, and associated data into a database, whether from catalog/ledger sheets, specimen labels, or from images of those paper records (Figure 2) (Farrell, 2014). Specific data entry protocols will largely depend on the database software in use. In the majority of cases data are entered by hand on a record-by-record basis, or by uploading previously prepared spreadsheets (see, for example, Module 3: Proactive Digitization below). Incorporating voice recognition software into the workflow may help speed up data entry, although it has been used primarily for data capture during imaging (Butts, 2013) and to streamline file renaming after image capture. Currently, optical character recognition (OCR) software is not widely used in paleontological collections, and is not included in this module due to the difficulties associated with handwritten labels and the need to handle specimens to remove the label, which is generally located underneath the specimen. OCR has been successfully implemented in other disciplines where labels are easily visible, such as herbarium sheets (Barber, 2012; Barber et al., 2013; Haston et al., 2012; Lafferty and Landrum, 2009), and may prove effective for some paper records in paleontological collections, such as typed locality cards or catalog ledgers. Additionally, once paper records have been imaged, collections could take advantage of crowdsourcing projects, such as Notes from Nature (www.notesfromnature.org/), which allow citizen scientists to transcribe data.

In addition to taxonomic and geographical data, paleontologists have to consider geological time and stratigraphic information. Relational databases that accommodate these data elements (e.g., Arctos (arctosdb.org/), EMu, Specify) deal with the relationships between the specimen record, the locality, and the geological context in different ways, in part a reflection of the lack of consensus within the paleontological community. When choosing a software program it is important to consider not only how closely the database structure matches local institutional records, but also which version will allow for the most efficient databasing and subsequent data querying (see Karim and Farrell, 2014; Morris, 2005).

Once data have been entered, they can be subsequently augmented and updated (Figure 2). In particular, many institutions have separate georeferencing workflows, which may take advantage of georeferencing tools that are integrated into the database (e.g., GEOLocate in Specify). In other cases, it can be more efficient to georeference localities before uploading spreadsheets of data in order to take advantage of collaborative georeferencing projects. Previously digitized source materials (e.g., field notebooks; see Module 2A) may also be attached once specimen records have been created (iDigBio, 2014a).

Maintaining clean data is a constant concern, and quality control steps are recommended at multiple stages in this module. In some collections, student databasing staff prepare spreadsheets, which can be checked by a qualified supervisor prior to upload. In other cases, or additionally, a combination of standard queries designed to catch mistakes (e.g., unexpectedly blank fields, contradictory information) and spot-checking are employed to catch and correct any errors within the database.

Imaging objects. The imaging module (Module 2) is designed for large-scale imaging projects that will allow researchers and the public to gain a better understanding of material held in a particular collection and the condition of that material. This module consequently does not cover detailed research-quality imaging (e.g., SEM). It does cover suggested equipment and methods for effective imaging of several types of objects as well as aspects of file naming, file formats, image processing, and file storage and archiving. Transporting and storing materials for imaging is included in the workflow. The module is subdivided into several sub-modules based on the type of object being imaged; these are listed in Figure 1. Each submodule requires similar steps to prepare and arrange the material for imaging, quality control, and return of specimens to their original location in the repository. Variations in workflow and methodology arise from the type of material imaged.

File naming protocols vary widely among collections, but should at least include the catalog number and ensure that there are no spaces, periods, or special characters in file names. A standardized file and folder naming system was implemented at CUMNH that includes that catalog number and the lens and magnification at which the image was acquired. This file name helps with semi-automated scale bar insertion, bulk image upload into Specify, and downstream data quality control issues. File names at the SNOMNH also include view for 3-D specimens (e.g., dorsal, ventral, lateral). UT-NPL also use a standardized file naming protocol, which includes the catalog number and view. Yale Peabody Museum employs the institutional plus division acronyms as a prefix, the catalog number, and an abbreviated object orientation as the suffix.

Module 2A covers the digitization of cards, ledgers, and field notebooks, which are typically stored separate from specimens and are often imaged independent of related specimens. Recommended equipment includes scanners or digital SLRs, depending upon whether the pages can be laid flat or perhaps cut and fed through a sheet feeder, or must be kept tightly bound. Whichever method is chosen, a file naming procedure, file format, and relevant metadata fields to capture should be determined before starting. A data quality check of both image and associated metadata is an important part of the workflow.

Module 2B addresses the digitization of labels stored with specimens/collection objects; they may be imaged with specimens or separately. The workflow assumes that only labels are being imaged although it recommends that label imaging should be coincident with specimen imaging to minimize specimen handling. Maintaining appropriate connection between label and specimen is a priority; ideally both label and specimen should carry the specimen number so that they can be easily reassociated. It is useful if the file naming protocol matches that used for the specimens. Labels, especially historic labels, may be fragile and may need to be repaired before imaging. They may also require cleaning to reveal more legible handwriting. Technicians should be trained in relevant conservation techniques.

Separate workflows are presented for imaging three-dimensionally (Module 2C) and two-dimensionally (Module 2D) preserved specimens, with the main differences being that the former includes steps for blackening and whitening, more complicated specimen mounting, and accounts for imaging the specimen in multiple views. The use of focus stacking (also referred to as focal plane merging, z-stacking, or focus blending) is included in both workflows and will depend on several variables, including the size and three-dimensional relief of the specimen, equipment, and final desired image quality. While focus stacking is primarily utilized for three-dimensionally preserved specimens, it can also be useful for capturing fine details on two-dimensionally preserved compression fossils (e.g., fossil insects and plants). We realize that these two workflow scenarios do not fully cover all types of fossil specimens, but they do cover many types, and they are intended to serve as a guide for developing workflows for additional preparation types.

Before specimen imaging is undertaken it is important to consider the types of specimens being imaged and what type of images are required for the digitization project, as this will inform many tasks in the final workflow. For instance, should the specimens be imaged from multiple angles? Should relief and other details be enhanced (such as whitening with ammonium chloride sublimate, see Teichert, 1948; Kier et al., 1965; Jago, 1973; Marsh and Marsh, 1975; Feldmann, 1989; Zachos, personal commun., 2008; Hegna, 2010) prior to imaging? Will all specimens in a lot be imaged or just a single exemplar? Will individual specimens on a slab be imaged separately or will one wider shot of the whole slab suffice? Will focus stacking be utilized for more three-dimensionally preserved specimens? How will the lighting be arranged? Strongly directed light can bring out surface relief in specimens, but can also introduce artifacts (Haug and Haug, 2014). There are obvious advantages to having clear written guidelines on such questions before the imaging begins for each project.

The initial steps of the specimen imaging workflows guide the user through some of these basic questions, and continue with how specimens will be transported to the imaging area, camera and software setup, specimen selection and cleaning, and image capture, and rehousing. Tasks such as specimen transport might seem trivial, but can become important if the collection is housed on a different floor, or even building, from the imaging station; a well thought out plan is necessary from the start of the digitization project. Imaging labels can also be done in conjunction with imaging specimens, as is noted in these workflow modules, and might be a more efficient way of capturing two different types of collections data at once. The KUMIP and UT-NPL rely heavily on label data for database record entry and quality control, and label images have become an important part of quality control in their data entry module.

We also recognize the opportunity for using voice recognition software to capture data (e.g., catalog numbers and specimen views) concurrently with specimen imaging as part of this module. Relevant information can be read into a digital spreadsheet for later reuse in file naming or potentially primary data capture. An added bonus to using voice recognition software for data capture while imaging is that the ability to enter data “hands-free” facilitates manipulation of the specimen during imaging. The Yale Peabody Museum (YPM) has pioneered the use of voice recognition software in paleontology collections (Butts, 2013). Data types captured include: image number, catalog number, and suffix abbreviation indicating the orientation of the specimen (ventral, dorsal, hinge, etc.). The data collected on the spreadsheet are processed in a macro that generates (1) composite images of specimens at different orientations and (2) two spreadsheets for automated import to EMu (Butts, 2013). The first spreadsheet is used for importing images to the EMu Multimedia module, and the second spreadsheet associates each multimedia record to its object record.

We reiterate that while many photographic techniques are available to researchers, these modules are designed for capturing a large number of high-quality images in a short amount of time that can be of use to researchers, teachers, and the general public. They are not necessarily intended to be publication quality, although in some cases they are. We expect that researchers will assess the type and condition of material held using these images; higher quality images can then be requested as "virtual loans" or the physical specimen can be requested so that more specialized imaging (e.g., stereo-pair images, reflectance transformation imaging (RTI), infrared imaging) can be undertaken by the researcher.

Module 2E outlines the tasks associated with image processing, such as processing image stacks, applying file names, inserting scale bars, archiving images, and creating derivatives. Possible variations in implementation of these tasks follow. If stacking software is being utilized, images are typically processed (i.e., stacked) immediately after the images are captured so that any specimens that are out of focus can be reimaged before they are put away. In some instances, stacks are batch processed overnight, with the caveat that some may need to be reimaged. Archiving "master" files and creating smaller derivatives to be quickly and easily distributed to users are critical steps in this module, and we recommend that institutions establish policies governing these processes to ensure efficiency and consistent results (Crick, 2005; iDigBio, 2012c). The use of digital scale bars and other text inserted into an image during this module varies widely. Many institutions skip this step by using a physical scale bar inserted at the time of image capture, while other institutions insert digital scale bars into images as part of derivative production. The UT-NPL uses a semiautomatic process to insert a digital scale bar based on the scale embedded in the image file. The CUMNH utilizes a script to retrieve the catalog number, lens, and magnification information, all contained in the file name, to insert a caption into the lower left corner of the image that include the catalog number and digital scale bar in black text over a small white rectangle. This task is partially automated by using actions and batch processing features in Adobe Photoshop. Reviewing existing community standards (see e.g., image guidelines from AntWeb [California Academy of Sciences, 2016]) can be helpful when developing this module at the institutional level, and we encourage other institutions to share this information so that broader community standards can be developed.

Proactive digitization. Given the trend toward greater field-based digitization through which at least some steps in the digitization process occur in the field (e.g., recording geographic coordinates of collection locality) and the sometimes extensive recording of digital images and data of in situ specimens, the workflows presented here include a module (Module 3) for what we term proactive digitization. We define this as digitization accomplished by collectors or field assistants at the time of specimen collection, and prior to deposition in a museum or academic collection, effectively moving digitization activities forward in the collecting and accessioning processes. In some cases, this might include recording text data into an electronic spreadsheet or other input device that can be easily imported into a collection database. This is a relatively new and rapidly developing method, but one that fits especially well within paleontology digitization workflows. Such strategies are currently in use in some domains as follows.

Ichthyology collection managers at the Museum of Comparative Zoology at Harvard, for example, attempt to collaborate with individual collectors prior to field excursions and generate a basic, pre-formatted data entry worksheet that matches the museum's database bulk loader and is also compatible with and complementary to the collector's workflow (Williston, personal commun., 2013). A similar project is underway at UT-NPL where spreadsheet formats currently used for pre-database data entry by UT-NPL students is being adapted for dispersal to all graduate students and faculty who are generating data for future addition to the repository. To encourage use, the template will be available on the web page (www.jsg.utexas.edu/npl/research/), the wiki (wikis.utexas.edu/display/specify6/Proactive+digitization-field+data), and physically distributed on a flash drive. Users will be able to add additional fields of relevance, which will further improve the data fields that are incorporated into the Specify database. It is a pre-Specify Workbench tool allowing for refinement of data prior to upload. The Workbench tool in Specify is a useful portable tool. For example, data for new locality records can be captured in the field and imported using the Workbench, making subsequent specimen databasing progress more rapid, especially in cases where the specimens themselves cannot be cataloged in the field. The Smithsonian NMNH is also utilizing similar methods with EMu FIMS (Field Information Management System) in conjunction with a workflow that efficiently transfers data from the field into the EMu database (Hollis, personal commun., 2015). The University of Kansas Entomology Collections (Short, 2014), VSC herbarium at Valdosta State University (Carter, personal commun., 2013), Southwestern Adventist University (Woods and Chadwick, 2014), and the Gray Fossil Site (Woodward and Compton, 2014) are also good examples of collections that are proactively capturing data in the field.

Proactive digitization can also utilize electronic interfaces designed for Trimble™ or other geolocation devices; custom interfaces for electronic tablets, smartphone applications, and similar devices; or direct entry into local copies of or wirelessly connected collection databases. In some instances this may entail assigning and recording catalog numbers and other identifiers in the field and creating or appropriating existing locality descriptions at the time of collection, depending on the type of collection object being acquired. We anticipate that technological improvements will enhance field collection protocols, significantly improve digitization efficiency by reducing the quantity of data to be digitized post-collection, and that such protocols are likely to become common practice in the near future. A recent study in Paleoanthropology demonstrates such developments in digital data collection in the field (Reed, et al., 2015).

Data Publishing

Although we do not specifically address this topic as a stand-alone module in the digitization workflows described herein, the publication of data is the logical follow up to all digitization workflows. The sharing of digitized collections data has long been a bottleneck in digitization workflows (MacLeod and Guralnick, 2000). Most museums today have a digital specimen catalog; however, in many cases these data are only searchable internally by museum staff. Museums rely heavily on IT staff (if they exist) to develop online data publishing tools and make data available for harvesting. Sharing datasets online with the research community has recently been made much easier due to the development of publishing tools (e.g., the Global

Biodiversity Informatics Facility's Integrated Publishing Toolkit [IPT]), the growth of online data aggregators (e.g., iDigBio, VertNet, GBIF), and the development of data sharing standards (e.g., DarwinCore). IPT is a public web server that is relatively easy for a systems administrator to install and can be hosted locally at an institution. Datasets exported from a local digital specimen catalog, typically in a .csv format, can be easily uploaded using the IPT web interface and detected and ingested by a variety of data aggregators, including those listed above. Aggregators make these data available for searching, which relieves the local institution of the need for a data portal and the burden of publishing.

Some institutions express concern about publishing data about sensitive fossil localities, including precise georeferences and detailed locality descriptions (Weiss, 2013; Barton, et al., 2006), based on concerns for site preservation, parameters stipulated by confidential agreements with landowners or donors, and legal restrictions. As digitization has become more prevalent, this has become an important issue across the natural history community (Chapman and Grafton, 2008) that continues to lack clear consensus. With some exceptions, Norris and Butts (2014) advocate in favor of making paleontological locality descriptions publicly available. We do not address efficacy of data redaction or steps to achieve it in Module 1 of the workflows, but recognize that institutional policy and federal law often governs the quantity and types of data an institution might choose to expose. For institutions who choose to redact, physical redaction might occur during data entry (e.g., embargoing certain pieces of locality data from public data exports), imaging (e.g., covering locality information on a label), or electronically at the time data are exposed on the web (e.g., only exporting data to county level or imposing a rule on the export query such that only precise data for non-sensitive localities are exported).

CONCLUSION

The advent of the U.S. National Science Foundation's Advancing Digitization of Biodiversity Collections (ADBC) initiative in 2011 has led to a rapid and significant increase in digitization activity in natural history museums and university-based biodiversity collections. This is especially true in the paleobiological sciences. Three Thematic Collections Networks (TCNs) encompassing 28 paleontological collections are currently funded through ADBC, all focusing largely on the digitization of marine and terrestrial invertebrate fossils. Documenting and sharing a combined set of the digitization workflows, protocols, and policies produced by members of the DROID working group, many of whom are also part of these earliest TCNs, ensures the availability of tested, vetted, and recommended practices as new digitization programs come online. The workflows documented herein will encourage coalescence of community-wide standardized processes and serve to alleviate data gaps that result from institutional variation in digitization practices (Nelson, 2015). We hope that the community will benefit from the well-designed and easily implementable set of flexible, modular workflows that captures what has been gleaned from the several successful paleo digitization programs.

ACKNOWLEDGMENTS

We thank iDigBio for sponsoring the DROID workshop, GSA digitization session, and use of Adobe Connect. We also thank our home institutions for their continued support of our digitization efforts. iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Barber, A.C. 2012. Advancing Access to Biodiversity Data Using the SALIX Method and Digital Field Guides. Unpublished Master's Thesis, Arizona State University, Tempe, Arizona, USA.

Barber, A.C., Lafferty, D., and Landrum, L.R. 2013. The SALIX Method: a semi-automated workflow for herbarium specimen digitization. *Taxon*, 62:581-590.

Barton, M.A., Frakes, B., and Meyer, H.W. 2006. Using a relational geodatabase to manage paleontological resources at Florissant Fossil Beds National Monument. *New Mexico Museum of Natural History and Science Bulletin,* 34:7.

Beaman, R. and Cellinese, N. 2012. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys*, 209:7-17.

Butts, S. 2013. YPM-IP workflow with voice recognition imaging. Presentation from the iDigBio paleontology digitization workshop, Yale Peabody Museum, New Haven, CT, September 2013. Document accessed 22 January 2015.
www.idigbio.org/content/ypm-ip-workflow-voice-recognition-imaging

California Academy of Sciences. 2016. AntWeb Participation. AntWeb version 6.9.2. Document accessed 22 July 2016.
www.antweb.org/documentation.do

Chapman, A.D. and Grafton, O. 2008. *Guide to Best Practices for Generalising Primary Species-Occurrence Data.* Global Biodiversity Information Facility, Copenhagen.

Crick, M. 2005. Image File Management, p. 41-56. In Häuser, C.L., Steiner, A., Holstein, J., and Scoble, M.J. (eds.), *Digital imaging of biological type specimens: a manual of best practice. Results from a study of the European Network for Biodiversity Information.* European Network for Biodiversity Information, Stuttgart.

Darwin Core Task Group. 2015. Darwin Core terms: a quick reference guide. Document accessed 19 February 2015.
rs.tdwg.org/dwc/terms/

Farrell, Ú.F. 2014. Label imaging workflow from the KU Invertebrate Paleontology collection – a quality control measure, p. 52. In (editor unknown), *Society for the Preservation of Natural History Collections 29th Annual Meeting, Cardiff, Programme and Abstracts*.

Feldmann, R.M. 1989. Whitening fossils for photographic purposes, p.342-346. In Feldmann, R.M., et al.**[AUTHOR: editors' names should be listed]** (eds.), *Paleotechniques*. *The Paleontological Society, Special Publication 4.*

Glenister, B.F. 1977. Innovative Programs & Electronic Data Processing in Collections Management, p.25-28. In Glenister, B.F. (ed.), *Fossil Invertebrates - Collections in North American Repositories 1976: A Report of the Paleontological Society Ad Hoc Committee on North American Resources in Invertebrate Paleontology (CONARIP)*. Paleontological Society, Columbus, OH.

Granzow-de la Cerda, I. and Beach, J.H. 2010. Semiautomated workflows for acquiring specimen data from label images in herbarium collections. *Taxon*, 59:1830-1842.

Guralnick, R.P., Cellinese, N., Deck, J., Pyle, R.L., Kunze, J., Penev, L., Wals, R., Hagedorn, G., Agosti, D., Wieczorek, J., Catapano, T., and Page, R.D. 2015. Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. *ZooKeys*, 494:133.
doi:10.3897/zookeys.494.9352

Guralnick, R., Conlin, T., Deck, J., Stucky, B.J., and Cellinese, N. 2014. The Trouble with Triplets in Biodiversity Informatics: A Data-Driven Case against Current Identifier Practices. *PLoS ONE,* 9(12): e114069. doi:10.1371/journal.pone.0114069

Haston, E., Cubey, R., Pullan, M., Atkins, H., and Harris, D.J. 2012. Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. *ZooKeys*, 209:92-102.

Haug, J.T. and Haug, C. 2014. *Eoprosopon klugi* (Brachyura) - the oldest unequivocal and most "primitive" crab reconsidered. *Palaeodiversity,* 7:149-158.

Häuser, C.L., Steiner, A., Holstein, J., and Scoble, M.J. (eds.). 2005. *Digital imaging of biological type specimens: a manual of best practice. Results from a study of the European Network for Biodiversity Information.* European Network for Biodiversity Information, Stuttgart.

Hegna, T.A. 2010. Photography of soft-bodied crustaceans via drying, whitening, and splicing. *Journal of Crustacean Biology,* 30:351-356.

iDigBio. 2012a. Digitization workflow workshop report. Report on the iDigBio Developing Robust Object-to-Image-to-Data (DROID) workflow workshop, Florida Museum of Natural History, Gainesville, FL, May 2012. Document accessed 05 February 2015.
www.idigbio.org/content/digitization-workflow-workshop-report

iDigBio. 2012b. Workflow modules and task lists. Protocol developed from the iDigBio Developing Robust Object-to-Image-to-Data (DROID) workflow workshop, Florida Museum of Natural History, Gainesville, FL, May 2012. Document accessed 05 February 2015.
www.idigbio.org/content/workflow-modules-and-task-lists

iDigBio 2012c. Policy on Acceptable Formats for iDigBio-hosted Images and Recommendations for the Acquisition, Processing, Storage, and Distribution of Digital Images. Document accessed 07 February 2015.
www.idigbio.org/sites/default/files/internal-docs/idigbio-standards/Image_File_Format_Recommendations_and_Standards_20120921.pdf

iDigBio. 2013. Paleo Digitization Workshop. Wiki developed for the iDigBio paleontology digitization workshop, Yale Peabody Museum, New Haven, CT, September 2013. Document accessed 05 February 2015.
www.idigbio.org/wiki/index.php/Paleo_Digitization_Workshop

iDigBio. 2014a. Digitizing from Source Materials Workshop. Wiki developed for the iDigBio digitizing from source materials workshop, Yale Peabody Museum, New Haven, CT, March 2014. Document accessed 12 February 2015.
www.idigbio.org/wiki/index.php/Digitizing-From-Source-Materials

iDigBio. 2014b. Paleo Imaging Workshop. Wiki developed for the iDigBio specimen imaging for paleontology workshop, University of Texas, Austin, TX, April-May, 2014. Document accessed 04 February 2015.
www.idigbio.org/wiki/index.php/Paleo_Imaging_Workshop

iDigBio, 2016. Workflow Modules and Task Lists. iDigBio website. Document accessed 22 July 2016.
www.idigbio.org/content/workflow-modules-and-task-lists

Jago, J.B. 1973. A hazard in the magnesium oxide method of whitening fossils. *Journal of Paleontology,* 47:591-592.

Karim, T.S. and Farrell, U.C. 2014. Specify Paleo Collections Workshop - training, new technology, and data models. Report on the iDigBio Specify for paleo collections workshop, University of Kansas, Lawrence, KS, May 2014. Document accessed 03 February 2015.
www.idigbio.org/content/specify-paleo-collections-workshop-training-new-technology-and-data-models

Kier, P.M., Grant, R.E., and Yochelson, E.L. 1965. Whitening fossils, p. 453-456. In Kummel, B. and Raup, D.M. (eds.), *Handbook of Paleontological Techniques*. W.H. Freeman and Company, San Francisco.

Lafferty, D. and Landrum, L. 2009. SALIX, a semiautomatic label information extraction system using OCR. Presentation given at the joint Annual Meeting of the Mycological Society of America, American Bryological and Lichenological Society, American Fern Society, American Society of Plant Taxonomists, Botanical Society of America, 25-29 July 2009, Snowbird, Utah.

MacLeod, N. and Guralnick, R. 2000. Paleoinformatics, p. 31-36. In Lane, R.H., Steininger, F.F., Kaesler, R.L., Ziegler, W., and Lipps, J. (eds.), *Fossils and the Future: Paleontology in the 21st Century*. W. Kramer, Frankfurt am Main.

Marsh, R.C. and Marsh, L.F. 1975. New techniques for coating paleontological specimens prior to photography. *Journal of Paleontology*, 49:565-566.

Mello, J.F. 1969. Paleontologic data storage and retrieval. *Proceedings of the North American Paleontologic Convention*, part *B*: 57-71.

Mello, J.F. and Collier, F.J. 1972. New procedures in recording specimen-related data on fossils. *Journal of Paleontology,* 46:776-777.

Morris, P.J. 2000. A data model for invertebrate paleontological collections information, p. 105-108. In Allmon, W.D. and White, R.D. (eds.), *Guidelines for the Management and Curation of Invertebrate Fossil Collections*. *The Paleontological Society, Special Publication 10.* The Paleontological Society, Boulder.

Morris, P.J. 2005. Relational database design and implementation for biodiversity informatics. *PhyloInformatics*, 7:1-66.

Nelson, G. 2012. Workflow elements and concepts: common practices. Presentation at the iDigBio digitizing biological collections workshop. Document accessed 05 February 2015.
www.idigbio.org/content/workflow-elements-and-concepts-common-practices

Nelson, G. 2015. Gaps in Biodiversity Data: Challenges for Digitization. Presentation at the Global Change Workshop, Missouri Botanical Garden. Document accessed 05 February 2016.
www.idigbio.org/wiki/images/3/3-D/Nelson-DigitizationGaps.pdf

Nelson, G., Paul, D., Riccardi, G., and Mast, A.R. 2012. Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys*, 209:19-45.

Nelson, G., Sweeney, P., Wallace, L.E., Rabeler, R.K., Allard, D., Brown, H., Carter, J.R., Denslow, M.W., Ellwood, E.R., and Germain-Aubrey, C.C. 2015. Digitization workflows for flat sheets and packets of plants, algae, and fungi. *Applications in Plant Sciences*, 3(9):1-9.

New York Botanical Garden. 2004. New York Botanical Garden Virtual Herbarium Best Practices Guide. Document accessed 13 January 2015.
sciweb.nybg.org/science2/VirtualHerbarium.asp

Norris, C. and Butts, S. 2014. Let your data run free? The challenge of data redaction in paleontological collections. *Collection Forum,* 28(1-2):113-118. doi:10.14351/0831-4985-28.1.113

Page, R.D.M. 2008. Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics*, 9(5):345-354. doi:10.1093/bib/bbn022

Page, R.D.M. 2009. bioGUID: resolving, discovering, and minting identifiers for biodiversity informatics. *BMC Bioinformatics*, 10(suppl. 14):S5.

Reed, D., Barr, W.A., Mcpherson, R.B., Geraads, D. Wynn, J.G., and Alemseged, Z. 2015. Digital Data Collection in Paleoanthropology. *Evolutionary Anthropology,* 24:238-249.
doi: 10.1002/evan.21466
onlinelibrary.wiley.com/doi/10.1002/evan.21466/epdf

Richards, K., White, R., Nicolson, N., and Pyle, R. 2011. A Beginner's Guide to Persistent Identifiers, version 1.0. Released on 9 February 2011. Copenhagen: Global Biodiversity Information Facility, 33 pp, accessible online at
links.gbif.org/persistent_identifiers_guide_en_v1.pdf

Scoble, M. 2010. Natural history collections digitization: Rationale and value. *Biodiversity Informatics*, 7:77-80, 35KB. journals.ku.edu/index.php/jbi/article/viewArticle/3994

Short, A. 2014. Data before specimens: workflows for digital management and sharing of field notes and images, p.150. In (editor unknown), *Proceedings, 62nd Annual Meeting of the Entomological Society of America*. Entomological Society of America.

Smith, V.S. and Blagoderov, V. 2012. Bringing collections out of the dark. *Zookeys,* 209:1-6.
doi:10.3897/zookeys.209.3699

Teichert, C. 1948. A simple device for coating fossils with ammonium chloride. *Journal of Paleontology*, 22:102-104.

Vollmar, A., Macklin, J.A., and Ford, L. 2010. Natural history specimen digitization: challenges and concerns. *Biodiversity Informatics*, 7:93-112, 188KB.
journals.ku.edu/index.php/jbi/article/view/3992/3806

Weiss, A. 2013. Wrestling with mosasaurs: results of the Sternberg Museum of Natural History - Forsyth Library fossil digitization pilot project, p. 103-124. In

Cool, C. and Ng, K.B. (eds.), *Recent Developments in the Design, Construction, and Evaluation of Digital Libraries: Case Studies*. Information Science Reference, Hershey, PA. doi:10.4018/978-1-466-2991-2.ch007

Wolniewicz, P. 2009. Easily-accessible digital palaeontological databases - a new perspective for the storage of palaeontological information. *Geologos*, 15(3-4):181-188. doi:10.2478/v10118-009-0002-1

Woods, J.A. and Chadwick, A.V. 2014. Development and implementation of an efficient, accessible online museum site and database with open access, p. 701. In (editor unknown), *2014 GSA Annual Meeting in Vancouver, British Columbia, Abstracts with Programs 46(6)*. Geological Society of America, Boulder.

Woodward, B. and Compton, B. 2014. Using GIS to map and catalog paleontological and geological specimens of the Mid-Pliocene Gray Fossil Site, Washington County, Tennessee, USA, p. 823. In (editor unknown), *2014 GSA Annual Meeting in Vancouver, British Columbia, Abstracts with Programs*, *46(6)*. Geological Society of America, Boulder.