Applications in Plant Sciences

Applications in Plant Sciences 2015 3(9): 1500065

Review Article

Digitization workflows for flat sheets and packets of plants, algae, and fungi[1]

Gil Nelson[2,27], Patrick Sweeney[3], Lisa E. Wallace[4], Richard K. Rabeler[5], Dorothy Allard[6], Herrick Brown[7], J. Richard Carter[8], Michael W. Denslow[9], Elizabeth R. Ellwood[10], Charlotte C. Germain-Aubrey[11], Ed Gilbert[12], Emily Gillespie[13], Leslie R. Goertzen[14], Ben Legler[15], D. Blaine Marchant[11,16], Travis D. Marsico[17], Ashley B. Morris[18], Zack Murrell[9], Mare Nazaire[19], Chris Neefus[20], Shanna Oberreiter[21], Deborah Paul[2], Brad R. Ruhfel[22], Thomas Sasek[23], Joey Shaw[24], Pamela S. Soltis[11], Kimberly Watson[25], Andrea Weeks[26], and Austin R. Mast[10]

[2]Integrated Digitized Biocollections (iDigBio), Florida State University, Tallahassee, Florida 32306-2100 USA; [3]Peabody Museum of Natural History, Yale University, New Haven, Connecticut, USA; [4]Department of Biological Sciences, Mississippi State University, Mississippi State, Mississippi, USA; [5]University of Michigan Herbarium–EEB, Ann Arbor, Michigan, USA; [6]Department of Plant Biology, University of Vermont, Burlington, Vermont, USA; [7]Department of Biological Sciences, University of South Carolina, Columbia, South Carolina, USA; [8]Biology Department, Valdosta State University, Valdosta, Georgia, USA; [9]Department of Biology, Appalachian State University, Boone, North Carolina, USA; [10]Department of Biological Science, Florida State University, Tallahassee, Florida, USA; [11]Florida Museum of Natural History, University of Florida, Gainesville, Florida, USA; [12]School of Life Sciences, Arizona State University, Tempe, Arizona, USA; [13]Department of Biological Sciences, Marshall University, Huntington, West Virginia, USA; [14]Department of Biological Sciences, Auburn University, Auburn, Alabama, USA; [15]Burke Museum, University of Washington, Seattle, Washington, USA; [16]Department of Biology, University of Florida, Gainesville, Florida, USA; [17]Department of Biological Sciences, Arkansas State University, Jonesboro, Arkansas, USA; [18]Department of Biology, Middle Tennessee State University, Murfreesboro, Tennessee, USA; [19]Rancho Santa Ana Botanic Garden, Claremont, California, USA; [20]Department of Biological Sciences, University of New Hampshire, Durham, New Hampshire, USA; [21]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; [22]Department of Biological Sciences, Eastern Kentucky University, Richmond, Kentucky, USA; [23]Department of Biology, University of Louisiana at Monroe, Monroe, Louisiana, USA; [24]Biological and Environmental Sciences, University of Tennessee, Chattanooga, Tennessee, USA; [25]William and Lynda Steere Herbarium, The New York Botanical Garden, Bronx, New York, USA; and [26]Department of Biology and the Ted R. Bradley Herbarium, George Mason University, Fairfax, Virginia, USA

Effective workflows are essential components in the digitization of biodiversity specimen collections. To date, no comprehensive, community-vetted workflows have been published for digitizing flat sheets and packets of plants, algae, and fungi, even though latest estimates suggest that only 33% of herbarium specimens have been digitally transcribed, 54% of herbaria use a specimen database, and 24% are imaging specimens. In 2012, iDigBio, the U.S. National Science Foundation's (NSF) coordinating center and national resource for the digitization of public, nonfederal U.S. collections, launched several working groups to address this deficiency. Here, we report the development of 14 workflow modules with 7–36 tasks each. These workflows represent the combined work of approximately 35 curators, directors, and collections managers representing more than 30 herbaria, including 15 NSF-supported plant-related Thematic Collections Networks and collaboratives. The workflows are provided for download as Portable Document Format (PDF) and Microsoft Word files. Customization of these workflows for specific institutional implementation is encouraged.

Key words: citizen science; digital imaging; digitization; herbarium; specimen database; workflow.

[1]Manuscript received 8 June 2015; revision accepted 30 July 2015.

The authors thank the herbarium digitization efforts with which the authors are associated (Box 2) and previous work of iDigBio's Flat Sheets and Packets Digitization Working Group. This material is based on work supported by the National Science Foundation under Cooperative Agreement no. EF-1115210. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[27]Author for correspondence: gnelson@bio.fsu.edu

doi:10.3732/apps.1500065

The world's 3400 herbaria curate 350 million specimens of plants, algae, and fungi (Thiers, 2015) and represent a critical big-data resource for pressing questions related to the environment, human health, biosecurity, commerce, and the biological sciences (Beach et al., 2010). However, only a modest fraction of their specimen data are digitally available to the scientific community, educators, and policy-makers, and many herbaria have not yet begun digitization. For example, in a recent survey of U.S. herbaria, Barkworth and Murrell (2012) found that about one third of U.S. herbarium specimens had been databased

(herein defined as having transcribed textual label data into a retrievable electronic format), half (46%) of the herbaria had yet to begin databasing their specimens, and just 24% of the herbaria had begun imaging specimens. With the support of strong national funding initiatives (e.g., the U.S. National Science Foundation's [NSF] Advancing Digitization of Biodiversity Collections program [ADBC] and Collections in Support of Biological Research) and national resources (e.g., iDigBio and the U.S. Geological Survey's Biodiversity Information Serving Our Nation [BISON]), the rate of digitization in the United States is expected to increase. This trend is likely to be true for other countries. One notable example is Australia, where 71% of the 7 million specimens have been databased to date (Thiele, 2014). In a more general survey of biodiversity research collections, Vollmar et al. (2010) found that funding, time, and staff limitations represented the top three challenges to digitization, but that issues that could be addressed by guidelines and suggestions ranked next. Here, we seek to reduce the impediment to herbarium digitization—specifically digital imaging, databasing, and georeferencing—by introducing a set of how-to workflow modules that reflect our collective practical experience with digitization.

Workflows have been produced by most larger herbaria (e.g., as documented for New York Botanical Garden in Tulig et al. [2012]), but few of these are available online, and smaller and resource-limited collections are lagging in the creation of workflows. Furthermore, some promising developments (e.g., public engagement in digitization and high-throughput imaging enabled by conveyor systems) are sufficiently new that few protocols exist. The recent establishment of regional, national, and international networks of digitizing collections (e.g., ADBC's 15 current thematic collections networks [TCNs]) has necessitated broad community dialogue regarding workflows and aggregation of relevant documents (e.g., Consortium of Northeastern Herbaria, Global Biodiversity Information Facility [GBIF], and iDigBio). The workflow modules described here are a next step in the aggregation of workflows, representing a synthesis of workflows across relevant TCNs with good representation of small collections. The data resources housed in those small or otherwise resource-challenged collections are particularly valuable because they often contain records from areas or taxa that are underrepresented in larger collections (Snow, 2005; Casas-Marce et al., 2012; A. Monfils, Central Michigan University, personal communication, 2015).

Our goal is to provide herbarium staff and administrators with a foundation for starting new, and enhancing existing, digitization projects of flat sheets and packets, but many of the modules are transferable to other specimen types (e.g., pinned specimens, specimens in jars). The modularity of the collective workflows makes implementation more flexible (i.e., an institution can use a selection that meets its needs) and scalable (as resource availability ebbs and flows; Haston et al., 2012). The workflow modules could also form a starting point for community-agreed-upon best practices, which are lacking for some, but not all, of the modules. GBIF has published more than 10 best-practice documents in the past decade (e.g., Hauser et al., 2005 [for digital imaging]; and Chapman and Wieczorek, 2006 [for georeferencing]). The workflow modules presented here represent a balance between describing tasks in general terms—to ensure broad applicability—and providing specific successful solutions developed at individual institutions.

MATERIALS AND METHODS

We adopted a collaborative strategy to address this goal. The concept of developing collections digitization workflows was initiated during the Developing Robust Object-to-Image-to-Data Workflows Workshop (DROID) held at the University of Florida (Gainesville, Florida) in May 2012, which brought together approximately 30 participants representing a range of disciplines, including botany, paleontology, entomology, and vertebrate zoology. Leading up to, and overlapping with, DROID, staff from iDigBio visited 28 digitization programs in 10 institutions for the purpose of reviewing and assessing successful workflows across the collections community (Nelson et al., 2012). Findings from these visits provided a development framework for DROID and the several working groups that followed. It became clear that broad disparities in digitization starting points, institutional infrastructure, curatorial practices, and precise digitization tasks among and within these groups focused on different taxa make the development of a single, consensus object-to-digitized-content workflow impractical. The diagrams presented in Figs. 1 and 2 demonstrate alternative implementations of a digitizing workflow: a workflow in which data entry precedes image capture (Fig. 1) and the DROID planners' original concept of an object-to-image-to-data workflow (Fig. 2). These represent just two of several successful workflow organizations and underscore our rationale here for modularity. Several smaller working groups emerged during the workshop representing various preparation types. Each of these groups met regularly following DROID to flesh out preparation-specific modules. The DROID Flat Sheets and Packets Working Group (https://www.idigbio.org/wiki/index.php/Developing_Robust_Object_to_Image_to_Data_%28DROID1%29) was charged with developing orderly task lists for digitizing specimens of that type, including

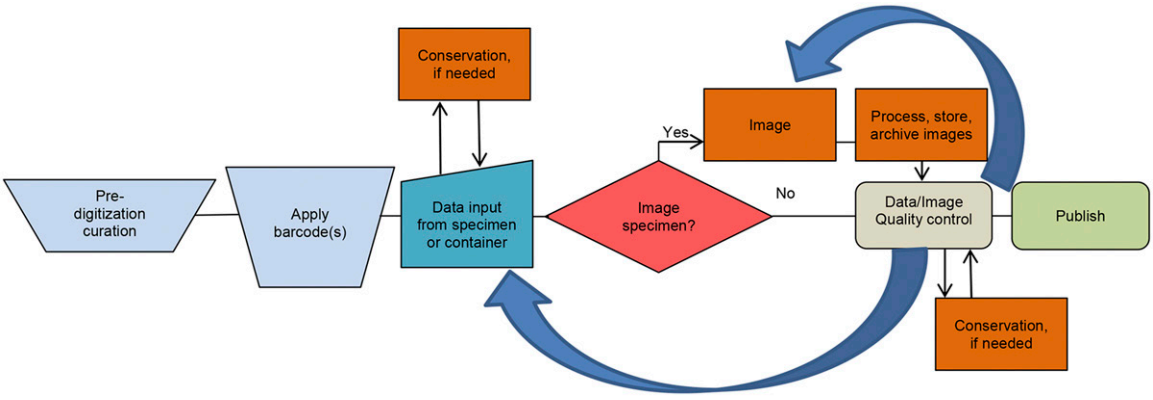

Fig. 1. Example object-to-data-to-image workflow. This workflow captures data directly from labels on physical specimens. Images of specimens may or may not be captured. Barcodes are usually applied inline or as an iterative step through which dozens or hundreds of barcodes are affixed, immediately preceding data entry. Pre-digitization curation, including nomenclatural annotations and specimen organization, is usually important in this workflow. The need for specimen conservation may be discovered and remedied as physical specimens are passed to data entry technicians or following the specimen handling associated with imaging procedures.

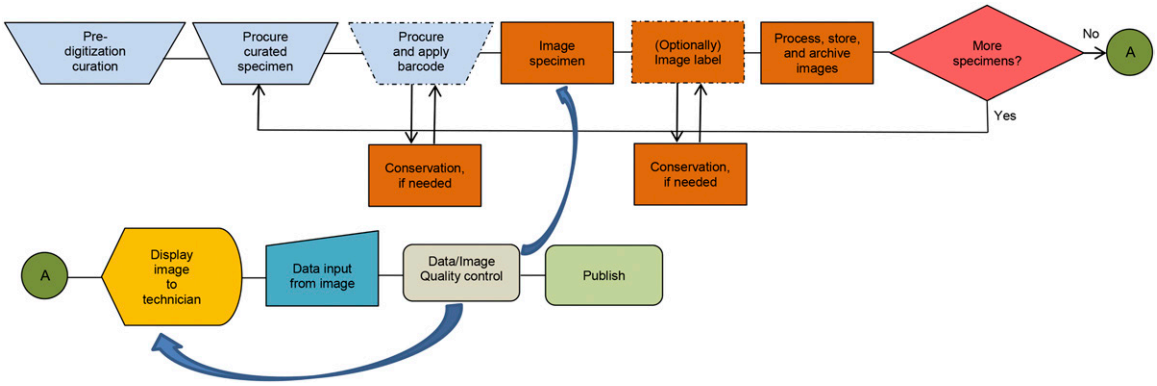

Fig. 2. Example object-to-image-to-data workflow. This workflow captures specimen images and uses these images as the basis for data capture. Barcodes are sometimes applied inline as the step immediately previous to imaging (shown optionally) and other times through an iterative process during which several dozen or several hundred barcodes are applied. Nomenclatural annotation during pre-digitization ensures synchronization of name-on-folder with name-on-specimen. The need for specimen conservation may be discovered and remedied before or after imaging.

those of plants, algae, and fungi. Initially, 11 working group members developed six modules (see Box 1) for tasks associated with these collections. Using a modular approach allowed the working group to accommodate the broad range of extant workflow implementations within the collections community and to assemble orderly, comprehensive task lists as foundations from which institutionally specific workflows could be created. Modules for flat sheets and packets from the DROID effort were published online (iDigBio, 2012).

Due to the rapid increase in herbarium digitization projects in the past few years—raw data from the most recently available survey (U.S. Virtual Herbarium, 2015) suggests that about 80% of responding U.S. herbaria have now initiated digitization—and workflow innovations since 2012, iDigBio sponsored a January 2015 Herbarium Workflows Workshop at Valdosta State University (Valdosta, Georgia, USA) to update the modules in collaboration with 15 digitizing projects (Box 2). Workshop attendees, which included representatives of all current TCNs and other consortia currently digitizing flat sheets and/or packets, examined and updated the initial Flat Sheets and Packets modules and created eight new modules (Box 1), maintaining the modularity and format of the original DROID products.

RESULTS

The latest set of modules, along with a glossary of relevant terms, is provided for download as a Portable Document Format (PDF) and editable word processing files on GitHub (https://github.com/iDigBioWorkflows/FlatSheetsDigitizationWorkflows), as PDF files at iDigBio (https://www.idigbio.org/content/workflow-modules-and-task-lists); the modules are also available as supplementary data with this article. We produced 14 modules presented in 4-column tables with columns for Task ID, Task Description, Explanations and Comments, and Resources. Task Description consists of a brief description of the task to be accomplished. Explanations and Comments offer additional notes and variations to assist in implementation, and Resources refers to, or lists, resources commonly used by members of the Flat Sheets and Packets Working Group or the participants in the January 2015 workshop.

DISCUSSION

The modules are best viewed as templates for customization presented in a reasonable, but not absolute, sequence and potentially including more tasks than a particular institution might choose to implement. Some of the workflow documents consist largely of unordered lists not dependent on sequence. We anticipate that order of execution and selection of tasks to implement will vary among herbaria based on facility configuration, personnel,

Box 1. Flat sheets and packets workflow modules.

Module 1: Pre-digitization Curation[1]
Module 2: Selecting Components for an Imaging Station
Module 3: Imaging Station Setup, Camera/Copy Stand[1]
Module 4: Imaging Station Setup, Light Box
Module 5: Imaging Station Setup, Scanner[1]
Module 6: Imaging[1]
Module 7: Image Processing[1]
Module 8: Organizing and Implementing a Public Participation Imaging Blitz
Module 9: Image Archiving
Module 10: Selecting a Database
Module 11: Data Capture[1]
Module 12: Organizing and Implementing a Public Participation Transcription Blitz
Module 13: Georeferencing
Module 14: Proactive Digitization

[1]Modules initially completed by the DROID Flat Sheets and Packets Working Group.

Box 2. NSF-funded digitization initiatives participating in the January 2015 Herbarium Workflows Workshop at Valdosta State University (Valdosta, Georgia, USA).

Plants, Herbivores, and Parasitoids: A Model System for the Study of Tri-Trophic Associations[1]
North American Lichens and Bryophytes: Sensitive Indicators of Environmental Quality and Change[1]
Mobilizing New England Vascular Plant Specimen Data to Track Environmental Change[1]
The Macrofungi Collection Consortium: Unlocking a Biodiversity Resource for Understanding Biotic Interactions, Nutrient Cycling and Human Affairs[1]
The Macroalgal Herbarium Consortium: Accessing 150 Years of Specimen Data to Understand Changes in the Marine/Aquatic Environment[1]
Documenting the Occurrence through Space and Time of Aquatic Non-indigenous Fish, Mollusks, Algae, and Plants Threatening North America's Great Lakes[1]
The Key to the Cabinets: Building and Sustaining a Research Database for a Global Biodiversity Hotspot (SERNEC)[1]
SEINet: North American Virtual Flora Network
Consortium of California Herbaria
Consortium of Pacific Northwest Herbaria
Magnolia grandiFLORA: The Digital Herbarium for Mississippi
CyberFlora Louisiana
The GA-VSC Herbaria Collaborative: Phase I of a Statewide Consortium
Imaging the Tall Timbers Research Station's Biological Research Collections
The Deep South Plant Specimen Imaging Project

[1]A Thematic Collection Network funded by NSF's Advancing Digitization of Biodiversity Collections Program.

equipment, institutional and research project goals, personal preferences of curators, etc., and that not all herbaria will implement every task or module. Although the presentation of task lists for most workflows presented here follows a linear format, some tasks or groups of tasks represent iterative processes that, in practice, might be repeated several to many times before progressing to a succeeding task. Instances in which a workflow document consists largely of an unordered task list are noted below.

Institutions implementing customized versions of these workflows should critically consider which tasks to include. We consider tasks such as quality assurance and control, specimen conservation, assignment of globally unique identifiers (GUIDs), and the recording of at least a basic set of data per specimen (i.e., skeletal records) to be essential. Many other tasks are elective, depending on institutional parameters, policies, and research needs. We also recognize that the workflows a particular institution implements might vary considerably from the ones presented here. However, we encourage institutions to universally assess the fitness of their digital outputs for use in research, especially such common uses as species distribution modeling for climate change (e.g., Loarie et al., 2008; Johnson et al., 2011), evolutionary studies (e.g., Soltis et al., 2014), and taxonomy.

In several places in the workflow documents, we refer in the Resources column to digitization policy manuals or project management plans. Implied is the suggestion that the success of a digitization program and its attendant workflows is dependent upon the development of written policies or plans that identify the overall goals of digitization for an institution or herbarium as well as specific projects (including research) within that institution or program. These might include expected outcomes and standards for (1) data entry; (2) image acquisition, processing, management, and archiving; (3) the ruling authority for the assignment of taxonomic names; (4) the processing of annotations and determination histories; and (5) file naming conventions. When projects involve many participants (e.g., students, volunteers, general public), it is essential to have available detailed written instructions for digitization practices and protocols. These may be developed using the outlines of the modules presented here. A discussion of each of the workflow modules follows.

Pre-digitization curation (module 1)—Pre-digitization curation involves tasks that occur prior to databasing or imaging (Nelson et al., 2012) and are presented in the workflow document as an unordered list (Appendix S1). Digitization provides motivation for attending to important curatorial requirements, such as organizing specimens, applying annotation labels to synchronize the name on the folder with the name on the specimen, and examining specimens for required conservation attention. Synchronizing the name on the folder with the name on the specimen is especially important if the institution will database from images through distributed downstream data entry processes, such as through public participation activities, or where physical specimens become separated from their enclosing folder.

The pre-digitization curation module is iterative and usually involves the processing and staging of numerous specimens to be moved at one time to cabinets or staging areas in proximity to data entry or imaging stations. Pre-digitization curation may also include the application of barcodes, assignment of GUIDs, and creation of a skeletal database record, if these are not completed during proactive digitization or in a later module. Occasionally, trained taxonomists provide determination services at this stage, although remote, postimaging determination is an important benefit of digitization and data exposure and may be part of an institution's overall digitization objectives. This stage does not ordinarily include the development of policies or management plans, tasks that we recommend precede implementation of a digitization program.

Selecting components for an imaging station (module 2)—Selecting an imaging station essentially means selecting and

combining components. Given the disparities among herbaria in available resources, physical space, types of materials to be digitized, and project goals, there is no single solution and few complete, preconfigured options. The workflow document includes an unordered list of considerations (Appendix S2). In general, we recommend purchasing the highest-quality components that budget constraints allow, particularly for the camera/lens or scanner and the light source (see iDigBio, 2014, for equipment suggestions).

Cameras and lenses are especially important and often complicated to evaluate. For example, digital single-lens reflex (DSLR) cameras with full-frame sensors record greater amounts of image data than crop-frame cameras and usually provide better results for herbarium digitization, but they can be considerably more expensive and produce images that require greater storage capacity. Some institutions use relatively inexpensive crop-frame cameras with success. For institutions using an ORTech Photo e-Box Bio Photographic Lighting System (M. K. Digital Direct, Chula Vista, California, USA [http://www.or-tech.com/photo-e-box-bio.html]), full-frame cameras should be fitted with a 50-mm lens for best results. This module provides a list of common considerations, components, and evaluative steps for selecting one component over another, including smaller stations with a single camera, lens, copy stand, light source, and computer. The stations discussed are intended for use indoors to digitize herbarium specimens and cannot be readily used in the field. Removing cameras from copy stands for field use is possible, but risks the introduction of foreign particles to the camera's sensor and may entail a lengthy process to reinstall and calibrate the camera upon return. For these reasons, we do not recommend dual use of digitization cameras.

Imaging station setup (modules 3–5)—There are three widely used imaging station alternatives for herbarium digitization: (1) copy stand with fluorescent lighting (Appendix S3), (2) light box with internal lighting (Appendix S4), and (3) inverted flatbed scanner (Appendix S5). We provide modules for each of these, all of which include ordered lists that guide users through station setup. The workflows begin with basic station assembly and conclude with preliminary image quality control steps and standards for adequate exposure, focus, and color balance. Camera settings are discussed briefly. Given the close association between alternatives 1 and 2, especially in relation to camera and copy stand setup, a thorough review of alternative 1 will be helpful in accomplishing alternative 2.

Selection of an appropriate light source is essential for high-quality image capture. Three practical image lighting setup options are available for specimen imaging: copy stand with light box; copy stand with fluorescent lights; and copy stand with strobe lights. We focus our workflows on the first two of these. Mechanical limitations of the light box and copy stand are discussed, with recommendations for modifications provided. The workflow emphasizes the Photo eBox, pioneered by the New York Botanical Garden (Tulig et al., 2012).

Inverted flatbed scanners have been widely used as a practical means of imaging specimens and have served as the primary imaging station for the Global Plants Initiative (GPI), an international collaborative project aimed at digitizing and making available plant type specimens. Inverted scanners are very easy to operate and can produce high-resolution images with consistent results. Some of the drawbacks include (1) slow scanning process (e.g., one scan may take 5–6 min); (2) only limited material can be scanned (flat herbarium specimens; bulky specimens are not recommended for scanning, as the scanner has a limited focal range and the integrity of the specimen may be compromised [i.e., crushed] when raising the specimen up to the glass); and (3) with the completion of the GPI project, it is likely that the production of imaging stations with inverted flatbed scanners (HerbScans) will be discontinued. Regardless of the disadvantages, inverted flatbed scanners may be a practical solution for many herbaria with limited resources or a small number of specimens. The scanner workflow provides guidelines for setting up a flatbed scanner station based on the scanning protocol in the *JSTOR Plants Handbook*. A few institutions have fabricated replicas of the HerbScan using the same or a similar model scanner, although inverting the device might void its warranty.

Automated, high-throughput conveyor systems are quite common in the manufacturing and food processing industry. These systems and technologies have a role to play in the digitization of natural history specimens and can increase the efficiency of parts of the digitization workflow. Some domains within the natural history community are particularly well-positioned to incorporate these approaches and are borrowing technologies and concepts from the manufacturing industry to increase the efficiency of the specimen imaging step. Indeed, conveyor belt systems are being used to image herbarium specimens (e.g., Tegelberg et al., 2012, 2014; and with the New England Vascular Plants [NEVP] TCN [http://nevp.org/projsummary]), and robotic camera systems are being used to image trays of insects (Buffington et al., 2005; Blagoderov et al., 2010, 2012; Dietrich et al., 2012; Mantle et al., 2012; Schmidt et al., 2012).

Although automated, high-throughput approaches have a role to play in the digitization of herbarium specimens, their use makes sense only within certain institutional, project, economic, and logistical contexts. For example, whether building a conveyor system from scratch (e.g., NEVP approach) or contracting with an outside entity (e.g., Pignal and Michiels, 2011; Digitarium [http://digitarium.fi/en]) to implement a system, there is a significant monetary investment required to install such a system, to train the digitizing staff, and to maintain the system. Such investments only make economic sense when imaging hundreds of thousands of specimens, and a more traditional system (e.g., a light box setup) will be more economical at smaller scales. We do not attempt to provide generalized guidance for setting up and using a conveyor system, as the steps involved in assembling such systems vary depending on the particular system being implemented.

Imaging and image processing (modules 6 and 7)—The imaging workflow is an ordered set of steps through which a specimen sheet is removed from an enclosing folder, imaged, returned to the original folder for refiling, and the resulting image examined for targeted quality. Given the extent of specimen handling required, specimen conservation practices are integral. Specimens should be checked for damage before and after imaging, whether using a copy stand or light box. Those with damage severe enough to detract from the quality or accuracy of the image should be routed for conservation prior to imaging. Less damaged specimens might be imaged before being repaired.

Barcodes should be affixed to sheets prior to recording an image, which might be during pre-digitization curation or in an early step within the imaging workflow, and should be unobstructed and clearly visible in the resulting image to ensure they can be easily and accurately scanned from the image by a barcode reader or read by an optical character recognition (OCR) application. Each image file should be associated with a database record. Although there are several methods for achieving this, many herbaria use the barcode value as the name of the

image file and scan the embedded barcode value into the database (e.g., as the Darwin Core field catalog number) to serve as part of a skeletal record immediately after imaging. Other approaches are described in the workflow document (Appendix S6).

Implementing an image processing workflow (Appendix S7) includes advance planning for information flow, including provision for temporary and longer-term image storage (i.e., archiving strategy), specifications for downstream images, and plans for how and where images will be hosted online. We discuss various ways to link the specimen image file name to an actual physical specimen (specifically using the barcode), perform basic quality control spot checks, and prepare derivative images for day-to-day use. When converting raw image files into other formats, some software (e.g., Adobe Lightroom; Adobe Systems, San Jose, California, USA) preserve the original camera metadata and provide processes for creating derivative files. Additional adjustments may be required (e.g., lens-specific spherical aberrations, white balance, etc.), and these are best done nondestructively in a batch process.

Organizing and implementing a public participation imaging blitz (module 8)—We consider a "blitz" to be a short period of intense effort involving more than the average number of people involved in digitization at a given herbarium (Appendix S8). The blitz might attract members of the public and have informal education objectives, which is what we focus on here, but it could also involve students in an event that has formal education objectives. Public engagement in digitization need not be limited to the short duration of a blitz—it could involve on-site volunteers contributing time to an effort over a much longer duration. In those non-blitz arrangements, the management of volunteers can look very similar to the management of paid digitization technicians. While the longer-term involvement of volunteers can mean greater pay-off for time spent training, there is also a clear tradeoff between the number of volunteers engaged and the time commitment expected. Training can be less extensive and involve more people simultaneously in a blitz. Greater outreach can lead to greater understanding and support in the local community for a herbarium, which can become an important strategy for sustainability of digitization beyond a typical grant funding cycle and make crowdfunding initiatives more successful (e.g., Florida State University's Robert K. Godfrey Herbarium [http://spark.fsu.edu/Projects/121/Blazing-a-New-Trail-for-Sustainability-with-Citizen-Science]). This and the other public participation module (module 12) include advertisement, education, and postevent evaluation considerations that differentiate them from the other modules.

The location of an imaging blitz is constrained by the location of the specimens and the imaging station(s). This typically means that at least a subset of the participants is spending time in the herbarium, which leads to a deeper understanding of the collection but also potentially exposes them to legacy pest-management compounds (e.g., naphthalene can infuse the specimen sheets and persist well after mothballs are removed from cabinets). If the pest-control compound produces a gas, the intense activities of a blitz (opening and closing many cabinets and removing and refiling many folders during the day) can lead to greater ambient levels than compared to an average day of digitization and to discomfort for the digitizers.

Three of us (A.R.M., E.R.E., and G.N.) conducted an imaging blitz at Florida State University's Robert K. Godfrey Herbarium in September 2014, engaging 22 volunteers at three imaging stations over two 4-h shifts to image 3000 specimens (125 specimens per hour per station). Preparation of the specimens for the blitz involved about 60 h of personnel time barcoding the targeted specimens, writing the family name on each folder (to make refiling easier), and application of annotation labels with the currently accepted name where needed. A herbarium technician also marked the beginning and ending of a stretch of targeted folders in the collection with green and red tabs so that they could be easily found in the cabinets on blitz day. After a brief introduction to the collection, schedule, and proper ways to handle specimens, as well as a big "thank you," the volunteers were divided into three 4-person imaging teams, and each team was assigned a coach—a staff member who could train them at an imaging station and check image quality throughout the day. The teams largely settled on a division of labor involving a courier, a photographer, a barcode-scanner, and a folder-compiler. The courier interacted with a staff member whose sole job was to remove targeted folders from the cabinets and refile them once imaged. The imaging teams could have consisted of three, rather than four, people if a single courier had floated among imaging stations. Participants received an event-branded water bottle in appreciation; the bottle was mentioned as an incentive in advertising for the event. After the event, a staff member double-checked image quality and file names for a sampling of the image files (e.g., every 20th) and corrected any errors. Results from a separate, previous blitz are reported in an iDigBio blog (https://www.idigbio.org/content/weekend-digitization-blitz-yields-4276-specimen-images-archbold-biological-station).

Image archiving (module 9)—Sophisticated strategies for digital preservation and archiving are relatively new to biodiversity museums and academic collections. Most institutions back up their digital data and images regularly, but few, except some larger institutions, have developed digital preservation protocols that mirror those in use in the library sciences. Our workflow (Appendix S9) consists of an ordered list of considerations for achieving a true digital asset management system (DAMS). Included are links to numerous documents that provide guidance as well as strong encouragement for building collaborations with an institution's library or DAMS. Perhaps chief among our recommendations is the development of a written digital preservation policy and/or plan that details the institutional goals of biodiversity-related digital asset preservation and the specific methods by which the goals will be attained. Other recommendations in the workflow document include a broad range of issues, including determining archival file formats, recording and preserving data consistent with several metadata standards, asset identification, determining and attributing ownership, and image transfer protocols and strategies.

Determination of file type for image preservation is of immediate need to most institutions just launching a digitization program. We list several possibilities. Different opinions exist on the retention of camera or converted raw vs. TIFF (Tagged Image File Format) images. While the uncompressed TIFF format has been the standard in the library community, other openly documented raw formats (e.g., DNG [Digital NeGative]) are also in wide use and now compete with the convention of relying on TIFF as the archival format. Conversion of proprietary camera raw files to these or other publicly documented formats is recommended and can be accomplished with several existing software applications (e.g., Adobe DNG Converter and Adobe Lightroom).

***Selecting a database (module 10)*—**Database selection typically depends on an institution's size, digitization goals, data sharing policies, financial resources, and available on-site IT infrastructure and support. This module is chiefly an unordered list of considerations for selecting an appropriate database for a given situation (Appendix S10). We address a variety of options, with a major focus on NSF-funded open access systems that can be freely downloaded or accessed online, such as Specify (http://specifyx.specifysoftware.org/) and Symbiota (Gries et al., 2014; http://symbiota.org/docs/). Symbiota provides the backbone for several existing herbarium digitization networks (http://swbiodiversity.org/portal/index.php, http://nansh.org/portal/, http://sernecportal.org/portal/, http://bryophyteportal.org/portal/, http://lichenportal.org/portal/, http://portal.neherbaria.org, http://macroalgae.org/portal/, http://mycoportal.org/portal/, http://greatlakesinvasives.org/portal/index.php) and has proven especially relevant to small herbaria. Specify is a complete collections management system in use with approximately 100 plant-related collections (Theresa Miller, University of Kansas, personal communication, 2015), with support for accessioning, loan management, storage tree definitions, and other collections management tasks.

Lack of IT infrastructure and technical support is a major impediment to digitization for smaller, resource-challenged collections, as is preparing data for publication to the Web through data aggregators (e.g., iDigBio, GBIF). We include recommendations that mitigate these obstacles. For larger institutions with adequate IT support and the need for a sophisticated collections management system, we address other options within the workflow document.

***Data capture (module 11)*—**Data capture is ostensibly the most important component of any herbarium digitization program. Accurately transcribing label data in sufficient detail to facilitate searching across several dimensions, especially in the absence of associated specimen images, is essential to discovery and research uses (e.g., Shanmughavel, 2007; Scoble and Bourgoin, 2010; Feeley and Silman, 2011; Schuh, 2012). Effective transcription of specimen labels to a permanent database is dependent on answers to several critical decisions to be made at the institutional level prior to launching a digitization program (Box 3).

Whether sensitive data will be redacted from published records is an institutional decision. There is, as of yet, no clear consensus across collection domains or within the herbarium community. Chapman and Grafton (2008) presented a set of best practices governing this issue, asserting that "Wherever possible, environmental information should be made available to all," cautioning that in cases where release might result in harm, the presumption remains in favor of release and that the need for restriction should be rigorously reviewed. Whether an institution chooses to redact locality data for sensitive species often depends on curator preference; land manager, landowner, or heritage program request; and concern for the conservation of endangered species. We have not addressed this issue in the workflows (Appendix S11), instead leaving it as an institutional decision.

***Organizing and implementing a public participation transcription blitz (module 12)*—**An on-site transcription blitz can look very much like an imaging blitz when the specimens are the source of the information being transcribed, or it can be quite different, if the source of the information is a digital specimen image (Appendix S12). The latter relieves potential space constraints in the herbarium by opening up venues such as campus computer laboratories. Furthermore, when the images and platform for transcription are online (e.g., using one of the platforms reviewed by Ellwood et al., 2015), software requirements are limited to now-ubiquitous Web browsers. Use of an online platform (e.g., Notes from Nature; Hill et al., 2012) can, of course, open up new, Internet-scale public engagement possibilities largely unconstrained by the space available on-site and the timing of a workweek. We do not provide a module for decentralized digitization activities of this type but instead focus on activities for an on-site transcription blitz that could make use

Box 3. Critical decisions to be made at the institutional level prior to launching a digitization program.

- Will data be captured from physical specimens or images of specimens?
- Will populated database records include all data recorded on the label or an abbreviated set of label data (often called skeletal records; Tulig, 2014; Rabeler, 2015)?
- Will data be entered directly into the permanent database or into an intermediate transitory format for later uploading (e.g., spreadsheet; Neefus, 2014)?
- Will sensitive data be redacted (Chapman and Grafton, 2008)?
- Will georeferencing and other enrichment data be recorded concurrently with label data, in batch through processes integrated into the permanent or transitory database, or as a separate activity (Chapman and Wieczorek, 2006)?
- Do clear instructions exist for handling entry of duplicate specimens? Can entry of the repeated information be made more efficient for duplicates held within an institution or between institutions using the same specimen data management system (e.g., Symbiota)?
- Will optical character recognition (OCR) or voice recognition be used (Haston et al., 2012; Butts, 2013; Neefus, 2014)?
- Will verification history or other annotation data be recorded?
- Are quality assurance and verification protocols in place to enhance accuracy (Chapman, 2005)?
- Are data entry technicians adequately selected and trained, and do they have at their disposal a detailed written protocol to guide decisions about how data should be parsed and entered?
- Are procedures in place to route damaged specimens for conservation (see Department of the Interior, 2009)?
- Are procedures in place for handling misfiled specimens?

of online platforms while gaining the benefits of education and outreach to the herbarium's local community. There are many parallels between an on-site blitz involving transcription with online tools and blitzes involving the other two core areas of online digitization discussed by Ellwood et al. (2015)—georeferencing specimen collection localities and annotating specimen images—and this workflow should largely be transferrable to those other activities. The potential monotony of transcriptions at a computer monitor can be ameliorated by adding games to the event that do not distract from the specimens, but instead require closer attention to them. For example, two of us (A.R.M. and E.R.E.) developed a set of game cards (e.g., bingo with common habitat terms) that can be used at an event with small prizes. The use of these games is described in a recent blog post about the March 2015 transcription blitz held simultaneously at Florida State University and Valdosta State University.

***Georeferencing (module 13)*—**The vast majority of herbarium specimens have some form of locality information incorporated into the specimen label. These data typically consist of country, state, county, and a written locale description. These descriptions provide critical information for pinpointing the collection site but cannot be readily incorporated into most digital maps or analyses. Georeferencing consists of transforming descriptive locality information into numerical coordinates with associated extent, geospatial datum, and uncertainty measures (Chapman and Wieczorek, 2006). These numerical coordinates increase the ease of digitally referencing and relocating the specimen locality, which permits a wide array of biodiversity and geographical analyses (range distributions, species distribution modeling, etc.).

Due to the variety, breadth, and age of locality information that may be present on a specimen label, georeferencing is far from a simple task. Before embarking on an effort to georeference your collection, we highly recommend that you and your digitizing staff review the various training materials that are available online or participate in a georeferencing training workshop (See http://georeferencing.org/online-training-resources.html; http://georeferencing.org/index.html). Due to the vagaries and imprecision inherent in locality descriptions, the georeferenced point assigned to a locality is often only a rough approximation of a physical collecting site's geographical location. Hence, describing the method used during the georeferencing process and the estimated precision of the derived point are essential.

We have outlined a basic protocol for efficiently georeferencing a specimen database en masse after digitization of the specimen labels (Appendix S13). The precise georeferencing protocol employed by an institution should be customized to that institution's collections depending on resources and data. Even in ideal situations, georeferencing a single locality can take minutes, making georeferencing a complete collection a formidable task, especially when investing the time to visually ensure accuracy.

***Proactive digitization (module 14)*—**We refer to proactive digitization as the act of moving digitization activities upstream in the collecting process to eliminate the creation of new legacy data and to encourage the submission of digitized data concurrent with the deposition of physical specimens (Appendix S14). In most cases where proactive digitization is employed, researchers and collectors use preformatted spreadsheets (Karim et al., unpublished) and electronic devices (tablets, smartphones, data loggers such as those by Trimble [Sunnyvale, California, USA], etc.) to record data while in the field. Strides have been made in capturing and downloading geographic coordinates directly from Global Positioning Systems for immediate transfer to a database, reducing the likelihood of transcription errors that incorrectly report the geographic position of collecting localities. As field-based use of existing and emerging electronic technology becomes more common, similar processes will become commonplace for other types of collections data. Procedures being developed for the Field Information Management System at the Smithsonian Institution (Gamble and Whitacre, 2014) and Museum of Comparative Zoology at Harvard (A. Williston, Harvard University, personal communication) exemplify this trend.

***Conclusions*—**Efficient workflows provide the foundation for successful digitization of biodiversity collections and foster the mobilization of increased quantities of specimen data for scientific research, natural resource management, education, and policy-making. The 14 workflow modules detailed here include substantial revision to the original six modules (iDigBio, 2012) and the addition of eight new modules. We believe these refinements and additions will further increase the availability of digitized data and enhance the opportunities for specimen-based botanical research. Nevertheless, we recognize that implementation of these workflows will lead to further refinement and expansion. Although some of the workflows presented here (e.g., pre-digitization curation, imaging, image processing, and data capture) represent relatively mature protocols based on extensive experience, others, especially image archiving, proactive digitization, and organizing and executing public participation blitzes, are ripe for testing and enhancement.

This is but one source of important resources for digitizing, or soon-to-be digitizing, herbaria. Several professional organizations serve as forums for discussion about digitization and produce resources for the digitizing community, including the Society for the Preservation of Natural History Collections, Biodiversity Information Standards, Society of Herbarium Curators, and Small Collections Network. Herbaria should also consider participating in relevant training programs sponsored by, e.g., iDigBio, Data Carpentry, and relevant software tools.

While we strongly encourage institutions in the process of customizing workflows to keep in mind fitness of use of their data products for common research applications (e.g., species distribution modeling), we recognize that greater availability of the data online could drive exciting, novel research uses that are difficult to anticipate. New integrations of diverse data, including specimen locality data, phenology, phylogeny, genetic variation, tissue isotope ratios, metagenomic sequences, microbial function, plant functional traits, environmental data, climate models, etc., could fuel new waves of specimen data sampling requiring new modules and changes to institutional strategies. New scientific innovations requiring the widespread creation of less common or yet-to-be-imagined derivative data from specimens reinforce the importance of the co-curation of the physical specimens along with the digital data.

LITERATURE CITED

Barkworth, M. E., and Z. E. Murrell. 2012. The US Virtual Herbarium: Working with individual herbaria to build a national resource. *ZooKeys* 209: 55–73.

Beach, J., S. Blum, M. Donoghue, L. Ford, R. Guralnick, M. Mares, B. Thiers, et al. 2010. A strategic plan for establishing a network

integrated biocollections alliance [online]. Website http://digbiocol.wordpress.com/brochure [accessed 27 April 2015].

Blagoderov, V., I. Kitching, T. Simonsen, and V. S. Smith. 2010. Report on trial of SatScan tray scanner system by SmartDrive Ltd. [online]. Website http://precedings.nature.com/documents/4486/version/1 [accessed 27 April 2015].

Blagoderov, V., I. J. Kitching, L. Livermore, T. J. Simonsen, and V. S. Smith. 2012. No specimen left behind: Industrial scale digitization of natural history collections. *ZooKeys* 209: 133–146.

Buffington, M. L., R. A. Burks, and L. McNeil. 2005. Advanced techniques for imaging parasitic Hymenoptera (Insecta). *American Entomologist* 51: 50–56.

Butts, S. 2013. YPM-IP Workflow with Voice Recognition Imaging [online]. Website https://www.idigbio.org/content/ypm-ip-workflow-voice-recognition-imaging [accessed 27 April 2015].

Casas-Marce, M., E. Revilla, M. Fernandes, A. Rodríguez, M. Delibes, and J. A. Godoy. 2012. The value of hidden scientific resources: Preserved animal specimens from private collections and small museums. *BioScience* 62: 1077–1082.

Chapman, A. D. 2005. Principles and methods of data cleaning—Primary species and species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen, Denmark. Website http://www.gbif.org/resource/80528 [accessed 27 April 2015].

Chapman, A. D., and O. Grafton. 2008. Guide to best practices for generalising sensitive species occurrence data. Report for the Global Biodiversity Information Facility, Copenhagen, Denmark. Website http://www.gbif.org/resource/80512 [accessed 27 April 2015].

Chapman, A. D., and J. Wieczorek. 2006. Guide to best practices for georeferencing. Report for the Global Biodiversity Information Facility, Copenhagen, Denmark. Website http://www.gbif.org/resource/80536 [accessed 27 April 2015].

Department of the Interior. 2009. Interior collections management system, Chapter 4: Associated Modules [online]. Website http://www.nps.gov/museum/publications/ICMSmanual/14-Chapter%204-Associated%20Modules.pdf [accessed 27 April 2015].

Dietrich, C., J. Hart, D. Raila, U. Ravaioli, N. Sobh, O. Sobh, and C. Taylor. 2012. InvertNet: A new paradigm for digital access to invertebrate collections. *ZooKeys* 209: 165–181.

Ellwood, E. R., B. A. Dunckel, P. Flemons, R. Guralnick, G. Nelson, G. Newman, S. Newman, et al. 2015. Accelerating the digitization of biodiversity research specimens through online public participation. *BioScience* 65: 383–396.

Feeley, K. J., and M. R. Silman. 2011. Keep collecting: Accurate species distribution modeling requires more collections than previously thought. *Diversity & Distributions* 17: 1132–1140.

Gamble, B., and J. Whitacre. 2014. Genetic sampling: From field to freezer and sharing the data [online]. Website https://emu.kesoftware.com/news-and-events/conferences/2146-global-emu-user-conference-15-17-october-2014 [accessed 27 April 2015].

Gries, C., E. E. Gilbert, and N. M. Franz. 2014. Symbiota–A virtual platform for creating voucher-based biodiversity information communities. *Biodiversity Data Journal* 2: e1114.

Haston, E., R. Cubey, M. Pullan, H. Atkins, and D. J. Harris. 2012. Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. *ZooKeys* 209: 93–102.

Hauser, C., A. Steiner, J. Holstein, and M. Scoble. 2005. Digital imaging of biological type specimens: A manual of best practice. European Network for Biodiversity Information, Stuttgart, Germany. Website http://www.gbif.org/resource/80576 [accessed 27 April 2015].

Hill, A., R. Guralnick, A. Smith, A. Sallans, R. Gillespie, M. Denslow, J. Gross, et al. 2012. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. *ZooKeys* 209: 219–233.

iDigBio. 2012. Workflow modules and task lists [online]. Website https://www.idigbio.org/content/workflow-modules-and-task-lists [accessed 27 April 2015].

iDigBio. 2014. iDigBio imaging equipment recommendations. Version 2.0 [online]. Website https://www.idigbio.org/wiki/images/8/86/IDigBioImagingGeneralEquipmentRecommendations1_0.pdf [accessed 27 April 2015].

Johnson, K. G., S. J. Brooks, P. B. Fenbert, A. G. Glover, K. E. James, A. M. Lister, E. Michel, et al. 2011. Climate change and biosphere response: Unlocking the collections vault. *BioScience* 61: 147–153.

Loarie, S. R., B. E. Carter, K. Hayhoe, S. McMahon, R. Moe, C. A. Knight, and D. D. Ackerly. 2008. Climate change and the future of California's endemic flora. *PLoS One* 3: e2502.

Mantle, B. L., J. La Salle, and N. Fisher. 2012. Whole-drawer imaging for digital management and curation of a large entomological collection. *ZooKeys* 209: 147–163.

Neefus, C. 2014. Digitization workflow for a small herbarium [online]. Website https://www.idigbio.org/content/digitization-workflow-small-herbarium [accessed 27 April 2015].

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

Pignal, M., and H. Michiels. 2011. Switching to the fast track: Rapid digitization of the world's largest herbarium [online]. Website http://collections.mnhn.fr/wiki/attach/Visit_October2012/Paris-Herbarium-Digitization_2012-07-12.pdf [accessed 27 May 2015].

Rabeler, R. 2015. Skeletal records accompanying images: Efficiency vs later utility. Presentation made to the annual meeting of the Society for the Preservation of Natural History Collections. Website https://www.idigbio.org/content/skeletal-records-accompanying-images-efficiency-vs-later-utility [accessed 27 April 2015].

Schmidt, S., M. Balke, and S. Lafogler. 2012. DScan—A high-performance digital scanning system for entomological collections. *ZooKeys* 209: 183–191.

Schuh, R. 2012. Integrating specimen databases and revisionary systematics. *ZooKeys* 209: 255–267.

Scoble, M. J., and T. Bourgoin. 2010. Natural history collections digitization: Rationale and value. *Biodiversity Informatics* 7: 77–80.

Shanmughavel, P. 2007. An overview on biodiversity information in databases. *Bioinformation* 1: 367–369.

Snow, N. 2005. Successfully curating smaller herbarium and natural history collections in academic settings. *BioScience* 55: 771–779.

Soltis, P. S., X. Liu, D. B. Marchant, C. J. Visger, and D. S. Soltis. 2014. Polyploidy and novelty: Gottlieb's legacy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 369: 20130351.

Tegelberg, R., J. Haapala, T. Mononen, M. Pajari, and H. Saarenmaa. 2012. The development of a digitising service centre for natural history collections. *ZooKeys* 209: 75–86.

Tegelberg, R., T. Mononen, and H. Saarenmaa. 2014. High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. *Taxon* 63: 1307–1313.

Thiele, K. 2014. Australia's virtual herbarium: 5 million records and counting [online]. Website http://avh.chah.org.au/index.php/2014/08/12/press-release/ [accessed 27 April 2015].

Thiers, B. 2015. Index Herbariorum [online]. Website http://sciweb.nybg.org/science2/IndexHerbariorum.asp [accessed 27 April 2015].

Tulig, M. 2014. Data capture from images [online]. Website http://biodiversity-informatics-training.org/wp-content/uploads/2014/03/D3_P3_MT_OCR.pdf [accessed 27 April 2015].

Tulig, M., N. Tarnowsky, M. Bevans, A. Kirchgessner, and B. M. Thiers. 2012. Increasing the efficiency of digitization workflows for herbarium specimens. *ZooKeys* 209: 103–113.

U.S. Virtual Herbarium. 2015. U.S. Virtual Herbarium 2014 survey data [online]. Website http://usvhproject.org/#/progress [accessed 27 April 2015].

Vollmar, A., J. A. Macklin, and L. S. Ford. 2010. Natural history specimen digitization: Challenges and concerns. *Biodiversity Informatics* 7: 93–112.