METHODS AND TECHNIQUES

Large-scale digitization of herbarium specimens: Development and usage of an automated, high-throughput conveyor system

Patrick W. Sweeney,[1] Binil Starly,[2] Paul J. Morris,[3] Yiming Xu,[4] Aimee Jones,[5] Sridhar Radhakrishnan,[6] Christopher J. Grassa[7] & Charles C. Davis[8]

1 *Division of Botany, Peabody Museum of Natural History, Yale University, P.O. Box 208118, New Haven, Connecticut 06520, U.S.A.*
2 *Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina 27607, U.S.A.*
3 *Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138, U.S.A.*
4 *School of Computer Science, University of Oklahoma, Norman, Oklahoma 73019, U.S.A.*
5 *School of Industrial and Systems Engineering, University of Oklahoma, Norman, Oklahoma 73019, U.S.A.*
6 *School of Computer Science, University of Oklahoma, Norman, Oklahoma 73019, U.S.A.*
7 *Harvard University Herbaria, Harvard University, Cambridge, Massachusetts 02138, U.S.A.*
8 *Department of Organismic and Evolutionary Biology and Harvard University Herbaria, Harvard University, Cambridge, Massachusetts 02138, U.S.A.*
Author for correspondence: *Patrick W. Sweeney, patrick.sweeney@yale.edu*

DOI https://doi.org/10.12705/671.9

Abstract The billions of specimens housed in natural science collections provide a tremendous source of under-utilized data that are useful for scientific research, conservation, commerce, and education. Digitization and mobilization of specimen data and images promises to greatly accelerate their utilization. While digitization of natural science collection specimens has been occurring for decades, the vast majority of specimens remain un-digitized. If the digitization task is to be completed in the near future, innovative, high-throughput approaches are needed. To create a dataset for the study of global change in New England, we designed and implemented an industrial-scale, conveyor-based digitization workflow for herbarium specimen sheets. The workflow is a variation of an object-to-image-to-data workflow that prioritizes imaging and the capture of storage container-level data. The workflow utilizes a novel conveyor system developed specifically for the task of imaging flattened herbarium specimens. Using our workflow, we imaged and transcribed specimen-level data for almost 350,000 specimens over a 131-week period; an additional 56 weeks was required for storage container-level data capture. Our project has demonstrated that it is possible to capture both an image of a specimen and a core database record in 35 seconds per herbarium sheet (for intervals between images of 30 minutes or less) plus some additional overhead for container-level data capture. This rate was in line with the pre-project expectations for our approach. Our throughput rates are comparable with some other similar, high-throughput approaches focused on digitizing herbarium sheets and is as much as three times faster than rates achieved with more conventional non-automated approaches used during the project. We report on challenges encountered during development and use of our system and discuss ways in which our workflow could be improved. The conveyor apparatus software, database schema, configuration files, hardware list, and conveyor schematics are available for download on GitHub.

Keywords automation; biodiversity informatics; digitization; herbarium specimens; imaging; New England; transcription; workflows

■ INTRODUCTION

The world's natural science collections house as many as three billion specimens (Alberch, 1993; Soberon, 1999; Ariño, 2010) and form the basis of our understanding of the biodiversity on our planet. These collections provide a wealth of data that are useful for scientific research, conservation, commerce, and education (Suarez & Tsutsui, 2004; Pyke & Ehrlich, 2010; Drew, 2011; Lister & al., 2011; Powers & al., 2014; Tewksbury & al., 2014).

To increase the accessibility of these collections, there have been numerous calls for their digitization (i.e., capturing specimen-level data and specimen images into digital form) and mobilization through the Internet (e.g., Crovello, 1967; Creighton & Crockett, 1971; Whitehead, 1971; Croft, 1989; Berendsohn & al., 1997, 2010; Chavan & Krishnan, 2003; Baird, 2010; Drew, 2011; Johnson & al., 2011; Beaman & Cellinese, 2012; Balke & al., 2013). To this end, collections digitization has been in progress since at least the early 1970s (Sunderland, 2013) and hundreds of millions of digitized specimen records and images have been made available online (e.g., http://www.gbif.org; http://portal.idigbio.org). However, the vast majority of specimens are yet to be digitized (e.g., Ariño, 2010; Barkworth & Murrell, 2012) and remain as dark data (sensu Heidorn, 2008). The kinds of scientific research that can make use of these digitized specimen records is extensive and includes

Article history: Received: 11 Apr 2017 | returned for (first) revision: 29 May 2017 | (last) revision received: 19 Sep 2017 | accepted: 20 Sep 2017 | published: online fast track, 12 Feb 2018; in print and online issues, 6 Mar 2018 || **Associate Editor:** Jürg Schönenberger

research on global change (e.g., Lavoie, 2013; Vellend & al., 2013; Davis & al., 2015; Harsch & HilleRisLambers, 2016), evolution (e.g., Doudna & Danielson, 2015; Davis Rabosky & al., 2016), and conservation (e.g., Greve & al., 2016), to name a few.

If digitization of the world's natural science collections is to be completed in the foreseeable future, fresh approaches are required (Hanken & al., 2013). It is well understood that the historical approach of moving through a single collection, specimen by specimen, capturing and validating data in depth at the time of specimen handling, does not scale to data capture on the order of hundreds of thousands, let alone millions, of specimens. In the 1990s and early 2000s, very significant progress was made in digitizing specimen occurrence data from bulk records such as handwritten ledgers, without specimen handling (e.g., large-scale data capture projects involving ichthyological, mammalogical, herpetological, and ornithological collections, which had mobilized some 52 million vertebrate specimen records by 2010; Stein & Wieczorek, 2004; Peterson & al., 2006; Constable & al., 2010). For disciplines where bulk records have historically been kept, a first pass at creating occurrence records is straightforward. However, for other disciplines, in particular entomology and botany, where frequently the only place to find the specimen data is attached to the specimens themselves (but see Tulig & al., 2012), very large gaps remain. Achieving very high digitization throughput rates in these disciplines will require innovative approaches based on a clear understanding of effective workflows for digitization (Nelson & al., 2012), potentially incorporating technologies such as the use of industrial hardware and software automation systems (Blagoderov & al., 2012). Such large-scale digitization and mobilization of natural science collection data offers significant informatics challenges; nevertheless, many groups are making headway. Along these lines, conveyor belt systems have been used to greatly increase the imaging rates of herbarium specimens and/or insects (Tegelberg & al., 2014; Heerlien & al., 2015), and there are even vended solutions geared towards the digitization of herbarium specimens (Picturae, https://picturae.com/uk/). Robotic cameras have been used to increase the efficiency of imaging insect drawers (Blagoderov & al., 2012; Dietrich & al., 2012; Schmidt & al., 2012), and software-based approaches have been developed to increase efficiency of various aspects of the overall digitization process (e.g., Granzow-de la Cerda & Beach, 2010; Bertone & al., 2012; Barber & al., 2013; Schmidt & al., 2012; Hudson & al., 2015).

Here, we report on the development, testing, use, and efficiency of an industrial-scale, high-throughput digitization system for vascular plant herbarium specimens mounted to sheets. A central feature of our system is an automated conveyor belt apparatus, which allows image capture and specimen handling to occur simultaneously, and which prioritizes imaging and the capture of storage container-level data. This system was developed as part of the Mobilizing New England Vascular Plant Specimen Data to Track Environmental Changes project (NEVP), a Thematic Collections Network (TCN) funded through the U.S. National Science Foundation's Advancing the Digitization of Biodiversity Collections (ADBC) program. The conveyor system was utilized within a larger context that aimed to digitize and mobilize vascular plant specimen data for over one million New England vascular plant specimens distributed among 17 large to small herbaria located across the region. A key aim of this project was to create a massive collection of herbarium specimen records and images that could be used for research on the effects of climate change and land use history in New England. More detailed information about this and other aspects of this project can be found on the NEVP project website (http://nevp.org/resources). And all of the specimen data and images from this effort can be downloaded at the Consortium of Northeastern Herbaria (CNH) portal (http://portal.neherbaria.org).

Table 1. Occurrence data elements and phase of the workflow in which they were captured.

Darwin Core Term	Captured in phase	Controlled vocabulary
dwc:scientificName	Pre-capture	Yes
dwc:scientificNameAuthorship	Pre-capture	Yes
dwc:stateProvince	Pre-capture or specimen-level data/image capture	Yes
dwc:county	Specimen-level data/image capture	Yes
dwc:municipality	Specimen-level data/image capture	Yes
dwc:locality	Specimen-level data/image capture, if not a municipality	No
dwc:recordedBy [collector]	Specimen-level data/image capture	Yes
dwc:recordNumber [collector's number]	Specimen-level data/image capture	No
dwc:eventDate	Specimen-level data/image capture	Interpreted from verbatim
dwc:verbatimEventDate	Specimen-level data/image capture	No
dwc:decimalLatitude	Enhancement	Gazetteer
dwc:decimalLongitude	Enhancement	Gazetteer
dwc:geodeticDatum	Enhancement	Gazetteer
dwc:coordinateUncertantyInMeters	Enhancement	Gazetteer
dwc:reproductiveState	Enhancement	Yes

■ MATERIALS AND METHODS

Overview. — The overall workflow we developed for the project was a variation of an object-to-image-to-data workflow that partly, but not completely, separates data transcription from specimen handling and imaging. In this workflow, a core set of specimen occurrence data was captured for each herbarium sheet by the time of imaging. The inclusion of particular terms in this core set was based on the science goals of the project and to maximize subsequent uses of the data. The specimen occurrence data captured during this project is given in Table 1. The workflow consisted of four phases. Figure 1 presents a diagram for the overall workflow. In the first phase (pre-capture), data were captured from storage containers (herbarium folders) without individual specimen handling. In the second phase (specimen-level data capture and imaging), the specimens were handled and imaged, the storage-level data were associated with specimen records, and additional core specimen-level data were captured. In the third phase (data and image transport), data were transported to both a remote portal and the database of record, images were placed in a digital asset management system, and links were made between the specimen records and the images. In the fourth phase (data enhancement), the specimen data was augmented through crowdsourcing and by adding georeferences.

Phase 1, Selection of specimens and storage container data capture (pre-capture). — New England vascular plant specimens housed in five collections of the Harvard University Herbaria (A, AMES, ECON, GH, NEBC) were targeted for digitization with the automated conveyor system described below. In a "pre-capture" phase (Morris & al., 2010a, b, 2014), the storage organization of the collection was exploited to capture scientific names and higher geographic information in bulk. The vascular plant collections of the Harvard University Herbaria (HUH) were generally organized in cabinets and shelves by higher taxon and geography, and at the folder level by current identification, generally at the species rank or lower. In an unavoidable inefficiency imposed by the scope of the project, folders of target specimens were selected by moving through the collections, cabinet-by-cabinet, and flagging sets of folders of specimens collected in New England (although within each cabinet these folders were grouped together in large batches, making this a less onerous task). An existing Java application was productized and generalized to facilitate the capture of folder-level data (Morris, 2011, 2013). While the application allowed for the capture of several folder-level data elements, scientific name and state were the main fields captured, and both of these fields utilized controlled vocabularies. The end result of this process was a folder label with the folder-level data encoded in machine readable form as structured data in JSON in a QR code barcode and in human readable form. Figure 2 shows a folder of specimens with a QR code label affixed. The QR code labels were specifically for the purpose of facilitating container-level data capture; the data present in the QR code folder labels were associated with specimen-level records as individual herbarium specimens entered into the specimen-level data and image capture phase.

Fig. 2. An open folder of specimens. **A,** The left panel of the folder has a QR code label affixed (white arrow). The data represented on the QR code label applies to all specimens within the folder. On the right, is the stack of specimens contained within the folder. On the topmost specimen, the catalog number barcode label is visible (black arrow). **B,** Close-up view of QR code label shown in A.

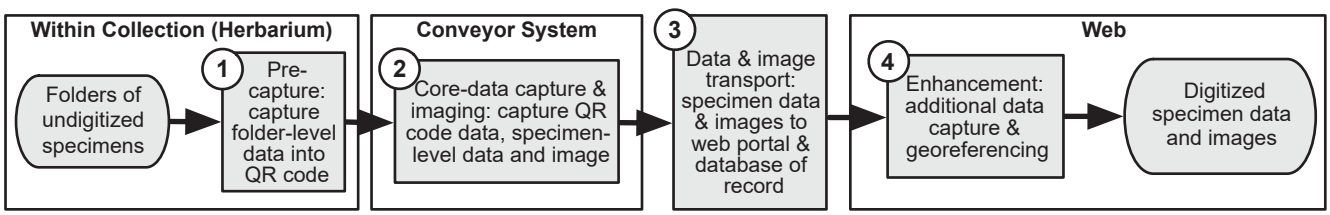

Fig. 1. Overview of overall workflow with four phases labeled 1–4. Phase **1,** pre-capture; Phase **2,** specimen-level data capture and imaging; Phase **3,** data and image transport; Phase **4,** data enhancement.

Phase 2, Specimen-level data and image capture. — We designed a high-throughput conveyor apparatus and accompanying workflow to capture a specimen image and a sub-set of specimen-level label data and to link folder-level data to specimen occurrence records. Development of the apparatus and workflow was conducted by engineering teams (lead by co-author Starly) based at North Carolina State University (NCSU) and the University of Oklahoma (OU) with guidance from curatorial and informatics personnel at Yale and Harvard.

The hardware and software system. – A system was developed with four major sub-system components. Figure 3 portrays a system diagram with the major hardware components, and Fig. 4 shows how the system components interacted. The first sub-component, the Data Entry User Interface system (DEUI), was a graphical user interface created to capture a specimen barcode number and folder- and specimen-level data. Figure 5 shows a screenshot of the DEUI form. Bar-code readers (Honeywell Zenon 1900), keyboard entry and touch screen mode operations were the primary input forms of data entry. The DEUI was run on a dedicated Hewlett Packard Touchsmart HP520-1070 touchscreen computer. The second sub-component, the System Controller (henceforth referred to as "Controller"), was responsible for the overall control and operation of the system, including conveyor control, imaging, and local database storage. The Controller software ran on a Hewlett Packard Z220 computer and operated without direct user control. The Controller computer had a graphical interface that allowed for a user to monitor image and data quality. The third sub-component, the camera and lighting system, captured a high-quality digital image of each herbarium specimen. The camera was a Canon EOS 5D Mark II with a 50 mm f/2.5 macro lens. The lighting system was custom built by Tarsia Technical Industries (Hauppauge, New York) and utilized LED lighting (TTI-LED Lighting 2100-78 lamp system, 12 lamps (6 per side), 5700K color temperature). The conveyor material handling system (see Fig. 3 for component details), the fourth sub-component, was responsible for safely securing, transporting, and delivering herbarium sheets from one user to another. A data management system based on a MySQL server backend was designed and built to capture all data generated from each herbarium specimen. The MySQL database was designed to facilitate the mapping of all specimen occurrence data to the Darwin Core standard (Wieczorek & al., 2012; see below, "Phase 3").

Local area Wi-Fi was used to transmit data from the Data Entry User Interface computer to the System Controller. Hi-speed data transfer of images from the camera to the System Controller was achieved using a USB 2.0 cable. A National Instruments Data Acquisition Board (DAQ) acquired the digital signals to be converted to analog to instruct the conveyor belt

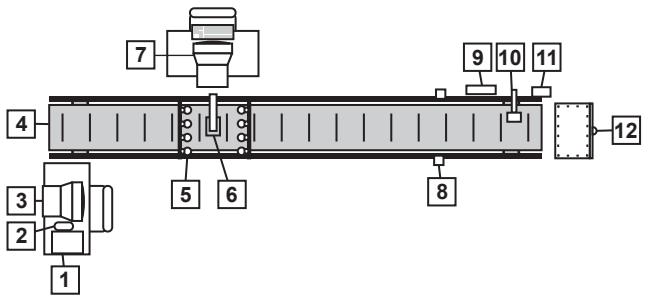

Fig. 3. System diagram with layout and named hardware components. **1,** Initial specimen staging area; **2,** Barcode scanner (Honeywell Zenon 1900); **3,** Data Entry User Interface (DEUI) touchscreen computer (Hewlett Packard Touchsmart HP520-1070), **4,** Conveyor belt (Dorner Mfg. Corp.); **5,** Lighting system (Tarsia Technical Industries); **6,** DSLR camera (Canon EOS 5D Mark II); **7,** Main (System Controller) computer (Hewlett Packard Z220); **8,** Photo-eye (Dorner Mfg. Corp.); **9,** Conveyor controller (Dorner Mfg. Corp.); **10,** Contrast laser sensor (SICK, Inc. model KT8L); **11,** Conveyor motor (Dorner Mfg. Corp.); **12,** Specimen return tray.

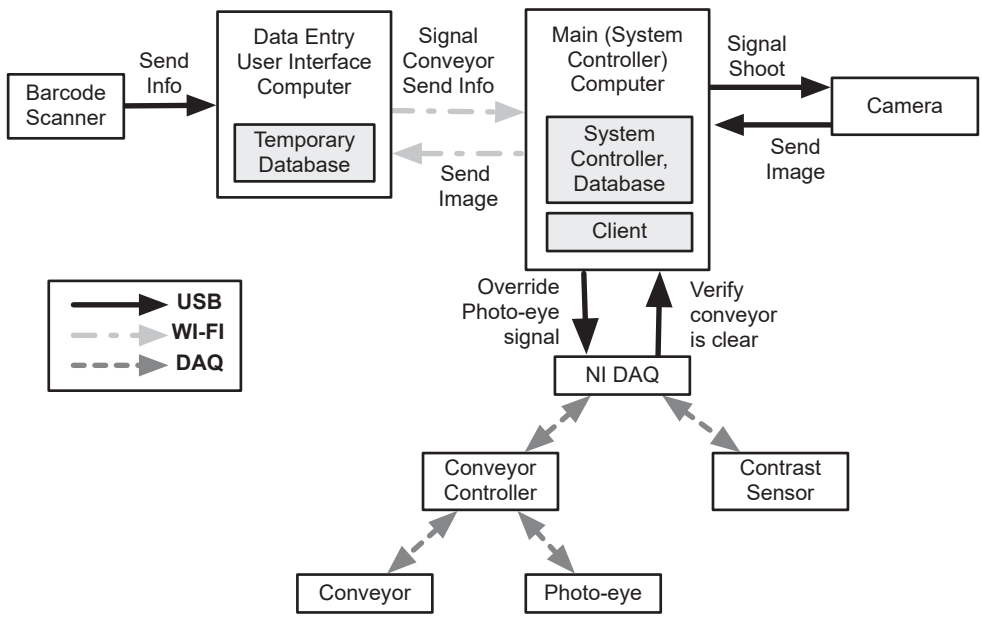

Fig. 4. System interaction diagram. The chart shows how the data is transmitted across systems.

to move when signaled by the operator. The DAQ was also responsible for linking the photo-eye sensors, the contrast sensors, and the conveyor motor to the System Controller.

To aid in development of the system described above, a series of preliminary time studies were performed to analyze the time elapsed considering different system configurations. Two kinds of time studies were conducted, one focused on the specimen-level data entry step of the workflow, and the second on conveyor belt length. To achieve maximum efficiency of specimen-level data entry, a series of tests, before and after various improvements, of the data entry user interface were conducted (Table 2, Data Entry User Interface tests). Both typing the full verbatim date and using a date selection process from the form were timed. Full data entry (i.e., capturing all targeted specimen-level data) was compared to only capturing the QR code and barcode information. To determine how conveyor belt length influenced efficiency, a simulation of a copy stand imaging station (i.e., no conveyor; 1 user), a 1.8-meter conveyor (1 user), and a conveyor longer than 3 meters (2 users) were timed (Table 2, Conveyor length efficiency tests). A final test that simulated production usage on a 3-meter conveyor belt with full data entry was conducted.

Choice of camera equipment was driven by the research aims of the project and other practical considerations. Our aim was to generate specimen images that would allow for transcription of label data and allow for scoring of reproductive phenology (e.g., Willis & al., 2017). We were also interested in generating images that would allow for as many other downstream uses as possible, without significantly slowing conveyor throughput rates. We settled on a full frame, 21-megapixel, DSLR camera (Canon EOS 5D Mark II) with a 50mm f/2.5 macro lens (Canon EF 50mm f/2.5 Compact Macro Lens). At the time that this project was initiated, this camera and lens combination was in widespread use by the United States herbarium community and had a proven history of delivering

Table 2. Preliminary time study results.

Test	Number of users	Collector name entered?	Simulated barcode attachment?	Average time/specimen
Data Entry User Interface tests				
Manually enter all dates	–	No	–	~45 s
Manually enter date by button selection	–	No	–	~35 s
Manually enter verbatim date	–	No	–	~25 s
Scan and place only	–	No	–	~8 s
Conveyor length efficiency tests				
Copystand simulation	1	No	No	~33 s
1.8-meter conveyor	1	No	No	~45 s
3-meter conveyor	2	No	No	~24 s
1.8-meter conveyor	1	Yes	Yes	~48 s
3-meter conveyor	2	Yes	Yes	~39 s

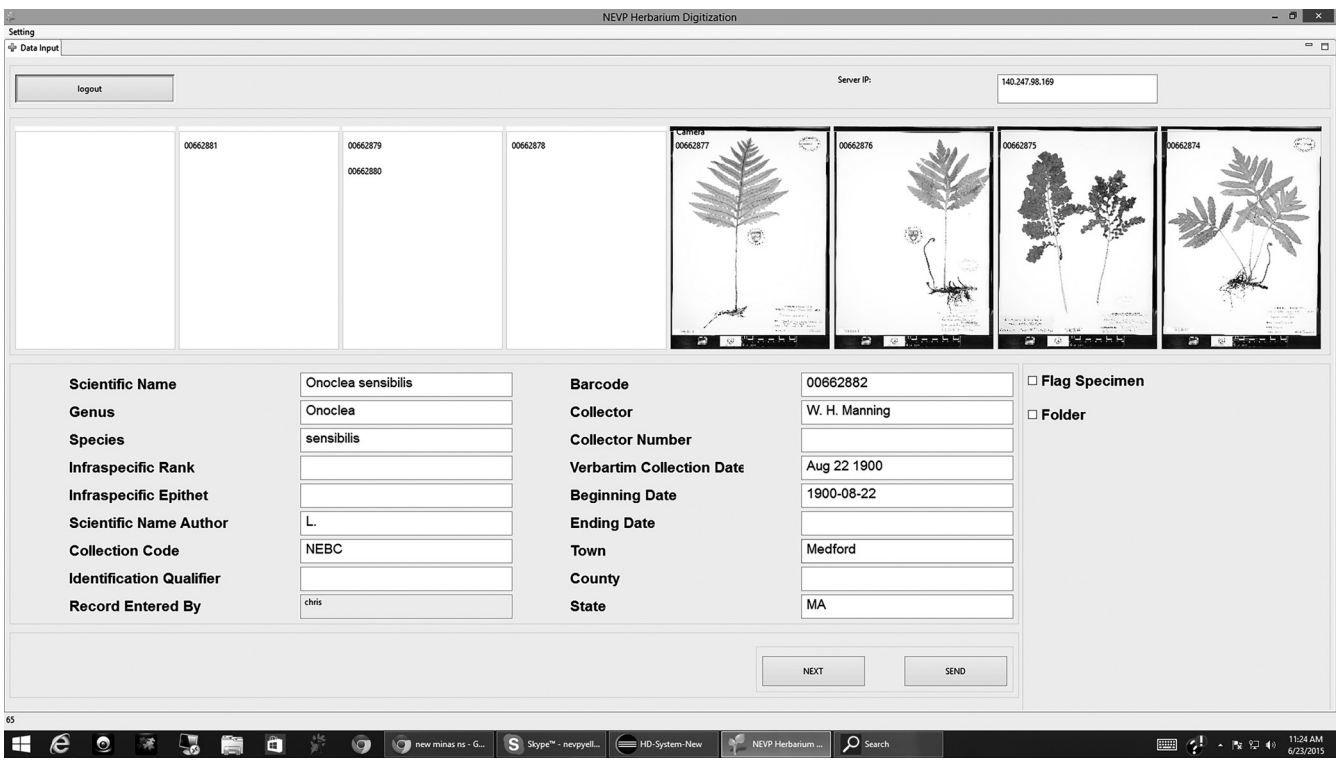

Fig. 5. Data Entry User Interface (DEUI).

images that were useful for a variety of research and curatorial purposes. Images produced by our camera set-up were more than adequate resolution and depth of field for the project's research aims.

The specimen handling and imaging workflow. – Prior to the labeled folders of specimens entering into the conveyor workflow, barcodes were affixed to each specimen sheet; this process was conducted in batch for several folders of specimens at a time. As is standard practice in herbarium digitization workflows, each specimen received an adhesive barcode label whose value was an institutionally unique catalog number (Nelson & al., 2015). This value was used as the catalog number in the HUH database, in the CNH and iDigBio portals, and accompanied all downloaded representations of the data. Images had the barcode value imbedded in their name.

The workflow we designed began with a QR code labeled folder (generated during the pre-capture step) of barcoded specimens. Figure 6 depicts how specimens and data moved through the conveyor system (Phase 2) and Fig. 7 shows images of the system and its components. During operation, the DEUI operator (located at the entry point of the conveyor, Fig. 7B) first scanned the QR code on the folder, which, containing machine readable structured data, populated folder-level data in the DEUI (Fig. 5). Next, the entry operator scanned the catalog number barcode of the first specimen of the folder and entered any required specimen-level information through the DEUI form (Fig. 5). The specimen-level label data that were targeted for capture are presented in Table 1. As a quality control measure, and to increase efficiency, the interface provided look-up lists for collector (pre-populated with names of individual botanists and teams from the HUH botanists database, http://kiki.huh.harvard.edu/databases/botanist_index.html) and for state, county, and town names from the NEVP gazetteer (Sweeney, 2015). When a gazetteer town was not present on the label, finer-level locality data were captured. The interface auto-filled the interpreted date by parsing the entered verbatim date (Fig. 5, Verbatim Collecting Date, and Beginning and Ending Date fields). The data entered into the DEUI were stored in the MySQL database on the Controller computer. After manual data capture, the specimen was placed on the belt. The operator then indicated on the DEUI computer screen to move the conveyor one step forward. This process continued until all specimens within a folder were completed. As the specimen moved down the conveyor belt, the camera (Fig. 7C) captured an image of the specimen and transmitted it to the Controller. Linkages between images and specimen records were stored in the Controller computer database. Some herbarium sheets contained more than one specimen (collection/unit/gathering) per sheet. These cases were handled in the workflow by applying a barcode for each specimen on the sheet and entering data for each specimen, with each specimen record eventually being linked to a separate, duplicate image of the sheet (multiple images of the sheet were automatically captured). During the imaging process when the camera was unable to focus on a specimen due to a lack of material within the autofocus area, a small cardboard disk with a Harvard University logo was placed in the center of the sheet to aid focus. The alignment of the specimen under the camera was achieved by using both hardware and software elements. The belt was divided into 12-inch spaces using raised cleats (Fig.

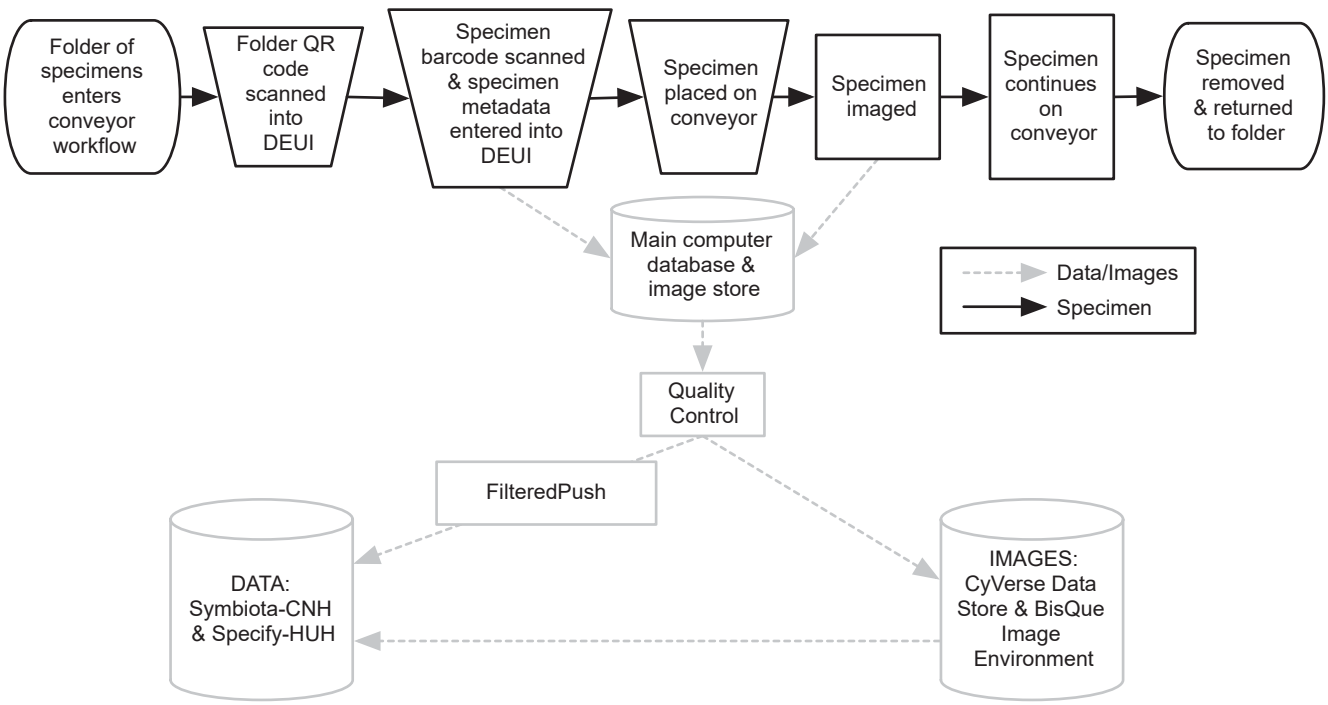

Fig. 6. Overview of Phase 2 of the workflow showing movement of specimens, data, and images.

7A) to ensure that the specimens were properly located within the belt. Along the length of the conveyor belt, an infrared sensor (i.e., the "photo-eye", Fig. 7F) detected movement of the raised cleats on the belt. The conveyor was programmed to move precisely between each cleat. This ensured that the image position did not "drift" away from the field of view of the camera due to position errors of the cleat or the conveyor motor. The operator at the exit portion of the conveyor retrieved specimens from the belt and carefully placed them back within the folder (Fig. 7D). The exit operator was also responsible for randomly checking the Controller computer for the quality of the images processed by both the imaging system and the entry user (Fig. 7E). Using the Controller computer interface, the exit operator also monitored the quality of data input by the DEUI operator and corrected data entry errors by modifying specimen records in the MySQL database; this process was performed without versioning. As images were linked to each specimen record, a thumbnail image of the specimen was sent back to the DEUI system. This allowed the entry operator to gain information about the status of operation and to flag any errors if not discovered by the exit operator. A contrast sensor at the end of the conveyor belt (Fig. 7G) prevented the belt from advancing in the event that a specimen was not discovered at the end of the conveyor line, ensuring that the specimen did not fall off the conveyor if the exit operator was pre-occupied during the retrieval stage. This process was repeated for each folder.

Installation. – After the initial development and testing phase, one conveyor system was installed at the Harvard HUH in July of 2013 and underwent further testing and workflow development until production began during November 2013. The footprint of this conveyor was approximately 5.5 meters long by 0.9 meters wide. A second, smaller conveyor system was installed during May 2014, and production runs using the second system began in June. The footprint of the second conveyor was approximately 4.5 meters long by 0.9 meters wide. The conveyors were installed on opposite walls of a single 148 square meter room.

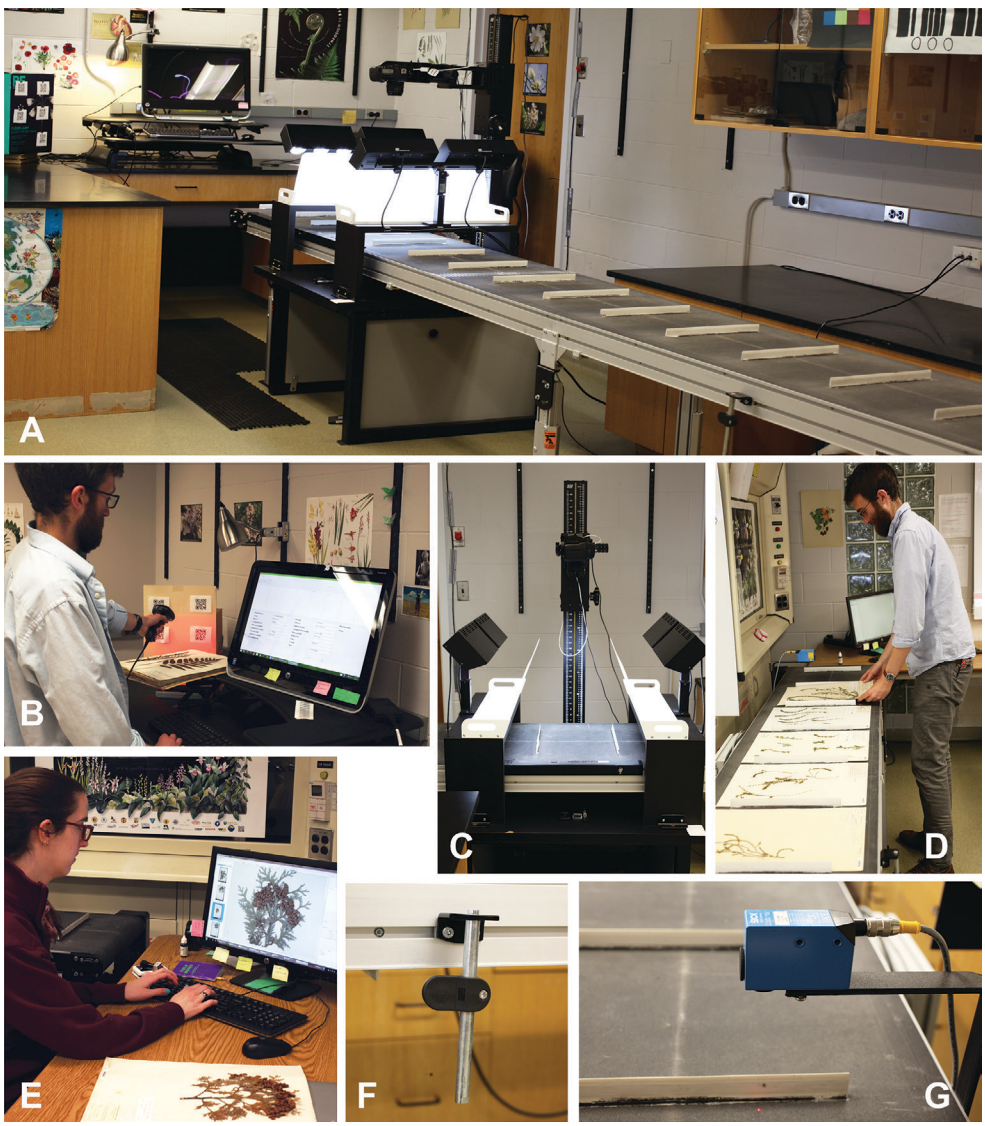

Fig. 7. Images of conveyor system and sub-components. **A,** View of conveyor system; **B,** Data Entry User Interface (DEUI); **C,** Lighting system and camera; **D,** End of conveyor showing manual removal of specimens; **E,** Main (System Controller) computer quality control interface; **F,** Photo-eye to control advance of conveyor belt; **G,** Contrast sensor to detect specimens at end of belt, preventing specimens from falling off of belt.

Phase 3, Data and image storage and transport. — Occurrence records and images were stored locally on the Controller computer of each apparatus as they were created. On a weekly basis following minimal quality control, Adobe DNG (Digital NeGative) images and specimen occurrence data were bundled into folders for processing and export. Before images were transferred from the apparatus computers, Adobe Lightroom (Adobe Systems, San Jose, California, U.S.A.) was used to perform image and metadata adjustments and to generate a JPG for each specimen. The DNG and JPG images were uploaded to the CyVerse (formerly iPlant, http://www.cyverse.org; Goff & al., 2011; Merchant & al., 2016) Data Store, the primary image repository for the NEVP project. CyVerse platforms and tools provided many benefits, including basic digital preservation, on-the-fly image derivative generation and adjustment, parallel data transfer, and configurable user permissions. A set of images was also archived locally at the HUH.

The specimen occurrence data was exported from the apparatus database in the form of RDF transport documents (representing the data as new occurrence annotations using the open annotation ontology [Sanderson & al., 2012, 2016], the oad extension for annotating data [Morris & al., 2013]), and the dwcFP owl representation of Darwin Core (Morris & al., 2013) serialized as RDF/XML using a template (Sweeney, 2013), which were then used to import records into the CNH portal (a Symbiota instance, Gries & al., 2014) and the HUH Specify (http://specifyx.specifysoftware.org) instance. Specify and Symbiota both allow for export of specimen occurrence data according to the Darwin Core data exchange standard (e.g., as Darwin Core Archives, Remsen & al., 2017).

As the specimen records were serialized they were each assigned a globally unique identifier (a version 4 UUID), which was represented in the annotations as the Darwin Core occurrenceID. The UUIDs were generated using the uuid4() function of the python uuid module. These occurrenceID values were present in all downstream representations of the data (e.g., within the HUH Specify database and iDigBio and CNH portals). Thus, each specimen record received an institutionally unique catalog number (the barcode value, see section above titled "The specimen handling and imaging workflow") and a globally unique occurrenceID. Both of these numbers reside in the HUH database and accompanied records when they were shared with aggregators (e.g., the CNH and iDigBio portals). Both of these values can be used to retrieve specimen data. The globally unique property and format of the occurrenceID value has the additional benefit of providing an unambiguous way of referring to the specimen and specimen record, which is essential for linking similar records and related assets across the internet (e.g., in the Semantic Web, Berners-Lee & al., 2001; Berners-Lee, 2006). For a more in depth discussion of the use of globally unique identifiers within biological collections see Guralnick & al. (2015) and Nelson & al. (in press).

Ingest of the RDF/XML documents into Symbiota and Specify was handled, in part, by code developed as part of the FilteredPush Project (Wang & al., 2009). Images of specimens (via links to CyVerse) were disseminated through the CNH portal and the HUH website. After ingest of the new occurrence annotation documents into HUH Specify, the images were analyzed with software designed to read barcodes in the image, verify that the image is linked to the correct specimen record, and to record assertions in the database of which specimens are found on the same herbarium sheet. This software used the zxing barcode reading library and took advantage of the CyVerse environment's BisQue application's ability to adjust image levels on-the-fly to make barcodes more readily detectable if one is not found on a first check.

Phase 4, Enhancement. — Occurrence records were augmented with additional data outside of the conveyor apparatus workflow. Within the CNH portal database, town-level geographic coordinates were applied in batch using backend database queries and a New England town gazetteer developed for the project (Sweeney, 2015). Reproductive phenology was scored for a subset of specimens using the CURIO platform (Willis & al., 2017) and through a trait scoring module available within the CNH portal (this latter task is in progress, see Yost & al., in press).

■ RESULTS

Storage container data capture. — A count of the number of specimens per folder was conducted for a random sample of 67 folders. The average number of specimens per folder was 19.8 (standard deviation = 10.4). Extrapolating to the entire set of digitized specimens suggests that approximately 20,000 QR code labels were generated. Additionally, we estimate that state was captured for about 90% of the folders, introducing additional throughput gains beyond those achieved by capturing scientific name at the folder level.

Specimen-level data and image capture. — The primary development phase of the conveyor system took place between July 2012 and June 2013, with additional development and testing taking place until November 2013. The conveyor apparatus software, database schema, configuration files, hardware list, and conveyor schematics are available for download on GitHub (https://github.com/psweeney-YU/NEVP-conveyor).

Table 2 shows the results of the preliminary time study tests. The Data Entry User Interface tests demonstrated that typing the verbatim date took less time than trying to manipulate the date controls on the user forms. Full data entry (with verbatim date only) took 25 seconds while just scanning the codes (i.e., no specimen-level data recorded) required 8 seconds. The Conveyor length efficiency tests showed that increasing conveyor length resulted in increased efficiency. A test that simulated production usage on a 3-meter conveyor belt with full data entry demonstrated that the average time of digitization was about 39 seconds. The hourly throughput rate of the test system with full data entry was estimated to be approximately 92 specimens per hour, with the bottleneck of the system being the user input process.

Digitization took place from November 2013 to May 2016. Both conveyors were used concurrently unless there was a technical issue with an apparatus or a staff absence. During production, each conveyor was operated by at least two individuals.

One individual was stationed at the DEUI computer (Figs. 5, 7B) and was responsible for scanning the QR code labels on folders, scanning specimen barcodes, capturing specimen-level label data, placing sheets on the conveyor for imaging, and controlling advancement of the conveyor. The second person was stationed at the end of the conveyor at the Main computer (System Controller) interface and was responsible for unloading specimens from the conveyor and performing quality control of data and images via a data review and editing interface that was part of the Controller (Fig. 7E). Additional quality control of records was conducted by interacting directly with the apparatus database. At times, a fifth person (a "floater") was available to assist the individuals operating the conveyor apparatuses, apply barcodes, and gather and return specimens from/to the collection (Fig. 7D).

During the production period of the conveyor, approximately 350,000 specimens were imaged and transcribed over a period of 131 work weeks (Fig. 8A). About 20,000 of these specimens already had occurrence data captured during previous projects. Excluding holiday and vacation periods, the maximum number of specimens digitized in a month was 15,545 and the minimum was 2715 (Fig. 8B). The average was 10,621. The weekly minimum and maximum were 1 and 7777, respectively, with an average of 2513 (Fig. 8C). The weekly and monthly variation in digitization rate had several causes, including variation in personnel availability and technical issues with the operation of the system (see below). Most records had scientific name captured (99.7% with current identification at rank of species and below). Collecting event date was captured least, with 96.9% of the records having verbatim and/or ISO interpreted date. 98% had town-level (or equivalent) data and 99.5% had data captured for collector. A total of 13.08 TB of images were generated (4.10 TB of JPGs and 8.98 TB of DNGs). The average JPG size was 12.09 MB and the average DNG size was 26.48 MB. Image dimensions were 3744 by 5616 pixels; the imaged area averaged about 30.4 cm by 45.6 cm. Records and images are available through the CNH portal and the HUH website. An example of a full resolution image can be found at this URL: https://bisque.cyverse.org/image_service/00-RJcQigNjHrRaHKx5ezDpbU.

Throughput. – Overall, the average time between successive images/records was 35.6 seconds (for intervals between images/records of 30 minutes or less, which captures most ongoing operation of the system, but excludes lunch breaks and other long stops of the system). Where the maximum time between a pair of images/records is less than one minute the average time drops to was 26.3 seconds. Figure 9 shows the distribution of times between images/records for times less than one minute. The distribution of times is strongly positively skewed, with 25% of the times below 20 seconds, the median at 25 seconds, 75% of the data below 33 seconds, and a mean of 35.6 seconds (Fig. 9). A different way to calculate throughput is on a per operator basis. The average number of specimens digitized per operator hour was 22.5 (standard deviation = 6.80), on days when both conveyors were in operation. This calculation underestimates the per operator hour conveyor digitization rate. On some days, only four operators were working throughout the day, and some of an operator's time was dedicated to non-conveyor digitization tasks such as barcoding and specimen retrieval and return.

In summary, we could describe this system's throughput examining the lag time between images/records as having a maximum throughput rate of about 15 seconds per specimen, a median rate of about 25 seconds per specimen, an average rate of 35.6 seconds per specimen, or, examining overall throughput with respect to operator effort as 22.5 specimens per operator hour, or 158 specimens per operator 7-hour day, or 788 specimens per operator 5-day week.

■ DISCUSSION

Storage container data capture. — Our "pre-capture" approach exploits the typical organization of natural science collections by taxonomy using the well understood principle that repeated information can simply be carried forward (e.g., Sarasan, 1978), but by capturing information about storage

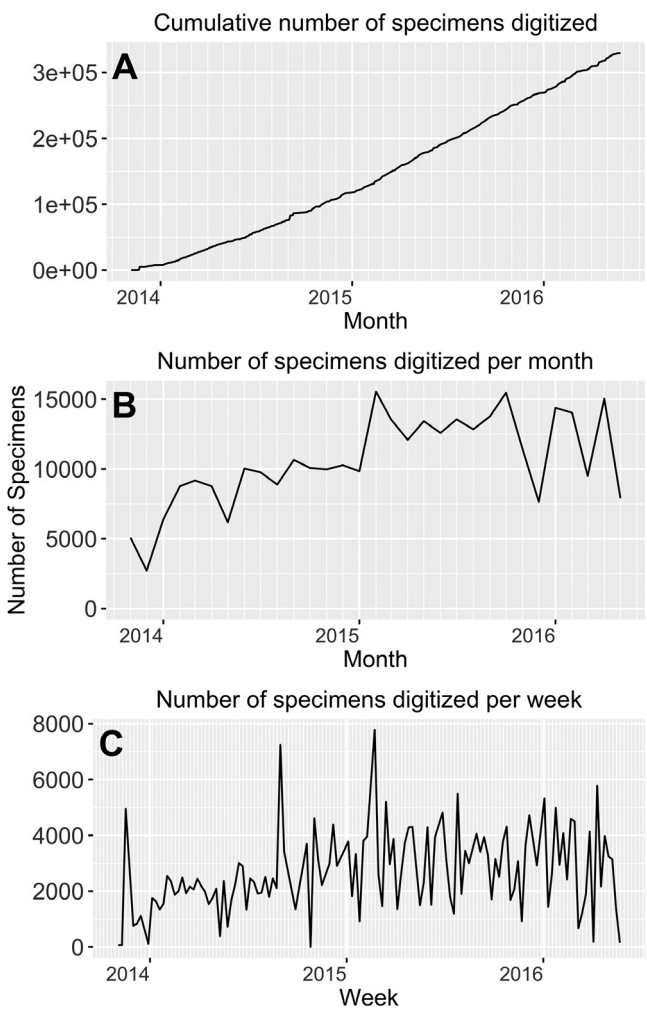

Fig. 8. Digitization (specimen-level label transcription & imaging) progress and throughput over time (two apparatuses). **A,** Cumulative number of specimens digitized; **B,** Monthly throughput rate; **C,** Weekly throughput rate.

containers before examining the individual specimens within them, thus minimizing data entry efforts. Capturing data that reflects the physical storage structure of the collection (i.e., storage-level data and metadata) before imaging and databasing individual specimens, and then associating these data with images and specimen-level data can be a very efficient approach to capturing a subset of the specimen data. This innovation offers considerable throughput gains because certain data (e.g., folder- or cabinet-level data) is entered only once for a batch of specimens (rather than individually) with the data being automatically associated with each specimen in the batch. Indeed, in our workflow the capture of folder-level data provided considerable efficiencies – scientific name, for example, was entered roughly 330,000 fewer times than it would have been with a specimen-level data capture approach.

The capture of container-level data need not be conducted separate from the capture of specimen-level data and imaging. For example, for each folder, container-level data could be captured as part of the specimen-level data transcription and imaging capture workflow, and this would lessen the number of times each folder was handled (one vs. two). However, for our project, conducting the process of container-level data capture as a separate phase from the specimen-level data transcription and imaging phase was not only more efficient, but also allowed for the digitization process to begin while the conveyor system was being developed and tested.

While the primary purpose of the QR codes labels was to facilitate digitization during the course of the current project, it is conceivable that the QR code labels could have potential curatorial uses beyond the digitization process (e.g., quick cabinet inventories, folder relocation, etc.). Future use of the labels would require that they be updated when specimen data changes (e.g., scientific name changes).

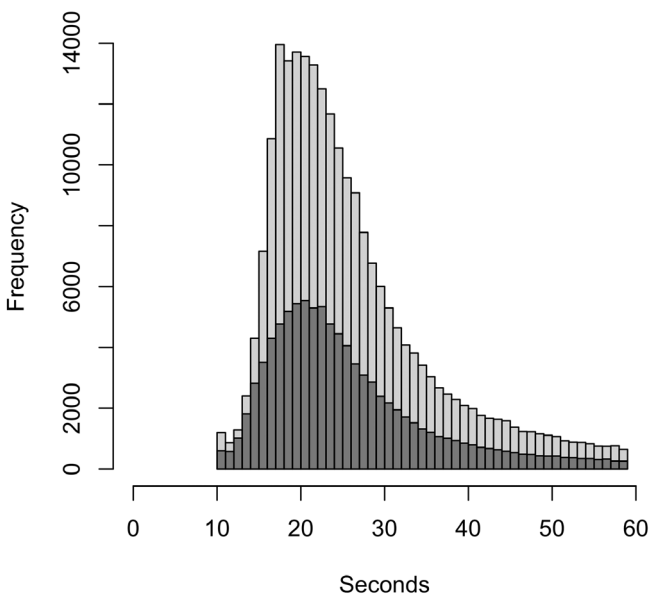

Fig. 9. Distribution of the times between one image and the next on the primary digitization apparatus, system 1 in light grey, system 2 in dark grey, for times less than 60 seconds between successive images. Median throughput is about 25 seconds per specimen image.

Specimen-level data and image capture. — *Throughput.* – The average time between successive images/records (35.6 seconds) was in line with our pre-project expectations for our approach. The distribution of times is strongly positively skewed (Fig. 9). An interpretation of this is that the throughput rate when the conveyors are fully staffed and everything is going smoothly is at about 24 to 25 seconds per specimen (with a maximum rate of about 15 seconds per specimen, and a lower limit imposed by the apparatus of 10 seconds per specimen), but the overall throughput rate was decreased when staff were absent (e.g., no "floater" staff person) or when problems (in data capture, with specimen conservation issues, or with the operation of the system) were encountered. Cumulatively, the long tail of slower (problem) cases accounted for 20%–30% of the time.

It is worthwhile mentioning here that the overall, realized digitization rates for our project are lower than the raw throughput rate of the apparatus. Other parts of the workflow decrease the overall digitization rate; these include the pre-capture process, barcoding specimens, moving specimens between the collection area and digitation area, conducting quality control, processing and exporting images, etc. In addition, occasional technical hardware and software issues interrupted the use of our systems (see "Challenges and improvements" below), and these also lowered our realized throughput rates.

Comparison to other approaches. – Other projects have taken an industrial hardware and software automation system oriented approach to aid in the digitization of herbarium specimens (e.g., Heerlien & al., 2015; Tegelberg & al., 2014). It is difficult to make direct throughput comparisons between our approach and these other conveyor-based approaches. The other approaches were primarily focused on capturing an image with minimal specimen-level data captured, while our approach yielded a significant amount of specimen-level data in addition to an image. There are also differences in the number of conveyors, hours of work per day, and numbers of personnel. The Tegelberg & al. (2014) approach achieved average daily throughput rates of around 700 specimens using one imaging line and between two and five personnel. The Heerlien & al. (2015) approach achieved daily rates between 22,000 and 24,000 specimens using three conveyors and 12 personnel. Of these two approaches, the Tegelberg & al. (2014) approach is most similar to ours in terms of number personnel and specimen data captured, and differed by using only one conveyor versus the two used by this project. Our approach achieved a comparable, albeit lower, average daily rate (ca. 500 vs. ca. 700 specimens per day); however, the Tegelberg & al. (2014) approach captured a subset of the data (i.e., scientific name, collection number, and broad geographical region) that was captured by our project. More typical herbarium imaging workflows that image specimens without using an automated approach are diverse (e.g., Nelson & al., 2015) and there are few published reports of throughput rates. Thiers & al. (2016) and Nelson & al. (2012) report rates of around 100 *images* (does not include label data capture) per hour (or ca. 36 seconds per specimen) per imaging station.

To provide a way to evaluate our system against these less data-capture intensive approaches, we conducted a set of test

"image-only" runs on a single apparatus where only QR data, a barcode number and an image were captured. We conducted a complementary set of runs of the same specimens capturing the same information as in the imaging-only runs plus specimen-level data (i.e., our core specimen-level data approach). The average throughput rate per specimen using the imaging-only approach in this test was about 20 seconds, versus 28 seconds when core specimen-level data was also captured. The rates for both the imaging-only approach (ca. 180 specimens an hour) and the core-data approach (ca. 129 specimens an hour) exceeds those reported from non-conveyor, imaging-only approaches. It should be mentioned here that our system was designed with the expectation that a subset of specimen-level data would be captured for each specimen just prior to it being placed on the conveyor for imaging. Thus, as specimens were imaged, some automated tasks were performed that placed an upper boundary on the imaging rate beyond the camera shutter speed (e.g., image derivative generation, transfer of images and data between camera, DEUI and Controller computers, etc.). A system explicitly engineered for a more object-to-image-to-data approach could achieve faster "imaging-only" rates.

Thus far we have been comparing our system throughput primarily to imaging rates reported by other approaches. In addition to an image, our ca. 35 seconds per specimen rate includes capture of a subset of specimen-level data (Table 1). What can we say about transcription rates? Comparable, non-conveyor object-to-image-to-data herbarium label transcription workflow rates achieved at Yale during the course of the NEVP project average about 80 seconds per specimen. Adding the 36 second per specimen imaging rate reported from the literature to the non-conveyor transcription rate, yields an overall non-conveyor-based digitization rate of about 116 seconds per specimen, about 3 times more than the average rate achieved by our conveyor apparatus.

Challenges and improvements. – While the conveyor systems were generally robust, various software and hardware issues arose during routine operation. Minor issues could often be resolved by the digitizing, curatorial, or informatics personnel at Harvard and Yale; however, more serious issues required assistance from the engineering teams at NCSU and OU, who were committed to provide support for the duration of the project. Our experiences suggest that use of our system will require that adequate informatics and engineering support is available (and "on call") to ensure uninterrupted production use of the system.

We recognize several ways in which our system could be improved. Interviews of the digitization apparatus operators indicated that, subjectively, a large portion of the long tail of longer times between successive images consisted of specimens with problematic data (e.g., difficult to read collector and/or locality data, especially). A modification of the workflow process that allows core data capture at the point of imaging, but flags and splits of problematic specimens for only container-level data capture and imaging, with subsequent data transcription from the images might achieve a 20% to 30% higher throughput rate. Another way that throughput might be increased is to manually focus the camera each session instead of using autofocus, which led to very brief interruptions in the workflow when the camera was unable to focus.

Very minor increases in handling rate per-specimen or reducing or eliminating repetitive tasks, when scaled to many specimens can produce significant overall improvements and cost savings. The total cost of a digitization project is dependent, in part, on the time it takes to digitize each specimen. When the number of specimens is large, small reductions in time can lead to a large overall cost reduction because the reduction is multiplied by the number of specimens being digitized. For example, if there are 1,000,000 specimens and time is reduced by 1 second per specimen, the total time is reduced by ~278 hours. If time is valued by $10 per hour, the cost savings of decreasing the time by 1 second is $2,777. A 10 second decrease will save $27,778.00.

One potential minor gain might be to put the container-level QR code barcodes into the imaging flow and associate the container-level data with the specimens by processing the images instead of having the apparatus operators scan the container-level barcodes by hand (that is, to image the folders instead of hand scanning their barcodes), as is done in standard DataShot workflow (Morris & al., 2010a, b; Morris, 2011, 2013). Given that this change would be per-folder (thus 20,000 repetitions in this project, instead of 350,000 per-specimen repetitions), and the folders would have to be imaged (probably at about 10–12 seconds each), this might not result in an improvement in throughput. We advocate such thinking, that is, analysis of the actual system in production, and the ability to provide an agile software development approach that actively observes the system in operation and tweaks the system to make small changes that improve the throughput.

For our project, we captured a subset of specimen-level data from specimen labels just prior to them being loaded on the conveyor to be imaged (more of an object-to-data-to-image approach). Our motivation for taking this approach was to capture, as soon as possible, specimen-level data that was of high-value to fulfilling project goals, rather than waiting until later in the project when time would be more at a premium. However, one potential way to improve throughput is to more fully separate the specimen-level data capture process from the image capture and storage container-level data capture processes (i.e., adopt a more object-to-image-to-data workflow). This would move the specimen level data transcription effort elsewhere, so that that effort could be parallelized independently of the specimen handling, and thus generate higher conveyor throughput rates. However, specimen-level data would still need to be captured at a later stage, possibly at a significantly greater effort than the 10-15 seconds per specimens that appears typical in our workflow (for example, non-automated NEVP project transcription rates average over 80 seconds a specimen). Many platforms already provide specimen data transcription from images functionality, including Symbiota (Gries & al., 2014), Specify (http://specifyx.specifysoftware.org), and the DataShot desktop and web applications from which the pre-capture application was taken (Morris & al., 2010a, b; Morris, 2011, 2013). This approach would also allow for subsequent transcription and data augmentation to be outsourced using a crowdsourcing

(e.g., Notes from Nature – Hill & al., 2012; Law & al., 2013; Willis & al., accepted) or community sourcing (Morris & al., 2010a, b) approach or to third-parties that leverage less expensive labor markets (e.g., http://alembo.nl/transcriptie).

In general, hardware, software, and methods are constantly advancing and improving. Incorporating such improvements and new innovations into our system would undoubtedly lead to further throughput gains and other improvements. For example, data transmission speeds could be increased by replacing USB 2.0 cables with USB 3.0. Optical Character Recognition (OCR) approaches are in a constant state of advancement and at some point in the future could be incorporated into the workflow to reduce manual data entry. Data read and write speeds could be increased by the use of solid state hard drives. Automated quality control workflows are being developed by the biodiversity informatics community and could speed-up quality control steps and improve the overall quality of the data. The use of DSLRs with greater resolution will increase the image resolution and the possible uses of the images. Crowdsourcing approaches for data capture are becoming more commonplace (e.g., Willis & al., 2017) and have the potential to lessen labor costs.

■ CONCLUSIONS

We have shown that it is possible to use a container-level pre-capture of current identification and an object-to-image-to-data workflow that involves an imaging step running at about 15 seconds per specimen to obtain an image and current identification of each specimen on each herbarium sheet. At an added time of about 10 to 15 seconds per specimen, locality down to municipality, date collected, collector name, and collector number can be transcribed from a very large proportion of sheets. In less than 30 seconds per specimen, an image and the core science data of "what taxon occurred where and when" can be captured for a very large number of sheets. Some sheets, however, are problematic, and capturing these core data for problem cases increases the average time per specimen.

The experiences, workflows, and high-throughput digitization infrastructure resulting from this project will be invaluable in helping to meet the grand challenge of digitizing the world's natural science museum collections and has resulted in the digitization and mobilization of nearly 350,000 specimens. Our work demonstrates that automated, conveyor approaches provide throughput rate improvements over non-automated approaches. Our project has demonstrated that it is possible to capture both an image of a specimen and a core database record in 35 seconds per herbarium sheet (with additional overhead for container-level data capture). The workflows and infrastructure developed as part of this project are available for refinement and continued use to digitize even more specimens in the region. In addition, this project was conducted under the umbrella of the CNH, helping this organization to fulfill its goals and providing a framework for the continued use of the data. Fulfillment of the project objectives improves access to the target collections, and thus will provide benefits to the biological sciences, conservation and land management efforts, as well as the general public.

■ AUTHOR CONTRIBUTIONS

PWS and PJM conceived the overall project. PJM designed and developed the pre-capture software and workflow, with PWS providing guidance on the application of the latter to herbarium collections. BS, YX, AJ, and SR developed, tested, and/or installed the conveyor software and hardware systems, and PWS, PJM, BS, YX, AJ, and SR designed and developed the conveyor workflow. PWS, PJM, and CCD conducted additional post-installation testing and troubleshooting of the conveyor system and workflow. PWS and PJM designed and developed software to transfer data from the conveyor system. PWS, CJG, and PJM compiled and analyzed throughput data. PWS wrote the paper with significant comments and editing from PJM, BS, CJG, and CCD. — PS, https://orcid.org/0000-0003-1239-189X; BS, https://orcid.org/0000-0002-8527-1269; SR, https://orcid.org/0000-0002-0327-8134

■ ACKNOWLEDGEMENTS

We would like to thank the digitization and curatorial staff at Harvard University Herbaria, in particular: Rebecca Bernardos, Lian Bruno, Simone Cappellari, Anne Marie Countie, Claire Hopkins, Michaela Schmull, Chris Schorn, and Ella Weber. In addition, we thank Robert A. Morris for assistance designing the new occurrence annotation documents; the Biota of North American Program (BONAP) for providing a standardized set of scientific names that was used in the pre-capture application. This work was funded by the ADBC program of the U.S. National Science Foundation (Awards 1208835 and 1209149).

■ LITERATURE CITED

Alberch, P. 1993. Museums, collections and biodiversity inventories. *Trends Ecol. Evol.* 8: 372–375. https://doi.org/10.1016/0169-5347(93)90222-B

Ariño, A.H. 2010. Approaches to estimating the universe of natural history collections data. *Biodivers. Inf.* 7: 81–92. https://doi.org/10.17161/bi.v7i2.3991

Baird, R.C. 2010. Leveraging the fullest potential of scientific collections through digitization. *Biodivers. Inf.* 7: 130–136. https://doi.org/10.17161/bi.v7i2.3987

Balke, M., Schmidt, S., Hausmann, A., Toussaint, E.F., Bergsten, J., Buffington, M., Häuser, C.L., Kroupa, A., Hagedorn, G., Riedel, A., Polaszek, A., Ubaidillah, R., Krogmann, L., Zwick, A., Fikáček, M., Hájek, J., Michat, M.C., Dietrich, C., La Salle, J., Mantle, B., Ng, P.K. & Hobern, D. 2013. Biodiversity into your hands - A call for a virtual global natural history "metacollection". *Frontiers Zool.* 10: 55. https://doi.org/10.1186/1742-9994-10-55

Barber, A., Lafferty, D. & Landrum, L.R. 2013. The SALIX Method: A semi-automated workflow for herbarium specimen digitization. *Taxon* 62: 581–590. https://doi.org/10.12705/623.16

Barkworth, M.E. & Murrell, Z.E. 2012. The US Virtual Herbarium: Working with individual herbaria to build a national resource. *ZooKeys* 209: 55–73. https://doi.org/10.3897/zookeys.209.3205

Beaman, R.S. & 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. https://doi.org/10.3897/zookeys.209.3313

Berendsohn, W.G., Anagnostopoulos, G., Hagedorn, G., Jakupovic, J., Nimis, P.L. & Valdés, B. 1997. A framework for biological information models. *Lagascalia* 19: 667–672.

Berendsohn, W., Chavan, V. & Macklin, J. 2010. Summary of Recommendations of the GBIF Task Group on the Global Strategy

and Action Plan for the digitisation of natural history collections. *Biodivers. Inf.* 7: 67–71. https://doi.org/10.17161/bi.v7i2.3989

Berners-Lee, T. 2006. Linked Data. https://www.w3.org/DesignIssues/LinkedData.html (accessed 22 Jun 2017).

Berners-Lee, T., Hendler, J. & Lassila, O. 2001. The Semantic Web. *Sci. Amer.* 284: 1–9. https://doi.org/10.1038/scientificamerican0501-34

Bertone, M., Blinn, R., Stanfield, T., Dew, K., Seltmann, K. & Deans, A. 2012. Results and insights from the NCSU Insect Museum GigaPan project. *ZooKeys* 209: 115–132. https://doi.org/10.3897/zookeys.209.3083

Blagoderov, V., Kitching, I.J., Livermore, L., Simonsen, T.J. & Smith, V.S. 2012. No specimen left behind: Industrial scale digitization of natural history collections. *ZooKeys* 209: 133–146. https://doi.org/10.3897/zookeys.209.3178

Chavan, V. & Krishnan, S. 2003. Natural history collections: A call for national information infrastructure. *Curr. Sci.* 84: 34–42.

Constable, H., Guralnick, R., Wieczorek, J., Spencer, C., Peterson, A.T. & The VertNet Steering Committee. 2010. VertNet: A new model for biodiversity data sharing. *PLoS Biol.* 8: e1000309. https://doi.org/10.1371/journal.pbio.1000309

Creighton, R.A. & Crockett, J.J. 1971. SELGEM: A system for collection management. *Smithsonian Inst. Inform. Systems Innov.* 2: 1–24.

Croft, J.R. (ed.) 1989. *HISPID – Herbarium information standards and protocols for interchange of data* [Version 1]. Canberra: Australian National Botanic Gardens.

Crovello, T.J. 1967. Problems in the use of electronic data processing in biological collections. *Taxon* 16: 481–494. https://doi.org/10.2307/1216951

Davis, C.C., Willis, C.G., Connolly, B., Kelly, C. & Ellison, A.M. 2015. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *Amer. J. Bot.* 102: 1599–1609. https://doi.org/10.3732/ajb.1500237

Davis Rabosky, A.R., Cox, C.L., Rabosky, D.L., Title, P.O., Holmes, I.A., Feldman, A. & McGuire, J.A. 2016. Coral snakes predict the evolution of mimicry across New World snakes. *Nature, Commun.* 7: 11484. https://doi.org/10.1038/ncomms11484

Dietrich, C., Hart, J., Raila, D., Ravaioli, U., Sobh, N., Sobh, O. & Taylor, C. 2012. InvertNet: A new paradigm for digital access to invertebrate collections. *ZooKeys* 209: 165–81. https://doi.org/10.3897/zookeys.209.3571

Doudna, J.W. & Danielson, B.J. 2015. Rapid morphological change in the masticatory structures of an important ecosystem service provider. *PLoS ONE* 10: e0127218. https://doi.org/10.1371/journal.pone.0127218

Drew, J. 2011. The role of natural history institutions and bioinformatics in conservation biology. *Conservation Biol.* 25: 1250–1252. https://doi.org/10.1111/j.1523-1739.2011.01725.x

Goff, S.A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A.E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., Muir, A., Merchant, N., Lowry, S., Mock, S., Helmke, M., Kubach, A., Narro, M., Hopkins, N., Micklos, D., Hilgert, U., Gonzales, M., Jordan, C., Skidmore, E., Dooley, R., Cazes, J., McLay, R., Lu, Z., Pasternak, S., Koesterke, L., Piel, W.H., Grene, R., Noutsos, C., Gendler, K., Feng, X., Tang, C., Lent, M., Kim, S.-J., Kvilekval, K., Manjunath, B.S., Tannen, V., Stamatakis, A., Sanderson, M., Welch, S.M., Cranston, K.A., Soltis, P., Soltis, D., O'Meara, B., Ane, C., Brutnell, T., Kleibenstein, D.J., White, J.W., Leebens-Mack, J., Donoghue, M.J., Spalding, E.P., Vision, T.J., Myers, C.R., Lowenthal, D., Enquist, B.J., Boyle, B., Akoglu, A., Andrews, G., Ram, S., Ware, D., Stein, L. & Stanzione, D. 2011. The iPlant Collaborative: Cyberinfrastructure for plant biology. *Frontiers Pl. Sci.* 2: 1–16. https://doi.org/10.3389/fpls.2011.00034

Granzow-de la Cerda, Í. & Beach, J.H. 2010. Semi-automated workflows for acquiring specimen data from label images in herbarium collections. *Taxon* 59: 1830–1842. http://www.jstor.org/stable/41059876

Greve, M., Lykke, A.M., Fagg, C.W., Gereau, R.E., Lewis, G.P., Marchant, R., Marshall, A.R., Ndayishimiye, J., Bogaert, J. & Svenning, J.-C. 2016. Realising the potential of herbarium records for conservation biology. *S. African J. Bot.* 105: 317–323. https://doi.org/10.1016/j.sajb.2016.03.017

Gries, C., Gilbert, E.E. & Franz, N.M. 2014. Symbiota – A virtual platform for creating voucher-based biodiversity information communities. *Biodivers. Data J.* 2: e1114. https://doi.org/10.3897/BDJ.2.e1114

Guralnick, R.P., Cellinese, N., Deck, J., Pyle, R.L., Kunze, J., Penev, L., Walls, R., Hagedorn, G., Agosti, D., Wieczorek, J., Catapano, T. & Page, R.D.M. 2015. Community next steps for making globally unique identifiers work for biocollections data. *ZooKeys* 154: 133–154. https://doi.org/10.3897/zookeys.494.9352

Hanken, J.M., McDade, L., Beach, J., Cook, J., Ford, L.S., Gropp, R., Joyce, K. & Thiers, B. 2013. Implementation plan for Network Integrated Biocollections Alliance. http://www.aibs.org/public-policy/biocollections.html (accessed 10 Apr 2017).

Harsch, M.A. & HilleRisLambers, J. 2016. Climate warming and seasonal precipitation change interact to limit species distribution shifts across Western North America. *PLoS ONE* 11: 1–17. https://doi.org/10.1371/journal.pone.0159184

Heerlien, M., Van Leusen, J., Schnörr, S., De Jong-Kole, S., Raes, N. & Van Hulsen, K. 2015. The natural history production line: An industrial approach to the digitization of scientific collections. *J. Comput. Cult. Herit.* 8: 3. https://doi.org/10.1145/2644822

Heidorn, P.B. 2008. Shedding light on the dark data in the long tail of science. *Library Trends* 57: 280–299. https://doi.org/10.1353/lib.0.0036

Hill, A., Guralnick, R., Smith, A., Sallans, A., Gillespie, R., Denslow, M., Gross, J., Murrell, Z., Conyers, T., Oboyski, P., Ball, J., Thomer, A., Prys-Jones, R., de la Torre, J., Kociolek, P. & Fortson, L. 2012. The Notes from Nature tool for unlocking biodiversity records from museum records through citizen science. *ZooKeys* 209: 219–233. https://doi.org/10.3897/zookeys.209.3472

Hudson, L.N., Blagoderov, V., Heaton, A., Holtzhausen, P., Livermore, L., Price, B.W., Van der Walt, S. & Smith, V.S. 2015. Inselect: Automating the digitization of natural history collections. *PLoS ONE* 10: e0143402. https://doi.org/10.1371/journal.pone.0143402

Johnson, K.G., Brooks, S.J., Fenberg, P.B., Glover, A.G., James, K.E., Lister, A.M., Michel, E., Spencer, M., Todd, J.A., Valsami-Jones, E., Young, J.R. & Stewart, J.R. 2011. Climate change and biosphere response: Unlocking the collections vault. *BioScience* 61: 147–153. https://doi.org/10.1525/bio.2011.61.2.10

Lavoie, C. 2013. Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspect. Pl. Ecol. Evol. Syst.* 15: 68–76. https://doi.org/10.1016/j.ppees.2012.10.002

Law, E., Dalton, C., Merrill, N., Young, A. & Gajos, K.Z. 2013. Curio: A platform for supporting mixed-expertise crowdsourcing. Pp. 99–100 in: *Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts.* AAAI. Technical Report CR-13-01. Palo Alto: AAAI. http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7534

Lister, A.M. & Climate Change Research Group. 2011. Natural history collections as sources of long-term datasets. *Trends Ecol. Evol.* 26: 153–154. https://doi.org/10.1016/j.tree.2010.12.009

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D. & Antin, P. 2016. The iPlant Collaborative: Cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol.* 14: e1002342. https://doi.org/10.1371/journal.pbio.1002342

Morris, P.J. 2011 DataShot. Program distributed by the author. https://sourceforge.net/projects/datashot/

Morris, P.J. 2013. DataShot PreCapture Application. Program distributed by the author. https://doi.org/10.5281/zenodo.154176

Morris, P.J., Eastwood, R., Ford, L.S., Haley, B. & Goldman-Huertas,

B. 2010a. Imaging and innovative workflows for efficient data capture in an entomological collection: The MCZ Lepidoptera Rapid Data Capture Project. Pp. 87–88 in: *Biodiversity 2010 and beyond: Science and collections: Program and abstracts*. http://www.spnhc.org/media/assets/SPNHC-CBA2010ProgramandAbstracts_FinalplusErrata.pdf

Morris, P.J., Eastwood, R., Ford, L.S., Haley, B. & Pierce, N.E. 2010b. Innovative Workflows for efficient data capture in an Entomological collection: The MCZ Rhopalocera (Lepidoptera) Rapid Data Capture Project. *Entomological Collections Network Annual Meeting*, Dec. 11 & 12, 2010. http://www.ecnweb.org/pdf/ECN_program2010.pdf

Morris, P.J., Hanken, J., Lowery, D.B., Ludäscher, B., Macklin, J.A., Morris, P.J., Morris, R.A., Song, T. & Sweeney, P. 2014. Capturing inventory level information about collections as a step in object to image to data workflows. *TDWG 2014 Annual Conference. #697 Symposium S05. Access to digitization tools and methods.* https://mbgserv18.mobot.org/ocs/index.php/tdwg/2014/paper/view/697

Morris, R.A., Dou, L., Hanken, J., Kelly, M., Lowery, D.B., Ludaescher, B., Macklin, J.A. & Morris, P.J. 2013. Semantic annotation of mutable data. *PLoS ONE* 8: e76093. https://doi.org/10.1371/journal.pone.0076093

Nelson, G., Paul, D., Riccardi, G. & Mast, A.R. 2012. Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys* 209: 19–45. https://doi.org/10.3897/zookeys.209.3135

Nelson, G., Sweeney, P., Wallace, L.E., Rabeler, R.K., Allard, D., Brown, H., Carter, J.R., Denslow, M.W., Ellwood, E.R., Germain-Aubrey, C.C., Gilbert, E., Gillespie, E., Goertzen, L.R., Legler, B., Marchant, D.B., Marsico, T.D., Morris, A.B., Murrell, Z., Nazaire, M., Neefus, C., Oberreiter, S., Paul, D., Ruhfel, B.R., Sasek, T., Shaw, J., Soltis, P.S., Watson, K., Weeks, A. & Mast, A.R. 2015. Digitization workflows for flat sheets and packets of plants, algae, and fungi. *Applic. Pl. Sci.* 3: 1500065. https://doi.org/10.3732/apps.1500065

Nelson, G., Sweeney, P. & Gilbert, E. In press. Use of Globally Unique Identifiers (GUIDs) to link herbarium specimen records to physical specimens. *Applic. Pl. Sci.* 6.

Peterson, A.T., Cicero, C. & Wieczorek, J. 2006. The ORNIS network: A broad virtual data resource with tools for the ornithological community. *J. Ornithol.* 147: 228–229. https://doi.org/10.1007/s10336-006-0093-1

Powers, K.E., Prather, A.L., Cook, J.A., Woolley, J., Bart, H.L., Monfils, A.K. & Sierwald, P. 2014. Revolutionizing the use of natural history collections in education. *Sci. Edu. Rev.* 13: 24–33. http://files.eric.ed.gov/fulltext/EJ1057153.pdf

Pyke, G.H. & Ehrlich, P.R. 2010. Biological collections and ecological/environmental research: A review, some observations and a look to the future. *Biol. Rev. Cambridge Philos. Soc.* 85: 247–266. https://doi.org/10.1111/j.1469-185X.2009.00098.x

Remsen, D., Braak, K., Döring, M. & Robertson, T. 2017 (released on 9 May 2011). Darwin Core Archives – How-to Guide, version 2.0. http://github.com/gbif/ipt/wiki/DwCAHowToGuide (accessed 1 Feb 2018).

Sanderson, R., Ciccarese, P. & Van de Sompel, H. (eds.) 2012. Open Annotation Core Data Model, Community Draft. W3C Open Annotation Community Group http://www.openannotation.org/spec/core/ (accessed 10 Apr 2017).

Sanderson, R., Ciccarese, P. & Young, B. (eds.) 2016. Web Annotation Data Model: W3C Candidate Recommendation 05 July 2016. W3C. http://www.w3.org/TR/2016/CR-annotation-model-20160705/ (accessed 10 Apr 2017).

Sarasan, L. 1978. AIMS: A rapid, economic approach to museum catalogue computerization and collection inventory. *Council for Museum Anthropology Newsletter* 2: 8–13. https://doi.org/10.1525/mua.1978.2.3.8

Schmidt, S., Balke, M. & Lafogler, S. 2012. DScan – A high-performance digital scanning system for entomological collections. *ZooKeys* 209: 183–191. https://doi.org/10.3897/zookeys.209.3115

Soberon, J. 1999. Linking biodiversity information sources. *Trends Ecol. Evol.* 14: 291. https://doi.org/10.1016/S0169-5347(99)01617-1

Stein, B.R. & Wieczorek, J. 2004. Mammals of the World: MaNIS as an example of data integration in a distributed network environment. *Biodivers. Inf.* 1: 14–22. https://doi.org/10.17161/bi.v1i0.7

Suarez, A.V. & Tsutsui, N.D. 2004. The value of museum collections for research and society. *BioScience* 54: 66–74. https://doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

Sunderland, M.E. 2013. Computerizing natural history collections. *Endeavour* 37: 150–161. https://doi.org/10.1016/j.endeavour.2013.04.001

Sweeney, P. 2013. NEVP-rdf/xml template. Program distributed by the author. https://github.com/psweeney-YU/NEVP-rdfXml

Sweeney, P. 2015. NEVP-gazetteer: First release. https://doi.org/10.5281/zenodo.16401 (accessed 10 Apr 2017)

Tegelberg, R., Mononen, T. & Saarenmaa, H. 2014. High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. *Taxon* 63: 1307–1313. https://doi.org/10.12705/636.13

Tewksbury, J.J., Anderson, J.G.T., Bakker, J.D., Billo, T.J., Dunwiddie, P.W., Groom, M.J., Hampton, S.E., Herman, S.G., Levey, D.J., Machnicki, N.J., del Rio, C.M., Power, M.E., Rowell, K., Salomon, A.K., Stacey, L., Trombulak, S.C. & Wheeler, T.A. 2014. Natural history's place in science and society. *BioScience* 64: 300–310. https://doi.org/10.1093/biosci/biu032

Thiers, B.M., Tulig, M. & Watson, K. 2016. Digitization of The New York Botanical Garden Herbarium. *Brittonia* 68: 324–333. https://doi.org/10.1007/s12228-016-9423-7

Tulig, M., Tarnowsky, N., Bevans, M., Kirchgessner, A. & Thiers, B.M. 2012. Increasing the efficiency of digitization workflows for herbarium specimens. *ZooKeys* 209: 103–113. https://doi.org/10.3897/zookeys.209.3125

Vellend, M., Brown, C.D., Kharouba, H.M., Mccune, J.L. & Myers-Smith, I.H. 2013. Historical ecology: Using unconventional data sources to test for effects of global environmental change. *Amer. J. Bot.* 100: 1294–1305. https://doi.org/10.3732/ajb.1200503

Wang, Z., Dong, H., Kelly, M., Macklin, J.A., Morris, P.J. & Morris, R.A. 2009. Filtered-Push: A map-reduce platform for collaborative taxonomic data management. Pp. 731–735 in: Burgin, M., Chowdhury, M.H., Ham, C.H., Ludwig, S., Su, W. & Yenduri, S. (eds.), *Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering*, 31 March–2 April 2009, Los Angeles, Calif. U.S.A., vol. 7. Los Alamitos: IEEE. https://doi.org/10.1109/CSIE.2009.948

Whitehead, P.J.P. 1971. Storage and retrieval of information in systematic zoology. *Biol. J. Linn. Soc.* 3: 211–220. https://doi.org/10.1111/j.1095-8312.1971.tb00181.x

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. & Vieglais, D. 2012. Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE* 7: e29715. https://doi.org/10.1371/journal.pone.0029715

Willis, C.G., Law, E., Williams, A.C., Franzone, B.F., Bernardos, R., Bruno, L., Hopkins, C., Schorn, C., Weber, E., Park, D.S. & Davis, C.C. 2017. CrowdCurio: An online crowdsourcing platform to facilitate climate change studies using herbarium specimens. *New Phytol.* 215: 479–488. https://doi.org/10.1111/nph.14535

Yost, J.M., Sweeney, P.W., Gilbert, E., Nelson, G., Guralnick, R., Gallinat, A.S., Ellwood, E.R., Rossington, N., Willis, C.G., Blum, S.D., Walls, R.L., Haston, E.M., Denslow, M.W., Zohner, C.M., Morris, A.B., Stucky, B.J., Carter, J.R., Baxter, D.G., Bolmgren, K., Denny, E.G., Dean, E., Davis, C.C., Mishler, B.M., Soltis, P.S. & Mazer, S.M. In press. Digitization protocol for scoring reproductive phenology from herbarium specimens of seed plants. *Applic. Pl. Sci.* 6.