METHODS AND TECHNIQUES

Semi-automated workflows for acquiring specimen data from label images in herbarium collections

Íñigo Granzow-de la Cerda[1,3] & James H. Beach[2]

1 *University of Michigan Herbarium, 3600 Varsity Dr., Ann Arbor, Michigan 48108, U.S.A.*
2 *Biodiversity Institute, University of Kansas, 1345 Jayhawk Boulevard, Lawrence, Kansas 66045, U.S.A.*
3 Current address: *Departament de Biologia Animal, Biologia Vegetal i Ecologia, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain*
Author for correspondence: *Íñigo Granzow-de la Cerda, inyigo.delacerda@uab.cat*

Abstract Computational workflow environments are an active area of computer science and informatics research; they promise to be effective for automating biological information processing for increasing research efficiency and impact. In this project, semi-automated data processing workflows were developed to test the efficiency of computerizing information contained in herbarium plant specimen labels. Our test sample consisted of Mexican and Central American plant specimens held in the University of Michigan Herbarium (MICH). The initial data acquisition process consisted of two parts: (1) the capture of digital images of specimen labels and of full-specimen herbarium sheets, and (2) creation of a minimal field database, or "pre-catalog", of records that contain only information necessary to uniquely identify specimens. For entering "pre-catalog" data, two methods were tested: key-stroking the information (a) from the specimen labels directly, or (b) from digital images of specimen labels. In a second step, locality and latitude/longitude data fields were filled in if the values were present on the labels or images. If values were not available, geo-coordinates were assigned based on further analysis of the descriptive locality information on the label. Time and effort for the various steps were measured and recorded. Our analysis demonstrates a clear efficiency benefit of articulating a biological specimen data acquisition workflow into discrete steps, which in turn could be individually optimized. First, we separated the step of capturing data from the specimen from most keystroke data entry tasks. We did this by capturing a digital image of the specimen for the first step, and also by limiting initial key-stroking of data to create only a minimal "pre-catalog" database for the latter tasks. By doing this, specimen handling logistics were streamlined to minimize staff time and cost. Second, by then obtaining most of the specimen data from the label images, the more intellectually challenging task of label data interpretation could be moved electronically out of the herbarium to the location of more highly trained specialists for greater efficiency and accuracy. This project used experts in the plants' country of origin, Mexico, to verify localities, geography, and to derive geo-coordinates. Third, with careful choice of data fields for the "pre-catalog" database, specimen image files linked to the minimal tracking records could be sorted by collector and date of collection to minimize key-stroking of redundant data in a continuous series of labels, resulting in improved data entry efficiency and data quality.

Keywords biological specimens; Central America; data acquisition; databases; digital images; digitization; geo-referencing; herbarium collections; Mexico; scientific workflows

■ INTRODUCTION

Biological collections have acquired new research relevance as the software tools and services of biodiversity informatics are providing unprecedented capabilities for discovery, retrieval, integration, and synthesis of information associated with specimen vouchers (Ertter, 2000; Graham & al., 2004; Soberón & Peterson, 2004). Federated database architectures and caches of collection data exemplified by GBIF, REMIB, MaNIS, HerpNet, FishNet, OBIS, ORNIS, and other projects, have become well established and have transformed accessibility of biological collection-based information (MacRander & Haynes, 1990; Edwards & al., 2000; Kirkup & al., 2005). Because of increasing network availability of species occurrence data and the software applications that use them, a larger community of researchers beyond taxonomy has been using this information to address broader questions in ecology, conservation, and environmental change (for reviews see: Guisan & Thuiller, 2005; Elith & Leathwick, 2009). Although herbaria are proceeding with database efforts on a substantial scale, the amount of data yet to be digitized is large. More than 60 million specimens are thought to exist in U.S. herbaria, and an estimated 350 million worldwide (Thiers, 2010). New networked and computationally driven applications in biodiversity informatics have identified the primary bottleneck preventing complete utilization of collection information: only a small fraction of specimens from collections around the world have been databased. For example, data as important as geo-coordinates are available for ca. 39 million plant specimens (Catalogue of Life, 2007, as of July 2010). Making biological collections data digitally available on the Internet is still comparatively slow and costly, and the largest collections institutions face daunting backlogs of historically important and geographically diverse holdings that need to be computerized.

More efficient methods to acquire data are needed if botanical collections hope to mobilize information stored in their repositories for applications not only to systematics research but also to broader global environmental issues of importance to science and society.

Networked software architectures and applications that recognize the inherent properties of botanical collections, such as the ubiquity of duplicate specimens across institutions and the non-random distribution of specimens of some plant groups skewed toward herbaria with staff specialists, would be a good place to implement the use of cybertools for accelerating data entry, mobilization of the data, and minimization of associated cost. Given funding constraints, priorities must be established for collection data capture and marshalling that information to the Internet.

The design and function of automated workflow environments is an active area of investigation in computer science. The application of workflow tools to acquisition, management, and analysis of biological research data will enable powerful new modes of knowledge discovery (Singh & Vouk, 1996; Berry, 1998; Cavalcanti & al., 2005; Greenwood, 2005; McPhillips & Bowers, 2005; Shankar & al., 2005; Versteeg & al., 2006). The present effort evaluates options for semi-automating and optimizing specimen data entry workflows in the computerization of botanical collections by assigning resources and refining methods in discretely organized steps. This analysis will inform the design of software and workflows for the automation of botanical specimen data acquisition on a larger scale.

The role of images in collection computerization. — The ease of use, availability and low cost of digital photography has created opportunities for new imaging methods and research applications. Biological collection institutions have made digital images of important museum and herbarium specimens available on the web. Not only have herbaria been offering high-resolution images of their holdings through their portals since at least the 1990s, but online images have also been used in other venues, such as regional digital flora projects (MacRander & Haynes, 1990; Wolf & Holland, 2000; Davies & al., 2002; Schaub & Dunn, 2002; Pankhurst, 2004; Schmull & al., 2005). There are ongoing investments being made to scan most botanical type collections and put them online (e.g., Aluka: API and LAPI; http://plants.jstor.org/, http://aluka.ithaka.org/plants/index.html; Morphbank, http://www.morphbank.net; Smith, 2004; Guthrie & Nygren, 2007). Several institutions have been creating database records from images of specimen labels. One prominent example is CONABIO's *Sistema Nacional de Información sobre Biodiversidad de México* (SNIB), where images of Mexican plant specimens held in herbaria outside of Maxico are remitted over the Internet to the institution's facility in Mexico City. Database records are created by reading the label images (see http://www.conabio.gob.mx/institucion/snib/doctos/acerca.html). Some systems have been developed to automate data acquisition from label images by means of optical character recognition (Lafferty & Landrum, 2009) and a combination of natural handwriting recognition and natural language processing (Beaman & al., 2006; Heidorn & Wei, 2008; http://www.herbis.org/index.php; http://www.yale.edu/peabody/collections/bot/botcurr_db.html). Although these technologies have great potential, the role they will play in efficient herbarium specimen data acquisition is not yet clear.

This project, carried out at the University of Michigan Herbarium (MICH), was focused on the use of workflows and specimen label images for creating specimen database records including latitude and longitude values (geo-coordinates) for terrestrial plant specimens from Mexico and Mesoamerica. Mexico ranks 4th worldwide in overall plant species richness. It contains up to 11% of the word's seed plant diversity, which includes an estimated 21,300–24,600 species of angiosperms (Espejo-Serna & al., 2004), of which 50%–60% are endemics (CONABIO, 2006; Sarukhán & al., 2009). Ranked after the U.S. National Museum of Natural History, the MICH herbarium probably holds the world's second largest collection of Mexican vascular plants outside of Mexico (Rzedowski, 1976). MICH has an estimated 260,000 Mexican and Mesoamerican specimens including those from H.H. Bartlett, C.L. Lundell, R. McVaugh, W.R. Anderson, and their students' collections, as well as important historical collections by J.J. Linden and C.G. Pringle. Part of the volume and value of this material is also the result of recent research activity like *Flora Novo-Galiciana* by R. McVaugh (1983–2001) and the *Moss Flora of Mexico* by Sharp & al. (1994).

Two strategies for specimen data acquisition. — Given the low cost of data storage, computer workstations, and networked communications (assuming appropriate software is available for the task), the limiting cost for collections computerization is the human labor needed to capture data from specimen labels into a structured database (Bart, 2005). Minimizing the cost of labor must be a high priority in any computerization project seeking to maximize output of specimen records. Historically, retrospective collection computerization projects have chosen methods that fall at some point along a gradient between two contrasting strategies:

Strategy 1: No data left behind. – This data-intensive method prioritizes the completeness of database records and aims to capture as much scientific data as are available from a specimen label for as many specimens as possible until project resources are spent. Essentially no label information is left un-captured. Because the kind, amount, and format of information on labels can vary considerably from specimen to specimen, capturing all information for thousands or tens of thousands of specimens requires that the set of data fields be exhaustive and that the variation in label data semantics and syntax be identified and accommodated. Data-entry staff must be well-trained to interpret, encode, and parse data that can have various semantics and syntax (e.g., differing formats of date and time or collectors' names), can be incomplete (e.g., locality descriptions or dates), or can have multiple values (e.g., determinations, qualifiers, or comments). The rationale for capturing all data is the assurance that specimens will never need to be handled a second time for the purposes of label data capture. Physically extracting, moving, and manipulating specimens are substantial contributors to the cost of digitization of specimens.

Strategy 2: Capture minimal data. – In this scenario, highest priority is given to computerizing the largest number of specimens within the project's budget, so that the resulting database represents the largest sample of the underlying collection—typically emphasizing some aspect of the collection such as the broadest diversity of taxa or widest geographical representation. This approach minimizes the initial cost of specimen data entry by reducing the amount of information captured from each label to an essential minimum. Identifying *a priori* which data can be ignored and which are needed for research use is one intellectual challenge of this strategy. Recording too much data per specimen reduces efficiency; capturing too little produces incomplete data records with limited research utility. Priority usually is given to the label data elements that would be most valuable to users. For taxonomic use, a minimal database record might include all taxonomic determinations, type status, collector name and number, date, and locality information to the nearest town or named place. For biogeographers, macroecologists, and environmental niche modelers who employ species occurrence data for analyzing species ranges or for species distribution modeling, the most recent determination, geo-coordinates, an ID number, and collection date would be sufficient. For ecologists, habitat information and the names of associated taxa, pollinators, or predators might be the most informative data.

The conundrum, given the varied research uses of specimen data, is that the ultimate research requirements that inform which label data to prioritize for capture may not be completely knowable in advance (Greenberg & al., 2006). But with finite time and financial resources for collections computerization, a data acquisition strategy must be efficient and the resulting database useful. A minimum data entry approach that maximizes number of database records generated does so at the expense of reduced completeness for individual records. The risk that insufficient information may be recorded at the time of data capture through this approach is minimized if the cost of re-accessing all information present on a specimen label can be driven toward zero. Capturing high-resolution digital images of labels at the time of minimal data entry is one way to accomplish that.

We assert that workflows for data capture of botanical specimens that are based on the acquisition of digital label images, optimize database development capacity, minimize the risk of not meeting research data requirements, and reduce cost. Herbaria universally file specimens based on taxonomic identity. With specimens shelved, sorted and grouped by species name, the kinds of label data and format of one specimen bear little in common with that of the subsequent specimen in the stack (other than taxon and perhaps geographical region). Although the semantics of plant specimen data concepts are fairly well standardized, the choice of content and label layout styles vary from collector to collector, and even more so over the 200+ year span of biological collecting history. Exhaustively populating all fields in a specimen database record from a single, sequential pass through specimens in a collection where labels of consecutive specimens may greatly differ in layout and data content is bound to be inefficient. We will argue that the effect of heterogeneity in the structure and content of the source material on efficiency can be minimized if prior to complete label data entry records are first sorted by collector name, then secondarily by collection date or collector number.

Retrospectively, georeferencing specimen localities is time-consuming and labor-intensive. As with the choice of which specimen data to computerize, deciding the level of precision and accuracy needed for georeferencing specimens depends on the specific requirements of future analyses. For example, studies of climate change or niche modeling on continental or global scales may only require resolution of localities to one or a few degrees, whereas conservation or studies at a habitat scale may require localities resolved to a scale of tens of meters (Wieczorek, 2001; GBIF, 2002; Wieczorek & al., 2004). Generally, higher levels of resolution and accuracy are better. Several recent research analyses have employed experts and students with extensive knowledge of a country's geography and history to conduct the georefencing. Some of the major difficulties specific to this project are the result of Mexico's highly diffuse rural population with finely granular distribution and a history of changes in toponymy. This is the case when dealing with old collecting events, some from 150 years ago or from pre-revolutionary times (before 1910), after which many place names were changed. Under these circumstances the benefits of capturing locality data and georeferencing by technicians in the country of origin are evident. When records in a partially-populated database (where records are not yet populated with locality data) are ordered by collectors' names and within collector, by date, specimens from a given locality are presented in sequence. Locality data common to those in groups can be entered once and copied to subsequent records. One determination of coordinates suffices to georeference all from that shared locality. Sequences of specimens sorted in this way are also often useful for resolving the geo-coordinates of vague or uninformative locality descriptions, as localities can be matched or inferred from the spatial and temporal context of collector itineraries.

In addition to choosing which label data to capture and the level of precision and accuracy for georeferencing collecting localities, the process of transferring the information from a label into the appropriate database fields may be ultimately optimized by optical character recognition, natural handwriting recognition, and natural language processing technologies. Schemes for doing this have recently been studied (Beaman & Conn, 2003; Conn, 2003; Heidorn & Wei, 2008), and projects aimed at optimizing throughput for label data acquisition are underway (Best & al., 2009; A. Neill, Botanical Research Institute of Texas, pers. comm., 2009). SilverBiology, Inc., has marketed interactive software for identifying regions of text in images of specimen labels and for semi-automatically mapping the text to the database schema (http://www.silverbiology.com/products/silverarchive/). There are insufficient data to demonstrate whether these automated approaches can approximate the efficiency of human-mediated mapping and keystroke data entry of specimen information.

Since 2002, the University of Michigan Herbarium, with support from the U.S. National Science Foundation (awards: DBI 0138621, DBI 0646301), has embarked on a project to computerize its holdings of Mexican and Mesoamerican land plant collections. In this paper, we report on the efficiency of the workflows tested and implemented during this project.

■ MATERIALS AND METHODS

1. DATA ACQUISITION

The scope of this study of the efficiency of data acquisition from herbarium specimens starts with the capture of images of specimen labels through the creation of complete specimen database records (Fig. 1). We did not analyze the cost of hardware, the recurring expense of software maintenance, or the technical administration of database server hardware. Of the estimated 260,000 specimens of land plants from Mexico, Mesoamerica, and the West Indies at the MICH herbarium, ca. 122,000 have been "pre-cataloged" and their images captured, using one of the workflows described below. By pre-catalog we refer to the initial working database in which only the minimal set of 11–12 fields (as defined in Task D, below) were populated. In the pre-catalog, fields for locality, geo-coordinate or habitat were not be populated. The taxonomic groups included were mosses, gymnosperms, monocots (except orchids), and over 25 families of dicots, including some of the larger ones (Fabaceae, Euphorbiaceae, Lamiaceae, Scrophulariaceae) and those particularly diverse in the region or well represented at MICH (Piperaceae, Solanaceae, Cucurbitaceae, Cactaceae, Myrtaceae, and the genus *Quercus*).

Three independent workflows for acquiring herbarium specimen information were designed and the results of each evaluated.

Workflow 1. — Workflow 1 was implemented during the first funding period (2002–2005). During Stage I (see below) ca. 86,000 specimens were pre-cataloged (Table 1). During Stage II a subset of 42,000 records, all specimens collected in Mexico, was georeferenced.

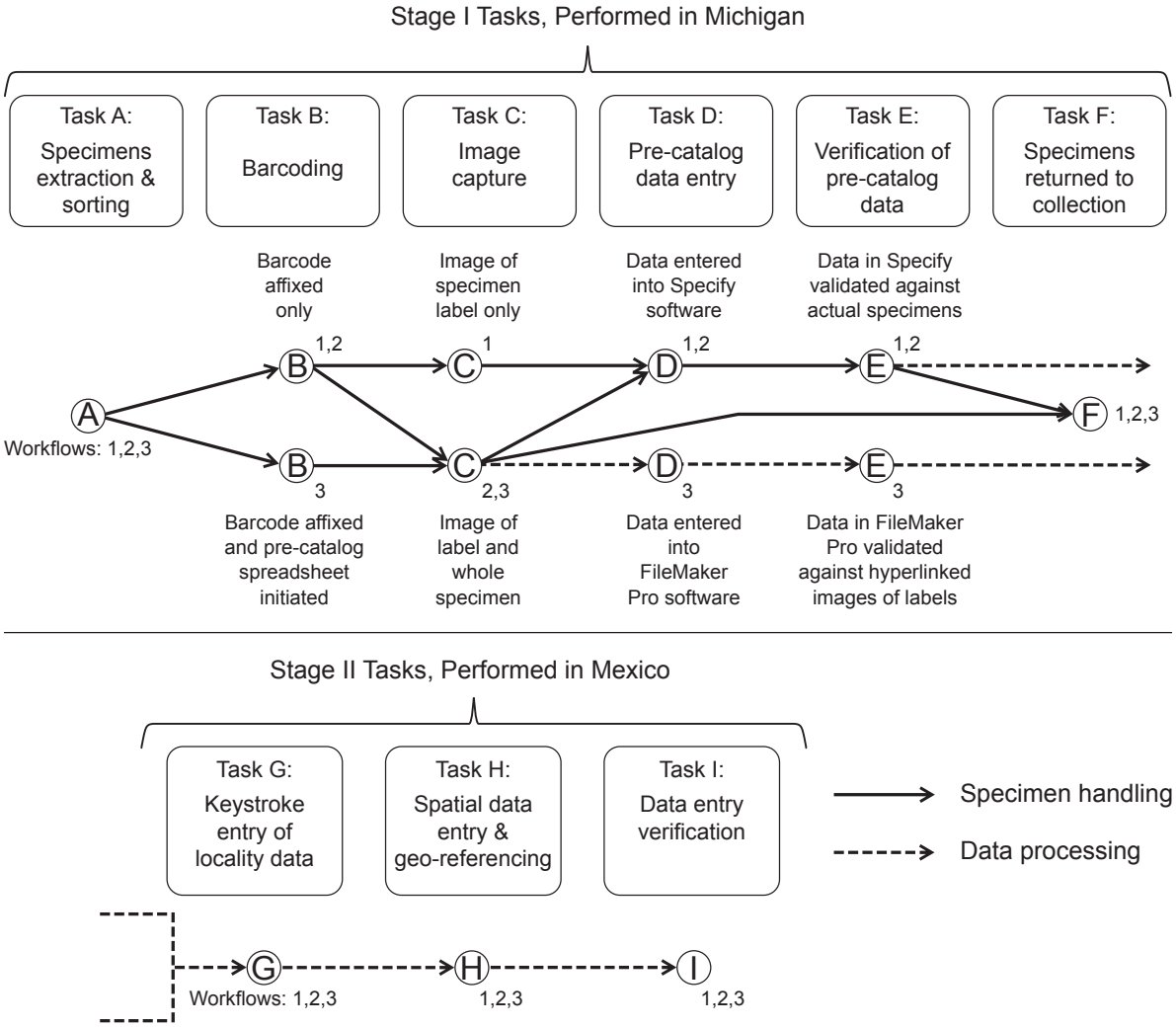

Fig. 1. Diagrammatic representation of workflow tasks. Numbers (1, 2, 3) refer to Workflow IDs.

Workflows 2 and 3. — Workflows 2 and 3 were implemented during the second funding period (2007–2008), creating ca. 38,300 database records with their corresponding images. Workflow 2: specimens were pre-cataloged in a SPECIFY database, directly from the specimen label by personnel at MICH. Workflow 3: records were pre-cataloged into a FileMaker Pro database, using hyperlinks to label images. Both images and the FMP database were hosted on MICH herbarium servers so data capture could be keystroked by personnel located outside MICH.

All workflows consist of three independent stages, each comprised of individual modular tasks (Fig. 1). The three workflows differ in how Tasks B, C, D, and E of Stage I were carried out (described below). Only Workflow 1 went through Stage II.

Description of modular tasks

Stage I: Specimen handling and pre-catalog database creation
- Task A: Specimen Extraction and Sorting
- Task B: Barcoding
- Task C: Image Capture
- Task D. Pre-catalog data Entry
- Task E: Verification of Pre-catalog Data
- Task F: Specimen Returned to Collection

Stage II: Georeferencing and Record Completion
- Task G: Keystroke Entry of Locality Data (verbatim)
- Task H: Spatial Data Entry and Georeferencing
- Task I: Data Entry Verification

Stage I: Specimen handling and pre-catalog database creation. — During this stage digital label images and pre-catalog were created. It was conducted at MICH and it consisted of six independent modular tasks. Time rates were recorded for each (Table 1). Specimens were moved through the Stage I workflow from task to task in batches, contained in 12-partition steel half-cases mounted on casters ("rolling cases"). Batch sizes were variable, mainly depending on the taxonomic group; the median being ca. 370 (±21) specimens per case, but sometimes containing as many as 1400 specimens. Holding batches in rolling cases during the entire processing of Stage I served as a buffer when workflow processing stalled.

Stage I Tasks. — *Task A. Specimen Extraction and Sorting.* – For each family selected, entire folders of Mexican and Central American specimens were pulled from collection cabinets to a rolling case, maintaining the order in which they were filed: alphabetically by taxon name and geographic region. All specimens belonging to a species were sorted by country and collector in order to spot duplicates, to flag those specimens consisting of more than one sheet, or to match specimens with corresponding fruit boxes when these existed. Markers were left in each shelf from which specimen folders were removed to maximize the efficiency of re-filing and to minimize error when specimens were returned to the collection (Task F, below). Task A did not vary between workflows.

Task B. Barcoding. – A barcode label was affixed 1–2 cm from the lower edge of the specimen, between the center of the sheet and the specimen label. When not possible, it was affixed within a ca. 160 × 10.5 mm² area that corresponded to a preset photograph frame. Specimens consisting of more than

Table 1. Productivity for the six modular tasks in Stage I (Specimen Handling and Pre-catalog Database Creation) of the three workflows tested. Standard errors are given for specimens/hour rates. Workflow 1 productivity figures are the median for batches sampled during the last year of the project.

	Workflow 1 (2004–05)		Workflow 2 (2007–08)		Workflow 3 (2007–08)	
Number of specimens in study	6627		23,638		11,512	
Modular Task	time (min:s)/ specimen	specimens/ hour	time (min:s)/ specimen	specimens/ hour	time (min:s)/ specimen	specimens/ hour
A. Specimen Extraction and Sorting	00:24	169	00:10	380±175	00:09	388±251
B. Barcoding	00:23	158	00:21	171±8	00:33	109±8[a]
C. Image Capture	00:33[b]	108	00:38	94±7	00:42	86±6
D. Pre-catalog Data Entry	02:27[b]	24[b]	01:09[b] (01:53)	52.2±1.7[b] (32±2)	00:57[b] (01:27)	63[b] (43±3)
E. Verification of Pre-catalog Data	00:34[b]	106[b]	00:43[b] (00:49)	85±0.5[b] (74±4)	00:32[b] (01:02)	113[b] (58±10)
F. Specimens Returned to Collection	00:01[c]		00:01.4	2634±213	00:01	2631±338
Median, entire batches	04:21[b]		03:13[b] (04:02)	19[b] (15±0.7)	03:00[b] (04:04)	20[b] (15±2)
Worst-case scenario: max. time/specimen	07:27		06:57		04:37	

[a] In Workflow 3 Task B (Barcoding) included the creation of a spreadsheet with only four data fields (columns) populated for tracking purposes.
[b] Values based on software timestamp data; to maintain consistency among tasks, time data from workers' time logs were used instead, when available (in parenthesis), for statistical analyses. For Task C, Workflow 1 captured images only of labels; Workflows 2 and 3 captured images of labels and entire specimens.
[c] Times for Task F were not measured in Workflow 1, but are not expected to differ as a result of workflow design; value is estimated from the mean of Task F in Workflows 2 and 3—which turned out to be almost identical.

one sheet or those with ancillary material such as fruit boxes were barcoded with a dedicated series that contained repeated numbers followed by sequential letters and flagged with a strip of colored paper to warn the person capturing images (Task C, below) of the need to photograph that additional material and the change in numbering sequence.

Task C. Image Capture. – Images of the specimen labels were taken using digital cameras, each mounted on a camera stand. In Workflow 1 images were taken with one camera per image capture station. In Workflows 2 and 3, the image capture station consisted of two cameras, each capturing one type of image. Both cameras were operated simultaneously by a single computer via USB or IEEE 1934 (Firewire) ports, each through its own proprietary image capture software (note that in order to use two cameras simultaneously, each needed to be controlled through a different capture application, which was not a problem since they were different brands, Fuji and Nikon, respectively). Image files were named consecutively through the capture software and directly saved onto hard drives. Necessary frame adjustments could be made using the zoom lens (at pre-set focal distances) to capture any annotation labels that fell out of the standard frame area. When necessary, an additional image (rarely more) was taken to capture labels that were affixed elsewhere on the specimen sheet. If an additional image was taken, the specimen was flagged with a strip of colored paper to indicate the need to record the existence of an additional image during pre-catalog data entry (next task). Obvious reduction in efficiency occurred when having to zoom out to capture a larger frame or, alternatively, when taking additional images. File names were assigned based on the specimen's barcode number. In order to optimize automation, every effort was made to run specimens sequentially by barcode number, as camera capture software assigned file names sequentially by default. In the case of specimens with more than one part, image file names needed to be entered manually (or with a barcode scanner) after each frame because barcode numbers with letters were in a different series. Image files were not manipulated further, other than running a batch automatic brightness, contrast, and color optimization, and compression with Adobe Photoshop to produce smaller JPEG files that could be hosted on the Herbarium server.

Task D. Pre-catalog Data Entry. – Specimen records were created to act as a "pre-catalog" database. The data fields populated for each record were:

1. Catalog (= barcode) number
2. Family
3. Taxon name (determination(s), including all previous determinations in Workflow 1, only the most recent and the original if different in Workflows 2 and 3)
4. Name(s) of determiner(s) ("agents")
5. Year of determination
6. Name(s) of collector(s) ("agents")
7. Collection number
8. Collection date or range of dates
9. Country of origin
10. Primary administrative unit (State, Province, or *Departamento*)
11. Presence/absence of ancillary material (i.e., a code indicating more than one sheet or additional material, such as a separate fruit specimen or wood sample)
12. Herbarium of origin when on label (collection from which the specimen was distributed, only in Workflow 1)

Taxonomic fields were populated from related authority tables through drop-down lists. The initial sources for names were CONABIO's Catálogo de Autoridades Taxonómicas for bryophytes (http://www.conabio.gob.mx/informacion/catalogo_autoridades/doctos/briofitas.html; Delgadillo, 2003), and Flora Mesoamericana (http://mobot.mobot.org/W3T/Search/index/mesoa.html) for vascular plants. For names that may appear in specimens but not in authority tables, its status as an accepted name, orthography, and the name's authority were verified by comparison with entries in the International Plant Name Index (www.ipni.org) and TROPICOS (www.tropicos.org) databases. Fields for standardized names of collectors and determiners ("agents") were also populated using related authority tables from CONABIO's Catálogo de Autores de Plantas Mexicanas (http://www.conabio.gob.mx/informacion/catalogo_autoridades/doctos/autplanvasmex.html; Villaseñor & al., 2008). Newly encountered agent names were verified through the Index of Botanists database (http://asaweb.huh.harvard.edu:8080/databases/botanist_index.html).

Task E. Verification of Pre-catalog Data. – Data in the pre-catalog records were compared with those in specimen labels to check for data entry errors and inconsistencies. Errors were corrected directly in the pre-catalog records.

Task F. Specimens Returned to Collection. – Specimens in an entire rolling case were re-filed in their original cabinet locations. In Workflows 1 and 2 this task was done after pre-catalog records were checked, but in Workflow 3 it was usually done after specimens were imaged.

Stage II: Georeferencing and Record Completion. — Label images were used to complete the contents of the locality by populating locality fields and georeferencing pre-catalog records. This stage was carried out between 2004 and 2005 by collaborators in Mexico City as part of Workflow 1 and only involved specimens from Mexico. (As of this writing, specimens in neither Workflow 2 nor 3 have been processed through Stage II.) MICH provided both images and the pre-catalog. Images were compressed (SID encoded, http://lizardtech.com) and loaded on Windows machines. Specify v.4.6 and MS SQL Server databases containing the pre-catalog data generated through Task D were installed on local workstations. SID-encoded images were invoked via a hyperlink in each specimen record. Records from Mexico were selected and then sorted by collector and collection date to create sequences of records that would share locality information and, thus, geo-coordinates.

Stage II Tasks. — *Task G. Keystroke Entry of Locality Data (Verbatim).* – For each batch of records, fields were populated by transcribing data verbatim from label images into the corresponding fields:

13. Verbatim locality
14. Longitude (if provided on label, verbatim)

15. Latitude (ibid.)
16. Altitude (ibid.) and units
17. Verbatim habitat

Error check for captured verbatim data was done through double entry.

Task H. Spatial Data Entry and Georeferencing. – Subsequently, data from the verbatim locality field (13, above) were read and interpreted using gazetteers (INEGI, 2000) to populate the standardized fields for state and municipality (18 to 20, below), and localities were assigned geo-coordinates. This required interpretation of locality information presented in labels in order to georeference. Geo-coordinates were calculated separately by means of Voronoi-generated polygons (Gold & al., 1996) with MapInfo and the resulting data used to populate fields 21 to 23, below.

18. State (if not populated in Stage I)
19. Municipality
20. Locality in current nomenclature (from gazetteer)
21. Latitude (decimal degrees, as per gazetteer)
22. Longitude (ibid.)
23. Altitude (in meters)

Task I. Data Entry Verification. – Before submitting database back to MICH a random check of data captured for fields 18–23 was conducted manually.

Data for locality fields (listed above) were populated from images. It was realized early in the project that sending a discrete batch of records at a time (e.g., a group of families as they were completed through Stage I) to collaborators in Mexico City was impractical. The most efficient was to send the whole pre-catalog after specimens of all families had been run through Stage I entirely, discontinuing all operations for Stage I once the pre-catalog was submitted. Ultimately, the 42,000 records that were suitable were georeferenced.

Cost for populating records in Stage II was based on a flat rate of US$1 per record populated, georeferenced, and verified (Tasks G, H, and I). Because of this model of a flat fee charge per record, which realistically is most likely to be the scenario for conducting Stage II in future applications, no time-motion data were kept.

Differences among Workflows. — *Workflow 1.* – Tasks A, B, and F were conducted in the manner described above without major modification. Task C: only label images were captured using either a Fuji FinePix S1 or a S2 Pro camera mounted on a photo stand. Two independent camera/computer workstations were run concurrently, each processing specimens from separate families. A pre-determined frame size of 16 × 10 cm was captured as to include the barcode and collection labels and, if present and when possible, an annotation label in just one image. An image resolution of 2300 × 1600 pixels produced 0.6–1.2 MB JPEG files that were uploaded to the computer that operated each camera. No further manipulation of images was conducted. Image files were backed up, compressed (encoded as SIDs), and shipped as a batch to our collaborators in Mexico for populating locality fields and georeferencing (Stage II). Task D: records were created using a Specify database (www.specifysoftware.org) v.4.6 by key-stroking directly from physical specimens in hand. A Microsoft SQL Server database computer was hosted on campus as the data repository, with locally networked MS Windows XP workstations. Task E: all pre-cataloged records in the Specify database were checked against the label information on the actual specimen sheets by a project manager.

Workflow 2. – Tasks A, B, and F were also conducted unmodified. Task C: the main difference with Workflow 1 is that, in addition to the specimen label, an image of the entire sheet at high resolution was captured for each specimen. Two cameras, a Fuji S2 Pro and a Nikon D80 were mounted in one photo station, both operated through a single computer running Mac OSX. Cameras could be operated simultaneously because each ran on their respective image capture software (Fuji Studio Utility v.1.2 and Nikon Camera Control Pro v.1.3.1, respectively). The Fuji S2 Pro camera mounted at one side of the stand captured only images of specimen labels. The Nikon D80 mounted on the center of the stand was oriented on a plane perpendicular to the Fuji S2, to capture images of the whole specimen sheet—placed sideways with respect to the camera frame—at the camera's highest resolution (3872 × 2592 pixels). The Nikon Capture software, operated from the same computer, generated a RAW image file (Nikon proprietary NEF format) plus a low-compression JPEG. Original JPEG image files of labels were backed up onto external hard-drives. Task D: similarly to Workflow 1, records were created in a Specify database, but using version 5.2. Fields populated were the same as for Workflow 1 except for 12 (Herbarium of origin), and no more than two determinations (field 3) were recorded: the most recent and, if different, the original on the specimen label. The Specify database was hosted at a local server at the herbarium, networked with up to three other Windows computers as clients. Task E: verification of data entered was conducted in the same manner as in Workflow 1.

Workflow 3. – Tasks A and F remained unmodified. Task B: as a batch of specimens (a rolling case) were being barcoded, a preliminary list was created as a Microsoft Excel spreadsheet, which consisted of the sequential series of barcode numbers that would be assigned to those specimens. Only data for just four fields, (1) catalog number (barcode), (2) family, (3) taxon name, and (8) country of origin, were recorded for each specimen. Task C: image capture was conducted in an identical manner as in Workflow 2, but at the completion of this task the batch was returned directly to the collection cabinets, i.e., to Task F. Task D: data entry differed substantially from the other two workflows as records were created in a FileMaker Pro database (www.filemaker.com). Importing the Excel spreadsheets previously generated in Task B (barcoding) was the basis for the FileMaker Pro pre-catalog. Hyperlinks to the corresponding label images hosted at the MICH server were automatically generated for each record. The database was uploaded to a Herbarium server. A single data-entry person working off-site would download the FMP pre-catalog from the MICH server and populate the eight remaining fields by keystroking from label images. The images, hosted at the MICH data server, were invoked through URL hyperlinks in the FMP pre-catalog. Task E: pre-catalog records were checked against label images by the project director by invoking the corresponding hyperlink

for each record. In this workflow, records were usually checked long after data entry, as specimens were returned to the collection cabinets after Task D.

2. MEASUREMENT OF WORKFLOW EFFICIENCY

Stage I. — Time and motion data were collected for each batch of specimens (each rolling case) as they were run through each of the modular tasks of Stage I (Specimen Handling and Pre-catalog Database Creation). In all three workflows, time data for pre-catalog data entry and verification of pre-catalog tasks, as well as for the image capture task in Workflow 1, were all taken from Specify and image capture software timestamps, respectively. Otherwise, data was taken from time logs kept by workers, including pre-catalog entry and verification tasks in Workflows 2 and 3, for which either source—software timestamps or workers' timelogs—was considered. Data used for comparing efficiencies between different workflows was the median time per specimen to complete each task. One-way analysis of variance and bivariate fit tests were conducted to determine significance of differences in efficiencies using the JMP package (v.5.0 for MacOSX, SAS Institute). Time data recorded includes time spent moving batches from one task's location to the next, consulting with supervisors or curators, loading novel taxonomic or agent names onto their respective data tables (including checks against nomenclatural resources), and being distracted from the task. Breaks lasting ≥20 minutes that interrupted the workflow were excluded from calculations (i.e., the clock was stopped), but those <20 minutes were regarded as part of the natural workflow.

Workflow 1. – Workflow samples of elapsed times for individual tasks were taken from those conducted by fully trained personnel in the last year of the project, when hardware, software and remote data entry issues were mostly resolved and efficiency optimized. Data samples for analyzing productivity were taken to include rather long stretches of time (2–4 hours), so the medians given are more realistic by accounting for all motions derived from normal activity. Time-motion data for Task C (image capture) were taken from file creation times, and for Tasks D and E from record created and record modified timestamps, respectively (stored by Specify with each record, along with person who last modified it). Recorded timestamps were taken for a given person's entire day of operation.

Workflows 2 and 3. – Time-motion data for all modular tasks were recorded for 93 of the 102 batches of specimens that were processed during the second iteration of the project (2007–08). Data from 56 batches (24,600 specimens) were processed through Workflow 2, and 36 batches (11,500 specimens) through Workflow 3 were suitable for the analysis. Personnel conducting each task filled data logs with the time spent on each task and the specimens processed (based on sequences of barcode numbers) for every batch. Data were reduced to minutes spent per specimen, or to specimens processed per hour on each task, A to F. Work sessions were defined as stretches of continuous time that a worker would spend on a given task, as long as idle time did not exceed 20 minutes. Additionally, for Tasks D and E (Pre-catalog Data Entry and Verification) time elapsed was also taken from record creation/modification timestamps provided by the software (Specify in Workflow 2, FMP in Workflow 3).

Stage II. — Data-entry and georeferencing specialists in Mexico City conducted Stage II (Georeferencing and Record Completion) for ca. 45,000 records that were pre-cataloged at MICH during Stage I of Workflow 1. Only the pre-catalog and label images generated during Stage I of Workflow 1 were processed through this Stage II, thus no comparison with other workflows was drawn for this stage.

■ RESULTS

Comparison among Workflows. — The total time to process a specimen through the Specimen Handling and Pre-catalog Database Creation Stage was 4:21 min for Workflow 1. This is the sum of the median times for each of modular tasks that constitute the workflows' Stage I (Table 1). The statistic is based on task data from the last year of the project, when efficiency was optimized. In a worst-case scenario (adding the times for the slowest performance of each modular task), 7:27 min would be required to process a specimen in full, while in a best-case scenario (adding the fastest performance of each modular task), only 2:56 min would be required.

In Workflow 2 per-specimen median time across all batches was 4:02 min (15 specimens per hour), and in Workflow 3 4:04 min. Median times to process a specimen through Stage I showed negligible difference between Workflow 2 and Workflow 3 (Table 1). These results are based on data obtained from workers' time logs, and ought to be regarded as rather conservative figures. When software timestamps were used for calculating time spent conducting pre-catalog data entry, and verification—the two most time-intensive tasks—the median times for Workflows 2 and 3 were 3:13 min and 3:00 min per specimen, respectively, with the slowest batches at 6:57 min and 4:37 min per specimen, respectively.

Effect of modular task design on efficiency. — On a per-task basis, no statistically significant differences were found for tasks that remained unchanged between Workflows 2 and 3 (Tasks A, C, and F, Table 1). In Workflows 2 and 3 Task A (Specimen Extraction and Sorting) seem to have been conducted at a faster pace than in Workflow 1, but the amount of data gathered during implementation of this task for Workflow 1 was insufficient to determine whether such difference is significant. In Workflow 1, on the other hand, a slightly faster pace for image capture (Task C) was seen with respect to Workflows 2 and 3 (Table 1), but not in a significant manner ($P = 0.35$). This indicates that the added burden of taking the additional high-resolution image of the entire specimen (in Workflows 2 and 3) had little effect on workflow efficiency or cost.

Comparing the pace of conducting pre-catalog data entry (Task D) between Workflow 1 and Workflow 2, an analysis of variance shows that Task D runs significantly faster ($F = 11.59$, $P = 0.001$) in Workflow 2 than in Workflow 1 (32 and 24 specimens/hour, respectively). In both workflows records were

populated into a Specify database for a very similar set of fields, with two main differences. First, the taxon name, determiner's name (agent), and year of determination that accompanied all successive determinations and annotations for a given specimen—when stated—were recorded during Workflow 1 (fields 2, 3, and 4, above). During Workflow 2, however, only the most recent determination was recorded and, if different, also the original one (the one given in the collection label, supposedly that under which the specimen was distributed). Second, the field for the specimen's herbarium of origin (i.e., herbarium that distributed those duplicates or *exsiccatae*, as indicated in label headers, if one of the approx. 20 for which MICH holds a large number of Mesoamerican duplicates) was recorded in Workflow 1, but the practice was abandoned during implementation of Workflows 2 and 3. Medians of per-specimen times for the pre-catalog data entry task were reduced by 23% in Workflow 2, and 41% in Workflow 3. As expected, reducing number of fields populated seems to be a major factor in the faster pace of this task for Workflows 2 and 3.

The purpose of conducting Workflow 3 was to test the scalability of using images hosted in a remote data server so keystroke data entry could be performed off-site. When comparing pace of Task D (Pre-catalog Data Entry) between Workflows 2 and 3, populating fields in a FileMaker Pro pre-catalog using remote images (Workflow 3) was significantly faster ($P = 0.002$) than populating the Specify database (Workflow 2). Although the Workflow 3 task of pre-catalog data entry from images in itself takes 23% less (27 s) than from data entered from actual specimens (Workflow 2), when the time to generate a preliminary spreadsheet during Workflow 3 Task B (Barcoding) is accounted for, the combined time difference drops to a mere 11%. Data entry verification (Task E), on the other hand, is quite variable. When the source of time data were the logs kept by workers, Workflow 3 was 21% slower when run on FMP by invoking images on-screen, as compared to Workflow 2 where verification was done in Specify from the actual specimen.

Effect of human variability on efficiency. — An analysis was conducted to determine how much variability is due to differential efficiency of individual workers carrying out each task. Data were collected from 17 individuals involved in one or more tasks of either Workflow 2 or 3. In order not to skew results, no data from individuals were excluded on the basis of their experience, or lack thereof, with any given task (i.e., from the first day on the job, even if it could be considered a training session). Due to the nature of the high turnover of the student workforce in a university setting—some were involved for less than three weeks—it would be unrealistic not to account for the low productivity of less trained workers. Tasks A, B, C, and D were conducted by all workers rather indiscriminately, but for Tasks B and D there was limited or no crossing over of workers between Workflows 2 and 3. The fastest 1–2 workers were significantly more efficient for most tasks than the slowest 1–2. Two of these workers were consistently the slowest for most tasks, but for the rest of workers their comparative efficiency differed from task to task (e.g., the fastest at pre-catalog data entry was among the slowest at image capture).

There is some variability in efficiency as a function of the individual worker for Tasks B, C, and D. Pre-catalog data entry (Task D), apart from being the bottleneck for all workflows, was also the task showing greatest variability among data-entry personnel. Significant differences in efficiency were seen between the 1–2 fastest individuals and the 2–3 slowest for Task D. Variability for the other tasks was much smaller. For most tasks, other than Task D, length of individual's involvement in the project appears to have little effect on their work efficiency. However, Tasks B, C, and D were rarely all conducted by the same worker throughout an entire batch, while Tasks A, E, and F were specifically performed by the project manager.

Increase in efficiency through the life of the project was observed during implementation of Workflow 1 (first funding period, from 2002 to 2005; Fig. 2). It almost doubled between the first and third year, to the point that two independent iterations of the workflow were run simultaneously by the second year of the project with the same level of staffing. In the space of 17 months (May 2007 to September 2008, second funding period), overall efficiency of Workflow 2 steadily increased by a factor of ca. 2.5 (from ca. 8 specimens/hour to more than 25 specimens/hour, statistically significant: $F = 34.46$, $P < 0.0001$), showing no indication of flattening out. This increase in efficiency was the result of the significantly increased pace of Tasks B (barcoding, $F = 10.05$, $P = 0.0025$), C (image capture, $F = 17.91$, $P < 0.0001$), D (pre-catalog, $F = 11.15$, $P = 0.0016$), and E (verification, $F = 21.73$, $P < 0.0001$). Workflow 3 showed great heterogeneity overall, despite the fact that two thirds of all pre-cataloging was conducted by just two persons, and a much more modest increase in overall efficiency throughout its implementation. The latter may be the result of having reached a ceiling in efficiency, precisely because during the later months fully trained individuals performed most of Task D.

Estimated per-specimen cost for Stage I. — Based on the times, the median labor cost to fully pre-catalog a record through Stage I of Workflow 2 and 3 was ca. US$1.00 (US$0.92 and US$0.63, respectively, if we base the calculation of pace for Tasks D and E on the presumably more accurate timestamps provided by the database software). This cost is calculated on a standard US$12/h rate for student workers and US$18/h for the project manager, plus benefits.

■ DISCUSSION AND CONCLUSIONS

Variability in efficiency for Task A was rather high. The large heterogeneity in the number of specimens for a given taxon factors in the efficiency for this task. Sorting specimens by collector takes more effort in species with much material, while it goes much faster in those with fewer representatives. This pattern is similar, although to a lesser extent, for number of species belonging to a genus. Adding this step to the task showed some effectiveness, in part by speeding up pre-catalog data entry in subsequent Task D, as collectors' names appeared in sequence, but mainly by reducing errors (as evidenced by a somewhat reduced time required for Task D, Verification of pre-catalog data). Although no data are available to back it up,

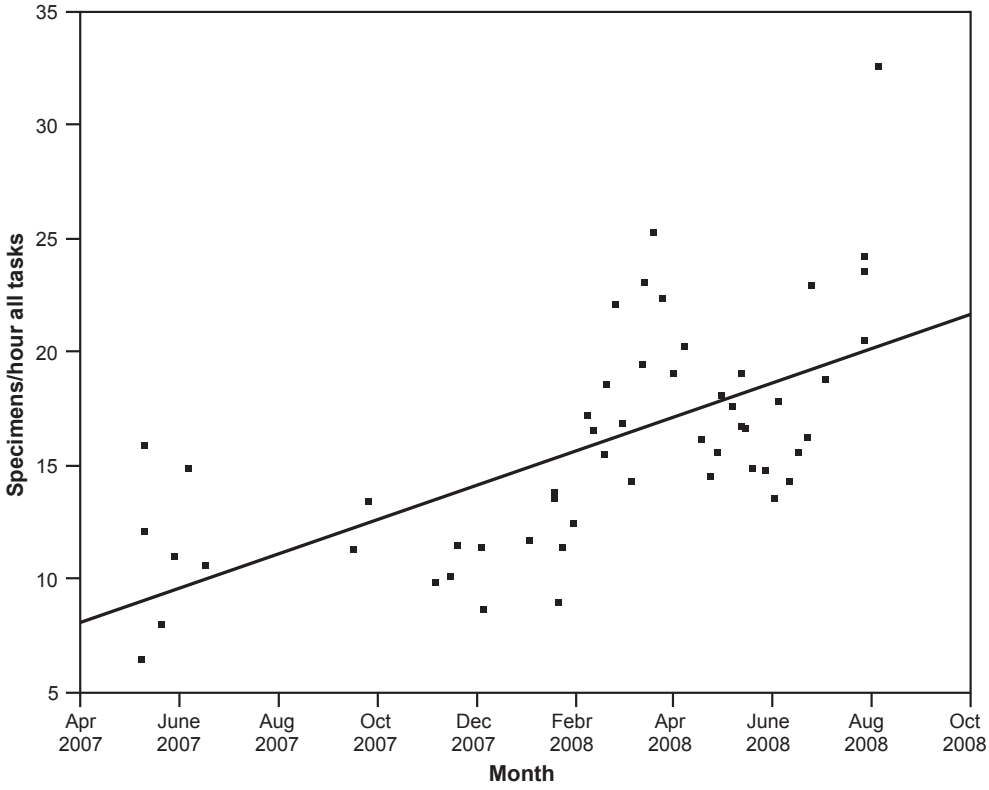

Fig. 2. Increase in efficiency of Workflow 1 with time. Efficiency is given as the number of specimens per hour processed for an entire batch (a single rolling case) as it went through all five tasks of Stage I (Specimen Handling and Pre-catalog Database Creation). The date corresponds to when a batch was started as the project proceeded. $R^2 = 0.412$.

sorting specimens by country and collector was strongly and unanimously perceived by data entry personnel as reducing time, effort, and errors during Task D (pre-catalog data entry). Groups in which many specimens are made of multiple sheets or include ancillary material (fruit boxes) require additional time for matching all the sheets that make up a whole specimen, as these sheets are often found not in sequence. The same applies when searching for fruit boxes—filed separately in the collection—associated with some specimen sheets. Recurrence of duplicates also slows down the process. Realization that a specimen is a duplicate often does not happen until Task B, D, or even E (barcoding, pre-catalog data entry, or verification), which transfers the burden—and the added time—of removing all duplication from the system to those tasks (only one of the duplicates will be databased and the rest removed from the collection).

The comparison of the relative efficiencies of Workflows 2 and 3 yields an interesting outcome: The purpose was to determine whether populating fields from label images accessed online (Workflow 3) would result in increased efficiency. Because data entry verification (Task E) is also conducted by accessing label images online, Workflow 3 also reduces in half the amount of handling of herbarium sheets and the times of rolling cabinets have to be moved from task to task by returning specimens to the collection after label images have been captured (Task C). Time for uploading image files to the Herbarium server is negligible when done in large batches. Workflow 3, however, requires the additional step of creating a preliminary spreadsheet during barcoding (Task B). This includes creating fields for taxon name that was populated quasi-automatically with the name by which specimens are filed in the collection. Despite this obvious efficiency, the procedure still slows down the pace of barcoding. Obviously, the data entry verification task is also carried out through images accessed online. Only a few fields needed to be populated in the pre-catalog in this workflow, and by highly competent data entry staff, so the error rate was very low. For this reason it was deemed appropriate to check a sample of ca. 20% of records. However, the results are inconclusive on whether the data entry verification task proceeds at a faster or slower pace though one workflow or the other. Because the pre-catalog data entry and verification tasks in both workflows are performed using two very different software (Specify and FMP, respectively), different user interfaces, and different data entry workflows, we cannot determine whether differences in pace are due to the latter or to keystroke data from actual specimens vs. images. Using FMP as a platform for specimen data acquisition cannot, by any means, be seen as an alternative to the greater power and scalability of Specify, but as a proxy, it was an easy to implement workflow to test the effectiveness of data handling and acquisition using images when off-site data entry becomes necessary. Ultimately, overall efficiency of a workflow that relies on populating the pre-catalog directly from actual specimens into Specify remains competitive.

The primary objective of our effort was to computerize the label information for as many Mexican (and some Central American) specimens in the MICH herbarium within the constraint of the available financial resources. We assumed, from previous experience, that data entry workflows should be composed of discrete tasks in order to maximize productivity.

These tasks include: (1) data acquisition from the physical specimen should be separated from keystroke data entry through imaging of specimen labels, (2) a pre-catalog database should be created to capture the least amount of information necessary to further identify and sort linked specimen label images of as many specimens as possible, and (3) the pre-cataloged label data and associated images should then be sorted prior to keystroke data entry on collector name, collection date, and collector number, with data from all pre-cataloged specimens (and taxa) pooled together. A methodology that maximizes the number of chronologically adjacent collection records for a given collector or collecting team would most likely significantly improve the quality and speed of data entry by optimizing the homogeneity in layout and content of successive labels. However, a further study would be needed to determine whether this is the case and to what extent. Also, when specimen labels are thus sorted, locality data shared among specimens become apparent, and full geographical data entry and georeferencing tasks become more efficient later in the process.

In order to optimize a workflow that relies on digital label images, Stage I (Specimen Handling and Pre-catalog Database Creation) should be completed for the maximum number of specimens from as many taxa as possible across the entire collection (or a particular project's entire circumscription). Then, the entire pre-catalog database, with all specimens (and taxa) in the target collection pooled, irrespective of taxonomic identity, is sorted by collector name and collection date and number. This will maximize the number and adjacency of records that share collector or collecting team. In this way not only data format consistency and the ability to interpret label data is increased, but also the frequency of a collector's itinerary is increased when records are sorted by collection date. Also, when 'collecting events' are thus sorted, locality data shared among specimens that were collected sequentially become apparent, and data entry and georeferencing in workflows' Stage II (Georeferencing and Record Completion) become more efficient.

Our study shows that the time and effort required to complete a workflow and, consequently, the cost of specimen record creation is greatly influenced by individual task optimization and the experience and competency of specimen and label data handlers. However, expertise for tasks such as specimen sorting, barcoding, and image capture (Tasks A, B and C) can be reached within 2–3 days, and no more than 4–5 days for the more demanding pre-catalog data entry (Task D).

The mean labor cost of pre-cataloging a botanical specimen in the workflows tested (Stage I) was US$1.36 per specimen for Workflow 1 (cost figures are based on gross salary rates of US$13/h for student and hourly-paid workers, and US$23.50/h for a half-time project manager, including fringe benefits). Populating locality fields and georeferencing (Tasks G through I of Stage II) were completed by collaborators in Mexico at an additional cost of US$1.00 per record, which brought the total cost per specimen to ca. US$2.36. These costs, however, do not include hardware, supplies, P.I. salary and benefits or facility and host institution overhead charges. When the entire project budget (NSF, BRC Program 2002–05) is computed, the total cost for a specimen pre-cataloged in Stage I (Specimen Handling and Pre-catalog Database Creation) was US$5.06 when processed through Workflow 1 (US$6.13 once Stage II—Georeferencing and Record Completion—was completed). Efficiency in completing Stage I through Workflows 2 and 3 increased slightly, bringing down the labor cost per specimen to between US$0.63 and US$1.20. The cost of completing Stage II is expected to be the same as that for Workflow 1, to total not more than US$2.20 per specimen. However, the most important increase in efficiency conducting Stage I is seen when considering the overall budget (NSF, BRC Program 2007), as the cost came down to US$3.80 per specimen (expected to remain well below US$5 after Stage II is implemented by CONABIO). The faster pace in conducting Stage I through Workflow 2 is the result of knowledge acquired during implementation of the first project (NSF, BRC Program 2002–05: i.e., development of Workflow 1). This, along with improved and less costly hardware have contributed to reducing the overall cost per specimen.

Dr. Pascal Chesselet (Muséum National d'Histoire Naturelle, Paris) estimated the cost for full cataloging of herbarium type specimens for the API and LAPI projects in the range of €12 to €20 (about US$17 to US$28 in 2009, presentation at TDWG, 2009 Annual Meeting, Montpellier, France). The cost was eventually brought down to €5 to €9 (ca. US$7.00 to US$12.50 in 2009) with increasingly experienced personnel, as she indicates that personnel constitute ca. 89% of costs in the budget of the Global Plants Initiative project (P. Chesselet, pers. comm., 2009). In our project at MICH, however, personnel costs for conducting Stages I and II (P.I. salary and all fringe benefits included) constituted only 65% of direct costs (46% of overall budget, when including institution's indirect costs).

Once a project has run for several years, with hardware and set-up costs largely amortized and less hands-on management of P.I. required, specimen handling becomes the primary direct cost. At that point, fully digitizing a specimen for US$2.00 to US$2.50 would be a realistic target. We summarize below the factors that contributed most to optimizing efficiency.

- Processing specimens through a workflow that consists of independent tasks reduces the effect of bottlenecks by allowing two or more batches of specimens to be run simultaneously. This approach also provides great flexibility when depending on a student workforce, characterized by irregular work schedules. The workflow can operate almost as effectively whether with just one worker or with as many as six to seven working in parallel.
- The rate limitation for the Pre-catalog Data Entry task is the time required for data entry personnel to recognize and interpret information on typescript and handwritten labels and, in the case of collector and determiner names, the effort needed to recognize an agent's correct identity. We observed that sharing knowledge among data-entry personnel working in close proximity to each other led to increased efficiency and quality.
- Spelling and typing errors are less frequent and data entry more consistent in fields that are populated by

controlled vocabularies of pick-lists. We used this well-known technique successfully for taxonomic, geographic, and agent name fields.

- Workflows that handle label image files that are relatively small (0.6 to 1.2 MB) keeps disk access delays to a minimum, thus using mid-resolution images of specimen labels is a practical and efficient technique for populating data fields and later for georeferencing localities from label image text. The use of dedicated images of just specimen labels avoids having to use image software to manipulate and magnify the label contents from a whole specimen image file.
- The motion of capturing two images simultaneously: one of the data label and one of the full specimen sheet during Task C (as conducted in Workflows 2 and 3), only requires 13% to 19% more time (5 to 9 seconds per specimen) than just capturing one image (that of the label, as in Workflow 1). It is therefore recommended to capture both types of images, label and full sheet, as long as it is done in the same motion, i.e., with two cameras operated simultaneously.

Future research on the cost and optimization of plant specimen digitization would benefit from additional comparative analyses of alternative procedures, outsourcing, and additional automation of the various component steps.

■ ACKNOWLEDGMENTS

This project was funded by NSF/BRC grants DBI 0138621 and 0646301. The authors are thankful for the intellectual and technical support provided by the Specify Software Project personnel, University of Kansas Biodiversity Institute, especially G. Garneau. We are greatly indebted to J. Panero, C. Delgadillo M. and R. Magill for providing essential authority tables and invaluable advice. Colleagues and co-PIs at the University of Michigan Herbarium and EEB Department, S. Lindsay, P.E. Berry, D. Goldberg, C. and W.R. Anderson, A. Reznicek, and G. Smith, and at CONABIO, mainly J. Soberón, its former director, and P. Koleff, R. Jiménez R., and P. Ramos R., were a major source of comments, encouragement, and quality assurance. The project was possible only because of the effort of J.C. Gómez M., J.L. García C., V. Cruz M., and V. Jiménez E. in México D.F. with data capture and georeferencing and that of the very committed project managers and students at MICH, especially S. Arruda, Y. Dolev, C. Richter, B. Mason, K. Fultz, and L. Wines. The detailed comments and suggestions of dedicated anonymous reviewers greatly improved this paper.

■ LITERATURE CITED

Bart, H.L. 2005. *Geolocate*. http://www.museum.tulane.edu/geolocate (accessed January 2010).

Beaman, R.S., Cellinese, N., Heidorn, P.B., Guo, Y., Green, A.M. & Thiers, B. 2006. HERBIS: Integrating digital imaging and label data capture for herbaria [Abstract]. *Botany 2006, California State University – Chico. 28 July–2August 2006*. http://www.2006.botanyconference.org/engine/search/index.php?func=detail&aid=402.

Beaman, R.S. & Conn, B.J. 2003. Automated geoparsing and georeferencing of Malaesian collection locality data. *Telopea* 10: 43–52.

Berry, P.M. 1998. Intelligent workflow for collection management: Workflow Management Systems (Workflow Engines). SRI International. http://www.ai.sri.com/~swim/resources/SOA-workflow.html (accessed January 2010).

Best, J.H., Moen, W.E. & Neill, A.K. 2009. A framework and workflow for extraction and parsing of herbarium specimen data [Abstract]. *Proceedings of the Taxonomic Database Working Group (TDWG) 2009*. http://www.tdwg.org/proceedings/article/view/567 (accessed July 2010).

Catalogue of Life. 2007. Annual Checklist, Species 2000 & ITIS Catalogue of Life Hierarchy, Edition 1 (2007). http://data.gbif.org/datasets/resource/1542 (accessed through GBIF data portal, July 2010).

Cavalcanti, M.C., Targino, R., Baião, F., Rössle, S.C., Bisch, P.M., Pires, P.F., Campos, M. L.M. & Mattoso, M. 2005. Managing structural genomic workflows using Web services. *Data Knowl. Engin.* 53: 45–74.

CONABIO. 2006. *Capital natural y bienestar social.* México D.F.: Comisión Nacional para el Conocimiento y Uso de la Biodiversidad.

Conn, B.J. 2003. Information standards in botanical databases: The limits to data interchange. *Telopea* 10: 53–60.

Davies, A.M.R., Bodensteiner, P., Pillukat, A. & Grau, J. 2002. INFOCOMP: The Compositae types digital imaging project in Munich. *Sendtnera* 8: 9–20.

Delgadillo, C. 2003. *Catálogo de la Colección Briológica del Herbario Nacional de México. Actualización 2003*. Herbario MEXU, Instituto de Biología, UNAM. Base de datos SNIB-Conabio, proyecto U006. Mexico.

Edwards, J.L., Lane, M.A. & Nielsen, E.S. 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science* 289: 2312–2314.

Elith, J. & Leathwick, J.R. 2009. Species distribution models: Ecological explanation and prediction across space and time. *Annual Rev. Ecol. Syst.* 40: 677–697.

Ertter, B. 2000. Our undiscovered heritage: Past and future prospects for species-level botanical inventory. *Madroño* 47: 237–252.

Espejo-Serna, A., López-Ferrari, A.R. & Ugarte, I.S. 2004. A current estimate of angiosperm diversity in Mexico. *Taxon* 53: 127–130.

GBIF (Global Biodiversity Information Facility). 2002. Draft Report of the Meeting of the Digitization of Natural History Collections Scientific and Technical Advisory Group of the Global Biodiversity Information Facility. Copenhagen, Denmark.

Gold, C.M., Nantel, J. & Yang, W. 1996. Outside-in: An alternative approach to forest map digitizing. *Int. J. Geogr. Inform. Syst.* 10: 291–310.

Graham, G.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. 2004. New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol. Evol.* 19: 497–503.

Greenberg, J., Spurgin, K. & Crystal, A. 2006. Functionalities for automatic metadata generation applications: A survey of expert's opinions. *Int. J. Metadata, Semantics and Ontologies* 1: 3–20.

Greenwood, M. 2005. MyGrid. http://www.mygrid.org.uk/wiki/bin/view/Mygrid/WorkflowLinks (accessed July 2010).

Guisan, A. & Thuiller, W. 2005. Predicting species distribution: Offering more than simple habitat models. *Ecol. Lett.* 8: 993–1009.

Guthrie, K. & Nygren, T. 2007. Aluka. Building a digital library of scholarly resources from Africa. http://www.aluka.org (accessed July 2010).

Heidorn, P.B. & Wei, Q. 2008. Automatic metadata extraction from museum specimen labels. Pp. 57–68 in: Greenberg, J. & Klas, W. (eds.), *Metadata for semantic and social applications: Proceedings of the International Conference on Dublin Core and Metadata Applications, Berlin, 22–26 September 2008, DC 2008: Berlin, Germany*. Göttingen: Universitätsverlag Göttingen.

INEGI. 2000. *Principales resultados por localidad. XII Censo General de Población y Vivienda 2000.* CD-ROM. Mexico Sistemas Nacionales Estadístico y de Información Geográfica, Instituto Nacional de Estadística, Geografía e Informática. http://www.inegi.org.mx/sistemas/biblioteca/detalle.aspx?c=14051&upc=0&s=est&tg=55&f=2&pf=Pob, accessed July 2010.

Kirkup, D., Malcolm, P., Christian, G. & Paton, A. 2005. Towards a digital African Flora. *Taxon* 54: 457–466.

Lafferty, D. & Landrum, L.R. 2009. SALIX, a semi-automatic label information extraction system using OCR [Abstract]. *Botany & Mycology 2009, Snowbird, Utah, 25–29 July 2009.* http://2009.botanyconference.org/engine/search/index.php?func=detail&aid=130 (accessed August 2010).

MacRander, A.M. & Haynes, R.R. 1990. SERFIS, a methodology for making multi-herbaria specimen databases a reality. *Taxon* 39: 433–441.

McPhillips, T.M. & Bowers, S. 2005. An approach for pipelining nested collections in scientific workflows. *SIGMOD Record* 34: 12–17.

McVaugh, R. 1983–2001. *Flora Novo-Galiciana*, vols. 1–7. Ann Arbor: Univ. of Michigan Press.

Pankhurst, R.J. 2004. Computer technology for the future of SW Asiatic botany. *Turkish J. Bot.* 28: 129–138.

Rzedowski, J. 1976. *Catálogo de los herbarios institucionales mexicanos.* México, D. F.: Sociedad Botánica de México.

Sarukhán, J., Koleff, P., Carabias, J., Soberón, J., Dirzo, R., Llorente-Bousquets, J., Halffter, G., González, R., March, I., Mohar, A., Anta, S. & de la Maza, J. 2009. *Capital natural de México. Síntesis: Conocimiento actual, evaluación y perspectivas de sustentabilidad*. Mexico: Comisión Nacional para el Conocimiento y Uso de la Biodiversidad.

Schaub, M. & Dunn, C.P. 2002. vPlants: A virtual herbarium of the Chicago region. *First Monday* 7(5). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/issue/view/145 (accessed January 2010).

Schmull, M., Heinrichs, J., Baier, R., Ullrich, D., Wagenitz, G., Groth, H., Hourticolon, S. & Gradstein, S.R. 2005. The type database at Göttingen (GOET)—a virtual herbarium online. *Taxon* 54: 251–254.

Shankar, S., Kini, A., DeWitt, D.J. & Naughton, J. 2005. Integrating databases and workflow systems. *SIGMOD Record* 34: 5–11.

Sharp, A.J., Crum, H. & Eckel, P. 1994. *The moss flora of Mexico.* Bronx, New York: New York Botanical Garden Press.

Singh, M.P. & Vouk, M.A. [1996]. Scientific workflows: Scientific computing meets transactional workflows. http://www.csc.ncsu.edu/faculty/mpsingh/papers/databases/workflows/sciworkflows.html (accessed August 2010).

Smith, G.F. 2004. The African Plant Initiative: A big step for continental taxonomy. *Taxon* 53: 1023–1025.

Soberón, J. & Peterson, A.T. 2004. Biodiversity informatics: Managing and applying primary biodiversity data. *Philos. Trans., Ser. B* 359: 689–698.

Thiers, B. 2010. Index herbariorum database. http://sweetgum.nybg.org/ih/ (accessed July 2010).

Versteeg, R.J., Richardson, A.N. & Rowe, T. 2006. Web-accessible scientific workflow system for performance monitoring. *Environm. Sci. Technol.* 40: 2692–2698.

Villaseñor, J.L., Ortiz, E. & Redonda-Martínez, R. 2008. *Catálogo de autores de plantas vasculares de México.* México, D.F.: Instituto de Biología, UNAM, CONABIO. http://www.conabio.gob.mx/informacion/catalogo_autoridades/plantas/AutoresPlantas/AutoresPlantasMexicanas.pdf

Wieczorek, J.R. 2001. *Georeferencing calculator.* http://manisnet.org/GeorefGuide.html (accessed July 2010).

Wieczorek, J., Guo, Q. & Hijmans, R. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *Int. J. Geogr. Inform. Sci.* 18: 745–767.

Wolf, C. & Holland, D. 2000. Digitizing and preserving plant images: Linking plant images and databases for public access. *First Monday* 5(6). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/issue/view/120 (accessed January 2010).