Advancing Access to Biodiversity Data

Using the SALIX Method and Digital Field Guides

by

Anne Christine Barber

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2012 by the
Graduate Supervisory Committee:

Leslie R. Landrum, Co-Chair
Martin F. Wojciechowski, Co-Chair
Daryl Lafferty
Edward Gilbert

ARIZONA STATE UNIVERSITY

August 2012

ABSTRACT

The Arizona State University Herbarium began in 1896 when Professor Fredrick Irish collected the first recorded Arizona specimen for what was then called the Tempe Normal School – a *Parkinsonia microphylla*. Since then, the collection has grown to approximately 400,000 specimens of vascular plants and lichens. The most recent project includes the digitization – both the imaging and databasing – of approximately 55,000 vascular plant specimens from Latin America. To accomplish this efficiently, possibilities in non-traditional methods, including both new and existing technologies, were explored. SALIX (semi-automatic label information extraction) was developed as the central tool to handle automatic parsing, along with BarcodeRenamer (BCR) to automate image file renaming by barcode. These two developments, combined with existing technologies, make up the SALIX Method.

The SALIX Method provides a way to digitize herbarium specimens more efficiently than the traditional approach of entering data solely through keystroking. Using digital imaging, optical character recognition, and automatic parsing, I found that the SALIX Method processes data at an average rate that is 30% faster than typing. Data entry speed is dependent on user proficiency, label quality, and to a lesser degree, label length. This method is used to capture full specimen records, including close-up images where applicable. Access to biodiversity data is limited by

the time and resources required to digitize, but I have found that it is possible to do so at a rate that is faster than typing.

Finally, I experiment with the use of digital field guides in advancing access to biodiversity data, to stimulate public engagement in natural history collections.

DEDICATION

For my father, who has unconditionally supported me in everything that I do. For my mother, who has incidentally provided me with the tenacity to follow my bliss. And for Joe, who has been my loving partner through everything.

ACKNOWLEDGMENTS

I am grateful for the opportunity to thank the following people for their invaluable help and support, without which this project would not have been possible. First, I would like to thank Dr. Les Landrum for his advice and leadership over the last few years. I couldn't have asked for a mentor more patient and kind than him. I am also thankful for the excellent programming expertise provided by Daryl Lafferty and Ed Gilbert. This project surely would never have seen the light of day without them.

I am grateful to Dr. Marty Wojciechowski for his service as my committee co-chair. Both Kathy Rice and Liz Makings have generously provided me with assistance in naming plants, and for this I am forever grateful. I also wish to express my appreciation for the students at the ASU Herbarium, who have worked tirelessly at helping to perfect the SALIX Method.

Finally, I am thankful for the American Recovery and Reinvestment Act of 2009, which ultimately provided me with funding through the Division of Biological Infrastructure at the National Science Foundation.

TABLE OF CONTENTS

CHAPTERS Page

LIST OF TABLES

Table Page

LIST OF FIGURES

PREFACE

The information stored in museums and herbaria forms the basis of what we know about life. Our species concepts are based in collections, compiled from many years of study and fieldwork. From this fundamental knowledge, we are able to visualize distributional patterns and identify endangered species. This information is absolutely necessary for studying the potential effects of climate change on biodiversity loss. When occurrence records are easily accessible, important ecological studies can be conducted without first having to compile a comprehensive dataset. It has been estimated that there are over 1 billion specimens archived worldwide, but perhaps only 10-20% have been databased. Natural history institutions have the enormous task of collecting, digitizing, and mobilizing collections data. For my research, I wanted to focus on ways to efficiently accomplish these tasks.

My motivation for undertaking this project was mainly experimental. I knew there were general digitization standards recommended by the collections community, but guessed that the actual application of them was more tricky than it appeared. My objective was to develop an efficient workflow that included best practices in data capture and imaging. I also wanted to explore new possibilities in accessibility – mainly through the use of digital field guides.

For the digitization portion of the project, I worked with optical character recognition to speed the process of data capture. Once I had

optimized the software to work with our document types, the results were better than I had predicted. Open source OCR engines produced basically unusable text results, but the commercial software I used worked surprisingly well. Imaging was done through the use of high resolution digital photography. Once calibrated to work with the specific subjects and lighting conditions at the ASU Herbarium, I was able to produce acceptable specimen images. The value of associating records with images of the specimens goes beyond aiding in species identification. The image serves as a digital voucher of the specimen, providing immediate proof of its existence and of the accuracy of the associated collection data.

For the accessibility portion of the project, I focused on compiling a geotagged photo library of the ASU Arboretum on campus. Since my programming resources were limited, I took advantage of existing technologies. Uploading the image library to Panoramio provided a diverse audience with access to the information via Google Earth. I also made the guide available on SEINet as a vouchered checklist of photo observations. This site is tailored more to the needs of a specific, academic audience. Both Panoramio and SEINet provide a way to view the guide through Google Maps, with each marker linking to the full record.

In the future, I look forward to collaboration with software developers in creating an interactive mobile app[1] for biodiversity data.

[1] From application; specifically, a piece of software designed for use on mobile devices

Because the focus of doing so should be on public outreach and education, I think this would work best for living collections. However, herbarium specimens would work very well for creating distribution maps and interactive keys. The ability for users to submit their own observations is also important, for both public participation in science and the collections community. An enormous amount of data could potentially be harvested in this way – images of easily recognizable species associated with geographic coordinates can flesh out current distribution maps. There are a handful of existing mobile apps that are very similar to what I describe, except that the link between citizen and academic science is weak or non-existent. I believe that the future of natural history collections depends on public support, so a good relationship needs to be fostered between the two.

The future of digitization will continue to evolve and adapt to new technologies. Optical character recognition is a viable solution, but not the only one. There is a lot of potential in capturing data from field books and duplicate records. The way in which field books are organized makes them excellent resources for databasing. One locality is often associated with several records, so that the information would only need to be typed once and then duplicated. The herbarium at the New York Botanical Garden has achieved extremely good results using this method.

Duplicate records are often distributed to many other institutions following collection. At least one will end up being databased. If other

institutions were able to easily search and import this data, a specimen would only need to be databased once. At ASU, we have implemented this feature in Symbiota. It searches for existing duplicates at more than 30 participating institutions. DarwinCore compliance makes this method relatively simple – since the information is organized according to specific metadata standards, there are not usually any problems with non-compatibility across databases.

Probably the most promising method of digitizing extremely large collections is to initially capture skeletal records. These would include a specimen image, GUID, scientific name, collector, collection number, and coordinates. Other fields would be excluded until additional funding could be secured. The other fields require considerably more effort to capture, and for large collections, this could take many years to complete. A skeletal dataset would mobilize the most important information first, and could be completed using a combination of approaches at a later date.

The future is bright for natural history collections. Digitization projects may be the main focus at this time, but new opportunities in collection and accessibility can help bring natural history to the forefront of science. A positive feedback loop is becoming established, in which the public contributes to collections, and collections contribute to ecology. The more public participation that takes place, the more support collections receive, and the more information we have to help us solve real world

problems. Advancing access to biodiversity data plays a key role in this process.

CHAPTER 1

INTRODUCTION

Applications in Biodiversity Informatics. – The inestimable value of natural history collections is widely recognized, particularly since the more recent attention to pressing environmental and ecological problems. If we are committed to preserving our planet's natural heritage, we need to know how best to plan for the future and where to concentrate our efforts. The information housed and maintained in natural history collections forms the basis of what we know about species both living and extinct (Lane, 1996). Without access to this information, questions relating to climate change and biodiversity would be difficult to answer. As concerns over loss of biodiversity and its potential impact on humanity increase, the data we have collected as far back as ca. 2,000 years ago (Plinius, 77) becomes increasingly valuable and useful. Biological collections are able to provide a historical backdrop to what we know about the diversity of life – how it has changed over time, and how we have come to recognize distinct species. It has been estimated that a mere 1-15% of the world's species have been described, with perhaps only a limited window of time in which to discover the rest before they disappear (Lane, 1996). Our task is to collect, describe, and document species occurrences on a global scale. This task is immensely complex, however, and requires innovative approaches to tackling the huge amount of information to be processed. The holdings at Arizona State University alone number in the hundreds of

thousands – with estimates to approximately 1 billion in collections worldwide. The ability to access this information freely and easily is critical, and for many institutions, is a key part of their vision and mission.

The SALIX Method – The accessibility of biodiversity data is limited by the resources required to convert printed information to digital. Many tedious hours of labor are invested in the enormous job of data entry. If this process of digitization – both the imaging and databasing – could be at least partially automated, it may be able to move along at a faster pace. In an attempt to meet this need, SALIX (semi-automatic label information extraction) was developed under a grant to digitize ca. 55,000 botanical specimens from Latin America, housed at the Arizona State University Herbarium (ASU). Combining optical character recognition (OCR) with digital photography as ancillary technologies, SALIX works as an automatic parser to move specimen record information into a web-accessible database. Label images are captured during the imaging process, and batch processed in a commercial OCR program to create a text file. This information is edited by a user and moved through SALIX, where it is automatically parsed into the correct fields. The information is

then exported to a DarwinCore[2] compliant CSV[3] file and uploaded to SEINet[4], the online database at ASU.

What I refer to as The SALIX Method (Fig. 1) consists of a number of different software packages used in digitization. Both SALIX and BarcodeRenamer (BCR) were developed at ASU and designed to meet specific needs. BCR provides a way of automatically renaming image files to match GUIDs (global unique identifier), while SALIX works as an automatic parser. The other tools that are used are proprietary. I quickly discovered that in order to make OCR worthwhile, a higher functioning program than what was available open-source was needed. ABBYY FineReader Professional Edition is both affordable and reliable, and only one copy of the software license is needed. FineReader can output the text results in a number of different file formats, but I chose Microsoft Word for its ease of use and search-and-replace functionalities. This is the program that is used for all of the text editing. Finally, I chose Adobe Lightroom for image management and editing. This software package was designed with high-volume processing in mind, and works very well with large archives.

[2] Metadata standards in biological collections, as outlined by the Taxonomic Database Working Group, or TDWG (http://rs.tdwg.org/dwc/)

[3] Comma-separated values; a file type that stores data in plain text format

[4] The Southwest Environmental Information Network (http://swbiodiversity.org/seinet/index.php)

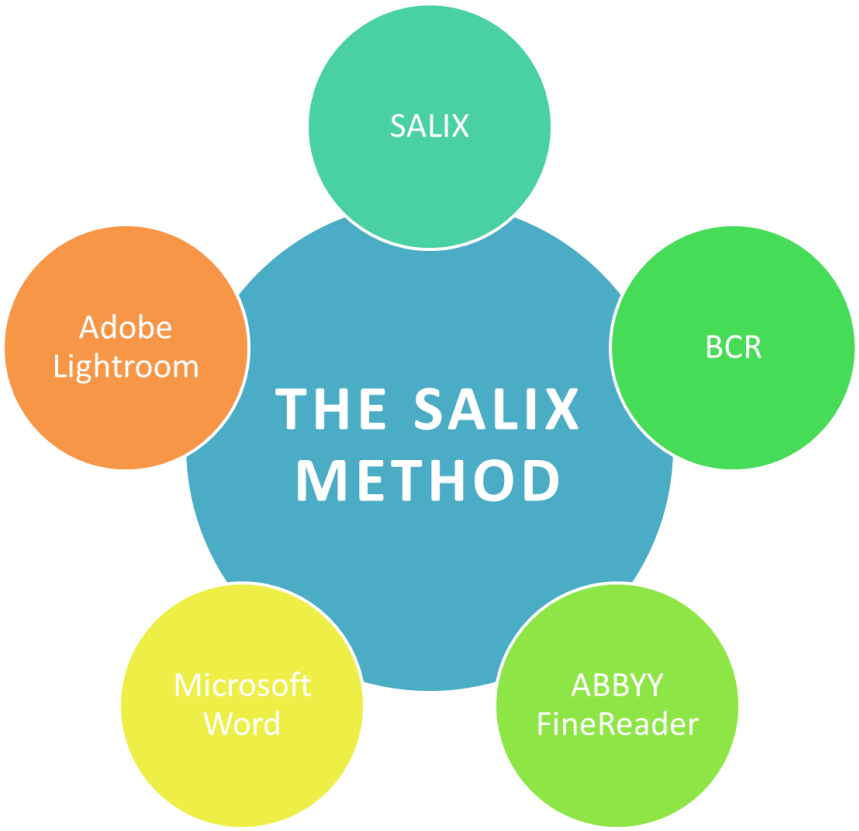


Fig. 1. The SALIX Method. The tools used in digitization consist of both commercially available software and applications developed at ASU.

I have found that the speed of data entry using the SALIX Method is dependent upon label quality and length, as well as user proficiency. When label quality is good, the SALIX Method can be up to 3 times faster than typing. On average, using the SALIX Method to database specimens has proven to be moderately faster than typing, but more importantly, it has opened up new possibilities in natural language processing (NLP) and the digitization of herbarium specimens.

The SALIX Method is not the only approach to semi-automatic data processing. HERBIS (erudite recorded botanical information synthesizer), developed initially at the Peabody Museum of Natural History, is also an

automatic parser that relies on OCR. Rather than parsing based on word statistics, as SALIX does, HERBIS uses label formatting to automate database field population (Heidorn & Wei, 2008). In a way, HERBIS was the inspiration for SALIX. It was thought that label formats might be too variable for automatic parsing to work reliably, and sought a different approach. Working with word statistics was chosen instead, so that parsing could function regardless of label formatting.

HERBIS is being used by the Botanical Research Institute of Texas (BRIT)[5] as a part of their digitization project, Apiary[6]. This method incorporates open-source OCR engines and parsing with HERBIS to aid in the automation of data entry (Moen & al., 2010). Unlike the SALIX Method, Apiary is web-based and the tools are accessible in a single application. SALIX is a stand-alone desktop application, and the other tools associated with it are used separately and behind the scenes by the project manager.

High-resolution digital photography and optical character recognition have only recently become affordable and practical. The use of these technologies in natural history collections has only just begun, and will likely continue to develop over the next several years.

Digital Field Guides. – The way in which information is accessed has changed dramatically over the last several years. Nearly half of all Americans are smartphone owners, that number having grown by 11% in

[5] The Botanical Research Institute of Texas (http://www.brit.org/)

[6] The Apiary Project (http://www.apiaryproject.org/)

the last year (Smith, 2012). Of these, 74% use their phones to find location-based information (Zickuhr, 2012). Adapting natural history data to work with this new demographic is an interesting challenge. As evidenced by the great number of natural history museums, botanical gardens, and paper field guides, we are very interested in the world around us. Easy, mobile access to location-based species occurrences can help answer common questions such as, *what is the name of this plant?* and *where does it grow?* Additionally, since 73% of cell phone owners use their phones to take pictures, there are also many opportunities for citizen science involvement (Smith, 2011).

There are a number of ongoing projects that attempt to combine natural history data with mobile technology. The National Audubon Society[7] has created quite a number of mobile apps[8] available for iOS and Android – these are targeted towards the general public and function very similarly to that of traditional paper guides. LeafSnap[9], developed by researchers at Columbia University, the University of Maryland, and the Smithsonian Institution, uses visual recognition software to automatically identify species from a photo of a leaf. It also includes an image library and a way for users to submit their own photographs of species. Currently, LeafSnap works only with the trees of New York City and Washington,

[7] Audubon Guides (http://www.audubonguides.com/field-guides/mobile-apps.html)

[8] From application; specifically, a piece of software designed for use on mobile devices

[9] LeafSnap (http://leafsnap.com/)

D.C. The Golden Gate Park Guide[10], developed at the California Academy of Sciences, provides location-based species information for Golden Gate Park in the city of San Francisco. The field guide portion of the mobile app is very similar to that of the Audubon guides, but with the added benefit of geotagged, user-submitted photos of species.

With these projects in mind, a digital field guide for the ASU Arboretum was created. Geotagged photographs of the living collection can be viewed either on SEINet as a vouchered checklist, or on Panoramio, Google's map-based photo sharing site. The project is being treated as a prototype for a fully-cataloged field guide that would be accessible on mobile devices. Instead of a paper map, visitors would be able to download an app that would function as an interactive guide to the Arboretum, viewable as images on a Google map layer. As it currently stands, these digital field guide projects are just scratching the surface in terms of data mobilization and outreach possibilities.

[10] Golden Gate Park Guide (http://www.calacademy.org/apps/ggp/)

CHAPTER 2

METHODS

Barcoding and Imaging. – Before any imaging occurs, each specimen is assigned a barcode (Fig. 2) that becomes its global unique identifier (GUID). Archival-quality, self-adhesive barcodes are used and placed directly above the label or as near to it as possible. These numbers do not match the accession numbers. Eventually, the GUIDs will replace the accession numbers as the entire collection comes nearer to being imaged. The GUID labels are preferable to the older, ink-stamped accession numbers in terms of longevity, readabilty, and risk of duplication. After a specimen is imaged, the GUID also becomes the image filename (ex: ASU0012345.jpg). This is valuable from an archival perspective – any record can be pulled up for reference by simply searching for the GUID in Windows Explorer or OS X Finder. Additionally, the SEINet database uses the GUID to automatically link images to specimen records, and vice versa.

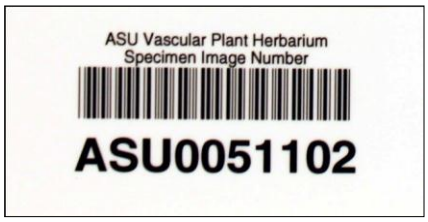


Fig. 2. An example of an image barcode. The barcodes used for this project are a set of catalog numbers that fulfill the DarwinCore element description of global unique identifier. These will eventually replace the older accession numbers.

The next step in the workflow is to photograph the specimens. A very simple platform surrounded by fluorescent lighting is used in place of a copy stand. The primary camera, an 18 MP DSLR[11], is mounted above the platform and connected to a remote shutter release. A second camera is positioned over the location of the label, and is set to Full Auto. The quality of the label image is not as important; the full auto setting will capture images that are perfectly acceptable for optical character recognition. This camera does not need to be very advanced. As long as it has at least a 5 MP sensor and the autofocus performs well, it will do the job.

It is commonly recommended to shoot in camera raw, which could be CR2 (Canon), NEF (Nikon), DNG (Adobe), or some other file extension. The reason for this is that JPGs tend to degrade rather rapidly with each adjustment and re-save. Raw images do a better job at preserving the original quality of the image. The downside to shooting in raw is that the files can be approximately 2-6 times larger than JPGs. The specimen images do not go through too many different adjustments before being put online, so it was decided to compromise shooting JPGs for server space. Especially since Adobe Lightroom was being used, this option seemed to make the most sense. Lightroom uses non-destructive editing processes, which means the original image always remains unaltered.

[11] 18 megapixel digital single-lens reflex camera

The standard, 18 MP full specimen images allow a visible magnification of about 3X when viewed at 100% on an electronic display. For specimens with important features that are rather small, separate close-up images are captured and associated with the specimen record online. For this, a 10 MP compact camera is used, which allows features to be viewed at a magnification of about 15X. This step is done later in the workflow.

Both the label and specimen image files from the photography step are then saved to a temporary location and processed through a program written by Daryl Lafferty, BarcodeRenamer (BCR). This program uses a scanner application to read each barcode, convert it to text, and then use it as the image file name. Rather than using OCR to read the text below the barcode (Fig. 2), BCR scans the actual barcode using code written specifically for this purpose. Thus, BCR is nearly all-automatic and about 99.5% accurate. The renaming results need to be edited where necessary and verified by a user, but this typically takes only a few minutes for each batch of 500 or so images. Renaming allows the label images to be sorted by barcode number and easily matched up with the specimen image during data entry. Next, the label images are run through ABBYY FineReader as a batch, which produces a Word document with each label separated by a page break. A user is given a Word document and a folder containing the corresponding full specimen images to database. All of the label images are deleted once the OCR process is completed.

The next step in the process is to permanently archive the renamed, but otherwise unaltered, specimen images. They are saved to a large capacity network drive and organized based on geographical location and family name. This file structure mirrors the way the physical herbarium is organized. For example, a folder named “MONOCOTS ETC” corresponds to a room in the ASU herbarium that houses monocots, gymnosperms, and the pteridophytes. Although families are often taxonomically rearranged, the changes are not reflected in the herbarium organizational structure as often. This file structure is unlikely to change much in the next few years, but it would be relatively simple to collapse the family file structure and to use only an organization based on geography. Copies of the images are backed up to an external disk using SyncToy[12], Microsoft’s free file synchronization application.

The final step of the process is to digitally enhance the images to be used online. I use Adobe Lightroom, which has excellent batch processing capabilities. Each family folder of images is imported into the program from the external disk, rotated to vertical, and automatically adjusted for white balance and tone. These adjustments are made on each image individually. They are then exported as JPGs at 10% compression. At ASU, we need to be conservative with our server space, and the 10% compression reduces file size while not visibly affecting the image quality. The compression mainly works on the white space in the

[12] SyncToy (http://www.microsoft.com/en-us/download/details.aspx?id=15155)

image. Lightroom offers a few different ways to import photos for editing. I chose “Add”, which acts like a window into the file location. With this option, Lightroom is pointed to the folder location on the external back-up disk, reads the image information, and creates the thumbnail images that you see while editing. The images are not moved from their location on the external disk. When the editing is finished, the images are exported as copies, and get put into a subfolder named “Web” located within the family folder. A total of three copies are archived: 1) the renamed originals as uncompressed JPGs located on the network drive, 2) back-ups of the renamed originals located on the external disk, and 3) compressed, web-ready versions stored in subfolders on the external disk. A fourth, but relatively insignificant copy is stored temporarily with the OCR text and used during data entry. These copies are not archived. Other than the renaming process, the original images are not altered in any way. The photo editing process is fully automatic and results in high quality, web publishable images. Once the specimen record has been processed using SALIX, the image can be uploaded and associated with the record based on its image barcode (GUID), and is now fully accessible on SEINet.

Rather than taking close-ups of every specimen tied to this project, just one or two good representatives of each species are selected. Of each of these, approximately 1-3 features (e.g., fruit, flower) are selected by a specially trained student. First, an image of only the barcode is taken, followed by the images of the features. Then, a barcode image of the next

specimen is taken, followed by the feature images, and so on. Each of these images is renamed using BCR. With the batch of images sorted in ascending order by timestamp, BCR begins with the first one, recognizes a barcode, and names that image to the barcode plus an “A” prefix, for example, AASU0012345. Subsequent images are renamed to match the barcode, followed by a lowercase letter suffix (e.g., ASU0012345a, ASU0012345b). When the next barcode image is found, BCR stops renaming with suffixes, and names that one AASU0012346 with all subsequent images as ASU0012346a, ASU0012346b, etc. These images are then processed through Lightroom in the same way as the specimen images, and uploaded to SEINet where they are available immediately. The barcode images - those named with an “A” prefix - are deleted.

Optical Character Recognition. – As previously mentioned, SALIX relies on ancillary technologies in order to function optimally. The system can be run with the most basic digital camera and the included open source OCR software, Tesseract[13], but will function at its highest capacity with more advanced tools. I use ABBYY FineReader Professional Edition for OCR processing, which supports documents in multiple languages and automatic batch processing. Several hundreds of label images can be run at a time, exporting the results in a Microsoft Word document with each label separated by a page break. The text results

[13] Tesseract (http://code.google.com/p/tesseract-ocr/)

obtained from FineReader are very good, certainly much better, on average, than what can be produced using Tesseract.

The error rate for the conversion of image to text is highly variable, depending on the quality of the print, and is thus the major bottleneck in the process. FineReader handles older font types and special characters with relative ease – results are dependent more on actual print quality. Faded, crooked, or handwritten labels pose much more of a problem. However, OCR technology is expected to improve with time, and for now, improvements in this data processing system are dependent on these forthcoming advancements.

The renamed label images are saved temporarily to the computer's Desktop, and are then opened through FineReader's Automation Manager. The application runs automatically through the batch of label images and produces a Microsoft Word document at the conclusion of the process[14]. The document is saved to the network drive and then the label images are deleted. Before this portion of the workflow produces consistent results, it would be wise to save the label images, should you need to re-run the OCR. Once consistently acceptable OCR results can be obtained, it is no longer necessary to keep them. The transcription results, in DOCX format, are saved in the same location as a folder of corresponding full specimen images, both of which are used during data processing with SALIX.

[14] A video tutorial showing the settings used at the ASU Herbarium can be found at http://vimeo.com/asuherbarium

Data Processing with SALIX. – With the technology currently available to us, a fully-automated natural language processing system is not possible. As it stands, there are too many errors associated with optical character recognition and automatic parsing for the system to work without an operator. Much of the cause for this lies within the primary data itself – the labels associated with the specimens are too variable in quality for any currently existing OCR software to process with an acceptable margin of error. However, I have found that it is possible to develop a system that is partially automated. SALIX is a user interface-based, Windows executable program written by Daryl Lafferty to handle automatic parsing of label information using machine learning. Label data are copied into a text box within the program window and an algorithm determines which pieces of information belong in which fields.

The algorithm is built on word statistics, compiled from repeated use by users. For example, the words “herb”, “yellow” and “flowers” appear 230, 999, and 687 times respectively, each with a 100% score in the Description field. The words “large” and “with” appear 147 and 474 times, but score less definitively for Description. The SALIX parsing algorithm analyzes the information word for word, and then combines the scores to come up with a score for the entire line. So a phrase such as, “Large herb with yellow flowers” would be analyzed word for word with the following results (Table 1), and then would be parsed into the Description field.

Table 1. An example of the word statistics used by SALIX in its parsing algorithm. Each word is scored individually, and then a total score for the line is calculated. Words with higher counts carry more weight.

Word	Count	Description	Locality	Habitat
flowers	687	100	0	0
herb	230	100	0	0
large	147	37	6	56
with	747	82	17	0
yellow	999	100	0	0

Parsing improves with use, as each confirmed addition to the database contributes to the word statistics. However, herbarium specimen labels are variable. Phrases such as, “Common shrub growing along the roadsides of Hwy 2 in Pitiquito, Sonora” are difficult for automatic parsing, and perhaps difficult even for the operator. Before each record is exported to the database, the accuracy of the parsing should be verified by a user. Parsing also improves when the user separates blocks of information by a new line in the Word document (Fig. 3). In the figure below, it can be seen how the locality, habitat, and description, are separated by line breaks. SALIX is programmed to consider this along with word statistics during parsing.

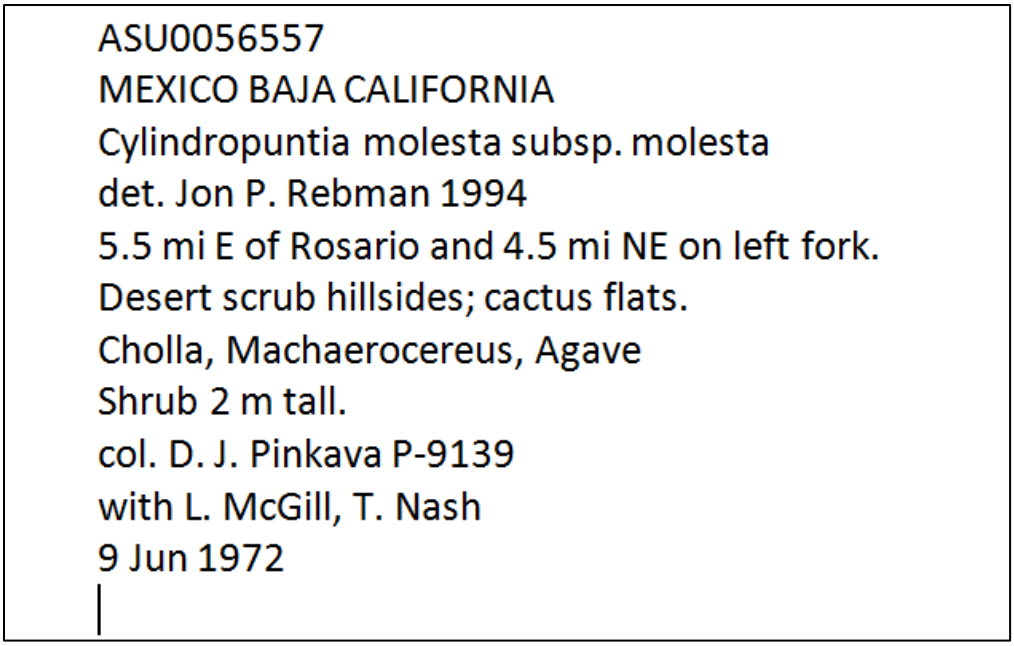


Fig. 3. An example of label text after editing. The locality, description, and habitat fields are separated by a line break. The determiner, collector, and associated collectors are prefaced by start words.

Collectors, associated collectors, and determiners also pose a problem in automatic parsing because any of the names could logically go into any of the three fields. To solve this problem, SALIX is programmed to recognize where names belong based on what are called “start words”. The collector is prefixed by “col.”, associated collectors by “with”, and the determiner by “det.” (Fig. **3**). The start words can be modified to fit any user’s needs, and multiple start words for one field are also permissible. For example, a user could enter in the following list of start words for the collector field: col, coll, colector, collector, collected by and SALIX would look for all of those when parsing to the collector field. This can be set up in the Tools menu under Field Definitions, and is customizable on a per user basis.

There are a couple systems in place that check the accuracy of the data before it gets exported. A problem faced early on was that misspelled, unpublished, or otherwise incorrect scientific names were being added to the taxonomic database. To fix this, an easy way of verifying new name additions was needed. The existing authority files are loaded into the SALIX program files, and those names appear in the taxa drop down menus. When a new name gets entered, SALIX opens a browser window and searches for the name on Tropicos[15]. If the name is found, it automatically approves the addition and the record gets exported. If the name is misspelled or otherwise incorrect, the user will need to find the right name. For example, say the specimen in question is labeled *Lupinus sparsiflora* rather than *Lupinus sparsiflorus*. SALIX would throw up an error message saying that the name was not found. The user would then begin searching Tropicos and would find *Lupinus sparsiflorus* and an author name matching the one on the label. Clicking on the correct name would bring up a message requesting verification, the user would verify, and the record would be exported to the CSV file.

Also built into the SALIX functions is a system for verifying the accuracy of geographic coordinates. A program file was built that contains geographic bounding boxes for all of the Latin American countries and some of the states of Mexico. When coordinates are present in a specimen record, SALIX checks those against this library during export. If

[15] A public database of nomenclatural specimen data, hosted at the Missouri Botanical Garden (http://www.tropicos.org/)

the coordinates fall outside of the geographic limits, an error is thrown up and the user must check the data against the specimen image. This greatly reduces the amount of georeference errors in the database and improves the reliability of the data.

SALIX also includes standardized drop-down menus for the names of countries and states/provinces, compiled by the project manager. This ensures that the correct spellings will be used, and that the correct states/provinces will be matched up with their corresponding countries. For example, a specimen collected in the state of Arizona, in the country of Mexico would not be allowed. Some counties are included for certain states in the USA, but since this project mainly includes only Latin American specimens, this list is not comprehensive. There are functions built into the SALIX user interface that allow easy editing and addition of political-geographical information.

The general workflow for using SALIX is fairly simple. The user begins by opening the first specimen image in a folder, comparing it to the text on the first page of the Word document, and then starts editing. It is recommended that the user remove any information that is irrelevant or unnecessary, such as the names of herbaria, so as to simplify the automatic parsing. When OCR results are very poor, label information is typed by hand into the Word document, copied into SALIX, and parsed as usual. The user should also look over the entire specimen image so as to not miss any important information, such as annotation labels or

accession numbers. A typical label before editing is represented in Fig. 4, followed by the edited text ready for SALIX in Fig. **5**. The order in which the information is presented is not important, but some users find it helpful to have a loose structure to follow. For example, in Fig. **5** you can see that the country and state were changed to all caps and moved to the top of the page. This makes it easier to verify that the information was parsed into SALIX correctly – you always look in the same spot, rather than scanning through the entire block of text.

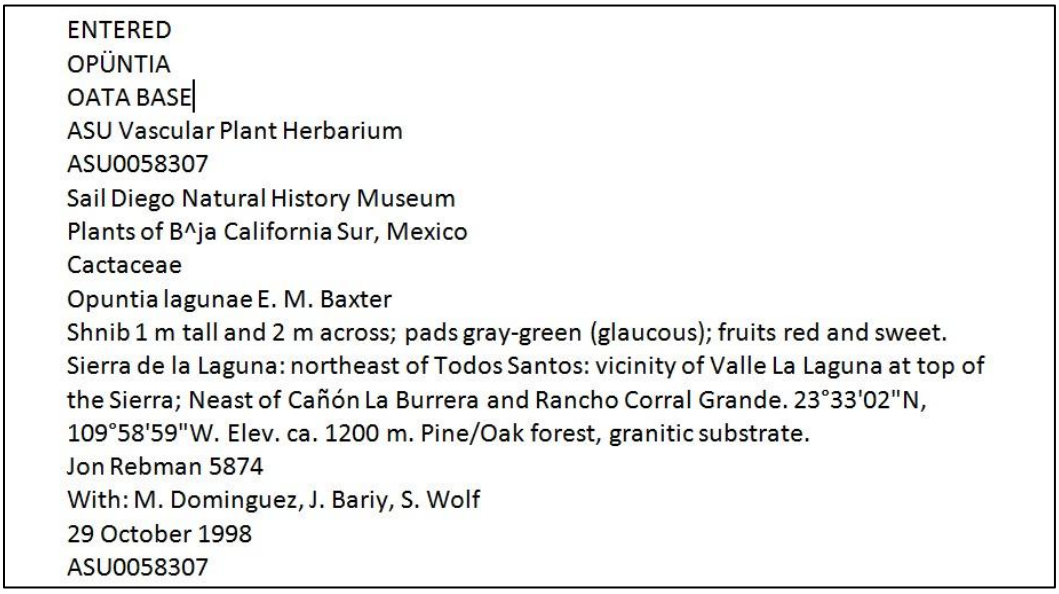

ENTERED
OPÜNTIA
OATA BASE
ASU Vascular Plant Herbarium
ASU0058307
Sail Diego Natural History Museum
Plants of B^ja California Sur, Mexico
Cactaceae
Opuntia lagunae E. M. Baxter
Shnib 1 m tall and 2 m across; pads gray-green (glaucous); fruits red and sweet.
Sierra de la Laguna: northeast of Todos Santos: vicinity of Valle La Laguna at top of the Sierra; Neast of Cañón La Burrera and Rancho Corral Grande. 23°33'02"N, 109°58'59"W. Elev. ca. 1200 m. Pine/Oak forest, granitic substrate.
Jon Rebman 5874
With: M. Dominguez, J. Bariy, S. Wolf
29 October 1998
ASU0058307

Fig. 4. An example of label text as produced by FineReader, before any editing has taken place.

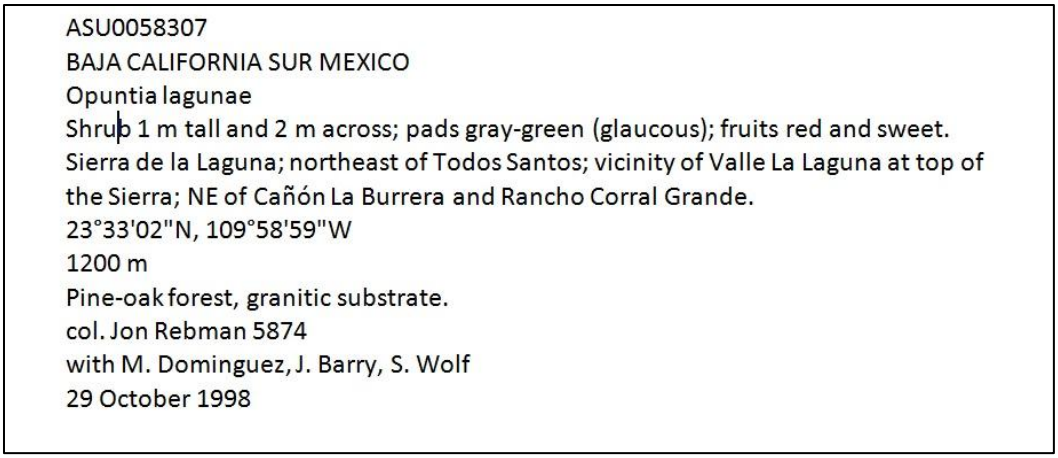


Fig. 5. An example of label text after editing. This process is the major bottleneck in the workflow. It is recommended to remove unnecessary information from the original text. In this case, that includes the name of the former owning institution (SDNHM) and the family name (Cactaceae). SALIX knows to which families species belong.

Once the block of text is edited, it is copied and pasted into the SALIX text box, and the user pushes the “Parse” button. The SALIX parsing algorithm is run, and the information gets moved to the appropriate fields. Further editing may be necessary, but it is minimal. In the example in Fig. 6, you can see the record information in SALIX after parsing but without any adjustments. The label text was all parsed correctly. The only field left unfilled is the accession (highlighted in red), which gets typed in by hand from the specimen image. Lastly, the “Export” button is pushed, and the label data are stored in a DarwinCore-compatible CSV file. Each new record that is exported from SALIX is added to this file, and at the end of the user’s shift, the file is uploaded to SEINet where it is immediately made public.

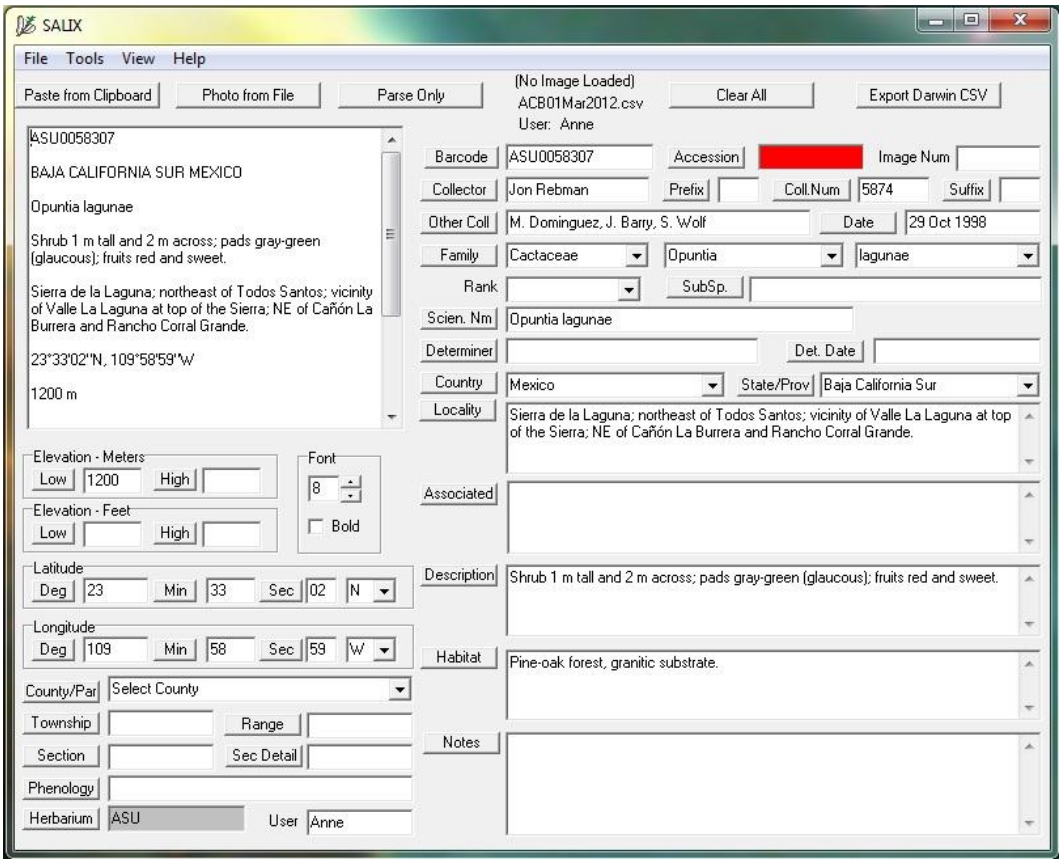


Fig. 6. A screenshot of the SALIX user interface. This shows the label text after pasting into the text box and parsing.

Workflow Summary. – The digitization process consists of two main workflows – the data processing done with SALIX and the project management. In the SALIX workflow (Fig. 7; right column), the OCR editing is the major bottleneck in terms of efficiency. Most of the time devoted to databasing is spent preparing the text for SALIX (Step 1). Steps 2-6 can be accomplished very quickly. The user begins by editing the OCR results (Step 1), and then copies and pastes the label text into the SALIX text box (Step 2). The user then pushes the Parse button (Step 3), and verifies that this was performed correctly by the parsing algorithm (Step 4). Once the label information is correct, the user pushes the Export

button and the label data get stored in a CSV file (Step 5). At the end of the shift, the user sends the CSV to the project manager, and it then gets uploaded to SEINet (Step 6).

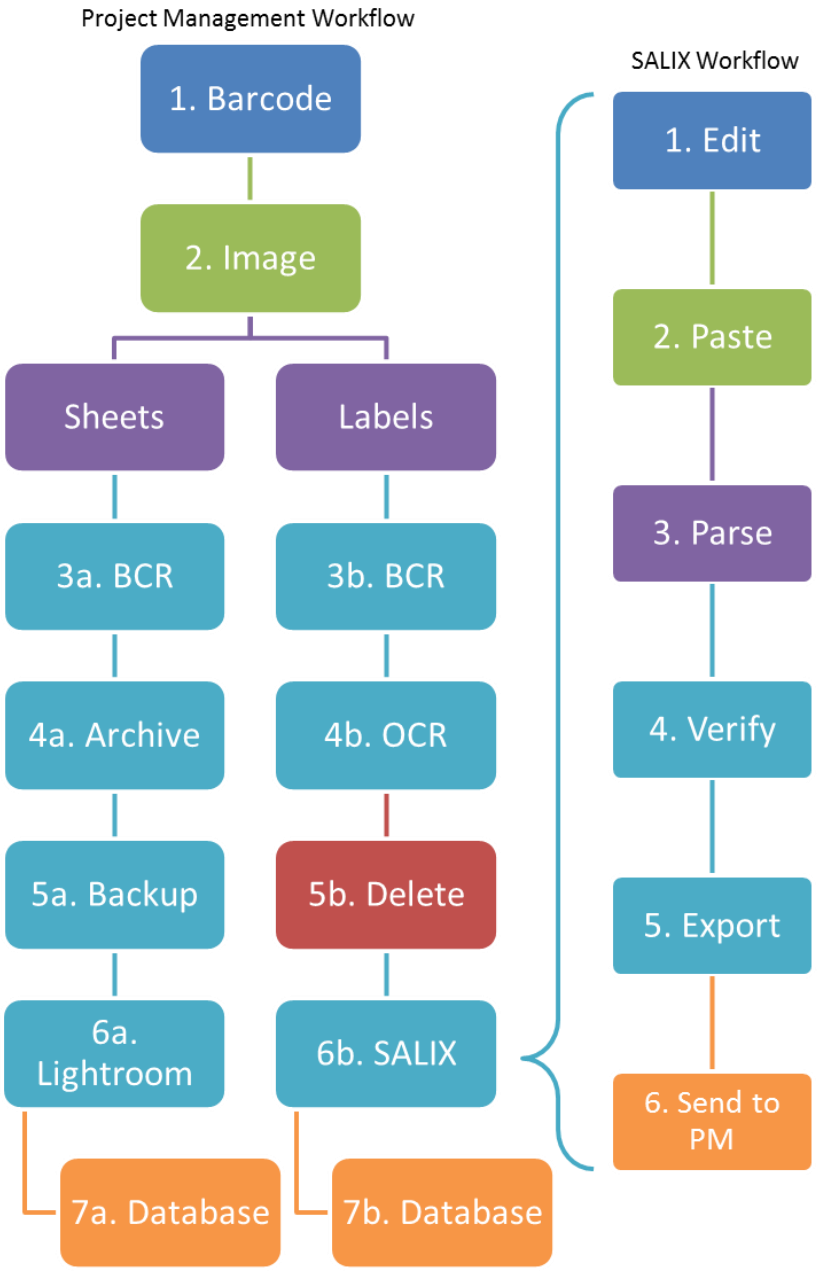


Fig. 7. The SALIX Method workflow. The left column outlines the project management portion of the process. The right column outlines the portion of the workflow performed by using SALIX.

The project management workflow (Fig. 7; left column) is a little more complex. The barcoding and imaging in Steps 1-2 are done by students, but the initial camera set up is done by the project manager. After the photographing is finished, the project manager starts processing the image files. Specimen and label images are saved in separate folders and then run through BCR (Steps 3a and 3b). After BCR is complete, the memory cards containing the original, unrenamed images are cleared. The renamed specimen images are archived on the network drive (Step 4a), while the renamed label images are run through OCR (Step 4b). Once OCR is finished, the resulting Word document is saved to the network drive and the renamed label images can be deleted (Step 5b). Student workers can now be given a Word document containing label text for editing, paired with a folder of corresponding specimen images (Step 6b). These folders of specimen images are copies of the archived versions that can be deleted once they are databased. When a user has finished databasing for the day, the CSV file is given to the project manager and uploaded to SEINet (Step 7b).

The archived specimen images are backed up to an external hard drive (Step 5a), where they are accessed by Lightroom (Step 6a). As previously mentioned, Lightroom does not move or copy these files, but works directly from the source folders on the external disk. The Lightroom catalog is stored here too, and this catalog is what contains all of the photo editing information. If you do some editing on an image, and then

close Lightroom, it doesn't change the original image at all. It stores the sequence of photo editing events that you performed in the catalog, and then shows you a thumbnail representation of the edited image. When the editing is finished, the specimen images are exported as web-ready copies and saved to a subfolder in the original folder from which they came. At this point, they can now be uploaded to SEINet (Step 7a), where they will be matched up, based on GUID, with the textual data from Step 6 in the SALIX workflow (Fig. 7; right column).

Additional project management is handled through SEINet, which includes finding missing data and routine data cleaning. Occasionally, a record without an image will exist in the database, or there will be images without collection data. These two scenarios need to be resolved by the project manager.

The workflow for capturing close-ups is fairly simple (Fig. 8). First, one or two good specimens are chosen for each species (Step 1). These specimens have already been barcoded. The user places the specimen on a specialized platform for close-up imaging, and takes a photograph of just the barcode (Step 2). Then, the user chooses 1-3 taxonomically important features and photographs each one separately (Step 3). This process is repeated for subsequent specimens. When the imaging is completed, the camera card is given to the project manager, and BCR is run (Step 4). After all of the images have been renamed to match the barcode, they are archived on the network drive (Step 5). Finally, the

close-ups are uploaded to SEINet and automatically associated with the specimen record online (Step 6). As previously mentioned, this association is done based on the GUID – an existing specimen record in the database will have the GUID listed in the catalog number field, and SEINet will use this and the image filename to match up both pieces.

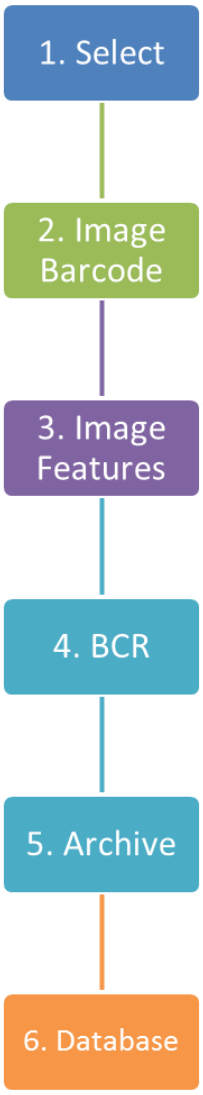


Fig. 8. The workflow for capturing close-up images. This is performed after the SALIX Method workflow.

Digital Field Guides. – The Southwest Environmental Information Network (SEINet) began as a way to integrate the vascular plant databases at Arizona State University, but now includes collections from all over the Southwest at 28 different institutions. A fairly new feature added to this web-accessible database is the ability for users to upload what is referred to as Photo Observations. These are photographs of plants taken in the field, and are associated with minimal collection data. Although these observations are no substitute for real specimens, they can provide valuable distributional information and are relatively easy to collect and share. A collection of images along with their associated geographic coordinates can also be made into a keyhole markup language (KML) file and viewed on Google Earth. Or, a dataset can be compiled within SEINet as a floristic project and viewed on Google Maps, with each marker a link to the record and associated image.

To demonstrate the benefits of this new capability, and also to serve as a sort of digital field guide for the ASU Arboretum, a portion of the living collection on campus has been cataloged as a vouchered checklist consisting of photo observations. Prior to this, a visitor would need to search the ASU website for a PDF map of the Arboretum trails to print and use in a self-guided tour. Now, users can follow links on SEINet, Facebook, Twitter, Panoramio and the ASU website that will lead them to the SEINet floristic project or the KML download on Panoramio. The KML can be viewed on mobile devices running Google Earth, or through the

device's map application. In order for this information to be viewed properly, the images need to contain geographic coordinates in the metadata – this is what is called "geotagging".

Few professional level digital cameras have GPS capabilities. Perhaps within the next several years this will become more of a standard technology, but for the present, the options are limited to a handful of point and shoots. Fortunately, there are many other good choices for easy geotagging with any digital camera. A wide variety of both adaptable and pocketable GPS trackers are available. The adaptable kinds work by connecting to the camera's hotshoe and immediately writing geographic data to the image metadata. The pocketable units record a continual stream of geographic information as you move, and then later sync[16] those data to your image files through a computer. Both types cost between $50 and $200, but the hotshoe adapters have the disadvantage of being huge drains on the camera's battery life. If you already have an iPhone or Android device, there are several good apps available for under $5 that work in the same way as a pocketable GPS tracker. These use the GPS capabilities of your phone to record continuous geographic data, and then sync them to your image files based on timestamps. Like the hotshoe adapters, these apps use a lot of battery power and will quickly drain your mobile device. Alternatively, the built-in cameras on most smartphones are becoming quite advanced, and these images are geotagged by

[16] From synchronize; the automatic process of updating or merging data from one device to another

default. If the purpose is to quickly and easily compile a geotagged image library, this would be the most effective approach.

The ASU Arboretum Digital Field Guide was compiled using a 15 MP DSLR and the app Geotag Photos Pro for the iPhone. There are two ways to access the field guide – first, through a vouchered checklist on SEINet[17], and second, through the map-based, photo-sharing site, Panoramio[18]. A vouchered checklist is one in which each species name on the list is linked directly to one or more specimen records – in this case, a photo observation of a living specimen. Navigating to the ASU Arboretum Digital Field Guide checklist page will show a list of the species as links that take you to the species profile page. This page consists of all available images and text descriptions for that species. Checking the box next to Notes & Vouchers and clicking on Rebuild List will display links to the actual photo observation records. Clicking on the map to the right of the checklist will open the Google Maps view and will allow the user to zoom in and change map layers. Each observation is displayed as a marker; clicking on one of these will open up the photo observation record and its associated image. At the top of the checklist page is a golden key icon; clicking on this will open up the interactive key application for the checklist. There is also a link in the checklist description to the Panoramio site.

[17] ASU Arboretum Digital Field Guide on SEINet (http://swbiodiversity.org/seinet/checklists/checklist.php?cl=2677&showvouchers=1)

[18] ASU Arboretum Digital Field Guide on Panoramio (http://www.panoramio.com/user/4926084)

Panoramio is a photo-sharing site developed by Google, with a focus on mapping. Geo-tagged images are evaluated for use in Google Earth, and if approved, show up publicly within 24 hours. Google Earth does not officially accept plant macros, but images of tree and shrub habits are often approved and can be seen by anyone using Google Earth to look around campus. A KML file is automatically generated for the user's entire collection, so that even rejected images can be viewed on a user by user basis. Currently, the Arboretum collection has had 4,979 views in Google Earth and 7,347 views overall. The benefit of using Panoramio is that it perfectly complements the more scientific-based collection on SEINet. It has more of a public audience, and is simpler to navigate for people unfamiliar with herbaria or SEINet. There are also easy links that work well for viewing on mobile devices.

CHAPTER 3

RESULTS

SALIX vs. Typing – Mixed Labels. – The results show that the average rate of data entry using the SALIX Method is 26.3 records/hr for Experts and 18.0 records/hr for Novices (Fig. 9). Experts are defined as having used the system for 31 days or more, while Novices are defined as having used the system for 1-30 days. Each day consists of a work shift ranging from 0.35-5.5 hours, or 2.25 hours on average. After 31 days, the average Novice will have worked 65.7 hours. The target rate (depicted in the following figures as “Target”) is defined as the average speed of data entry by Experts. Filing, barcoding, and photographing can be accomplished at an average rate of 100 records/hr for one person. The image post-processing steps and uploading are handled by the project manager and were not timed. The majority of these tasks are nearly all automatic and do not take any significant amount of time to accomplish. Rather, the main job of the project manager is to find and correct errors, manage student workers, and develop workflow techniques.

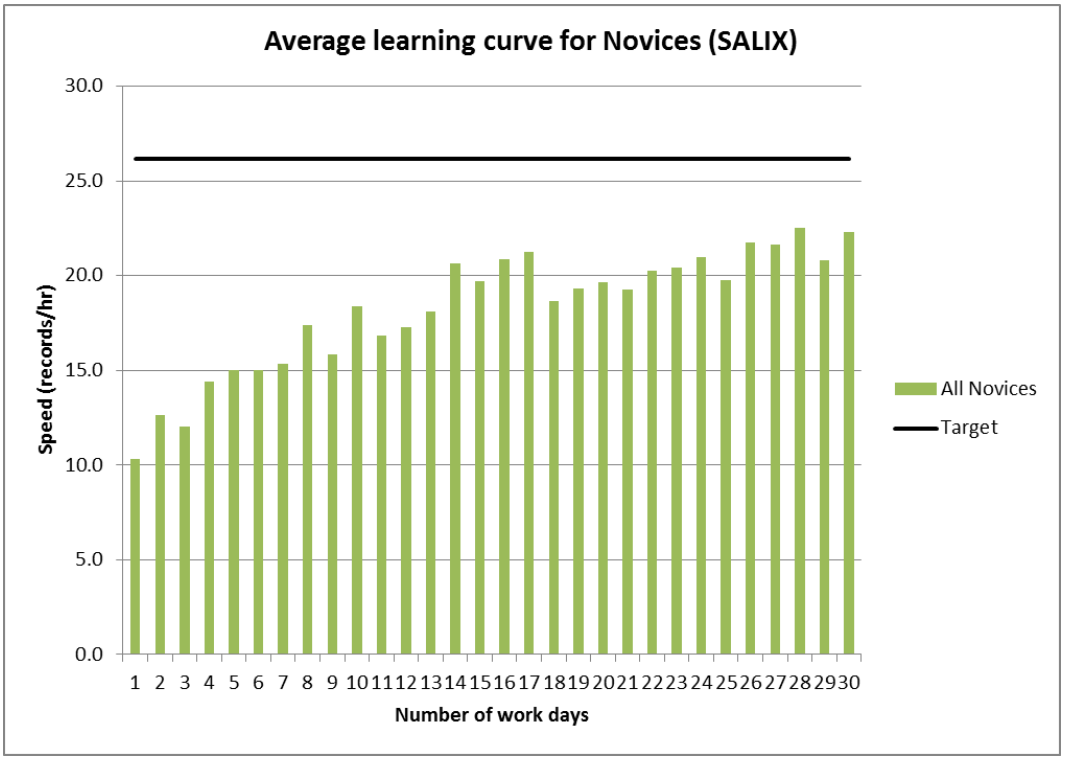


Fig. 9. Average rate of data entry for Novices using SALIX with mixed labels. This shows the general learning curve for new users. The data were collected from 6 different Novice users. The line marked “Target” is the average rate of data entry for Experts using SALIX.

Speeds and learning curves are highly variable and user-dependent (Fig. 10). Some individuals possess a natural ability to navigate through the complex world of natural history data, while others find it difficult to untangle the mess of OCR results and fit it into the orderly structure demanded by a database. Some Novices quickly reach the target rate and even exceed it, while others struggle throughout the first 30 days and reach a plateau early on (Fig. 10). All of the users, save myself, are undergraduate students with limited knowledge of botany and botanical collections.

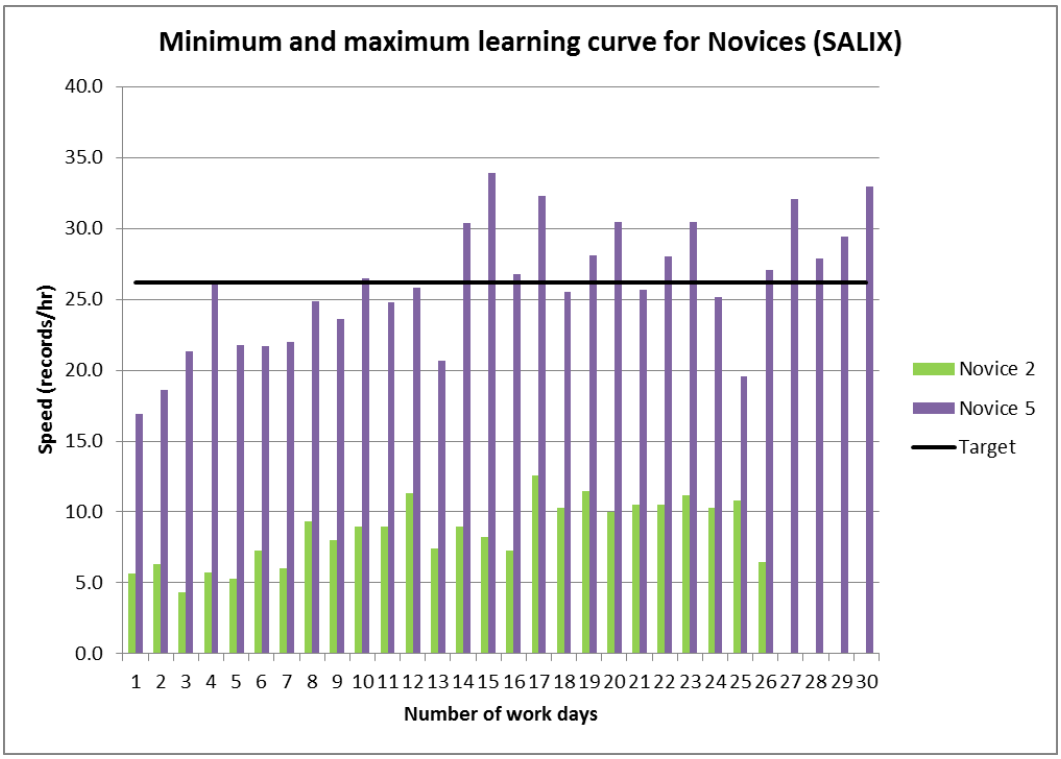


Fig. 10. Highest and lowest rates of data entry for Novices using SALIX with mixed labels. This shows the extreme differences in user proficiency. The data were collected from 2 different Novice users. The line marked "Target" is the average rate of data entry for Experts using SALIX.

In order to determine whether or not the SALIX Method is advantageous in terms of speed, data entry rates were collected by measuring the speed of Experts using the SEINet interface. The method of data entry in the online system is likely similar to many other institutions, in which data are entered by the user cell-by-cell while reading label information (Fig. 11). Instead of reading data from a physical label, users read from an image of the specimen that opens next to the empty record.

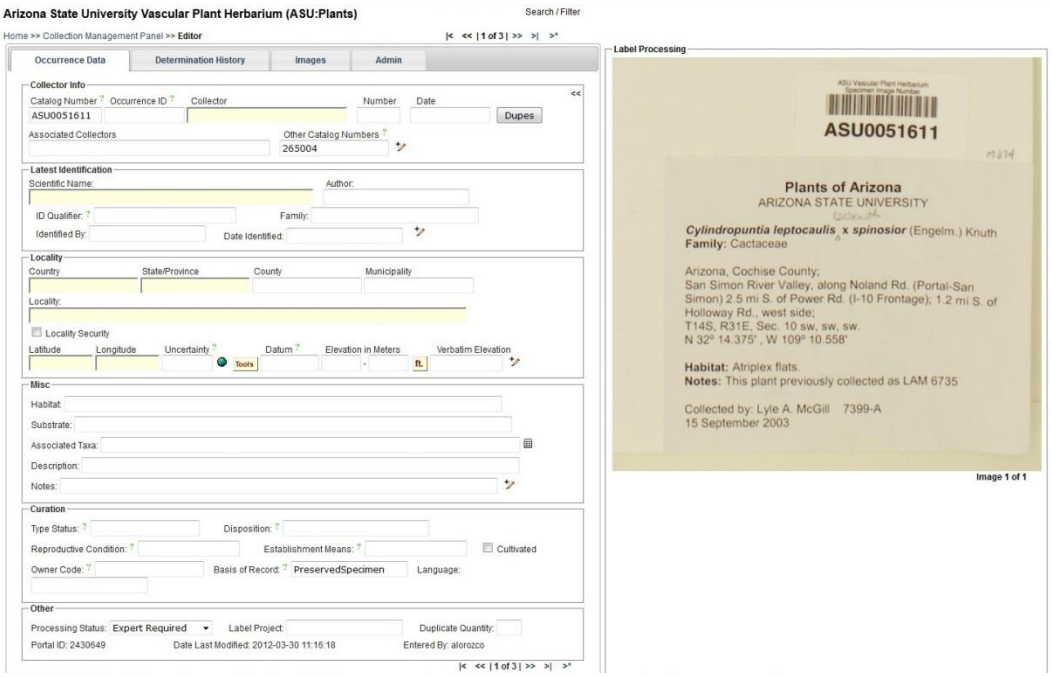


Fig. 11. An example of the data entry form on SEINet. This was used to measure the speed of data entry by typing.

The average rate of keystroke data entry for Experts using SEINet is 20.4 records/hr, compared to the faster 26.3 records/hr that can be achieved using the SALIX Method (Fig. 12).

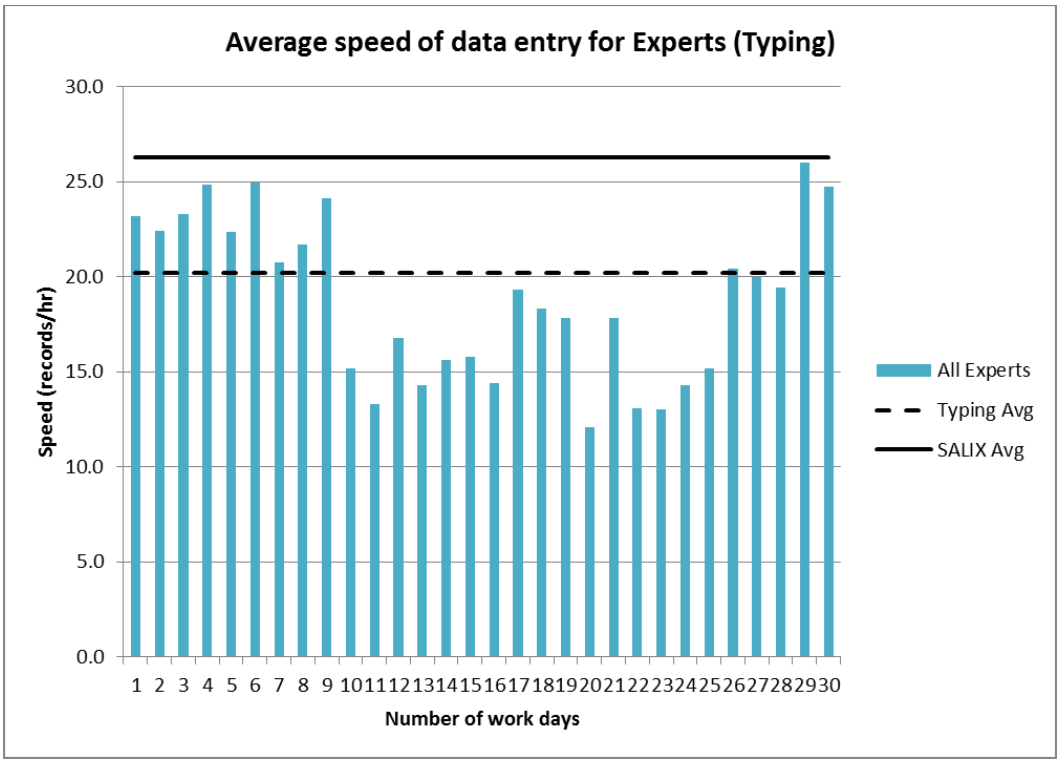


Fig. 12. Average rate of data entry for Experts by typing with mixed labels. This shows the significant speed advantage to using SALIX over typing. The data were collected from 4 different Expert users. The line marked "SALIX Avg" is the average rate of data entry for Experts using SALIX; the line marked "Typing Avg" is the average rate of data entry for Experts using typing.

SALIX vs. Typing – By Label Type. – A more detailed analysis was conducted in terms of label quality and length. Four separate datasets were compiled, with each set consisting of 10 label images (Table 2). Each dataset was used twice by each Expert user – once using SALIX, and once through typing into SEINet (Fig. 11).

Table 2. The dataset for speed analysis by label type. Label type includes both length in words and quality. The average length for Long labels is 127.4 words/label; the average length for Short labels is 44.1 words/label.

Length	**Quality**	
	Good	Poor
Long	10	10
Short	10	10

Fig. 13 illustrates the average number of words per label for the different label types in Table 2. The line labeled “Avg” shows the average length of 2,867 mixed labels, which were chosen at random from the archive. The sets of labels were run through FineReader, and the words were counted in Microsoft Word by dividing the entire word count by the number of pages in the document. The text was not edited before the words were counted.

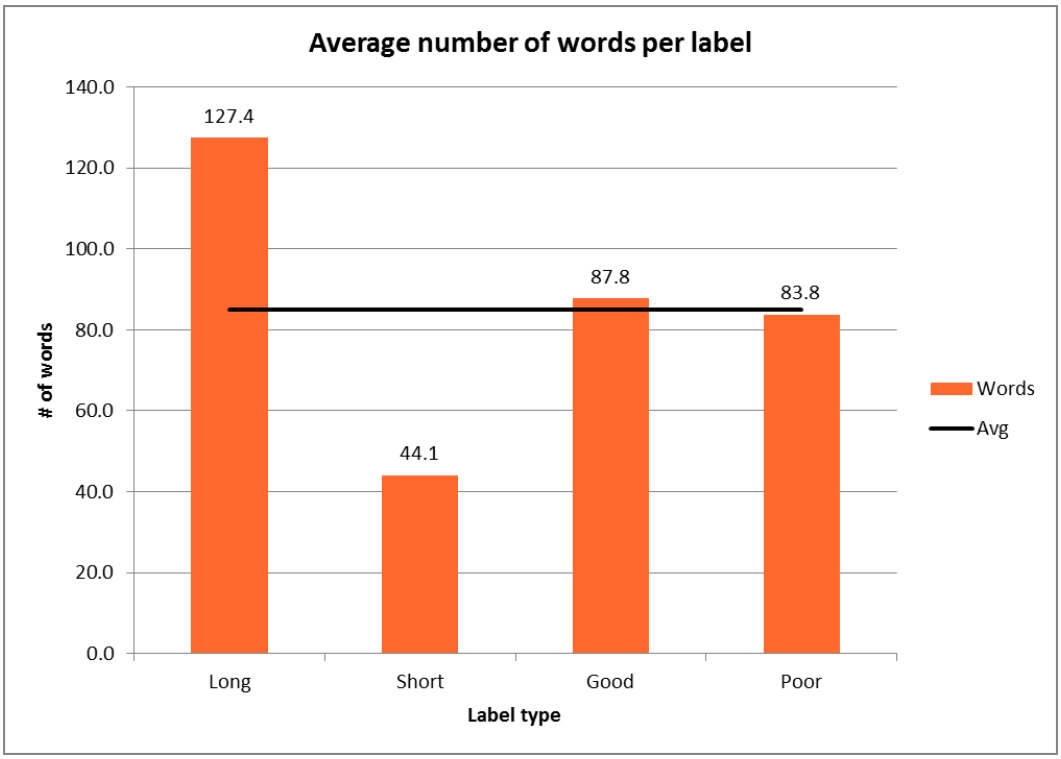
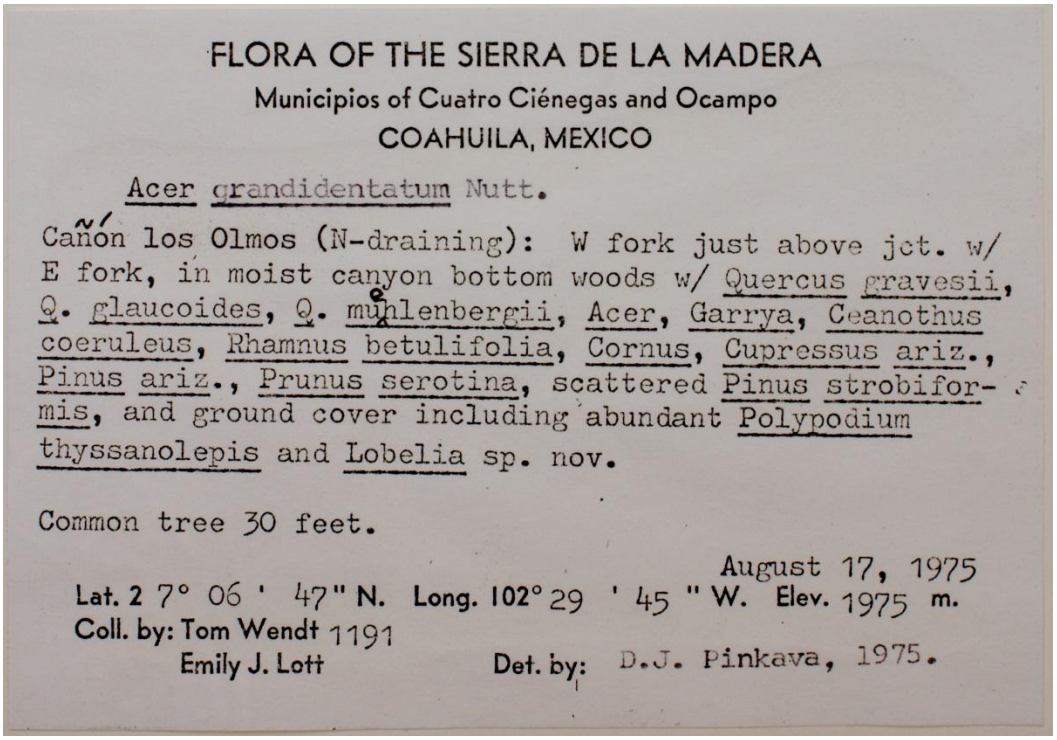


Fig. 13. Average number of words per label type in the test dataset (Table 2). This was calculated to assure that the test dataset was representative of all labels in the project. The line marked “Avg” is the average length for all label types, compiled from 2,867 randomly selected labels.

Labels that qualified as “Poor” in terms of quality typically consisted of faded ink, typewritten or handwritten information, and crooked lines (Fig. 14). Poor labels typically produced OCR results with only 10-20% accuracy. Labels qualifying as “Good” were generally newer and produced using modern word processing and printing technology (Fig. 15). Good labels typically produced OCR results with up to 80-90% accuracy.



FLORA OF THE SIERRA DE LA MADERA

Municipios of Cuatro Ciénegas and Ocampo

COAHUILA, MEXICO

Acer grandidentatum Nutt.

Cañón los Olmos (N-draining): W fork just above jct. w/ E fork, in moist canyon bottom woods w/ Quercus gravesii, Q. glaucoides, Q. muhlenbergii, Acer, Garrya, Ceanothus coeruleus, Rhamnus betulifolia, Cornus, Cupressus ariz., Pinus ariz., Prunus serotina, scattered Pinus strobiformis, and ground cover including abundant Polypodium thyssanolepis and Lobelia sp. nov.

Common tree 30 feet.

August 17, 1975

Lat. 27° 06 ' 47" N. Long. 102° 29 ' 45 " W. Elev. 1975 m.

Coll. by: Tom Wendt 1191

Emily J. Lott

Det. by: D.J. Pinkava, 1975.

Fig. 14. An example of a label designated as “Poor” quality. This label was probably produced from a typewriter. The print is faded, crooked, and includes some handwriting.

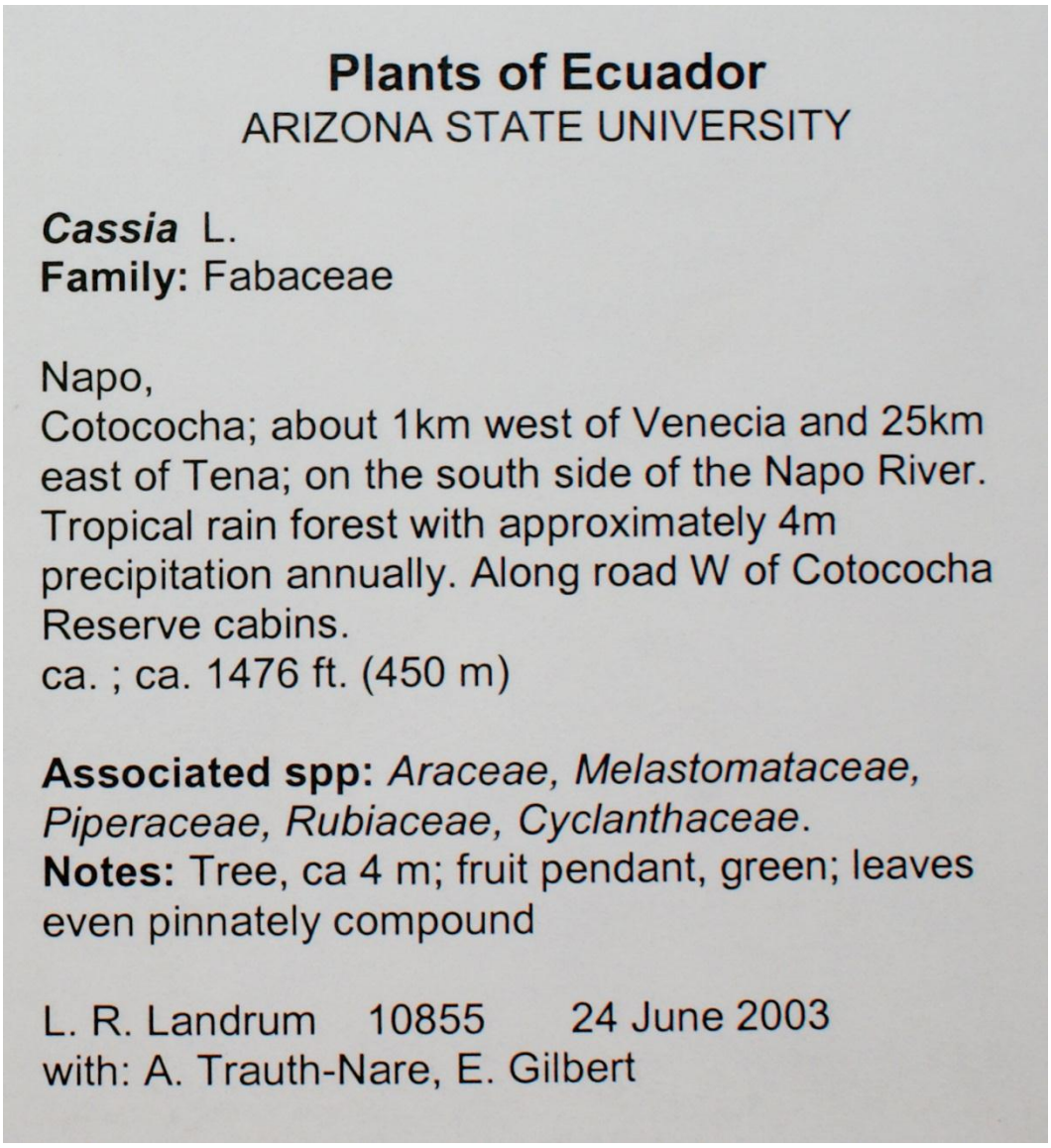


Fig. 15. An example of a label designated as “Good” quality. This label was produced from a modern computer and printer. The print is clear, straight, and not faded.

Consistent with the results obtained using mixed labels, speed according to label type is also highly user-dependent. Measuring SALIX data entry speed was done first (Fig. 16).

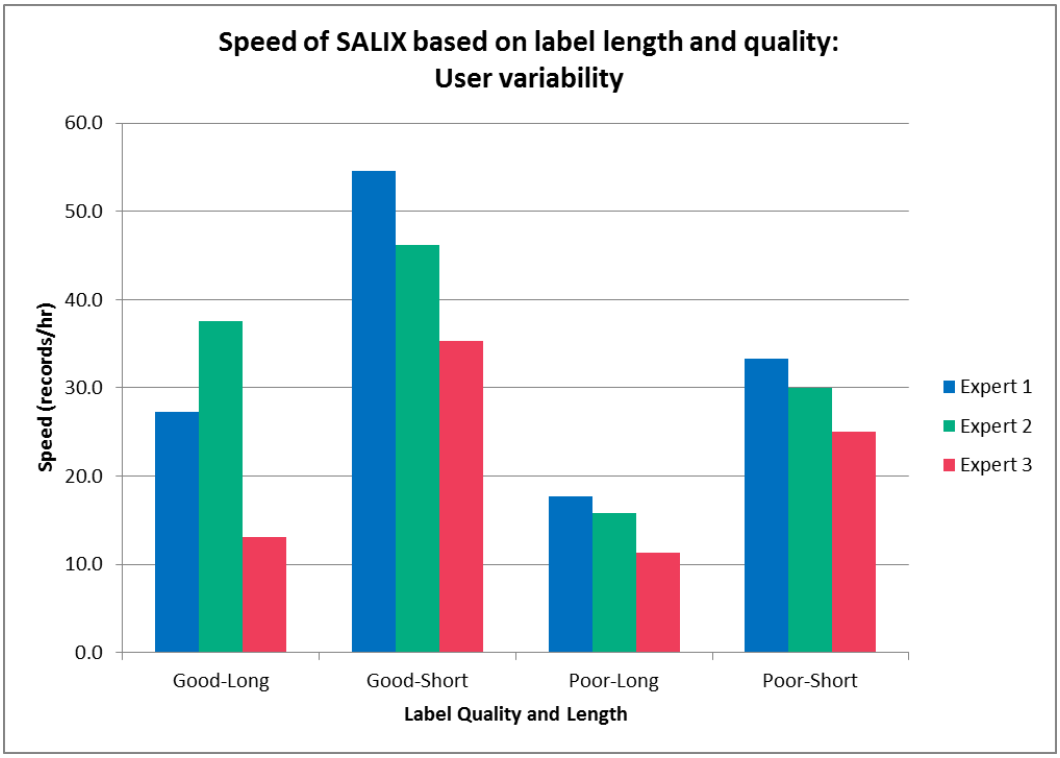


Fig. 16. Rate of data entry for Experts using SALIX, based on label type. This shows the differences in user proficiency. The data were collected from 3 different Expert users.

Next, data entry speed by typing was measured (Fig. 17). There was a timespan of about 1-2 weeks between the SALIX and typing tests, so users probably did not remember the label data well enough to affect the results.

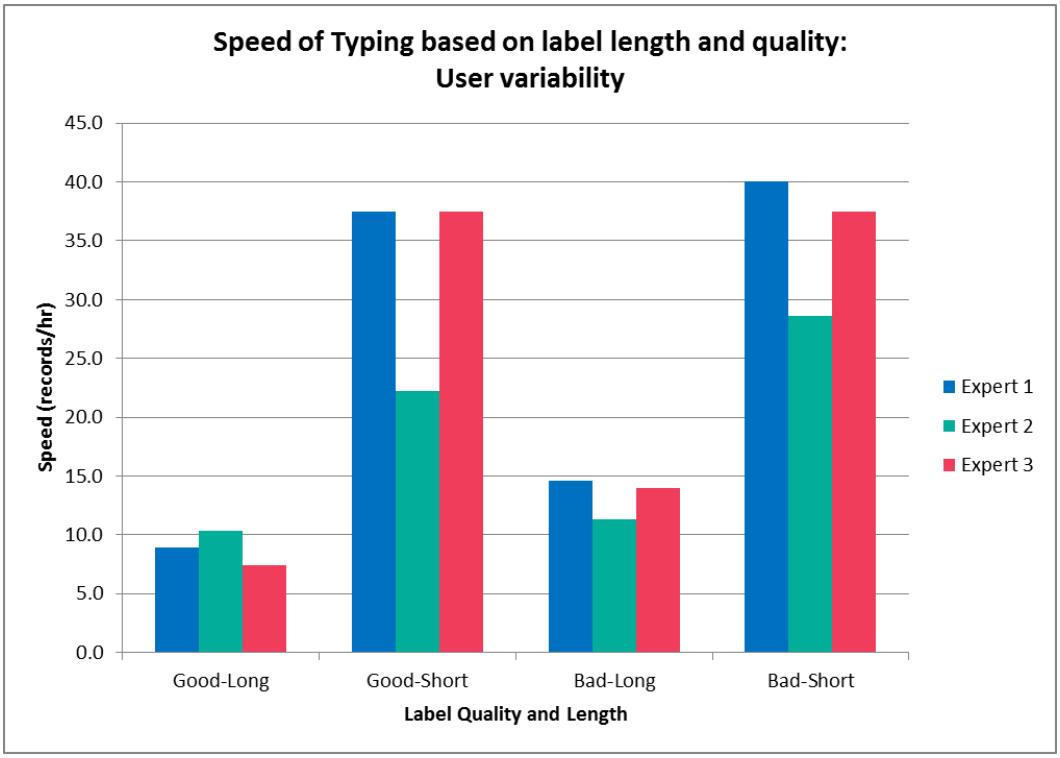


Fig. 17. Rate of data entry for Experts using typing, based on label type. This shows the differences in user proficiency. The data were collected from 3 different Expert users.

If we look at a bar chart showing user averages, the trend is easy to see (Fig. 18). When labels are long and are of good quality, SALIX is approximately 3 times faster than typing. SALIX is approximately 1.5 times faster than typing when labels are short and of good quality. There is little to no advantage of using SALIX over typing when label quality is poor, regardless of length.

Fig. 18. Average rate of data entry for Experts using SALIX, based on label type. This shows that using SALIX is much faster than typing when label quality is good. There is little to no difference in using SALIX over typing when label quality is poor, regardless of length. The data were collected from 3 different Expert users.

User Experience. – Users report that SALIX is more enjoyable than typing directly into database fields. Student workers have expressed that using SALIX is like solving a puzzle, is more engaging, and even relaxing. Depending on the individual, spending up to 3 or 4 hours at a time in front of the computer doing data entry does not appear to be as tedious as it once was. However, this feedback is also highly user-dependent. I have found that in order to achieve fast data processing with the minimal amount of fatigue, finding the right person for the job is the key.

Digital Field Guides. – At the conclusion of this project, the ASU Arboretum Digital Field Guide consists of 142 photo observations and 115 species. On Panoramio, the collection has been viewed 7,347 times in the last 16 months, with 4,979 of the views coming from Google Earth (Fig. 19). The number of views on SEINet is unknown.

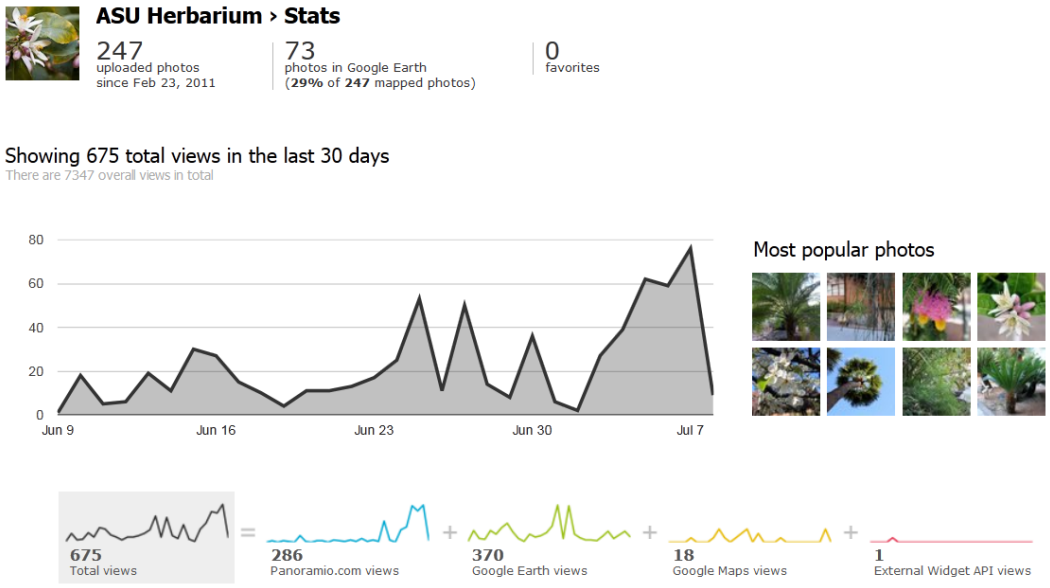


Fig. 19. Google-generated statistics of Panoramio views. This shows the number of times the ASU Arboretum Digital Field Guide was accessed on Panoramio, Google Earth, or Google Maps in the last 30 days.

CHAPTER 4
DISCUSSION

Optimization. – The results of this project show that it is possible to digitize herbarium specimens at a faster rate with the aid of optical character recognition and automatic parsing. However, the speed of data entry is highly user dependent and varies depending on label quality, and to a lesser degree, label length. Controlling for label variability may be impossible, or at least inconvenient. Rather, I recommend moving through a botanical collection from start to finish, without taking the time to sort out the best candidates for OCR. If some labels produce unusable transcription results, it is simple to incorporate them within the workflow. Instead, I recommend taking time to hire the most suitable people for doing the work. Thorough interview and training processes are imperative. The right person might not be the most obvious – students with backgrounds in Art, Psychology, or Literature should not be passed over in favor of those with experience in Biology or Computer Science. I found that the best workers were methodical, detail oriented, and self-motivated, regardless of academic major or background.

In addition to finding the right personnel for data entry, I found that the digitization project required day to day supervision. As project manager, I worked with the curator and programmer to establish a functional workflow, and supervised the majority of the project execution. This included the entire imaging process, from station set-up to image

post-processing and archiving, to quality control and the training of student workers.

Although I have found that the SALIX Method has significant speed advantages over typing, I am not able to compare these results with another, similar project. Specific speed data are currently unavailable for HERBIS and Apiary, the two projects that most closely resemble the SALIX Method.

Contemporaries. – HERBIS[19] was developed as a web-based tool for automated text extraction only; it does not directly incorporate OCR engines or include a user interface. Rather, users first submit specimen images to an off-site OCR processing location, formerly at the Yale Peabody Museum. The transcription results are then forwarded to another location, where they are processed by HERBIS and parsed into a DarwinCore XML file. As previously mentioned, the HERBIS algorithms are based on label formatting, and use supervised machine learning to assist in correct XML markup.

Apiary[20] is a web-based application that utilizes HERBIS and a variety of open source OCR engines. The user interface consists of integrated data entry forms, specimen images, OCR processing, text parsing, and data exporting features. The workflow begins with the manual identification of regions of interest (ROIs) – typically labels or barcodes

[19] HERBIS (http://www.herbis.org)

[20] The Apiary Project (http://www.apiaryproject.org)

that need to be individually identified by a user. Next, OCR is run by OCRAD[21], OCRopus[22], and GOCR[23] and the user is able to choose the best results or type by hand. Typically, the text results are not good enough to use, so most of the transcription is accomplished by hand. Parsing is then presumably automated by HERBIS, although it is not currently functional. Instead, parsing is performed manually by a user, and then the record is exported to the database at the Botanical Research Institute of Texas (BRIT).

The SALIX Method shares some similarities with both HERBIS and Apiary. All three use OCR to varying degrees of efficiency, as well as automatic parsing. Like HERBIS, it is possible to export data from SALIX as DarwinCore XML, although DarwinCore CSVs are used because the database on SEINet is built to work with these file types. Like Apiary, the SALIX Method utilizes a user interface and a specialized workflow to digitize herbarium specimens.

However, what makes SALIX unique is its approach to natural language processing, in which parsing is based on compiled word statistics that improve with use. SALIX uses machine learning to automatically recognize the semantic patterns in label data, and then decides in which field the pieces of data belong. Of course, HERBIS does

[21] OCRAD (http://www.gnu.org/software/ocrad/)

[22] OCRopus (http://code.google.com/p/ocropus/)

[23] GOCR (http://jocr.sourceforge.net/)

this too, but recognizes patterns based on label format rather than actual textual data. Additionally, SALIX includes a specialized user interface (UI) that allows users to edit settings and text, and create custom user profiles. I have also developed an efficient workflow that uses high end optical character recognition, digital photography, image post-processing, and project management techniques – together, what is referred to as the SALIX Method.

Since OCR editing is the major bottleneck in terms of speed, the quality of the transcription results is very important. I chose to use ABBYY FineReader because of its batch processing capabilities and the discernibly better transcription results. None of the currently available open source OCR engines are able to produce text results at a low enough error rate to be useful in this workflow.

Digital Field Guides. – Although still in its infancy, the ASU Arboretum Digital Field Guide has garnered interest from both the public and local institutions. The geotagged photo library, consisting of 142 photo observations and 115 species, has received a total of 7,347 views on Panoramio, Google Earth, and Google Maps in about 16 months. The potential applications in further development of this project are broad. From a public outreach perspective, this would be an excellent opportunity for engaging citizen science initiatives. A digital field guide should be a dynamic thing, where users can contribute their own observations and track changes in occurrences. On a smaller scale, such as in an

arboretum or botanical garden, a digital guide would allow visitors to interact with the collection in a technologically modern way. The information associated with individual specimens need not be only textual – it could consist of a combination of text, photographs, audio, and video. Wrapping all of these features into an application for mobile devices makes it relevant to the way in which data are now accessed (Smith, 2012).

A digital field guide for larger areas, such as nature preserves or state parks, could very easily accommodate observations submitted by the public. This crowdsourced[24] data would aid biologists in tracking changes in species distribution and phenology. In Arizona, this might be useful for easily recognizable species, such as the saguaro (*Carnegiea gigantea*) or ocotillo (*Fouquieria splendens*). Events similar to geocaching, perhaps called biocaching, could be organized with local native plant societies, in which members focus on a particular species or habitat. As a mobile app, the information could be easily accessed over wireless networks and contributions could be made via a simple user interface. If wireless service is not available, as is often the case in remote locations, the images could be queued up on the device and submitted once the device is within range of a signal.

The digitization of field guides opens up new possibilities in learning and public involvement, in which natural history data are presented in a

[24] Crowdsourcing is the process of distributing tasks to an undefined group of people, commonly online

relevant and interactive way. I believe that the future of natural history collections depends upon public support. As Joseph Campbell once said, "I don't believe in being interested in a subject just because it's said to be important. I believe in being caught by it somehow or other." (Campbell & Moyers, 1988). Natural history is one of the most interesting and inherently valuable subjects to me, but I realize that not everyone sees it the way I do. With the right introduction, though, a natural fascination will arise along with the realization of the value of collections. People will not value what we do just because we say that it's important; we need to demonstrate that it is so. Public engagement through digital field guides is one way to achieve this.

CHAPTER 5

CONCLUSION

Herbarium specimens can be digitized more efficiently with the use of optical character recognition and automatic parsing. The rate is highly user dependent, ranging between 21.0-39.4 records/hr for Experts, with an average rate of 26.3 records/hr. Based on this study, data entry using the SALIX Method is approximately 30% faster than typing. Speed is also dependent upon label quality, and to a lesser degree, label length. Using the SALIX Method to database good labels is approximately 140-300% faster than typing, compared to only 0-12% faster if the label quality is poor.

The collections community needs a quick and inexpensive way of digitizing their holdings. The SALIX Method was developed to be an efficient alternative to traditional methods of data entry, and fulfills this need under certain circumstances. I have shown that optical character recognition can be a useful tool, and that automatic parsing can work reliably if based on word statistics that improve with use. For collections that are largely historical, with mostly poor and handwritten labels, other approaches will be more efficient. However, for herbaria with a good portion of newer specimens, the SALIX Method would be a good choice. The entire workflow as outlined here could be used, or portions of it incorporated into an existing workflow.

Using the SALIX Method to database specimens has proven to be moderately faster than typing, but more importantly, it has opened up new possibilities in natural language processing and the digitization of herbarium specimens. As previously mentioned, the major bottleneck in the workflow is the editing of OCR results. Until OCR technology improves, this will remain a major obstacle. However, it might be possible to first capture a more skeletal dataset, and then go back to complete the data at a later time. This would reduce the amount of work spent on editing text, and give the technology some time to develop. Working with custom dictionaries and controlled vocabularies may prove to be useful for automated text editing. I do not think we should expect major advancements in natural handwriting recognition, but it is reasonable to look forward to improved accuracy with type.

The future of SALIX most likely involves full integration with the Symbiota[25] software, on which SEINet is based. There are currently 14 active portals using Symbiota; SALIX integration would allow them all the choice of experimenting with this new technology.

Development of data mobilization techniques using new technologies is particularly important at this time. Most people now obtain information via the web rather than print, and the demographic of smartphone users is rapidly growing. We need to generate interest and support through outreach in order to sustain our collections. If we can

[25] Symbiota (http://symbiota.org)

combine solid biodiversity data with good design and make it easily accessible on mobile devices, it will make our work relevant to the way in which information is now accessed.

Digitization does not end with the conversion of printed information to digital – it continues with advancing access to important biodiversity data through effective mobilization.

REFERENCES

Campbell, J. & Moyers, B. The Power of Myth. New York: Doubleday, 1988. Print.

Granzow-de la Cerda, Í. & Beach, J.H. 2010. Semi-automated workflows for acquiring specimen data from label images in herbarium collections. *Taxon* 59(6): 1830-1842.

Heidorn, P.B. & Wei, Q. 2008. Automatic metadata extraction from museum specimen labels. *Proceedings from the International Conference on Dublin Core and Metadata Applications*: 57-68.

Lane, M.A. 1996. Roles of natural history collections. *Annals of the Missouri Botanical Garden* 83: 536-545.

Moen, W.E., Huang, J., McCotter, M., Neill, A. & Best, J. 2010. *Extracting and parsing of herbarium specimen data: Exploring the use of the Dublin Core Application Profile framework.* In proceedings of iConference 2010, University of Illinois at Urbana-Champaign: 154-160.

Parmesan, C. & Yohe, G. 2003. A globally coherent fingerprint of climate change impacts across natural systems. *Nature* 421: 37-42.

Petersen, A.T. & Vieglais, D.A. 2001. Predicting species invasions using ecological niche modeling: New approaches from Bioinformatics attack a pressing problem. *BioScience* 51(5): 363-371.

Plinius, G. 77. *Naturalis Historiæ*. Rome.

Smith, A. 2011. Mobile devices help people solve problems and stave off boredom, but create some new challenges and annoyances. *Pew Internet and American Life Project*. Accessed 18 May 2012 (http://pewinternet.org/Reports/2011/Cell-Phones.aspx)

Smith, A. 2012. Nearly half of American adults are smartphone users. *Pew Internet and American Life Project*. Accessed 18 May 2012 (http://pewinternet.org/Reports/2012/Smartphone-Update-2012.aspx)

Thomas, W. Wayt. 1999. Conservation and monographic research on the flora of Tropical America. Biodiversity and Conservation 8: 1007-1015.

Wilson, E.O. 2003. The encyclopedia of life. *Trends in Ecology and Evolution* 18: 2.

Zickuhr, K. 2012. Three-quarters of smartphone owners use location-based services. *Pew Internet and American Life Project*. Accessed 18 May 2012 (http://pewinternet.org/Reports/2012/Location-based-services.aspx)

BIOGRAPHICAL SKETCH

Anne Barber began her journey into the world of plants as a student of Herbal Medicine at the Blue Heron Academy in Grand Rapids, Michigan. She quickly realized that the wide array of ethnobotanical uses mirrored the biodiversity of the plants themselves, and sought to understand botany beyond herbal medicine. This would eventually take her to Arizona State University, where, as an undergraduate, she focused her attention on lichens and collections management. Her graduate work allowed her to further explore her interests in archives, digitization, public outreach, and natural history collections on a broader scale. She takes to heart what the great mythologist, Joseph Campbell, once said, “Follow your bliss and the universe will open doors where there were only walls.