The SALIX Method- documentation

By Graduate student Anne Barber and advisors Les Landrum, Daryl Lafferty, and Ed Gilbert

The accessibility of biodiversity data is limited by the resources required to convert printed information to a digital form. Many tedious hours of labor are invested in the huge job of data entry. Our expectation has been that if this process could be at least partially automated, it may move at a faster pace. In an attempt to meet this need, SALIX (**s**emi-**a**utomatic **l**abel **i**nformation e**x**traction) was developed under a grant to digitize ca. 55,000 botanical specimens from Latin America, housed at the ASU Herbarium. 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 an upper-level OCR program to create a text file. This information is then 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 compliant CSV (comma-separated values) file and uploaded to SEINet (Southwest Environmental Information Network, http://swbiodiversity.org/seinet/index.php), our online database.

What we refer to as The SALIX Method (Figure 1) consists of a number of different software packages used in digitization, all revolving around SALIX. Both SALIX and BarcodeRenamer (BCR) were developed Lafferty at ASU and designed to meet specific needs. BCR provides a way of automatically renaming image files to match global unique identifier (GUID), while SALIX works as an automatic parser. The other tools we use are proprietary. We quickly discovered that in order to make OCR worthwhile, we needed a higher functioning program than what was available open-source. 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 we chose Microsoft Word for its ease of use and search-and-replace functionalities. This is the program we use for all of our text editing. Finally, we chose Adobe Lightroom for image management and editing. This software package was designed with high-volume processing in mind, and works very well with our large archive.

Figure 1. The SALIX Method: SALIX, BCR, ABBYY FineReader, Microsoft Word, and Adobe Lightroom

We have found that the speed of data entry using SALIX is dependent upon label quality and length, as well as user proficiency. When label quality is good, SALIX can be up to 3 times faster than typing. Using SALIX to database our specimens (with a mixture of good and bad quality and long and short labels) has proven to be moderately faster than typing, but more importantly, it has opened up new possibilities in data processing and digitization of herbarium specimens.

Methodology

Barcoding and Imaging

Before any imaging occurs, each specimen is assigned a barcode (Figure 2) and that acronym plus number becomes its GUID. We use archival-quality, self-adhesive barcodes and place them 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 we come nearer to having the entire collection imaged. The GUID labels are preferable to the older, ink-stamped accession numbers in terms of longevity, readabilty, and the risk of duplication. After a specimen is imaged, the GUID also becomes the image filename (e.g., ASU0012345.jpg). This has huge benefits from an archival perspective – any record can be pulled up for reference by simply searching for the GUID in Windows Explorer.

Figure 2. Image barcode (GUID)

The next step in the workflow is to photograph the specimens. We use a very simple platform surrounded by fluorescent lighting in place of a copy stand. The primary camera, an 18 MP DSLR, is mounted above the platform. The aperture is set to be low and the exposure set a little above normal because of the white background; auto-focusing is used. A second camera is positioned over the location of the label, and is set to Full Auto. Both are connected to remote shutter releases. 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 is about 10 MP and the auto-focus 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 2-6 times larger than JPGs. We guessed that our images would not go through too many different transformations before being put online, and decided to compromise shooting JPGs for server space. Especially since we started using Adobe Lightroom, this option seemed to make the most sense. Lightroom uses non-destructive editing processes, which means the original image always remains unaltered.

Our standard, 18 MP full sheet images allow a visible magnification on a computer screen of about 3X. For specimens with important features that are rather small, we are trying to capture separate close-up images of selected specimens (one or two per species) and associating those with the specimen record online. For this we use a 10 MP compact camera or a 14 mp DSLR camera, which allows characters to be viewed on a computer screen at a magnification of about 15X. This step is done later in the workflow.

Both the label and full sheet image files from the photography step are then saved to a temporary location and processed through BCR. This program uses an integrated barcode scanner to read each barcode, convert it to text, and then overwrite it as the file name. 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 couple of 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 full 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. The original, un-renamed full sheet image files on the camera's memory card are deleted, and all of the label images are deleted once the OCR process is completed.

The next step of 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 herbarium that houses monocots, gymnosperms, and the pteridophytes. Although families are often taxonomically rearranged, the changes are only reflected in the herbarium when practical. 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, Microsoft's free file synchronization application.

The final step of the process is to digitally enhance the full sheet images. We 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 adjusted for white balance and tone. 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 image. Lightroom offers a few different ways to import photos for editing. We 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 user is finished editing, 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, 2) copies of the renamed originals located on the external disk, and 3) web-ready versions stored in subfolders on the external disk. Other than the renaming process, the original images are not altered in any way. This process is fully automatic and results in high quality, web publishable images. Once the specimen record has been databased using SALIX, the image can be uploaded and associated with the record based on its image barcode (GUID), and is now fully accessible via the web.

Rather than taking close-ups of every specimen tied to this project, we are selecting just one or two good representatives of each species. 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 character images, and so on. Each of these images are 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. All 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 full sheet images, and are 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 (Google's tesseract), but will function at its highest capacity with more advanced tools. We use ABBYY FineReader Professional Edition for OCR processing, which supports documents in multiple languages and automatic batch processing. Several hundred label images can be run at a time, outputting the results in a Word document with each label separated by a page break. 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. OCR technology is expected to improve with time, and for now, our data processing system is dependent on these expected advancements.

The renamed label images are saved temporarily to the Desktop, and are then opened through FineReader's Automation Manager. The application runs automatically through the batch of label images and produces a Word document at the conclusion of the process[1]. The document is saved

[1] In order to optimize the quality of the OCR results, it is important to customize the settings in FineReader to work with herbarium labels. Here are our recommended settings for ABBYY FineReader 10 Professional Edition:
Tools>Options>Document
- click on *Edit Languages*, select *Specify languages manually*, and check the languages used on your institution's labels
- under *Document print type*, select *Autodetect*

Tools>Options>Scan/Open
- under *General*, select *Do not read and analyze acquired page images automatically*
- check *Enable image preprocessing*
- uncheck *Detect page orientation* and *Split dual pages*

Tools>Options>Read
- under *Reading mode*, select *Thorough reading*
- under *Training*, select *Do not use user patterns*

Tools>Options>Save>RTF/DOC/DOCX
- under *Retain layout*, choose *Plain text*
- under *Default paper size*, choose *Automatic*
- under *Text settings*, check *Keep page breaks*
- uncheck *Keep headers and footers*, *Keep line breaks*, and *Retain text and background colors*
- under *Picture settings*, uncheck *Keep pictures*
- under *Advanced*, uncheck *Highlight uncertain characters* and *Enable compatibility with other word processors*

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

to the network drive and then the label images are deleted or kept temporarily. Before this portion of the workflow is perfected, it would be wise to save the label images, should you need to re-run the OCR. After consistently acceptable OCR results can be obtained, it is no longer necessary to keep them for more than a few days. The OCR results are saved in the same location as a folder of corresponding full sheet images, both of which are used during data processing with SALIX.

Data Processing

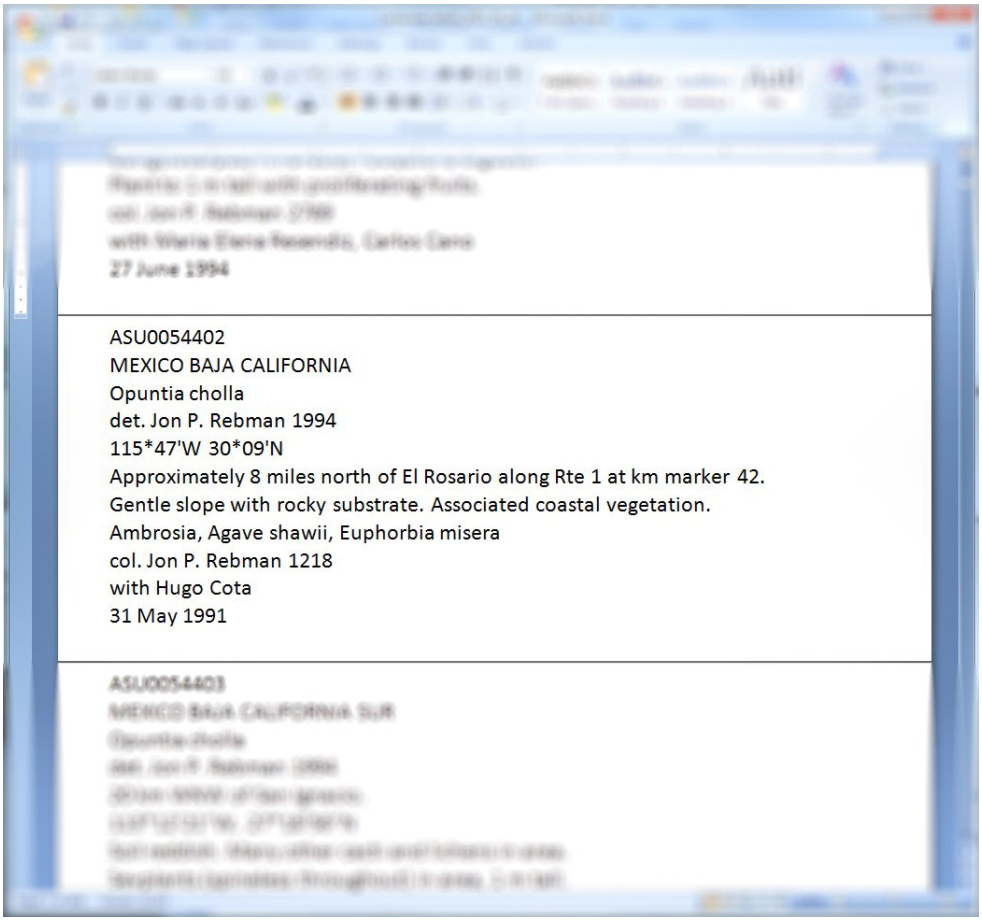
With the technology currently available to us, a fully-automated system for data processing is impossible. As it stands, there are too many errors associated with optical character recognition and automatic parsing for the system to work without an operator. Much of the cause for this lies within the primary data itself – the labels associated with the specimens are too highly variable in quality for any currently existing OCR software to process with an acceptable margin of error. However, we believe that it is possible to develop a system that is partially automated. With the aid of our programmers, principal investigator, and Barber as project manager, we have developed a system revolving around the central software package, SALIX. This is a Windows executable program written by Lafferty to handle automatic parsing of label information. Label data is 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. For example, the words “yellow” and “flowers” appear 916 and 603 times respectively, each with a 100% score in the Plant Description field. So a phrase such as, “Large herb with yellow flowers” would be analyzed word for word with the following results (Table 1), and then determined to belong in the Plant Description field.

Table 1. Word Stats

Word	Count	Description	Locality	Habitat
large	124	52	13	34
herb	170	100	0	0
yellow	916	100	0	0
flowers	603	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. The accuracy of the parsing should be verified by a user before each record is exported to the database. Parsing also improves when the user separates blocks of information by a new line in the Word document (Figure 3). In the figure below, you can see how the locality, habitat, and associated species are separated by line breaks. SALIX is programmed to consider this along with word statistics during parsing.

Figure 3. OCR editing



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 we call “start words”. The collector is prefixed by “col.”, associated collectors by “with”, and the determiner by “det.” (Figure 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, leg 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 we faced early on was that misspelled, unpublished, or otherwise incorrect scientific names were being added to the database. To fix this, we needed an easy way of verifying new name additions. Our existing taxonomic authority file is 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 the Tropicos database of names (http://www.tropicos.org/).. 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.

Also built in to the SALIX functions is a system for verifying the accuracy of geographic coordinates. A program file was built that contains geographic limits 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 the coordinates fall outside of the geographic limits, an error is thrown up and the user must check the data against the label image. This greatly reduces the amount of georeference errors in the database and improves the reliability of our data.

The general workflow for using SALIX is fairly simple. The user begins by opening the first sheet 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. The user should also look over the entire sheet image so as to not miss any important information, such as annotation labels or accession numbers. A typical label before editing is represented in Figure 4, followed by the editing text ready for SALIX in Figure 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 Figure 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.

Figure 4. OCR results before editing

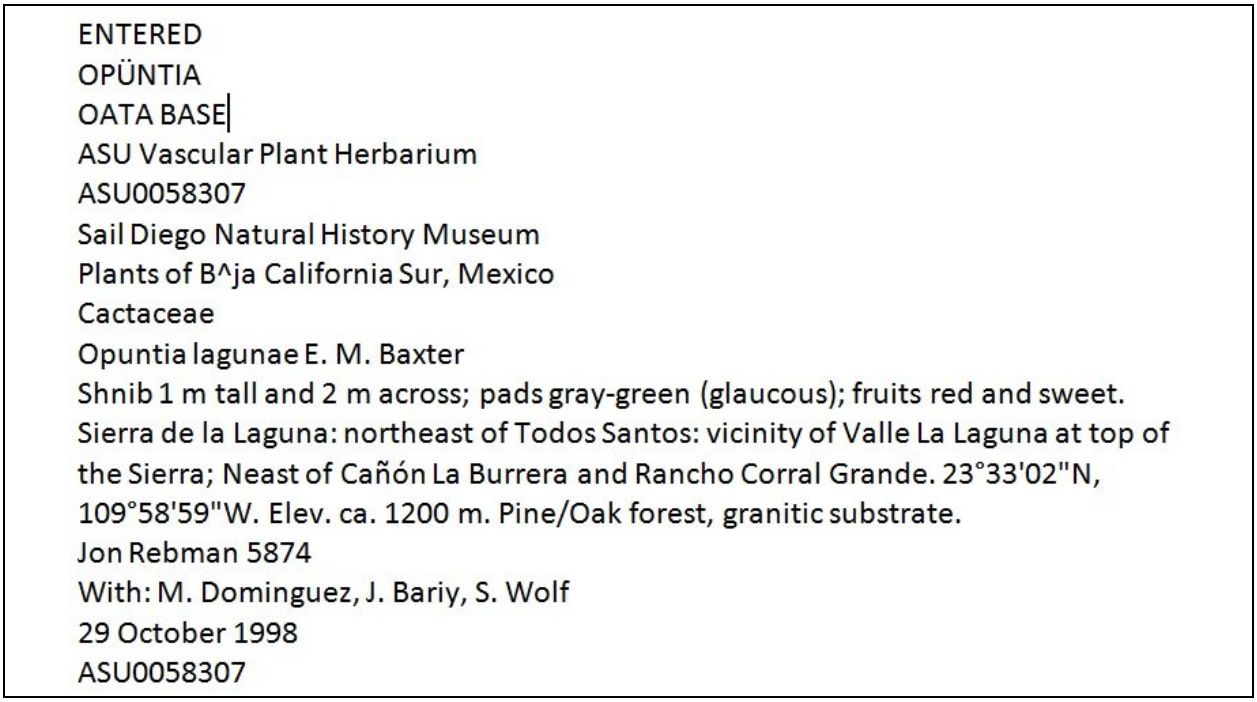
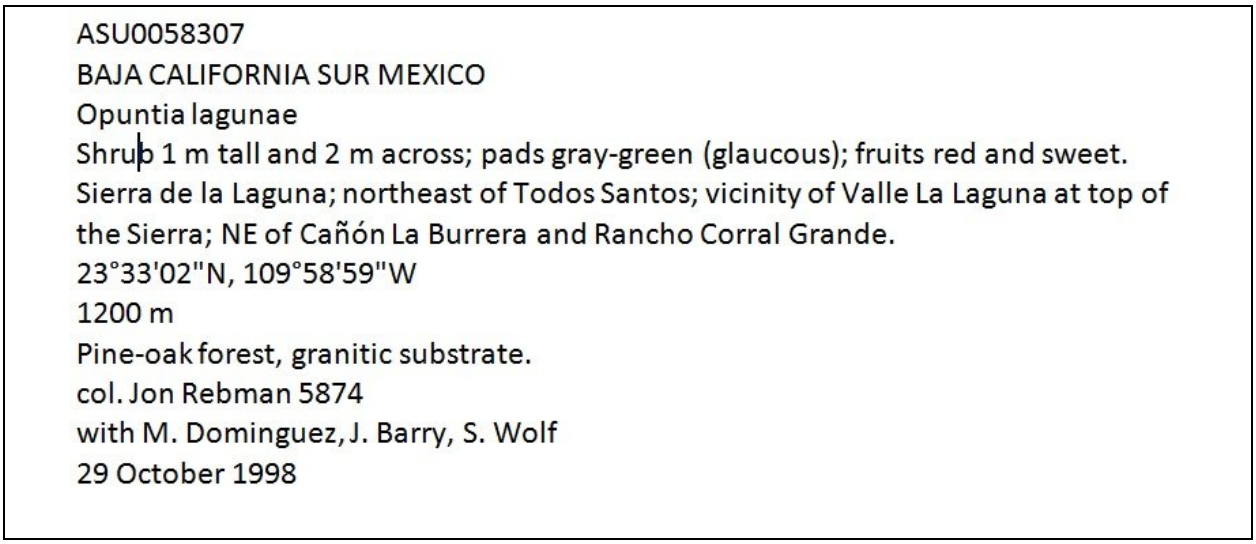


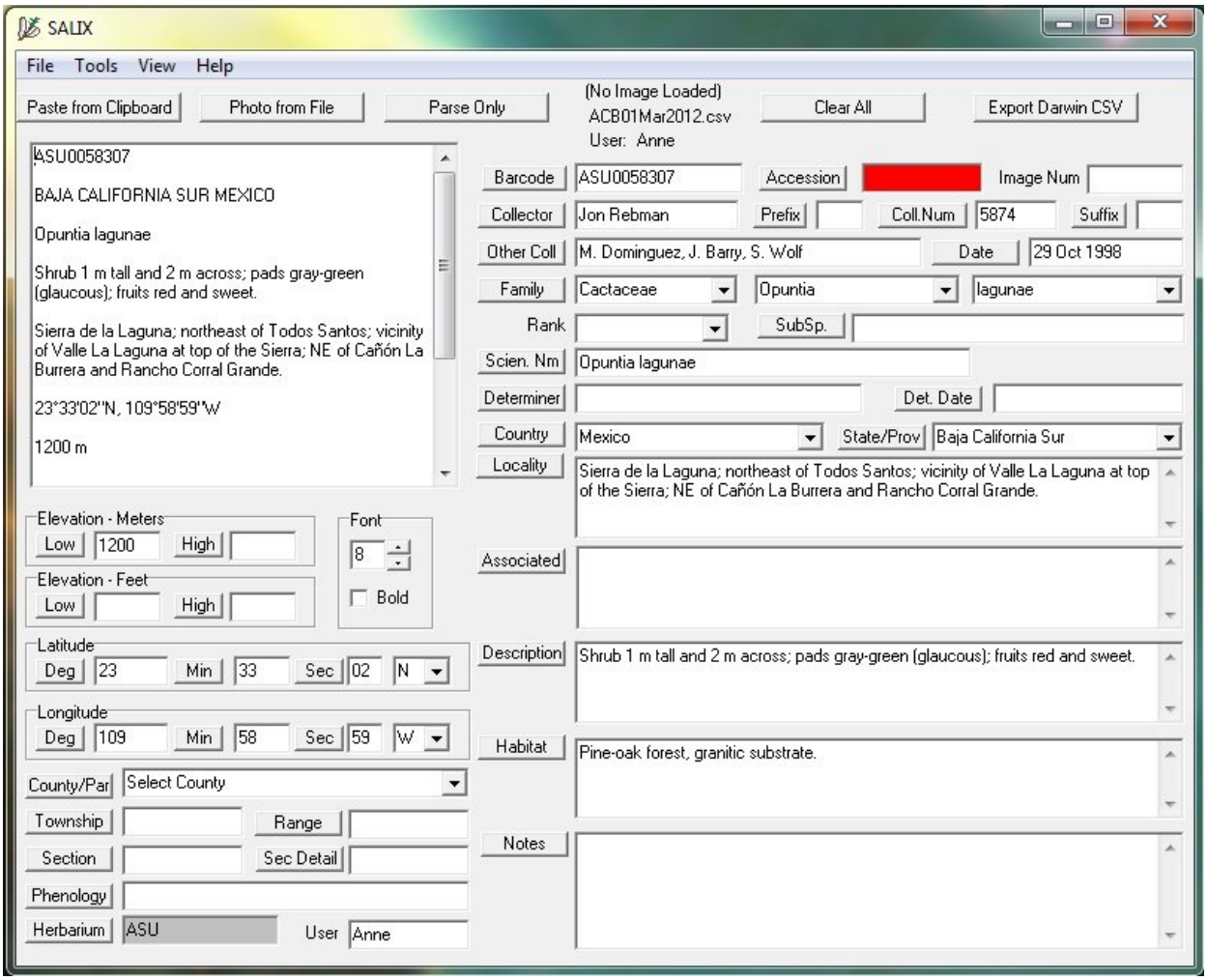
Figure 5. OCR results after editing



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 Figure 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, which gets typed in by hand from the sheet image. Lastly, the “Export” button is pushed, and the label data is stored in a DarwinCore-compatible CSV (comma separated values) 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.

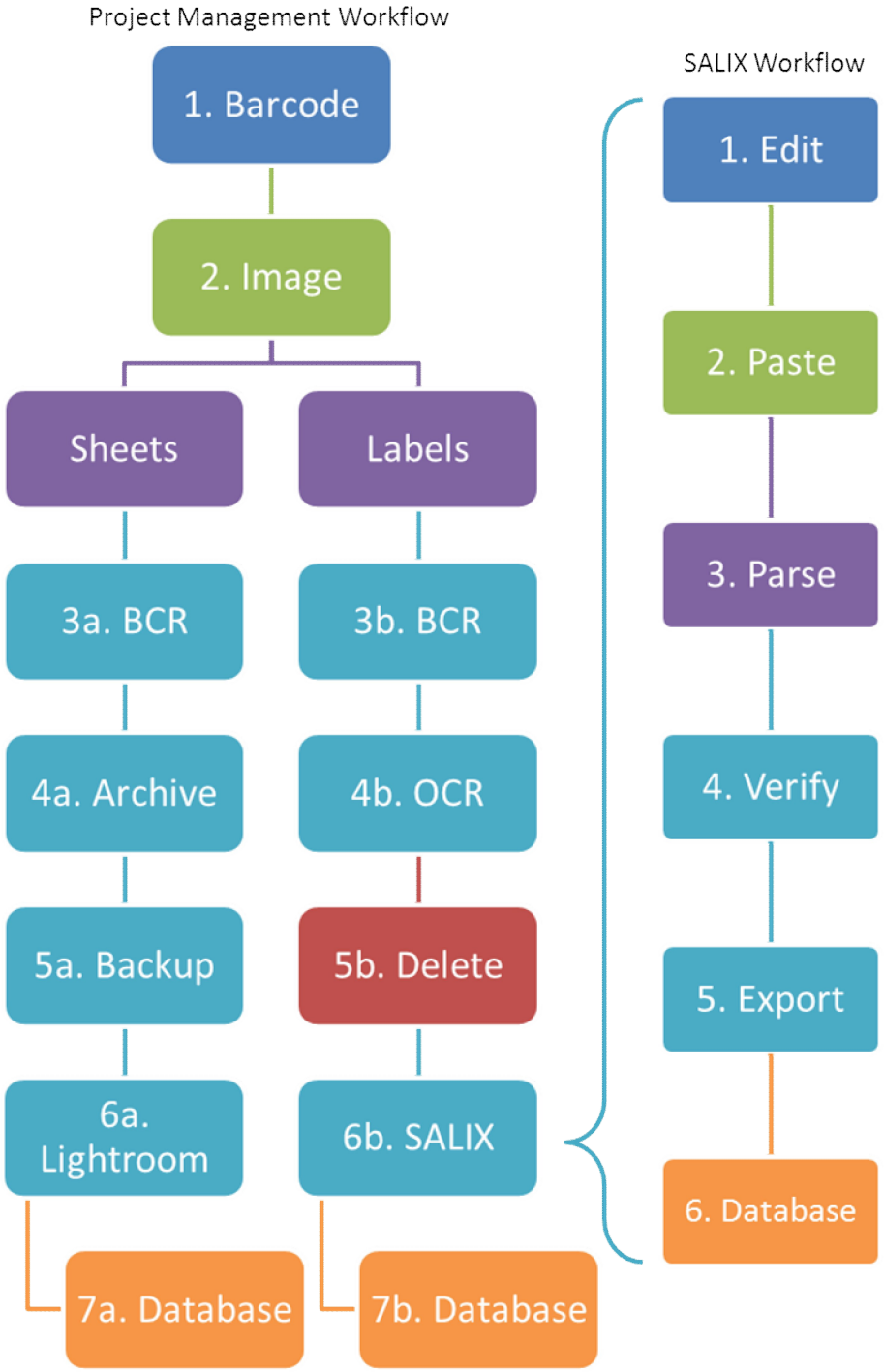
Figure 6. SALIX



Workflow Summary

Our digitization process consists of two main workflows – the data processing done with SALIX and the project management. In the SALIX workflow (Figure 7; right column), the OCR editing is the major bottleneck in terms of efficiency. Most of the time devoted to databasing is spent prepping the text for SALIX (Step 1). Steps 2-6 can be accomplished in relatively no time at all. 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 pushes the Parse button (Step 3), and then verifies that that step was performed correctly (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 then it gets uploaded to SEINet (Step 6).

Figure 7. The SALIX Method workflow



The project management workflow (Figure 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. Sheet 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 sheet 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 to be edited, paired with a folder of corresponding sheet images (Step 6b). These folders of sheet 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 sheet 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 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 sheet 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 with the textual data from Step 6 in the SALIX workflow (Figure 7; right column).

The workflow for capturing close-ups is fairly simple (Figure 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).

Figure 8. Close-ups workflow

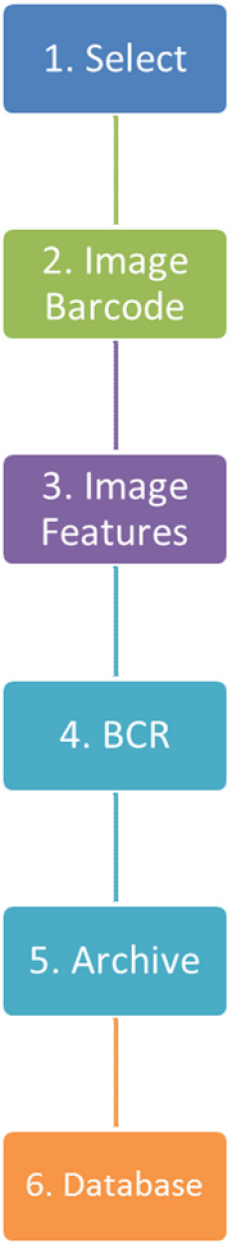
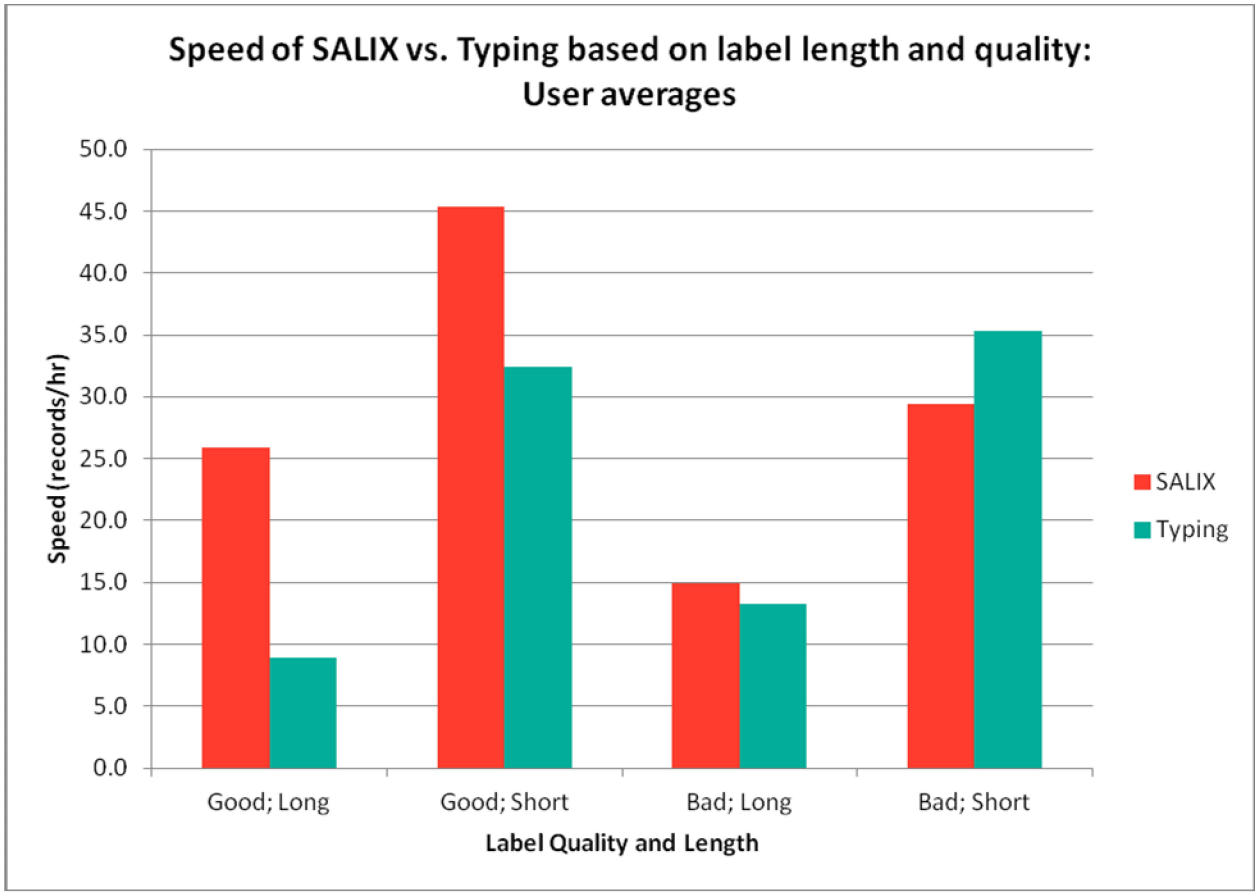


Figure 9. Preliminary results. Overall average length 86 word, short ave. 44 words, long ave. 127 words



Preliminary cost analysis.

Pay rate. $11/hour (we try to get work study students to reduce cost).

Photography including barcoding and filing. 100 specimens/hour [11 cents/specimen]

Databasing wide assortment of labels. 20/hour for average worker [55 cents per specimen]

Supervising graduate student [10 cents per specimen, a subjective estimate]

Cost to get an imaged and databased specimen on the web ca. 76 cents.