SALIX, the Semi-automatic Label Information Extraction system

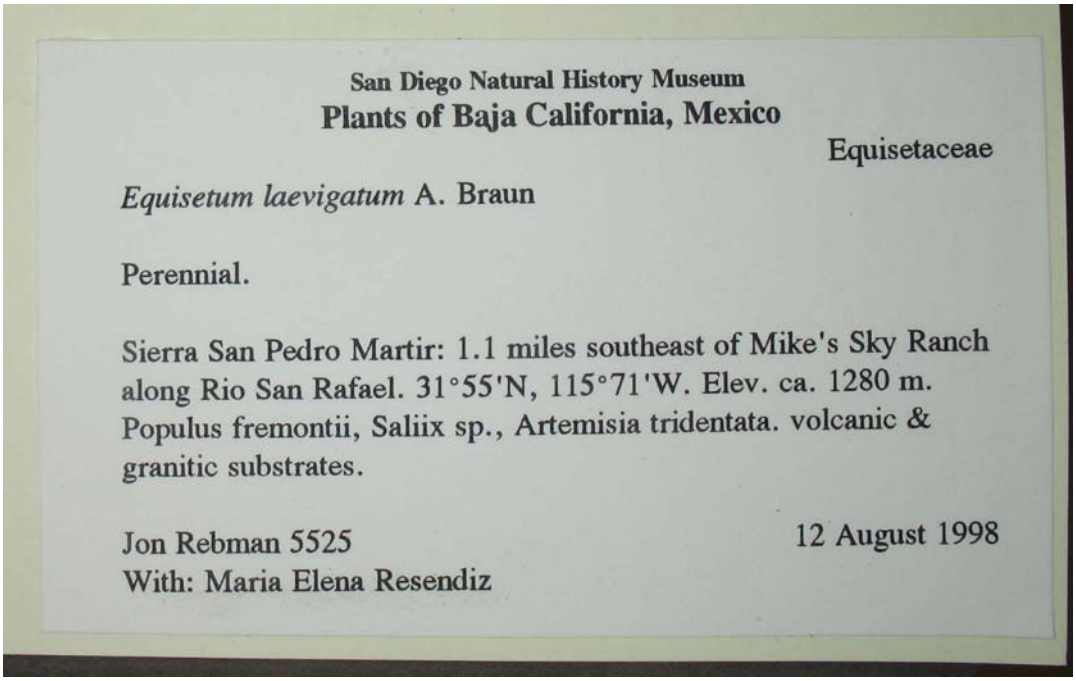
Daryl Lafferty and Leslie R. Landrum
School of Life Sciences
Arizona State University
Tempe, Arizona 85287-4501
USA

The use of Optical Character Recognition (OCR) software to read label data coupled with additional software to transfer those data into a database has been a goal in recent years (http://www.herbis.org/index.php). If the process of extracting data from specimens could become automated, then the process of databasing numerous specimens held in herbaria and museums could be greatly accelerated. We believe that full automation may be an impossible goal. Specimen labels are so variable in format and quality that there will always be a need to do some checking for accuracy at some stage of the process.

We have found that ABBYY OCR programs work well with label data and one of us (Lafferty) has developed a program that facilitates moving data from the OCR out-put files to a database file. We have been able to use this system, with numerous labels and have improved it to the point where it becomes a practical alternative to typing data into a database. So far the speed is not much greater than typing, but there are two advantages: 1) a photographic record of each label and specimen is made; and 2) when labels are information-rich (e.g., listing numerous associated species, or having an extensive habitat description) all the information can be transferred easily. When typing in label information, these fields sometimes have to be abbreviated for efficiency's sake. Furthermore, all the processing is in the control of the operator and the necessary equipment and software is relatively inexpensive. A moderately priced digital camera, ABBYY 5.0 software (available as packaged software with many scanners), and SALIX are all that is needed.

By doing the whole job in-house, one can proof-read the database as one goes along. There are several points at which errors can occur: 1) the original label may have spelling errors; 2) the OCR results can be flawed due to fuzzy or smudged letters; 3) the OCR results may be incorrect because the program is trying to read the label as English when it is Portuguese or Spanish (this kind of error can be corrected with ABBYY which has many language options); 4) the OCR program may parse the image into blocks incorrectly (blocks can also be reset with ABBYY); 5) parsing of the OCR out-put to a database automatically can introduce mistakes (e.g., by putting habitat data into the locality field). SALIX parses the OCR out-put to a database semi-automatically and the user is watching and facilitating the process, so mistakes can be corrected immediately.

The following is an example of how SALIX works with ABBYY.



San Diego Natural History Museum
Plants of Baja California, Mexico

Equisetaceae

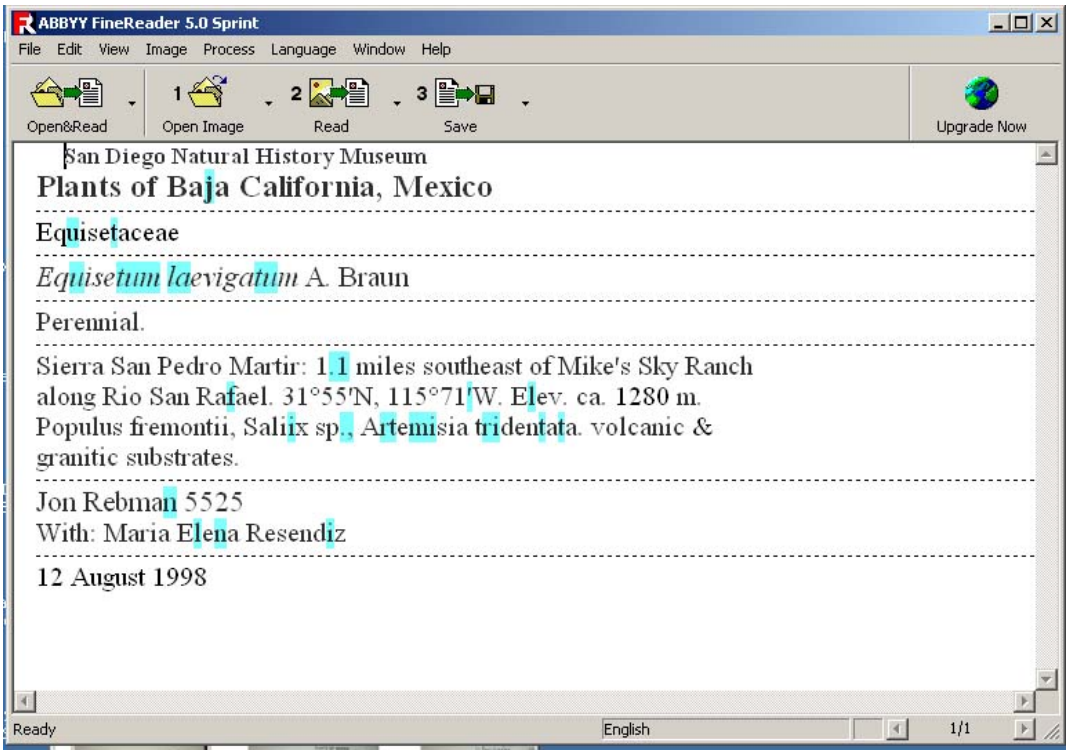
Equisetum laevigatum A. Braun

Perennial.

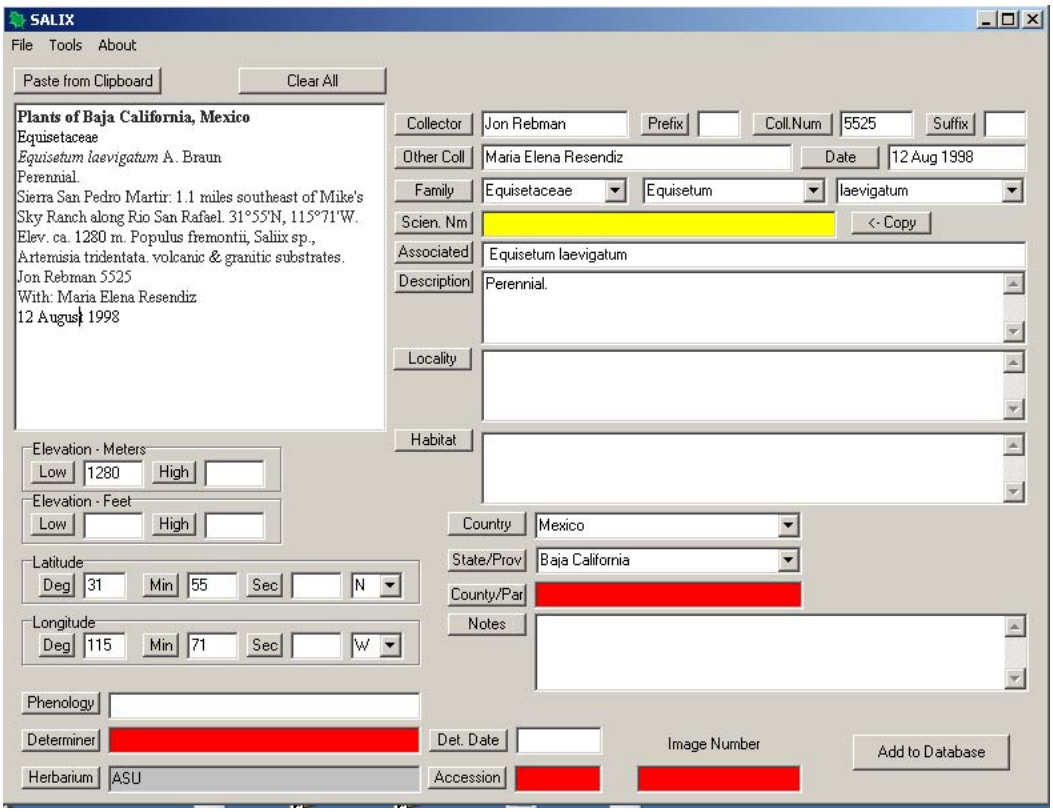
Sierra San Pedro Martir: 1.1 miles southeast of Mike's Sky Ranch along Rio San Rafael. 31°55'N, 115°71'W. Elev. ca. 1280 m. Populus fremontii, Saliix sp., Artemisia tridentata. volcanic & granitic substrates.

Jon Rebman 5525 12 August 1998
With: Maria Elena Resendiz

First, a photo of the label corner of the specimen is taken (or of the whole specimen if the camera is of high enough resolution). This photo was taken with a Canon PowerShot A620 (7.1 megapixels)



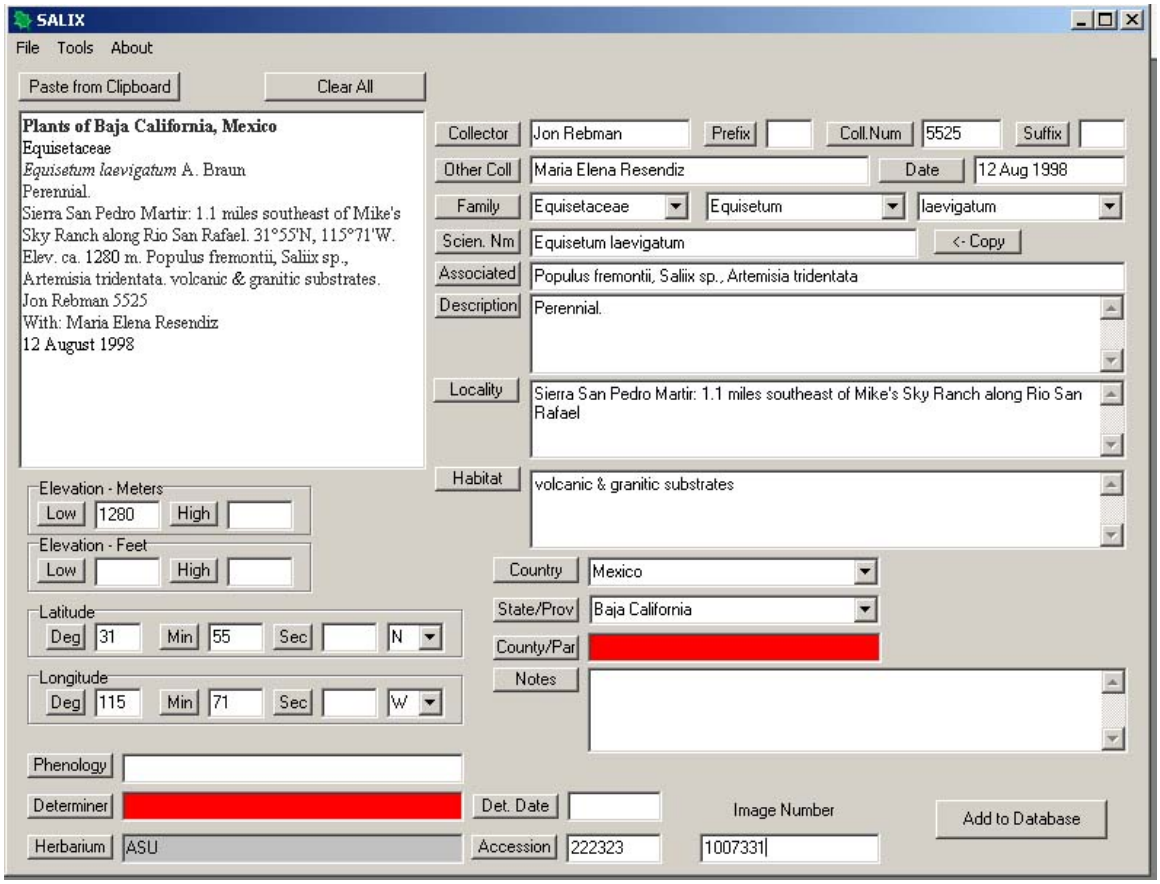
Next the image is opened and read with ABBYY 5.0 (click on “Open&Read” and choose and image). By clicking on “Language,” the operator can select different language options. “Window” allows one to see the image again after it has been processed and change the parsing of the OCR. In this case the OCR was excellent. Labels vary in the OCR results they give. The operator next highlights the text and copies it. It automatically goes to the clipboard.



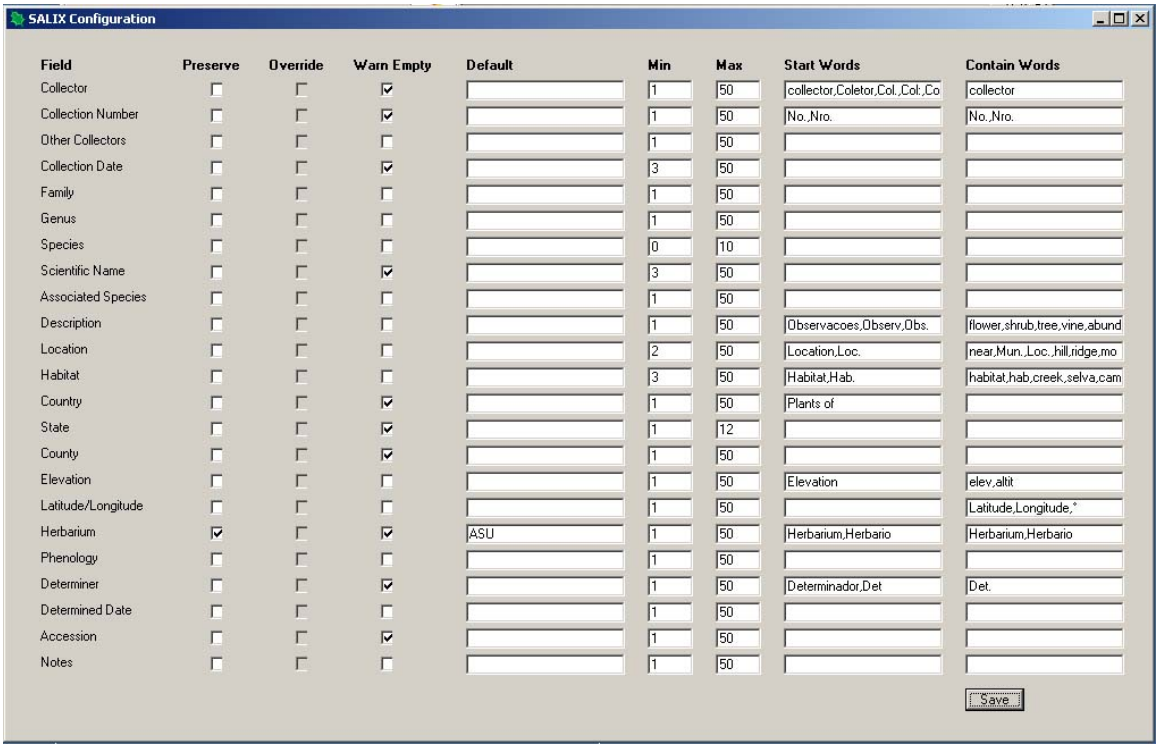
Next the operator opens SALIX and clicks on "Paste from Clipboard." The text fills in the window on the upper left, and parts of the text are recognized and go to other field windows. In this case SALIX did a fairly good job of recognizing portions of the text. The pull down menus for family, genus, and specific epithet have numerous names in them that are correct. In this case the family, genus and epithet were guessed correctly. If this does not happen, one can use the pull down menus to find the right name. If the name is not in the menus, the operator can write a name. Clicking on "Copy" transfers the name to "Scien. Nm." The yellow and red colored blocks indicate that these should be filled if possible. SALIX mistakenly put *Equisetum laevigatum* in "Associated" and did not put *Populus fremontii*, *Salix* sp., *Artemisia tridentata* there, so that would be corrected.

Next, the operator can highlight portions of the text in the large window and then click on the buttons (e.g., locality) to transfer the text. By right-clicking on a field window, one can designate if additional text is to be added before or after existing text. This becomes useful, for instance, if habitat data are divided into separate portions of the label. Finally, the operator clicks on "Add to Database." If some colored blocks still exist, as they do for this label, then the operator will be warned that data are missing. The operator will be able to ignore this warning if the data just do not exist, or have a chance to fill them in if the data do exist.

Data are then added to a tab delimited database file that can be transferred to other database formats.



By clicking on “Tools” in the upper left one opens another window that allows changes of some of the parameters of SALIX. Suppose the operator is databasing several specimens in a particular family and genus. These can be set to be preserved from label to label. New habitat words might be added to the existing list, or others might be deleted. The aim is to make SALIX a program that the user can modify as needed. Certain changes will require a programmer.



We must emphasize that labels vary in quality, so OCR results will also vary. ABBYY and SALIX will be less useful when labels are brief than when they are information rich. In some cases it will be easier to type in the label information than to try to use ABBYY and SALIX. But we believe that SALIX as it is right now is a good start and refinement of the program and modifications made by the users to vocabulary words will make it work better. Contact Daryl Lafferty (salix@daryllafferty.com) if you would like to try using SALIX.