TCN: Digitizing Lichens and Bryophytes: Sensitive Indicators of Environmental Quality and Change.

Specimen Processing Workflow

Data capture will involve a semi-automated process consisting of following steps: 1) Imaging all specimens; 2) Using automated pre-processing script to prepare images for label processing; 3) Using automated data capture processing involving Optical Character Recognition (OCR) and natural language data parsing; 4) having human assisted review of the label with the ability to edit, OCR, and reparse as needed; 5) Georeferencing; and 6) Data cleaning

I. Imaging

The imaging team will consist of 3 people working together (one photographer, one to bring specimens and one to return them) to capture an images of label, annotation, and specimen. For most cases, 2-3 images will have to be captured for each specimen. For increased speed and efficiency, a camera and light stand will be used to capture image rather than a scanner. All specimens will be barcoded at this time to avoid duplication across institutions and because some traditional barcodes may be too small for good optical resolution. The new barcode will identify each specimen using a Global Unique Identifier (GUID; InstitutionCode:CollectionCode: CatalogNumber?). The new barcode needs to be in every image, ideally in the same general location. After images are captured, the barcode will be affixed in a standard location. A scale and color chart will also be included in each image. Due to the heterogeneity of specimen storage methods, each image is likely to differ in size of relevant content. To avoid the need to manually crop each image or constantly readjust camera position, workflow would benefit from a system that would allow an auto-cropping script to be used after the images are captured. To facilitate this, the imaging team will position an adjustable frame to indicate active content within image. The scale rule, color chart, and barcode can be included on the inside of the frame. It is estimated that around 300-600 specimens can be imaged within a work day. [Someone within an herbarium might want to walk through the workflow of the image capturing process to test the specifics of the procedure and timing.] At the end of each day, a large TIF (probably around 20 megapixels?) will be stored on a local backup drive and a copy transferred over to the principal Image Processing Folder (IPF) maintained centrally.

II. Pre-processing of Images

The imaging process will finish with automated pre-processing involving the following steps:

1. A copy of the original TIF image will be transferred to an archival folder on a server (or multiple servers as a precaution)
2. The barcode will be captured using the OCR to read the barcode as well as the text of the barcode. These two barcode values will be compared as a verification step.
3. The image file will be renamed to contain the barcode as part of the name.
4. Three JPG versions of the image will be created and transferred to the web server for display within the TCN consortium portal: i) a large version for optional viewing by the user, ii) a web version for routine display, and iii) a “thumbnail” copy. The large versions may be skipped for imaging of label data only.

III. Automated Label Data Capture

Label data will only need to be captured for specimens that have not been previously processed (i.e. duplicates in a different herbarium may be already databased). Specimens that have previously been databased will only need to have their images linked to the existing data record. This will be done by matching the barcode with the accession number. For specimens that have not yet been digitized, label data capture will start with an automated processing script that will cycle through each image within the IPF folder and perform the following steps:

1. Barcode number (primary key) is obtained from the file name
2. The OCR is used to retrieve all text from each of the specimen images
3. Raw text blocks are stored in a MySQL database along with the specimen GUID (barcode)
4. Pre-parsing of the raw text is attempted. Current experience shows that the scientific name, collector, collector number, and collection date can be reliably obtained for roughly 60-75% of the typed specimen labels. Older font styles tend to be more difficult for OCR algorithms. Using OCR capture will not be possible for written labels.
5. For specimens where collector, number and date are retrieved, the consortium dataset will be queried to locate duplicate record data that has already been processed at another institution. If duplicate records are located, a Filter-Push (links below) type workflow will be implemented in the human assisted processing step to increase efficiency of label data capture.
    a. http://etaxonomy.org/mw/FilteredPush
    b. http://bdei2.cs.umb.edu/mw/Use_Case_Scenarios#Finding_Duplicates
6. Specimens are marked as preprocessed and ready for review
7. At this point, records that meet minimum requirements can be published as readable, yet still marked as needing further review. This will make these records immediately searchable (within the TCN consortium) with the possibility of review being done by trained portal users who have agreed to volunteer their services as reviewers.

IV. Human Assisted Label Processing

OCR character misinterpretations (e.g. the letter L interpreted as the number 1) and the fact that herbarium collection labels rarely follow a standard format make it extremely difficult to establish a fully automatic OCR-NLP process that does not produce a high error rate. To ensure data quality, each specimen will be required to go through a human reviewed label processing step. Labels that produce poor OCR output (e.g. written labels) will have to be key stroked by hand. Label processing will be handled using one of the two following methods.

The first label processing option involves making use of the Semi-Automatic Label Information extraction System (SALIX). SALIX is a well tested OCR/parsing desktop application that is actively being used and further developed at Arizona State University Herbarium. SALIX can be downloaded for free at http://daryllafferty.com/salix/. SALIX workflow and functionality has been well-documented through the following publication: http://nhc.asu.edu/vpherbarium/canotia/SALIX3.pdf. Current workflow at ASU of parsing up to 30 records/hr are being achieved.

The second label processing option **and our preferred one** involves making use of web applications currently being designed in conjunction with the SYMBIOTA software project. This application makes use of many of the concepts used by SALIX, though through a web user interface. Benefits of a browser-based, online "cloud" network review system include:

1. Reviewers can be located globally, given that there is web access.
2. Volunteer portal users can be approved to review data. A web training and review process would be needed to make sure it was being done correctly.
3. Review application will be platform independent since it will be designed to function within supported browsers.
4. Web APIs, such as Google Map, can be easily incorporated to aid georeferencing and data quality check (e.g. do coordinates fall within given state or county).
5. Taxonomic web services can be used to validate names.
6. AJAX and javascript are well suited for creating user friendly systems.
7. Data capture directly from images reduces the number of times specimens are physically handled.

Both SALIX and the SYMBIOTA web-based label processing system will possess the following capabilities:

- While label formats tend to differ greatly as a whole, individual collectors tend to follow a certain format for most of their collecting career. This makes it possible to develop a system (machine learning process) where the reviewer can train the parsing algorithms to perform better with specific label formats. SALIX is already capable of this functionality. Our TCN reviewers will have the ability to query the raw data by the collector name, thus enabling them to process labels in batches based on a similar label format.

- When duplicate records are located in the integrated consortium datasets, the reviewer will have the ability to make use of these data, with or without edits. If multiple records of the same collection are located, the reviewer will be able to use a FilterPush (Macklin 2010) mechanism to quickly review the field data that differ and select the more accurate values. In order to make duplicate specimen data available to other partners, all data will be published within the consortium dataset immediately after label processing.

In either of the above processes, label data will be transferred on a regular schedule to the herbarium source database through a procedure that makes use of the Darwin Core exchange standard. The individual herbarium data will then be harvested from the consortium database using standard data exchange protocols built into the portal (http://symbiota.org/tiki/tiki-index.php?page=Data+Interoperability).

V. Georeferencing

TDWG ratified georeferencing protocols and standards (http://wiki.tdwg.org/Geospatial) will be followed. Existing scripts will be used to obtain decimal coordinates for records with UTM and Township, Range, Section (TRS) information. Geomancer and GEOLocate georeferencing services will be used as

appropriate. Georeferencing will only need to be completed for *ca.* 60% of the specimens that have not been originally georeferenced by the collector. This process will take place during the final year of the project, thus increasing efficiency by allowing georeferencing to be done as a batch process preformed on the consortium's combined dataset. In this manner, coordinates can be accurately and efficiently assigned to multiple specimens that share the matching locality descriptions. Georeference data will be transferred back to source databases using standard exchange protocols. ASU and CCH both stated that they can do 12-15 per hour. **Note:** GEOLocate and Geomancer are currently cool, progressive ideas that hold promise, yet when it comes down to current implementation, there are potential problems although F has had some success. We will strongly recommend that the HUB make a priority of increasing the functionality of these procedures.

VI. Data cleaning

An emphasis will be put on developing a data collection and review process that will best avoid and reduce error from the beginning. Taxonomic names will be verified using a taxonomic thesaurus for lichens and bryophytes. Other errors that will be sought during the label processing step include coordinates that fall outside of political boundaries and invalid collection dates. Data entry errors are difficult to avoid entirely and errors created by the collector or due to political boundary changes will also have to be addressed. Scripts will be developed to locate records containing possible errors. Such scripts will look for improbable collection events, such as two events by the same collector on the same day that took place in localities too distant than one could practically travel within a day. These techniques have been implemented with success by the Consortium of California Herbaria project (CCH).

Some Issues for Discussion (but not limited to these, of course)

- Do we plan on imaging specimens that have already been databased? We recommend doing so since imaging goes relatively quickly and linking images to existing specimen records will not take much extra time.
- Some institutions will already have barcodes that follow their accession numbers. If they are fully barcoded, then this is not a big problem. If they are partially barcoded, should they be re-barcoded or continue with their barcoding scheme? Barcoding specimens by matching barcodes to accession numbers is much more time consuming than using continuous barcode numbers. Institutions without barcodes should be encouraged to use the new barcode as their unique identifier (accession number, catalog number).

- Brent Mishler has suggested strongly that imagining bryophyte specimens should be avoided to reduce potential damage to the specimens and because microscopic information is often required for IDs. However, the RFP clearly calls for imaging and people reviewing our proposal may well expect it. Grabbing an image of the specimen will take minimal extra time.