Image Capture and Processing: an Overview

Consortium of Pacific Northwest Herbaria

July 26, 2010

Ben Legler
Informatics Specialist, CPNWH
WTU Herbarium, Burke Museum of Natural History and Culture
University of Washington
blegler@u.washington.edu
1-206-221-5234

This document describes the general processes used for image capture, image processing, data capture, and data/image dissemination used by the Consortium of Pacific Northwest Herbaria, as carried out under the Consortium's 2010-2013 collaborative NSF Grant (DBI0956414). Details are omitted in an attempt to provide an overall understanding of the process.

Several large herbaria and about one dozen small herbaria within the Pacific Northwest are being imaged under this grant. Many of these collections do not wish to maintain their own database in-house, or lack the means to do so. For this reason, nearly all image processing, image storage, and database hosting is centralized on a web server maintained by the Consortium. Only imaging occurs directly within the collections, using a combination of hourly student help, work study, or volunteers.

The workflow described here (Figure 1) applies to collections within the region that have not been previously databased. These collections are imaged first, with data entry occuring later from the images. Database records are created automatically from the images; thus, there is no need to barcode specimens to link images to database records. However, for collections that are already databased, barcodes are used as the means to link images to existing database records.

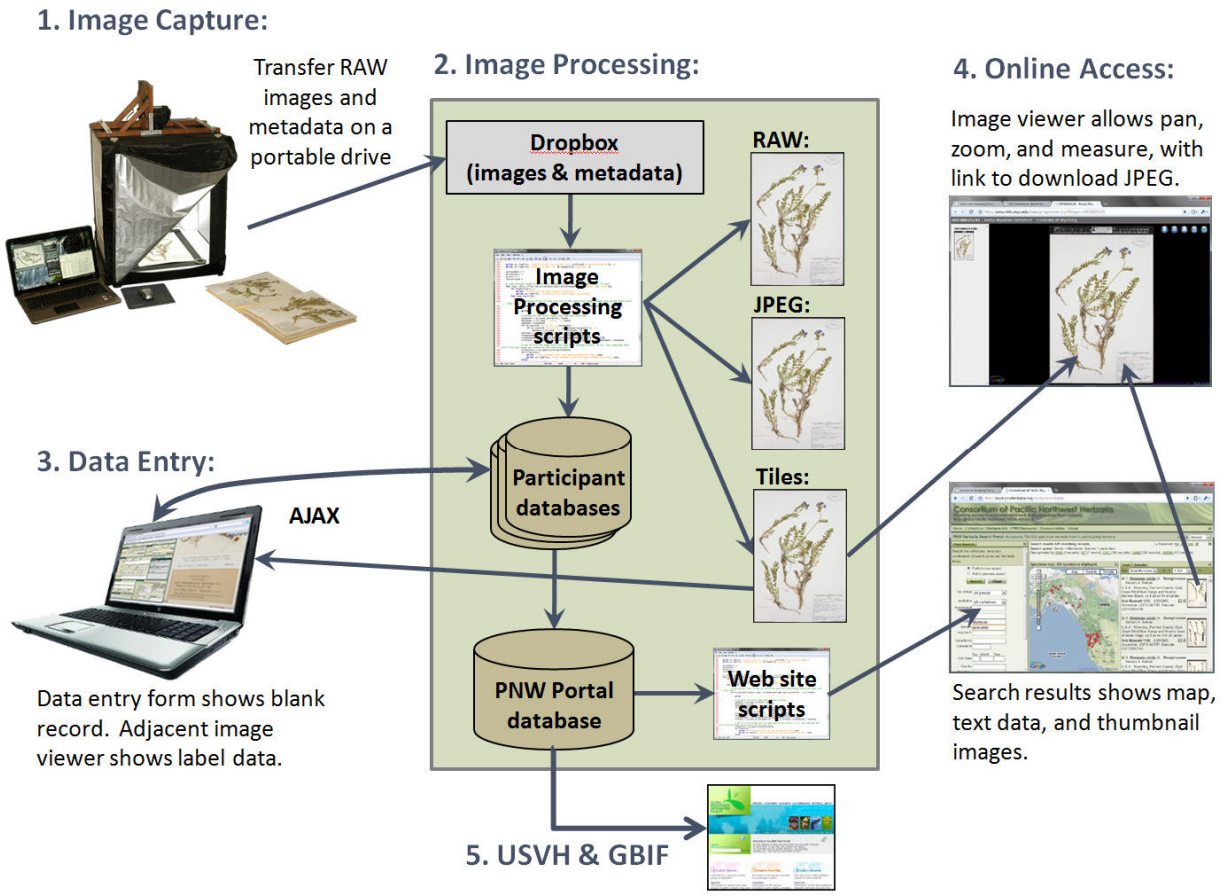

Figure 1. Generalized flow of images and data, from image capture to image and data dissemination.

1. Image Capture:

We are using a lightbox, digital SLR, and laptop computer. The lightbox provides uniform lighting for shadow-free images and involves fewer components than flash units (and thus fewer potential problems). The camera is controlled from the laptop using a USB cord and the software provided by the camera manufacturer. Images are captured in RAW format (in this case, Canon's CR2 format). Camera settings are configured to provide an image that requires minimal post-processing.

Images are stored on the camera's CF card and the laptop's hard drive, with periodic transfer to an external hard drive. The hard drive is occasionally mailed to WTU for transfer to the Consortium's server. Throughout the process there are always two copies of every image on two separate devices as protection against data loss.

Images are named during image capture to a standard format consisting of the collection acronym followed by a six-digit, zero-padded number that increments automatically. The acronym is sufficient to identify which collection the image came from.

Basic metadata for each image is captured during the imaging process using a standalone program written for this project. This metadata identifies the folder in which each specimen is stored, the name of the person who captured the image, and the collection where housed. Metadata is later linked to images indirectly using timestamps recorded with the metadata and in the image EXIF data.

2. Image Processing:

Images are mailed to WTU where they are transferred to a temporary storage directory on the server (called the Dropbox). We then run a batch program provided by Canon to convert the RAW images to high quality JPEG images. Although this part of the process is not fully automated, the higher quality images provided by Canon's conversion algorithms justifies the extra trouble.

The medata, also sent on the hard drive, is copied over to the server and a script is run that copies this data over to a table in the database that corresponds to the collection from which these images came.

These JPEG images are then picked up by a script that runs each night. This script fully automates the remaining image processing steps, which include:

1) Rotate the JPEG (this is the only image manipulation required – no cropping, exposure adjustments, sharpening, color balance corrections, etc.)
2) Create a tiled version of the image for use in the Consortium's online image viewer.
3) Move the JPEG to a permanent directory for later online access and general use.

4) Convert the original RAW image from CR2 to DNG (Digital Negative) and move this to a permanent directory for archiving.
5) Create a record for this image in an images table in the database that corresponds to the collection from which this image came. Each image record is linked at this time to the corresponding folder metadata record using the timestamp in the image EXIF data. At this point, the image is not linked to a specimen record (such record does not yet exist).

3. Data Entry:

Databasing occurs from the images following image capture and processing. A custom data entry interface is being created for this project that allows access to each collection's database over the internet. This interface is being designed as a Rich Internet Application using AJAX, MySQL, and PHP. The end result is a database that can be accessed from any computer without installing any client-side programs or plugins. Thus, managers at the collections being imaged will have full access to their own data without the need to manage their own server and database.

The core component of this database interface is a split-screen view of a data entry form and an image viewer. By default, the image viewer zooms in to the label region of the sheet, allowing the user to keystroke data into the corresponding record for the specimen. Images are displayed for data entry by clicking a button that finds an image without a corresponding specimen record, creates a new record for that specimen, and displays the image and record in the data entry interface. There is no need for the data entry personnel to keep track of which images have already been databased and which remain. Nor is there much chance for write collisions from two workers editing the same record.

Databasing occurs at the lead institutions involved with the grant (WTU, OSC, and ID) using personnel hired on for the project. The workload is split up by state, with each institution databasing the collections from within their own state.

Although OCR techniques show promise for automating the data capture process, there do not appear to be any current solutions ready for deployment here. However, OCR can, and likely will, eventually take the place of the data entry process described here. A more likely alternative in the short-term is the use of OCR to assist the manual data entry process. The image viewer we are using can be easily extended to integrate OCR assist.

4. Online Access:

Data is periodically copied from the individual institution databases to the main Portal database using scripts that run on a nightly basis. Because all databases reside on the Consortium's server and are fully under our control, we can bypass the usual data sharing protocols such as TapirLink and IPT in favor of a custom script that directly transfers data. We can also bypass the limitations of current data schemas and transfer formats, namely

DarwinCore, GBIF's Star Schema, and XML. The result is more frequent updates, richer data sets, and much faster performance.

This data transfer process only applies to databases hosted on the Portal server. For those herbaria that manage their own database in-house, we use existing protocols such as DiGIR, TapirLink, and IPT to transfer data to the Portal database. However, we may eventually bypass these solutions in favor of custom methods that provide richer data sets and faster performance.

Data access is provided by the Consortium's online search interface (http://www.pnwherbaria.org/portal/search.php). This interface will be extended to include thumbnail previews of each specimen image within the list of matching records returned by a search. The thumbnails will link to the image viewer which offers full zoom and pan functionality similar to Google Maps.

5. USVH & GBIF:

The Consortium will create and manage a data access point for those collections hosted on the Portal server. This access point will use IPT to provide data to GBIF and, eventually, USVH. Other herbaria in the region that manage their own databases in-house will remain responsible for maintaining their own data access points.