

SKELETAL RECORDS ACCOMPANYING IMAGES: EFFICIENCY VS LATER UTILITY

Richard K. Rabeler

University of Michigan Herbarium – EEB

Ann Arbor MI



IMAGING SPECIMENS – NOT THE ENDPOINT

- Imaging herbarium specimens is now a common activity
- In a 2014 survey of herbaria in the US, 143 of 248 respondents (57.6%) noted at least some of their specimens have been imaged.
- Of the 13 current Thematic Collection Network projects (NSF/ADBC program), seven involve herbarium specimen digitization.

DIGITIZATION – IT'S MORE THAN JUST IMAGING

- Images are *only one* of the highly desired results in any digitization effort.
- Specimen data needs to be collected (OCR/keyboard/crowdsourced) before the images can be effectively used as more than just an image.
 - Can't easily index them in a web portal.
 - Can't use specimen images directly in ecological analyses.
 - Can't directly map an image.

WHAT DATA TO RECORD WHEN IMAGING?

- The amount and type of data collected at the time of imaging varies among projects:
 - No data records created
 - Recording only the barcode and filed-by name
 - The “minimum” recommended standard
 - Recording additional pieces of data (collector, #, country, state, etc.)
- Comments on the latter two approaches are based on experience involving six imaging projects currently underway at MICH.

NO DATA RECORDS CREATED

- Original concept: all data, including the barcode, could be read via OCR interpretation of the image. Is *not* as reliable as originally expected.
- Not a good practice, especially if a database already exists.
- No easy way to match images with any extant database records.
- Any information only found on folders (i.e., the filed-by name) is lost unless recorded.
- Imager does not need to know how to interpret a specimen label.

BARCODE + FILED-BY NAME

- Approach used by Tri-Trophic TCN, Lichen/Bryo TCN, Macroalgae TCN, MICH CSBR
- Record barcode number of specimen and “filed by” name on folder into an Access database or Excel spreadsheet, often when barcode is applied.

8	1479192	Crataegus macrosperma
9	1465849	Crataegus macrosperma
10	1465848	Crataegus macrosperma
11	1465847	Crataegus macrosperma
12	1465821	Crataegus macrosperma
13	1465798	Crataegus macrosperma
14	1465822	Crataegus macrosperma
15	1465704	Crataegus intricata
16	1465705	Crataegus intricata
17	1465706	Crataegus intricata
18	1477291	Crataegus intricata
19	1474863	Crataegus intricata

- Imager does not need to know how to interpret a specimen label.
- When an image is loaded to a portal, only information: barcode + scientific name.
- Allows the scientific name to be locked during later label transcription.

CAPTURING ADDED FIELDS

- Two projects have created data entry screens where added info from the specimen label is captured.
 - Macrofungi: barcode, scientific name, collector name, collection number, date, associated collectors, additional numbers, exsiccati information (name of set, number)
 - Great Lakes Invasives: barcode, scientific name, collector name, collection number, Country, State/province, county
- The more data you expect an imager to enter, the longer it will take per record and imagers will require more training to be able to locate the data and record *correctly*.
- Any data initially captured will be available when images are loaded to a portal.
- Any additional data can be used to sort images for transcription/crowdsourcing.

MACROFUNGI AT MICH

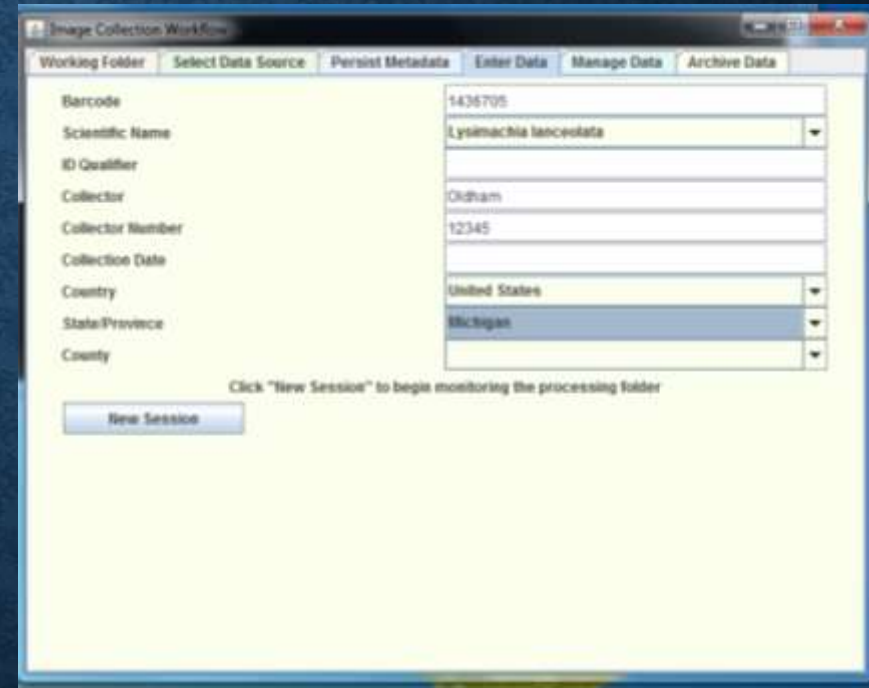
- Data are entered directly into a Symbiota portal record created at the time of imaging.
- Scientific name and geography fields are controlled by drop-down lists.
- Chrome auto-fill also functions to create pick lists.
- After verification/processing, the image is loaded to the portal and linked to the portal record.

The screenshot shows the University of Michigan Herbarium (MICH) data entry interface. The form is titled "University of Michigan Herbarium (MICH)" and includes a navigation bar with tabs for "Occurrence Data", "Determination History", "Images", "Search Lists", and "Admin". The main form is divided into several sections:

- Collector Info:** Fields for Catalog Number (280307), Other Numbers, Collector (A. H. Smith), Number (7345), Date (1988-08-17), Associated Collectors (K. A. Harrison), and Voucher Date.
- Local Identifications:** Fields for Scientific Name (Carthagenus obovatus), Author (K.), ID Number, Family (Carthagenaceae), Identified By, and Date Identified.
- Locality:** Fields for Country, State/Province, County, Municipality, Locality, Locality Security, Search Reason, Latitude, Longitude, Uncertainty, Datum, and UTM Coordinates.
- Notes:** Fields for Name, Synonym, Associated Type, Description, and Note.
- Other Fields:** Fields for Life Stage, Sex, Number of Clones, Sampling Protocol, Preservation, Phenology, and Establishment Means.

GREAT LAKES INVASIVES AT MICH

- Data entered into custom JAVA app
- Scientific name, Country, State/Province, and County are controlled via dropdown lists
- User can select which fields are locked
- Not all fields are required
- Output is a CSV file with DwC headers
- Loaded with image to portal



The screenshot shows a Java application window titled "Image Collection Workflow". The window has a menu bar with the following options: "Working Folder", "Select Data Source", "Persist Metadata", "Enter Data", "Manage Data", and "Archive Data". The main area contains a form with the following fields:

Barcode	1436705
Scientific Name	Lysimachia lanceolata
ID Qualifier	
Collector	Oldham
Collector Number	12345
Collection Date	
Country	United States
State/Province	Michigan
County	

Below the form, there is a text instruction: "Click 'New Session' to begin monitoring the processing folder". At the bottom left, there is a button labeled "New Session".

HOW WILL THE DATA RECORDS BE COMPLETED?

- Various strategies exist for completing a data record, usually tied to the workflow designed for a project and available resources.
 - Keyboard the entire label content from the image.
 - Keyboard selected fields to add to skeletal content.
 - Compare specimens against duplicates in web portals – cut and paste appropriate data.
 - Process a batch of labels via an OCR program (Tesseract, ABBYY, etc), compare/correct/transfer the result into a database.
 - Submit a batch of labels for crowdsourcing.

DATA COMPLETION AT MICH - 1

- CSBR project: minimal data entry at imaging, project manager transcribes all data from images into an Excel spreadsheet.
- Tri-Trophic TCN: minimal data entry at imaging, images loaded (and ABBYY OCR completed) to a project Symbiota portal at NY, additional data transcribed in Symbiota portal from label images by paid staff at MICH and/or volunteers at NY. Limited opportunities for duplicate matching/cut-paste data.
- Lichens/Bryophyte TCN: minimal data entry at imaging, images loaded at MICH to Lichen (<http://lichenportal.org/portal/>) or Bryophyte (<http://bryophyteportal.org/portal/>) Consortia Symbiota portals, additional data transcribed in portals from label images by paid staff at MICH. Duplicate matching often used for specimens from exsiccati sets.

DATA COMPLETION AT MICH - 2

- Macroalgae TCN: minimal data entry at imaging, images loaded at MICH to Macroalgae Consortia Symbiota portal (<http://macroalgae.org/portal/index.php>). Additional data transcribed in portal from label images by paid staff at MICH; minimum data transcribed: collector, collector #, locality. Use of duplicate matching is a personal preference; one transcriber uses extensively, another does not.
- Macrofungi TCN: enhanced data entry at imaging, images loaded at MICH to Macrofungi Consortium Symbiota portal (<http://mycoportal.org/portal/index.php>), locality information added at NY.

DATA COMPLETION AT MICH - 3

- Great Lakes Invasives TCN: enhanced data entry at imaging, images and data loaded at MICH to Consortium of Midwest Herbaria (CMWH) Symbiota portal (<http://midwestherbaria.org/portal/>), ABBYY OCR of images completed at WIS.
- Additional information transcribed in portal from label images by paid staff at MICH.
- Opportunities for duplicate matching/cut-paste data are greater than in Tri-Trophic since CMWH portal is part of broader SEINet portal (now includes 132 herbaria esp. from SW, SE, central USA).

MAGIC FORMULA?

- There is no magic formula as to which approach would be most efficient
- Items to consider:
 - Resources – what can you afford?
 - Abilities and training of your workforce:
 - photographic vs. botanical experience.
 - How long will someone be working on a project vs. time invested in training.
 - What data do you want later transcribers to *not* be able to edit?
 - What data would be needed to effectively benefit from any semi-automated approach?
 - Do you want to be able to sort images for more efficient transcription?
 - Data should be proofed – no real way to completely automate.
 - Learn from others – see what others are doing!

ACKNOWLEDGMENTS



- National Science Foundation for funding the following grants at MICH:
 - CSBR – 1349276
 - Lichens & Bryophytes TCN – 1115030
 - Tri-trophic TCN – 1115081
 - Macrofungi TCN – 1206134
 - Macroalgae TCN – 1303779
 - Great Lakes Invasives TCN – 1405032
- Project Managers for MICH projects:
 - Diego Barroso (Great Lakes Invasives)
 - Mackenzie Caple (CSBR)
 - Matthew Foltz (Bryo/Lichens, Macrofungi)
 - Samantha Winder (Macroalgae)
- Great Lakes Invasives TCN at WIS: NSF Award - 1410683
 - Robert Anglin (WIS) : development of Great Lakes Invasives Java App
 - Ed Gilbert (ARIZ): development of CMWH portal.