

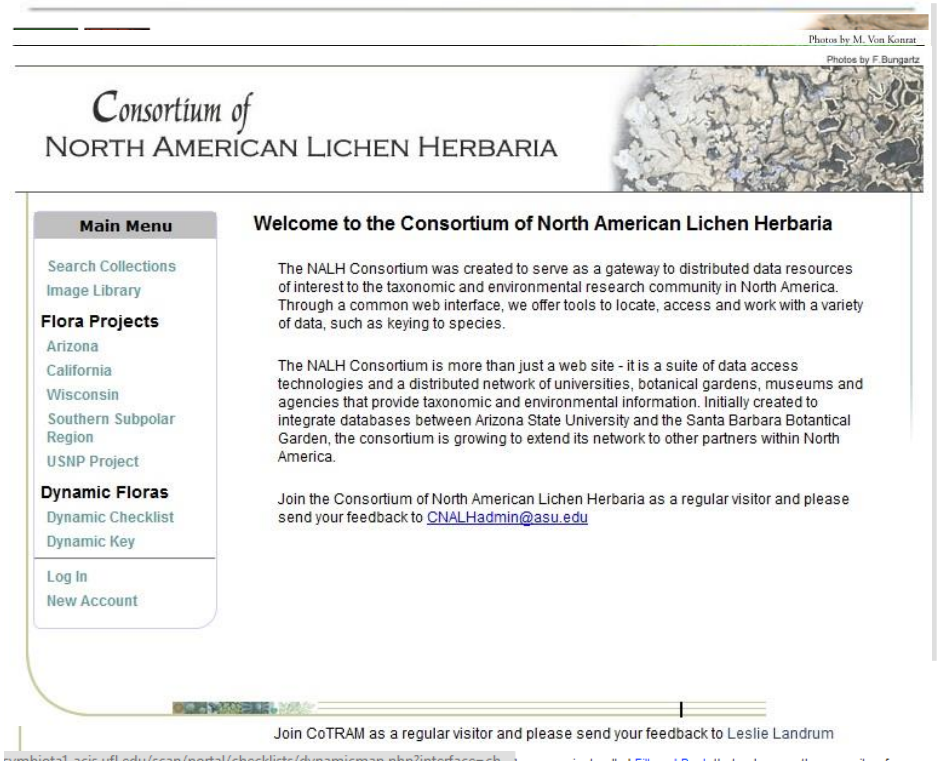

Symbiota: a specimen-based biodiversity portal platform

Edward Gilbert

National Science Foundation
WHERE DISCOVERIES BEGIN

Scientific Community Portals

- Community portals
- Distinct datasets
  - Taxonomic scope
  - Geographic scope
- Custom look & feel
  - CSS, config files
- SEINet
  - 3m specimens
  - 400k images



SERNEC

Southeast Regional Network of Expertise and Collections

Welcome to SERNEC

Herbaria are not simply repositories of plant specimens, they are repositories of a tremendous amount of information. Current technologies provide an opportunity to access this information at an unprecedented scale. The real power of herbaria as research tools can be fully realized when both large and small collections within a broad geographic region are electronically available and searchable. SERNEC (SouthEast Regional Network of Expertise and Collections) is designed to facilitate this process, by building partnerships, encouraging the utilization of the collective expertise of the network, and assisting herbaria in providing information to the public. SERNEC is 1) networking the 230 herbaria in 14 states in southeastern North America, 2) developing a strategy for advancing each state's ongoing databasing effort, and 3) working to publish online botanical resources that will be available to scientists, land managers, state and federal agencies, educators and the general public. These data will provide a greater understanding of one of the most botanically diverse regions of the earth and will lead to better research, better management planning and a more well-informed public. Development of a searchable collective database at a regional scale will provide a powerful research tool, and by combining 150 years of botanical information housed in herbaria in the Southeast with models of past plant migrations and current ecological parameters, we can revolutionize studies in biodiversity, evolution, ecology and systematics. We are also working to link our efforts with those of other regional herbarium groups through the US Virtual Herbarium and with the national biodiversity informatics effort, iDigBio.

Plant of the Day

What is this plant?

Click here to test your knowledge

General Data Usage Policy

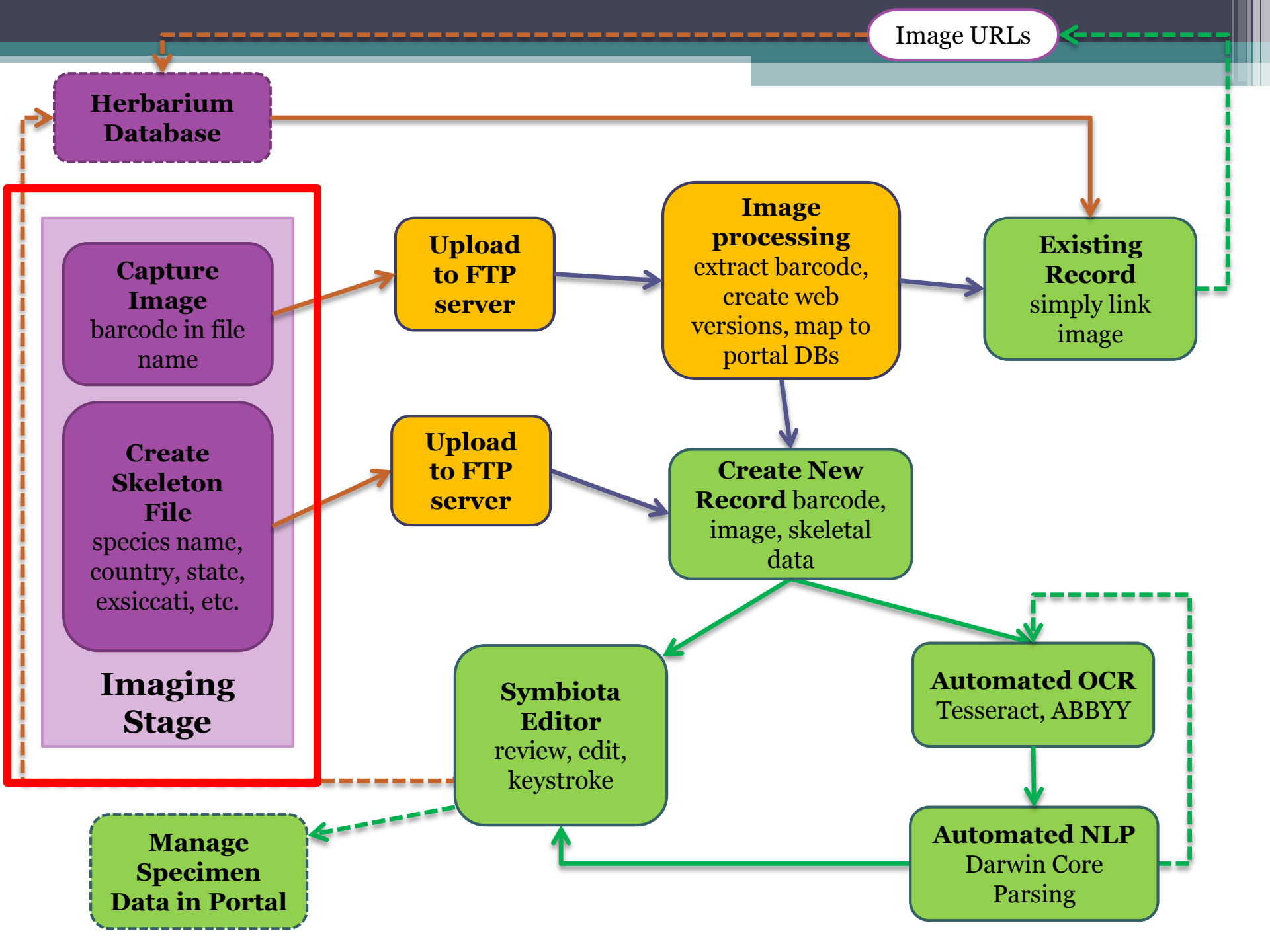


Image URLs
Herbarium Database
Capture Image barcode in file name
Create Skeleton File species name, country, state, exsiccati, etc.
Imaging Stage
Upload to FTP server
Upload to FTP server
Image processing extract barcode, create web versions, map to portal DBs
Existing Record simply link image
Create New Record barcode, image, skeletal data
Symbiota Editor review, edit, keystroke
Automated OCR Tesseract, ABBYY
Automated NLP Darwin Core Parsing
Manage Specimen Data in Portal

Skeletal Data File

- Imaging
  - File is named using barcode
  - Locally unique identifiers
  - JPGs
- Skeletal File / Stub data
  - CSV file containing:
    - Scientific name (filed by)
    - Country, state, county
    - Collector, number, date
  - LBCC imaging workflow Java Application

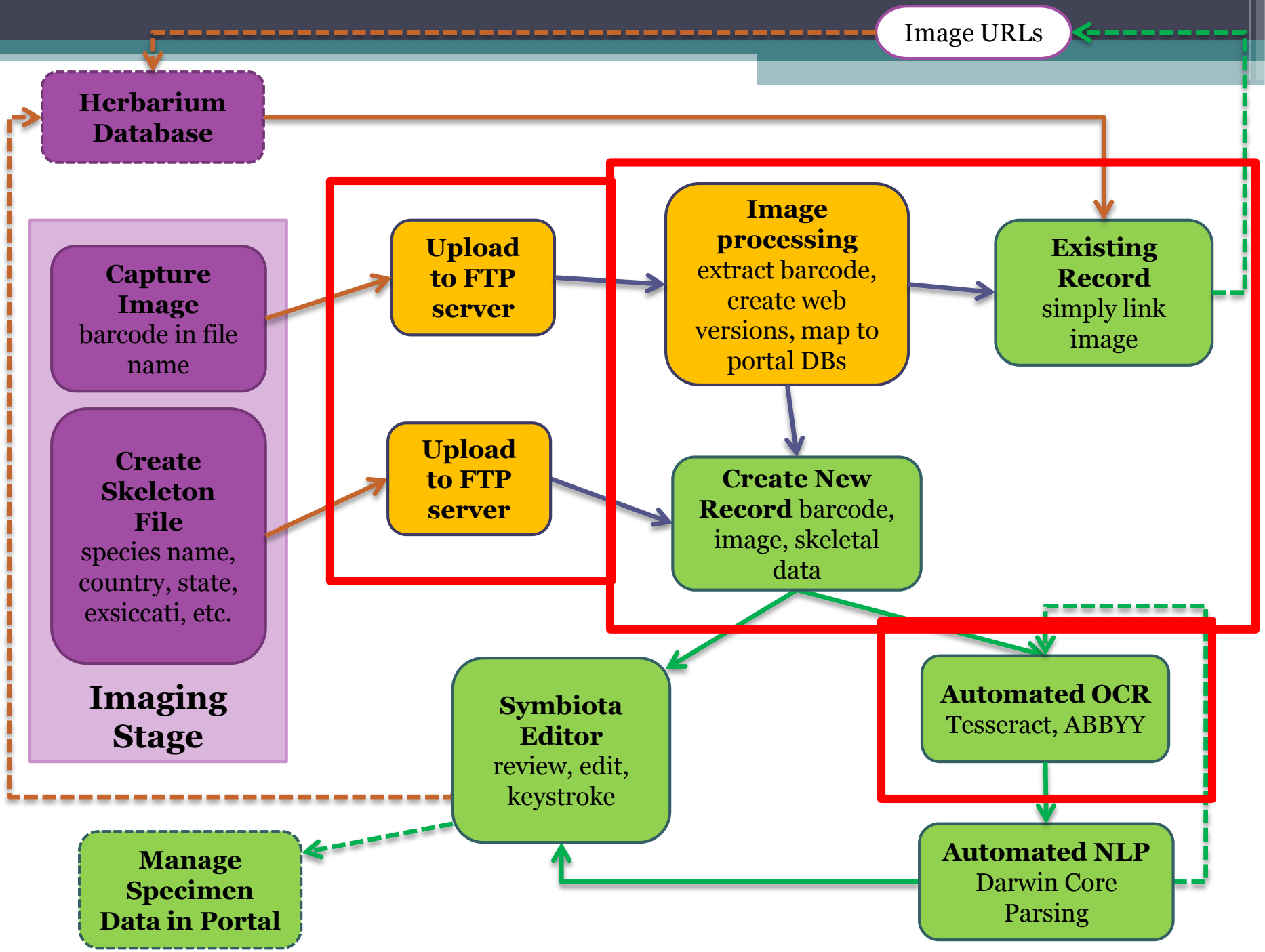


Image URLs
Herbarium Database
Capture Image barcode in file name
Create Skeleton File species name, country, state, exsiccati, etc.
Imaging Stage
Upload to FTP server
Upload to FTP server
Image processing extract barcode, create web versions, map to portal DBs
Existing Record simply link image
Create New Record barcode, image, skeletal data
Symbiota Editor review, edit, keystroke
Automated OCR Tesseract, ABBYY
Automated NLP Darwin Core Parsing
Manage Specimen Data in Portal

OCR: Optical Character Recognition

- Tesseract
  - Free
  - Integrated into portal
  - Good quality but with limits
- ABBYY
  - Windows only
  - Hotfolders – Corporate edition only
  - Batch processing scripts
  - Batch import

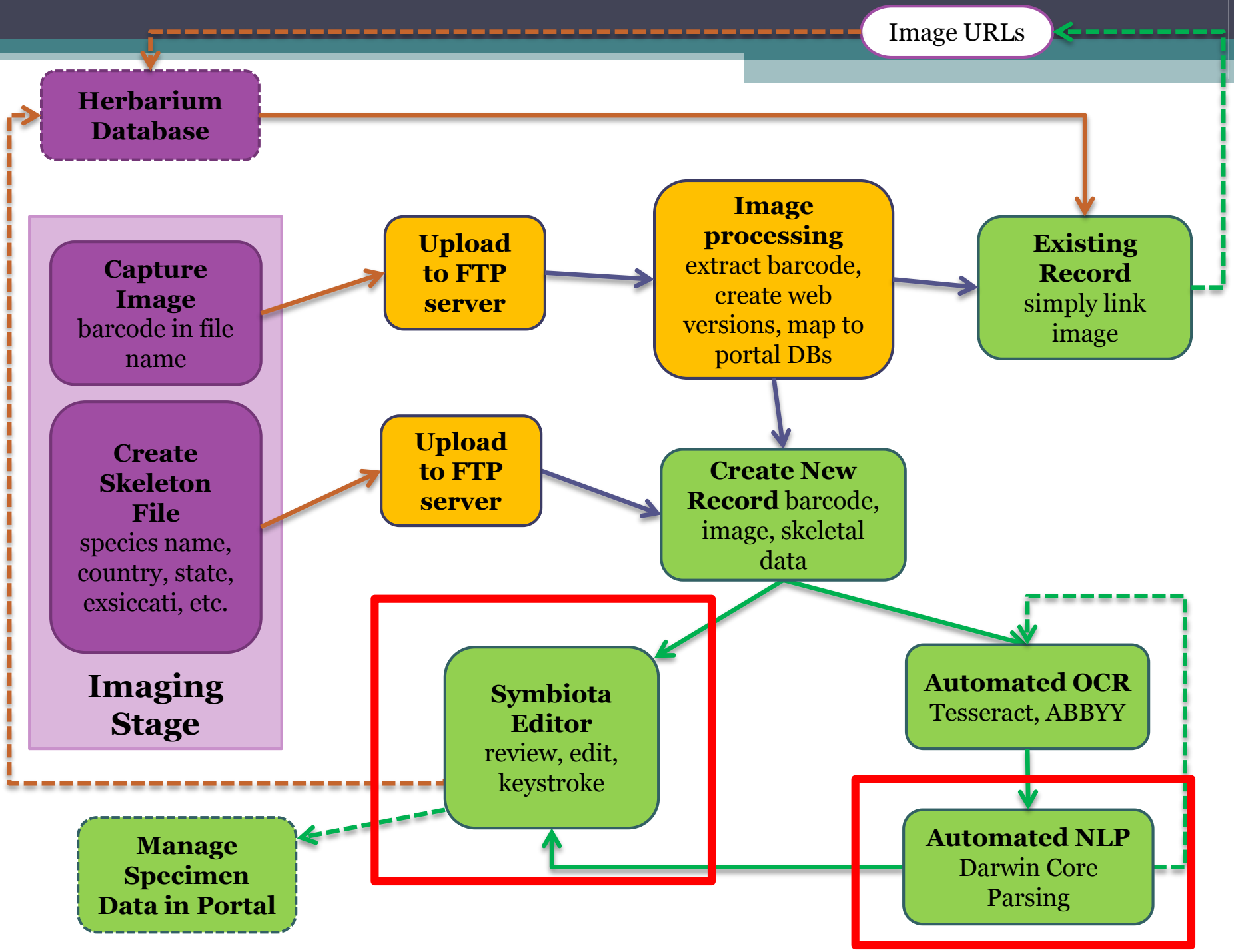
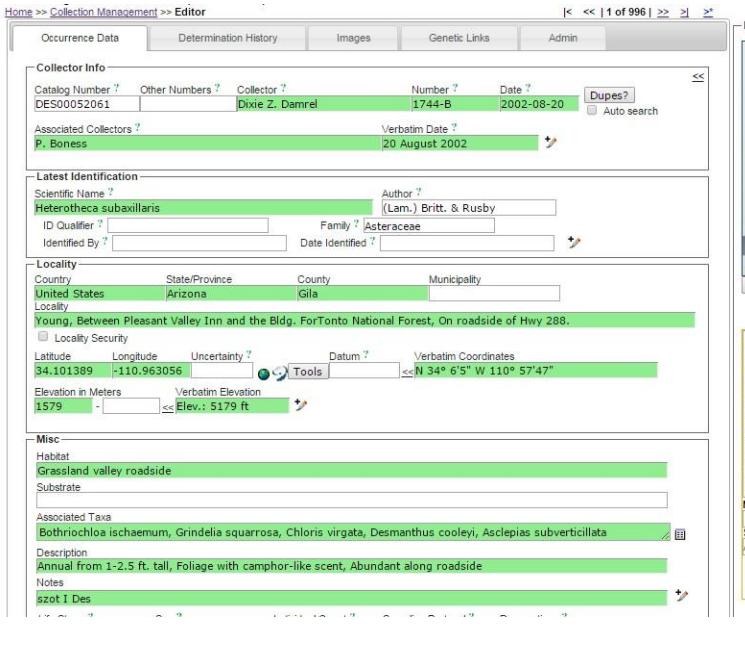
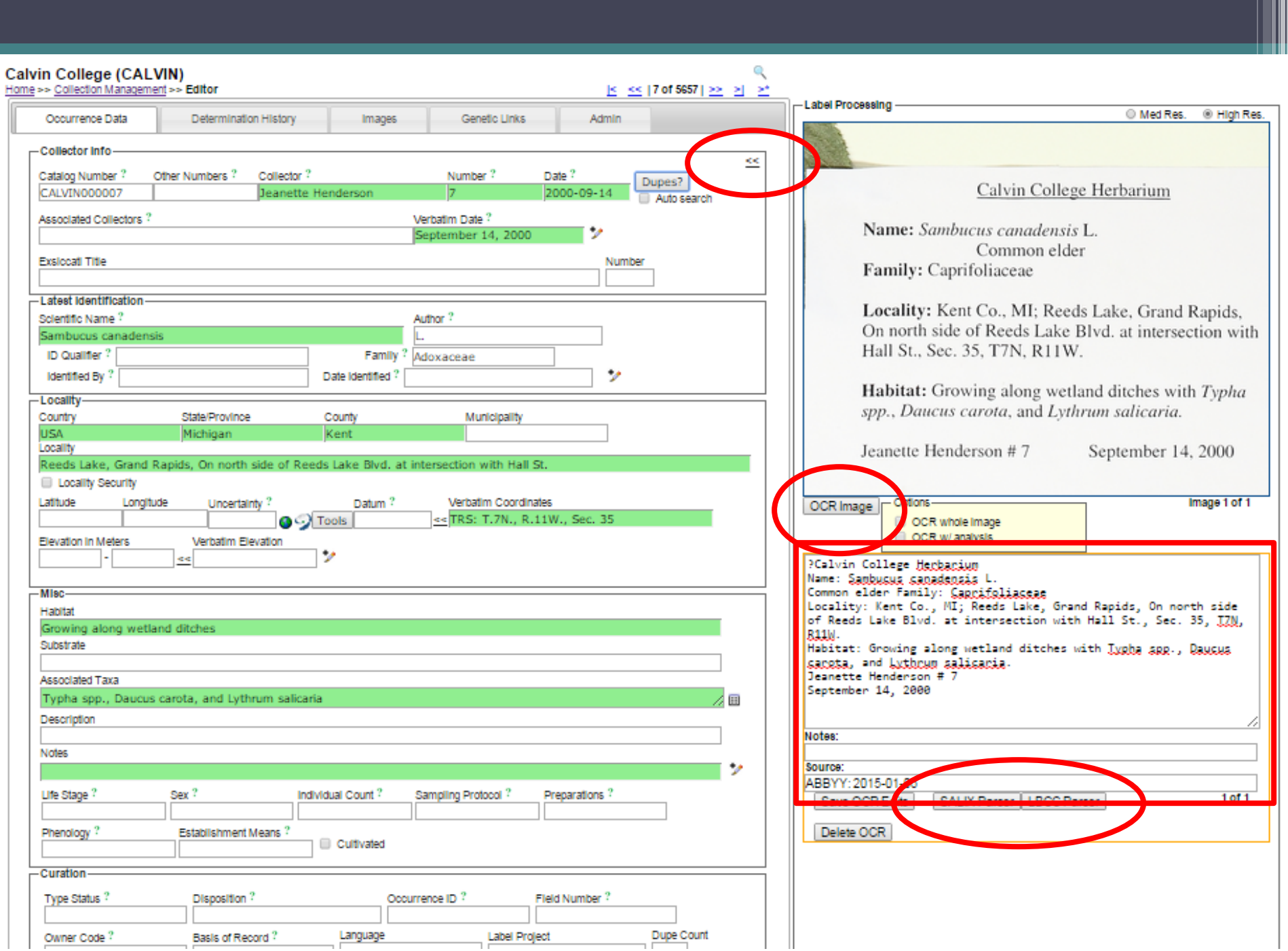


Image URLs
Herbarium Database
Capture Image barcode in file name
Create Skeleton File species name, country, state, exsiccati, etc.
Imaging Stage
Upload to FTP server
Upload to FTP server
Image processing extract barcode, create web versions, map to portal DBs
Existing Record simply link image
Create New Record barcode, image, skeletal data
Symbiota Editor review, edit, keystroke
Automated OCR Tesseract, ABBYY
Automated NLP Darwin Core Parsing
Manage Specimen Data in Portal

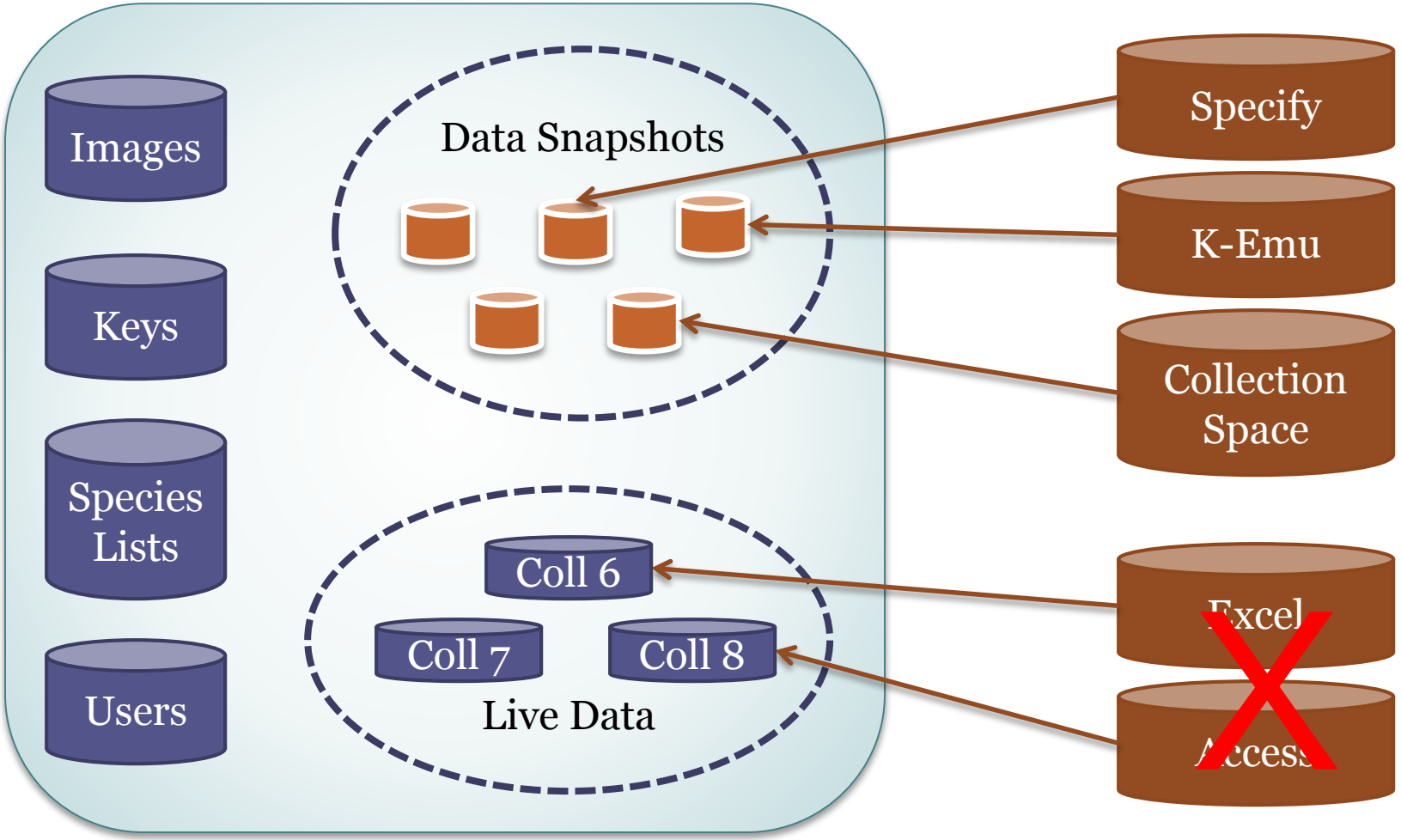
Symbiota - Biodiversity CMS

- Password Protected
- Online Browser-based application
- Platform independent
- Globally accessible
- No special software installation (free)
- Make use of web services




Calvin College (CALVIN)
Home >> Collection Management >> Editor
|< << | 7 of 5657 | >> >| >*
Occurrence Data
Determination History
Images
Genetic Links
Admin
Collector Info
Catalog Number ?
CALVIN000007
Other Numbers ?
Collector ?
Jeanette Henderson
Number ?
7
Date ?
2000-09-14
Dupes?
Auto search
Associated Collectors ?
Verbatim Date ?
September 14, 2000
Exsiccati Title
Number
Latest Identification
Scientific Name ?
Sambucus canadensis
Author ?
L.
ID Qualifier ?
Family ?
Adoxaceae
Identified By ?
Date Identified ?
Locality
Country
USA
State/Province
Michigan
County
Kent
Municipality
Locality
Reeds Lake, Grand Rapids, On north side of Reeds Lake Blvd. at intersection with Hall St.
Locality Security
Latitude
Longitude
Uncertainty ?
Tools
Datum ?
Verbatim Coordinates
TRS: T.7N., R.11W., Sec. 35
Elevation in Meters
Verbatim Elevation
Misc
Habitat
Growing along wetland ditches
Substrate
Associated Taxa
Typha spp., Daucus carota, and Lythrum salicaria
Description
Notes
Life Stage ?
Sex ?
Individual Count ?
Sampling Protocol ?
Preparations ?
Phenology ?
Establishment Means ?
Cultivated
Curation
Type Status ?
Disposition ?
Occurrence ID ?
Field Number ?
Owner Code ?
Basis of Record ?
Language
Label Project
Dupe Count
Label Processing
Med Res.
High Res.
Calvin College Herbarium
Name: Sambucus canadensis L.
Common elder
Family: Caprifoliaceae
Locality: Kent Co., MI; Reeds Lake, Grand Rapids, On north side of Reeds Lake Blvd. at intersection with Hall St., Sec. 35, T7N, R11W.
Habitat: Growing along wetland ditches with Typha spp., Daucus carota, and Lythrum salicaria.
Jeanette Henderson # 7
September 14, 2000
OCR Image
Options
OCR whole image
OCR w/ analysis
Image 1 of 1
?Calvin College Herbarium
Name: Sambucus canadensis L.
Common elder Family: Caprifoliaceae
Locality: Kent Co., MI; Reeds Lake, Grand Rapids, On north side of Reeds Lake Blvd. at intersection with Hall St., Sec. 35, T7N, R11W.
Habitat: Growing along wetland ditches with Typha spp., Daucus carota, and Lythrum salicaria.
Jeanette Henderson # 7
September 14, 2000
Notes:
Source:
ABBYY: 2015-01-
1 of 1
Delete OCR

Publishing: Snapshot vs Live



SERNEC

Southeast Regional Network of Expertise and Collections

Central Michigan University (CMC)

Data Editor Control Panel

- Add New Occurrence Record
- Edit Existing Occurrence Records
- Print Labels
- Batch Georeference Specimens
- Loan Management

Administration Control Panel

- Edit Metadata and Contact Information
- Update Statistics
- Manage Permissions
- Import/Update Specimen Records
  - Quick File Upload
- Processing Toolbox
- Darwin Core Archive Publishing
- Review/Verify General Specimen Edits
- Data Cleaning Tools
- Duplicate Clustering
- Download Backup Data File

The Central Michigan University Herbarium (CMC) strives to integrate research and education, providing centralized botanical resources, research facilities, and educational opportunities to the University and broader scientific community.

Contact: Anna Monfils, Director (monfi1ak<at>cmich.edu)

Home Page: http://cmcherbarium.bio.cmich.edu/

Central Michigan University

Introduction | Image Loading | Crowdsourcing Module | OCR | Reports | Exporter

Specimen Image Statistics

Total specimens with images: 18155
"unprocessed" specimens with images: 18154
with OCR: 18154
without OCR: 0

Custom Query:
Select Processing Status ▼ Reset Statistics

Batch OCR Images using the Tesseract OCR Engine

Processing Status: unprocessed ▼

Number of records to process: 100

Run Batch OCR

Note: This feature is dependent on the proper installation of the Tesseract OCR Engine on the hosting server

OCR Batch Processing

This inteface will process and load into the database OCR text files. ABBYY FineReader (Corporate Edition) includes the ABBYY HotFolder tool that can batch process multiple specimen label images to produce separate text file containing label text. However, this tool will upload OCR text obtained by other processes. OCR text is linked directly to the specimen image.

Requirements:

- OCR files must be in a text format with a .txt extension. Use the "Create a separate document for each file" and "Save as: Text (*.txt)" HotFolder settings.
- Files must be named using the Catalog Number

Following Actions

- Batch georeferencing
- Darwin Core Archive publishing
- Data quality cleaning
  - Scientific names, coordinates, dates
- Duplicate linking
- Darwin Core Archive publishing
- Data quality cleaning
- Duplicate linking
- Specimen comments
- Exsiccati