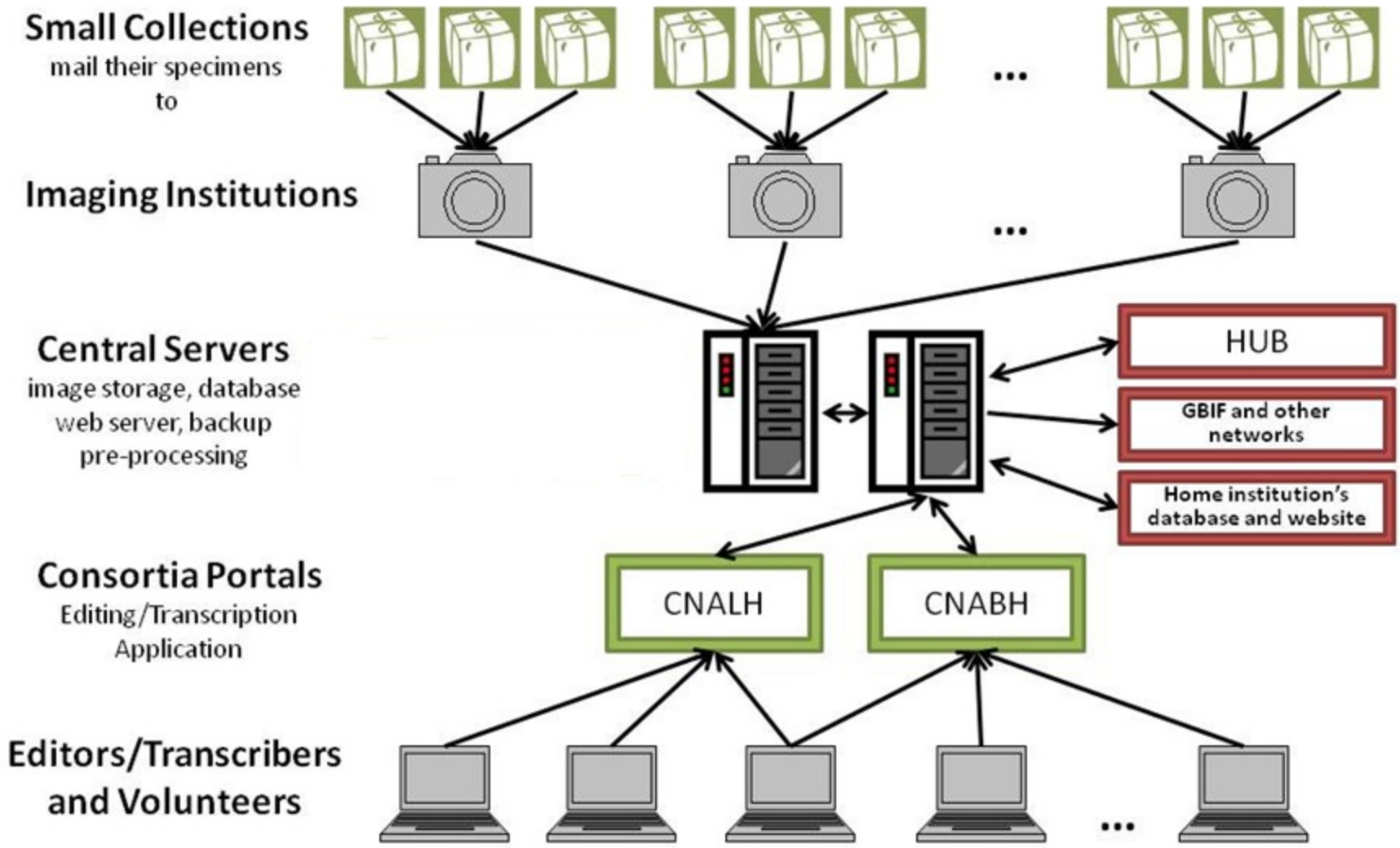
Symbiota and Specimen Label Digitization Workflows

National Science Foundation
WHERE DISCOVERIES BEGIN

Digitization Workflow



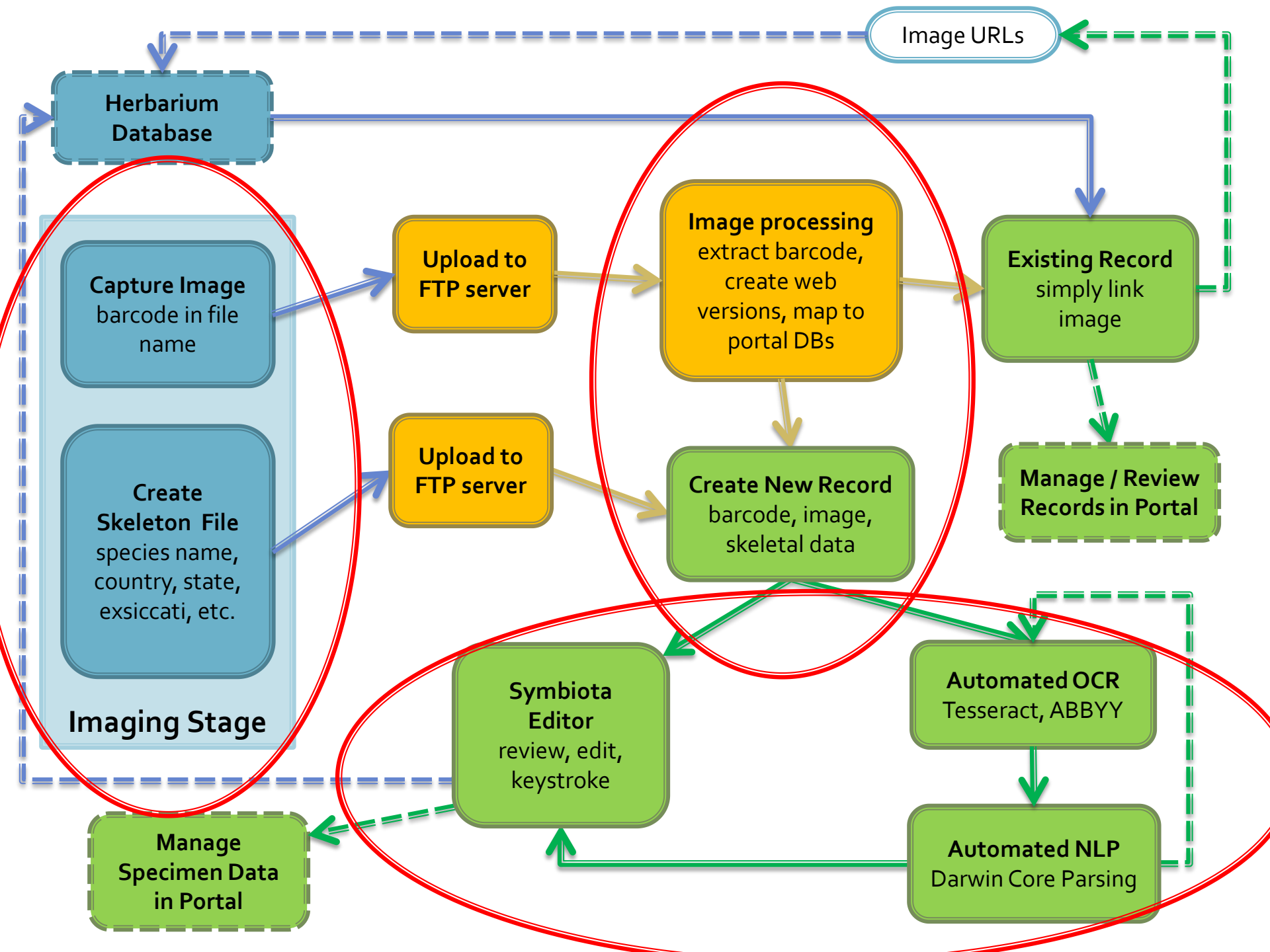


Image URLs
Herbarium Database
Capture Image barcode in file name
Create Skeleton File species name, country, state, exsiccati, etc.
Imaging Stage
Upload to FTP server
Upload to FTP server
Image processing extract barcode, create web versions, map to portal DBs
Existing Record simply link image
Create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

Symbiota - Biodiversity CMS

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

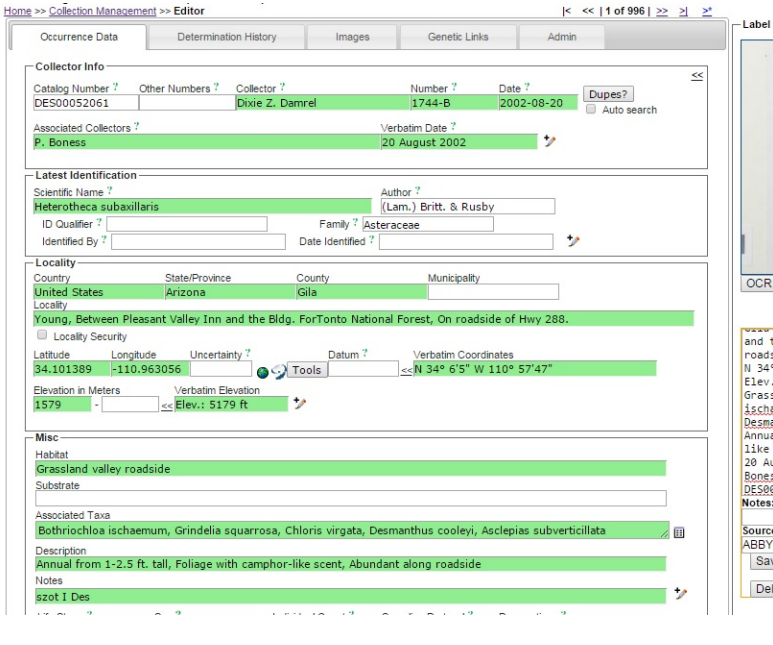


Image to Specimen Workflow

- Load an image
- Batch loading images
  - FTP image drop
  - iDigBio image loading
  - iPlant image loading
- Skeletal data capture
  - FTP skeletal file drop
  - Manual upload
  - LBCC application

Image to Specimen Workflow

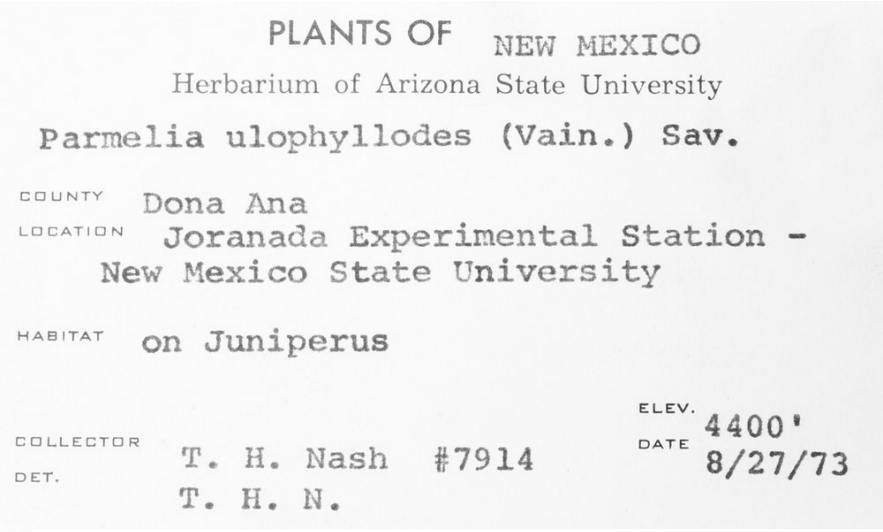
- OCR
  - Batch OCR
  - Tesseract vs ABBYY
- NLP (OCR parsing)
  - LBCC parser
  - SALIX
    - Wordstats tables

Miscellaneous Tasks

- Review management menu
- Batch uploading specimen data
- Specimen data entry
  - General review
  - Quality checks
  - Duplicate harvesting
  - OCR and OCR parsing
- Reviewing edits / data versioning
- Crowdsourcing

OCR - Introduction

- Optical Character Recognition
- Convert image of text into actual text
- OCR Engines
  - Tesseract
    - Google, open source, free
  - ABBYY
    - Proprietary, Windows or expensive
- Nightly Batch OCR

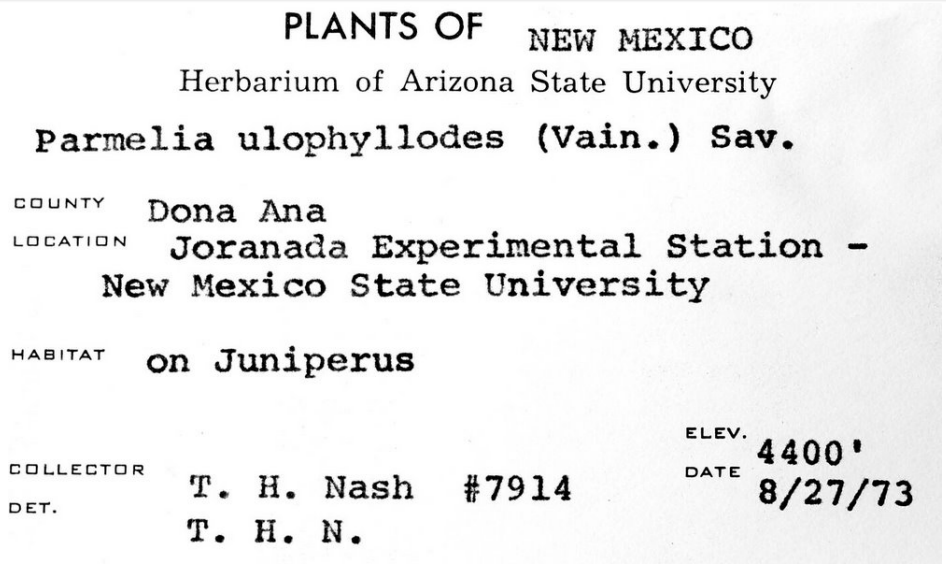


PLANTS OF NEW MEXICO
Herbarium of Arizona State University
Parmelia ulophyllodes (Vain.) Sav.
COUNTY Dona Ana
LOCATION Joranada Experimental Station - New Mexico State University
HABITAT on Juniperus
ELEV. 4400'
COLLECTOR T. H. Nash #7914 DATE 8/27/73
DET. T. H. N.

PLANTS OF New Mexico
Herbarium of Arizona State University
Parmelia ulophyllodes (Vain.) Sav.
COUNTY Dona Ana
Location Joranada Experimental Station - New Mexico State University
Habitat on Juniperus
ELEV. 4400'
Collector T. H. Nash #7914 DATE 8/27/73
Det. T. H. N.

OCR Challenges

- Issues
  - Old fonts
  - Faded labels
  - Form labels
  - Handwritten labels
  - Specialized terms
- Solutions
  - Image treatments
  - OCR tuning
  - Dictionaries
  - Consensus OCR



PLANTS OF NEW MEXICO
Herbarium of Arizona State University
Parmelia ulophyllodes (Vain.) Sav.
COUNTY Dona Ana
LOCATION Joranada Experimental Station -
New Mexico State University
HABITAT on Juniperus
ELEV. 4400'
COLLECTOR T. H. Nash #7914
DET. T. H. N.
DATE 8/27/73

Â PLANTS OF NEW r~1Exlco
fi Herbarium of Arizona State University
P Parmelia ulophyllodes (Vain.) Sav.
â COUNTY â€œÂ°â€â€œâ€œ
Z Joranada Experimental Station -
Â New Mexico State University
"' "â€œâ€œâ€œ' on Juniperus
V ELEV. ' 4400
n EEILLEETUR DATE
T DU T. H. Nash #7914 8/27/73

NLP Introduction

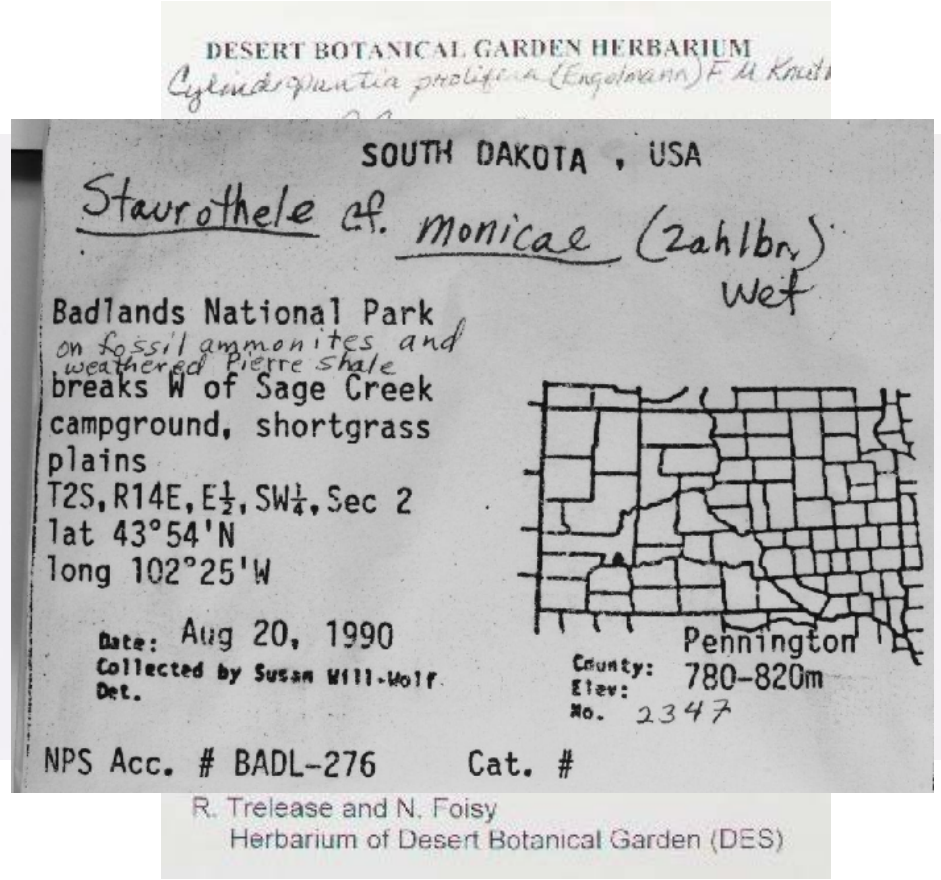
- Natural Language Processing
- Parse OCR text into target fields
- Augment / repair
  - OCR errors
  - Misspellings
  - Data type conversions

PLANTS OF New Mexico
Herbarium of Arizona State University
Parmelia ulophyllodes (Vain.) Sav.
COUNTY Dona Ana
Location Joranada Experimental Station -
New Mexico State University
Habitat on Juniperus
ELEV. 4400'
Collector T. H. Nash #7914 DATE 8/27/73
Det. T. H. N.

scientificName: Parmelia ulophyllodes (Vain.) Sav.
stateProvince: New Mexico
county: Dona Ana
Locality: Joranada Experimental Station
substrate: on Juniperus
verbatimElevation: 4400'
minimumElevationInMeters: 1341
recordedBy: T. H. Nash
recordNumber: 7914
eventDate: 1973-08-27
identifiedBy: T. H. Nash

NLP Challenges

- Issues
  - Variable layouts
  - Loose standards
  - OCR error
- Solutions
  - Authority tables
  - Levenshtein distance
  - Word stats
  - Format recognition
  - Parsing profiles
  - Duplicate harvesting



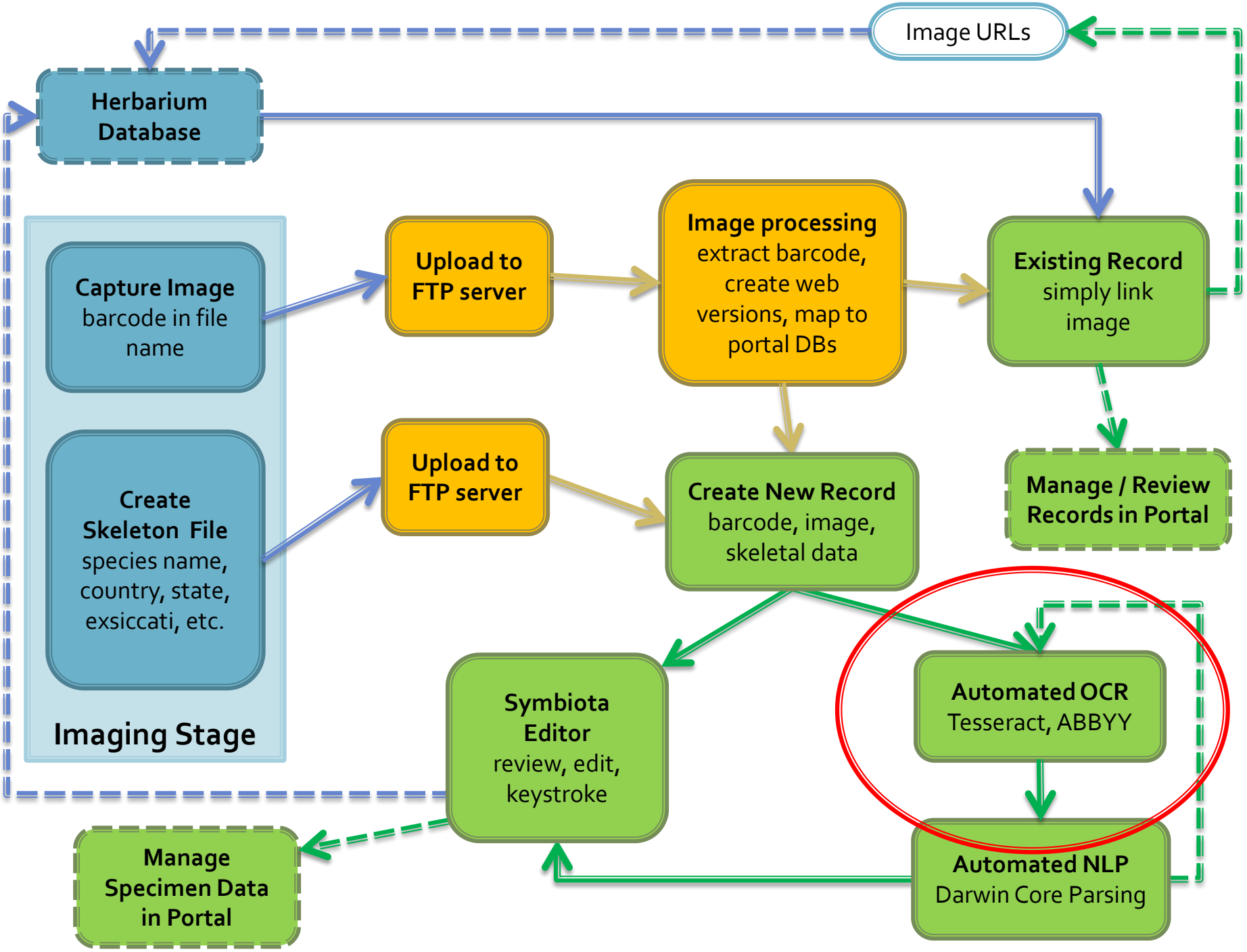
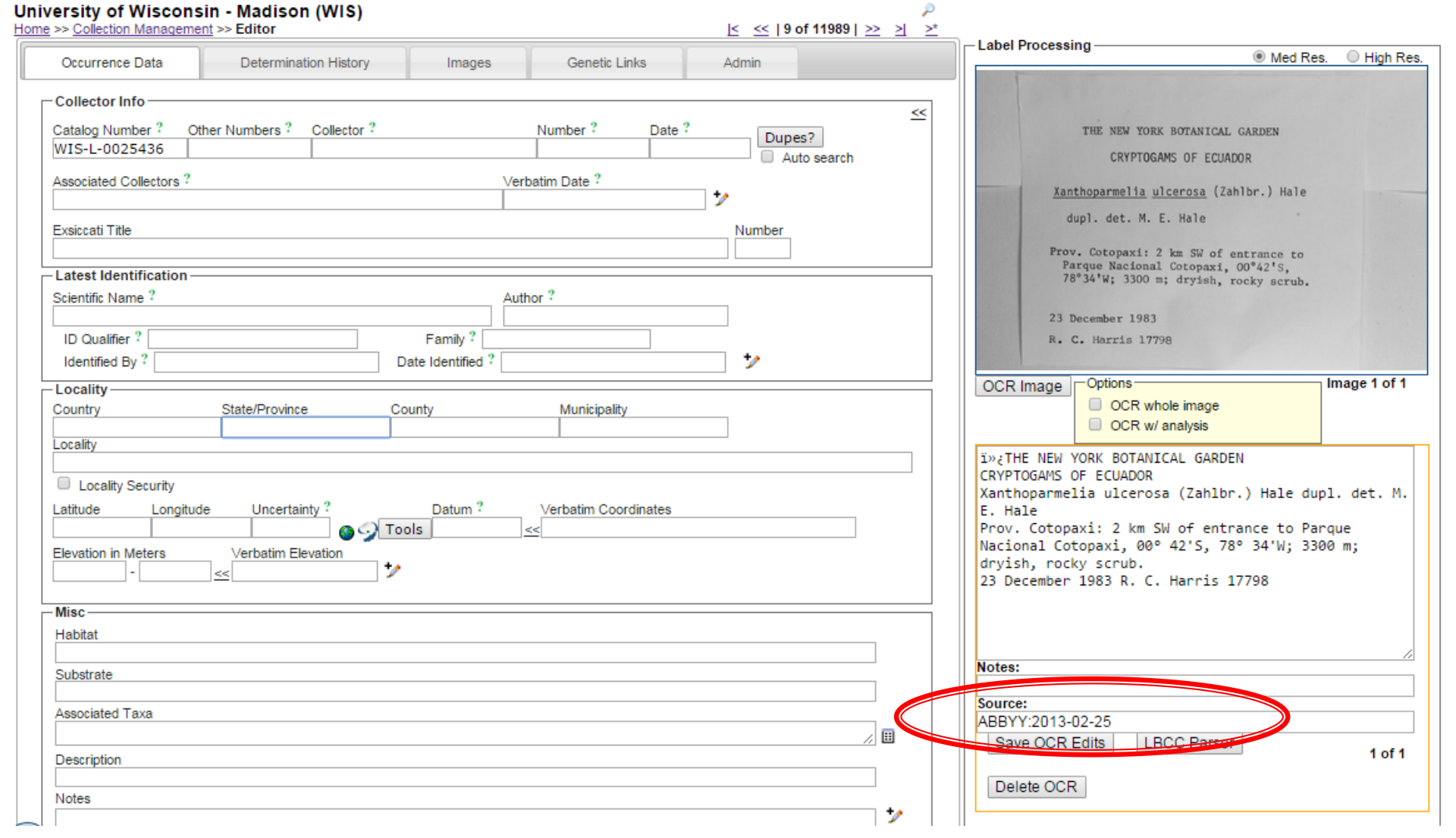


Image URLs
Herbarium Database
Capture Image
barcode in file name
Create Skeleton File
species name, country, state, exsiccati, etc.
Imaging Stage
Upload to FTP server
Upload to FTP server
Image processing
extract barcode, create web versions, map to portal DBs
Existing Record
simply link image
Create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

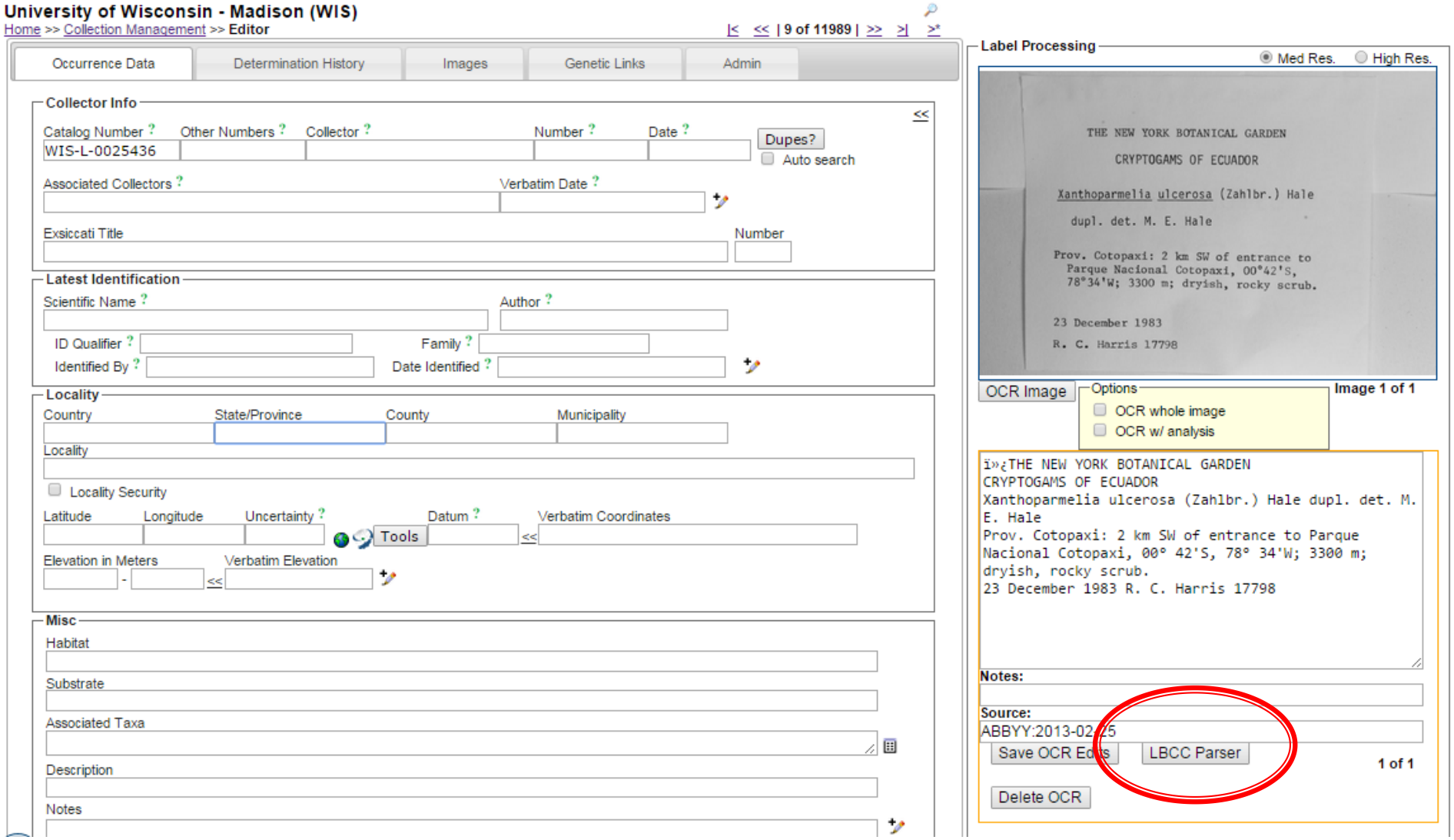
Ready for Processing



LBCC Parser

- Specifically tuned for lichen and bryophyte
- Logic, pattern matching, lookup tables
  - scientific name, collector, number, date , assoc. collectors, locality, coordinates, elevation, habitat, substrate, descriptions, general notes
- Recognizes label formats for specific collectors
- Programmer: Robert Anglin
  - UW-Madison

LBCC Parser


University of Wisconsin - Madison (WIS)
Home >> Collection Management >> Editor
|< << | 9 of 11989 | >> >| >*
Occurrence Data
Determination History
Images
Genetic Links
Admin
Collector Info
Catalog Number ?
Other Numbers ?
Collector ?
Number ?
Date ?
Dupes?
Auto search
WIS-L-0025436
Associated Collectors ?
Verbatim Date ?
Exsiccati Title
Number
Latest Identification
Scientific Name ?
Author ?
ID Qualifier ?
Family ?
Identified By ?
Date Identified ?
Locality
Country
State/Province
County
Municipality
Locality
Locality Security
Latitude
Longitude
Uncertainty ?
Tools
Datum ?
Verbatim Coordinates
Elevation in Meters
Verbatim Elevation
Misc
Habitat
Substrate
Associated Taxa
Description
Notes
Label Processing
Med Res.
High Res.
THE NEW YORK BOTANICAL GARDEN
CRYPTOGAMS OF ECUADOR
Xanthoparmelia ulcerosa (Zahlbr.) Hale
dupl. det. M. E. Hale
Prov. Cotopaxi: 2 km SW of entrance to Parque Nacional Cotopaxi, 00°42'S, 78°34'W; 3300 m; dryish, rocky scrub.
23 December 1983
R. C. Harris 17798
OCR Image
Options
OCR whole image
OCR w/ analysis
Image 1 of 1
ï»¿THE NEW YORK BOTANICAL GARDEN
CRYPTOGAMS OF ECUADOR
Xanthoparmelia ulcerosa (Zahlbr.) Hale dupl. det. M. E. Hale
Prov. Cotopaxi: 2 km SW of entrance to Parque Nacional Cotopaxi, 00° 42'S, 78° 34'W; 3300 m; dryish, rocky scrub.
23 December 1983 R. C. Harris 17798
Notes:
Source:
ABBYY:2013-02-25
Save OCR Edits
LBCC Parser
1 of 1
Delete OCR

LBCC Parser

University of Wisconsin - Madison (WIS)

Home >> Collection Management >> Editor

|< << | 9 of 11989 | >> >| >*

Occurrence Data | Determination History | Images | Genetic Links | Admin

Collector Info

Catalog Number	Other Numbers	Collector	Number	Date
WIS-L-0025436		Richard C Harris	17798	1983-12-23

Dupes? Auto search

Associated Collectors | Verbatim Date

Exsiccati Title | Number

Latest Identification

Scientific Name: Xanthoparmelia ulcerosa | Author: (Zahlbr.) Hale

ID Qualifier | Family: Parmeliaceae

Identified By: M. E. Hale | Date Identified

Locality

Country	State/Province	County	Municipality
Ecuador	Cotopaxi		

Locality: Prov. Cotopaxi: 2 km SW of entrance to Parque Nacional Cotopaxi

Locality Security

Latitude	Longitude	Uncertainty	Datum	Verbatim Coordinates
0.7	-78.566667		Tools	00° 42'S, 78° 34'W

Elevation in Meters: 3300 - | Verbatim Elevation: 3300 m

Misc

Habitat: dryish, rocky scrub.

Substrate

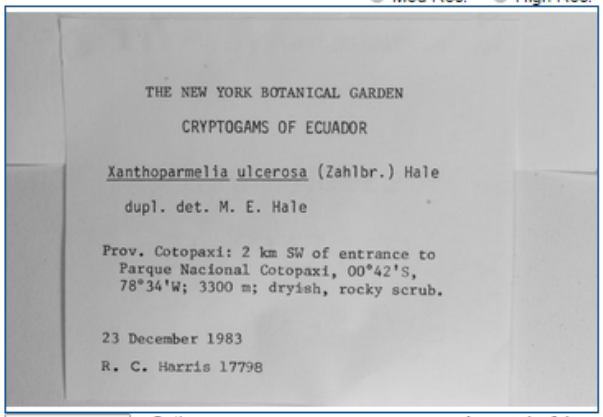
Associated Taxa

Description

Notes

Label Processing

Med Res. High Res.



OCR Image | Options: OCR whole image, OCR w/ analysis | Image 1 of 1

ï»¿THE NEW YORK BOTANICAL GARDEN
CRYPTOGAMS OF ECUADOR
Xanthoparmelia ulcerosa (Zahlbr.) Hale dupl. det. M. E. Hale
Prov. Cotopaxi: 2 km SW of entrance to Parque Nacional Cotopaxi, 00° 42'S, 78° 34'W; 3300 m; dryish, rocky scrub.
23 December 1983 R. C. Harris 17798

Notes:

Source: ABBYY:2013-02-25

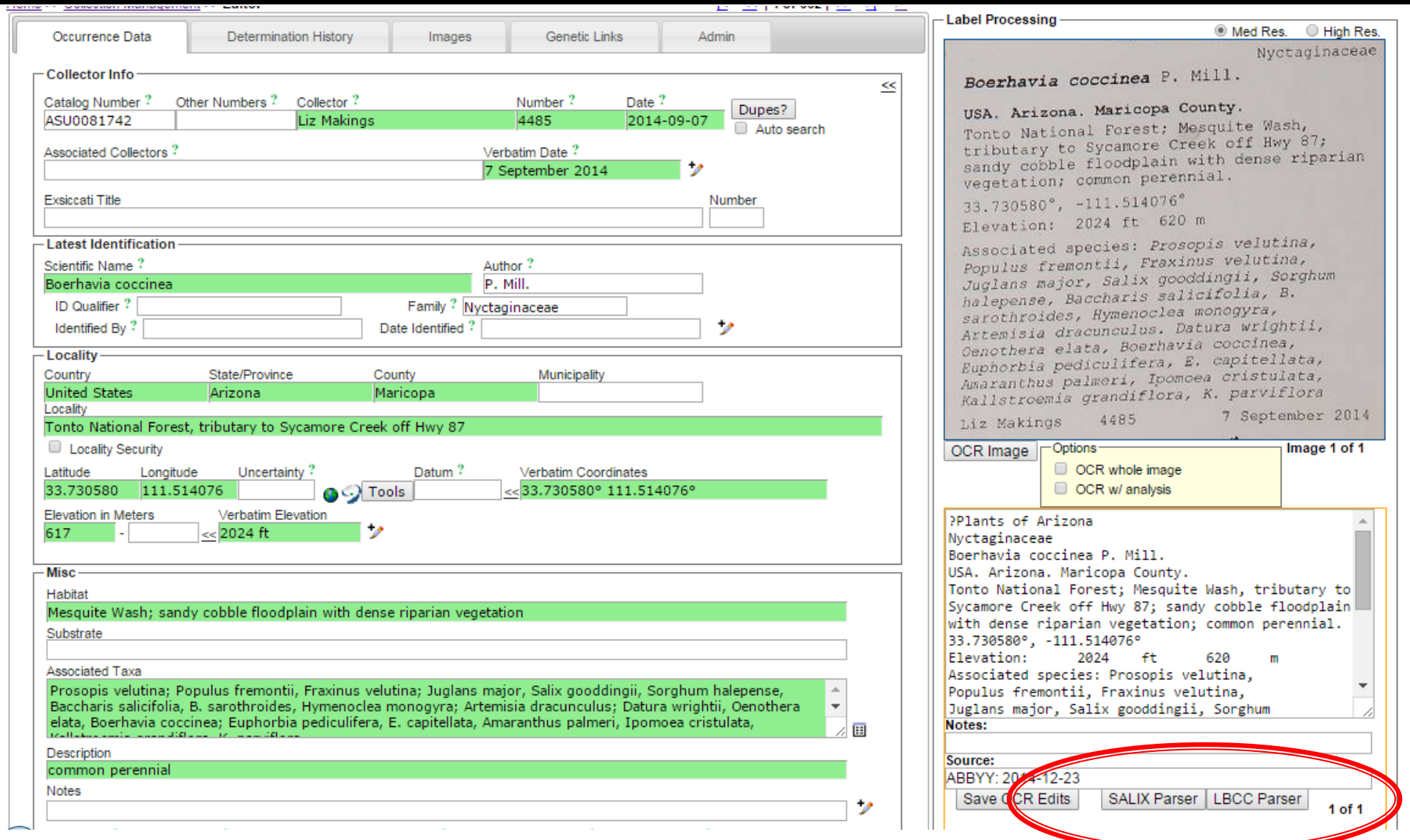
Save OCR Edits | LBCC Parser | 1 of 1

Delete OCR

SALIX - overview

- Logic, pattern matching, lookup tables
  - Scientific name, collector, number, date, coordinates, elevation
- Word frequency tables
  - locality, habitat, substrate, description, notes
- Daryl Lafferty
  - Arizona State University associate
- Open source PHP class
  - Input: OCR text block
  - Output: Darwin Core array

SALIX - example



SALIX – word frequency table

Fields: locality, habitat, substrate, description, notes

First Word	Second	Locality	habitat
creek		49992 (72%)	16990 (24)
eagle	creek	239 (98%)	5 (2%)
creek	bottom	85 (12%)	544 (76%)
along	creek	315 (14%)	1530 (69%)
sycamore		1453 (76%)	354 (19%)
under	sycamore	1 (8%)	9 (75%)
riparian		874 (7%)	9185 (73%)
pedro	riparian	242 (66%)	1 (0%)
mesquite		386 (14%)	1841 (66%)

SALIX - review

- Previously completed records required to build word frequency table
- Tuned and tailored to portal
  - Word stats specific to portal data
- Improves with use
  - Word stats adjusted with new records

OCR Filtering

- Theme filtering
- Word clouds
- Target similar label formats
- Use raw OCR to locate “Nash” labels
- Exclude:
  - Determined by Nash
  - Author of scientific name
  - Associated collector
  - County

Combined methods

- Batch processing
- Duplicate harvesting
  - Last name, number, date
  - Exact duplicates or duplicate events
- High similarity indexes
- OCR block comparison
- Consensus record