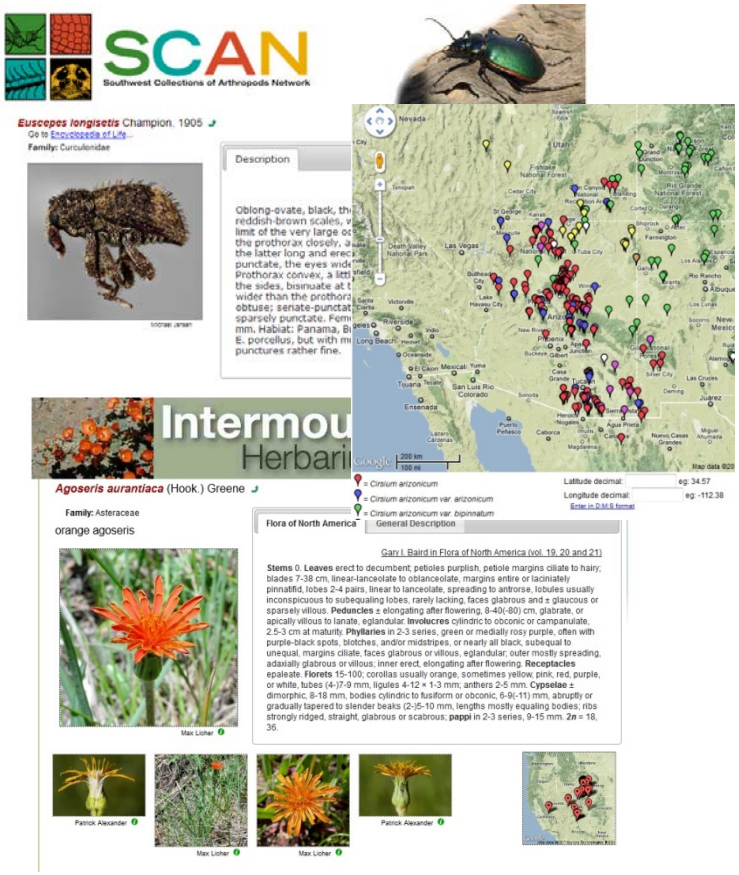
Edward Gilbert
Robert Anglin
Steven Gottschalk
Daryl Lafferty
Nico Franz
Corinna Gries
Leslie Landrum
Barbara Thiers

# Specimen label digitization using OCR/NLP tools integrated within the Symbiota processing toolkit

National Science Foundation
WHERE DISCOVERIES BEGIN

# Symbiota Software Project

- Open source
- Virtual flora/fauna
- Specimen search engine
- Biodiversity inventories
- Identification keys
- Images, maps, descriptions, common names, etc

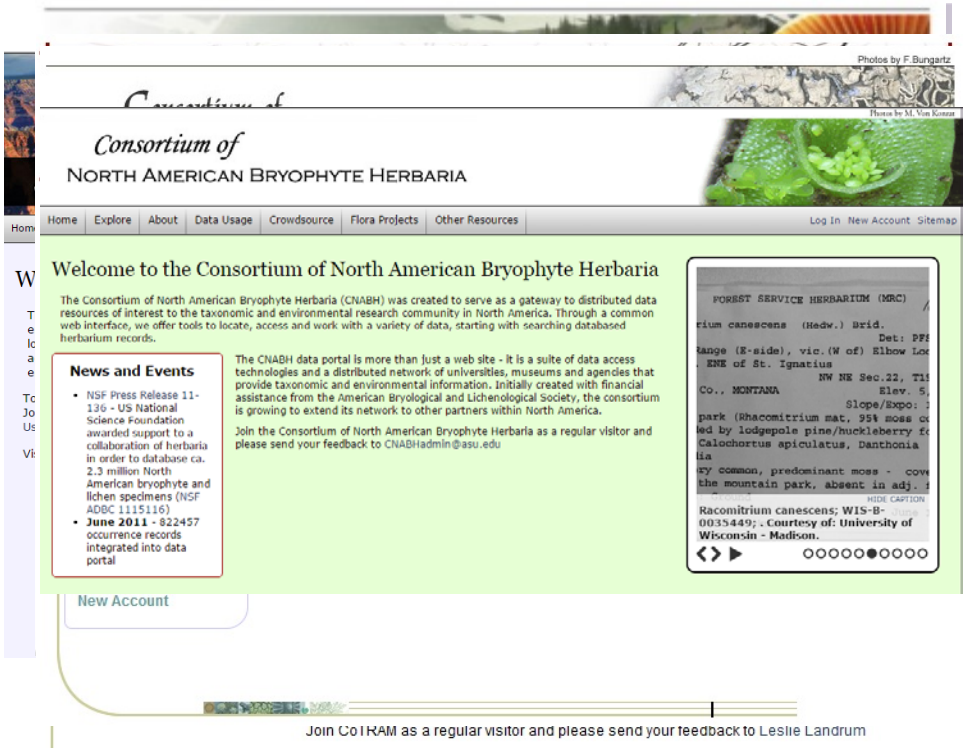


# Specimen Centric Model

- Backbone of Floristic Research
- Baseline data
- Proof of Occurrence
- Expert Reviewed
- Verifiable
- Millions occurrence records

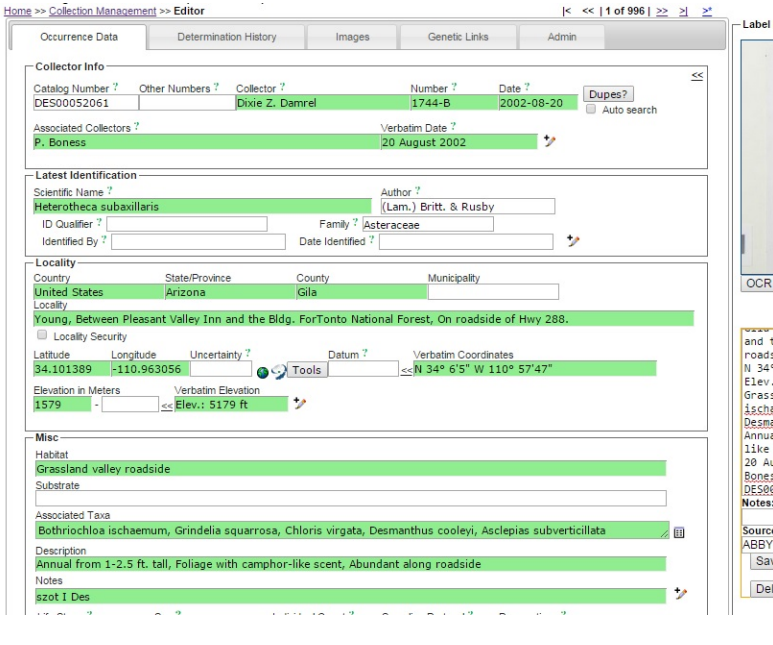
# Scientific Community Portals

- Community portals
- Distinct datasets
  - Taxonomic scope
  - Geographic scope
- Custom look & feel
  - CSS
  - configuration files

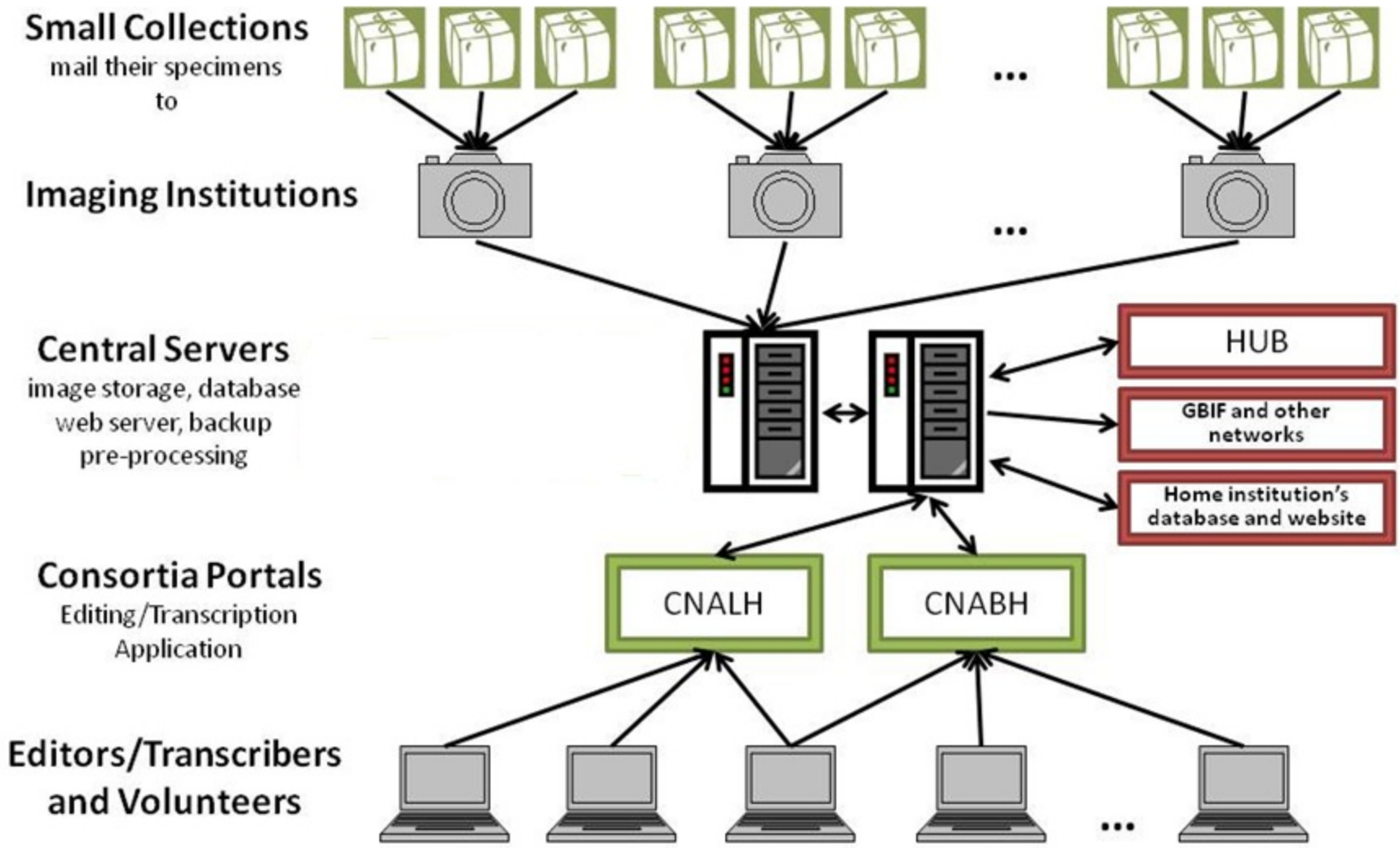


# Symbiota - Biodiversity CMS

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



# Digitization Workflow



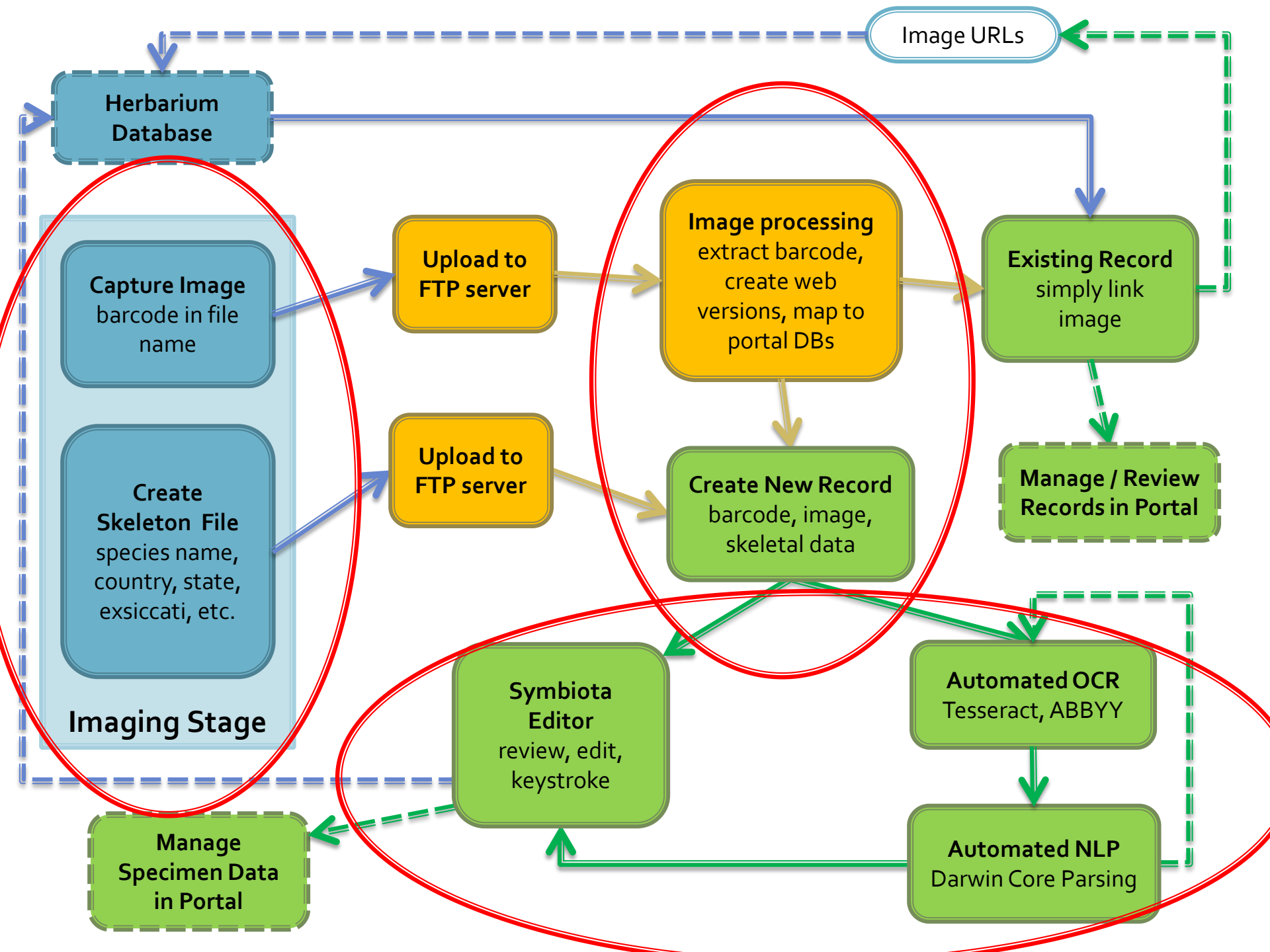


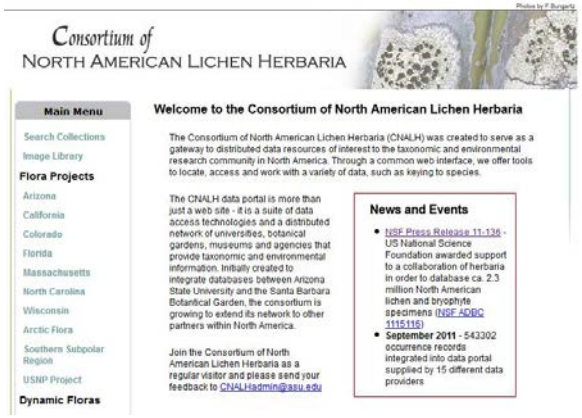
Image URLs
Herbarium Database
Capture Image barcode in file name
Create Skeleton File species name, country, state, exsiccati, etc.
Imaging Stage
Upload to FTP server
Upload to FTP server
Image processing extract barcode, create web versions, map to portal DBs
Existing Record simply link image
Create 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

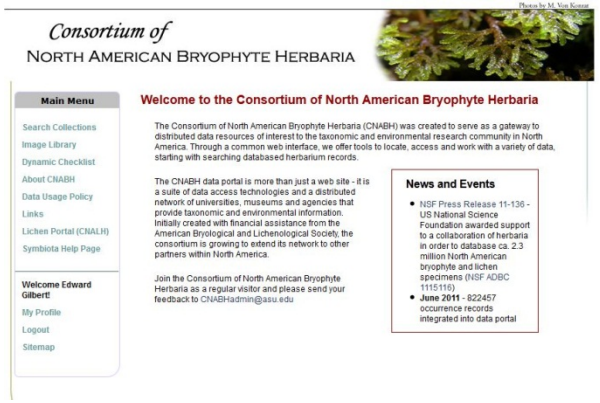
# Lichen and Bryophyte Parser

- NSF ADBC (#1115116)
- ~ 2.3 million specimen
  - 90% of all specimens
  - 900,000 lichens
  - 1.4 million bryophytes
- > 60 non-governmental US herbaria (95%)
  - Mexico, US, Canada
- 16 digitization centers

# National Portals

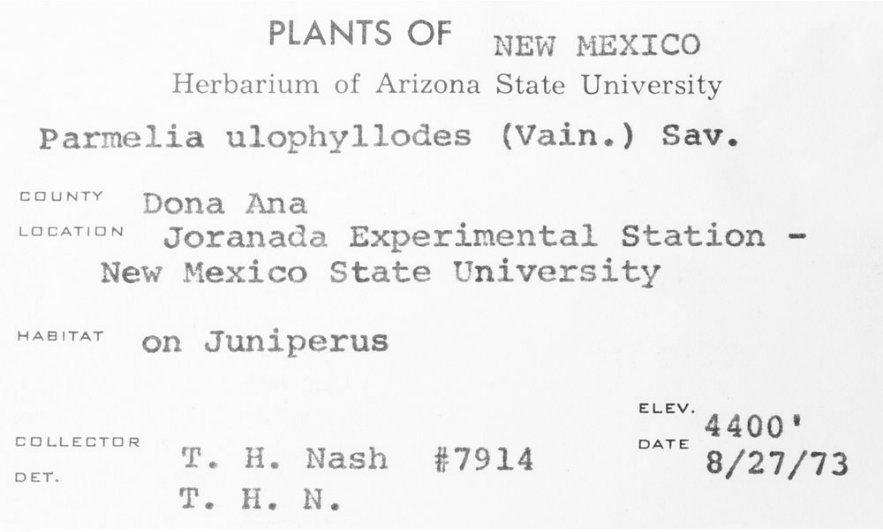
- Lichen Consortium
  - http://lichenportal.org
  - 73 collections
  - 1.7 million records
  - 338,000 unprocessed
- Bryophyte Consortium
  - http://bryophyteportal/
  - 71 collections
  - 2.2 million records
  - 725,000 unprocessed



# 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

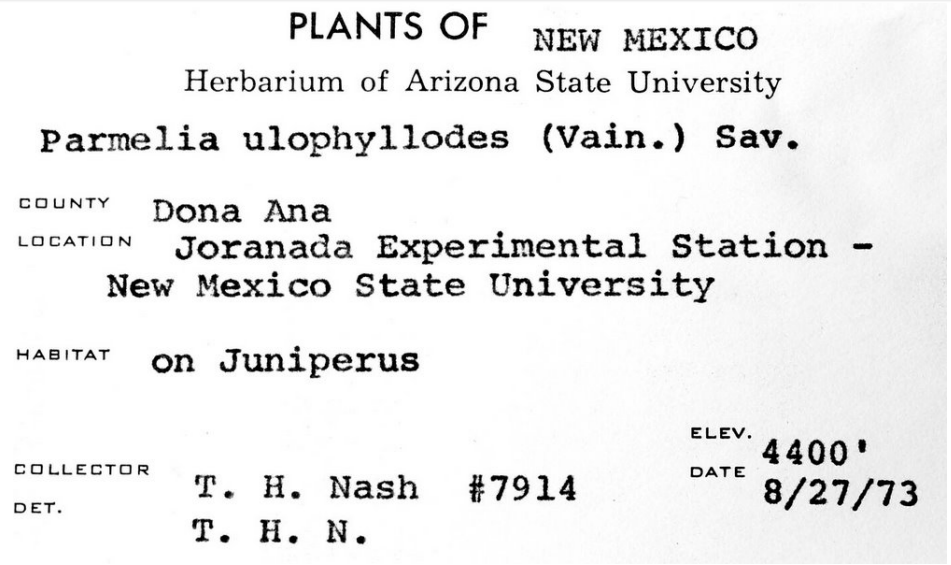


PLANTS OF NEW MEXICO
Herbarium of Arizona State University
Parmelia ulophyllodes (Vain.) Sav.
COUNTY Dona Ana
LOCATION Joranada Experimental Station - New Mexico State University
HABITAT on Juniperus
ELEV. 4400'
COLLECTOR T. H. Nash #7914 DATE 8/27/73
DET. T. H. N.

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

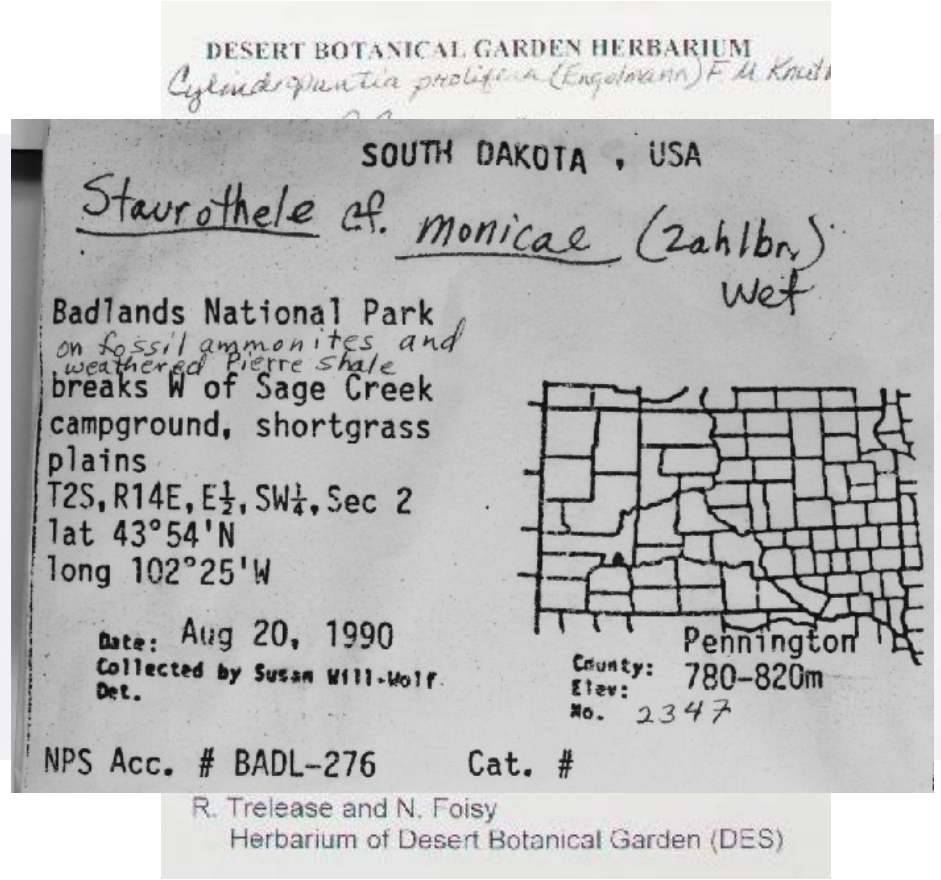
- Natural Language Processing
- Parse OCR text into target fields
- Augment / repair
  - OCR errors
  - Misspellings
  - Data type conversions

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



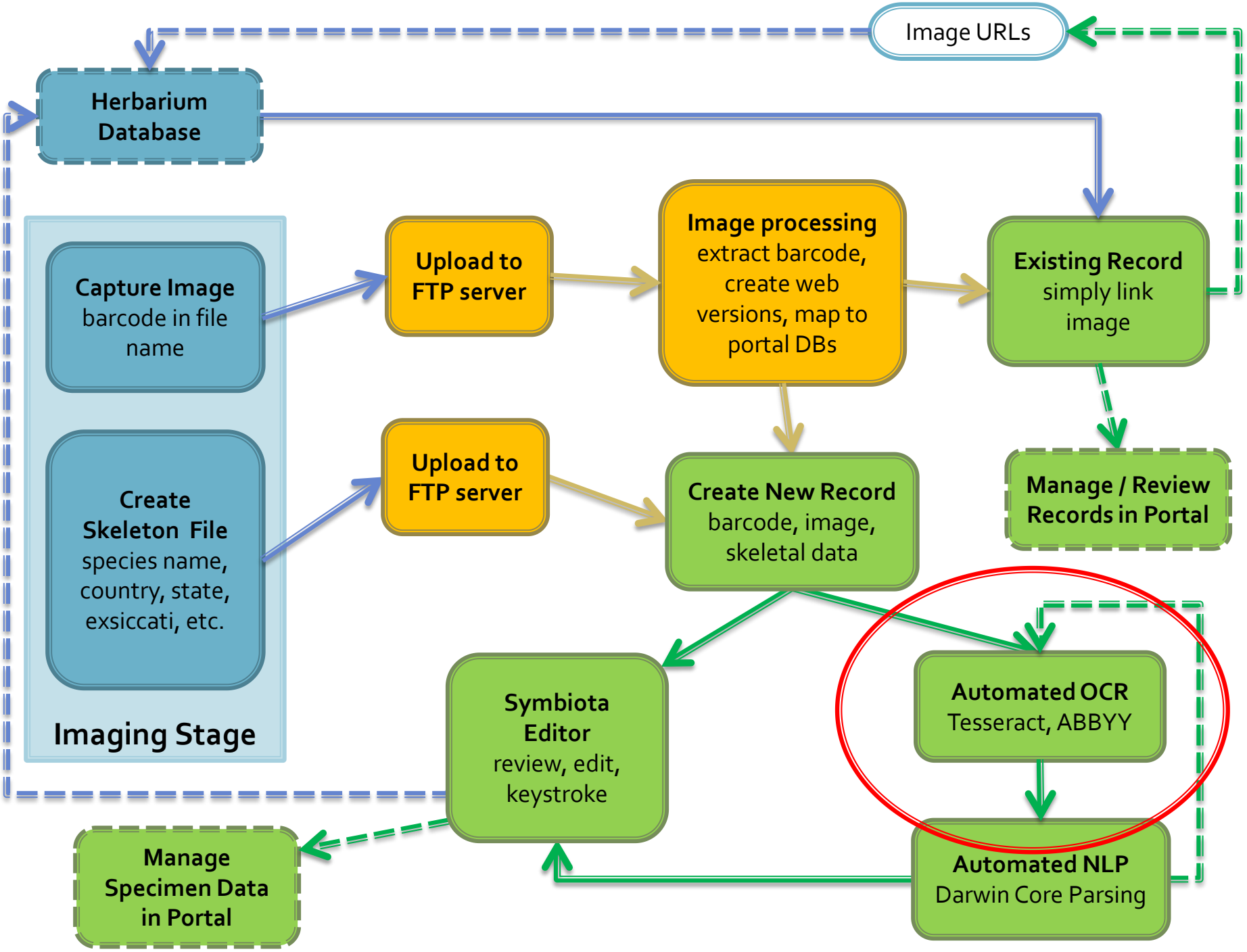
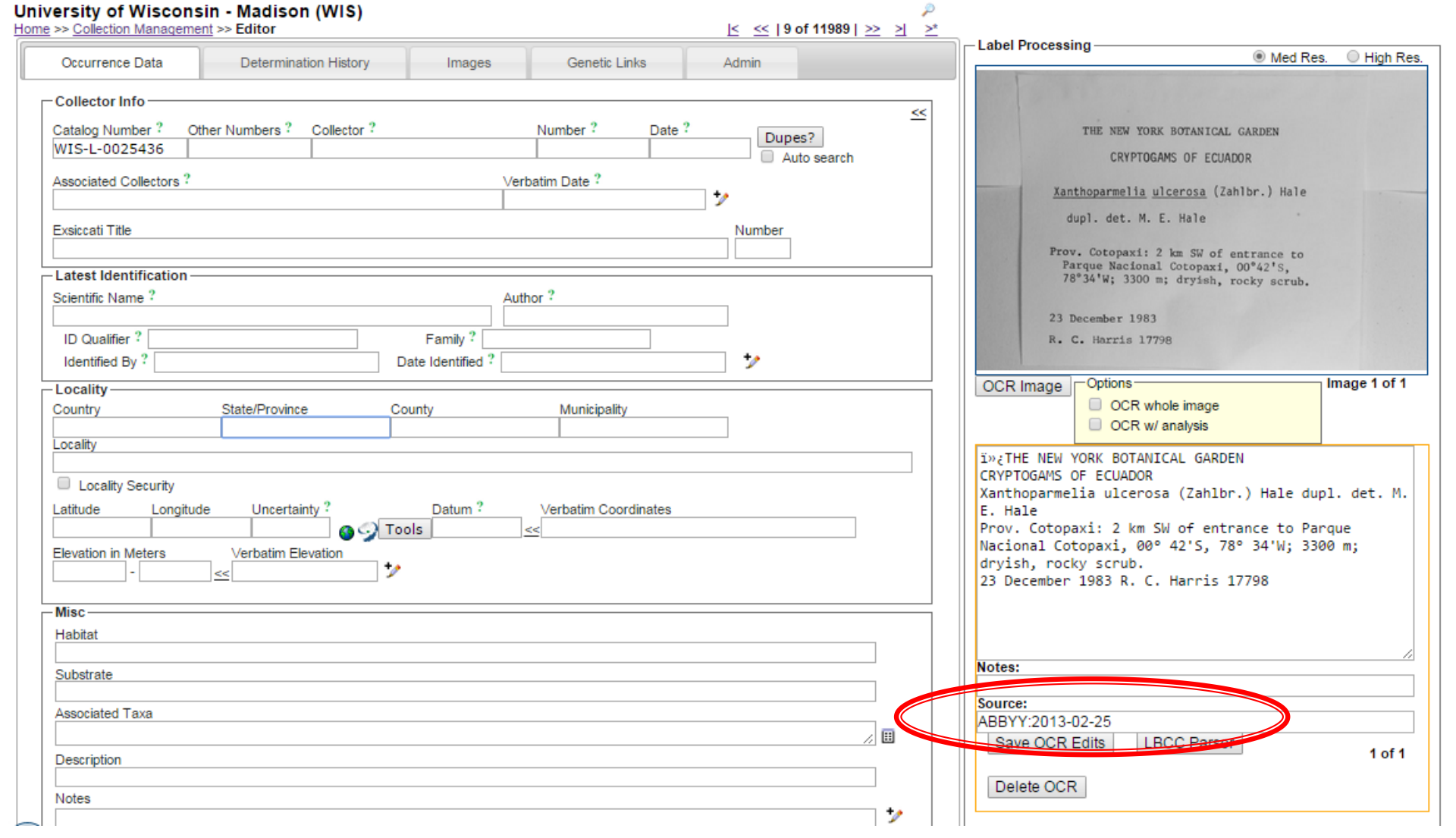


Image URLs
Herbarium Database
Capture Image
barcode in file name
Create Skeleton File
species name, country, state, exsiccati, etc.
Imaging Stage
Upload to FTP server
Upload to FTP server
Image processing
extract barcode, create web versions, map to portal DBs
Existing Record
simply link image
Create 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

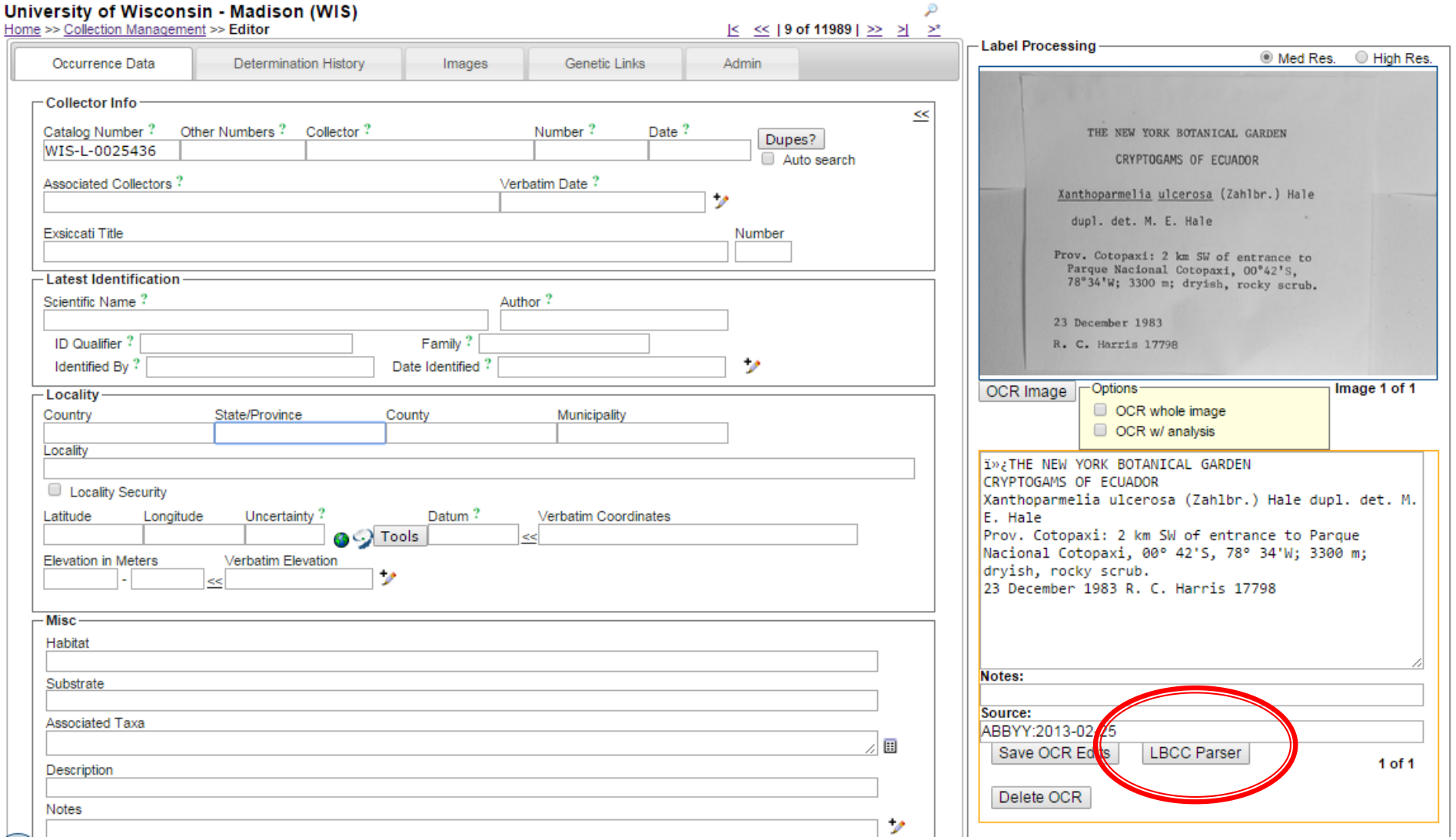
# Ready for Processing



# 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 | | Verbatim Elevation |
|---|---|---|
| 3300 | - | 3300 m |

**Misc**

Habitat: dryish, rocky scrub.

Substrate

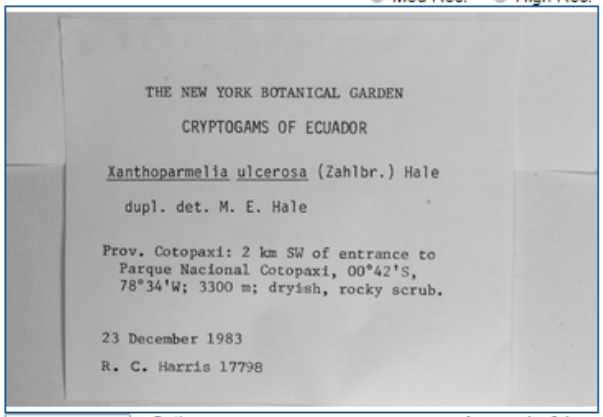
Associated Taxa

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

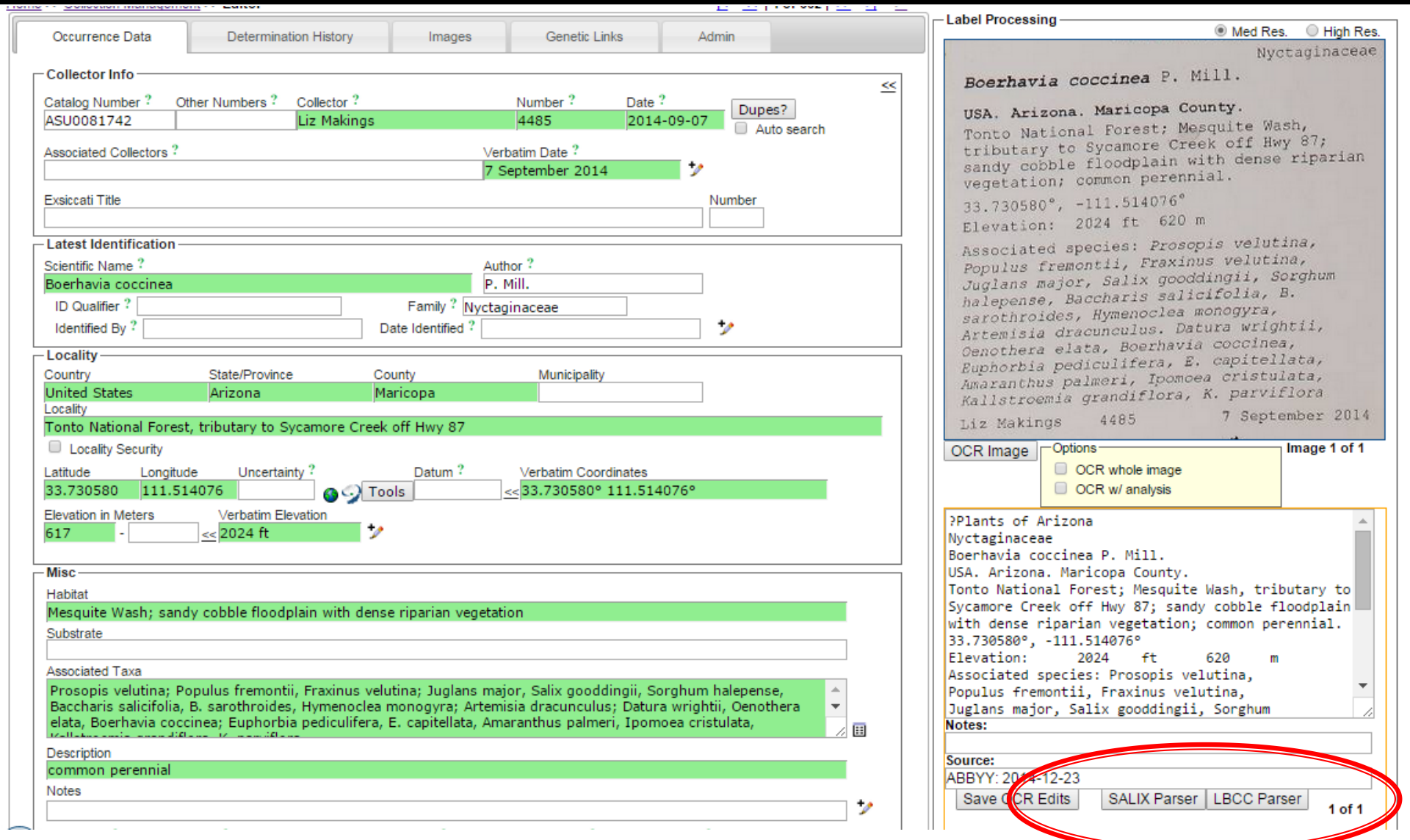
Save OCR Edits | LBCC Parser | 1 of 1

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

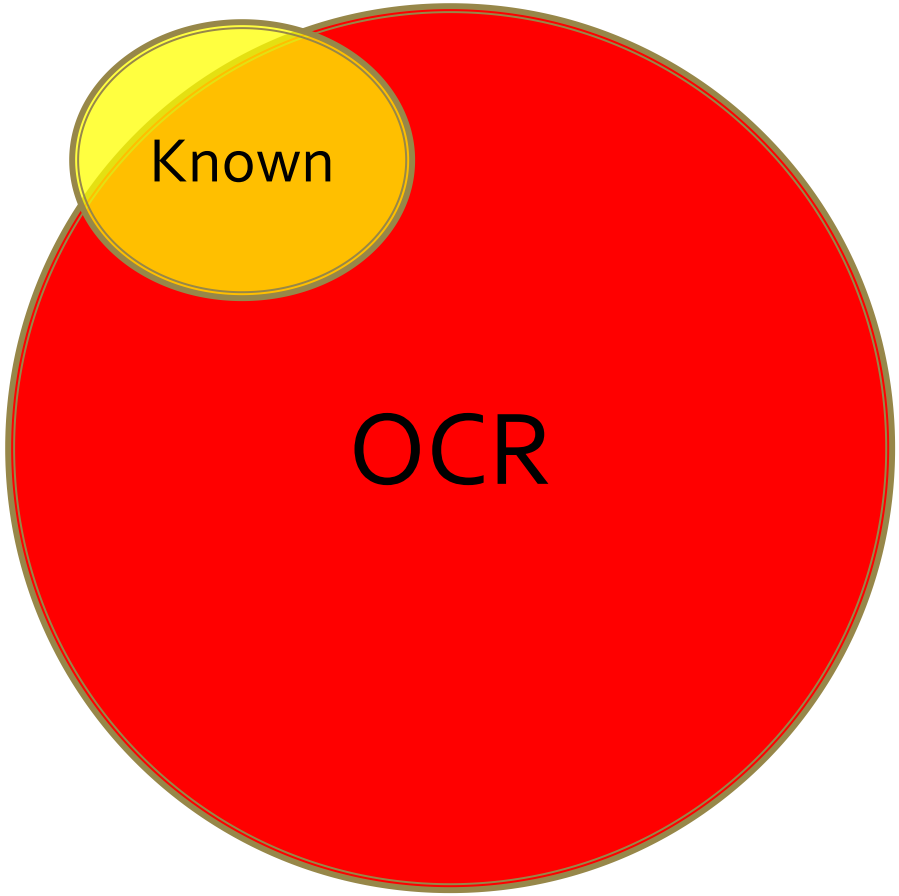
| First Word | Second | Locality | habitat |
|---|---|---|---|
| creek | | 49992 (72%) | 16990 (24) |
| eagle | creek | 239 (98%) | 5 (2%) |
| creek | bottom | 85 (12%) | 544 (76%) |
| along | creek | 315 (14%) | 1530 (69%) |
| sycamore | | 1453 (76%) | 354 (19%) |
| under | sycamore | 1 (8%) | 9 (75%) |
| 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

# Gottschalk Method

- No parsing
- Uses OCR to establish correlation between post processed and unprocessed



# Preprocessing: Create finger print

use only the left-most 15 characters: remove punctuation, special characters, spaces, and numbers

| **Raw Line of OCR** | **Fingerprint** |
|---|---|
| ■SANTO DOMINGO. REPUBLICA DOMINICANA | santodomingorep |
| • República Dominicana: Prov. El Seibo: | repúblicadomini |
| ¡ T. Zanoni, J. Pimentel § R. García | tzanonijpimente |
| ^ BOTANICAL | botanical JARDIN |
| BOTANICO NACIONAL DR. RAFAEL M. MOSCOSO" | jardinbotanicon |
| JARDIN BOTANICO NACIONAL'"DR. RAFAEL M. MOSCOSO" | jardinbotanicon |
| ■ ■ ■ JARDIN BOTANICO NACIONAL*"DR. RAFAEL M. MOSCOSO" | jardinbotanicon |
| JARDIN BOTANICO NACIONAL-"DR. RAFAEL M. MOSCOSO" | jardinbotanicon |

# Probability Matrix

| Catalog Number | fingerprint | Known Countries | Probs |
|---|---|---|---|
| 1442947 | tzanonirgarcíaj | Dominican Republic | 1.0000 |
| 1442947 | provindependenc | Dominican Republic | 1.0000 |
| 1442947 | nlesteelevm | Dominican Republic | 1.0000 |
| 1442947 | republicadomini | Dominican Republic | 1.0000 |
| 1442947 | mayo | Cuba | 0.3333 |
| 1442947 | mayo | Haiti | 0.3333 |
| 1442947 | mayo | Dominican Republic | 0.3333 |

# Gottschalk Method

- Dependent on post processed records
  - with images and OCR!
- Improves as specimens are processed
- Duplicate specimens are a plus
- Return corrected values
  - Translates abbreviations
  - OCR errors less of an issue
  - Better for country, state than substrate, locality, etc

# 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

# Exsiccati

**University of Wisconsin - Madison (WIS)**

Record Search Form

Collector: | Number: | Date:
Catalog Number: | Other Catalog Numbers:
Entered by: | Date entered: | Date modified:
Processing Status: | With images | Without images
Custom Field 1: OCR Fragment | CONTAINS | Exs
Display Editor | Display Table | Reset Form | Sort by: | ascending

Home >> Collection Management >> Editor  |< << | 1 of 5447 | >> >| >*

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

**Collector Info**

| Catalog Number | Other Numbers | Collector | Number | Date |
|---|---|---|---|---|
| WIS-L-0000188 | l0000188WIS | | | |

Dupes? Auto search

Associated Collectors | Verbatim Date

Exsiccati Title: Lichenes Wisconsinenses Exsiccati [Thomson: Lich. Wiscon. Exs.] | Number: 5

**Latest Identification**

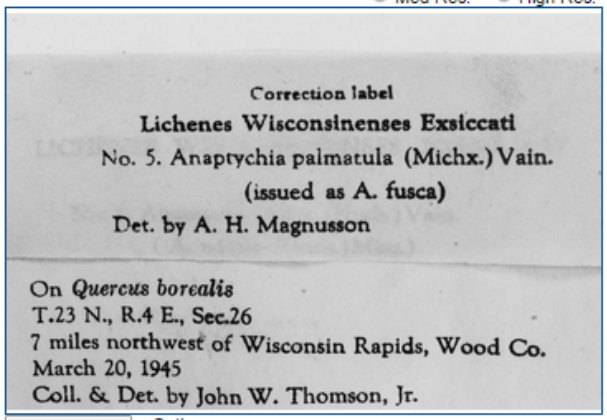
Scientific Name | Author
ID Qualifier | Family
Identified By | Date Identified

**Locality**

Country | State/Province | County | Municipality
Locality
Locality Security
Latitude | Longitude | Uncertainty | Tools | Datum | Verbatim Coordinates
Elevation in Meters | Verbatim Elevation
Georeferenced By | Georeference Sources | Georeference Remarks

**Label Processing**

Med Res. High Res.


Correction label
Lichenes Wisconsinenses Exsiccati
No. 5. Anaptychia palmatula (Michx.) Vain.
(issued as A. fusca)
Det. by A. H. Magnusson

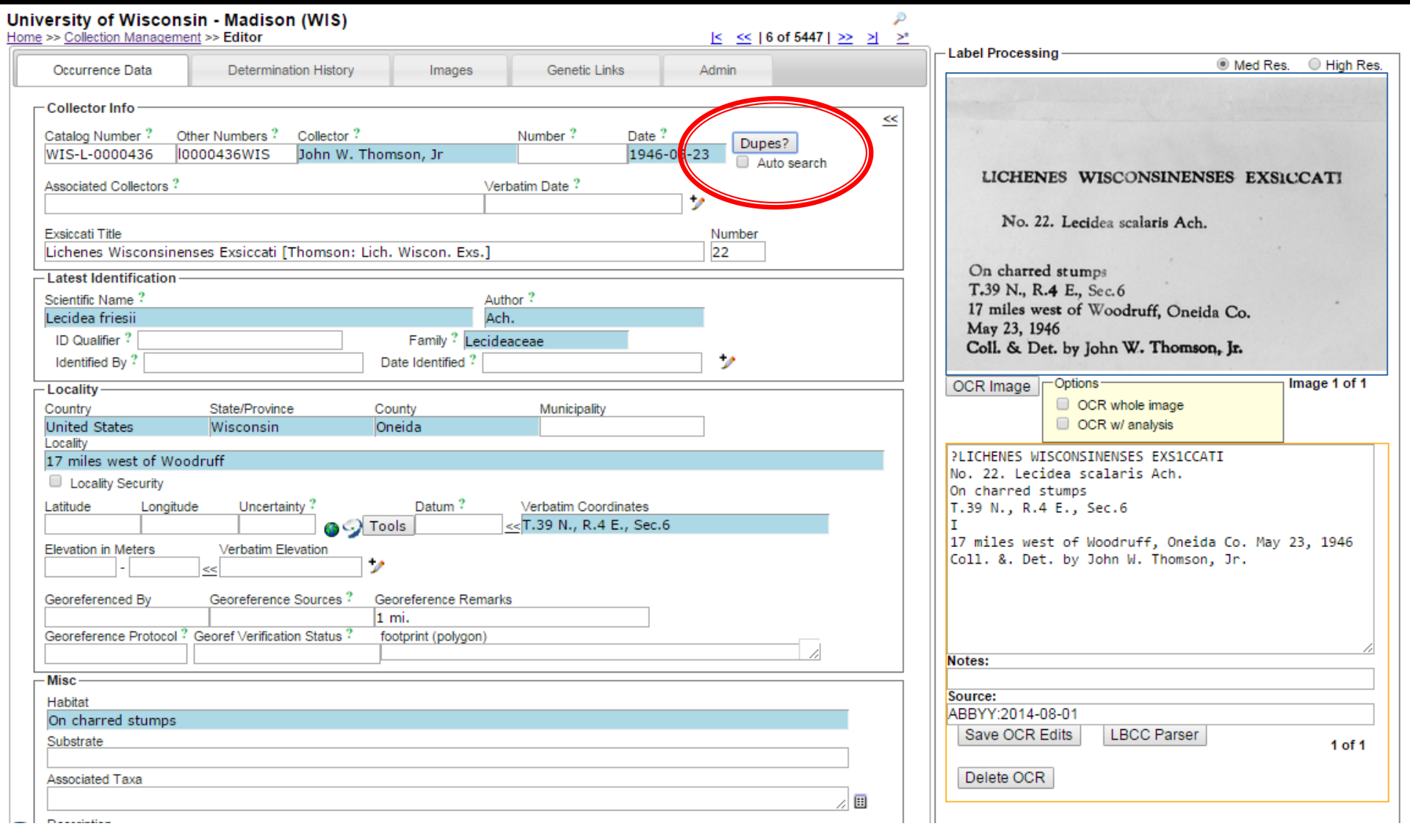
On *Quercus borealis*
T.23 N., R.4 E., Sec.26
7 miles northwest of Wisconsin Rapids, Wood Co.
March 20, 1945
Coll. & Det. by John W. Thomson, Jr.

OCR Image | Options: OCR whole image | OCR w/ analysis

Image 3 of 3=>>

```
?Correction label
Lichenes Wlsconsinenses Exsiccati
No. 5. Anaptychia palmatula (Michx.) Vain.
(issued as A. fusca)
Det. by A. H. Magnusson
wmm*m r ? , .n    _      ir_     _
On Quercus borealis T.23 N., R.4 E., Sec.26
7 miles northwest of Wisconsin Rapids, Wood Co.
```

# Exsiccati


University of Wisconsin - Madison (WIS)
Home >> Collection Management >> Editor
|< << | 6 of 5447 | >> >| >*
Occurrence Data
Determination History
Images
Genetic Links
Admin
Collector Info
Catalog Number ?
WIS-L-0000436
Other Numbers ?
l0000436WIS
Collector ?
John W. Thomson, Jr
Number ?
Date ?
1946-0?-23
Dupes?
Auto search
Associated Collectors ?
Verbatim Date ?
Exsiccati Title
Lichenes Wisconsinenses Exsiccati [Thomson: Lich. Wiscon. Exs.]
Number
22
Latest Identification
Scientific Name ?
Lecidea friesii
Author ?
Ach.
ID Qualifier ?
Family ?
Lecideaceae
Identified By ?
Date Identified ?
Locality
Country
United States
State/Province
Wisconsin
County
Oneida
Municipality
Locality
17 miles west of Woodruff
Locality Security
Latitude
Longitude
Uncertainty ?
Tools
Datum ?
Verbatim Coordinates
T.39 N., R.4 E., Sec.6
Elevation in Meters
Verbatim Elevation
Georeferenced By
Georeference Sources ?
Georeference Remarks
1 mi.
Georeference Protocol ?
Georef Verification Status ?
footprint (polygon)
Misc
Habitat
On charred stumps
Substrate
Associated Taxa
Label Processing
Med Res.
High Res.
LICHENES WISCONSINENSES EXSICCATI
No. 22. Lecidea scalaris Ach.
On charred stumps
T.39 N., R.4 E., Sec.6
17 miles west of Woodruff, Oneida Co.
May 23, 1946
Coll. & Det. by John W. Thomson, Jr.
OCR Image
Options
OCR whole image
OCR w/ analysis
Image 1 of 1
?LICHENES WISCONSINENSES EXSICCATI
No. 22. Lecidea scalaris Ach.
On charred stumps
T.39 N., R.4 E., Sec.6
I
17 miles west of Woodruff, Oneida Co. May 23, 1946
Coll. &. Det. by John W. Thomson, Jr.
Notes:
Source:
ABBYY:2014-08-01
Save OCR Edits
LBCC Parser
1 of 1
Delete OCR

# Expeditions defined by OCR

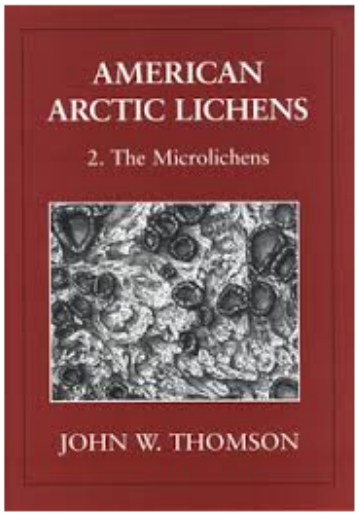
**ALCAN Expedition**


Map of the Alaska Highway (ALCAN Highway)

The ALCAN Highway was constructed during World War II for the purpose of connecting the contiguous United States to Alaska through Canada. It begins at the junction with several Canadian highways in Dawson Creek, British Columbia, and runs to Delta Junction, Alaska, via Whitehorse, Yukon. Completed in 1942 at a length of approximately 2,700 km (1,700 mi).

John Thomson

Teuvo Ahti




Flora of American Arctic Lichen

The eminent American lichenologist, John W. Thomson, and the young Finnish lichenologist, Teuvo Ahti, who has since become the world expert in the large lichen genus Cladonia., traveled the Alcan Highway together in 1967, after John had spent a sabbatical in Helsinki. The road was primarily gravel at that point in time, and they had to change tires 13 times. This was one of John's early expeditions to the American arctic. Across his 40 year career at the University of Wisconsin, he mounted 14 major expeditions across the Canadian and American arctic resulting in an American Arctic Lichen Flora published in two volumes. The Alcan trip of John and Teuvo cemented a lifelong friendship and spurred continued collaboration over several decades. They were among the first lichenologists to traverse the highway and the lichens collected provided an important contribution to the final flora treatments.

Find ALCAN Records | Find Thompson Records | Find Ahti Records

# Combined methods

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

# Acknowledgments


- National Science Foundation
- Arizona State University - GIOS and SOLS
- New York Botanical Garden
- University of Wisconsin - Madison
- Harvard University
- GeoLocate, iDigBio, STRI, Filtered Push, Specify


ASU SCHOOL OF Life Sciences
ARIZONA STATE UNIVERSITY

NSF National Science Foundation
WHERE DISCOVERIES BEGIN