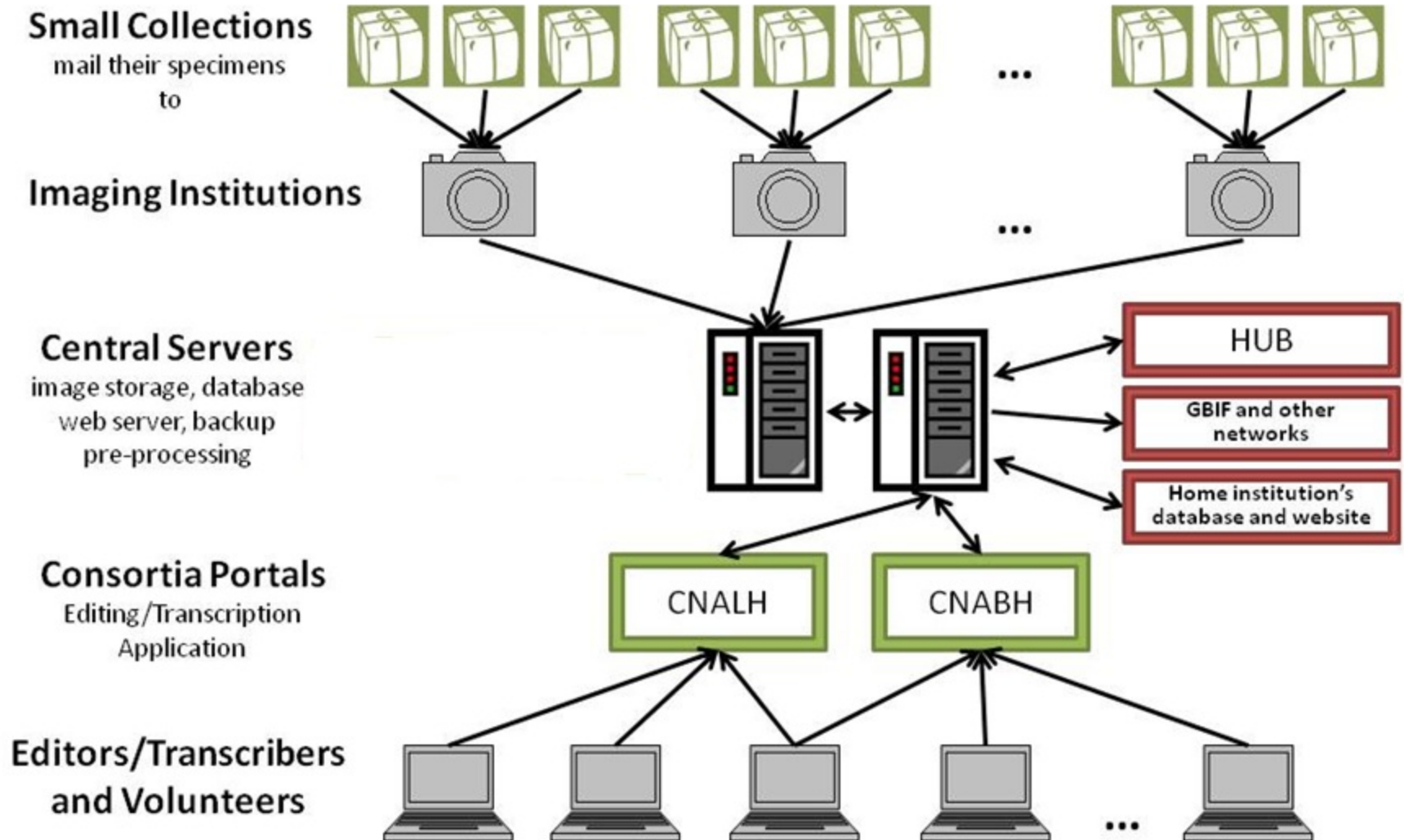


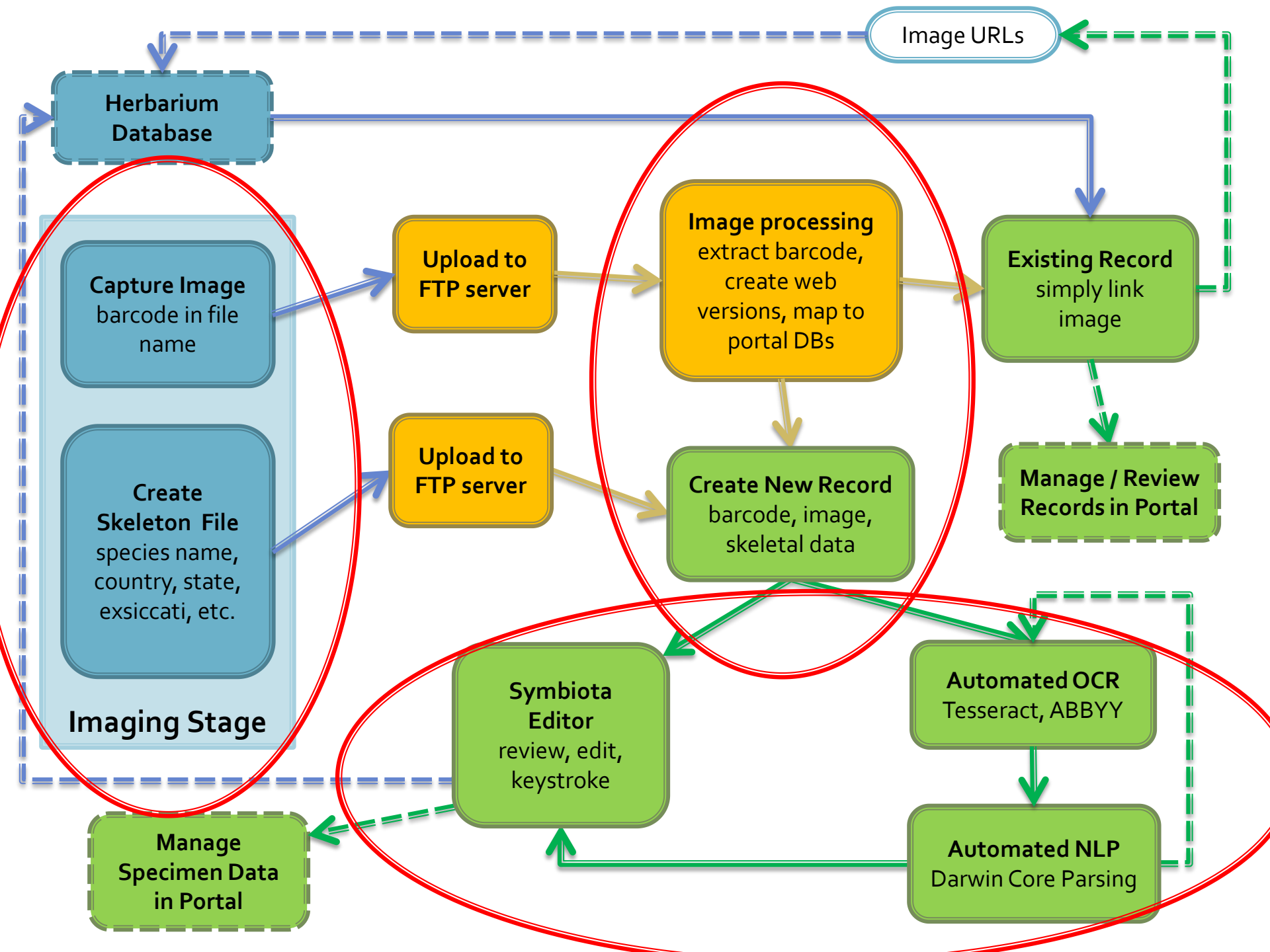
Symbiota and Specimen Label Digitization Workflows



National Science Foundation
WHERE DISCOVERIES BEGIN

Digitization Workflow





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

Home >> Collection Management >> Editor | 1 of 996

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

Collector Info

Catalog Number ?	Other Numbers ?	Collector ?	Number ?	Date ?	Dupes?
DES00052061		Dixie Z. Damrel	1744-B	2002-08-20	<input type="checkbox"/> Auto search

Associated Collectors ? Verbatim Date ?

P. Boness 20 August 2002

Latest Identification

Scientific Name ? Author ?

Heterotheca subaxillaris (Lam.) Britt. & Rusby

ID Qualifier ? Family ? Asteraceae

Identified By ? Date Identified ?

Locality

Country	State/Province	County	Municipality
United States	Arizona	Gila	

Locality: Young, Between Pleasant Valley Inn and the Bldg. ForTonto National Forest, On roadside of Hwy 288.

Locality Security

Latitude Longitude Uncertainty ? Datum ? Verbatim Coordinates

34.101389 -110.963056 Tools << N 34° 6' S W 110° 57' 47"

Elevation in Meters Verbatim Elevation

1579 - Elev.: 5179 R

Misc

Habitat: Grassland valley roadside

Substrate:

Associated Taxa: Bothriochloa ischaemum, Grindelia squarrosa, Chloris virgata, Desmanthus cooleyi, Asclepias subverticillata

Description: Annual from 1-2.5 ft. tall. Foliage with camphor-like scent, Abundant along roadside

Notes: szot I Des

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
COLLECTOR T. H. Nash #7914 ELEV. 4400'
DET. T. H. N. DATE 8/27/73

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
COLLECTOR T. H. Nash #7914
DET. T. H. N.
ELEV. 4400'
DATE 8/27/73

PLANTS OF NEW MEXICO
Herbarium of Arizona State University
Parmelia ulophyllodes (Vain.) Sav.
COUNTY Dona Ana
Joranada Experimental Station -
New Mexico State University
"on Juniperus
ELEV. 4400
DATE
T. H. Nash #7914 8/27/73

»..-1

æ

™

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

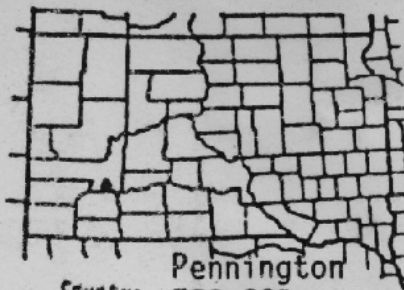
DESERT BOTANICAL GARDEN HERBARIUM
Cylindropuntia prolifera (Engelmann) F. M. Knuth

SOUTH DAKOTA, USA

Stavrothele cf. monicae (Zahlbr.)
Wet

Badlands National Park
on fossil ammonites and
weathered Pierre shale
breaks W of Sage Creek
campground, shortgrass
plains
T2S, R14E, E $\frac{1}{2}$, SW $\frac{1}{4}$, Sec 2
lat 43°54'N
long 102°25'W

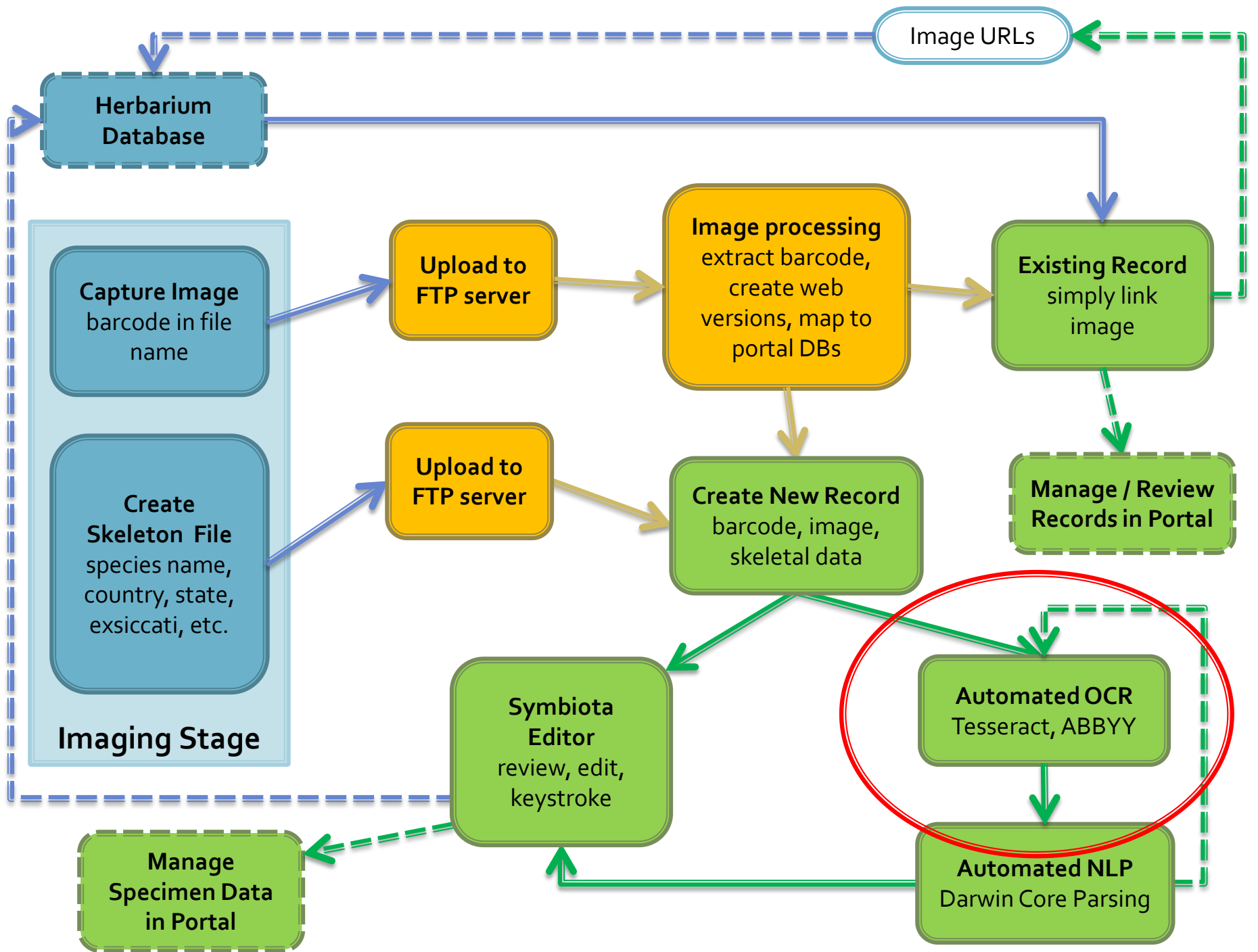
Date: Aug 20, 1990
Collected by Susan Will-Wolf
Det.



Pennington
County: Pennington
Elev: 780-820m
No. 2347

NPS Acc. # BADL-276 Cat. #

R. Trelease and N. Foisy
Herbarium of Desert Botanical Garden (DES)



Ready for Processing

University of Wisconsin - Madison (WIS)

Home >> Collection Management >> Editor

< << | 9 of 11989 | >> > | X

Occurrence Data

Determination History

Images

Genetic Links

Admin

Collector Info

Catalog Number ? Other Numbers ? Collector ? Number ? Date ? Dupes? Auto search

Associated Collectors ? Verbatim Date ?

Exsiccata 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 ? Datum ? Verbatim Coordinates

Elevation in Meters Verbatim Elevation

Misc

Habitat

Substrate

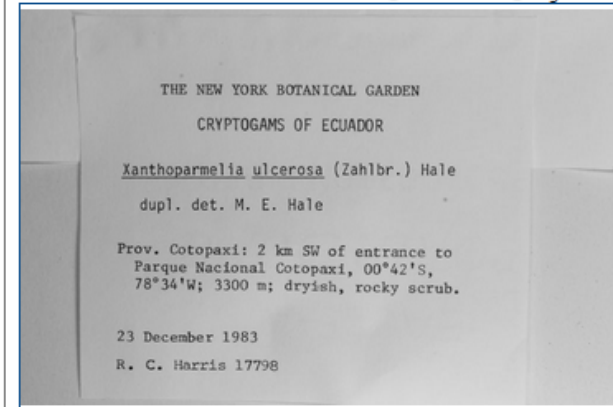
Associated Taxa

Description

Notes

Label Processing

Med Res. High Res.



OCR Image

Options

- OCR whole image
- OCR w/ analysis

Image 1 of 1

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:

ABBY:2013-02-25

Save OCR Edits

LCC Parser

Delete OCR

1 of 1

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 ? Datum ? Verbatim Coordinates

Elevation in Meters Verbatim Elevation

Misc

Habitat

Substrate

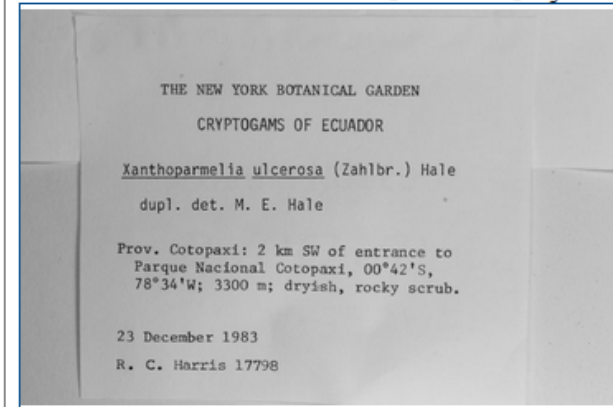
Associated Taxa

Description

Notes

Label Processing

Med Res. High Res.



OCR Image

Options

- OCR whole image
- OCR w/ analysis

Image 1 of 1

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:

ABBY:2013-02-15

Save OCR Edits

LBCC Parser

Delete OCR

1 of 1

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 Richard C Harris 17798 1983-12-23

Associated Collectors ? Verbatim Date ?

Exsiccati Title Number

Latest Identification

Scientific Name ? Author ?
Xanthoparmelia ulcerosa (Zahlbr.) Hale
 ID Qualifier ? Family ? *Parmeliaceae*
 Identified By ? M. E. Hale Date Identified ?

Locality

Country State/Province County Municipality
 Ecuador Cotopaxi

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

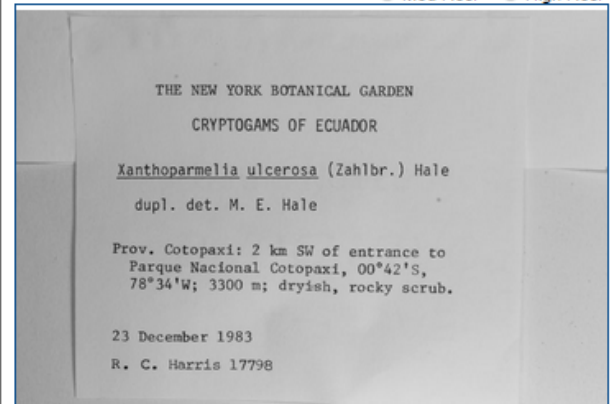
Associated Taxa

Description

Notes

Label Processing

Med Res. High Res.



OCR Image

Options

- OCR whole image
- OCR w/ analysis

Image 1 of 1

i>> THE NEW YORK BOTANICAL GARDEN
 CRYPTOGRAMS 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:

ABBY:2013-02-25

Save OCR Edits

LBCC Parser

Delete OCR

1 of 1

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

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

Collector Info

Catalog Number ? ASU0081742 | Other Numbers ? | Collector ? Liz Makings | Number ? 4485 | Date ? 2014-09-07 | Dupes? Auto search

Associated Collectors ? | Verbatim Date ? 7 September 2014

Exsiccati Title | Number

Latest Identification

Scientific Name ? Boerhavia coccinea | Author ? P. Mill.

ID Qualifier ? | Family ? Nyctaginaceae | Identified By ? | Date Identified ?

Locality

Country	State/Province	County	Municipality
United States	Arizona	Maricopa	

Locality Tonto National Forest, tributary to Sycamore Creek off Hwy 87

Locality Security

Latitude 33.730580 | Longitude 111.514076 | Uncertainty ? | Datum ? | Verbatim Coordinates 33.730580° 111.514076°

Elevation in Meters 617 | Verbatim Elevation 2024 ft

Misc

Habitat Mesquite Wash; sandy cobble floodplain with dense riparian vegetation

Substrate

Associated Taxa Prosopis velutina; Populus fremontii, Fraxinus velutina; Juglans major, Salix gooddingii, Sorghum halepense, Baccharis salicifolia, B. sarothroides, Hymenoclea monogyra; Artemisia dracuncululus; Datura wrightii, Oenothera elata, Boerhavia coccinea; Euphorbia pediculifera, E. capitellata, Amaranthus palmeri, Ipomoea cristulata, Kallstroemia grandiflora, K. parviflora

Description common perennial

Notes

Label Processing Med Res. High Res.

Nyctaginaceae

Boerhavia coccinea P. Mill.

USA. Arizona. Maricopa County.
Tonto National Forest; Mesquite Wash, tributary to Sycamore Creek off Hwy 87; sandy cobble floodplain with dense riparian vegetation; common perennial.
33.730580°, -111.514076°
Elevation: 2024 ft 620 m

Associated species: *Prosopis velutina*, *Populus fremontii*, *Fraxinus velutina*, *Juglans major*, *Salix gooddingii*, *Sorghum halepense*, *Baccharis salicifolia*, *B. sarothroides*, *Hymenoclea monogyra*, *Artemisia dracuncululus*, *Datura wrightii*, *Oenothera elata*, *Boerhavia coccinea*, *Euphorbia pediculifera*, *E. capitellata*, *Amaranthus palmeri*, *Ipomoea cristulata*, *Kallstroemia grandiflora*, *K. parviflora*

Liz Makings 4485 7 September 2014

OCR Image Options OCR whole image OCR w/ analysis **Image 1 of 1**

?Plants of Arizona
Nyctaginaceae
Boerhavia coccinea P. Mill.
USA. Arizona. Maricopa County.
Tonto National Forest; Mesquite Wash, tributary to Sycamore Creek off Hwy 87; sandy cobble floodplain with dense riparian vegetation; common perennial.
33.730580°, -111.514076°
Elevation: 2024 ft 620 m
Associated species: *Prosopis velutina*, *Populus fremontii*, *Fraxinus velutina*, *Juglans major*, *Salix gooddingii*, *Sorghum*

Notes:

Source:
ABBY: 2014-12-23

Save OCR Edits | SALIX Parser | LBCC Parser **1 of 1**

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