iDigBio
Integrated Digitized Biocollections

Database Tools & Techniques

Gil Nelson
September 16-18, 2012
Valdosta State University

iDigBio's Biological Collections Databases, Tools, and Data Publication Portals

https://www.idigbio.org/content/biological-collections-databases

(On the Wiki under Database Resources)

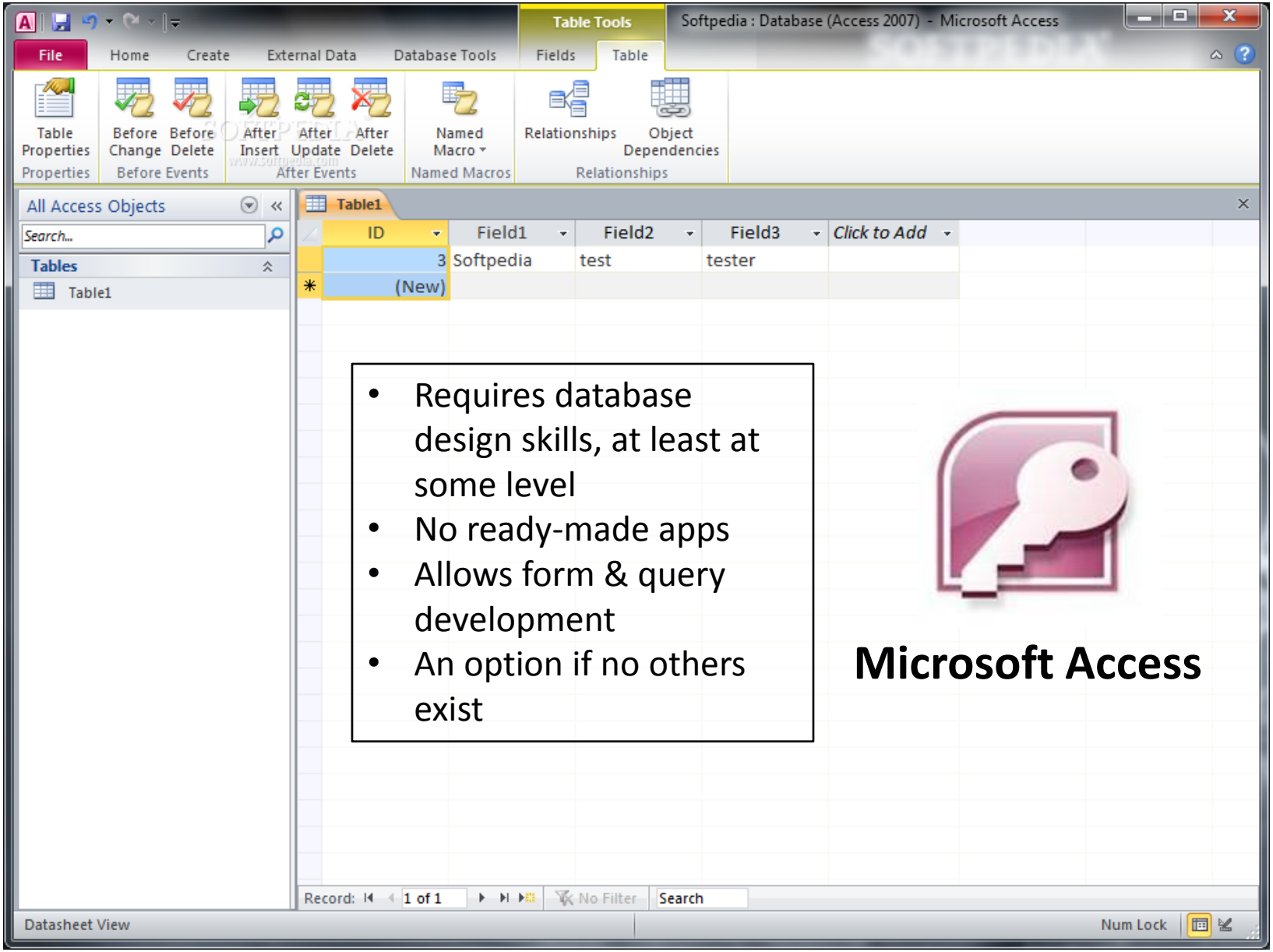
If there is something you'd like reviewed, let us know!

Spread Sheets: The Scientist's Buddy!

Ghita1.csv - Microsoft Excel

RecordID	accid	Taxon	fedstatus	statestatu	country	stapro	coupar	site	township	rangevar	section	quarter	sixteenth	quadnuml	notes	collectorn	geolat	geolon	datum	georefme	geoprecis	link	technician
27	26	Sarracenia flava			United Sta	Florida	Leon	In 2 or 3 inches of water along Forest Service Road 309, 1/4 mile SW of its jct with F								181:1983-(	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt
166	166	Drosera brevifolia			United Sta	Florida	Leon	Bloxham in open field just N of rte 267 (Bloxham Cutoff), 2.2 r						NULL		18575:199	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt
205	205	Drosera capillaris			United Sta	Florida	Leon	shores of cypress lake, uppermost lake in Lake Bradford chair						NULL								her	Ghita Heidt
262	262	Drosera tracyi			United Sta	Florida	Leon	Tallahassee bog, 9 miles W of Tallahassee, by Fla. Rt. 20. pinel						NULL								her	Ghita Heidt
264	264	Drosera tracyi			United Sta	Florida	Leon	Tallahassee bog, 9 miles W of Tallahassee, by Fla. Rt. 20. pinel						NULL								her	Ghita Heidt
273	273	Drosera tracyi			United Sta	Florida	Leon	Tallahassee flatwoods, about 12 miles west of Tallahassee on						NULL								her	Ghita Heidt
449	449	Pinguicula pumila			United Sta	Florida	Leon	Tallahassee roadside bordering savannah approx. 9 m. W of Ta						NULL								her	Ghita Heidt
505	507	Utricularia cornuta			United Sta	Florida	Leon	Tallahassee 13 mi. W of Tallahassee on Bristol Rd. wet sand.						NULL								her	Ghita Heidt
527	529	Utricularia floridana			United Sta	Florida	Leon	Woodville pond, 1 mile N of Woodville. in about 3.5 ft. of wat						NULL								her	Ghita Heidt
528	530	Utricularia floridana			United Sta	Florida	Leon	Apalachicola National Forest Clear Lake, Apalachicola Nationa						NULL								her	Ghita Heidt
536	540	Utricularia floridana			United Sta	Florida	Leon	Woodville pond, 1 mile N of Woodville. in ca. 3 ft of water, pc						NULL								her	Ghita Heidt
537	542	Utricularia floridana			United Sta	Florida	Leon	Tallahassee Collected at McFarlain's Pond in the vicinity of Lal						NULL								her	Ghita Heidt
538	543	Utricularia floridana			United Sta	Florida	Leon	Woodville pond, 1 mile N of Woodville. in about 3.5 ft. of wat						NULL								her	Ghita Heidt
539	544	Utricularia floridana			United Sta	Florida	Leon	Tallahassee Silver Lake, 12 miles W of Tallahassee. Anchored i						NULL								her	Ghita Heidt
545	551	Utricularia foliosa			United Sta	Florida	Leon	upper Lake Miccosukee. shallow water, upper Lake Miccosuk						NULL								her	Ghita Heidt
549	555	Utricularia foliosa			United Sta	Florida	Leon	Roadside pond near Lake Munson. roadside pond.						NULL								her	Ghita Heidt
550	556	Utricularia foliosa			United Sta	Florida	Leon	Roadside pond near Lake Munson. roadside pond.						NULL								her	Ghita Heidt
565	571	Utricularia foliosa			United Sta	Florida	Leon	Tallahassee in waters of Piney Z Lake in eastern sector of Talla						NULL	MP-							her	Ghita Heidt
588	610	Utricularia gibba			United Sta	Florida	Leon	Lake Miccosukee. Floating and intermingled with grass stems						NULL								her	Ghita Heidt
592	614	Utricularia gibba			United Sta	Florida	Leon	Woodville pond, 1 mile N of Woodville. in thick mats, shallow						NULL								her	Ghita Heidt
624	648	Utricularia inflata			United Sta	Florida	Leon	Woodville swamp 1/2 mile north of Woodville on Rt 319. cyp						NULL	[2nc							her	Ghita Heidt
625	650	Utricularia inflata			United Sta	Florida	Leon	Tallahassee Roadside pond about four miles west of Tallahass						NULL								her	Ghita Heidt
629	655	Utricularia inflata			United Sta	Florida	Leon	Tallahassee Swamp, 15 miles N of Tallahassee along [Meridiar						NULL		59409:196	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt
630	657	Utricularia inflata			United Sta	Florida	Leon	In Lake Jackson. In Lake Jackson.						NULL		59410:196	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt
635	663	Utricularia inflata			United Sta	Florida	Leon	Cypress pool along Aenon Church Road, between Rt. 20 and R						NULL		6236:1976	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt
659	701	Utricularia inflata			United Sta	Florida	Leon	Woodville edges of swamp forest, 1/2 mile N of Woodville. in						NULL		60863:196	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt
661	704	Utricularia inflata			United Sta	Florida	Leon	Tallahassee Nyssa swamp; About 15 miles north of Tallahasse						NULL		432:1960-(	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt
662	706	Utricularia inflata			United Sta	Florida	Leon	Tallahassee About 11 miles north of Tallahassee. In quiet wat						NULL		436:1960-(	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt
663	708	Utricularia inflata			United Sta	Florida	Leon	located along shore on Lake Jackson at the end of Faulk Dr. al						NULL		40939.79	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt
699	744	Utricularia olivacea			United Sta	Florida	Leon	Apalachicola National Forest Clear Lake, Apalachicola Nationa						NULL	[seeds att	75706:197	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt
700	745	Utricularia olivacea			United Sta	Florida	Leon	Tallahassee Apalachicola National Forest near the shore, Lost I						NULL		75239:197	30.45846	-84.2777	NAD83	County ce	704.17 squ	http://her	Ghita Heidt

- Not relational (flat, not normalized)
- Has a mind of its own!
- Data quality issues
- Accepts various data types in same column
- Useful as a tool for download/upload



- Requires database design skills, at least at some level
- No ready-made apps
- Allows form & query development
- An option if no others exist

Microsoft Access

Botanical Research and Herbarium Management System

Department of Plant Sciences, University of Oxford, UK

Home | Resources | Published Sites | News & Events | Software | Documentation | Sample databases | Login | English

Home | About | Advisory Group | Register | Contact

Botanical Research And Herbarium Management System

BRAHMS is used by research institutions globally supporting collection management in herbaria, botanic gardens and seed banks; taxonomic study; botanical survey; diversity analysis and many further categories of botanical research initiative.

Downloads

Version 7 update notes

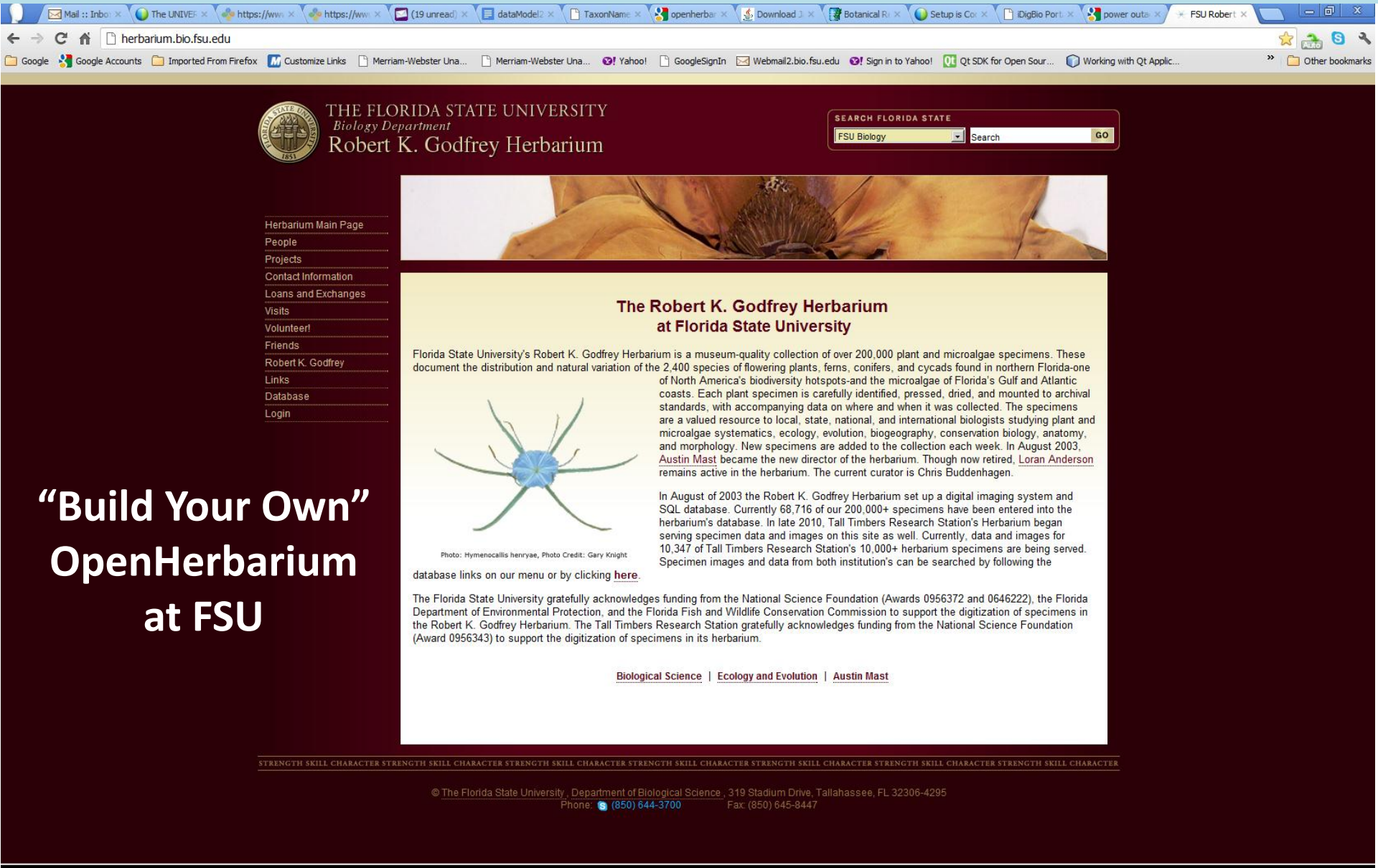
The largest single database to date with ca. 2 million specimens runs at the National Herbarium of the Netherlands, accessed via terminal services by all of the main Dutch herbaria and also published online. The country with the highest number of individual projects (2012) is Brazil.

The BRAHMS project is part of the plant diversity and systematics research group at the University of Oxford.

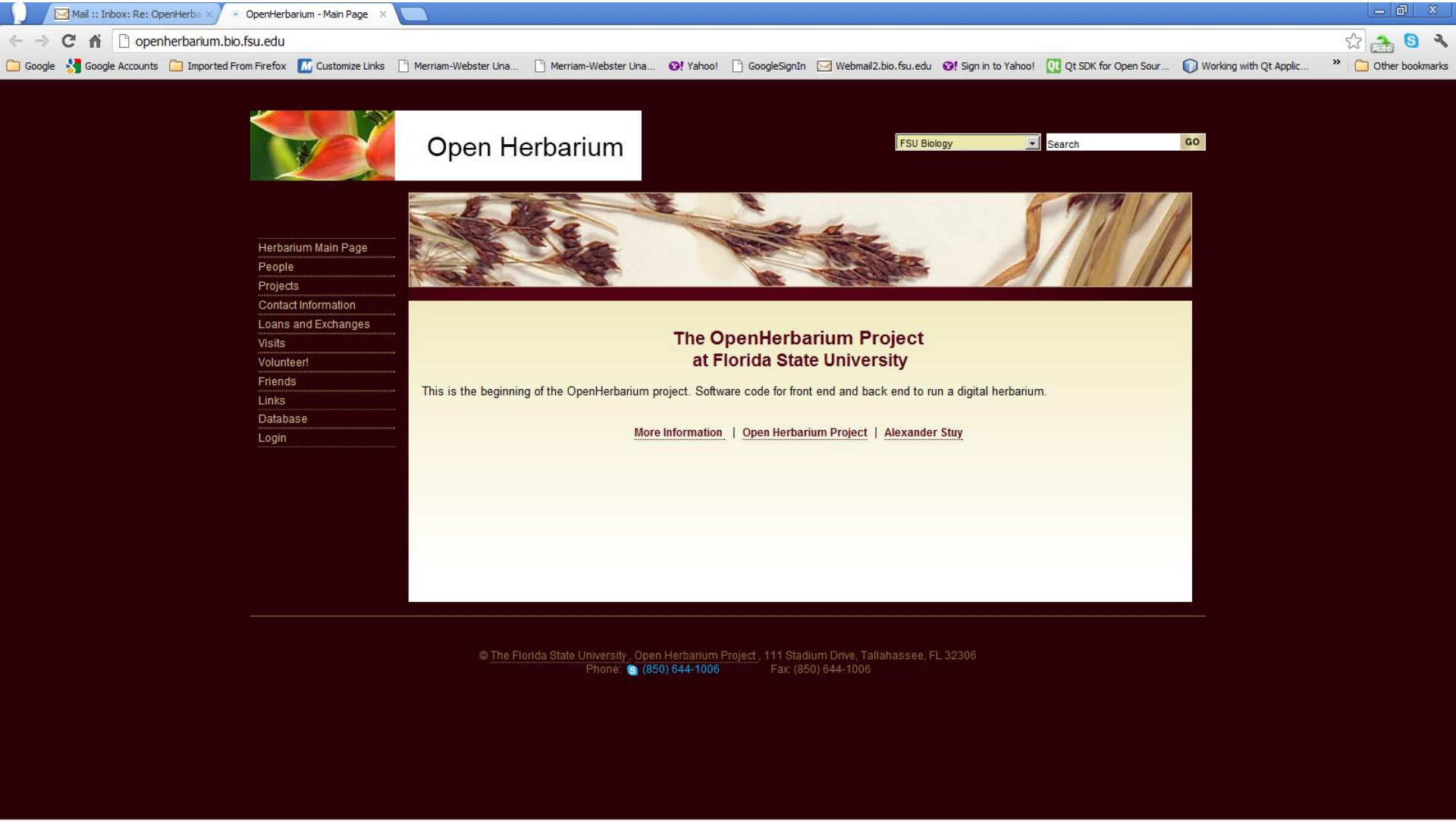
BRAHMS is a powerful database management system for botanical researchers and herbaria. It provides wide-ranging and innovative functionality to gather, edit, analyse and publish botanical data, optimizing its use for the widest possible range of curation services and research outputs. Find out more in English, French, Spanish, Portuguese and Russian.

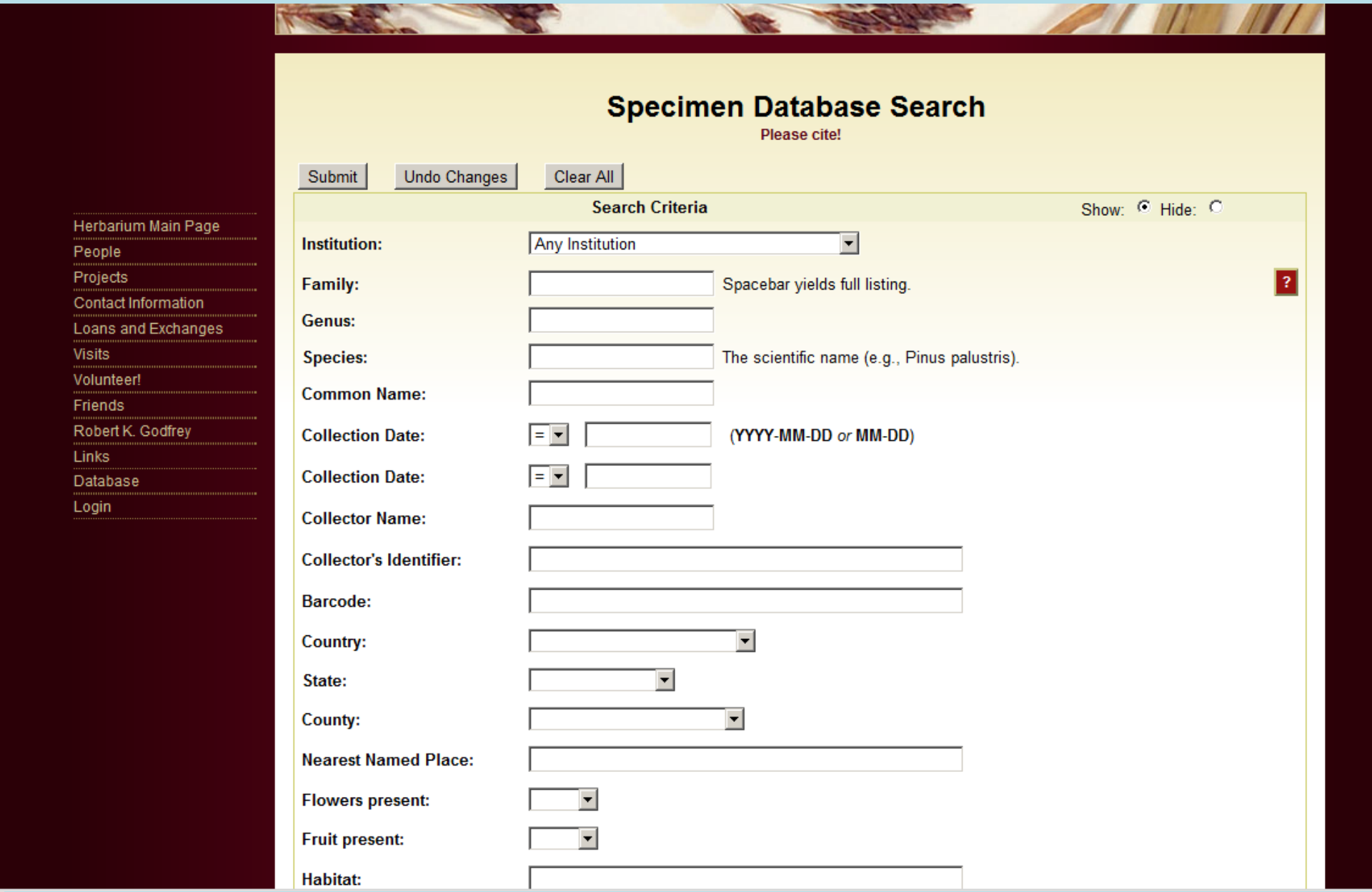
- FoxPro Files
- Mostly European
- Fairly easy to use and setup
- Good training manual
- Links to IPNI

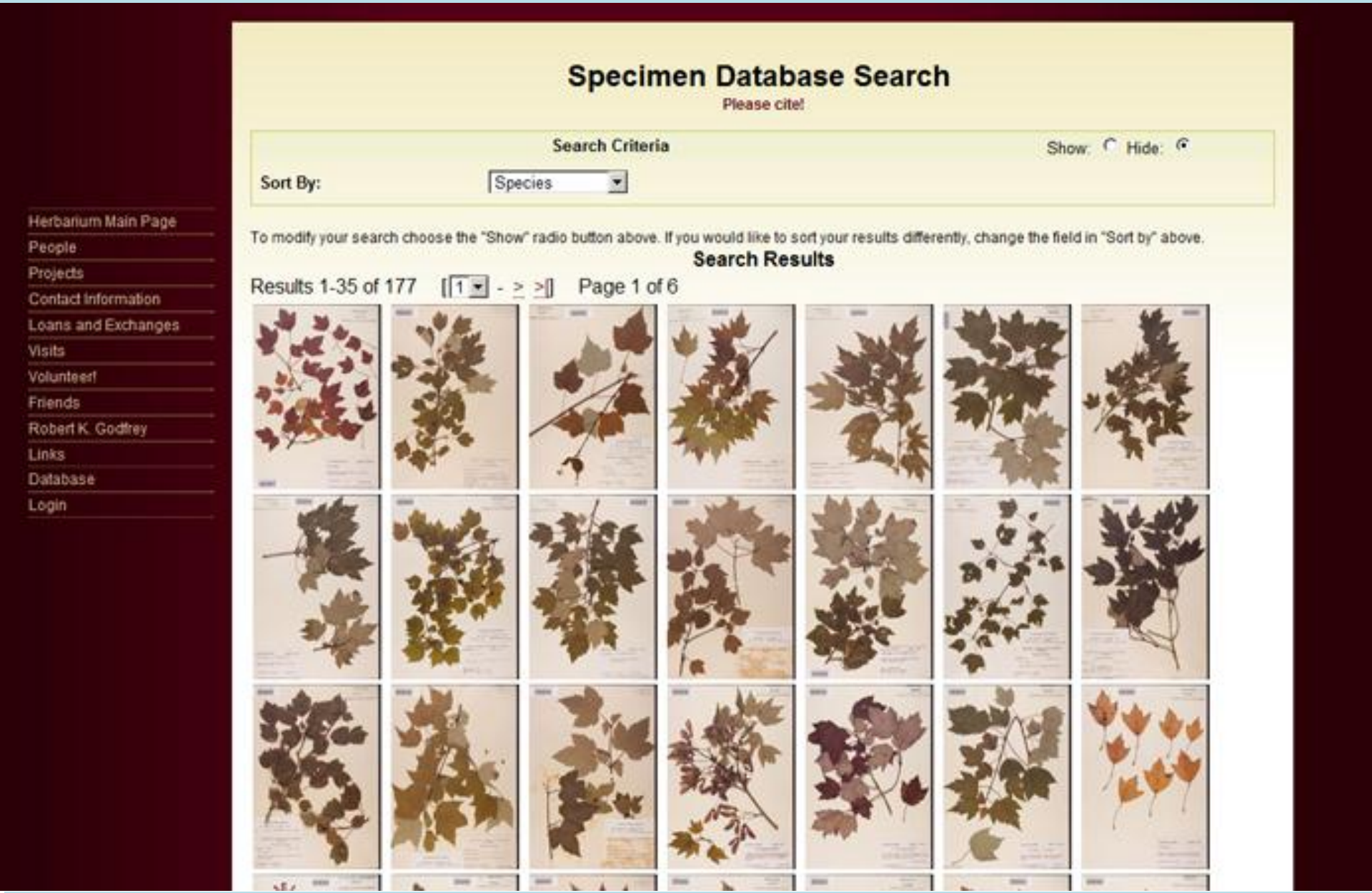
Herbarium Digitization Workshop

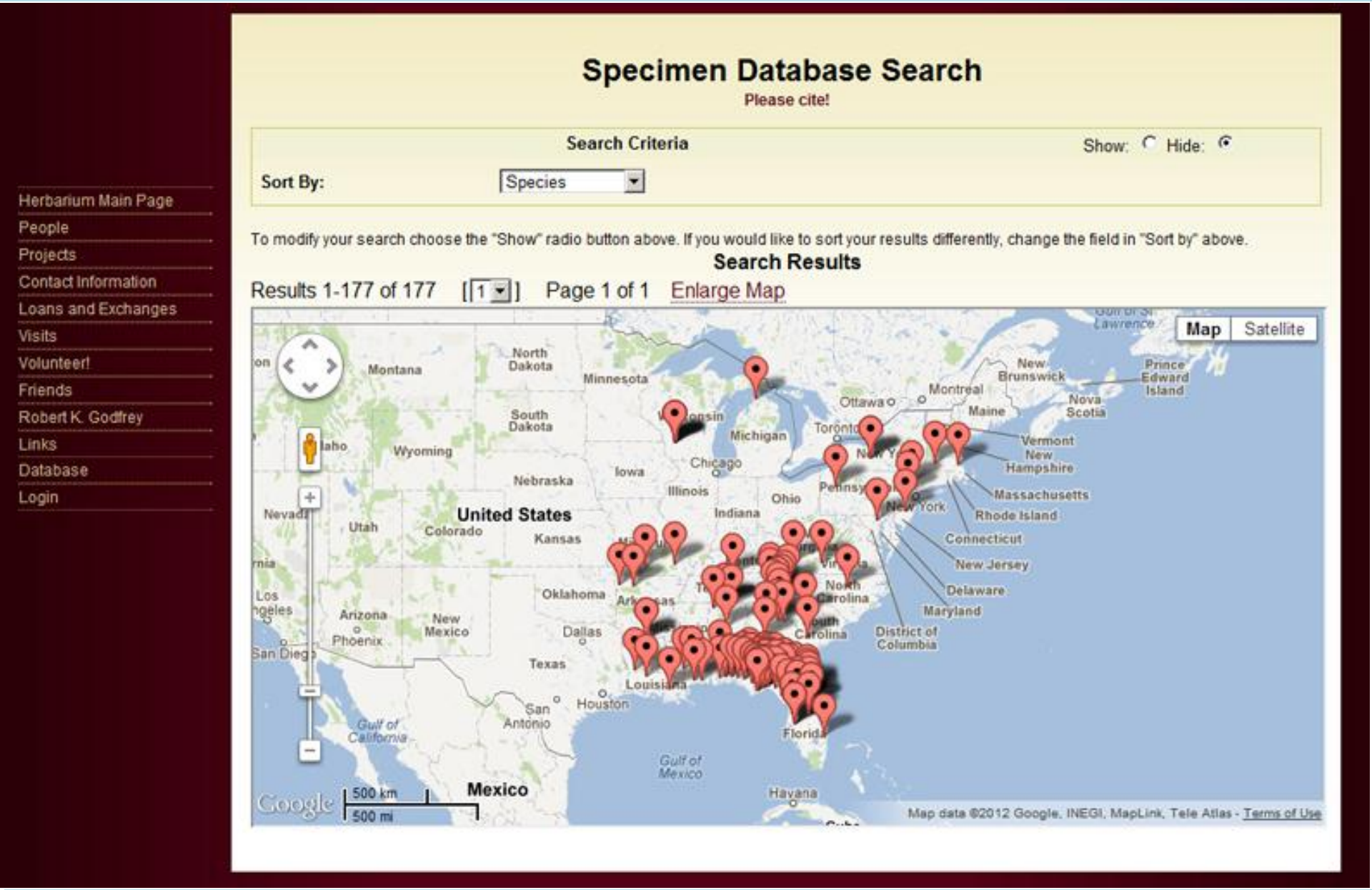


"Build Your Own" OpenHerbarium at FSU


Mail :: Inbox: Re: OpenHerba
OpenHerbarium - Main Page
openherbarium.bio.fsu.edu
Google
Google Accounts
Imported From Firefox
Customize Links
Merriam-Webster Una...
Merriam-Webster Una...
Yahoo!
GoogleSignIn
Webmail2.bio.fsu.edu
Sign in to Yahoo!
Qt SDK for Open Sour...
Working with Qt Applic...
Other bookmarks
Open Herbarium
FSU Biology
Search
GO
Herbarium Main Page
People
Projects
Contact Information
Loans and Exchanges
Visits
Volunteer!
Friends
Links
Database
Login
The OpenHerbarium Project
at Florida State University
This is the beginning of the OpenHerbarium project. Software code for front end and back end to run a digital herbarium.
More Information | Open Herbarium Project | Alexander Stuy
© The Florida State University , Open Herbarium Project , 111 Stadium Drive, Tallahassee, FL 32306
Phone: (850) 644-1006
Fax: (850) 644-1006


Herbarium Main Page
People
Projects
Contact Information
Loans and Exchanges
Visits
Volunteer!
Friends
Robert K. Godfrey
Links
Database
Login
Specimen Database Search
Please cite!
Submit
Undo Changes
Clear All
Search Criteria
Show:
Hide:
Institution:
Any Institution
Family:
Spacebar yields full listing.
Genus:
Species:
The scientific name (e.g., Pinus palustris).
Common Name:
Collection Date:
(YYYY-MM-DD or MM-DD)
Collection Date:
Collector Name:
Collector's Identifier:
Barcode:
Country:
State:
County:
Nearest Named Place:
Flowers present:
Fruit present:
Habitat:


Specimen Database Search
Please cite!
Search Criteria
Show: Hide:
Sort By:
Species
To modify your search choose the "Show" radio button above. If you would like to sort your results differently, change the field in "Sort by" above.
Search Results
Results 1-35 of 177
Page 1 of 6
Herbarium Main Page
People
Projects
Contact Information
Loans and Exchanges
Visits
Volunteer!
Friends
Robert K. Godfrey
Links
Database
Login


Herbarium Main Page
People
Projects
Contact Information
Loans and Exchanges
Visits
Volunteer!
Friends
Robert K. Godfrey
Links
Database
Login
Specimen Database Search
Please cite!
Search Criteria
Show: Hide:
Sort By:
Species
To modify your search choose the "Show" radio button above. If you would like to sort your results differently, change the field in "Sort by" above.
Search Results
Results 1-177 of 177
Page 1 of 1
Enlarge Map
Map
Satellite
United States
Mexico
Montana
North Dakota
Minnesota
South Dakota
Wyoming
Nebraska
Iowa
Illinois
Michigan
Chicago
Ohio
Indiana
Nevada
Utah
Colorado
Kansas
Oklahoma
Arizona
Phoenix
New Mexico
Dallas
Texas
Louisiana
Houston
San Antonio
Florida
Ottawa
Montreal
Toronto
Maine
New Brunswick
Nova Scotia
Prince Edward Island
Vermont
New Hampshire
Massachusetts
Rhode Island
Connecticut
New Jersey
Delaware
Maryland
District of Columbia
Gulf of California
Gulf of Mexico
Havana
500 km
500 mi
Map data ©2012 Google, INEGI, MapLink, Tele Atlas - Terms of Use

Herbarium Digitization Workshop

https://herbarium.bio.fsu.edu/phpadmin/index.php?db=herbarium&token=3fb3d9c965335740ac901236d0bff97c

phpMyAdmin

Database

herbarium (52)

localhost ▸ herbarium

Structure | SQL | Search | Query | Export | Import | Operations

Table	Records [1]	Type	Collation	Size	Overhead
AccessionRecords	455	MyISAM	latin1_swedish_ci	22.2 KiB	-
CollectorRecords	99,058	MyISAM	latin1_swedish_ci	4.3 MiB	-
Counties	6,138	MyISAM	latin1_swedish_ci	619.5 KiB	200 B
Countries	251	MyISAM	latin1_swedish_ci	18.9 KiB	-
FIPS	3,222	MyISAM	latin1_swedish_ci	146.6 KiB	-
FLGNIS	54,037	MyISAM	latin1_swedish_ci	6.1 MiB	-
FLMA	2,075	MyISAM	utf8_general_ci	233.3 KiB	-
FloridaSections	56,588	MyISAM	latin1_swedish_ci	5.7 MiB	-
geolocaterecords	1,668	MyISAM	latin1_swedish_ci	601.8 KiB	-
geolocateuploads	814	MyISAM	latin1_swedish_ci	268.4 KiB	-
Images	3	MyISAM	latin1_swedish_ci	4.2 KiB	-
iptTestView	~0[2]	View	---	-	-
MasterSpeciesList	10,283	MyISAM	latin1_swedish_ci	1.9 MiB	-
ModRecords	457,124	MyISAM	latin1_swedish_ci	27.3 MiB	-
Notes	9,147	MyISAM	latin1_swedish_ci	848.0 KiB	-
People	3,883	MyISAM	latin1_swedish_ci	467.3 KiB	-
Places	29,514	MyISAM	latin1_swedish_ci	2.0 MiB	-
PLSS	~84,492	InnoDB	utf8_general_ci	12.5 MiB	-
Projects	35	MyISAM	latin1_swedish_ci	7.8 KiB	-
Project_species	903	MyISAM	latin1_swedish_ci	38.4 KiB	-
QuadListFlorida	53,801	MyISAM	latin1_swedish_ci	4.9 MiB	-
Quads	53,765	MyISAM	utf8_general_ci	2.7 MiB	-
SelectedCounties	3,040	MyISAM	latin1_swedish_ci	293.8 KiB	-
SpecimenRecords	79,063	MyISAM	latin1_swedish_ci	33.2 MiB	-
SpecimenRecordsTmp	49,453	MyISAM	latin1_swedish_ci	18.9 MiB	-
SpecimenRecordsTmp2	49,397	MyISAM	latin1_swedish_ci	18.9 MiB	-
SpecimenRecordsTmp3	49,497	MyISAM	latin1_swedish_ci	18.9 MiB	-
SpecimenRecordsTmp4	49,397	MyISAM	latin1_swedish_ci	18.9 MiB	-
testIPTview	~0[2]	View	---	-	-
test_CollectorRecords	80,249	MyISAM	latin1_swedish_ci	3.5 MiB	-
test_SpecimenRecords	66,796	MyISAM	latin1_swedish_ci	30.4 MiB	-
test_VerificationPeople	79,727	MyISAM	latin1_swedish_ci	5.0 MiB	-
test_VerificationTable	74,126	MyISAM	latin1_swedish_ci	8.6 MiB	-
tmplation	5,981	MyISAM	latin1_swedish_ci	674.6 KiB	-
tmplation2	0	MyISAM	latin1_swedish_ci	1.0 KiB	-
tmpnative	114	MyISAM	latin1_swedish_ci	6.6 KiB	-
TTRS1FSUCollectorExport	100	MyISAM	latin1_swedish_ci	3.4 KiB	-
TTRS1FSUSpecimenExport	100	MyISAM	latin1_swedish_ci	27.3 KiB	-

- Open source
- Apache/IIS
- PHP
- Enterprise level

phpMyAdmin

- Can be installed on a workstation
- Requires database knowledge and skills

http://www.youtube.com/watch?v=UXvzZUlaB7I&feature=plcp

http://www.youtube.com/watch?v=faCP15wjc4g&feature=plcp

Silver BIOLOGY

Data Capture/Enrichment Techniques

(See link on Wiki to Workflow Modules and Tasks: Data Capture)

Keystroking:

- From images
- From specimen sheets
- Long vs. short (skeleton) records
- May be the quickest, most efficient method, especially if recording skeleton records

Optical Character Recognition (OCR)

Scanning electronic images with software designed to extract and make readable embedded text.

OCR Software

ABBYY Finereader 11, Corporate

- **Converts to Word or text, single files or multiple**
- **Provides a user interface**
- **Includes batch processing options**
- **Supports training to specific data sets**
- **Relatively inexpensive**
- **Relatively easy to configure**

tesseract-ocr

Tesseract open source OCR
Originally developed by HP in the 1980s
Now owned by Google
Focus of iDigBio OCR working group

Optical Character Recognition (OCR)

Potential Uses

Ingesting unedited OCR: Specify

Building robust searches of unedited text: VSU

Use as part of other software tools: Apiary, Symbiota

tesseract-ocr

The Apiary Project:

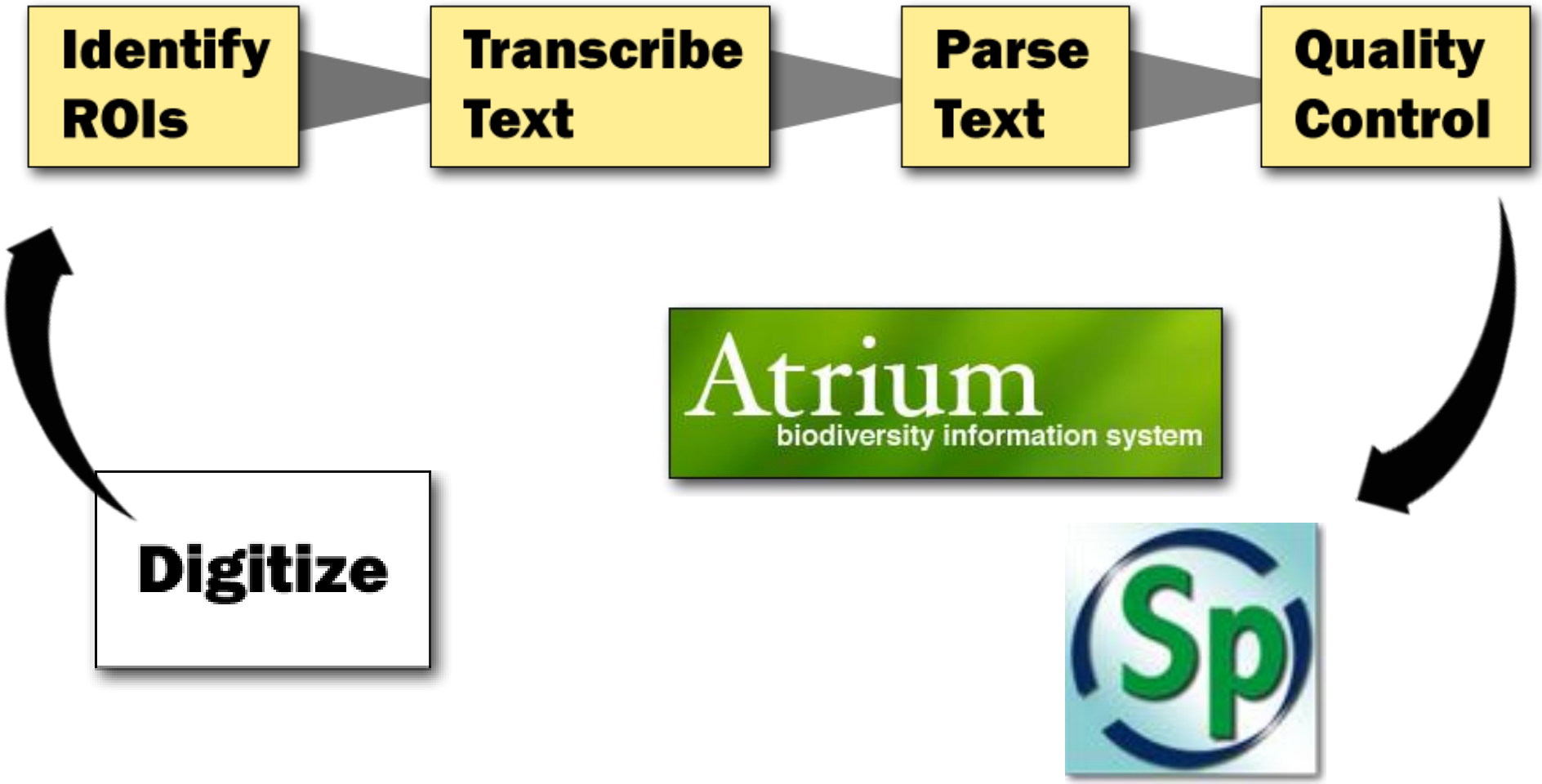
A collaborative workflow for extraction of herbarium label data

A project of BRIT and UNT's Texas Center for Digital Knowledge

Apiary Project – www.apiaryproject.org - Funded by IMLS National Leadership Grant # 06-08-0079-08
Botanical Research Institute of Texas / UNT TxCDK

Collector Name	Number	Scientific Name
J. C Taylor	31007	Solidago altiplanities
W. Hess	7283	Spiraea densiflora
H. St. John	14846	Sophora manejarevaensis
C.G. Pringle	13949	Polygala minutifolia
R.F. Hoover	3614	Senecio clevelandii
O. Degener	33,680	Wikstroemia perdita

The Technology and Workflow

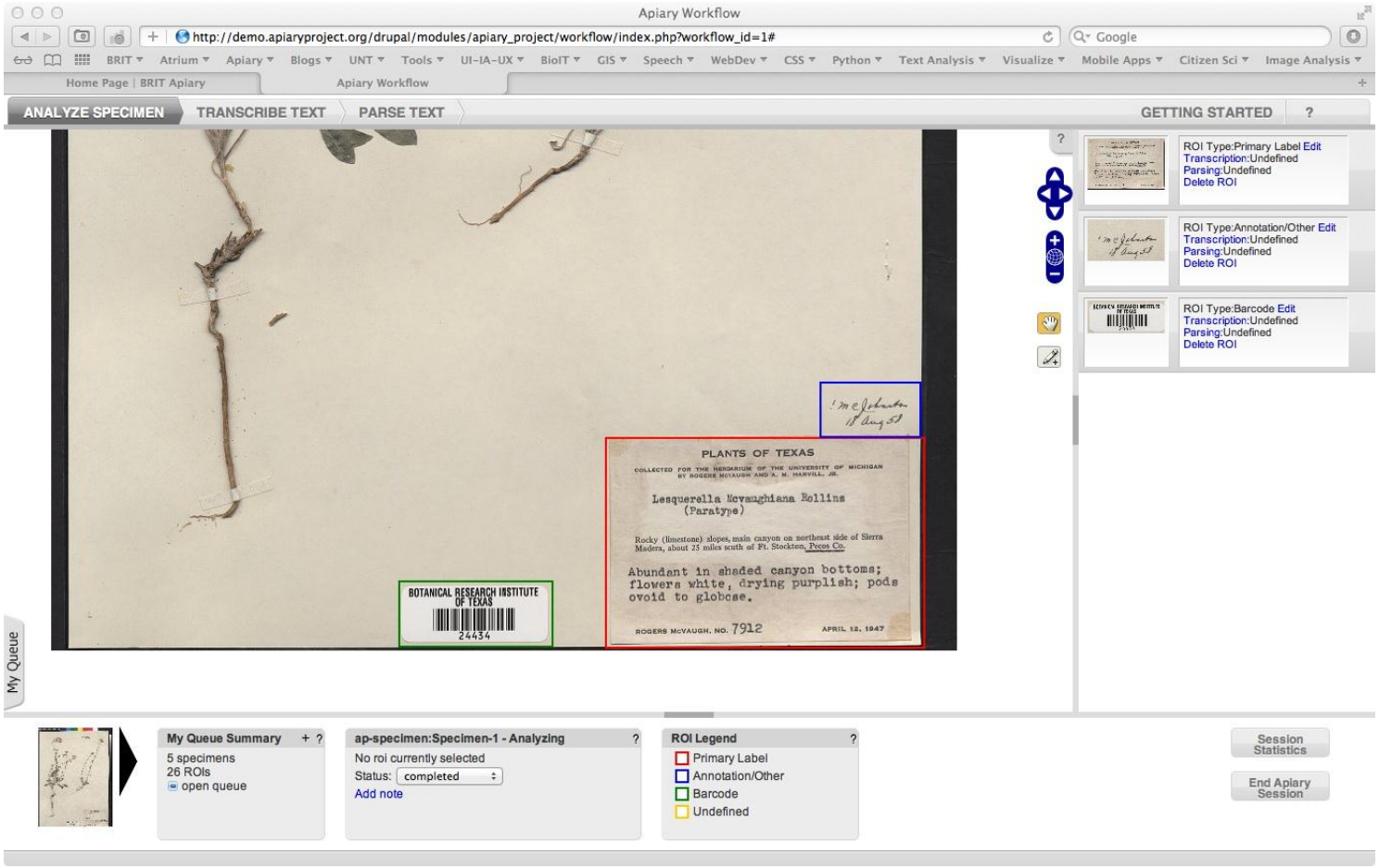


Digitize

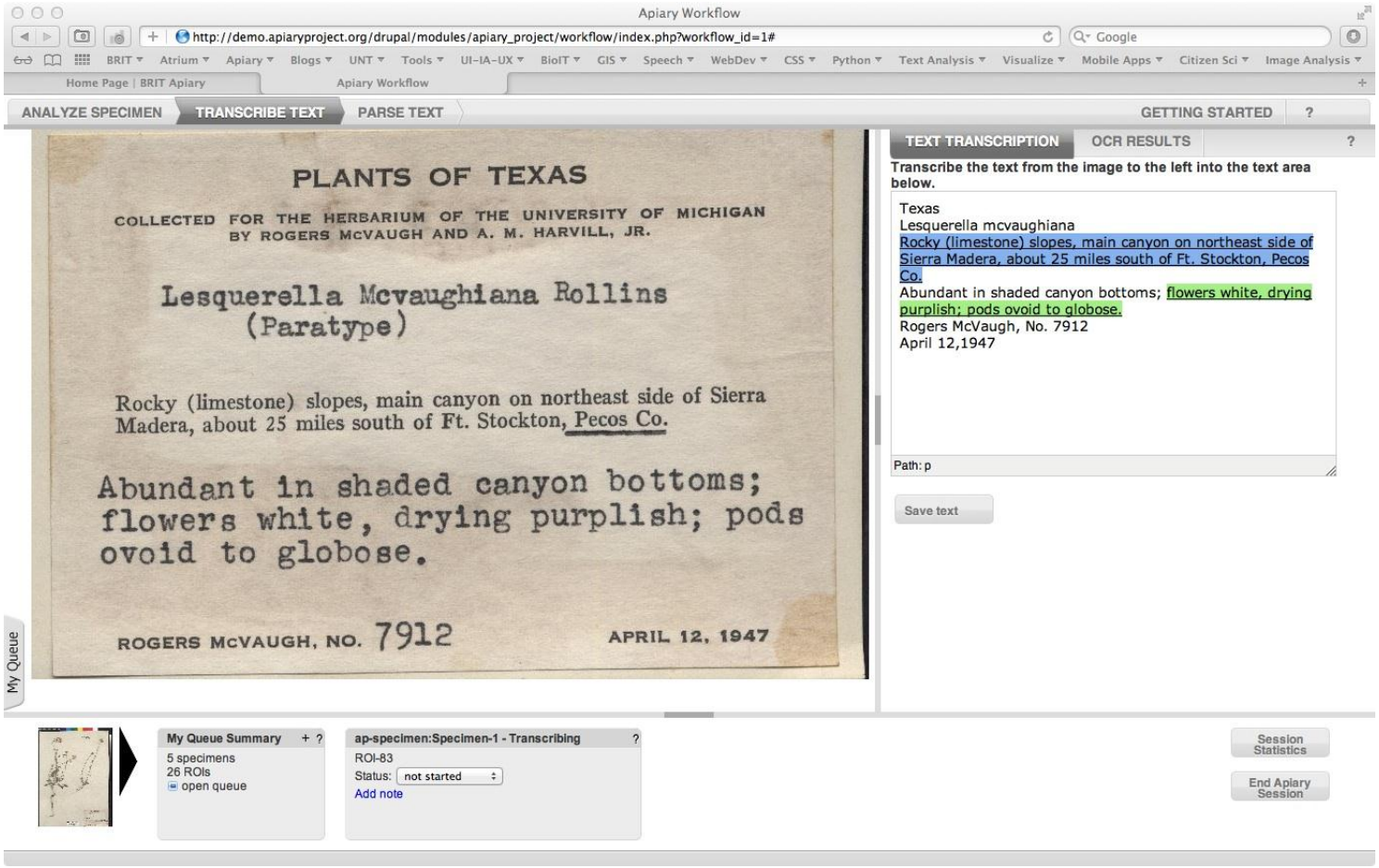
Identify ROIs → **Transcribe Text** → **Parse Text** → **Quality Control**

Apiary Project – www.apiaryproject.org - Funded by IMLS National Leadership Grant # 06-08-0079-08
Botanical Research Institute of Texas / UNT TxCDK

Finding Regions of Interest

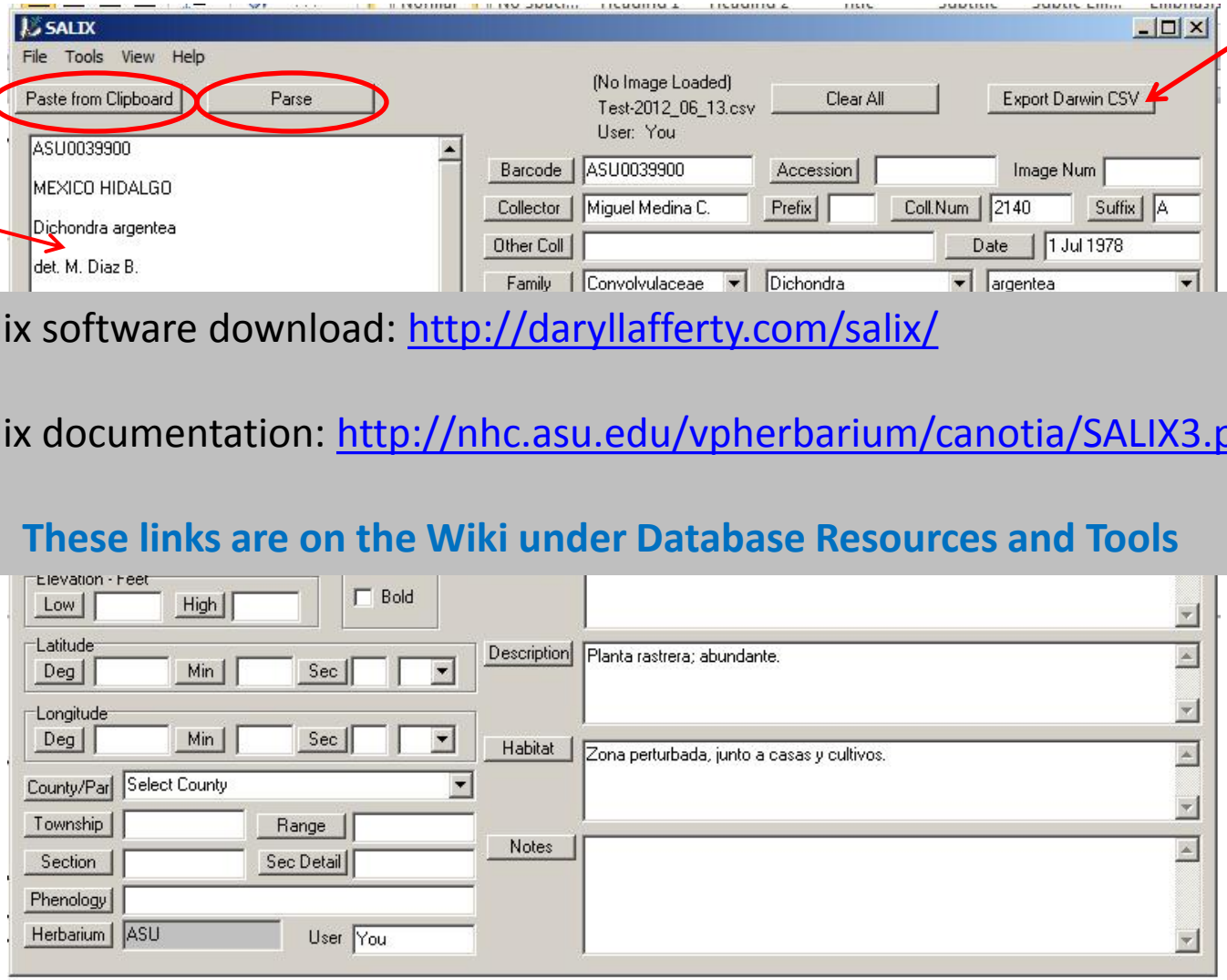


Transcription or OCR



Uploading a CSV in Salix: http://vimeo.com/42586885

Cleaned text



Salix software download: http://daryllafferty.com/salix/

Salix documentation: http://nhc.asu.edu/vpherbarium/canotia/SALIX3.pdf

These links are on the Wiki under Database Resources and Tools

Voice/Speech Recognition

Dragon Naturally Speaking
Nuance (now owns IBM's ViaVoice)
Mac & PC
Works better with a single user(?)
~$200.00 for premium version

Microsoft

Speech to text
Training
BRIT project (Windows API)
Included with Windows

Capturing Bar Code Values

Barcode scanning

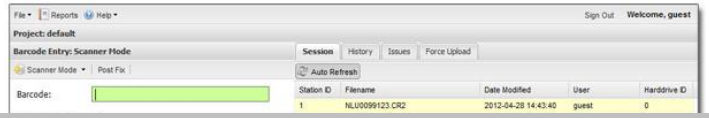
- Linear
- 2D
- Avoid data other than catalog number

Sync barcode value with camera-named files

Capturing Bar Code Values

FNIntercept

SilveImage



Barcode values can be capture at more than one place in the workflow.

- Pre-digitization curation
- Data capture
- Image capture

Bardecodefiler
BCRename

Renaming files to the barcode value

iDigBio
Integrated Digitized Biocollections

Thank You!