JACKSON
SCHOOL OF GEOSCIENCES

Creating workflows for effective data migration and imaging in an invertebrate paleontology collection

Ann Molineux, Liath Appleton, Angella Thompson, Louis G. Zachos*
Non-vertebrate Paleontology Laboratory, Jackson School of Geosciences, The University of Texas at Austin;
* Department of Geology & Geological Engineering, University of Mississippi

Acknowledgements

- Volunteers and associates of NPL especially: Gareth Cross, Christopher Garvie, Cissy Geigerman, Judy Humer, Keith Minor
- Melissa Winans, Data analyst at Texas Natural Science Center
- Many of the protocols are based upon work supported by the National Science Foundation under grants:
  - DBI-1057396: *Open Access: Conservation, Digitization and interoperability of the Historic Non-vertebrate Collections of the Texas Natural Science Center.*
  - EF-1305070: *Digitization PEN: Targeted digitization to expand and enhance the Paleoniches TCN.*
  - Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Non-vertebrate Paleontology Laboratory
TEXAS NATURAL SCIENCE CENTER
THE UNIVERSITY OF TEXAS AT AUSTIN

JACKSON
SCHOOL OF GEOSCIENCES

NSF

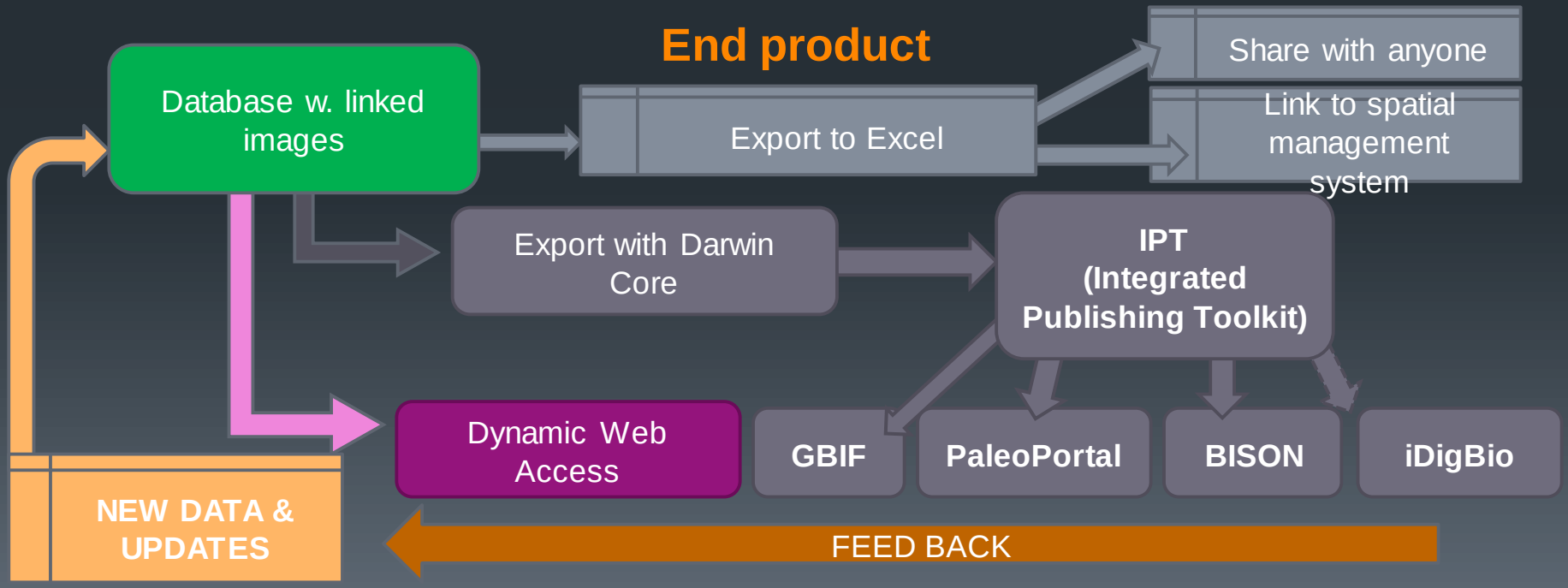
Summary

Goals

- Digital archive
- Useful data
- Accessible data
- ‘Short’ time frame
- ‘Lowest’ cost

Tasks

- Digitize
- Digitize accurately
- Digitize effectively
- Digitize efficiently



Major issues-

Data migration

- Common relational database
  - Migrate from MS Access into Specify6
- Normalized data
  - Build reference trees
  - Translate data
  - Refine data
  - Upload data

Specimen imaging

- High quality images
  - Cope with multi-scale objects
  - Create excellent depth of field
  - Provide analytical ability
- Reference images
  - Basic specimen image
  - Capture label data
  - Group images (drawers)

Migration resources

MAIN FOCUS

DATA CLEAN-UP

Translation tables

- Agent name
- BEG localities
- Chronostrat
- Lithostrat
- Storage [adding aisles]
- Taxon
- Titles [publications]

Open Refine

- Collection number
- Locality name
- Locality [verbatim]
- Storage location
- Old storage location
- Type status
- Preparation type
- Horizon and Series
- All dates
- Duplicate fields
- Catalog remarks

MAIN FOCUS

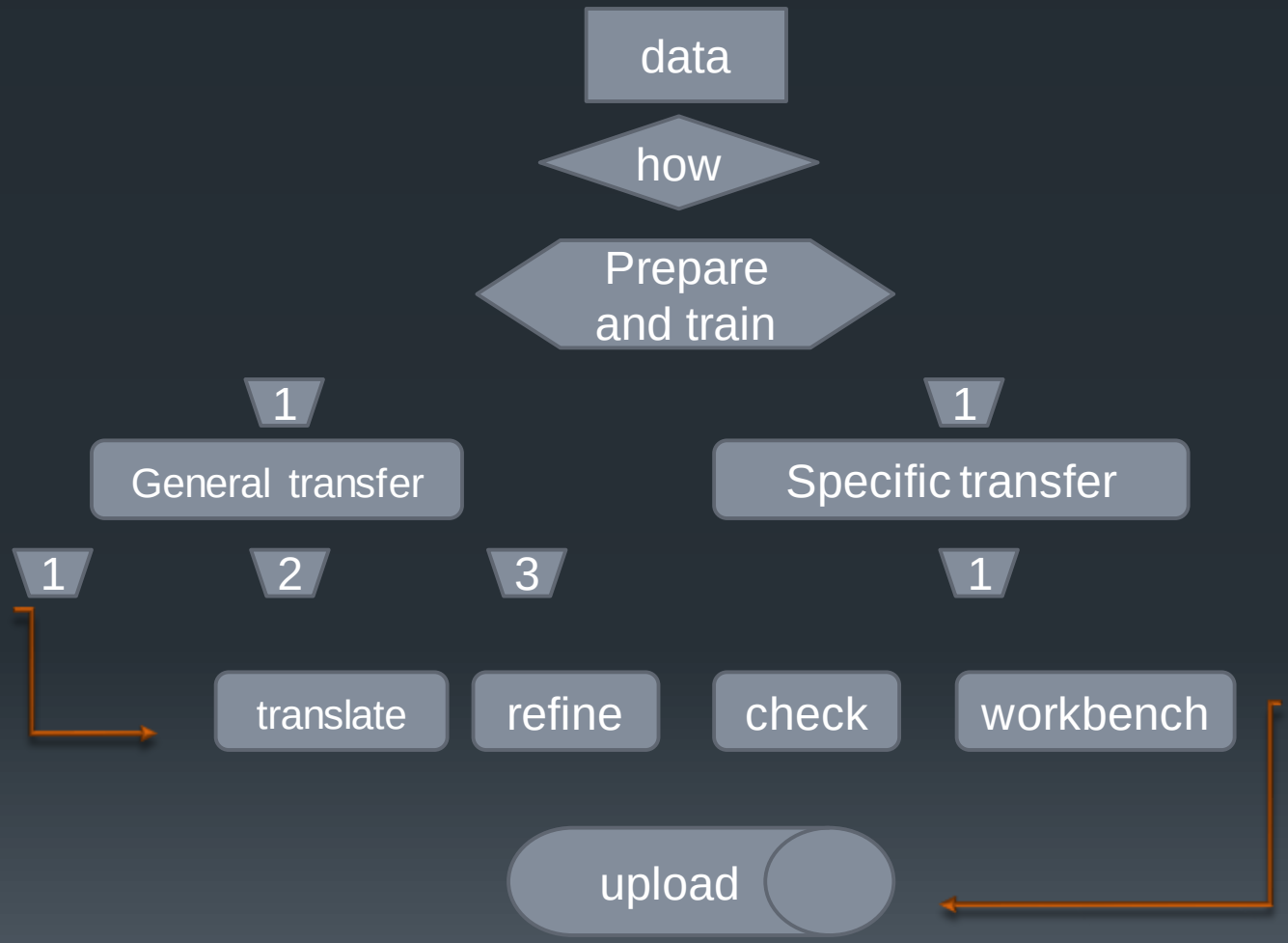
DATA STRUCTURE FOR WORKBENCH

Standardized [Normalized] data

- Import into Refine database as xls, csv,xml etc
- BUT export from Refine will be formatted in 97-2003 xls

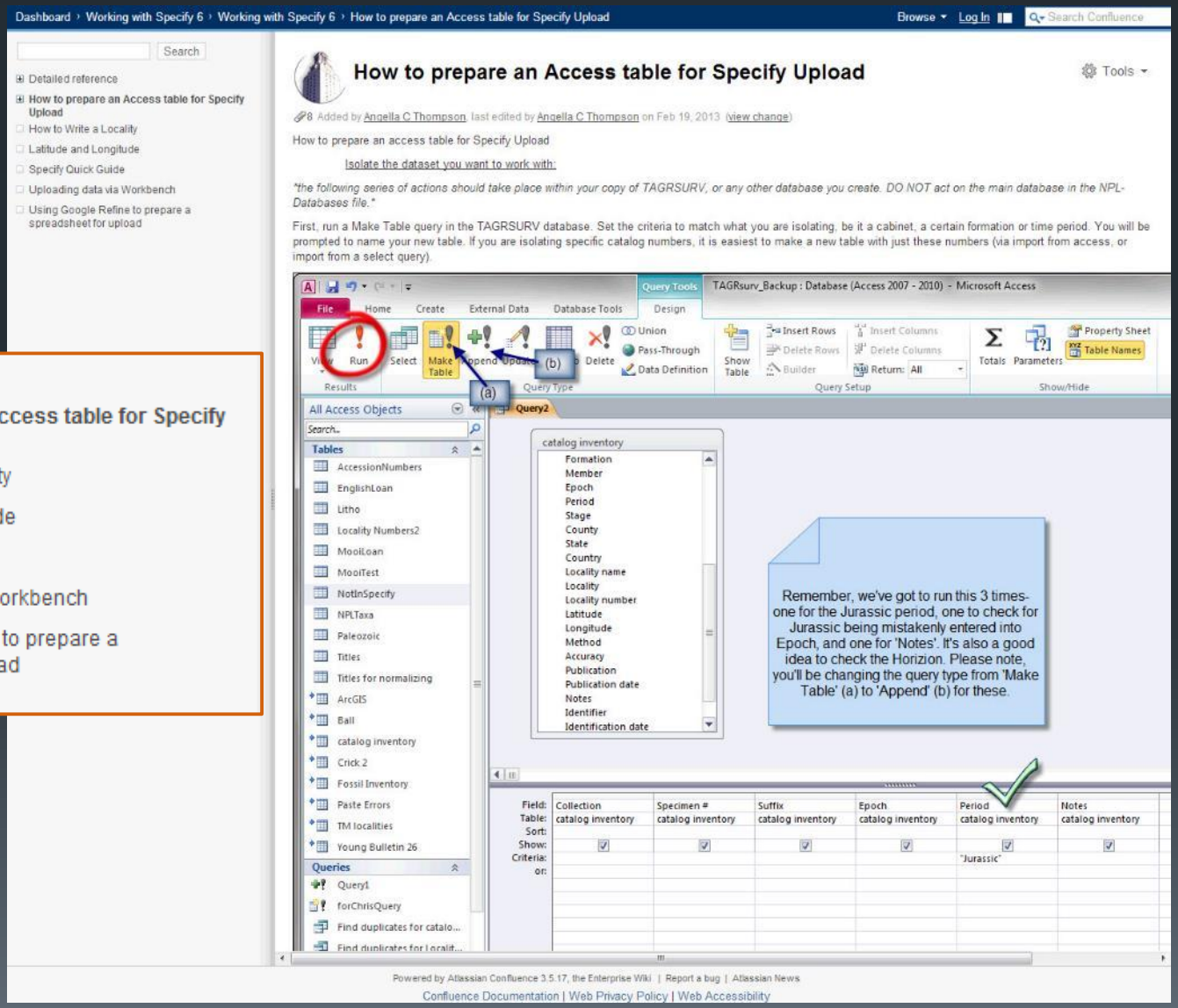
https://github.com/OpenRefine/OpenRefine/wiki

Migration flow

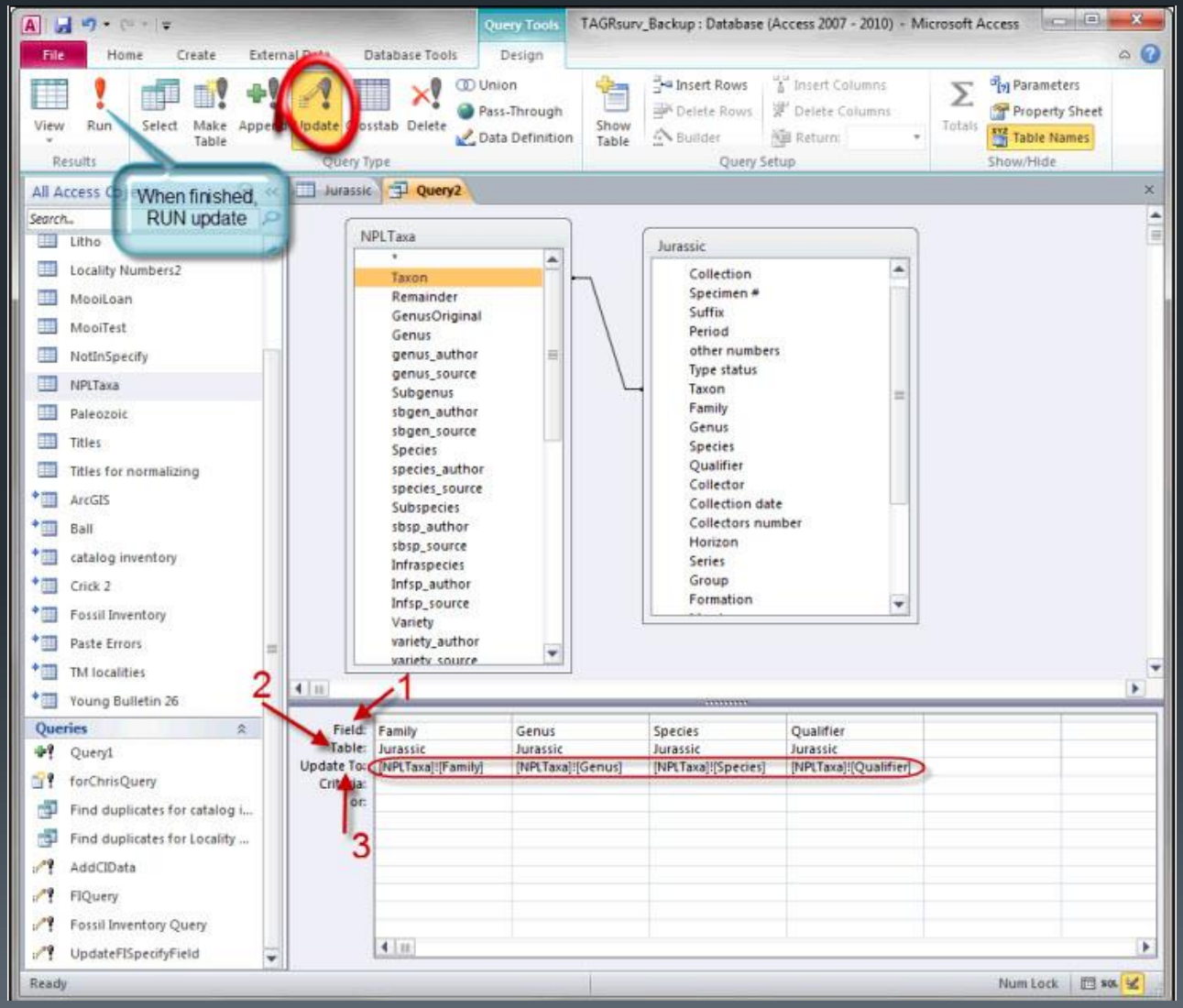


Training tool-wiki

- Detailed reference
- **How to prepare an Access table for Specify Upload**
- How to Write a Locality
- Latitude and Longitude
- Specify Quick Guide
- Uploading data via Workbench
- Using Google Refine to prepare a spreadsheet for upload



Standardize



Refine (Google/Open)

Google Refine recipe order.txt - Notepad

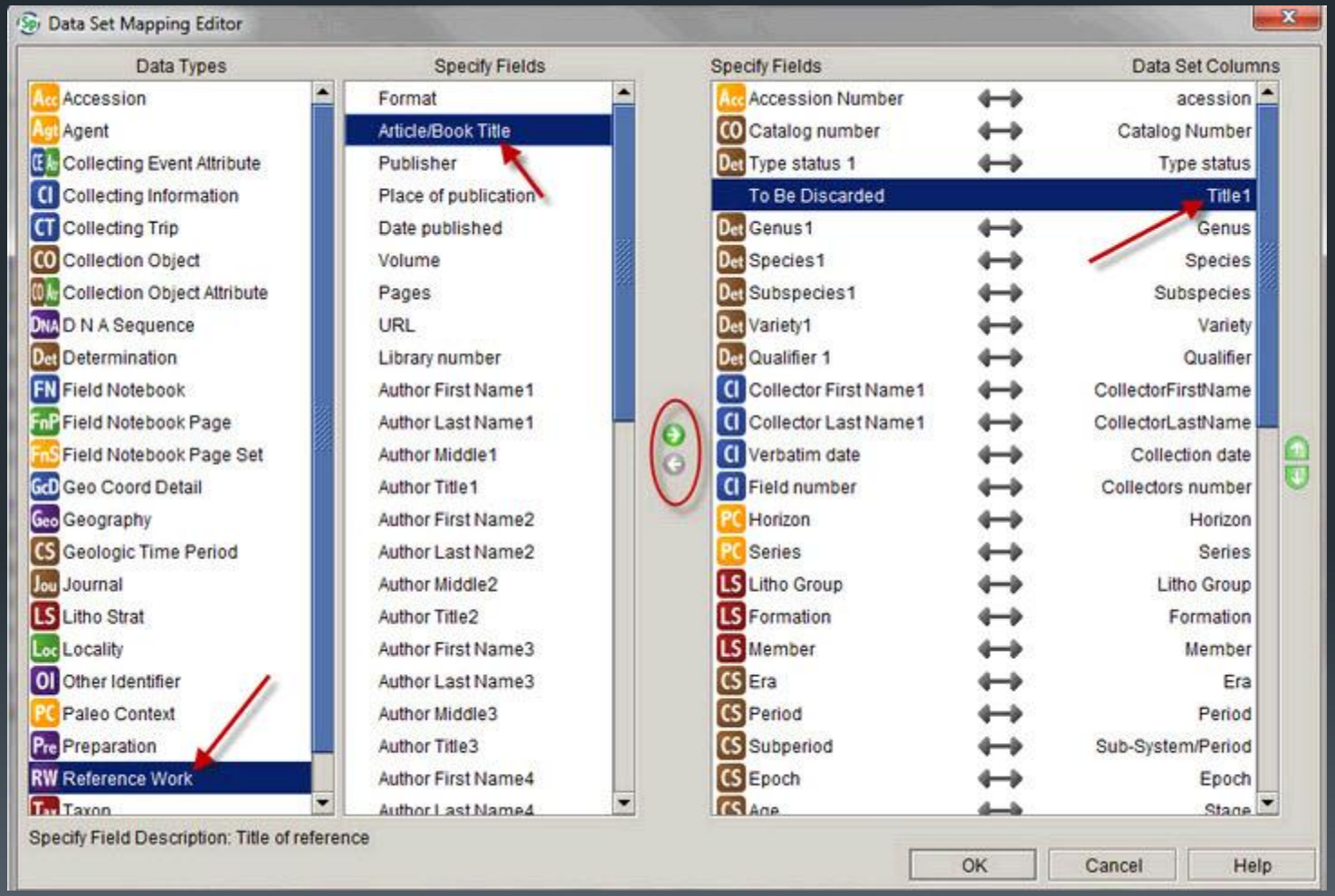
File Edit Format View Help

```
Run Translation Table steps 1-4 before uploading to Refine


1. delete the two exported to specify columns as well as any Access created index columns

2. Do a text facet clean up of Suffix (remove decimals, set nulls to "0" and make letters
uppercase)
   Do a text facet clean up of Collection as well.

3. -lumped1 (convertCatNum, combineOtherNums, combineRemarks, addColumns1...it also should
be moving columns to the correct place)

4. Find and add accession number

5. Add Columns: Qualifier 1 and Determination Remarks1

6. Check taxonomy, and add any necessary determination remarks from the Taxon Column (see
Liath for explanation) text filter taxon for "?", "aff", "cf" and "var"--many of the
varieties (subspecies) are not in the taxonomy translate table.

For any taxon names that are not translated, you can just move all of that data over to
determination remarks to be resolved later (or you can look up info and resolve the data
right away)

7. Check type columns and, then remove column Fossil Inventory
catalog inventory

8. Check accuracy column and Method for any useful info. In TA
listed as "low", "med" or "high". (if accuracy is really long,

ADD METHOD AND ACCURACY TO LONGITUDE
value + " " + cells["Method"].value + " - Accuracy: " + cells[

9. -lumped2 (renameMoveTypeStatus, combineLatLong, combineLoca
```

« Start Over | Configure Parsing Options

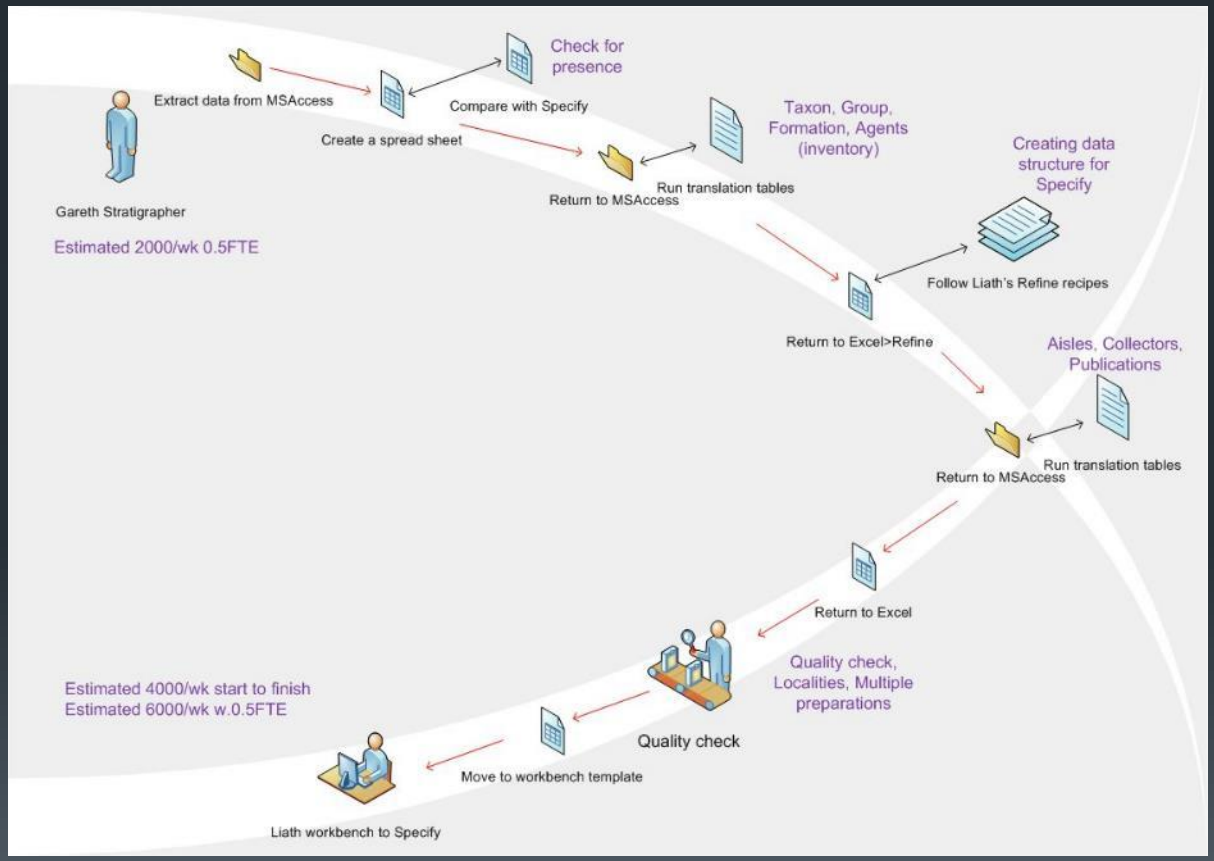
	Loan_ID	Collection	Specimen	Suffix	CatalogNumber
1.	1	TMM	822	96	TMM00000822.096
2.	2	BEG	19163	.	BEG00019163.000
3.	2	BEG	20928	.	BEG00020928.000
4.	2	BEG	19168	.	BEG00019168.000
5.	2	BEG	19166	.	BEG00019166.000
6.	2	BEG	19167	.	BEG00019167.000
7.	2	BEG	20938	.	BEG00020938.000
8.	2	BEG	19164	.	BEG00019164.000
9.	2	BEG	20926	.	BEG00020926.000

Workbench upload



Example workflow

- Content ‘expert’ to database ‘expert’
- ADD time to develop protocols
  - Translation tables from legacy data
  - Refine recipes to structure data



Workflow rates

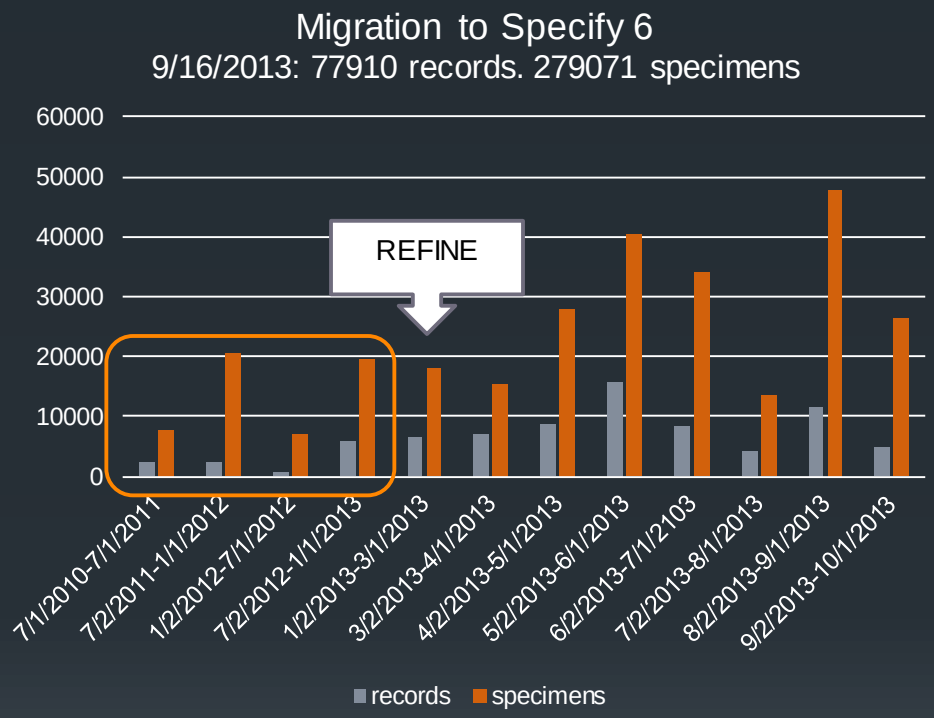
Data migration

- 4000 records per week- refine and migrate to Specify via workbench (1FTE)
- 6000 per week, with 1.5FTE
- 24000/month, 300,000/year

How long will it take?

- Current record count from all databases 500,000
- TWO years
- Assume no major glitches

Data migration rate



NB: Graph x-axis variable time scale

- iDigBio 'Train the trainers' georeferencing workshop 10/8/2012
- In-house transfer of training knowledge 11/2012
- Rate is FTE constrained and FTE knowledge base

Image flow

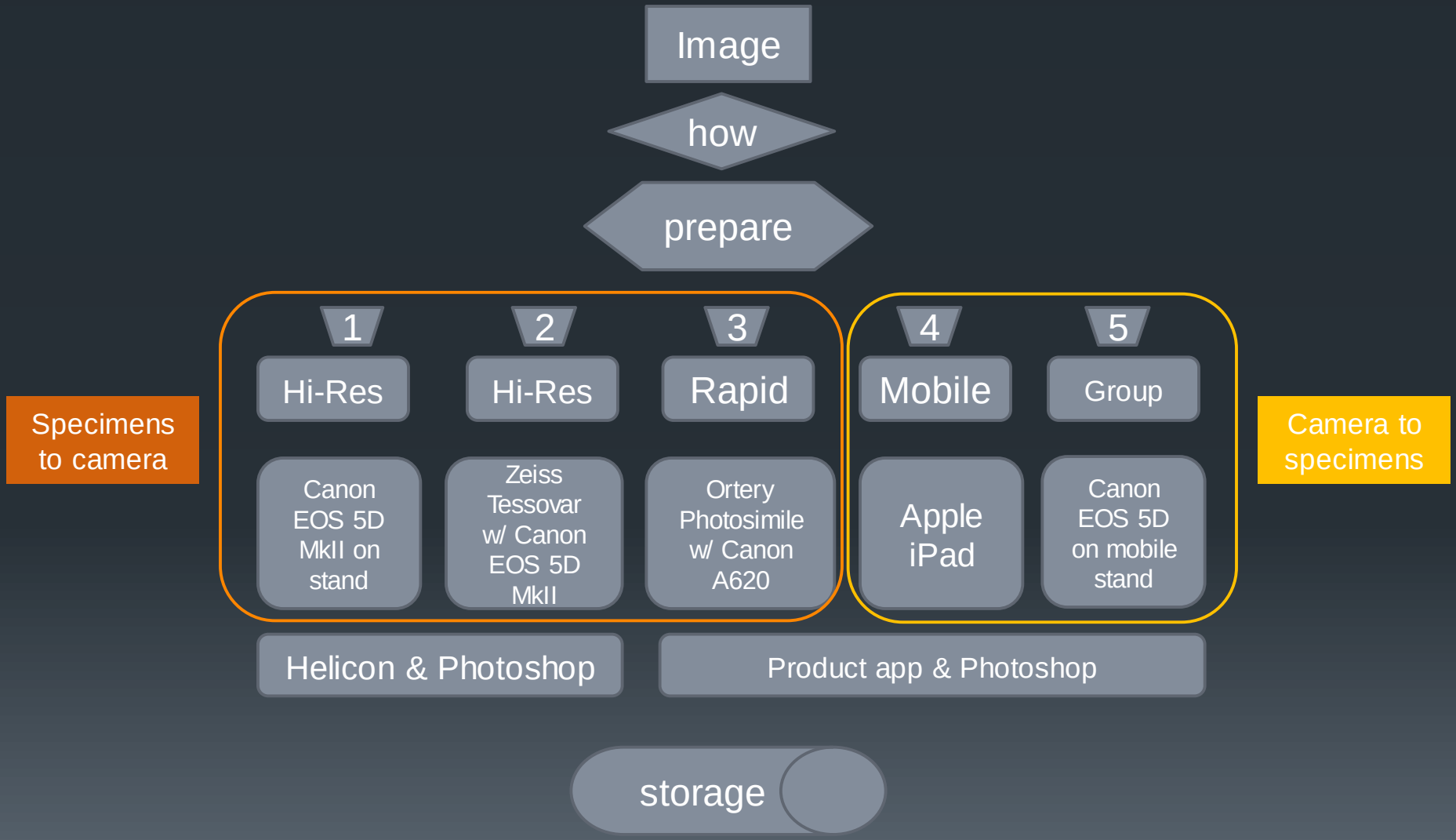


Image
how
prepare
1
2
3
4
5
Hi-Res
Hi-Res
Rapid
Mobile
Group
Specimens to camera
Camera to specimens
Canon EOS 5D MkII on stand
Zeiss Tessovar w/ Canon EOS 5D MkII
Ortery Photosimile w/ Canon A620
Apple iPad
Canon EOS 5D on mobile stand
Helicon & Photoshop
Product app & Photoshop
storage

Standardization

Common

- File naming convention
- One image views
- Standard lighting
- Scale bar
- JPEG and PDF

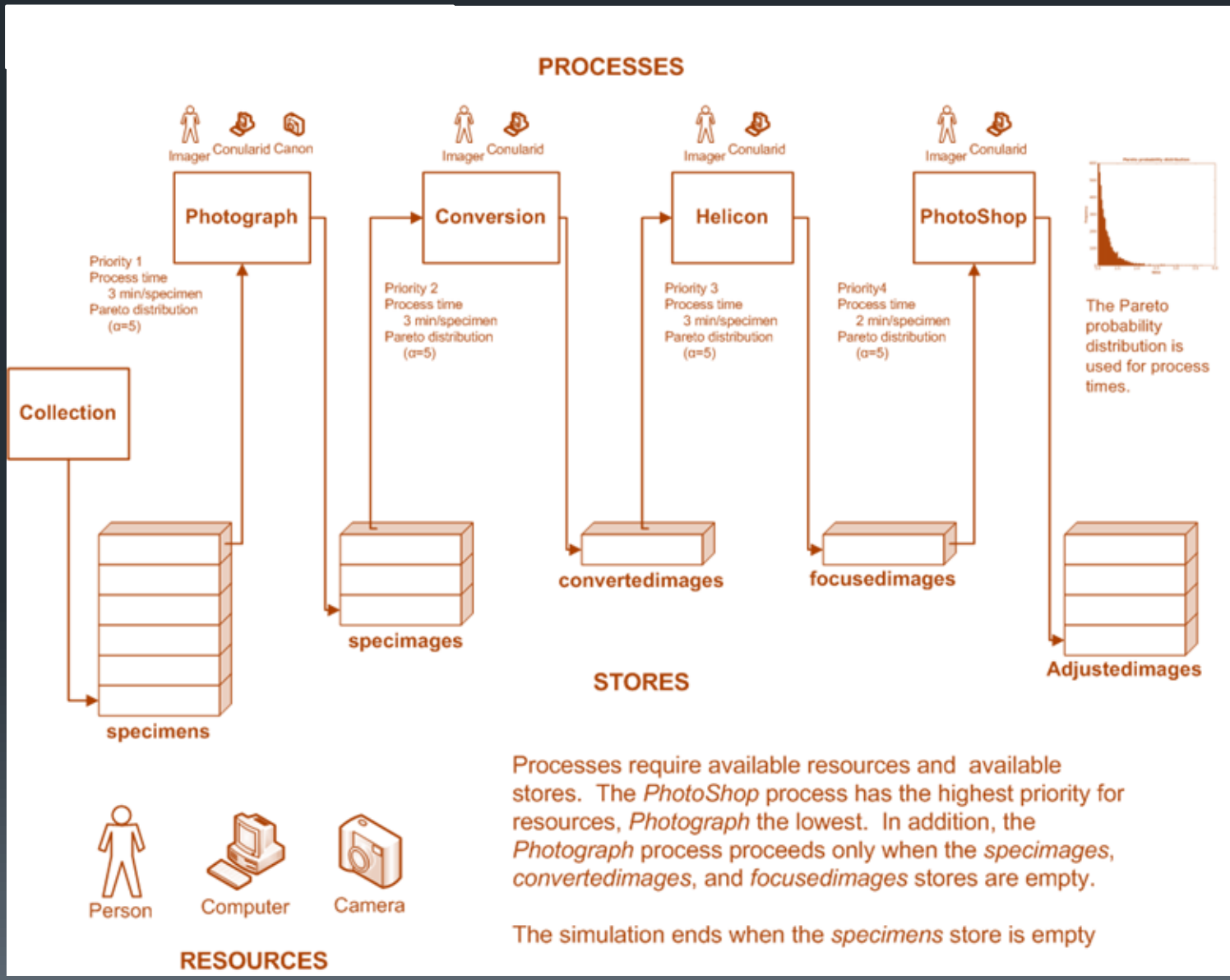
Specific to HiRes

- Multiple views
- Multi-focus composites
- Variable lighting
- Standard scale bars
- Embedded scaling
- TIFF and JPEG

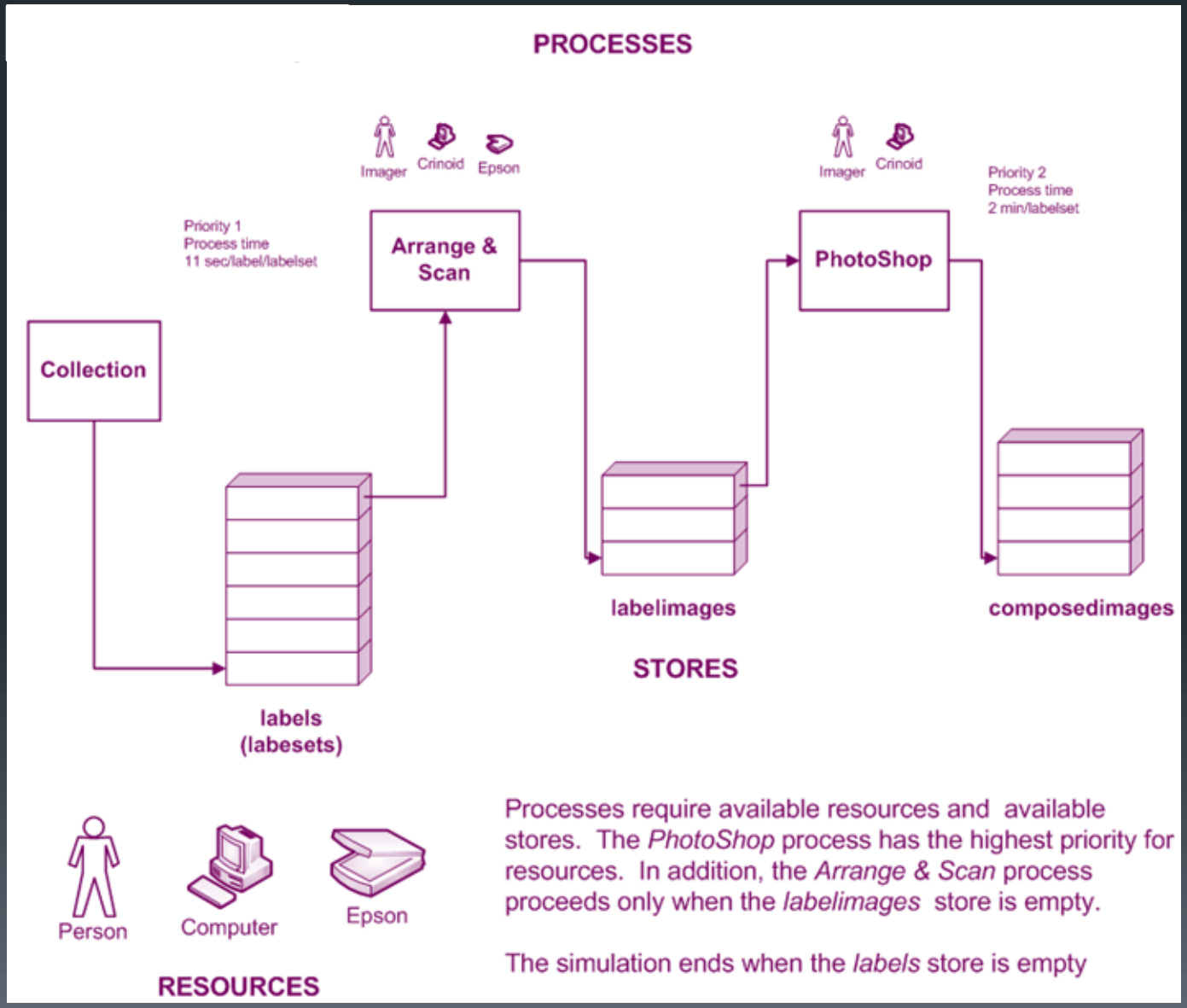
Variable

- Ammonium chloride smoked
- India ink or less permanent stain
- Emersion

3D object imaging



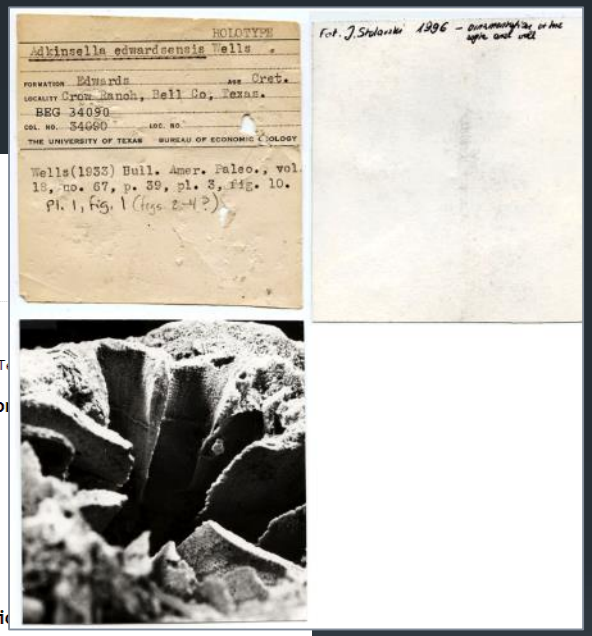
2D object scanning



Hi-Res product

3D- Specimen image

2D-Labels and photos



ion

Locality Verbal
Crow Ranch, Bell County, T

Geographic Locatio
Bell County, Texas, USA

Geologic Age
Era: Mesozoic
Period: Cretaceous
Epoch: Early
Age: Albian

Stratigraphic Positi
Group: Fredericksburg
Formation: Edwards
Member:

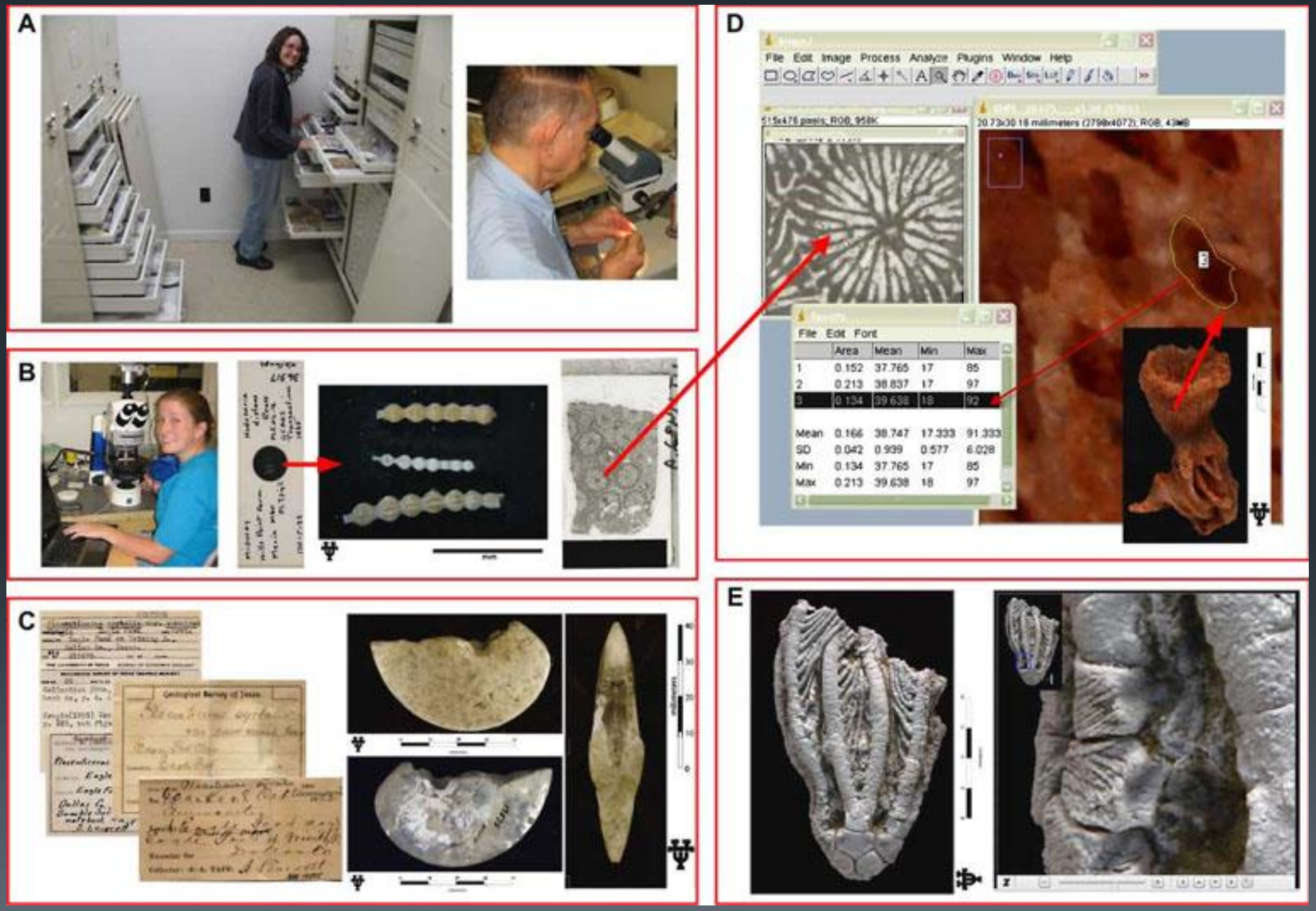
Collector
Collector: WS Adkins
Collection Date:

Publication
Wells (1933) Bull. Amer. Paleo., v. 18, no. 67, p. 39, pl. 3, fig. 10, pl. 1, figs. 1-4

Storage Location
PRC122, Types, 7, 182

Images

Image workflow



Research imaging

PaleoCentral

HOME MAP RECORDS REFERENCES TOOLS LOGIN

Types Collection

Records » TX 1121.66 « prev rec

Scientific Name
Eopinnacrinus pinnulatus Brower & Veinus n.gen., n

Common Name
crinoid

Taxonomic Hierarchy
Phylum: Echinodermata
Class: Crinoidea
Order:
Family:

Collector
Collector: Graffham, A.
Collection Date:

Publication
Brower & Veinus. 1982.Sprinkle, J. T (Ed.), University of Kansas Paleontological contributions, Monograph 1, pl. 11 fig. 3, 7, fig. 42a, fig. 42b, p. 141.

Locality Verbal
West Branch Sycamore Creek, Daube Ranch, Jo

Geographic Location
Johnston County, Oklahoma, USA

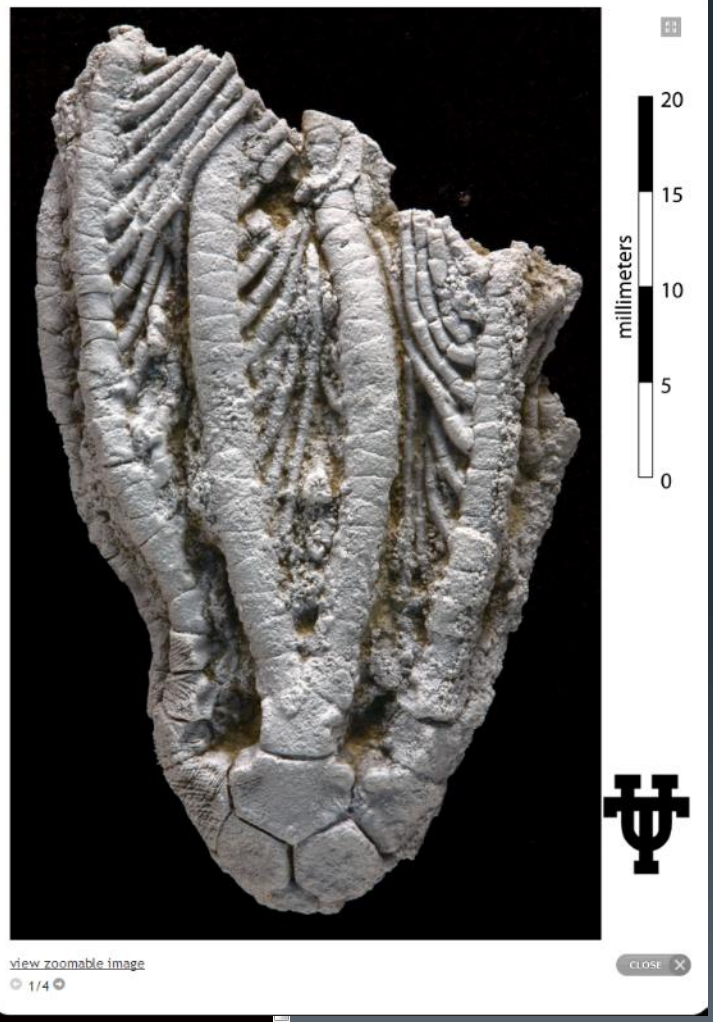
Geologic Age
Era: Paleozoic
Period: Ordovician
Epoch: Late
Age: Katian (Mohawkian)

Stratigraphic Position
Group: Simpson
Formation: Bromide
Member: Mountain Lake

Storage Location
PRC122, Types, 19, 50

Images

Eopinnacrinus pinnulatus Brower & Veinus n.gen., n



view zoomable image
1/4
CLOSE

HiRes imaging

Zeiss Tessovar

- Found adaptor on line for Canon EOS 5 Mark II
- Using the same software protocols
- Especially useful for small specimens

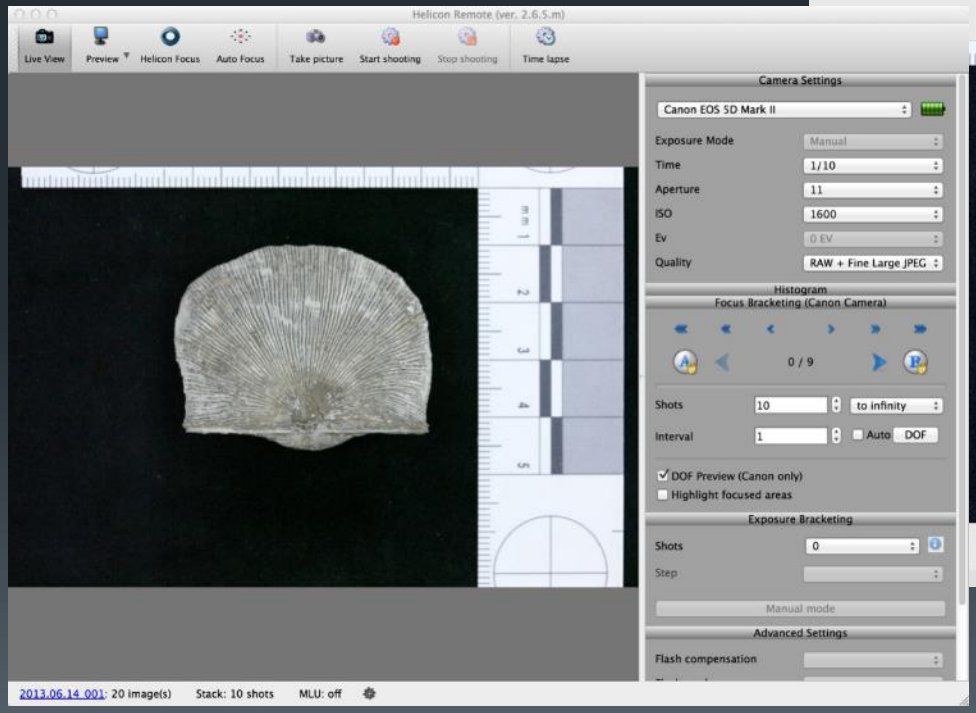
Improving throughput

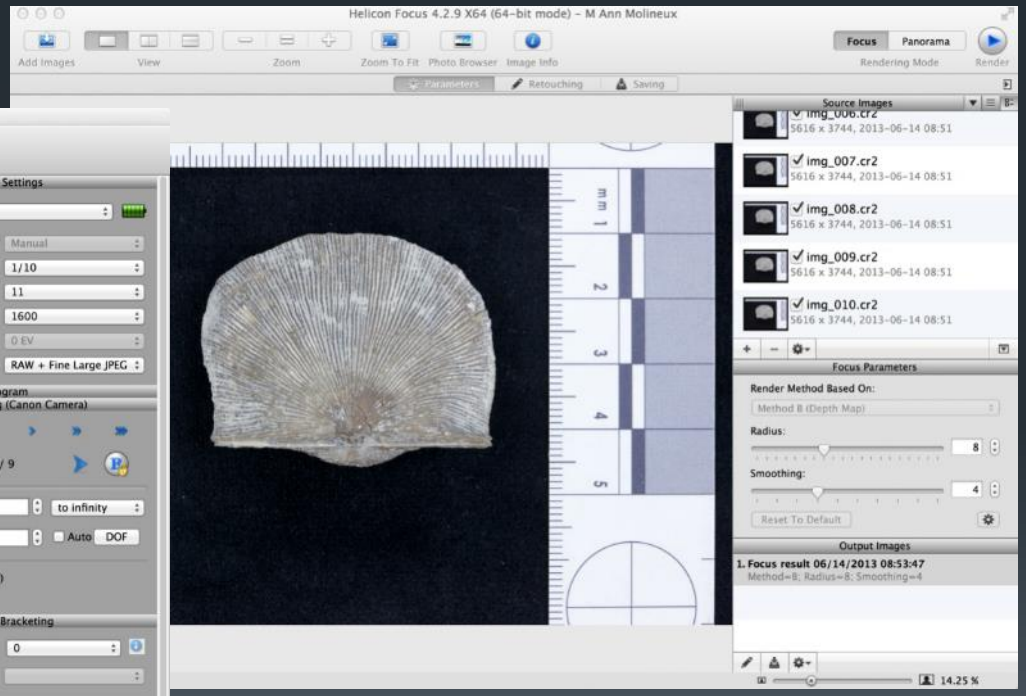
Remote shooting

- Helicon remote
- Camera lens is driven by the software

Multi-focus rendering

- Economic multi-focus shooting





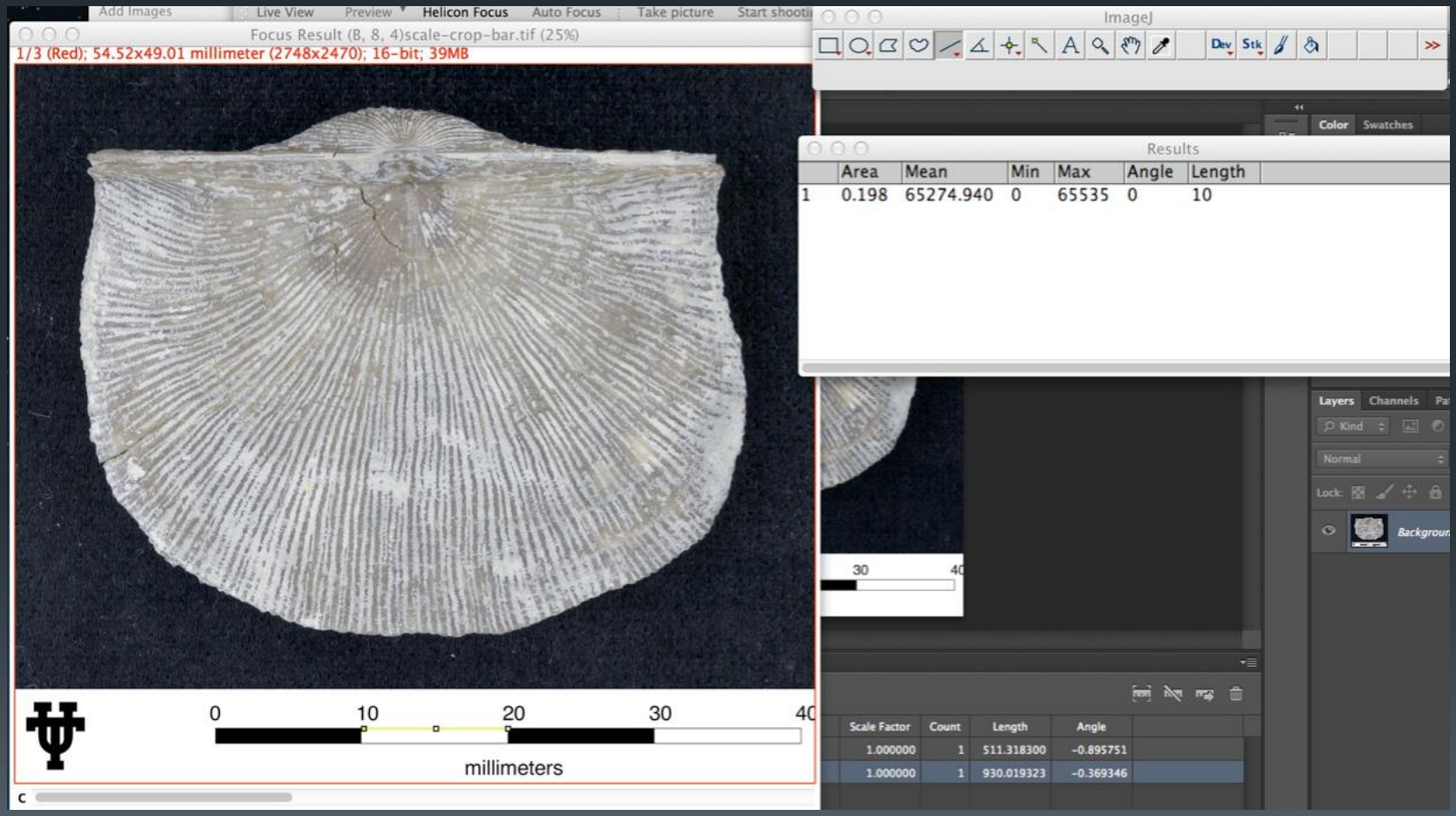
Improving throughput

Batch processing

- Final PhotoShop processing includes several JavaScripts, batch routines to generate scaled images, image metadata, standard backgrounds, scale bars, and trademarks.
- Scripts are available at http://www.utexas.edu/tmm/npl/projects/imaging

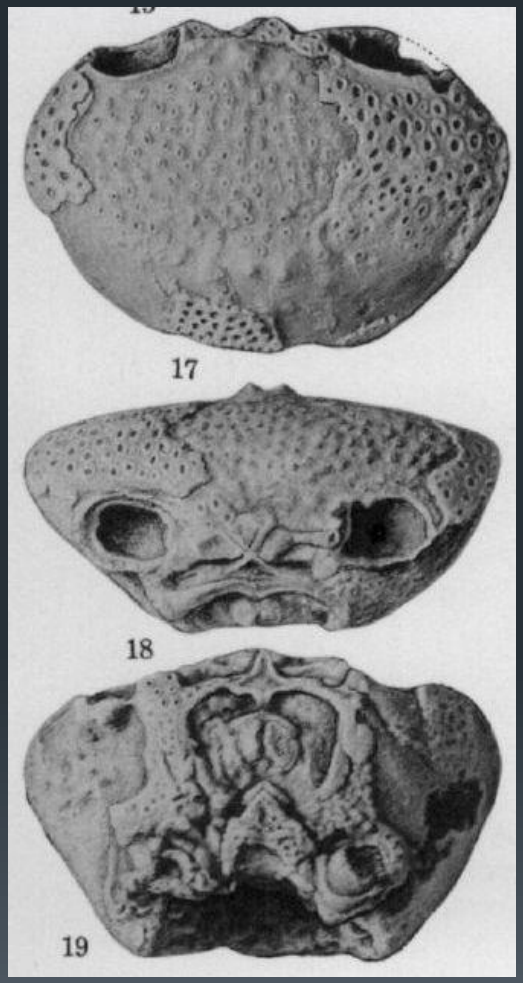
Analytical quality

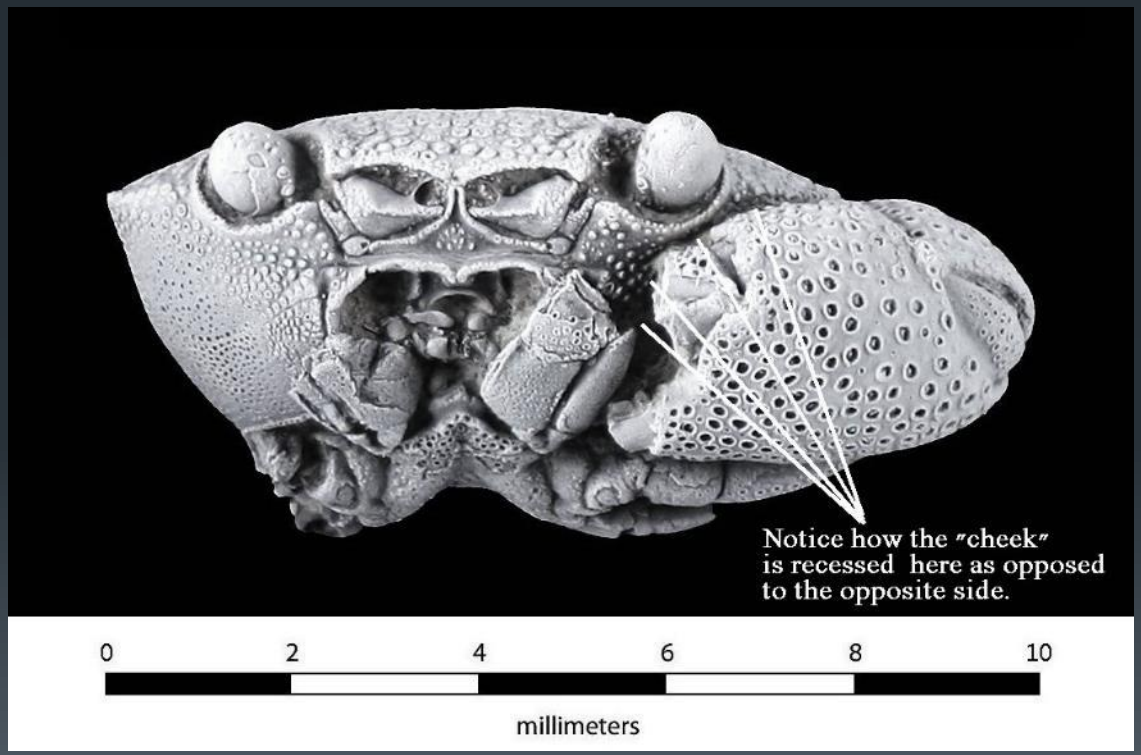
JavaScripts applied to rendered image



Example using ImageJ

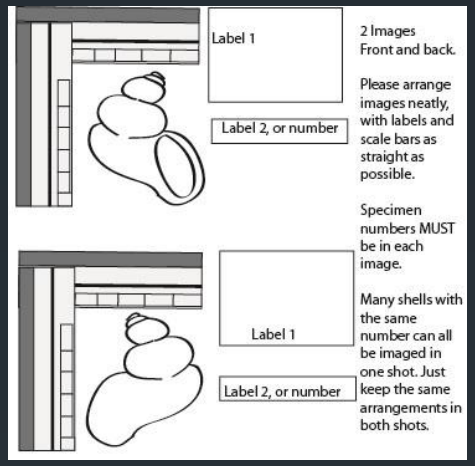
Images ‘loaned’



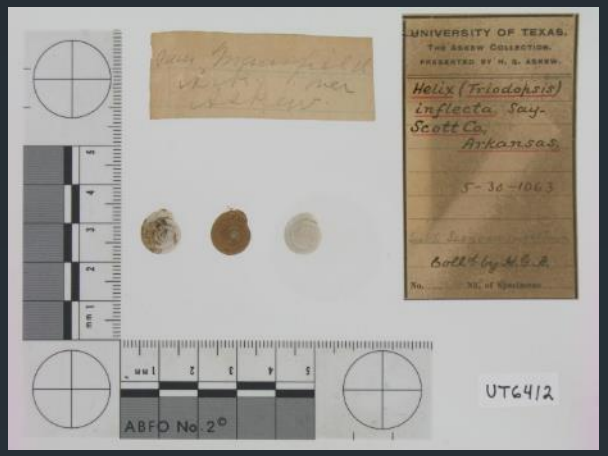


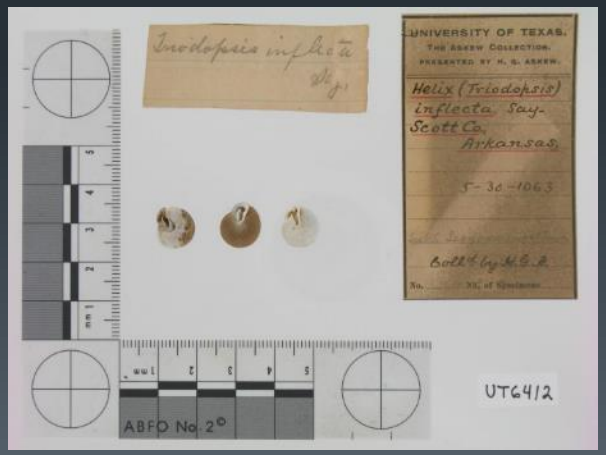
Improving throughput

Basic reference + label



Minimal processing

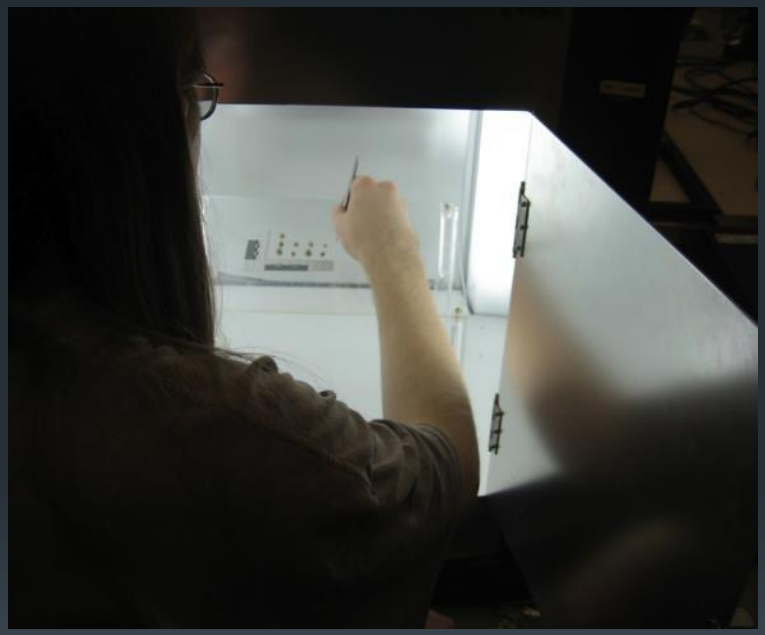


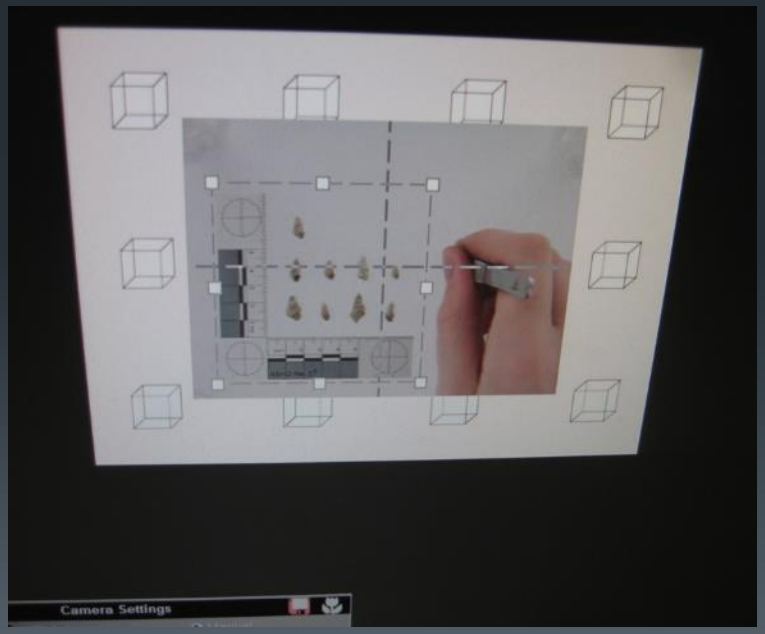


Label text is legible: Specimen detail adequate: Thumbnails for drawers

Rapid imaging

Stationary Photosimile light box





In situ imaging

Individual and drawer imaging

Basic inventory in situ

- Select drawer
- Clean and conserve specimens
- Count specimens
- Complete inventory sheet
- Make specimen # visible

Basic drawer imaging

- Image whole drawer
- Link to ArcOnline for browsing

Basic specimen imaging

- Select drawer
- Remove specimen from bag
- Flatten labels beneath glass
- Add specimens
- Image using iPad
- Bag historic labels
- Return drawer
- Attach image files to Specify record
- Create thumbnails to form ‘drawer’
- Link to ArcOnline for browsing

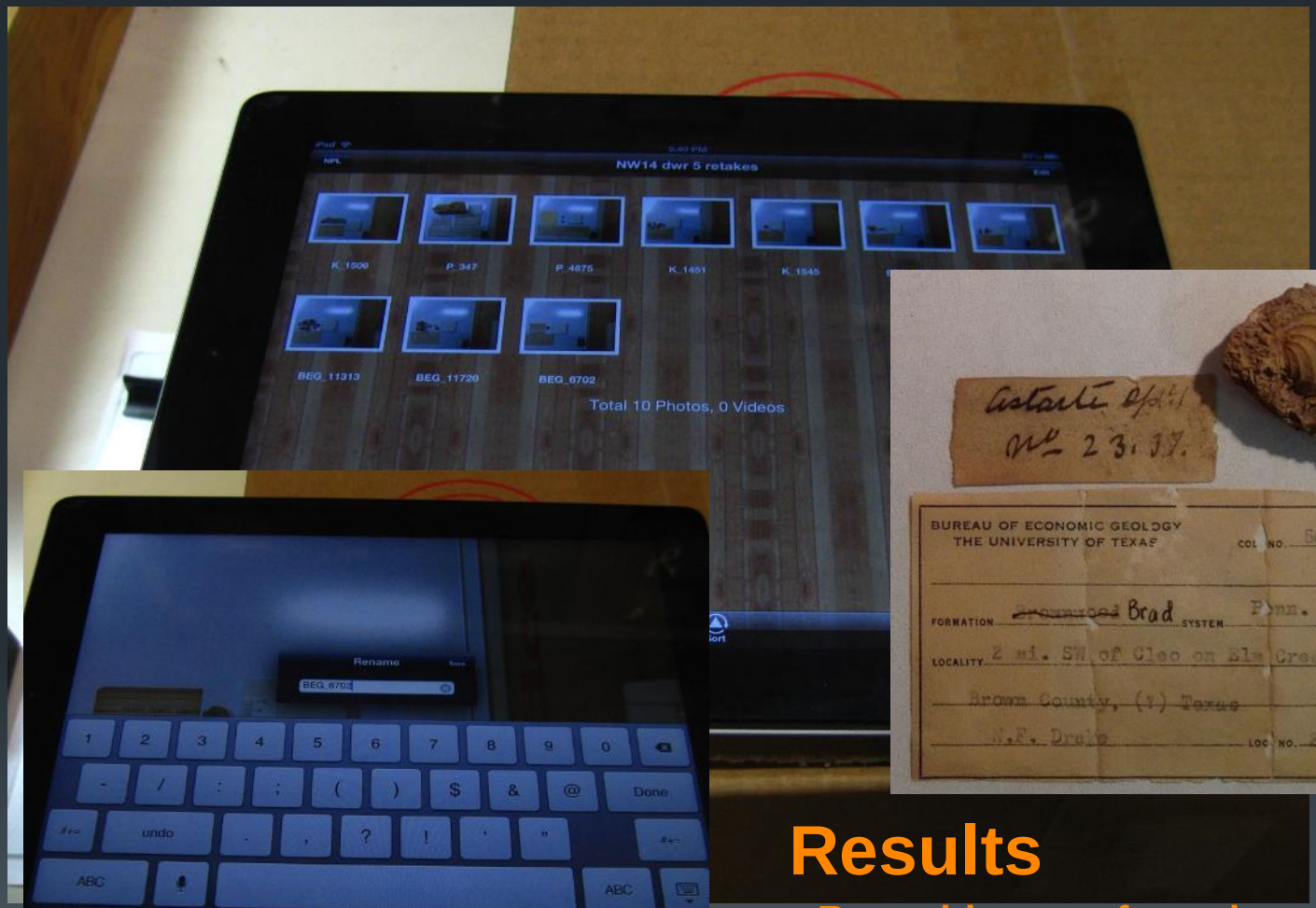
Rapid imaging flow

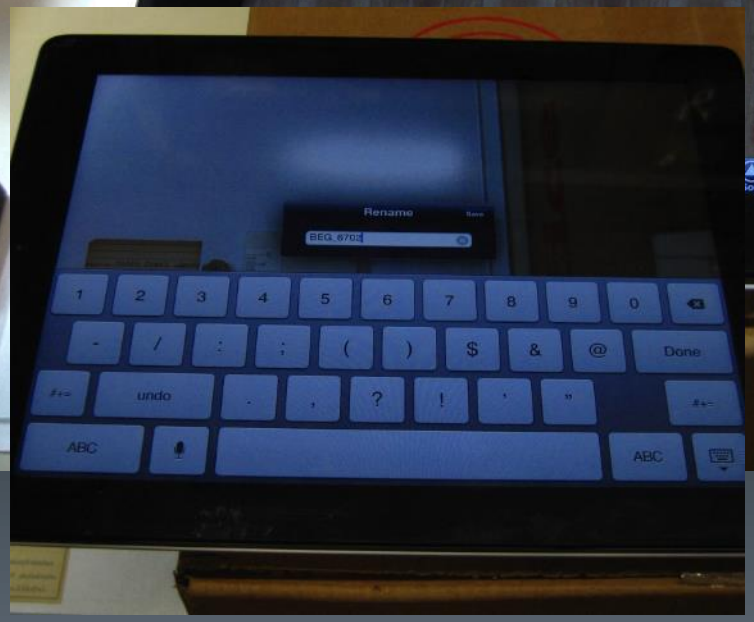
Inventory-conserve-image

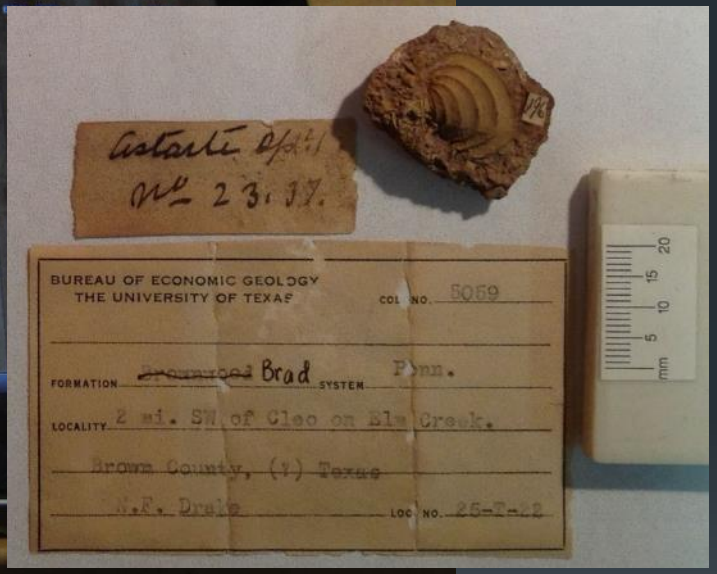
Results

- Record image of specimen
- Labels are legible and provide additional data for the database
- Can be linked to ArcOnline for browsing

In situ imaging







Results

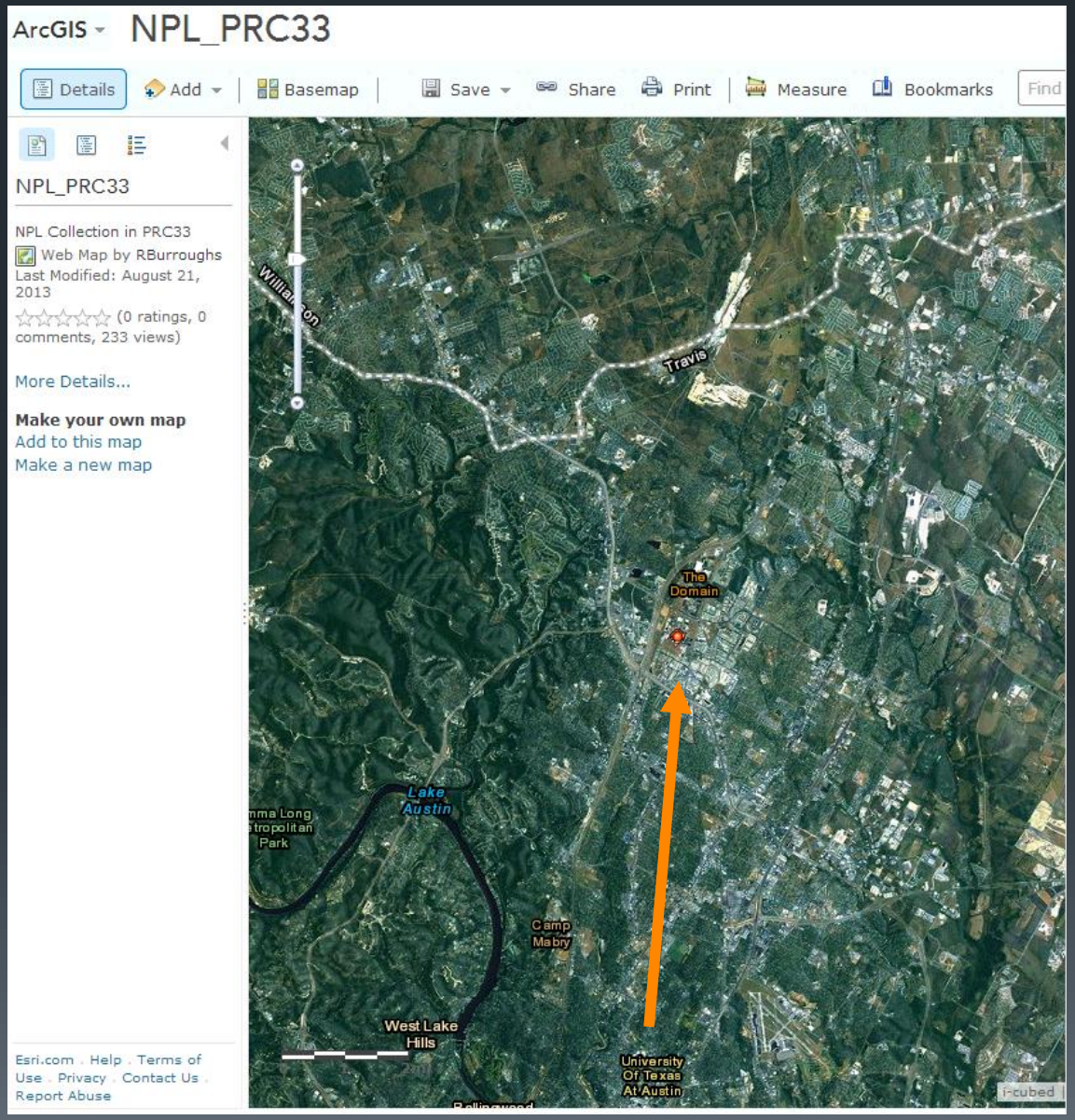
- Record image of specimen
- Labels are legible and provide additional data for the database
- Can be linked to ArcOnline for browsing

Workflow

- Imaging
  - High resolution- 5 per hour [on both Tessovar and Stand]
    - Processing-file numbering, composites, scaling
  - Basic imaging- 9-17 lots per hour
  - In situ (Mobile) imaging-17 lots per hour
- Trade-offs
  - Hi-Res- research quality- higher training curve
  - Rapid- inventory- low training curve
  - In situ- inventory- flexibility
- Results
  - All capture label data
  - All can be linked to the database as attachments
  - Complete drawer imaging provides a feasible way to virtually view via the GIS system

Online assets

- Main campus access
  - Teaching
  - Research
- Global access
  - Research
- Query access
  - Research
  - Public enquiry
- Inventory
- Conservation
- Space assets



Online accessbility

THE UNIVERSITY OF TEXAS AT AUSTIN
Jackson School of Geosciences
Non-vertebrate paleontology Lab

About NPL | History | Collections | Databases | Research | Projects | Exhibits | Links | Site Map

Home-NPL | Home-JSG

About NPL

Practices & Protocols

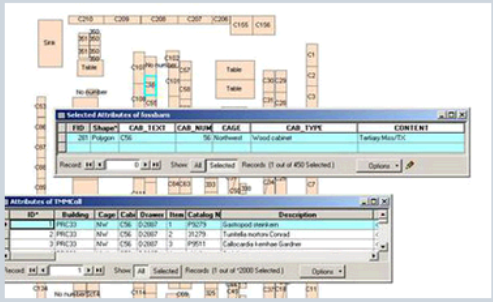
GIS-Geographic Information Systems

The databases are linked to a geographic information system (GIS) map of both repository buildings. This allows us to find information about each specimen when in the repository, and also to know where it is located within the buildings of the repository.

Current developments include on-line access to enable virtual browsing of the specimens with whole drawer images.

The GIS linked system improves our emergency management and will enhance our pest management. For example, we can locate zones susceptible to concrete moisture 'weeping' and ensure that any cabinets in those areas are raised off the floor surface. We can also monitor areas with noticeable pests and target mitigation procedures more effectively.

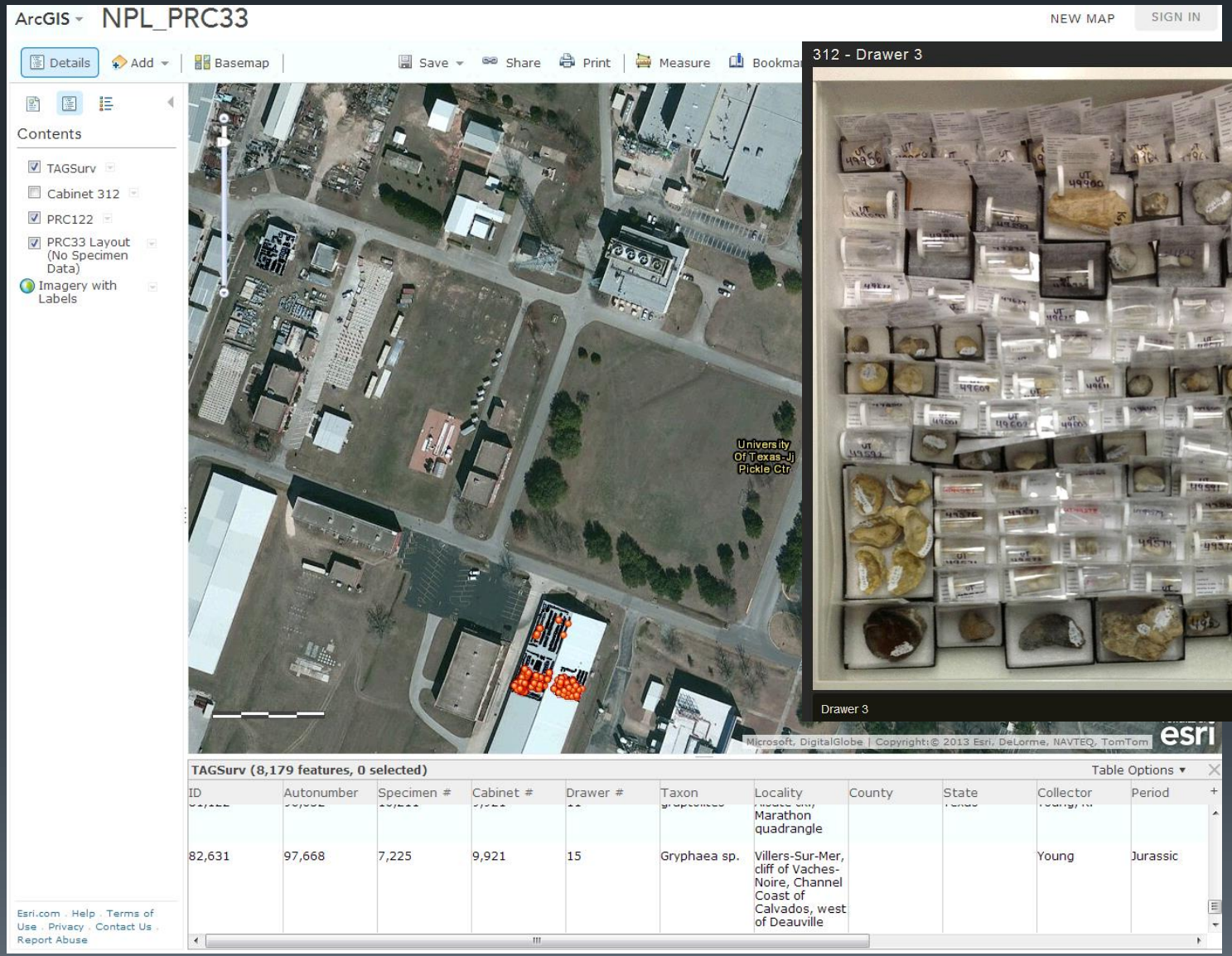
Search our site | Clear



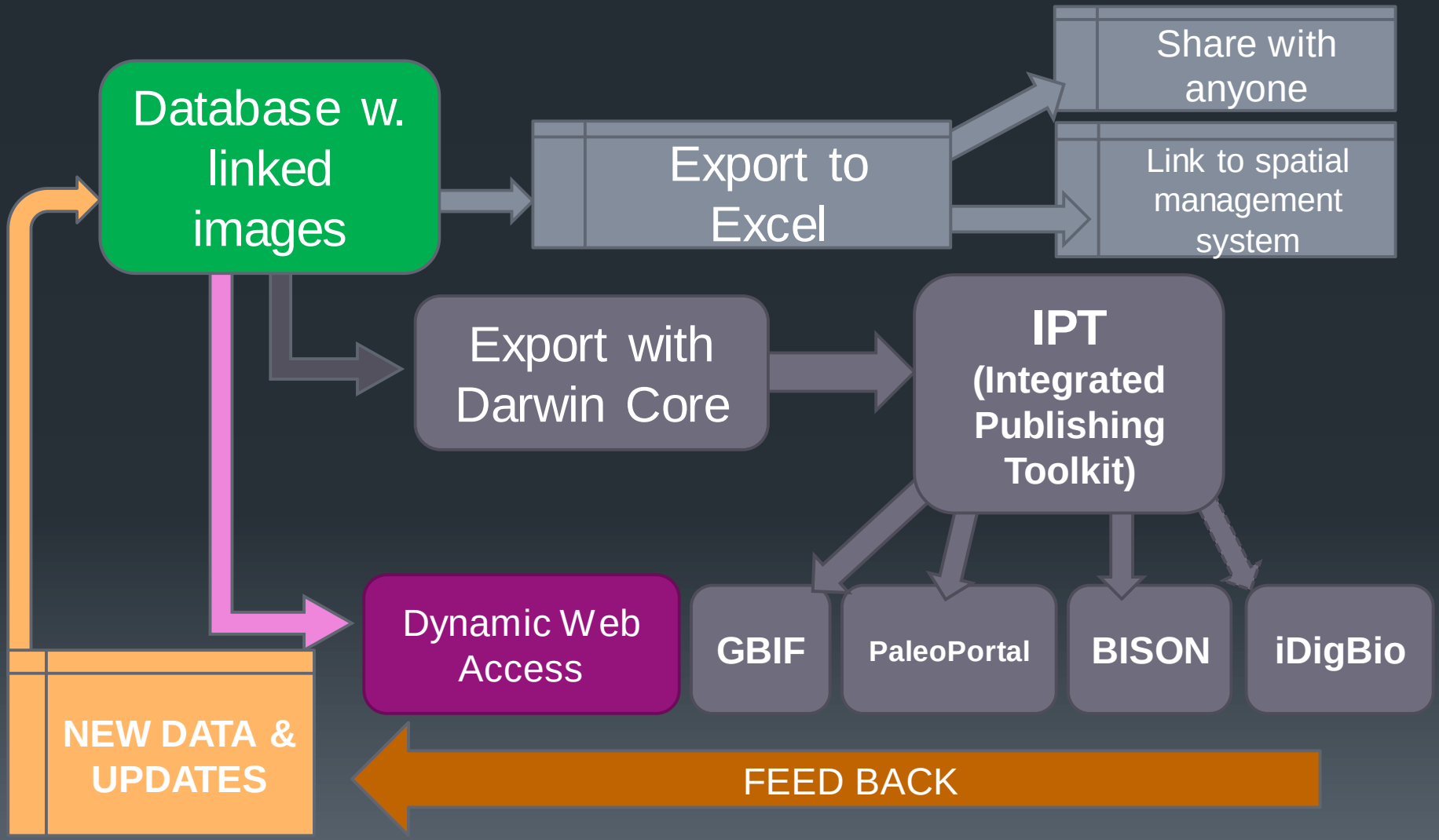
**Part of the GIS map of the NW cage in the main repository.
The tabular details refer to the highlighted cabinet on the map.
Other aspects of the collection environment are tied to this system.**

Visually browse our collections below or access more details via this test link

Online browsing



Goal


Database w. linked images
Export to Excel
Share with anyone
Link to spatial management system
Export with Darwin Core
IPT (Integrated Publishing Toolkit)
Dynamic Web Access
GBIF
PaleoPortal
BISON
iDigBio
NEW DATA & UPDATES
FEED BACK