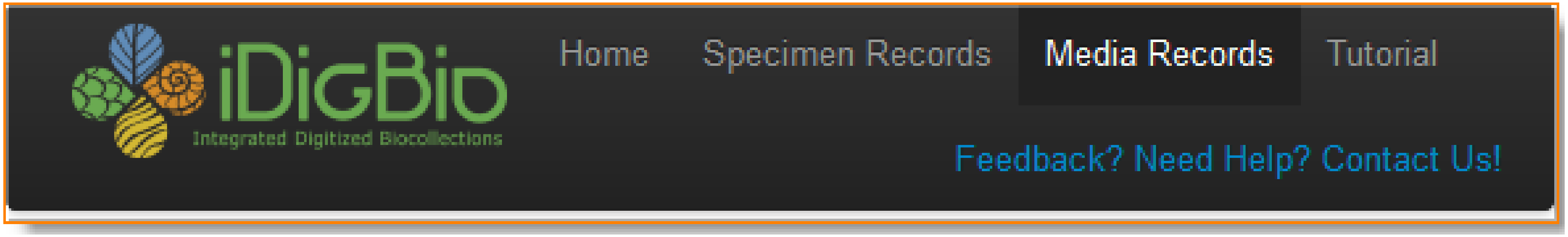

iDigBio
Integrated Digitized Biocollections
Home
Specimen Records
Media Records
Tutorial
Feedback? Need Help? Contact Us!

Pre & Post Digitization Curation

decisions - opportunities - options

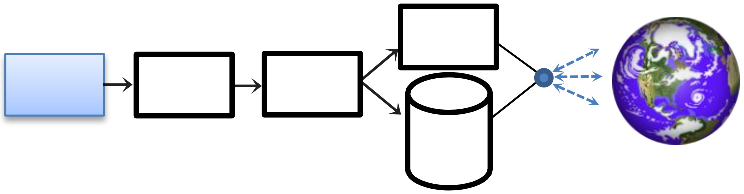
Deborah Paul, Gil Nelson

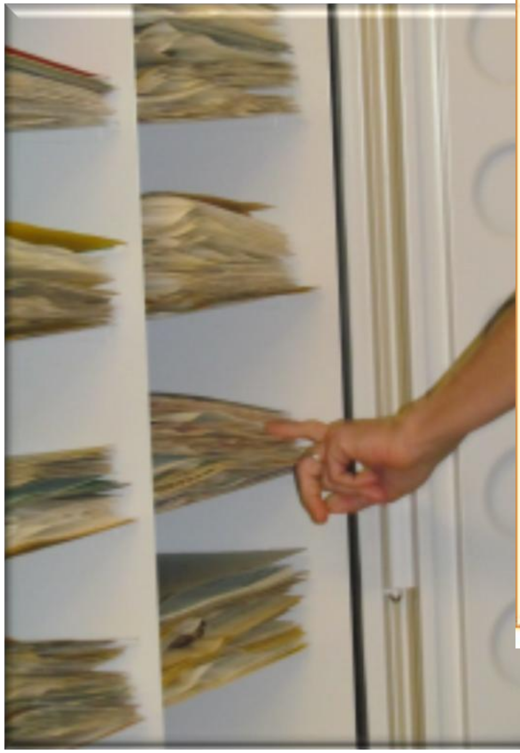
Valdosta State University, September 17 – 18, 2012

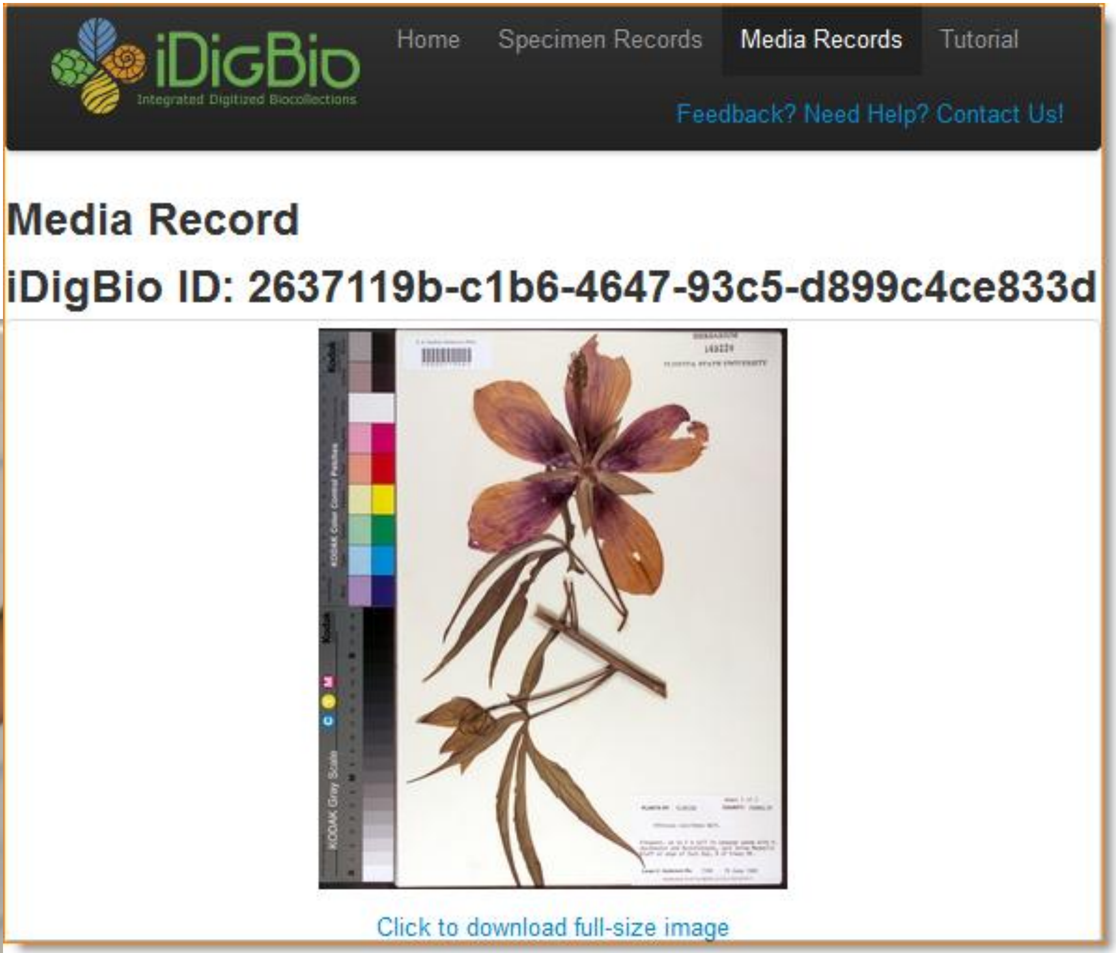
iDigBio Digitizing Vascular and Non-vascular Plant Collections Workshop

support from NSF grant: Advancing Digitization of Biological Collections Program (#EF1115210)

Pre & Post Digitization Curation





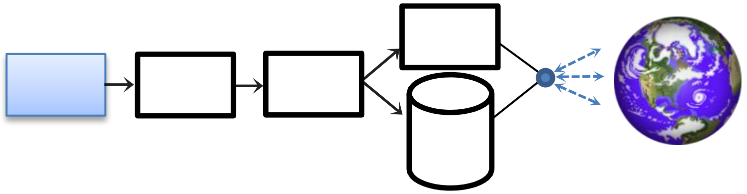

iDigBio
Integrated Digitized Biocollections
Home
Specimen Records
Media Records
Tutorial
Feedback? Need Help? Contact Us!

Media Record

iDigBio ID: 2637119b-c1b6-4647-93c5-d899c4ce833d

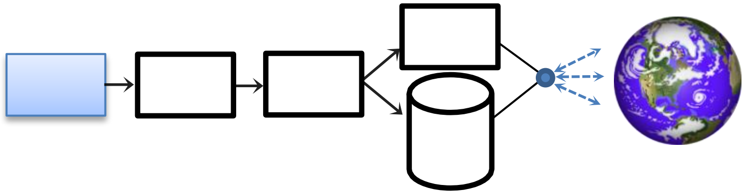
Click to download full-size image

Overview



- 5 task clusters
- Pre-digitization curation
  - – Decisions / Opportunities / Options
  - – Key point: specimen handling is an opportunity
- Post-digitization curation
  - – Revisualization is revealing
  - – Data Quality / Data Enhancement / Data Discovery

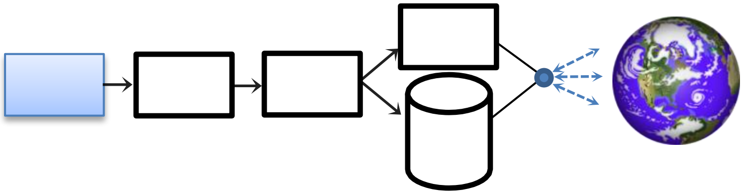
Characterizing Workflows



- We can divide activities into coherent groups Task Clusters
  - **pre-digitization curation**
  - data capture & processing
  - imaging capture
  - image processing
  - image storage
- The entire workflow process
  - Data Quality
  - Data Integrity

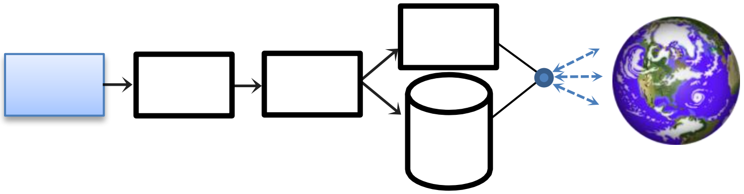
iDigBio Integrated Digitized Biocollections NSF UNIVERSITY OF FLORIDA FLORIDA STATE UNIVERSITY FLORIDA MUSEUM OF NATURAL HISTORY

DROID Workflows Workshop



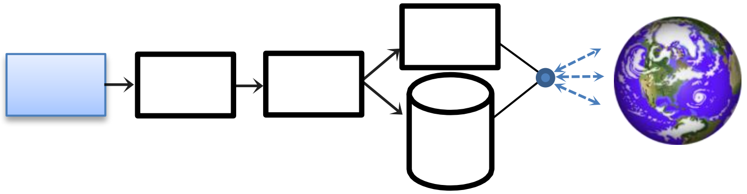
- **D**eveloping **R**obust **O**bject to **I**mage to **D**ata Workflows
  - Workflows by storage type
  - DROID1 – flat sheets
  - Module 1 – Pre-digitization Curation

DROID1 –
Pre-Digitization Module



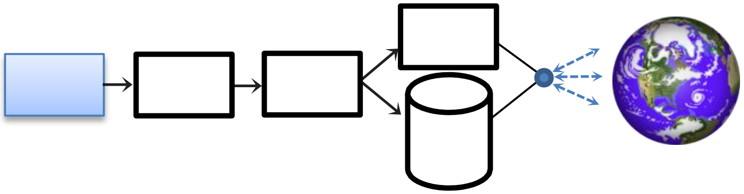
- Pre-digitization Curation for Flat Sheets
  - each module has tasks, T1, T2, T3, T4, …
  - **designed to help projects choose steps appropriate to their collection and digitization project**
  - ff-fb (again)

Pre-Digitization Module Tasks



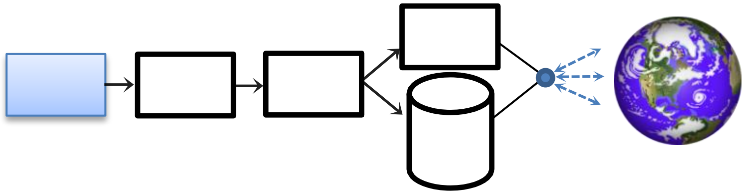
- T1 – apply storage locator barcodes
- T2 – selecting what to digitize
- T3 – apply machine readable barcodes at collection level
- T4 – locate specimens (flag cabinets)
- T5 – pull specimens from cabinet*
  - *(optional) sort by collector, date, geography
- T6 – curate collection in place (check nomenclature and annotations)

Pre-Digitization Module
Tasks



- T7 – transport specimen to imaging station
- T8 – placeholder to flag pulled specimens
- T9 – sort to remove any already imaged / barcoded
- T10 – separate specimens needing conservation work before imaging
- T11 – apply barcodes
- T12 – create skeletal database record

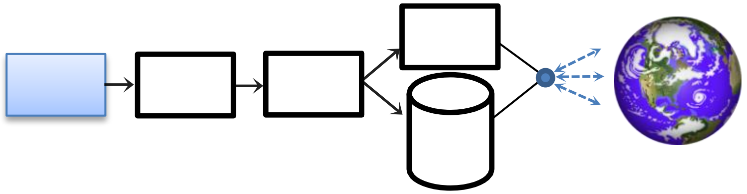
Pre-Digitization Opportunities



- evaluate collection health
  - aka "collection profiling"
  - Profiling Natural History Collections: A Method for Quantitative and Comparative Health Assessment
  - hard data for museum directors & administrators
  - "an important tool in reinvigorating collection management and in particular providing data to support funding requests."
- finding unknown unknowns and lost material
- experts or non-experts?
- high-hanging fruit (or tasks perhaps long put off)
  - cabinet reorganization
  - equipment updates
  - loan returns
  - specimen repair

iDigBio Integrated Digitized Biocollections · NSF · UNIVERSITY OF FLORIDA · FLORIDA STATE UNIVERSITY 1851 · FLORIDA MUSEUM OF NATURAL HISTORY

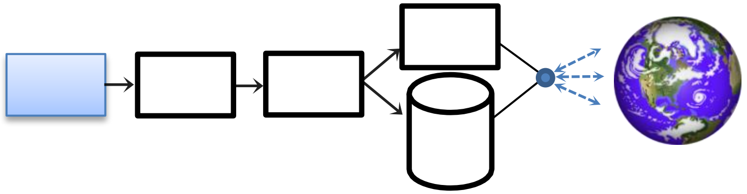
Evaluate Pre-Digitization Module Tasks



	Evaluate Pre-digitization tasks	Pre-digitization	Post-digitization	Expert	Non-expert	Crowd source	Automation	QA / QC	Authority files	Time \ Cost computation	Finding slowest steps
t1	apply storage locators										
t2	what to digitize										
t3	barcodes at collection level										
t4	locate specimens - flag cabinets										
t5	pull specimens										
t6	curate collection in place										
t7	transport specimens										
t8	placeholders in cabinets										
t9	sort already imaged / barcoded										
t10	conservation										
t11	apply barcodes										
t12	skeletal database record										

- Keep staffing in mind as well as
- new developments
- track issues / document
- decisions, opportunities, options

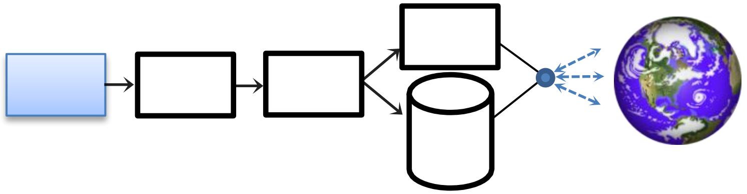
Factors for Task Order



- are experts needed?
- where will conservation fit in?
- how will materials from conservation get back into the digitization workflow?
- “filed as” or up-to-date taxonomy?
- collection size factor
- isolating steps that can be done after digitization
- reliance on the database
- “imaged” written in ink or pencil on specimen?
- *how many are actively using the collection?
- *how do those using the collection cooperate in this process?

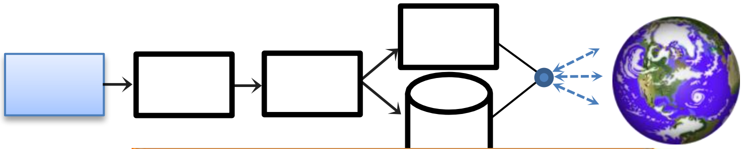
iDigBio Integrated Digitized Biocollections NSF UNIVERSITY OF FLORIDA FLORIDA STATE UNIVERSITY 1851 FLORIDA MUSEUM OF NATURAL HISTORY

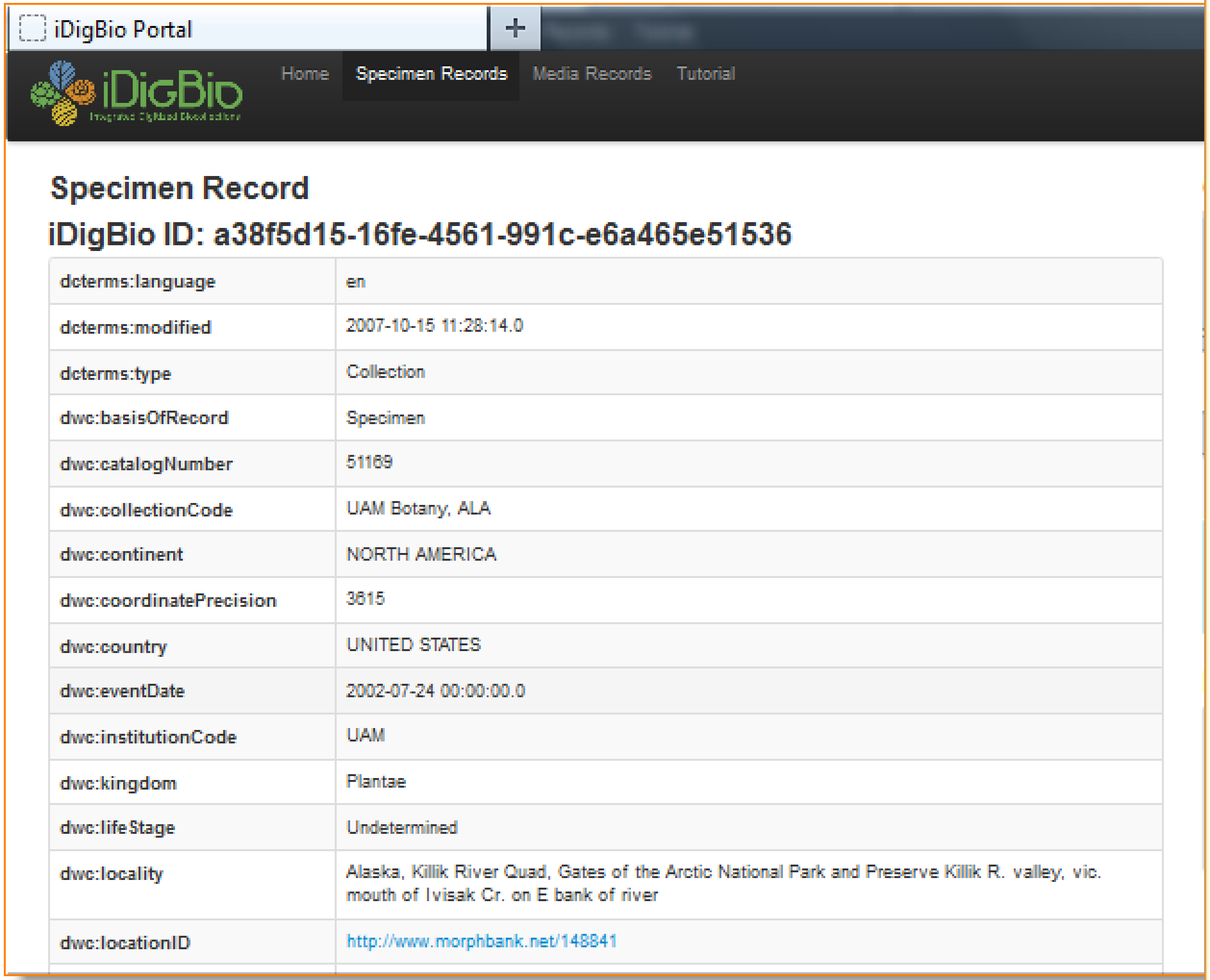
Overview



- 5 task clusters
- Pre-digitization curation
  - Decisions / Opportunities / Options
  - Key point: specimen handling is an opportunity
- **Post-digitization curation**
  - Revisualization is revealing
  - Data Management
  - Data Quality / Data Enhancement / Data Discovery

Revisualization

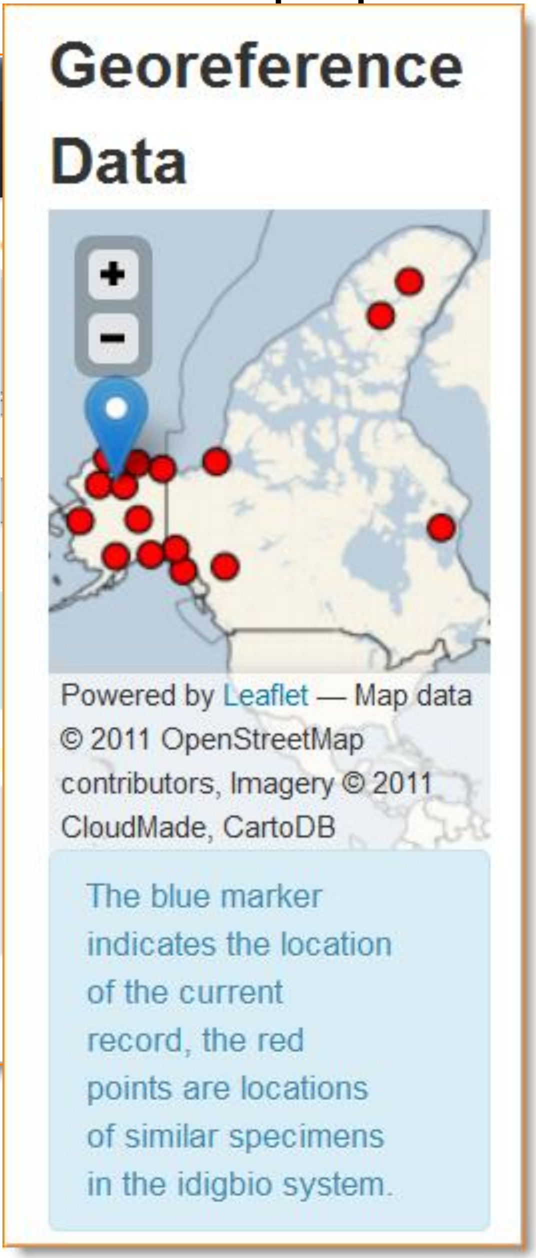




Specimen Record

iDigBio ID: a38f5d15-16fe-4561-991c-e6a465e51536

dcterms:language	en
dcterms:modified	2007-10-15 11:28:14.0
dcterms:type	Collection
dwc:basisOfRecord	Specimen
dwc:catalogNumber	51169
dwc:collectionCode	UAM Botany, ALA
dwc:continent	NORTH AMERICA
dwc:coordinatePrecision	3615
dwc:country	UNITED STATES
dwc:eventDate	2002-07-24 00:00:00.0
dwc:institutionCode	UAM
dwc:kingdom	Plantae
dwc:lifeStage	Undetermined
dwc:locality	Alaska, Killik River Quad, Gates of the Arctic National Park and Preserve Killik R. valley, vic. mouth of Ivisak Cr. on E bank of river
dwc:locationID	http://www.morphbank.net/148841

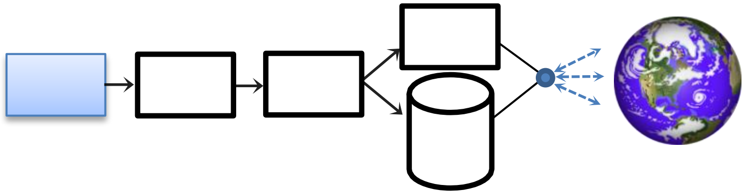


Georeference Data

Powered by Leaflet — Map data © 2011 OpenStreetMap contributors, Imagery © 2011 CloudMade, CartoDB

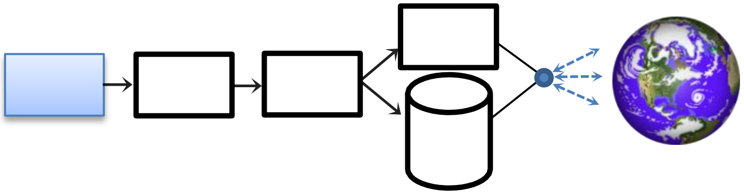
The blue marker indicates the location of the current record, the red points are locations of similar specimens in the idigbio system.

Post-Digitization



- Revisualization is revealing
- Querying dataset to find / fix errors
  - filename errors
  - typos
  - georeferencing errors
  - taxonomic errors
  - guid errors
  - format errors (dates)
  - mapping (from Workbench for example)

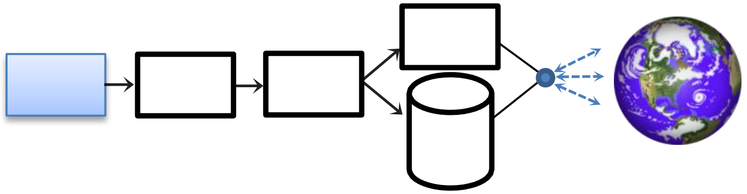
Post-Digitization



- Using new tools
  - Kepler Kurator – Data Cleaning, Data Enhancement
  - Google Refine, desktop app
    - from messy to marvelous
    - http://code.google.com/p/google-refine/
    - remove leading / trailing white spaces
    - standardize values
- Query / Report / Update features of Databases
  - learn SQL

iDigBio Integrated Digitized Biocollections NSF UNIVERSITY OF FLORIDA FLORIDA STATE UNIVERSITY FLORIDA MUSEUM OF NATURAL HISTORY

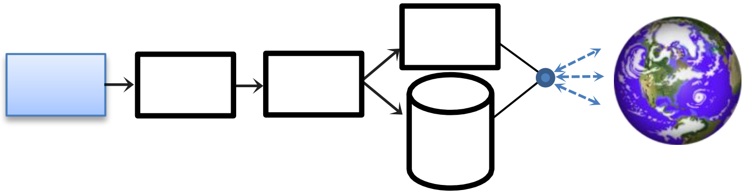
Post-Digitization



- Filtered PUSH Project
- Scatter, Gather, Reconcile – Specify
- Exposing Data to Outside Curation
- Planning for Ingestion of Feedback – Policy Decisions
- re-determinations
  - the annotation dilemma
  - to re-image or not to re-image
  - “annotated after imaged”
  - to attach a physical annotation label to the specimen from a digital annotation or not (Flora of North America)
- georeferences – new or different from existing
- data discovery
  - dupes, grey literature, annotations of many kinds

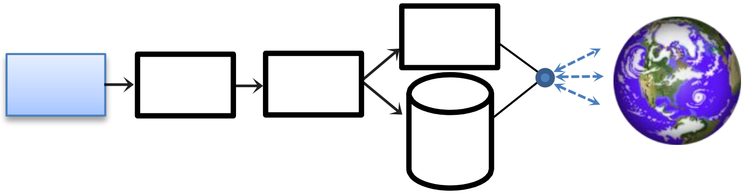
iDigBio Integrated Digitized Biocollections
NSF
UNIVERSITY OF FLORIDA
FLORIDA STATE UNIVERSITY 1851
FLORIDA MUSEUM OF NATURAL HISTORY

Post-Digitization



- Managing Data Enhancement
  - Crowd-Sourcing
    - completing skeletal records
    - georeferencing
    - multi-keying
  - Geolocate
  - Symbiota
- Opportunities for
  - Data Quality
  - Data Integrity
  - Data Enhancement
- Sharing the data (the other presentation).
- Everyone is looking forward to lots of new datasets!

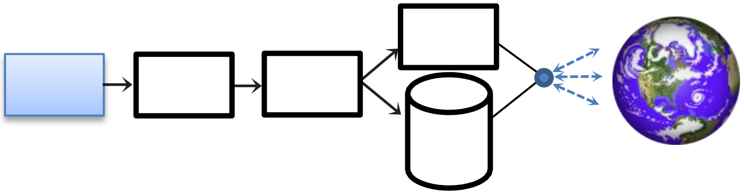
Assessing Digitization Tasks



- Reed Beaman, James Macklin, Michael Donoghue, James Hanken. 2007. Overcoming the Digitization Bottleneck in Natural History Collections: A summary report on a workshop held 7 – 9 September 2006 at Harvard University.
- Íñigo Granzow-de la Cerda and James H. Beach. December 2010. Semi-automated workflows for acquiring specimen data from label images in herbarium collections. Taxon 59 (6): 1830-1842
- Bryan Kalms. Digitisation: A strategic approach for natural history collections. Canberra, Australia, CSIRO, 2012.
- John Tann & Paul Flemons. 2008. Report: Data capture of specimen labels using volunteers. Australian Museum
- Ana Vollmar, James Alexander Macklin, Linda Ford. 2010. Natural History Specimen Digitization: Challenges and Concerns. Biodiversity Informatics 7 (1): 93 – 112
- Favret C, Cummings KS, McGinley RJ, Heske EJ, Johnson KP, Phillips CA, Phillippe LR, Retzer ME, Taylor CA, Wetzel MJ. 2007. Profiling Natural History Collections: A Method for Quantitative and Comparative Health Assessment. Collection Forum 22(1–2): 53 - 65
- Nelson G, Paul D, Riccardi G, Mast AR 2012. Five task clusters that enable efficient and effective digitization of biological collections. In: Blagoderov V, Smith VS (Ed) No specimen left behind: mass digitization of natural history collections. ZooKeys 209: 19–45. doi: 10.3897/zookeys.209.3135
- **iDigBio Developing Robust Object to Image to Data (iDigBio DROID) Workshop** – May 30 – 31, 2012

Thank You from iDigBio

Integrated Digitized Biocollections



American Museum of Natural History (AMNH)
Botanical Research Institute of Texas (BRIT)
Florida Museum of Natural History (FLMNH)
Florida State University (FSU)
Harvard Herbarium (HUH)
Museum of Comparative Zoology (Harvard)
New York Botanical Garden (NYBG)
Yale Peabody Museum (YPM)
Southeast Regional Network of Expertise and Collections (SERNEC)
Specify Software Project (University of Kansas)
Symbiota Software Project (Arizona State University)
Tall Timbers Research Station and Land Conservancy (TTRS)
Tulane University Museum of Natural History
University of Kansas Biodiversity Institute Entomology Department
Valdosta State University (VSU)

and ***all participants at the iDigBio* Digitizing Vascular and Non-vascular Plant Collections Workshop** hosted by Valdosta State University, September 17 – 18, 2012

iDigBio Integrated Digitized Biocollections NSF UNIVERSITY OF FLORIDA FLORIDA STATE UNIVERSITY 1851 FLORIDA MUSEUM OF NATURAL HISTORY