Data at iDigBio

Feeding the Research Data Pipeline:
long live the data

Deborah Paul, iDigBio Digitization and Training Specialist, @idbdeb
Missouri Botanical Garden, December 2 – 3, 2015

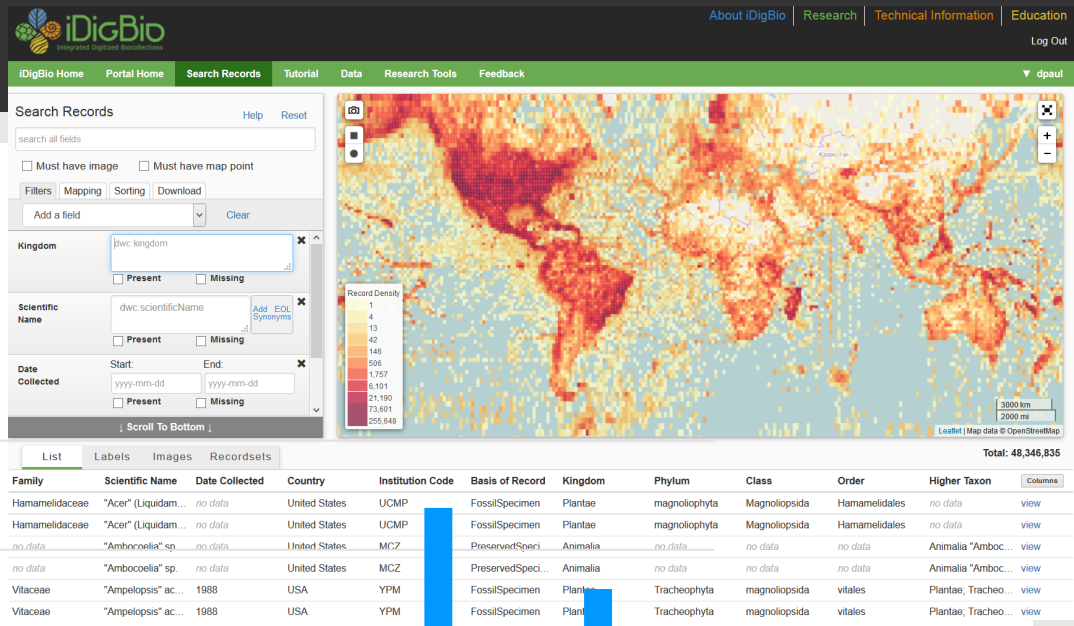
Content contributors: iDigBio staff - Alex Thompson, Joanna McCaffrey, Kevin Love, Matt Collins, Shelley James, Gil Nelson

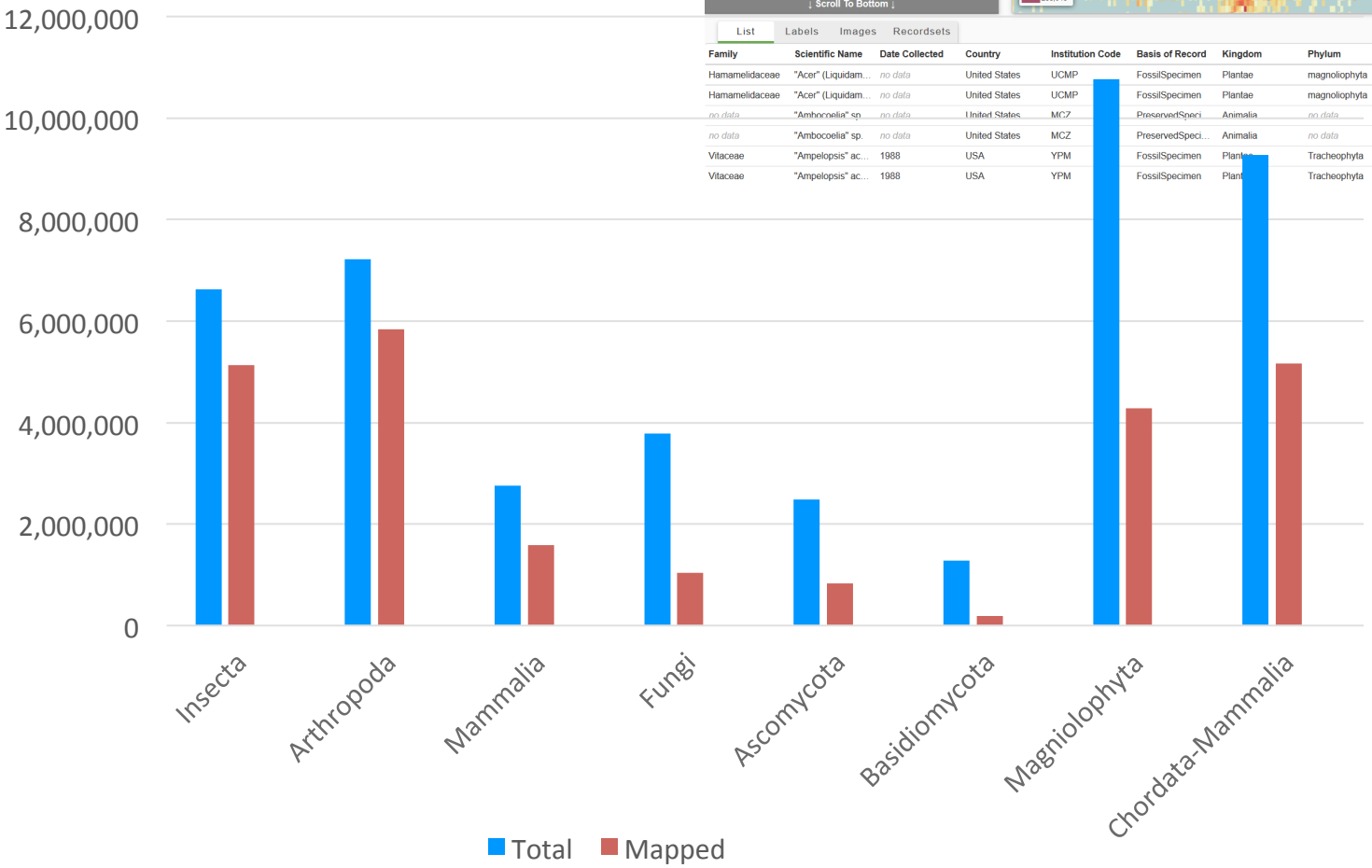
iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

https://www.idigbio.org
@iDigBio

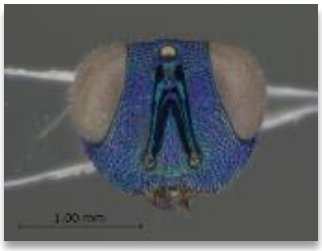
#valueOfAggregation

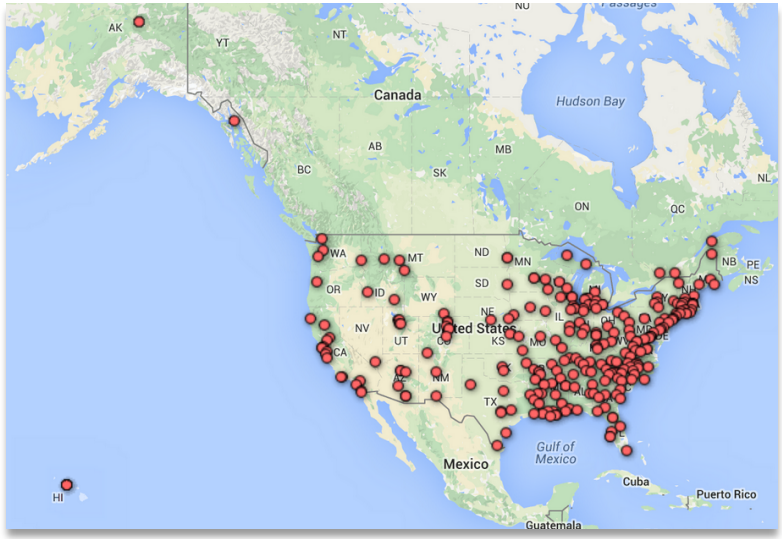
Specimen Records Overview

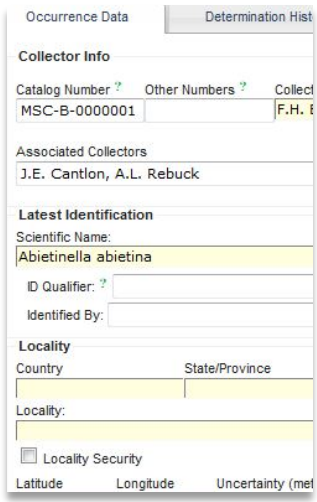




15 Thematic Collection Networks (TCNs)





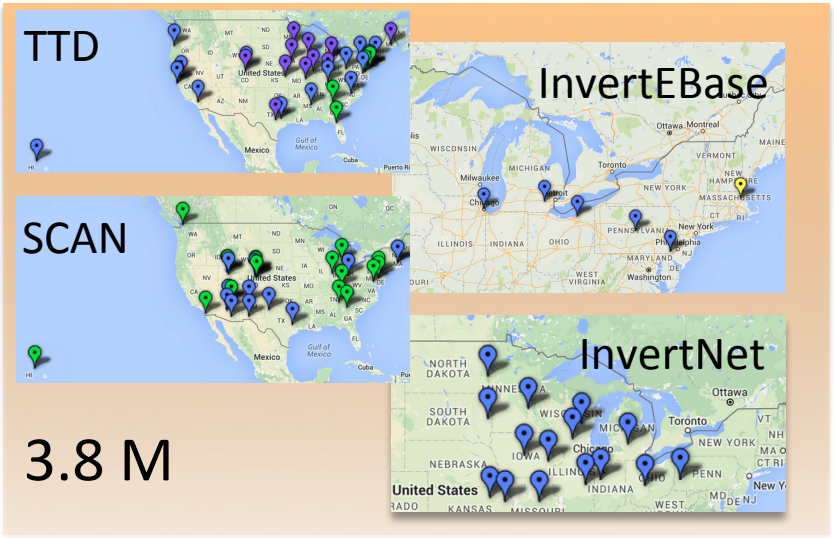


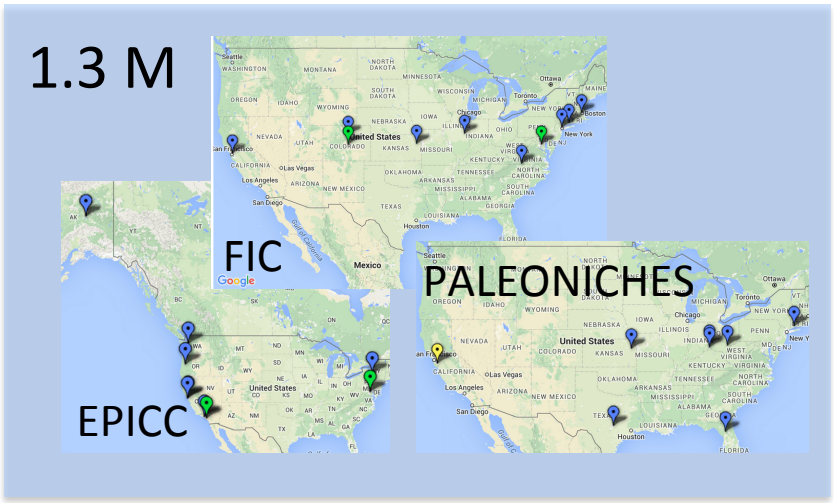
15 Thematic Collections Networks (TCNs), 15 PENs

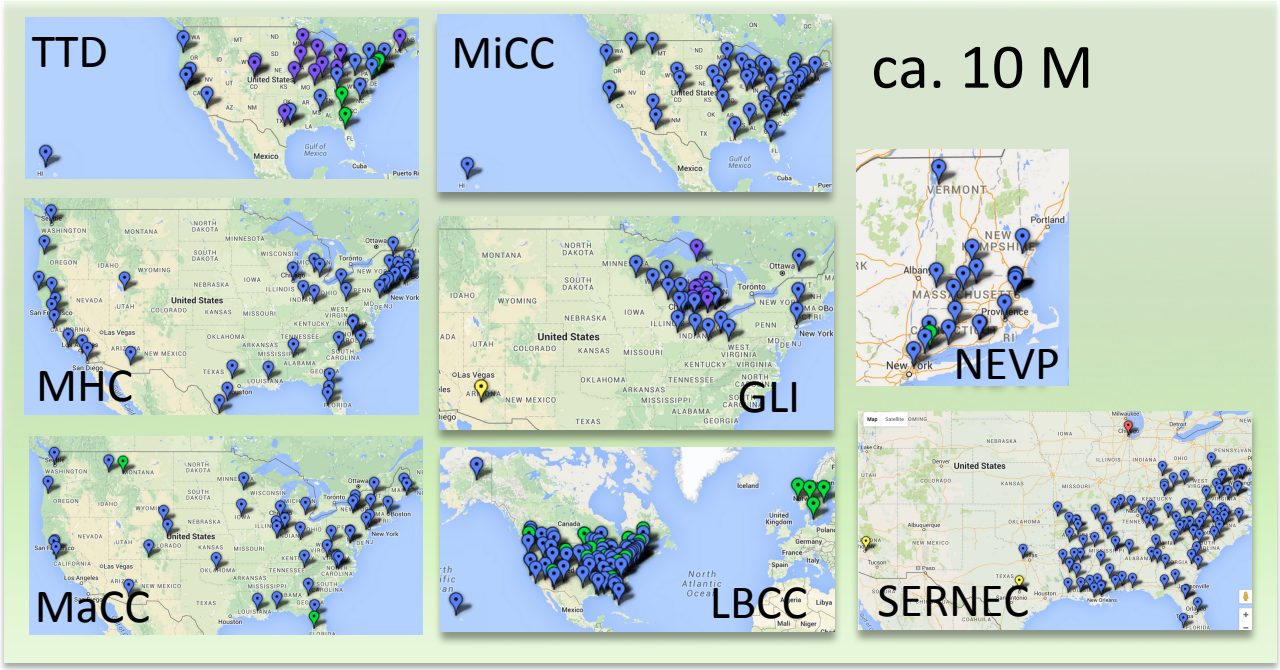
- InvertNet: An Integrative Platform for Research on Environmental Change, Species Discovery and Identification (Illinois Natural History Survey, University of Illinois) http://invertnet.org
- Plants, Herbivores, and Parasitoids: A Model System for the Study of Tri-Trophic Associations (American Museum of Natural History) http://tcn.amnh.org (TTD)
- North American Lichens and Bryophytes: Sensitive Indicators of Environmental Quality and Change (University of Wisconsin – Madison) http://symbiota.org/nalichens/index.php http://symbiota.org/bryophytes/index.php (LBCC)
- Digitizing Fossils to Enable New Syntheses in Biogeography - Creating a PALEONICHES-TCN (University of Kansas)
- The Macrofungi Collection Consortium: Unlocking a Biodiversity Resource for Understanding Biotic Interactions, Nutrient Cycling and Human Affairs (New York Botanical Garden)(MaCC)
- Mobilizing New England Vascular Plant Specimen Data to Track Environmental Change (Yale University) (NEVP)
- Southwest Collections of Arthropods Network (SCAN): A Model for Collections Digitization to Promote Taxonomic and Ecological Research (Northern Arizona University) http://hasbrouck.asu.edu/symbiota/portal/index.php
- iDigPaleo: Fossil Insect Collaborative: A Deep-Time Approach to Studying Diversification and Response to Environmental Change (FIC)
- Developing a Centralized Digital Archive of Vouchered Animal Communication Signals (VACS)
- The Macroalgal Herbarium Consortium: Accessing 150 Years of Specimen Data to Understand Changes in the Marine/ Aquatic Environment (MHC)
- Documenting the Occurrence through Space and Time of Aquatic Non-indigenous Fish, Mollusks, Algae, and Plants Threatening North America's Great Lakes (GLI)
- InvertEBase: Reaching Back to See the Future: Species-rich Invertebrate Faunas Document Causes and Consequences of Biodiversity Shifts
- The Key to the Cabinets: Building and Sustaining a Research Database for a Global Biodiversity Hotspot (**SERNEC**)
- The Microfungi Collections Consortium: A Networked Approach to Digitizing Small Fungi with Large Impacts on the Function and Health of Ecosystems (MiCC)
- Documenting Fossil Marine Invertebrate Communities of the Eastern Pacific - Faunal Responses to Environmental Change over the last 66 million years (EPICC)

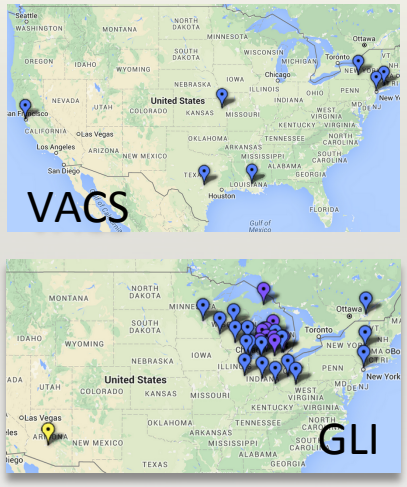
#valueOfAggregation

Present Data from some of the 15 Thematic Collection Networks (TCNs)





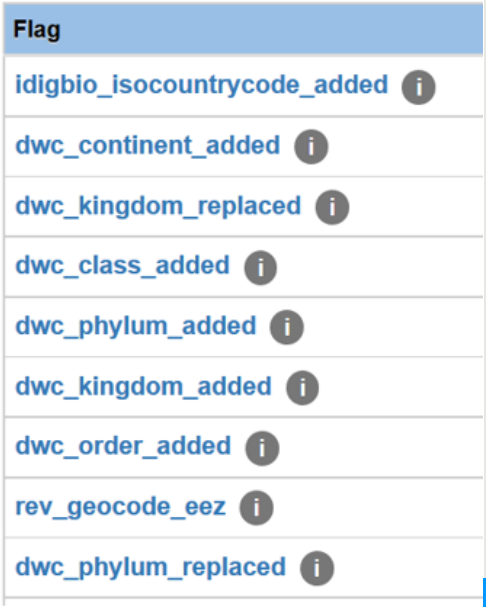




#valueOfAggregation

Kinds of data issues to think about

- 1100 unique strings in dwc:country field
- 1.7 M unique strings in dwc:scientificName
  - Using Apache Spark
- 6.7 M unique locality strings
  - The ones you need?
  - Georeferenced already?
  - With errors?
  - With metadata?



iDigBio
Integrated Digitized Biocollections

KINGDOMS

Taxa within GBIF backbone kingdoms.

- Animalia 1,913,825
- Archaea 463
- Bacteria 14,258
- Chromista 18,186
- Fungi 319,366
- Plantae 953,712
- Protozoa 41,539
- Viruses 5,724
- Other 6,248

RANKS

Number of accepted taxa by ranks.

- kingdom 9
- phylum 111
- class 272
- order 1,183
- superfamily 356
- family 20,760
- genus 390,790
- species 2,497,114
- infraspecific name 691
- subspecies 171,473
- infrasubspecific name 1
- variety 149,011
- subvariety 1,070
- form 40,473
- subform 7

Names

There are 1,143,026 synonyms in this dataset.

UNIQUE NAMES

There are 4,410,899 unique names in this dataset. On average 0.124% of the names are found in more than one taxon.

2,497,114	4,416,347
Species	Taxa

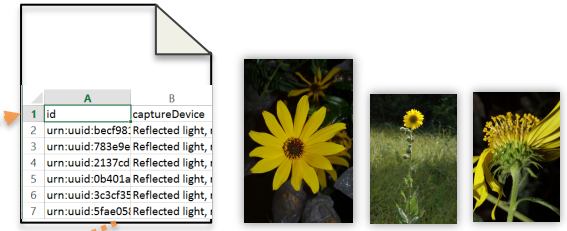
View species

When you *download data* from iDigBio, you get a bonus

What's Inside a DwC-A at iDigBio?

enhanced
sampleoccurrence.txt

samplemultimedia.txt
(extension)



describes

describes

describes

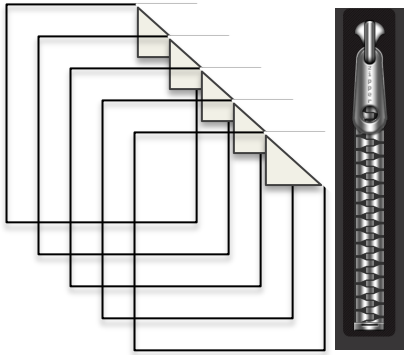
XML

meta.xml

raw sampleoccurrence.txt
(core)

extends

extends



eml.
XML

eml.xml
metadata about the dataset,
collection and contact information

coreid	identifiedBy
urn:uuid:1449	A. R. Diamond
urn:uuid:1449	I.M. Notsura
urn:uuid:982e	J.D. Freeman
urn:uuid:982e	Mia Best Gess
urn:uuid:982e	Justin Ideeah
urn:uuid:0b45	Hannelore Run

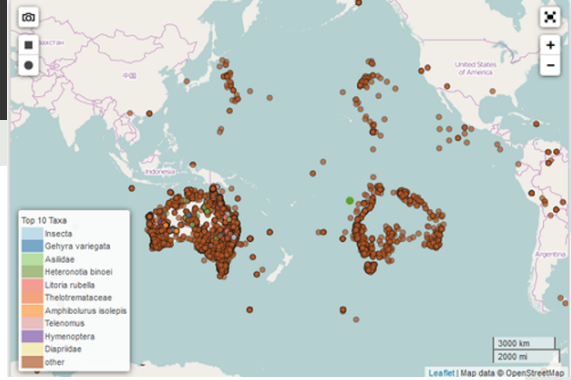
Determination 1

Determination 2

Determination n

sampledeterminations.txt
(extension)

#valueOfAggregation



More data → data quality flags

Data Corrected | Data Use | Raw

This table shows any data corrections that were performed on this recordset to improve the capabilities of iDigBio Search. The first column represents the correction performed. The last two columns represent the number and percentage of records that were corrected. A complete list of the data quality flags and their descriptions can be found here. Clicking on a data flag name will take you to a search for all records with this flag in this recordset.

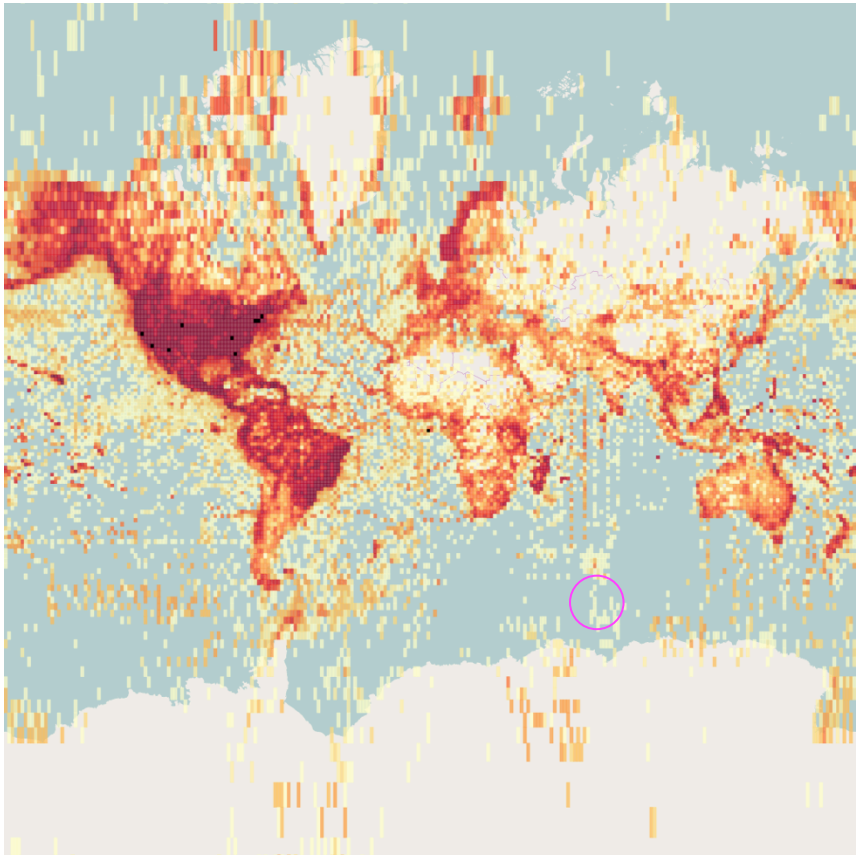
Flag	Records With This Flag	(%) Percent With This Flag
idigbio_isocountrycode_added	2370	98.709
dwc_continent_added	2369	98.667
dwc_kingdom_replaced	784	32.653
dwc_class_added	370	15.41
dwc_phylum_added	370	15.41
dwc_kingdom_added	365	15.202
dwc_order_added	360	14.994
rev_geocode_eez	339	14.119
dwc_phylum_replaced	337	14.036

Downstream Use Cases

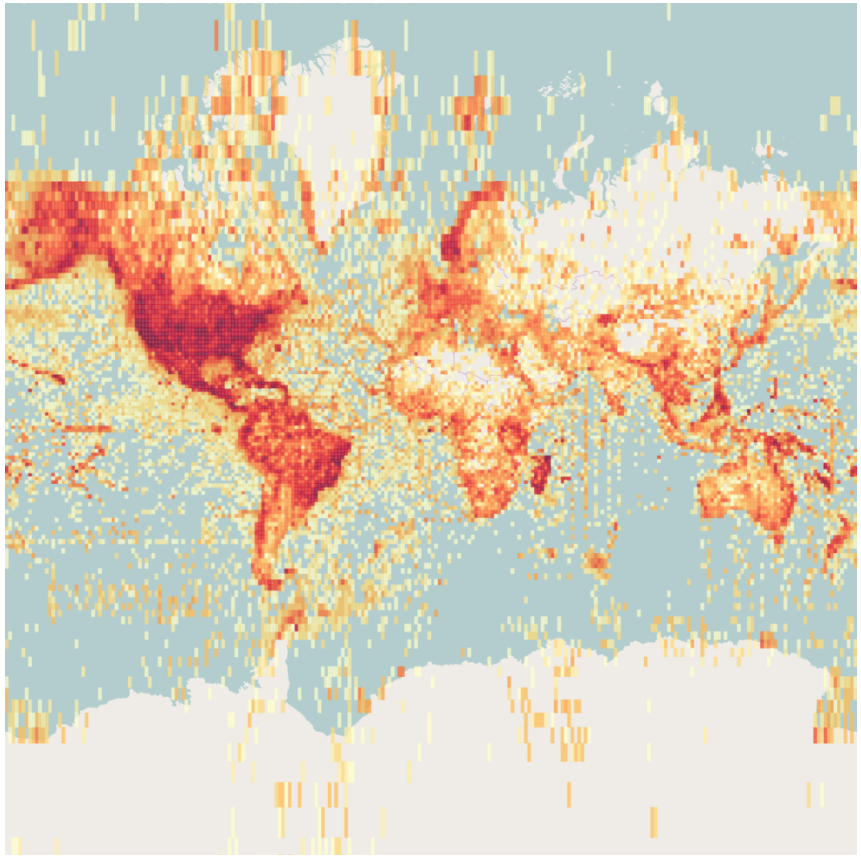
- Environmental Ontology Checking - Pier Buttigieg & Grant Godden https://www.idigbio.org/content/webinar-shaping-semantic-layer-mining-digitised-data-encounter-between-idigbios-plant
- Data Cleaning - Heather Appleby & Katja Seltmann https://www.idigbio.org/content/summer-learning-r-clean-data-idigbio-portal-recordset-correction-feature
- Lifemapper - KU & ACIS Collaboration http://lifemapper.acis.ufl.edu
- PhyloJive - Joe Miller & ACIS Collaboration http://phylojive.acis.ufl.edu
- Proximity and Correlation: Two new computer programs for mining phytosociological information held in herbarium databases using central Arizona as a test case. Daryl Lafferty & Les Landrum, Arizona State University http://dx.doi.org/10.12705/645.9 Taxon, Volume 64, Number 5, 28 October 2015, pp. 998-1016(19) https://www.idigbio.org/wiki/index.php/Specimens_Full_Circle_SPNHC_2015

Advanced Mapping API Capabilities

"Species Density" Map

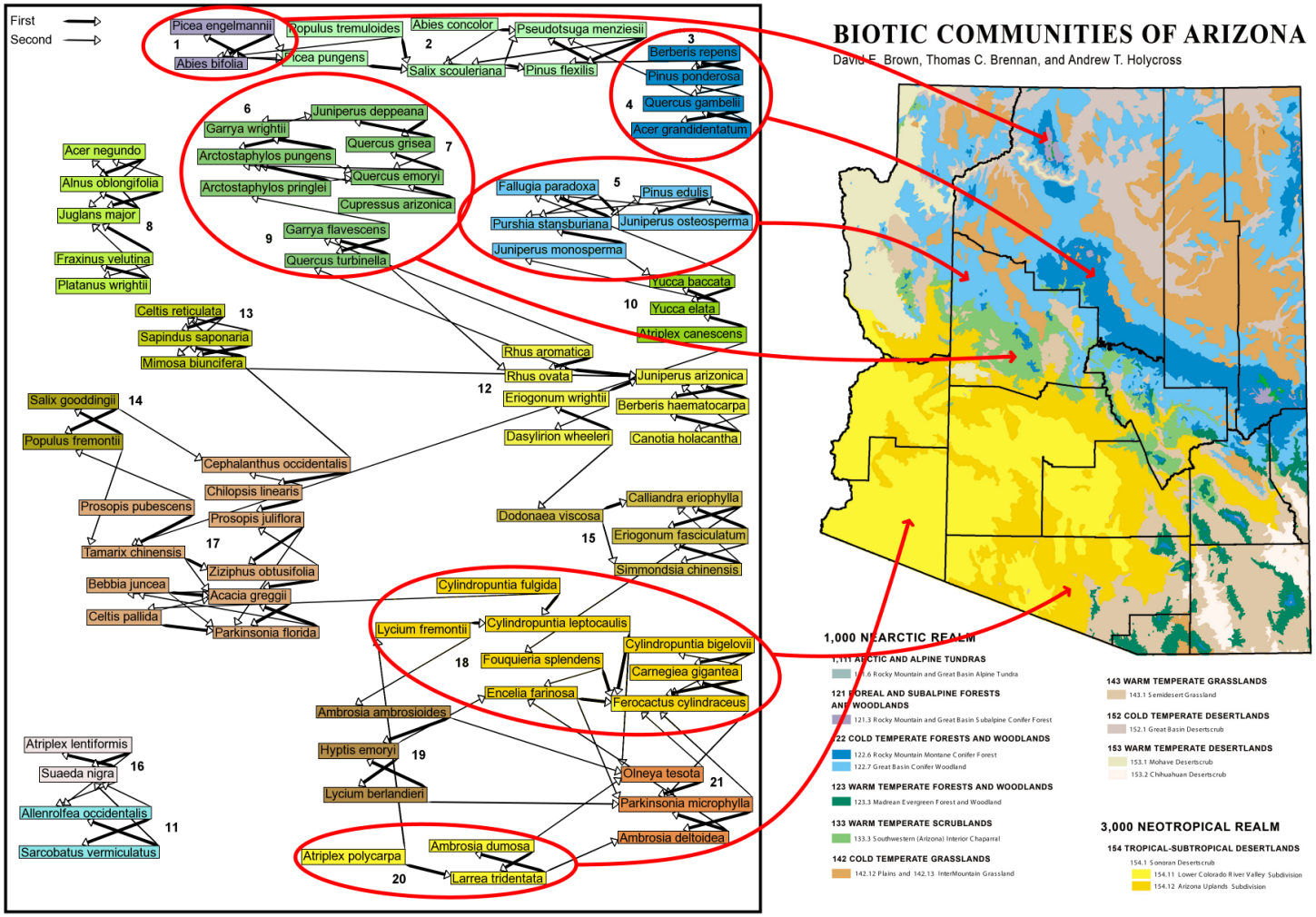


Specimen Density Map



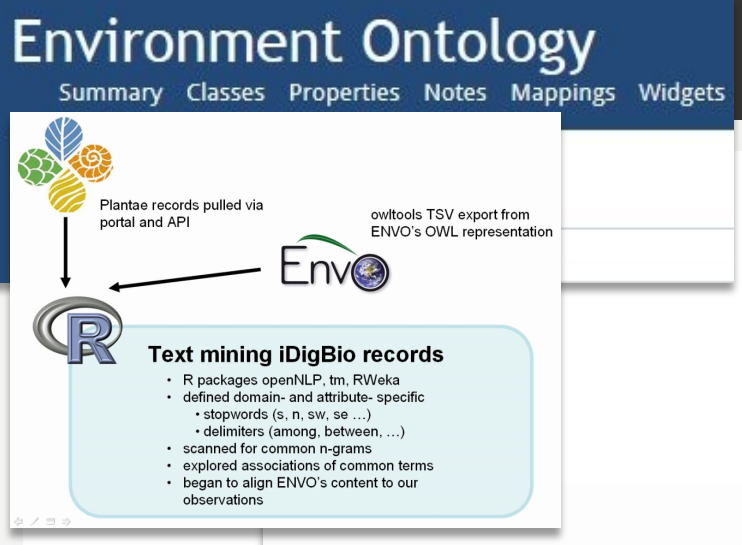
#valueOfAggregation

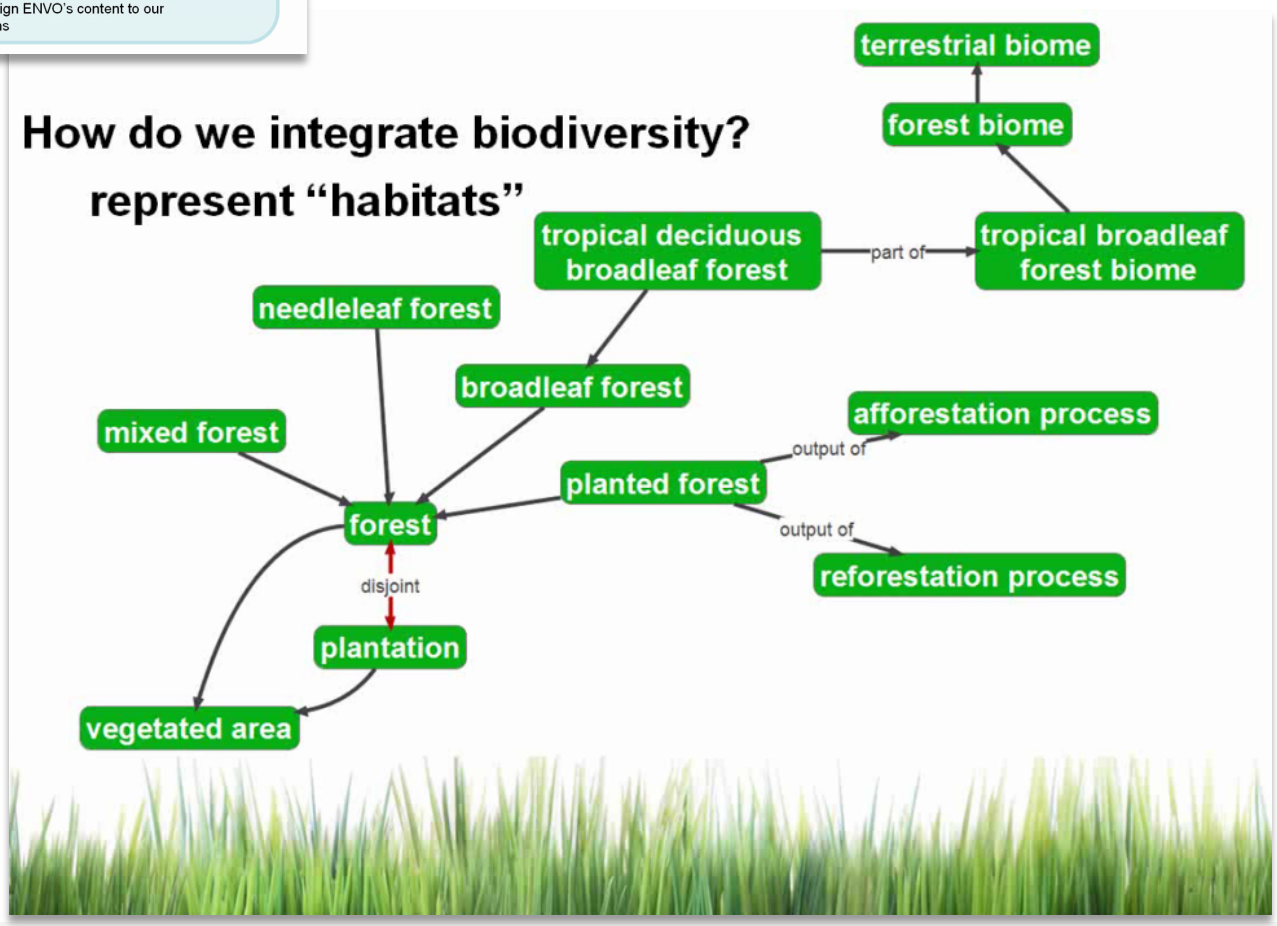
Mining and Using Associated Taxa Data



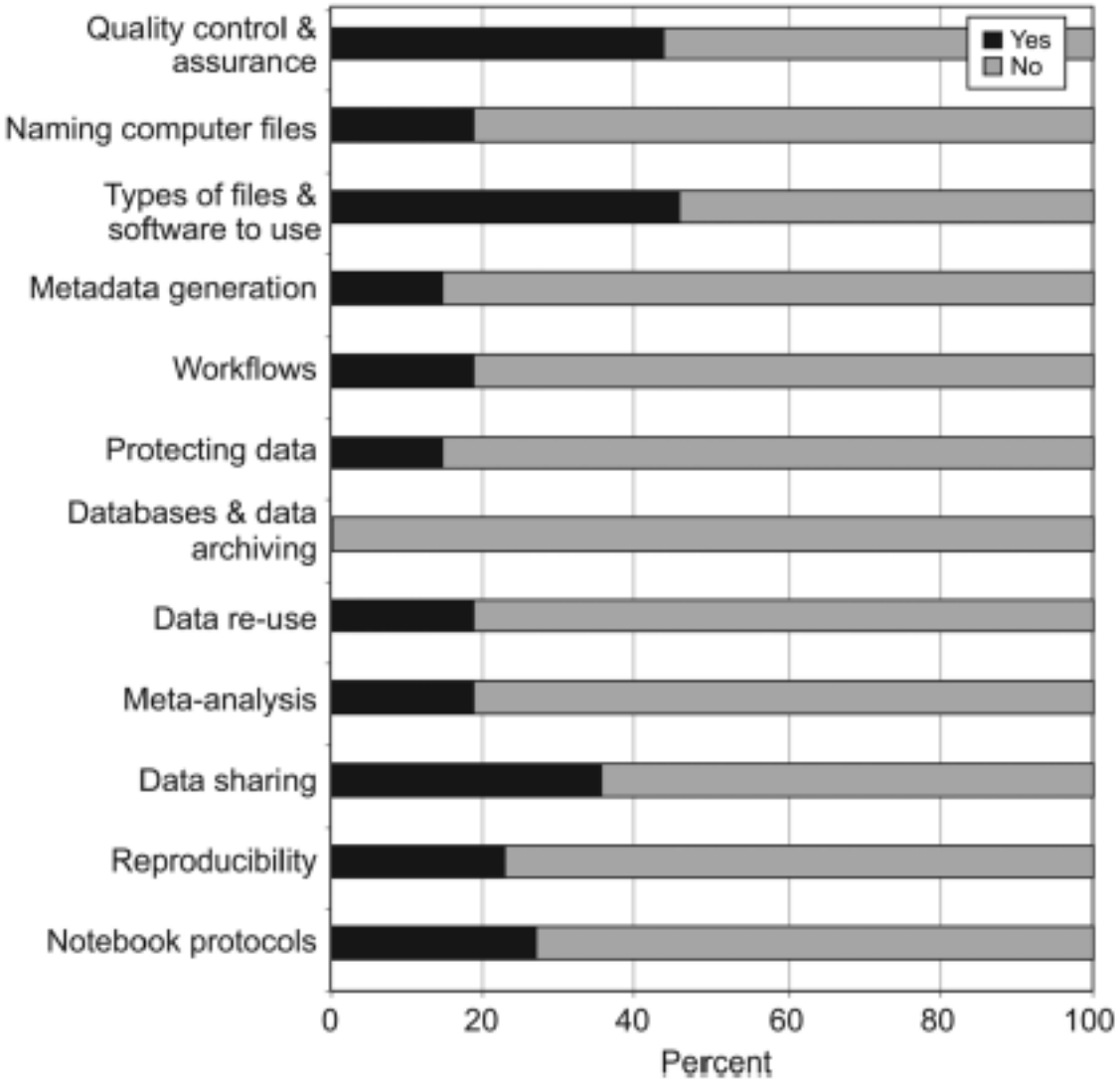
#valueOfAggregation

Mining Aggregated Data to Enhance Ontologies In support of future research





#valueOfAggregation



% Ecology courses that address / teach the listed data management principles

The path ahead

- refine our corpora and TM approaches to deliver robust and more directly actionable output
  - account for synonymy and, if feasible, multilingual issues
- engage experts on specific taxa to evaluate the representation of environments relevant to their work
  - create taxon-specific mining routines and link to efforts in the semantic representation of habitats
- Perform gap analyses using ENVO as a reference to find undersampled environment types
- Pre-compose classes in ENVO and related ontologies whicl align to the most frequent (or important) environment descriptions in the available corpora

3-gram	Frequency
montane rain forest	41
evergreen broadleaf forest	38
subtropical evergreen broadleaf	36
substrate loam disturbance	34
baserock granite substrate	22
granite substrate loam	20
disturbance agriculture slope	15
loam disturbance agriculture	15

Providing semantic feedback to iDigBio?

engaging botanists (and other biologists), text-miners, ontologists, a database engineers is the key to creating semantically-aware, approachable, and sustainable infrastructure

digitiser training is

Looking forward to more data and enhanced data being used in research! Up next, the TCNs…

www.idigbio.org

facebook.com/iDigBio

twitter.com/iDigBio

vimeo.com/idigbio

idigbio.org/rss-feed.xml

webcal://www.idigbio.org/events-calendar/export.ics

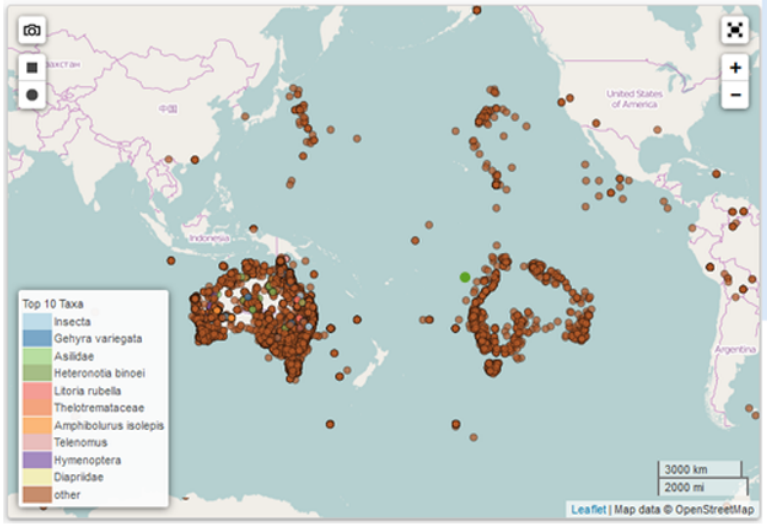
iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

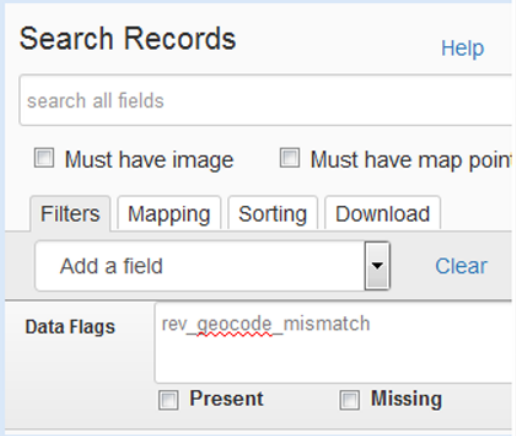
http://tinyurl.com/GWGmikeyostGPSapps

iDigBio Data Quality (DQ) Flags enhance Georeferencing Workflows!

iDigBio DQ flags making it easier to
spot and fix georeferencing issues.
Great for researchers, great for data providers

See https://www.idigbio.org/portal/publishers





Flag
dwc_class_added
dwc_kingdom_added
dwc_order_added
dwc_phylum_added
idigbio_isocountrycode_added
dwc_continent_added
geopoint_datum_missing
rev_geocode_eez
dwc_stateprovince_replaced
dwc_country_replaced
rev_geocode_mismatch
rev_geocode_lon_sign
rev_geocode_lat_sign
geopoint_similar_coord
rev_geocode_flip_lat_sign
datecollected_bounds

A Locality Service?

What is it? It can speed up georeferencing!
Ask us about it. Let's reduce re-georeferencing.

#valueOfAggregation

- Data by organismal group
  - Unique data
    - Measurement or fact
      - Host associations
    - Associated source materials, like recordings
  - Locality data for mining
- Mining aggregated data
  - Enriching ontologies to support research
    - ENVO
    - Other ontologies that could do this
- The data sets.
- Ecologists want to know
- Using data to test models
  - Historic (back-casting)
- Clarifying fit-for-useness
  - How to indicate what type of research a dataset is good for
    - Georef to centroid?
    - Fine-grain georef?
    - Tdwg
- Aggregators, Data Generators, Researchers – friends with benefits (#valueOfAggregation)
- Unique data
  - Habitat
  - Localities

What we hear

- *I've borrowed my colleague's computer.*
- Scaling Up?: I'm running analysis on three different computers.
- Excel is a database, isn't it?
- Should I learn R (Python,…)? Is it worth my time?
- How do I visualize this data?
- What's the best way to share data with colleagues?
- How do I work with a txt (csv, hdf5,…) file?
- It's difficult to replicate my (others) research.
- What is an API?
- How do I use APIs to enhance my research, …

What do we hear after the workshops?

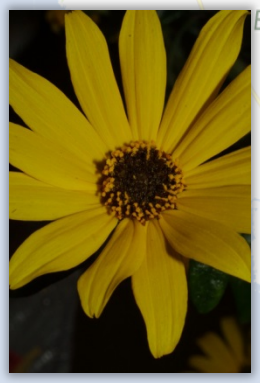
- “the course is well suited for beginners; I received the courage to venture and learn more about R”
- “I feel I can start using R again”
- “I appreciate all the hands-on work”
- “the class is well organized and easy to follow”
- “really great working through all the examples/challenges, lots of examples = lots of practice = understanding”
- wish I’d known this before graduate school!
- we need to hire a data manager.
- I can now give cleaner data to aggregators like iDigBio and GBIF

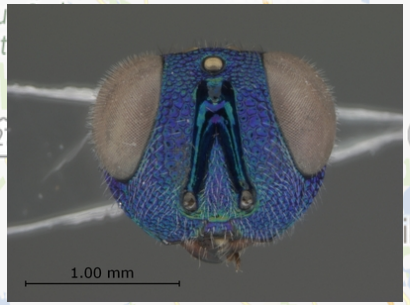
What are some of these biodiversity informatics skills and who needs them?

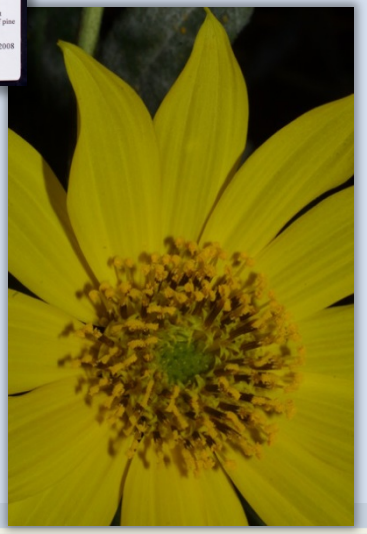
- For the researcher
- Collections Manager
- Data Manager
- Student

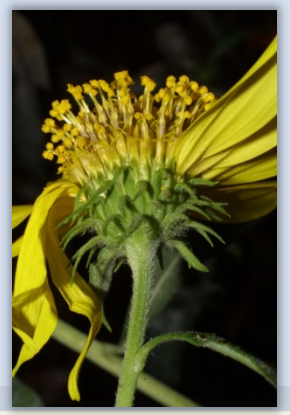
- Standards
- File management, organization, and literate programming
- Better Spreadsheet Skills!
- Workflows
  - Digitization
  - Research
- Data Analysis
- Data Visualization
- Data Mobilization


R. K. GODFREY HERBARIUM
202973
FLORIDA STATE UNIVERSITY
PLANTS OF FLORIDA, USA
Helianthus heterophyllus Nuttall

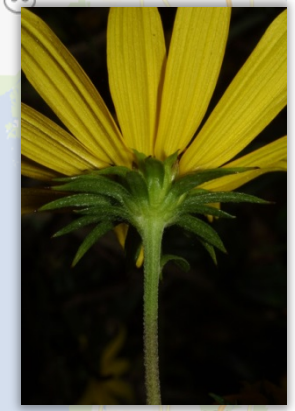



1.00 mm






°F
°C
120
100
80
60
40
20
0
20
40
60
50
40
30
20
10
0
10
20
30
40
50
60



Data and Metadata.
It's about discovery and data re/use.
It's about feedback and accountability.
It's about credit and attribution.
Make sure your data's not under a rock.