Biological collections data:

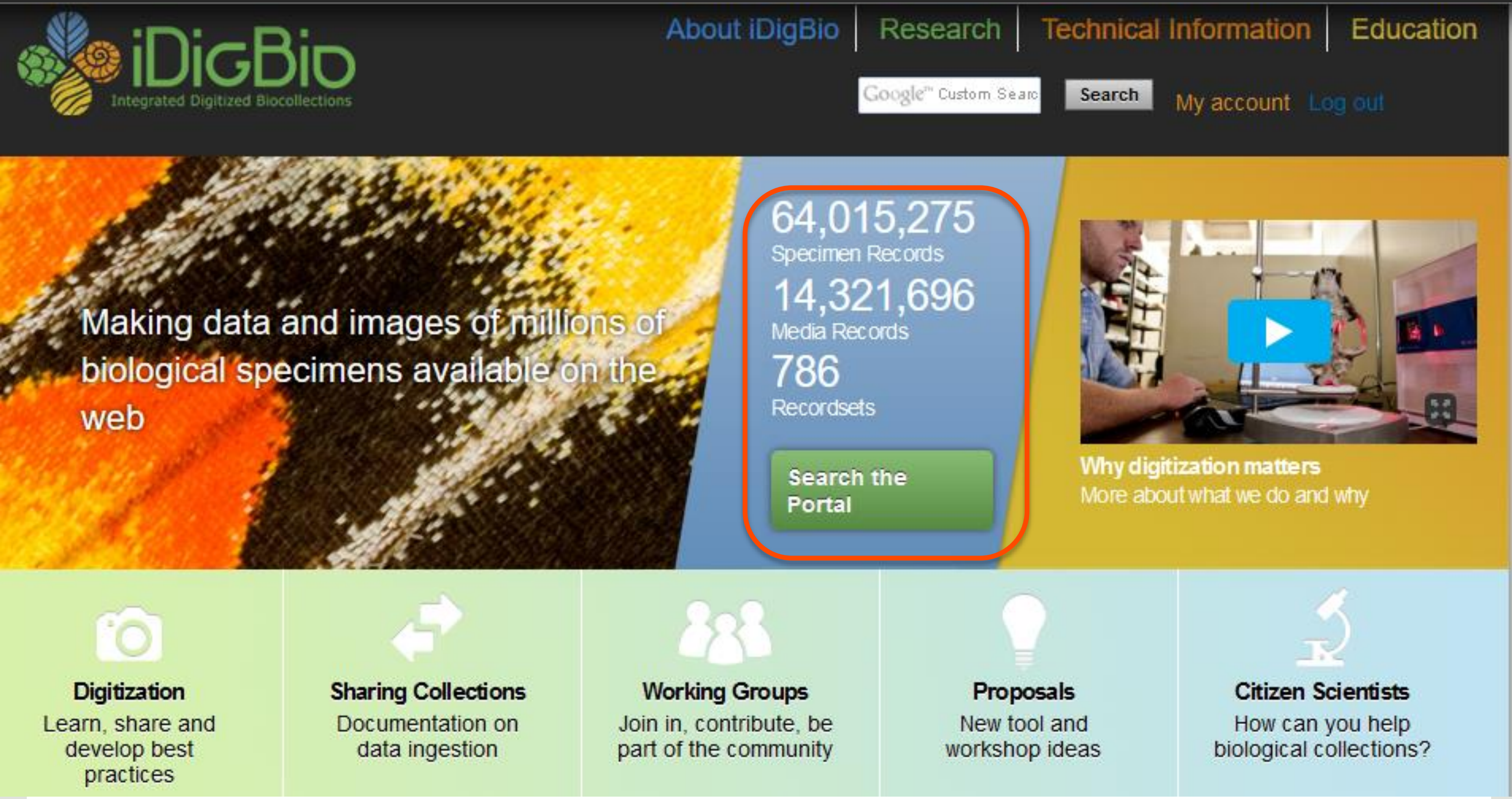
Best practices and trends for standards, digitization, and biodiversity informatics literacy for research use of collections data

Deborah Paul, Florida State University, iDigBio
Katja Seltmann, Cheadle Center for Biodiversity and Ecological Restoration (CCBER) #IslandBiology2016 University of the Azores at Angra do Heroísmo, Terceira Island, Azores, Portugal 19 July 2015
twitter @idbdeb @irene_moon

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.

An overview

- The need for high-quality data
- Digitizing collections and georeferencing
- Identifiers required
- Data sharing standards
- Researchers supplying, using data from aggregators
- Data gaps (Shelley)
- Biodiversity Informatics skills and literacy



Making data and images of millions of biological specimens available on the web

64,015,275
Specimen Records

14,321,696
Media Records

786
Recordsets

Why digitization matters
More about what we do and why

Digitization
Learn, share and develop best practices

Sharing Collections
Documentation on data ingestion

Working Groups
Join in, contribute, be part of the community

Proposals
New tool and workshop ideas

Citizen Scientists
How can you help biological collections?

https://www.idigbio.org

@iDigBio

What Happens in an **Internet Minute?**

60
Blogger
NETFLIX
Botnet
WIKIPEDIA
Identity
Email
Mobile
Apps
amazon
PANDORA
LinkedIn
flickr
twitter
facebook
Google
YouTube

639,800 GB of global IP data transferred

20
New victims of identity theft

204 million
Emails sent

47,000
App downloads

$83,000
In sales

61,141
Hours of music

20 million
Photo views

3,000
Photo uploads

320+
New Twitter accounts

100,000
New tweets

135
Botnet infections

6
New Wikipedia articles published

1,300
New mobile users

100+
New Linkedin accounts

277,000
Logins

6 million
Facebook views

2+ million
Search queries

30
Hours of video uploaded

1.3 million
Video views

And **Future Growth** is **Staggering**

Today, the number of networked devices =

the global population

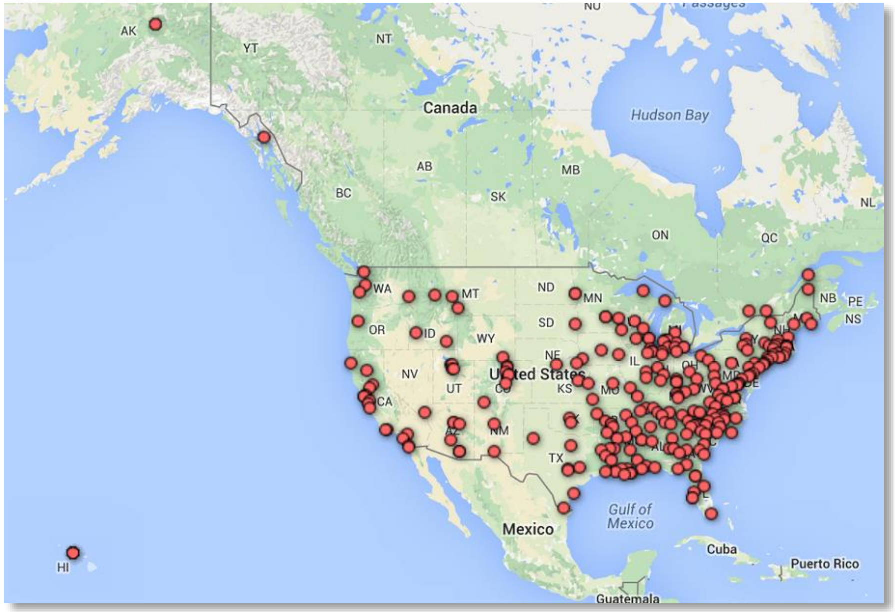
By **2015**, the number of networked devices = 2x

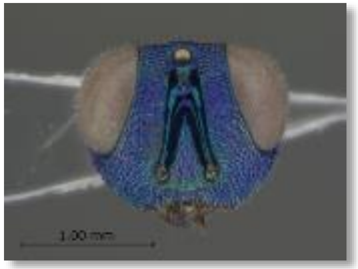
the global population

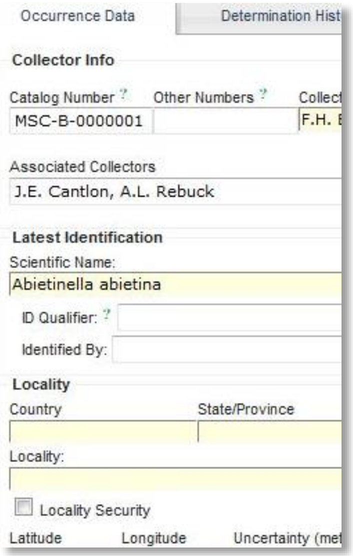
In **2015**, it would take you **5 years**

to view all video crossing IP networks each **second**

Where does the data in iDigBio come from?







Occurrence Data | Determination Hist

Collector Info

Catalog Number ?	Other Numbers ?	Collect
MSC-B-0000001		F.H. E

Associated Collectors
J.E. Cantlon, A.L. Rebuck

Latest Identification

Scientific Name:
Abietinella abietina

ID Qualifier: ?

Identified By:

Locality

Country | State/Province

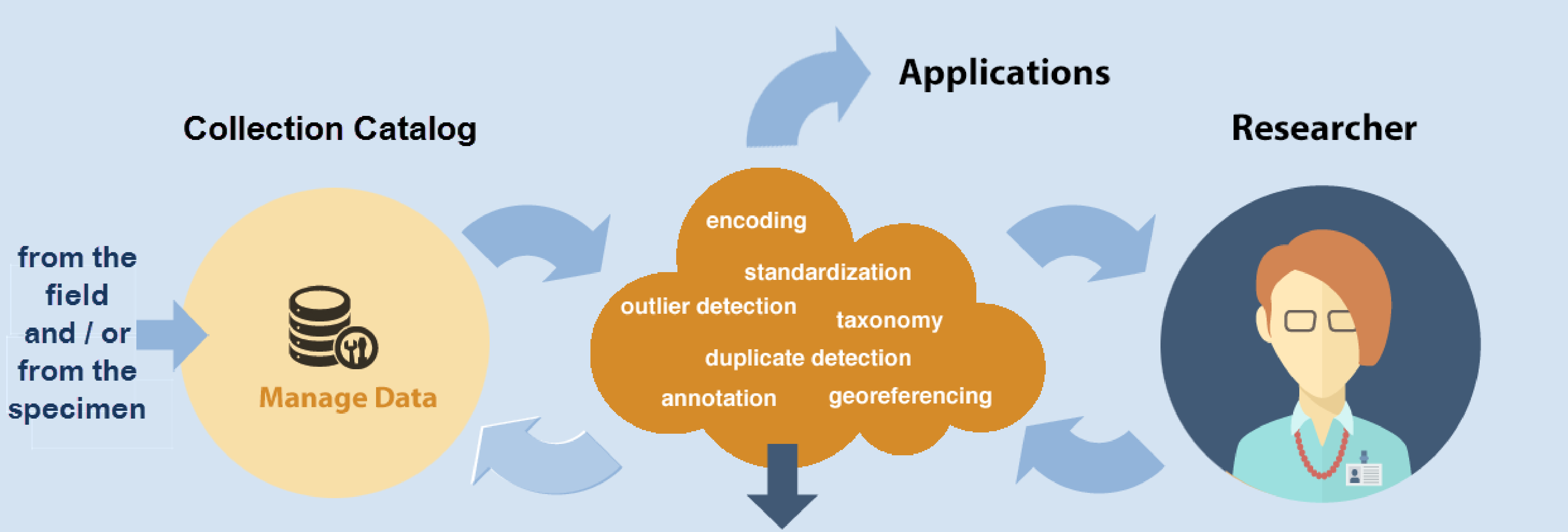
Locality:

□ Locality Security

Latitude | Longitude | Uncertainty (met

MARINE ALGAE OF El Salvador
HERBARIUM OF WM. RANDOLPH TAYLOR PRESENTED 1967 UNIVERSITY OF MICHIGAN
Padina crispata Thivy
Locality: Gulfe de Fonseca
Habitat: Playitas, near La Union, surfy, rocky intertidal area
Collected by E. Yale Dawson #21841 Sept. 6, 1950
det. by E. Y. D.

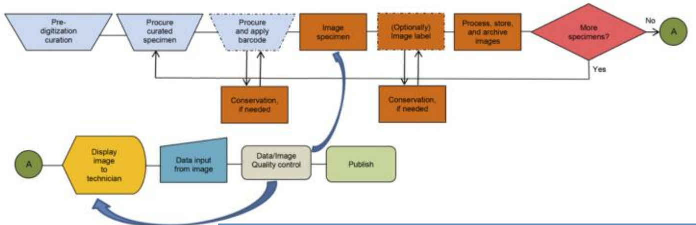
The Research Data Pipeline



Possibilities

species ranges, outlier discovery, new species, gaps in collecting, traits, relationships, predictive niche models, collector maps,...

Data > Knowledge > Application



Digitization & Imaging

- Workflows, Protocols, ZooKeys
  - Best practices
- Prioritization trend
  - Research "digitization on demand"
- Curation
  - Physical and digital collections
- Working groups, webinars, publications, …

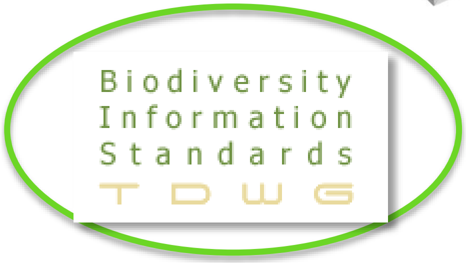
A peer-reviewed open-access journal
Launched to accelerate biodiversity research

- Fossil Insect Collaborative
- Fossil Marine Invertebrates (EPICC)
- Great Lakes Invasives
- InvertEBase
- InvertNet
- Lichens & Bryophytes
- Macroalgal Consortium
- Macrofungi Consortium
- Microfungi Consortium
- NEVP
- PALEONICHES
- SCAN
- SERNEC
- Tri-Trophic
- Vouchered Animal Communication Signals

Standards?

HTML

- All kinds of standards
- Data sharing standards



Biodiversity
Information
Standards
TDWG

What are some examples of standards used for sharing biodiversity data? Where do they come from?

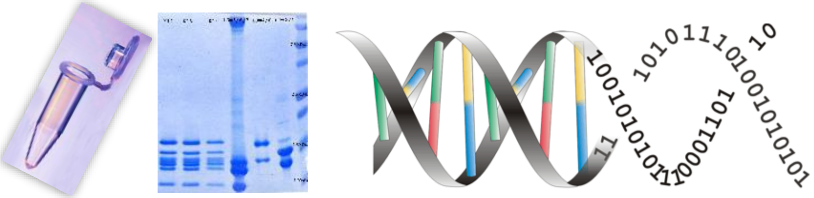
Data	Standards
specimens & observations	Darwin Core (DwC)
specimen & observation datasets	Ecological Metadata Language (EML)
media	Audubon Media Core
derivatives	Material Sample Core and GGBN Extensions, …

TDWG
Biodiversity Information Standards
Costa Rica 2016

creative commons

What's in the dataset?

1 of 3
2 of 3
3 of 3
Epistenia coeruleata, Westwood, 1832, 1991-05-14
Epistenia coeruleata, Westwood, 1832, 1991-05-14
Epistenia coeruleata, Westwood, 1832, 1991-05-14



Scientists want to share data

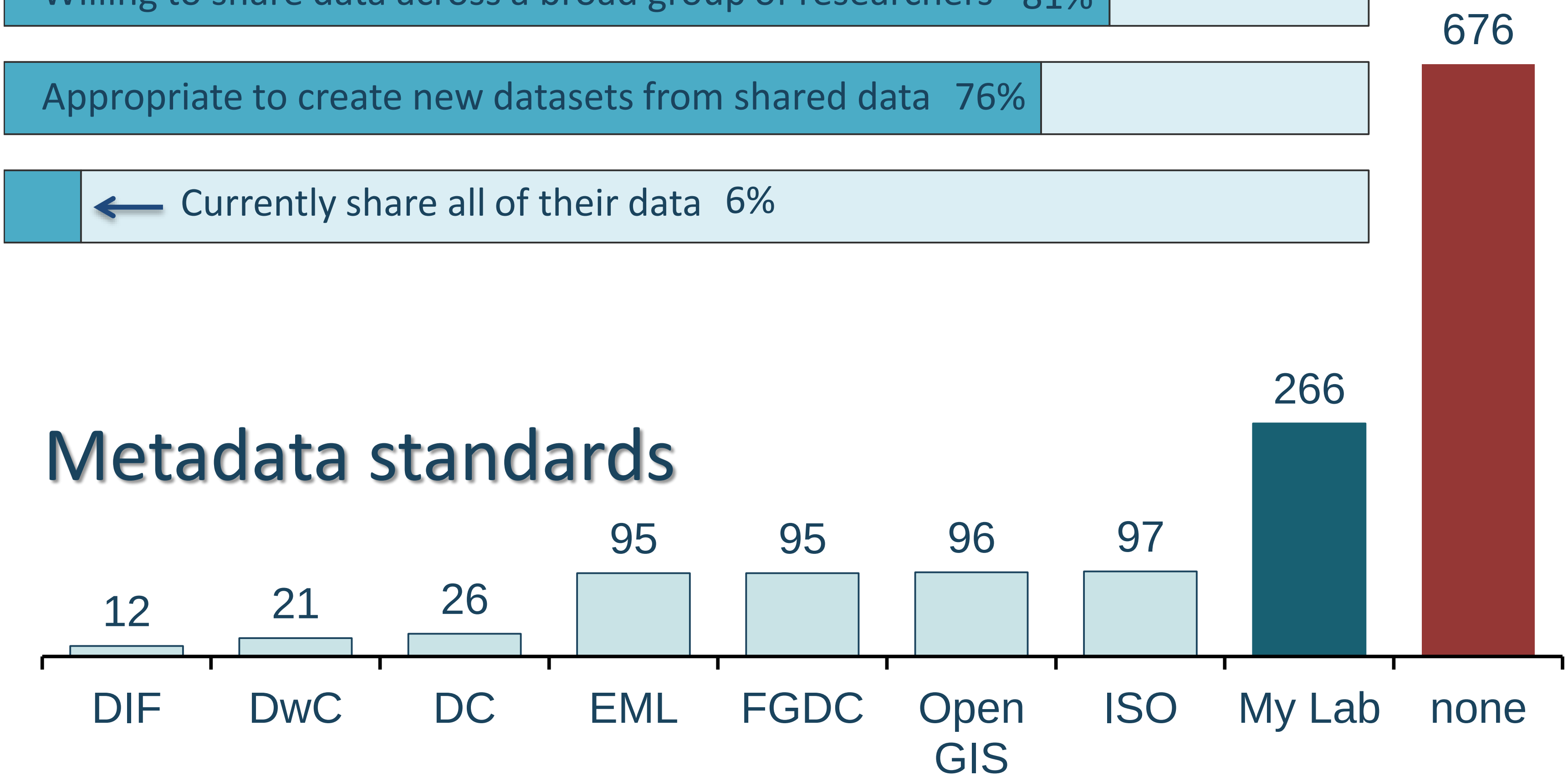
Use other researchers' datasets if easily accessible 84%

Willing to share data across a broad group of researchers 81%

Appropriate to create new datasets from shared data 76%

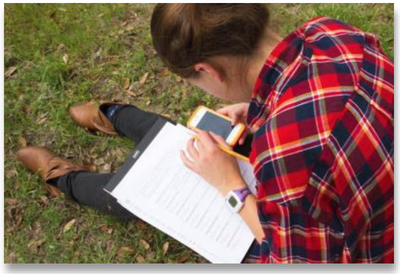
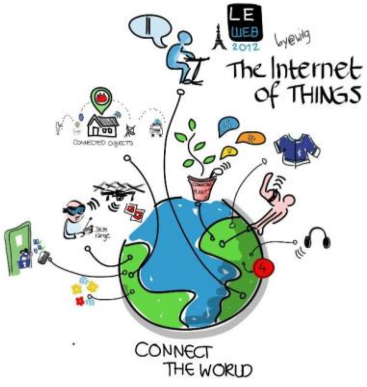
← Currently share all of their data 6%

Metadata standards



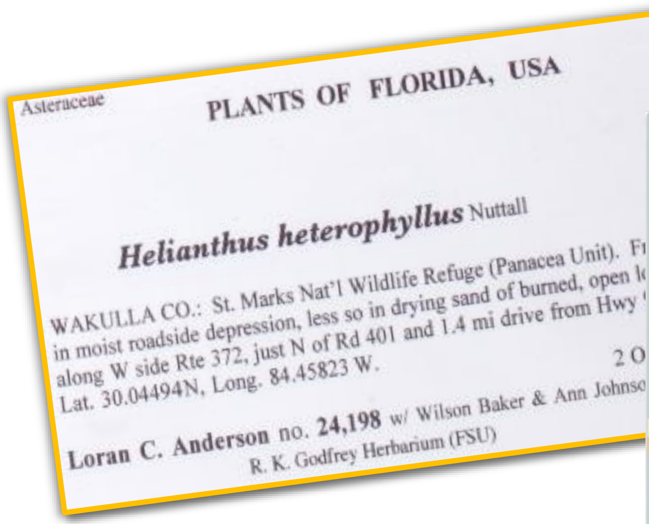
Standards – Why use them?

- Extend and expand data life
- Enhance sharing
- Facilitate re-use
- Increase likelihood of new uses
- Make linked-data initiatives possible
- Feedback, attribution

Georeferencing

- What’s your georeferencing workflow?
  - Legacy or New data
  - Use best practices
- Darwin Core (dwc)



Asteraceae

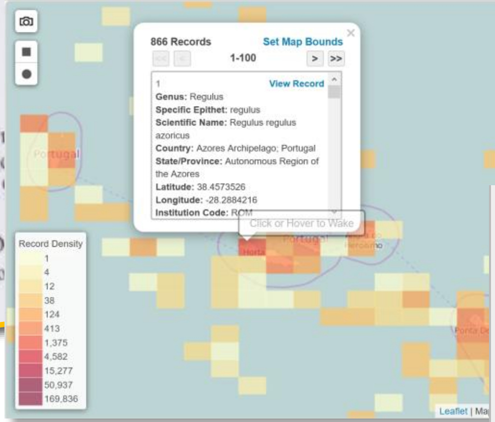
PLANTS OF FLORIDA, USA

Helianthus heterophyllus Nuttall

WAKULLA CO.: St. Marks Nat’l Wildlife Refuge (Panacea Unit). Fr
in moist roadside depression, less so in drying sand of burned, open lo
along W side Rte 372, just N of Rd 401 and 1.4 mi drive from Hwy
Lat. 30.04494N, Long. 84.45823 W.

2 O

Loran C. Anderson no. **24,198** w/ Wilson Baker & Ann Johnso
R. K. Godfrey Herbarium (FSU)



866 Records

Set Map Bounds

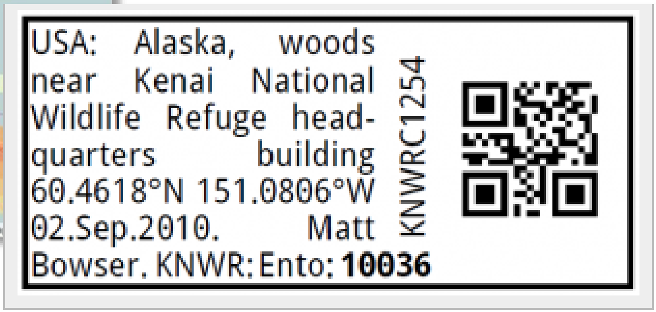
1-100

1 View Record

Genus: Regulus
Specific Epithet: regulus
Scientific Name: Regulus regulus azoricus
Country: Azores Archipelago; Portugal
State/Province: Autonomous Region of the Azores
Latitude: 38.4573526
Longitude: -28.2884216
Institution Code: ROM

Click or Hover to Wake

Record Density
1
4
12
38
124
413
1,375
4,582
15,277
50,937
169,836



USA: Alaska, woods near Kenai National Wildlife Refuge headquarters building 60.4618°N 151.0806°W 02.Sep.2010. Matt Bowser. KNWR:Ento:**10036**

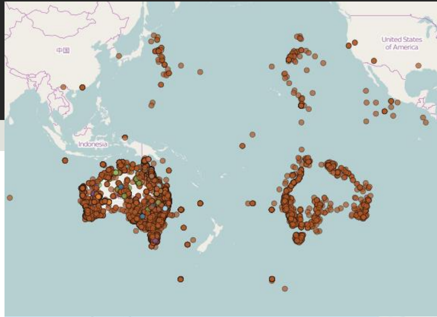
KNWRC1254

Darwin Core terms
Location and Geological Context

locationID | higherGeographyID | higherGeography | continent | waterBody | islandGroup | island | country | countryCode | stateProvince | county | municipality | locality | verbatimLocality | minimumElevationInMeters | maximumElevationInMeters |minimumDepthInMeters | maximumDepthInMeters | verbatimDepth | minimumDistanceAboveSurfaceInMeters | maximumDistanceAboveSurfaceInMeters | locationAccordingTo | locationRemarks | decimalLatitude | decimalLongitude | geodeticDatum | coordinateUncertaintyInMeters | coordinatePrecision | georeferencedBy | georeferencedDate | georeferenceProtocol | georeferenceSources | georeferenceVerificationStatus | georeferenceRemarks

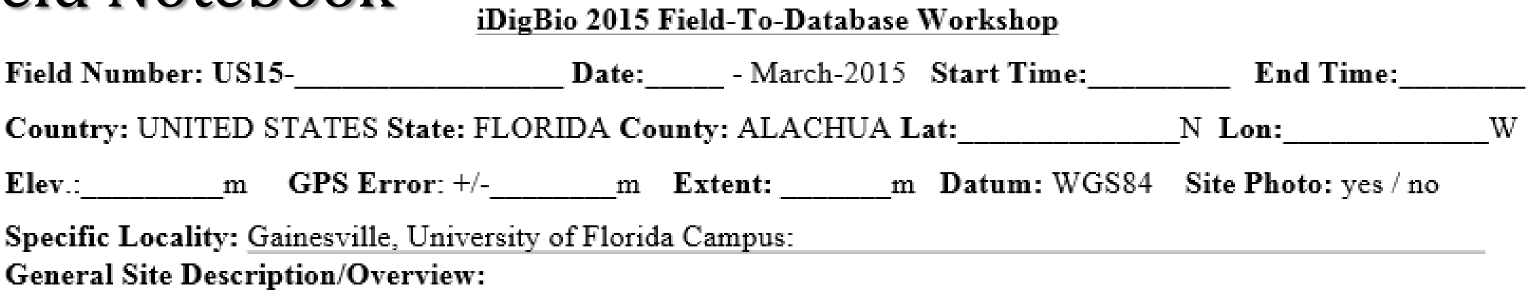
geologicalContextID | earliestEonOrLowestEonothem | latestEonOrHighestEonothem | earliestEraOrLowestErathem | latestEraOrHighestErathem | earliestPeriodOrLowestSystem | latestPeriodOrHighestSystem | earliestEpochOrLowestSeries | latestEpochOrHighestSeries | earliestAgeOrLowestStage | latestAgeOrHighestStage | lowestBiostratigraphicZone | highestBiostratigraphicZone | lithostratigraphicTerms | group | formation | member | bed

Georeferencing Tools, Materials, Workflows



- http://georeferencing.org/
- GEOLocate
- GWG at iDigBio
  - – Listserve, expertise
- Workshop materials, videos, powerpoints

- Good Localities Bad Localities: a Guide for your Field Notebook



iDigBio 2015 Field-To-Database Workshop

Field Number: US15-__________________ **Date:**_____ - March-2015 **Start Time:**__________ **End Time:**_________

Country: UNITED STATES **State:** FLORIDA **County:** ALACHUA **Lat:**_______________N **Lon:**______________W

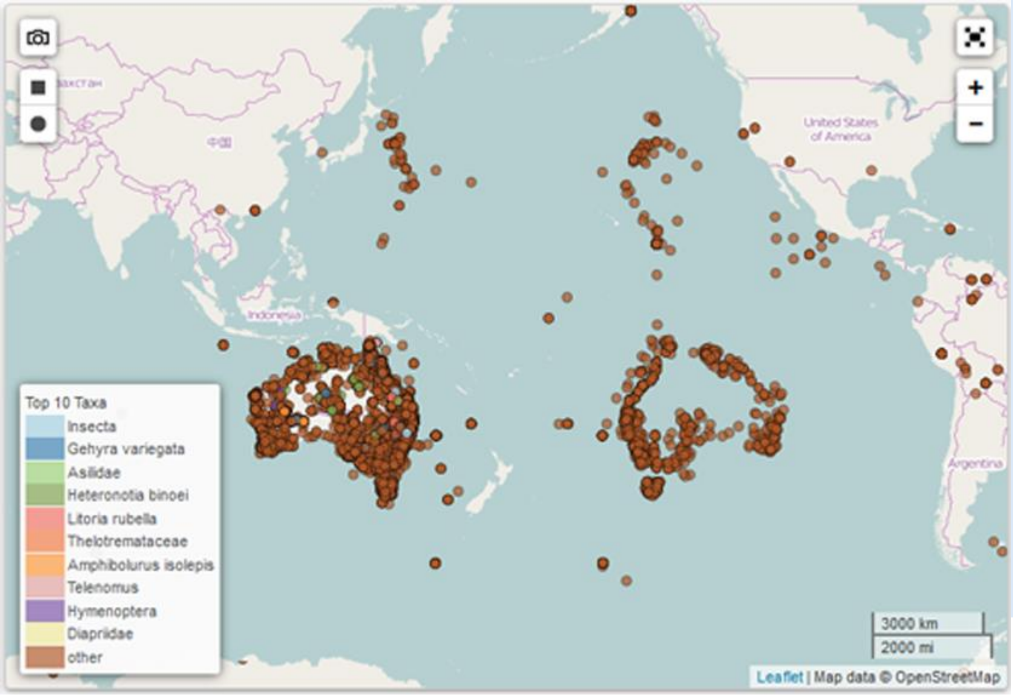
Elev.:__________m **GPS Error:** +/-_________m **Extent:** _______m **Datum:** WGS84 **Site Photo:** yes / no

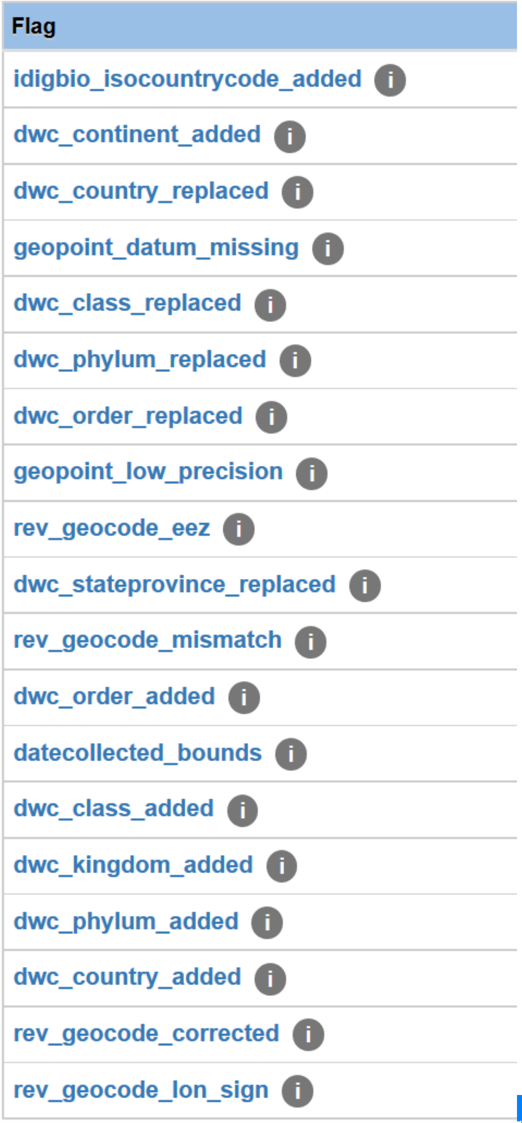
Specific Locality: Gainesville, University of Florida Campus:

General Site Description/Overview:

iDigBio Data Quality (DQ) Flags enhance Digitization and Research Workflows

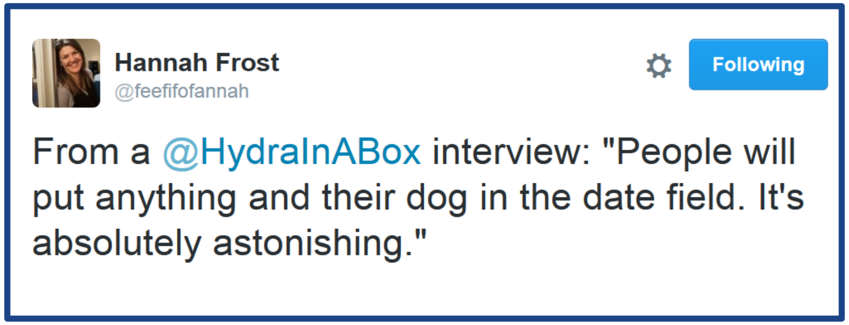
Example: spot and fix georeferencing issues.





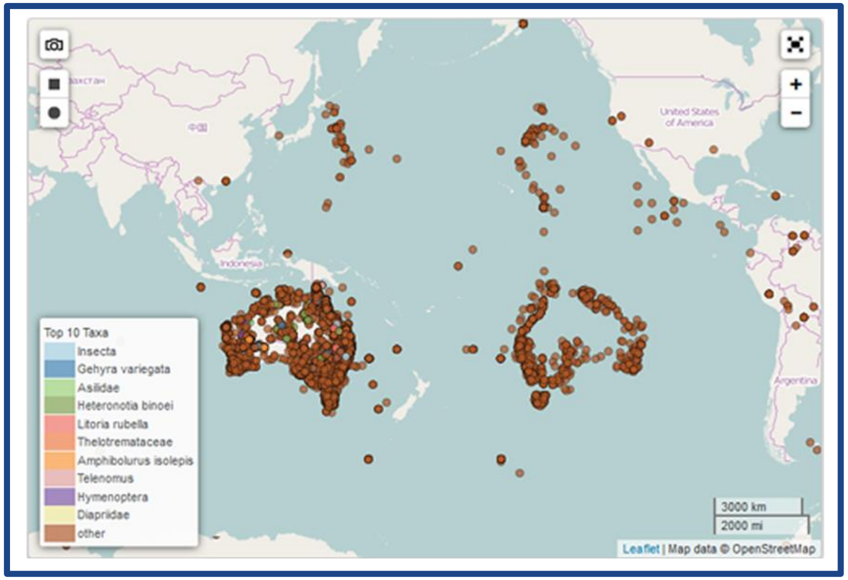
Flag
idigbio_isocountrycode_added ⓘ
dwc_continent_added ⓘ
dwc_country_replaced ⓘ
geopoint_datum_missing ⓘ
dwc_class_replaced ⓘ
dwc_phylum_replaced ⓘ
dwc_order_replaced ⓘ
geopoint_low_precision ⓘ
rev_geocode_eez ⓘ
dwc_stateprovince_replaced ⓘ
rev_geocode_mismatch ⓘ
dwc_order_added ⓘ
datecollected_bounds ⓘ
dwc_class_added ⓘ
dwc_kingdom_added ⓘ
dwc_phylum_added ⓘ
dwc_country_added ⓘ
rev_geocode_corrected ⓘ
rev_geocode_lon_sign ⓘ

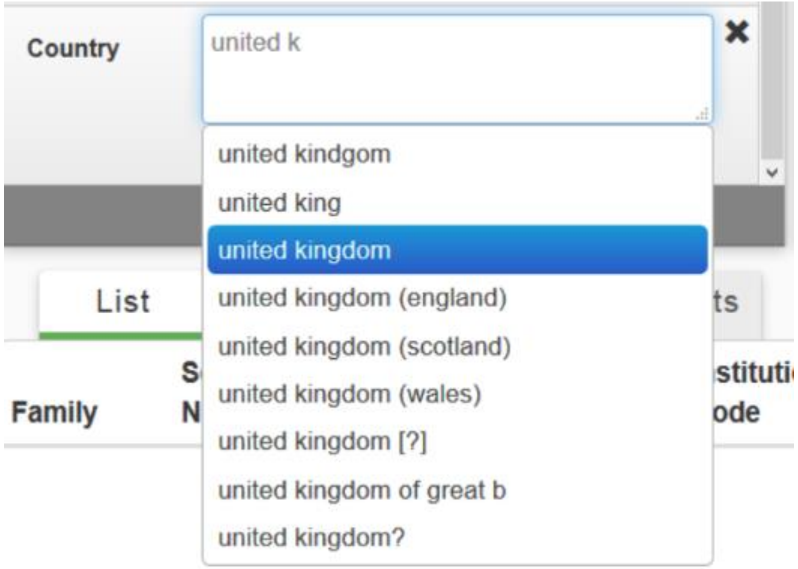
Data quality: *an issue at many levels*



Hannah Frost
@feefifofannah

From a @HydraInABox interview: "People will put anything and their dog in the date field. It's absolutely astonishing."





Country: united k

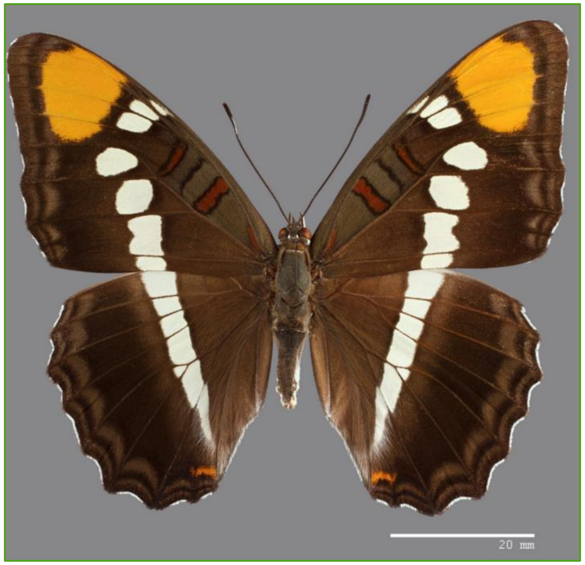
- united kindgom
- united king
- united kingdom
- united kingdom (england)
- united kingdom (scotland)
- united kingdom (wales)
- united kingdom [?]
- united kingdom of great b
- united kingdom?

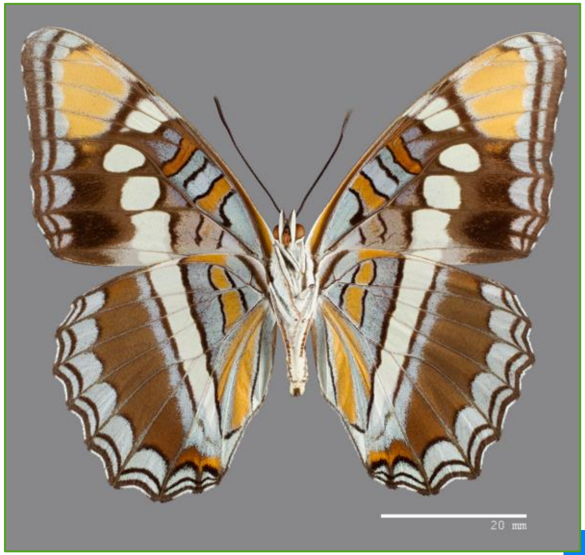
196 Countries in the world, but 1100 distinct values in the country field

Identifying specimens

- Determined by: Roman S. Wielgus
- Collected by: Roman S. Weilgus; Dalie Wielgus
- Date identified: 1968
- Sex: female
- Decimal Latitude: 34.7198786
- Georeferenced by: Jean-Batiste Quirino
- Date Collected: 1968-06-02
- Collector number: rsw1256
- Catalog number: ASUHIC0080642
- Institution code: ASU
- Collection code: ASUHIC
- Occurrence ID (GUID): afb73b66-ad30-40ea-bb86-8522448ad044

Adelpha bredowii eulalia (Geyer, 1837)





Identifying specimens globally

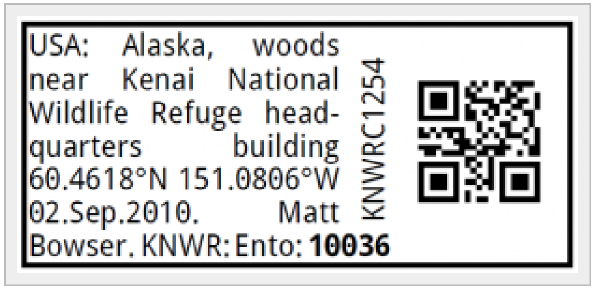
- Specimens (and other bits)
  - need Globally Unique Identifiers (GUIDs)
- Plan ahead – Shelley's example and FIMS
- Journals using identifiers – Pensoft
- Store identifiers and **no re-use**

- Why use them?
  - Support linked data
  - Data reusability
  - Data discovery

Examples

- e20d9229-4dfc-41c8-be05-afa7b40245a3
- urn:catalog:CAS:IZ:25668
- urn:uuid:6fdb535e-d43d-40c7-a387-e60c07cec5d1
- urn:lsid:biosci.ohio-state.edu:osuc_occurrences:OSUM__Birds__B__13589
- http://ucjeps.berkeley.edu/cgi-bin/new_detail.pl?RSA460997
- occurrenceID: afb73b66-ad30-40ea-bb86-8522448ad044

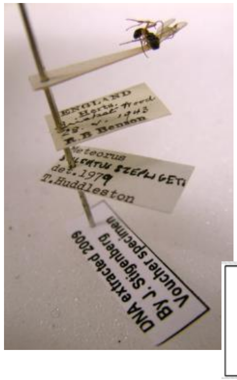
Adelpha bredowii eulalia (Geyer, 1837)

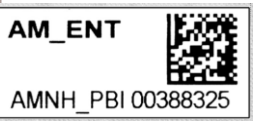


USA: Alaska, woods near Kenai National Wildlife Refuge headquarters building 60.4618°N 151.0806°W 02.Sep.2010. Matt Bowser. KNWR:Ento:**10036**

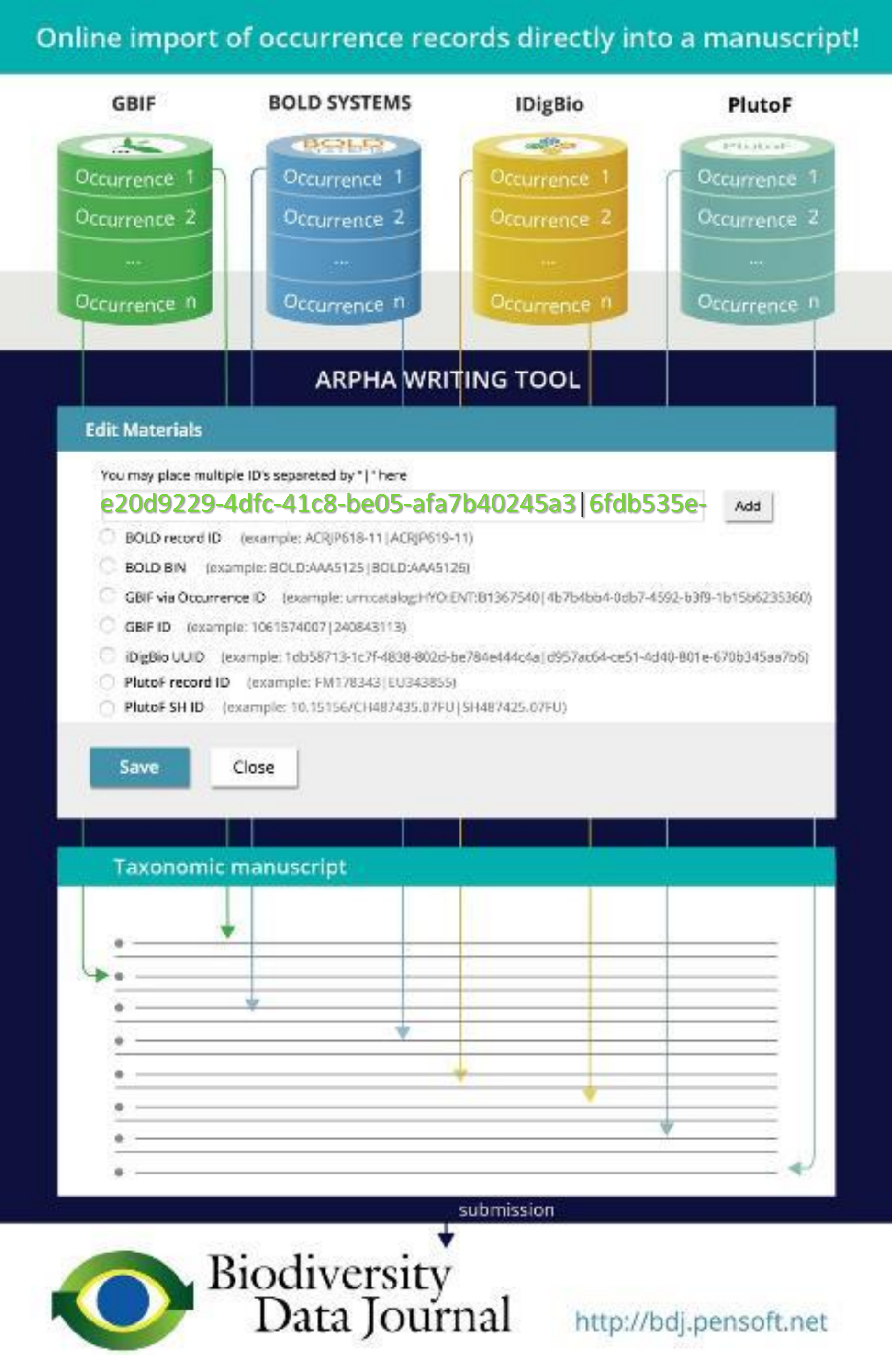
KNWRC1254

SEMC0993403
KUNHM-ENT





AM_ENT
AMNH_PBI 00388325



Edit Materials

You may place multiple ID's separeted by " | " here

e20d9229-4dfc-41c8-be05-afa7b40245a3|6fdb535e- [Add]

○ BOLD record ID (example: ACRJP618-11|ACRJP619-11)

○ BOLD BIN (example: BOLD:AAA5125|BOLD:AAA5126)

○ GBIF via Occurrence ID (example: urn:catalog:HYO:ENT:B1367540|4b7b4bb4-0db7-4592-b3f9-1b15b6235360)

○ GBIF ID (example: 1061574007|240843113)

○ iDigBio UUID (example: 1db58713-1c7f-4838-802d-be784e444c4a|d957ac64-ce51-4d40-801e-670b345aa7b6)

○ PlutoF record ID (example: FM178343|EU343855)

○ PlutoF SH ID (example: 10.15156/CH487435.07FU|SH487425.07FU)

[Save] [Close]

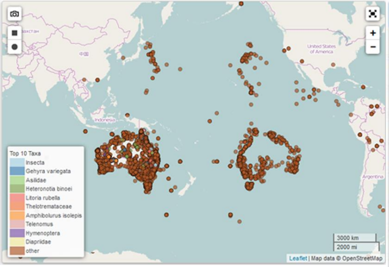
REPOSITORIES

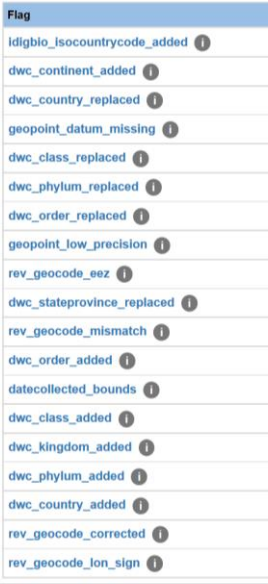
ARPHA WRITING TOOL

MANUSCRIPT

PUBLISHED ARTICLE

Natural History Museum
Specimen Data Aggregators and You





Flag

- idigbio_isocountrycode_added
- dwc_continent_added
- dwc_country_replaced
- geopoint_datum_missing
- dwc_class_replaced
- dwc_phylum_replaced
- dwc_order_replaced
- geopoint_low_precision
- rev_geocode_eez
- dwc_stateprovince_replaced
- rev_geocode_mismatch
- dwc_order_added
- datecollected_bounds
- dwc_class_added
- dwc_kingdom_added
- dwc_phylum_added
- dwc_country_added
- rev_geocode_corrected
- rev_geocode_lon_sign

- iDigBio, GBIF, VertNet, ALA, …
  - providing data quality information
- iDigBio dataset downloads
  - **original and enhanced data!**
  - **dataset citation** (Matt)
- when using museum specimen data
  - use / cite the guids (dwc:occurrenceID) provided
- share your data widely
- ask collections you work with if they are sharing their data
  - easy to do
    http://www.idigbio.org/wiki/index.php/Data_Ingestion_Guidance

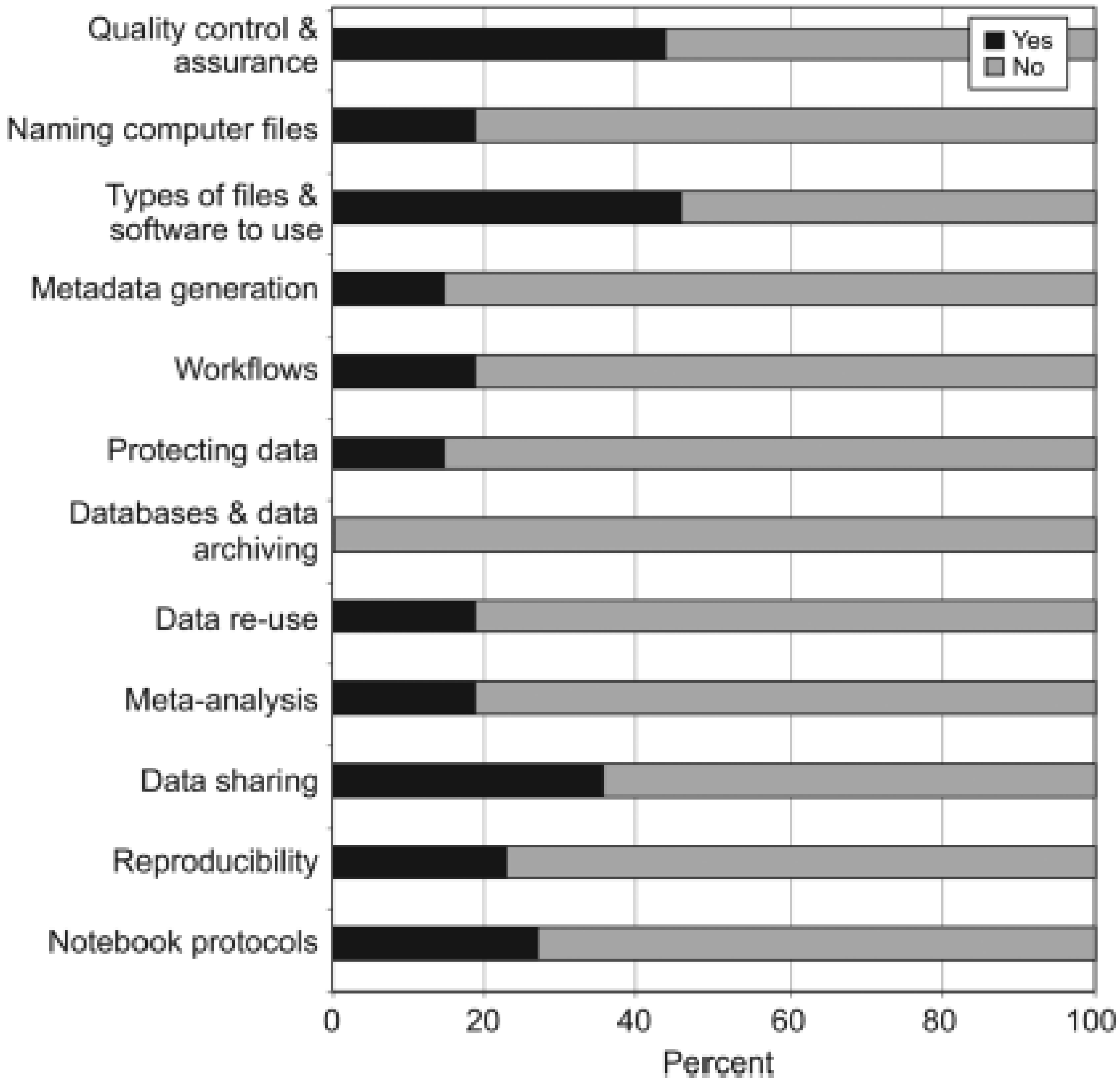
Once digitized, why share and publish your data? Can you share it in more than one place? Yes!

- power of aggregation
- accessibility (researchers, funders, collaborators)
- discoverability
- enhancement (scripts)
- tools and methods
- data quality checks
- off site copy
- visibility
- new uses for data

A researcher point-of-view

Researchers sentiments about data:
(BEACON, SESYNC, NESCent, iPlant, iDigBio)

- I usually manage data in Excel and it's terrible.
- I'm organizing GIS data and it's becoming a nightmare.
- I'm having a hard time analyzing microarray, SNP or multivariate data with Excel and Access.
- I want to use public data.
- I work with faculty at undergrad institutions and want to teach data practices, but I need to learn it myself first.
- I'm interested in going in to industry and companies are asking for data analysis experience.
- I'm trying to reboot my lab's workflow to manage data and analysis in a more sustainable way.
- I'm re-entering data over and over again by hand; there must be a better way.
- I have overwhelming amounts of data.
- I'm tired of feeling out of my depth on computation and want to increase my confidence.



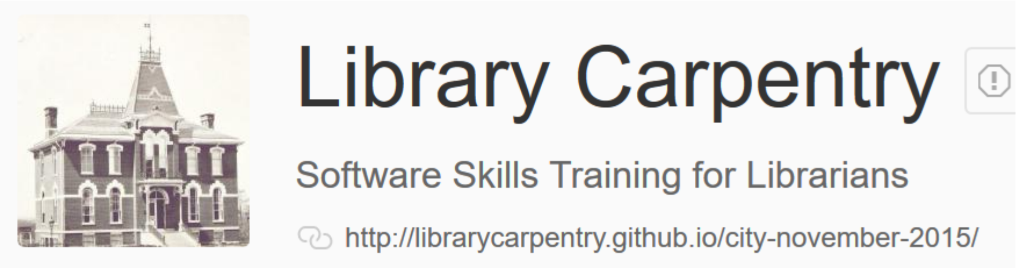
% Ecology courses that address / teach the listed data management principles

People seeking skills and efforts to scale up to meet needs

software carpentry

DATA CARPENTRY

Reproducible Science Curriculum

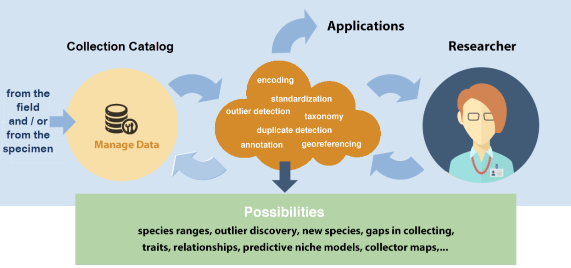


Library Carpentry

Software Skills Training for Librarians

http://librarycarpentry.github.io/city-november-2015/

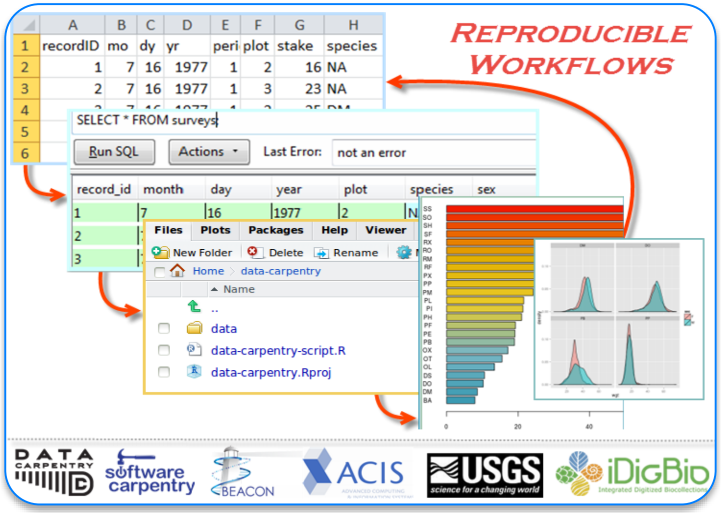
@GitHub



Possibilities

species ranges, outlier discovery, new species, gaps in collecting, traits, relationships, predictive niche models, collector maps,...

Some Biodiversity Informatics short courses at iDigBio



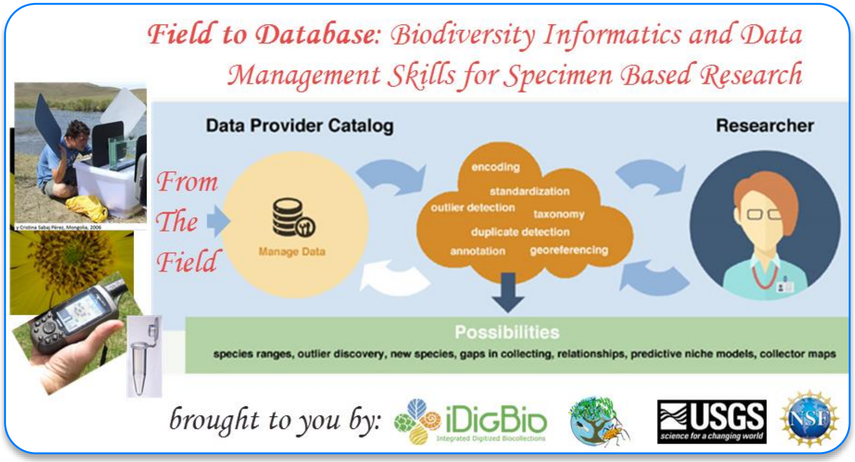
REPRODUCIBLE WORKFLOWS

	A	B	C	D	E	F	G	H
1	recordID	mo	dy	yr	peri	plot	stake	species
2	1	7	16	1977	1	2	16	NA
3	2	7	16	1977	1	3	23	NA

```
SELECT * FROM surveys
```

Run SQL | Actions | Last Error: not an error

record_id	month	day	year	plot	species	sex
1	7	16	1977	2	N	



Field to Database: Biodiversity Informatics and Data Management Skills for Specimen Based Research

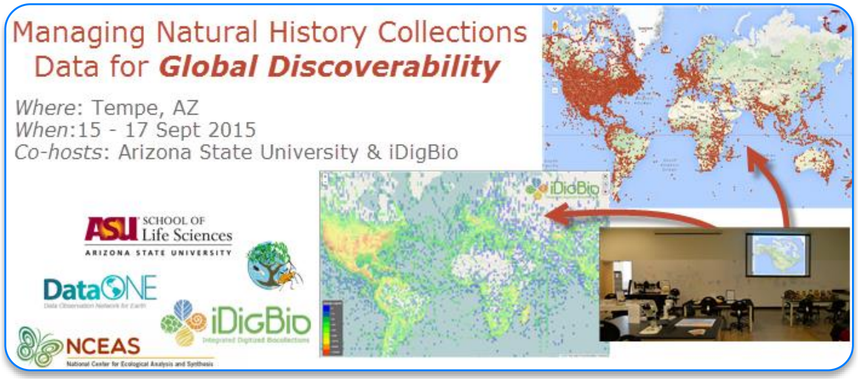
Data Provider Catalog
Researcher
From The Field
Manage Data
encoding
standardization
outlier detection
taxonomy
duplicate detection
annotation
georeferencing

Possibilities

species ranges, outlier discovery, new species, gaps in collecting, relationships, predictive niche models, collector maps

brought to you by:

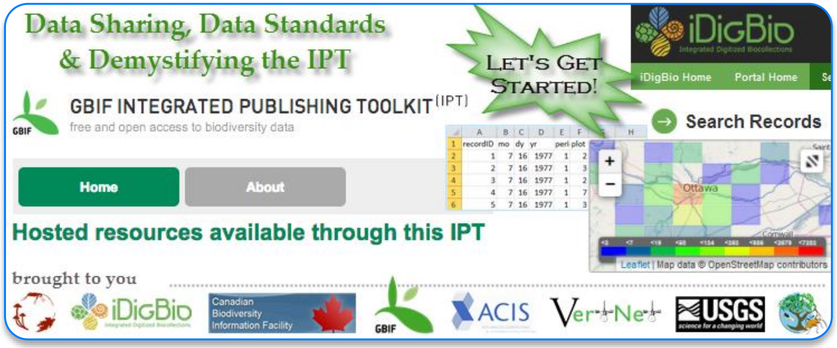
iDigBio Integrated Digitized Biocollections
USGS science for a changing world
NSF



Managing Natural History Collections Data for ***Global Discoverability***

Where: Tempe, AZ
When:15 - 17 Sept 2015
Co-hosts: Arizona State University & iDigBio

iDigBio
ASU SCHOOL OF Life Sciences ARIZONA STATE UNIVERSITY
DataONE Data Observation Network for Earth
iDigBio Integrated Digitized Biocollections
NCEAS National Center for Ecological Analysis and Synthesis



Data Sharing, Data Standards & Demystifying the IPT

LET'S GET STARTED!

GBIF INTEGRATED PUBLISHING TOOLKIT (IPT)
free and open access to biodiversity data

Home | About

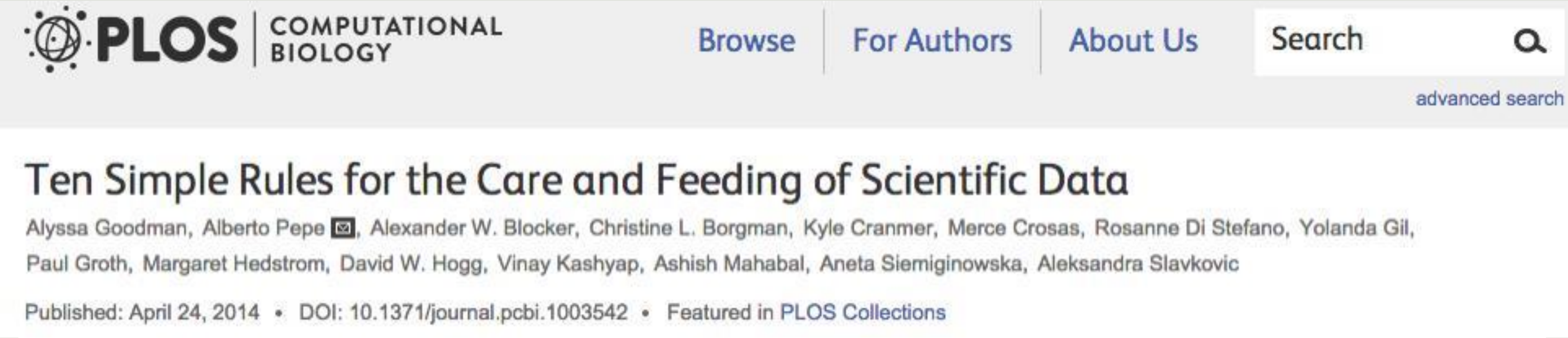
Hosted resources available through this IPT

	A	B	C	D	E	F
1	recordID	mo	dy	yr	peri	plot
2	1	7	16	1977	1	2
3	2	7	16	1977	1	3
4	3	7	16	1977	1	2
5	4	7	16	1977	1	7
6	5	7	16	1977	1	3

iDigBio Integrated Digitized Biocollections
iDigBio Home
Portal Home
Search Records
Ottawa
Leaflet | Map data © OpenStreetMap contributors
GBIF

brought to you

iDigBio
Canadian Biodiversity Information Facility
GBIF
ACIS
VertNet
USGS science for a changing world



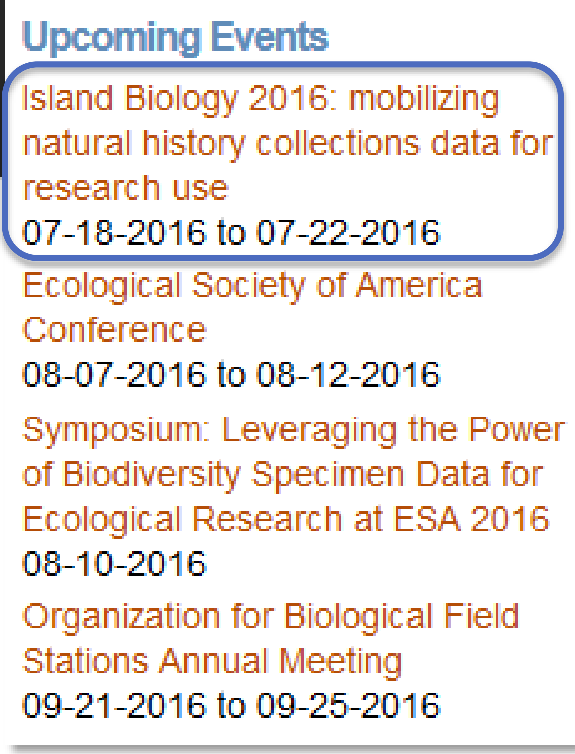
Ten Simple Rules for the Care and Feeding of Scientific Data

Alyssa Goodman, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, Yolanda Gil, Paul Groth, Margaret Hedstrom, David W. Hogg, Vinay Kashyap, Ashish Mahabal, Aneta Siemiginowska, Aleksandra Slavkovic

Published: April 24, 2014 • DOI: 10.1371/journal.pcbi.1003542 • Featured in PLOS Collections

- Love your data, and help others love it, too
- Share your data online, with a permanent identifier
- Conduct science with a particular level of reuse in mind
- Publish workflow as context
- Link your data to your publications as often as possible
- Publish your code (even the small bits)
- State how you want to get credit
- Foster and use data repositories
- Reward colleagues who share their data properly
- Be a booster for data science

And now, more about collections data fit-for-research use,



Upcoming Events

Island Biology 2016: mobilizing natural history collections data for research use
07-18-2016 to 07-22-2016

Ecological Society of America Conference
08-07-2016 to 08-12-2016

Symposium: Leveraging the Power of Biodiversity Specimen Data for Ecological Research at ESA 2016
08-10-2016

Organization for Biological Field Stations Annual Meeting
09-21-2016 to 09-25-2016

facebook.com/iDigBio

twitter.com/iDigBio

vimeo.com/idigbio

idigbio.org/rss-feed.xml

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

www.idigbio.org

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.

TASK GROUPS ON FITNESS FOR USE:
INVASIVE ALIEN SPECIES AND DNA EVIDENCE

2016 group on invasive alien species:
Melodie McGeoch + 5 experts

- meetings, teleconferences and survey
- report in November 2016

* * *

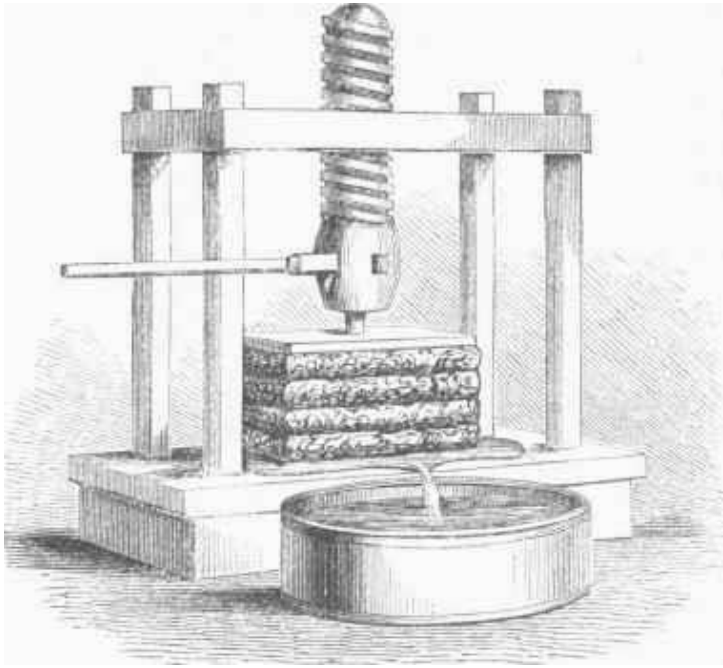
Exploring data use and data publishing needs of molecular biodiversity research: **DNA** survey, GGBN conference and TDWG

Interested? Contact Dmitry Schigel
dschigel@gbif.org

DATA QUALITY, ANALYSIS AND USE

PASSED

Promoting data quality culture:
talks, teaching, [publications]



TDWG / GBIF
data quality interest group, chairs:
Antonio Saraiva & **Arthur Chapman**

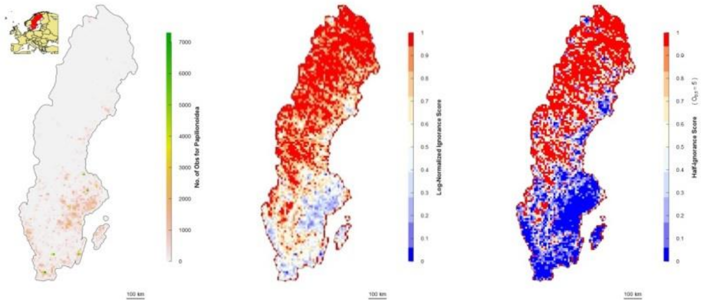
Biodiversity
Information
Standards
TDWG

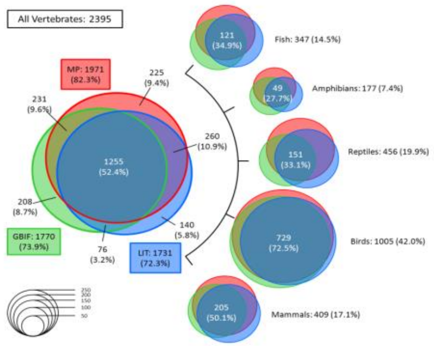
Join the work at the **GBIF Community Site!**

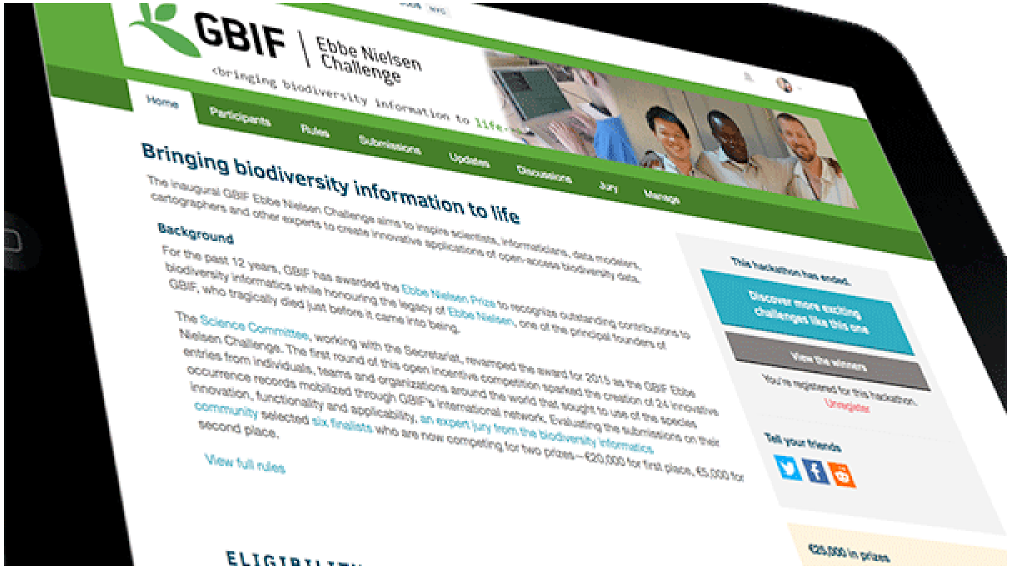
http://www.gbif.org/using-data

GBIF

GAPS AND EBBE NIELSEN CHALLENGE 2016







GBIF | Ebbe Nielsen Challenge

<bringing biodiversity information to life>

Home · Participants · Rules · Submissions · Updates · Discussions · Jury · Manage

Bringing biodiversity information to life

The inaugural GBIF Ebbe Nielsen Challenge aims to inspire scientists, informaticians, data modelers, cartographers and other experts to create innovative applications of open-access biodiversity data.

Background

For the past 12 years, GBIF has awarded the Ebbe Nielsen Prize to recognize outstanding contributions to biodiversity informatics while honouring the legacy of Ebbe Nielsen, one of the principal founders of GBIF, who tragically died just before it came into being.

The Science Committee, working with the Secretariat, revamped the award for 2015 as the GBIF Ebbe Nielsen Challenge. The first round of this open incentive competition sparked the creation of 24 innovative entries from individuals, teams and organizations around the world that sought to use of the species occurrence records mobilized through GBIF's international network. Evaluating the submissions on their innovation, functionality and applicability, an expert jury from the biodiversity informatics community selected six finalists who are now competing for two prizes—€20,000 for first place, €5,000 for second place.

View full rules

This hackathon has ended.

Discover more exciting challenges like this one

View the winners

You're registered for this hackathon. Unregister

Tell your friends

€25,000 in prizes

ELIGIBILITY

Challenge opens on 29 July 2016

First prize **€20,000**, second prize **€5,000**

Winners announced on 26 October 2016

Dealing with gaps is the theme of the Ebbe Nielsen Challenge 2016

http://www.gbif.org/about/awards

Public Participation in Digitization

Transcribe museum records to

TAKE NOTES FROM NATURE

BIOSPEX

WeDigBio

DigiVol

Transcription Blitz with FL Native Plant Society

Herbarium Imaging Blitz