From Biocollections to Global Change Biology: New Conceptual and Cyberinfrastructure Frameworks For Closing The Gap

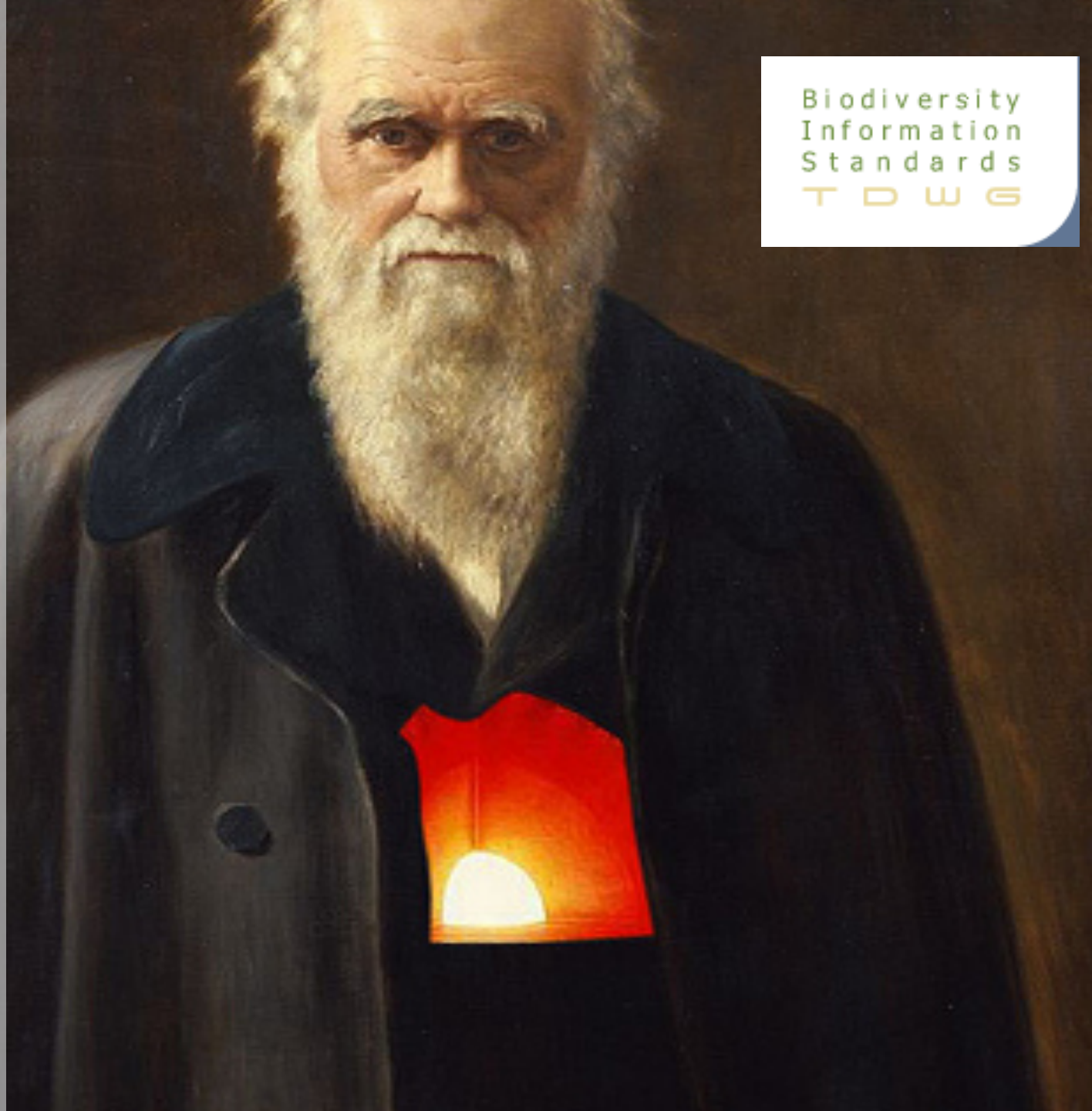
Collaborators: Javier Otegui (CU Boulder), Nico Cellinese (UF), John Deck (UC Berkeley), Ramona Walls (iPlant), John Wieczorek (Cal), Walter Jetz (Yale University).

Funding support: Global Biodiversity Information Facility, National Biological Information Infrastructure, National Science Foundation, NASA, iPlant, NCEAS, EOL-BioSync.

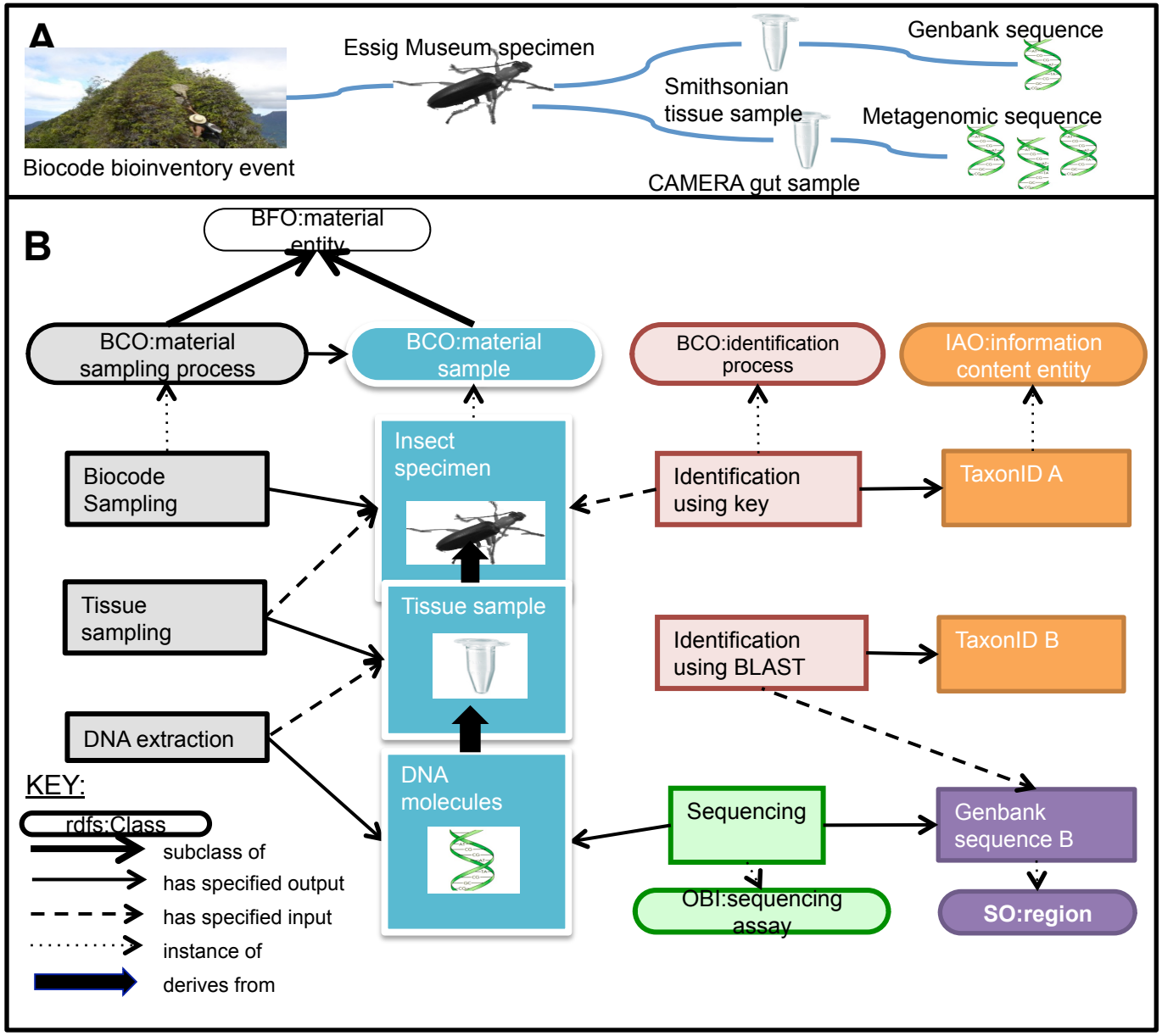
BIOCOLLECTIONS REPRESENT A FUNDAMENTAL SOURCE OF LEGACY DATA ABOUT OUR BIOSPHERE

Our collections practice is based on evolving but still (mostly) consistent best practices

Natural History Community Has Developed DarwinCore. And it is awesome BUT…

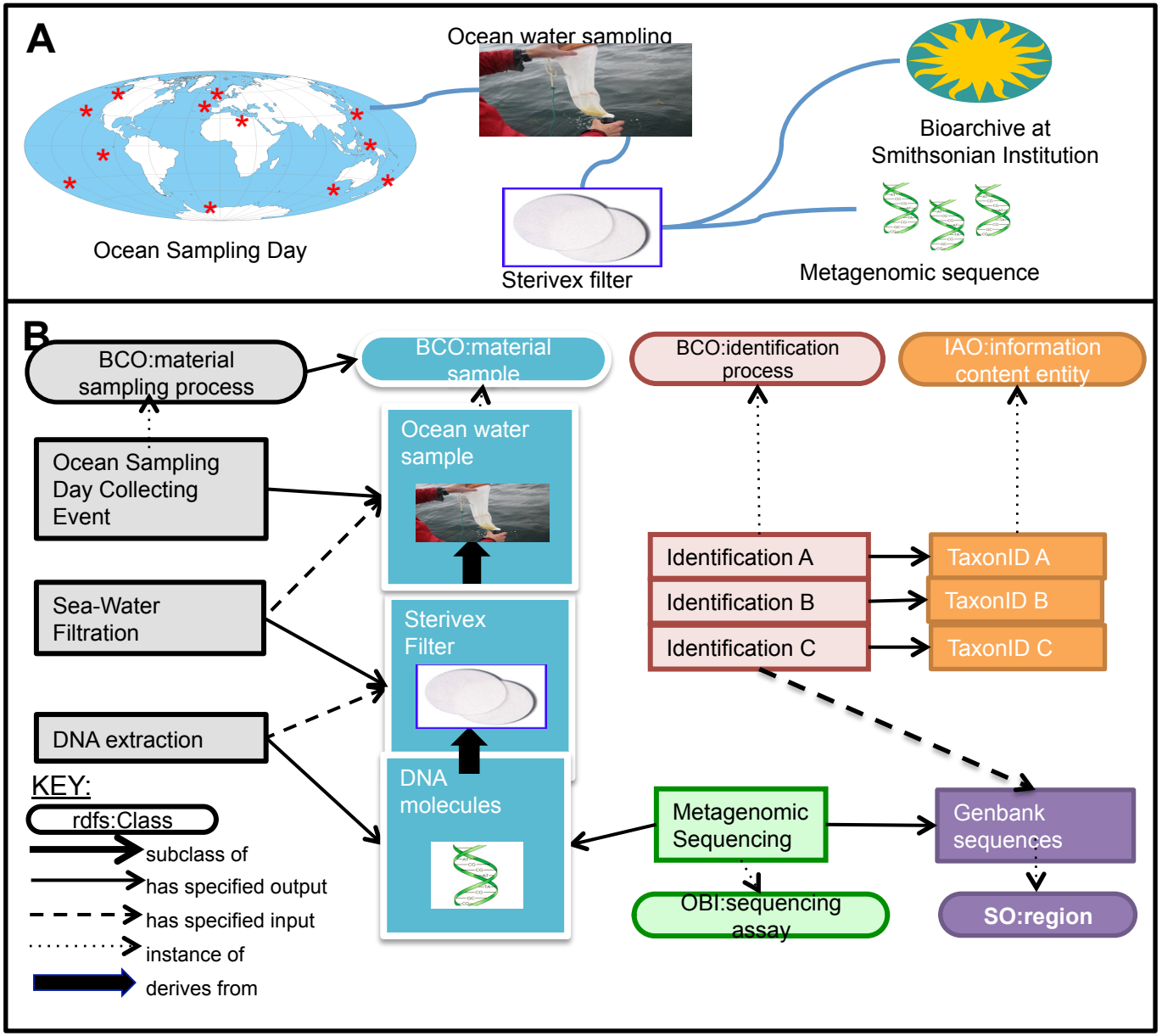


Biodiversity Information Standards TDWG



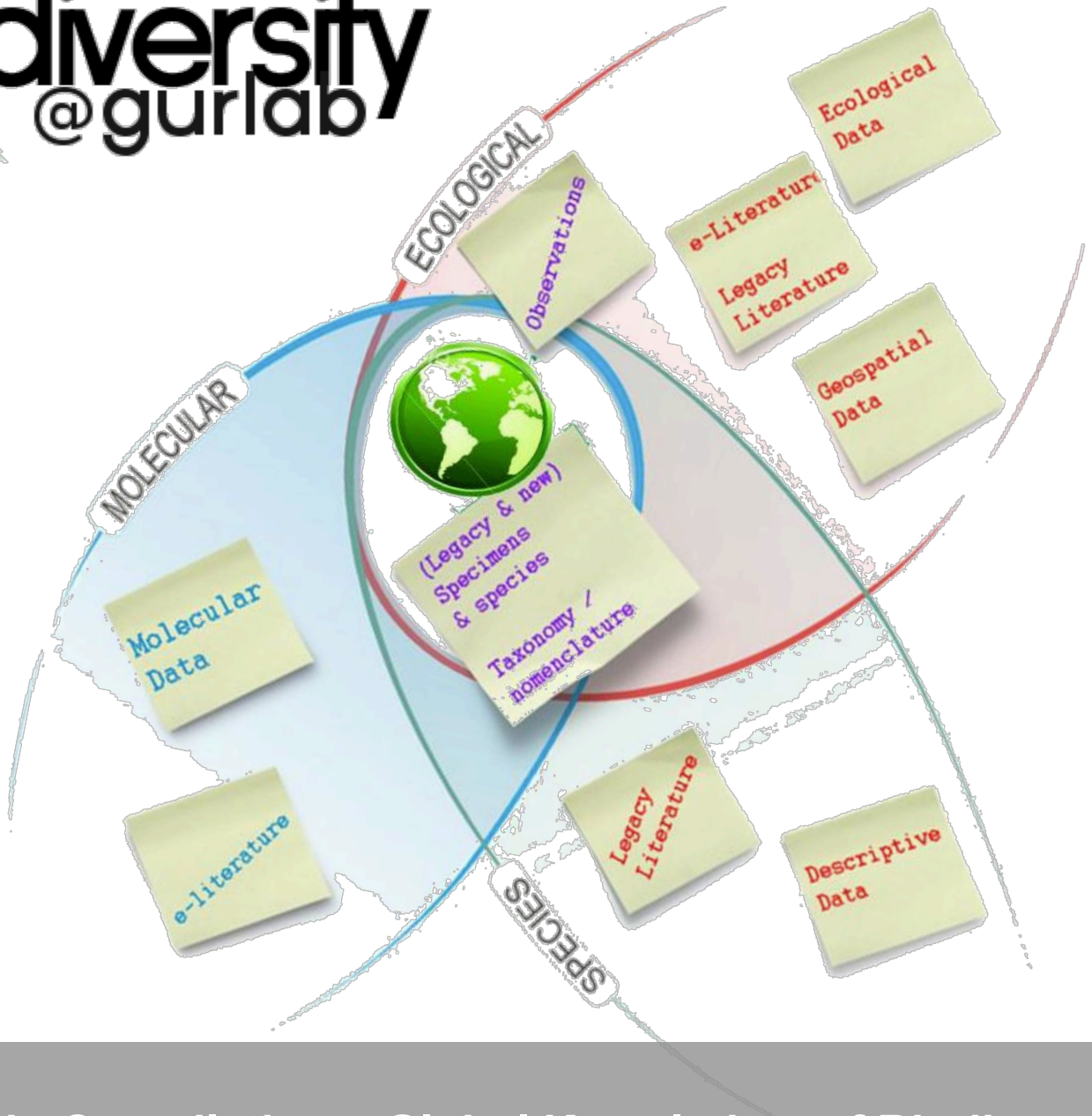
From Walls et al. 2014 PLOS ONE Biocollections Ontology

Biocollections in the 21st century are growing and changing –
We need ways to represent novel sampling and sub-sampling processes and keep them all linked together



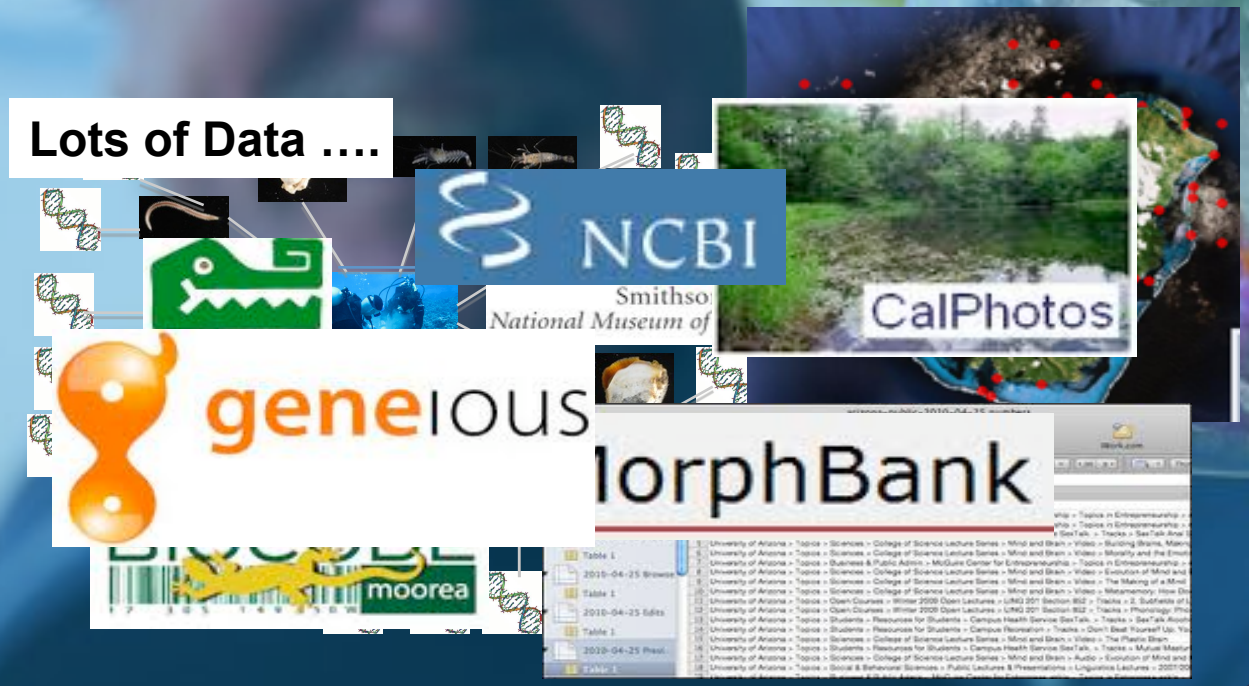
And some material sample biocollections yield information about organisms, but may not provide evidence other than genes

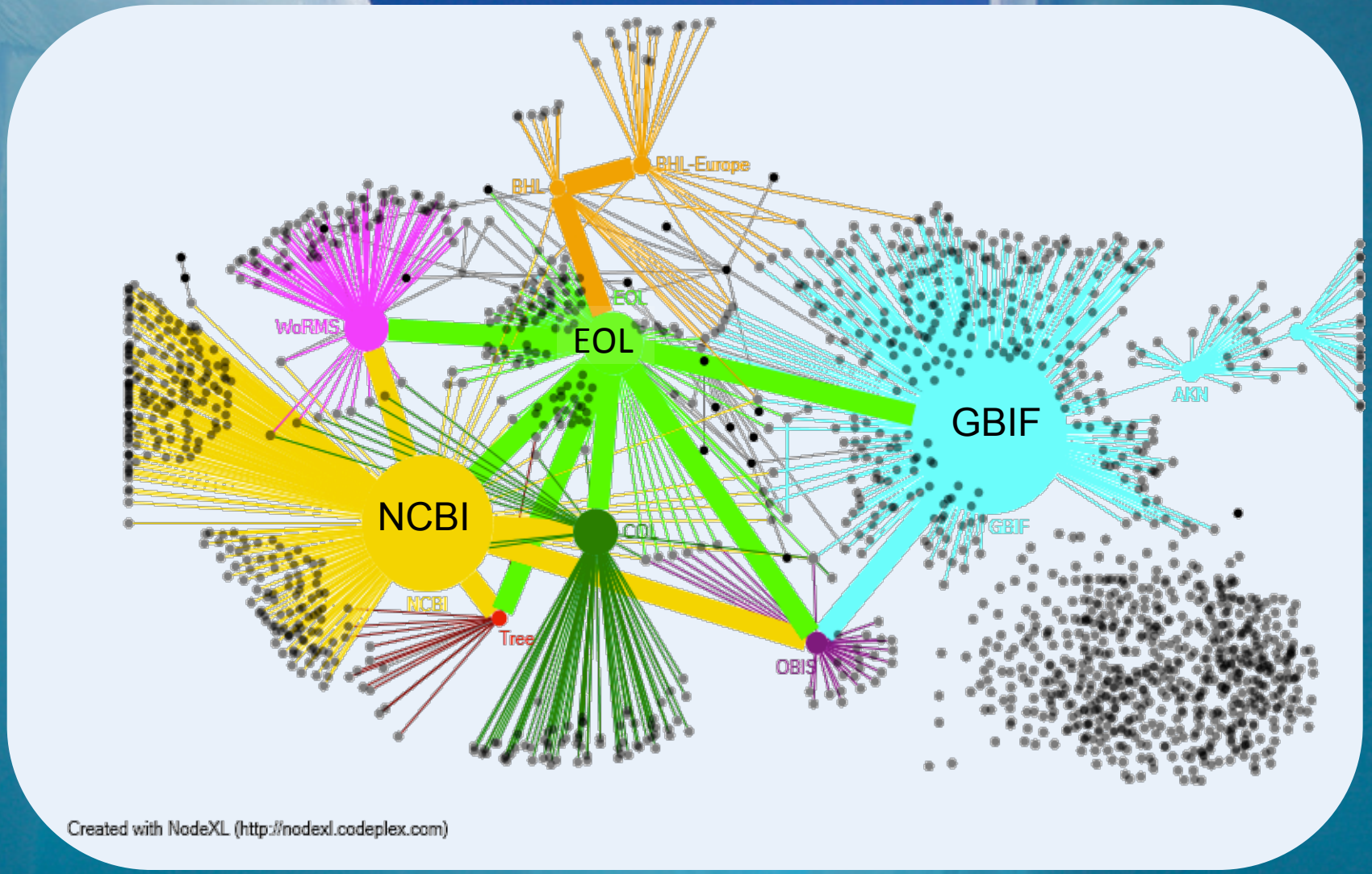
Biodiversity @gurlab



Big Compilation – Global Knowledge of Biodiversity

Here is the problem:



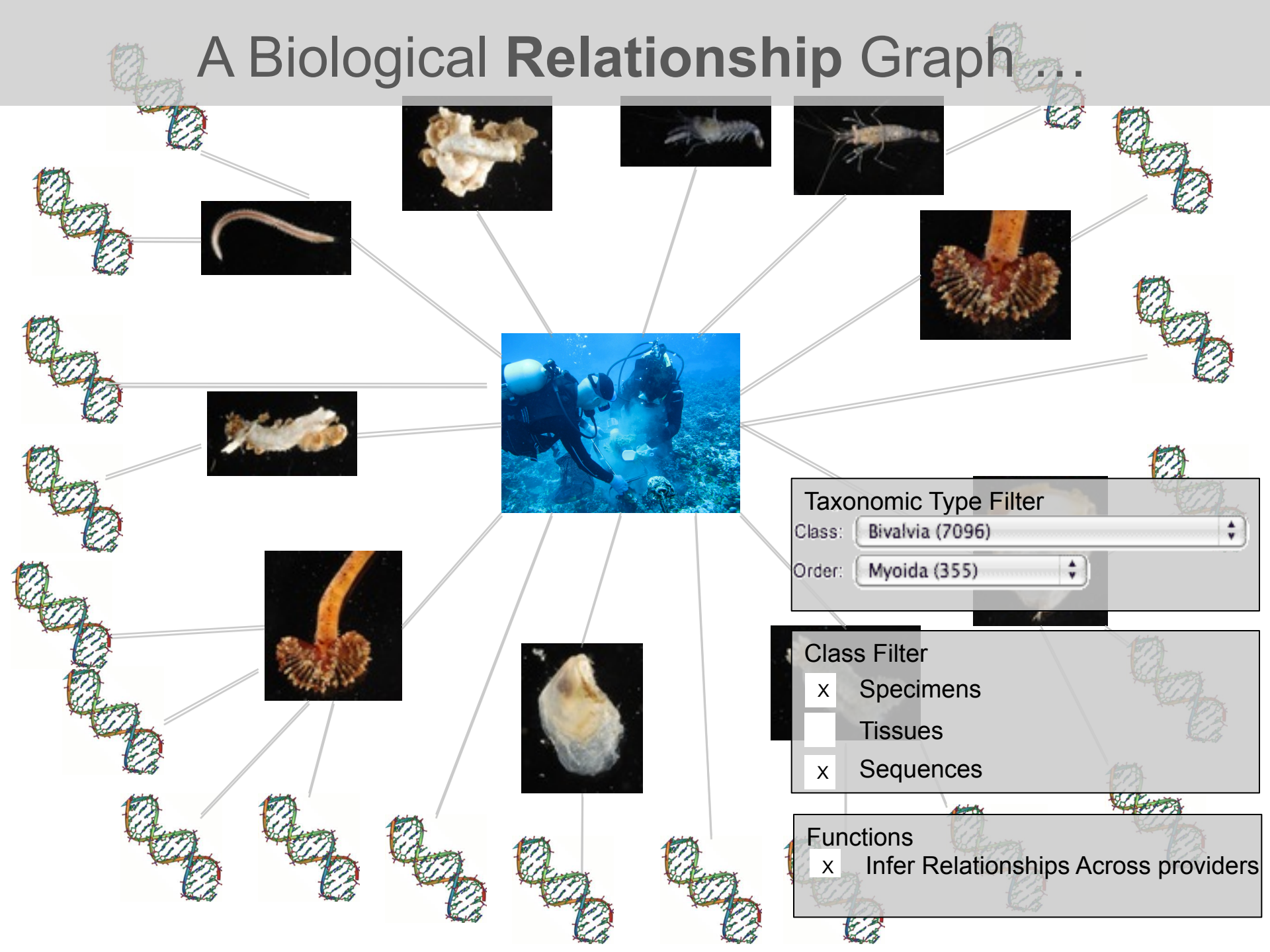


A Growing Constellation of Biodiversity Data and Knowledge

How do we link all these data together?

Borrowing from Facebook and social media…
Can we track relationships for Biological Objects as well?

A Biological **Relationship** Graph …



How to Guide: Tracking Biological Object **Relationships**

Utilize community standards built on class, object and data properties. This requires extending, or rethinking existing standards such as Darwin Core

Assign Identifiers to objects. Use globally unique, resolvable, persistent identifiers for each class or term.

Link Identifiers using relationship terms and specified classes. For example, "This object is related to that object."

Put this data on the Web.

Global Unique identifiers:

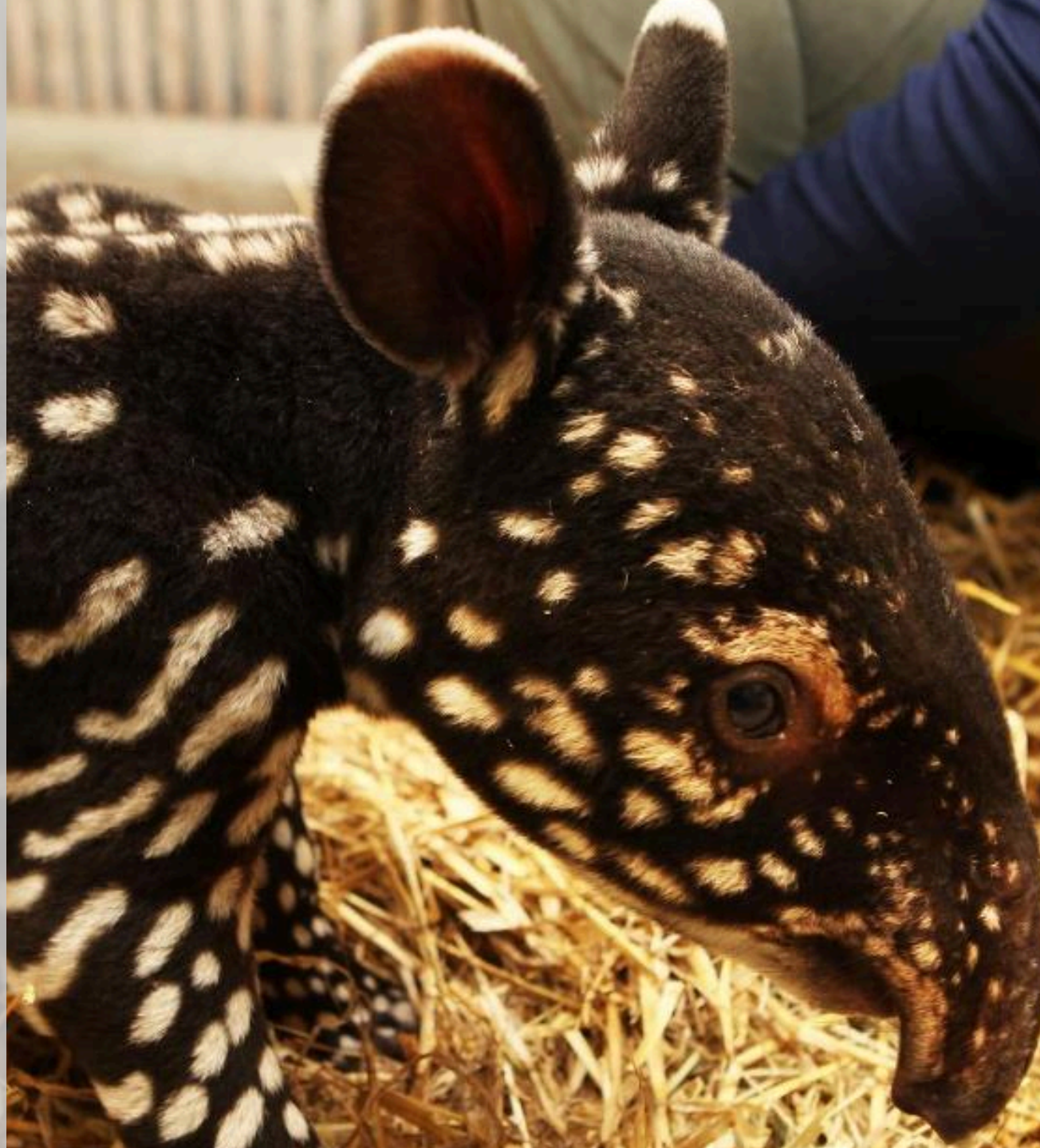
- Globally unique (mandatory)
- Persistent (not mandatory, but very helpful)
- Resolvable (not mandatory, but very helpful)

Examples:

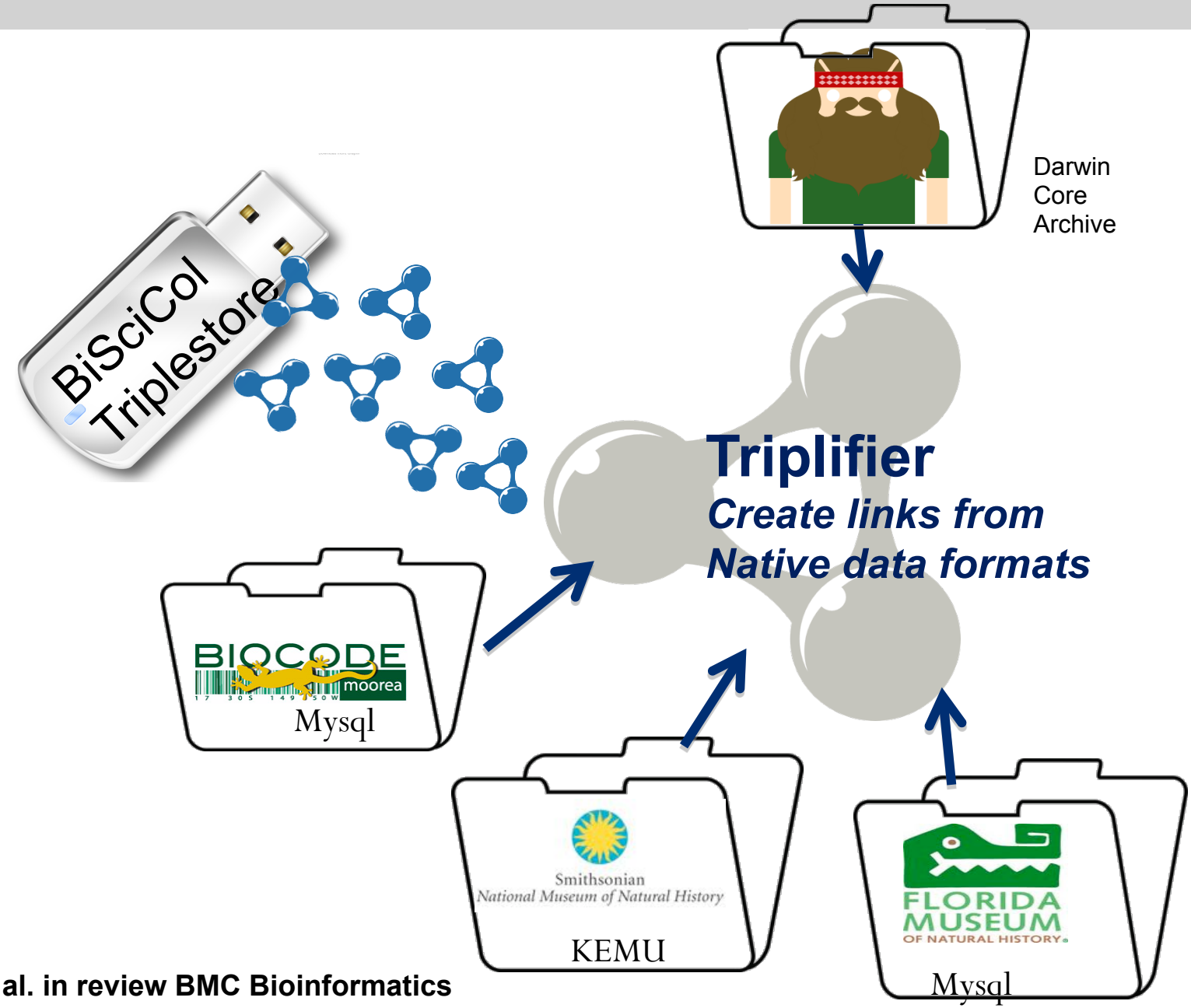
http://example.org/urn:lsid:example.org:specimen/7217D220-836A-11DF-8395-0800200C9A66
http://mycollection.org/specimen/JDeckSpecimen1
http://mycollection.org/specimen/uuid=7217D220-836A-11DF-8395-0800200C9A66
http://dx.doi.org/10.5072/FK2JW8GKM

ONE FINAL PIECE OF THE PUZZLE:

GIVING BIRTH TO DATA IN THE RIGHT FORMAT FOR LINKING



“Triplifier” - creating the format for linking biological objects



Stucky et al. in review BMC Bioinformatics

BISCICOL – EXAMPLE SEARCH

Client Interface:

Search Scientific Name: Aedes increpitus Run

Results:
OccurrenceID1 (*Aedes increpitus* **Dyar, 1916**)
OccurrenceID3 (*Aedes vittata* **Theobald, 1903**)

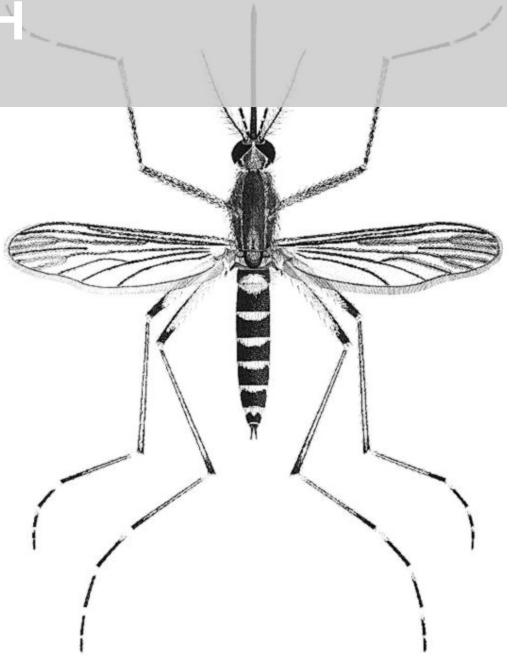


Plate 68. *Aedes increpitus* Dyar, female.

Taxon SERVICE (ITIS / GNUB)
http://lsid.itis.gov/urn:lsid:itis.gov:itis_tsn:126314
http://lsid.itis.gov/urn:lsid:itis.gov:itis_tsn:126317
http://gnub.org/8E19F1DC-74BA-47D4-A505-6498414B4CCE

BISCICOL SERVICE LOOKUP:
dwc:IdentificationID1 :relatedTo http://lsid.itis.gov/urn:lsid:itis.gov:itis_tsn:126314
dwc:IdentificationID1 :relatedTo dwc:OccurrenceID1
dwc:IdentificationID2 :relatedTo http://lsid.itis.gov/urn:lsid:itis.gov:itis_tsn:126317
dwc:IdentificationID2 :relatedTo dwc:OccurrenceID3

Data Impact Factor – Graph Metrics

Collectors

Gustav Paulay
(102,000 direct children)

Christopher Meyer
(83,000 direct children)

Craig Moritz
(523 direct children)

Events

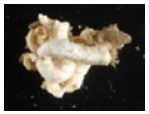
Biocode10234
(4234 direct children)

Expedition21234
(1023 direct children)

Graphs

[ ] GBIF Relations Graph
[X] Moorea Biocode
[X] SI MSNGR System
[+] Add New Graph

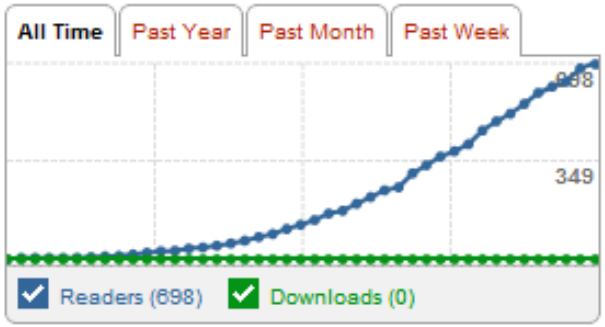
Occurrences



MBIO99999
(1024 total descendents)

IMBL8888888
(723 total descendents)

Cited occurrences over time



FROM LINKING DATA TO ASSESSING QUALITY

Feature	Total
Total records	**200183168 (100.00%)**
(A) Geospatial Errors	
Coordinates equal to zero	1456654 (0.73%)
Impossible coordinates	10117 (0.01%)
Low precision	16252100 (8.12%)
Out of the specified country	14040820 (7.01%)
Transposed coordinates	322734 (0.16%)
Negated Latitude	272829 (0.14%)
Negated Longitude	383919 (0.19%)
(B) Spatio-taxonomic errors	
Inside range map	146877631 (73.37%)
Less than 55Km	18408468 (9.20%)
55-111Km	2228994 (1.11%)
111- 555Km	3783218 (1.89%)
More than 555Km	5417136 (2.71%)
Without range map assessment	23467721 (11.72%)
Without RM assessment - taxon issues	12912456 (6.45%)

Data from Otegui and Guralnick, In prep

SPATIO-TAXONOMIC ERROR RATES IN GBIF (2012)

A global assessment of terrestrial vertebrates

Individual Explanatory Variable	McFadden Pseudo-R2 value
Dataset	0.9040
Publisher	0.7192
Publisher country	0.5511
Installation type	0.2908
Publisher continent	0.2626
Dataset creation year	0.0353
Basis of record	0.0320
GDP	0.0184
Income group	0.0128
Occurrence year	0.0074
Taxonomic class	0.0007

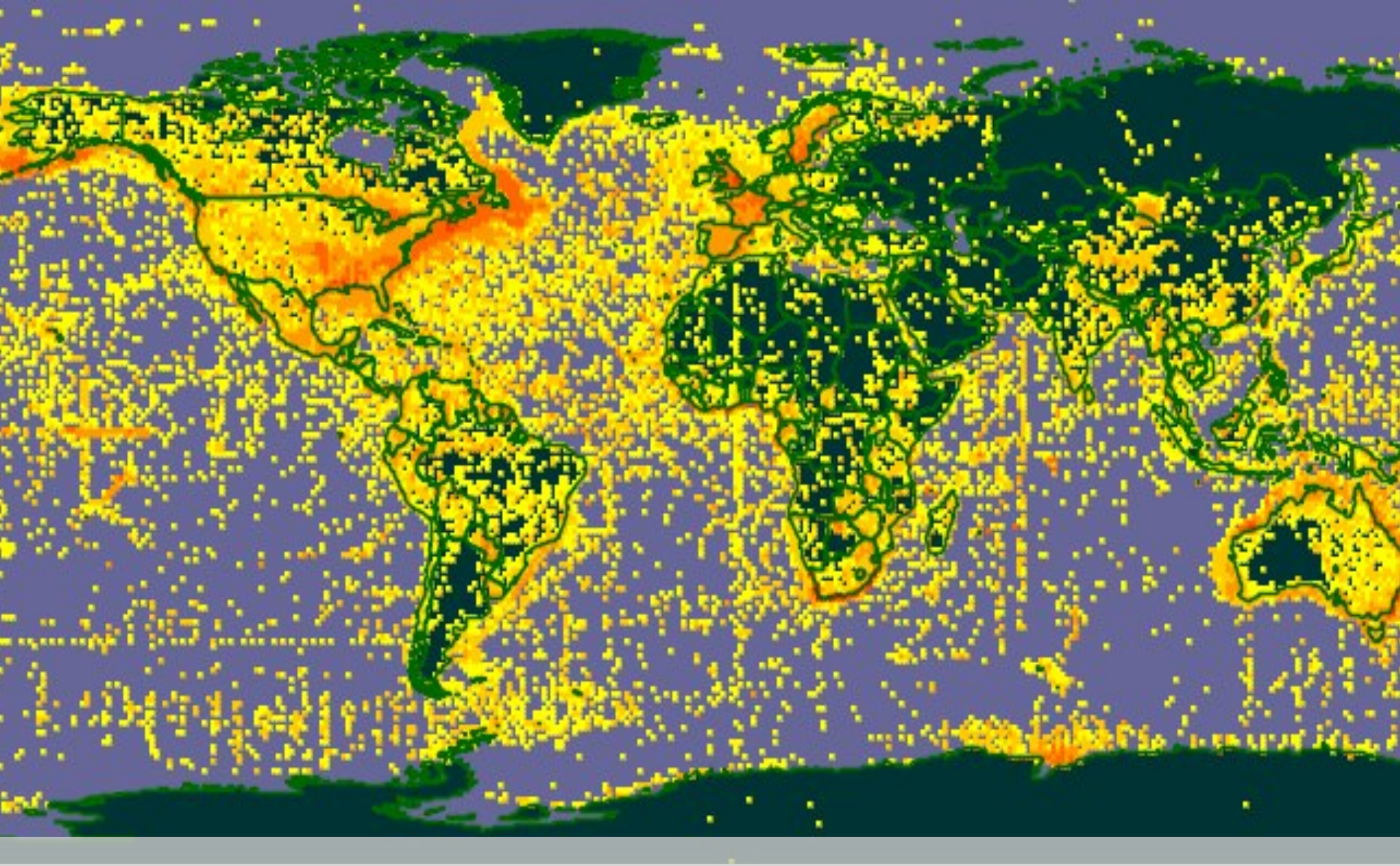
Data from Otegui and Guralnick, In prep

Single factor logistic models – Factors related to error rates

ALSO… Data quality tools we developed available as a prototype API …
For "typical error" checking and for "point-range map intersections"

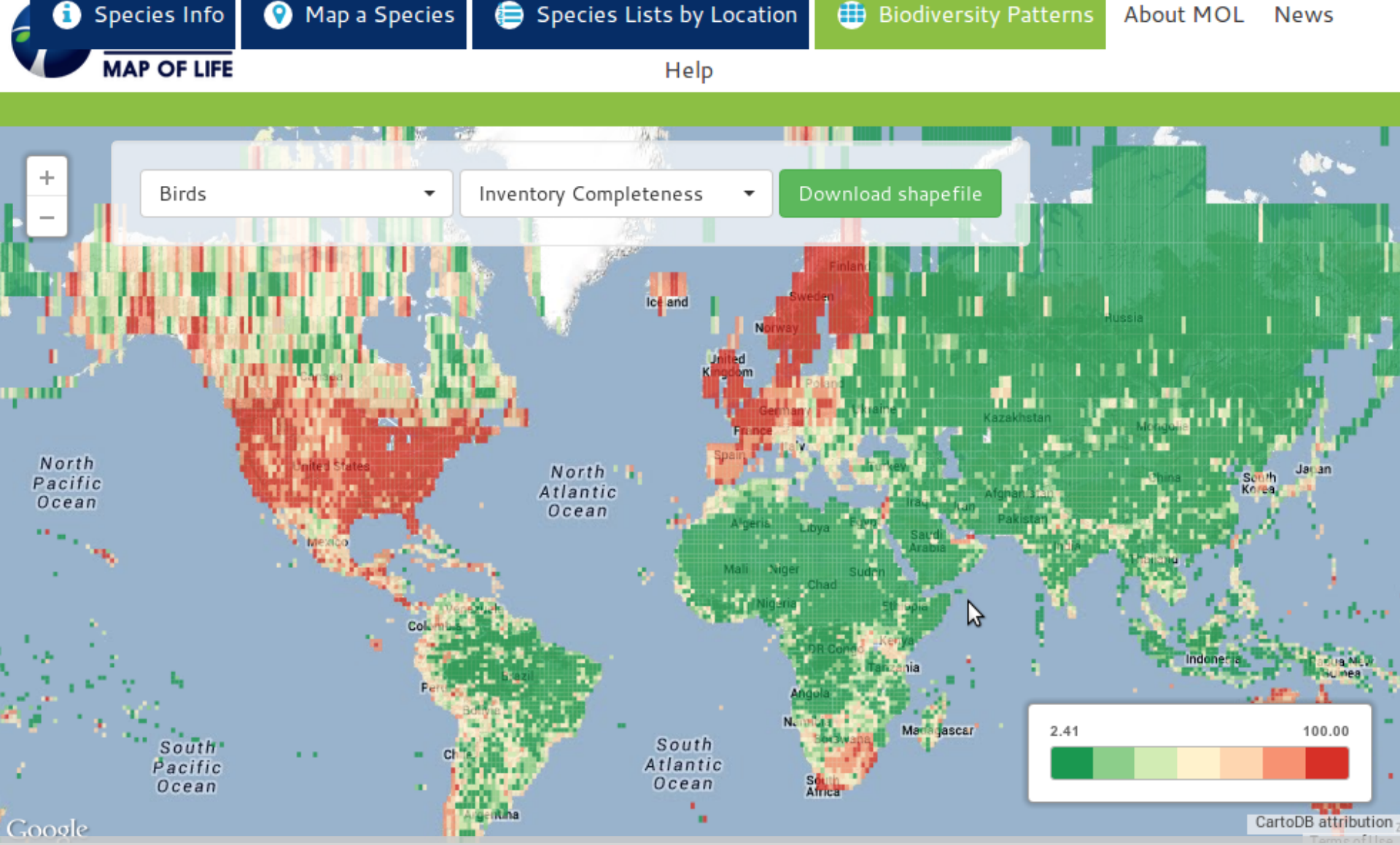
HOW WE CAN DOCUMENT KNOWLEDGE GAPS AND BETTER FORECAST CHANGE?

Two examples: Inventory Completeness, Niche/Distribution shifts



Provisioning of our collections data starts to provide a global view on biodiversity

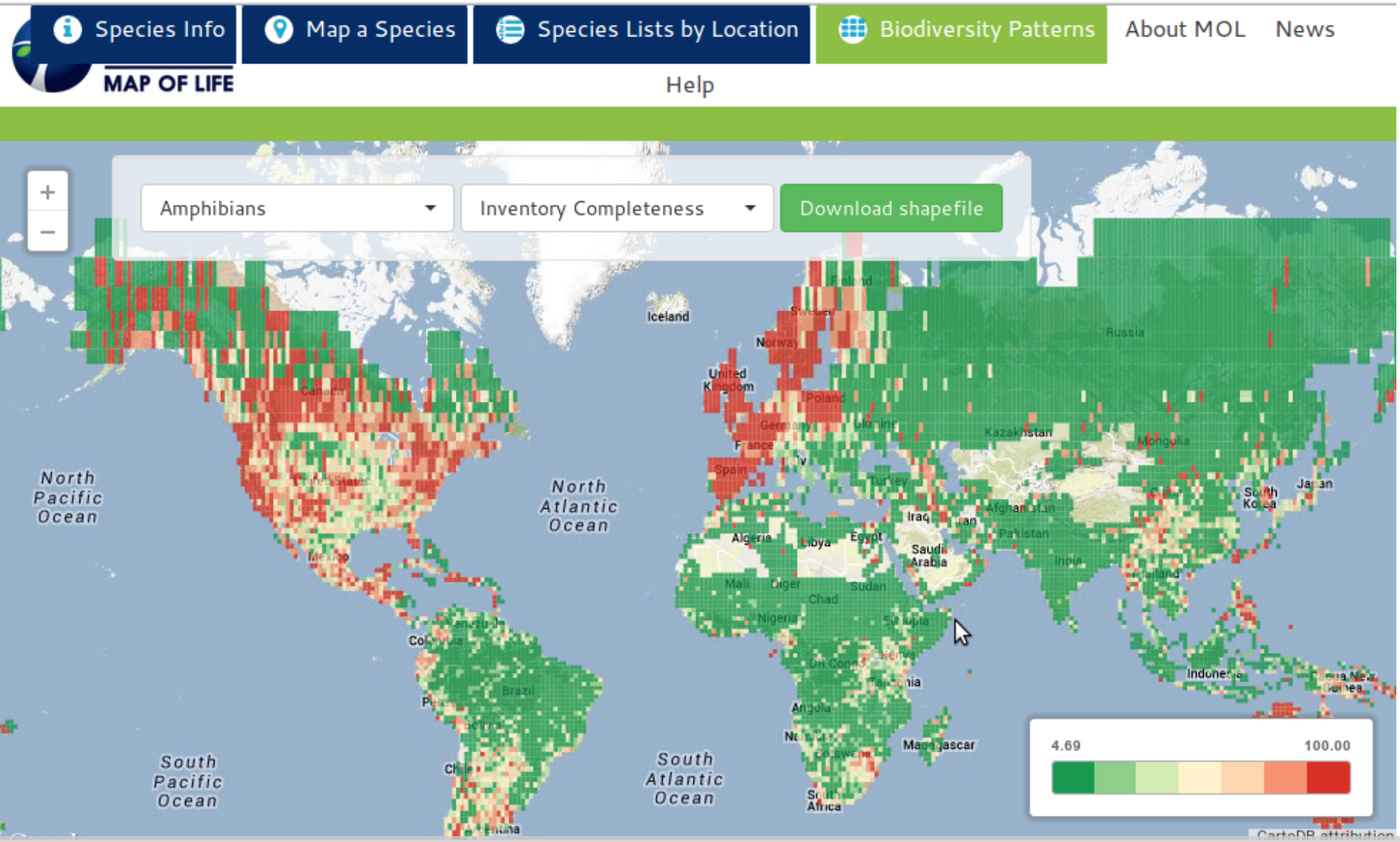
- ***We can start with record density,***
- ***<u>but what about other, more meaningful measures?</u>***



How well do existing global occurrence repositories reflect our knowledge of inventory completeness?

And how do we priortize gap filling efforts?

A new, growing tool on Map of Life



How well do existing global occurrence repositories reflect our knowledge of inventory completeness?

Does it vary in different taxonomic groups? What drives these patterns?

(Meyer, Kreft, Guralnick, Jetz in review Science)

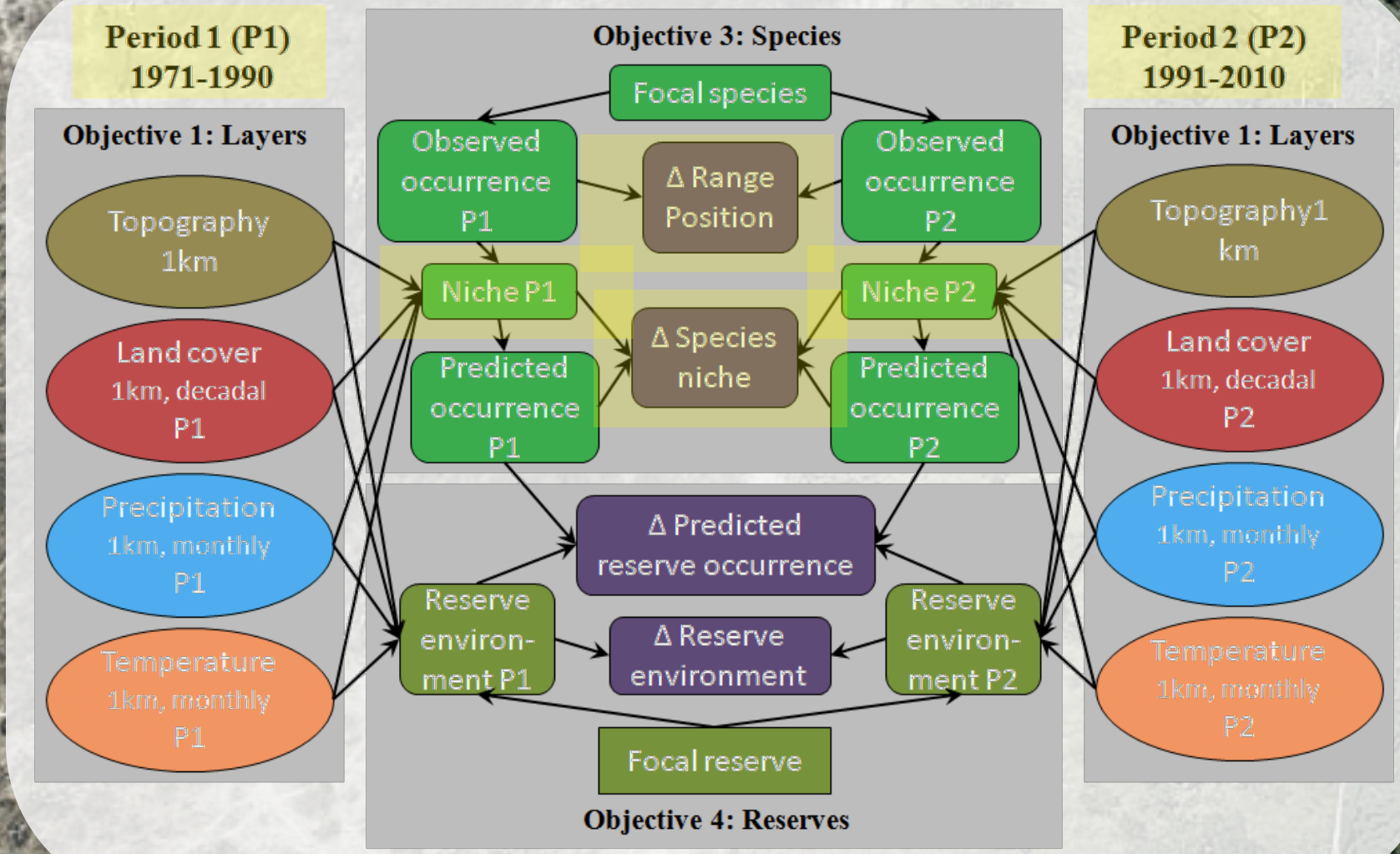
A BROAD-SCALE, CHANGE ASSESSMENT METHOD OVER RECENT PAST :

Objective 1: Develop of global environmental layers 1971-2010

Objective 2: Annotate species occurrences with environment at time of collection

Objective 3: Model niche and distribution at two different periods (‘70-’90 & ‘90-2010)

Objective 4: Change in environment, focal species occurrences in reserves 1971-2010.


Period 1 (P1)
1971-1990
Objective 1: Layers
Topography
1km
Land cover
1km, decadal
P1
Precipitation
1km, monthly
P1
Temperature
1km, monthly
P1
Objective 3: Species
Focal species
Observed
occurrence
P1
Δ Range
Position
Observed
occurrence
P2
Niche P1
Niche P2
Δ Species
niche
Predicted
occurrence
P1
Predicted
occurrence
P2
Δ Predicted
reserve occurrence
Reserve
environ-
ment P1
Δ Reserve
environment
Reserve
environ-
ment P2
Focal reserve
Objective 4: Reserves
Period 2 (P2)
1991-2010
Objective 1: Layers
Topography1
km
Land cover
1km, decadal
P2
Precipitation
1km, monthly
P2
Temperature
1km, monthly
P2

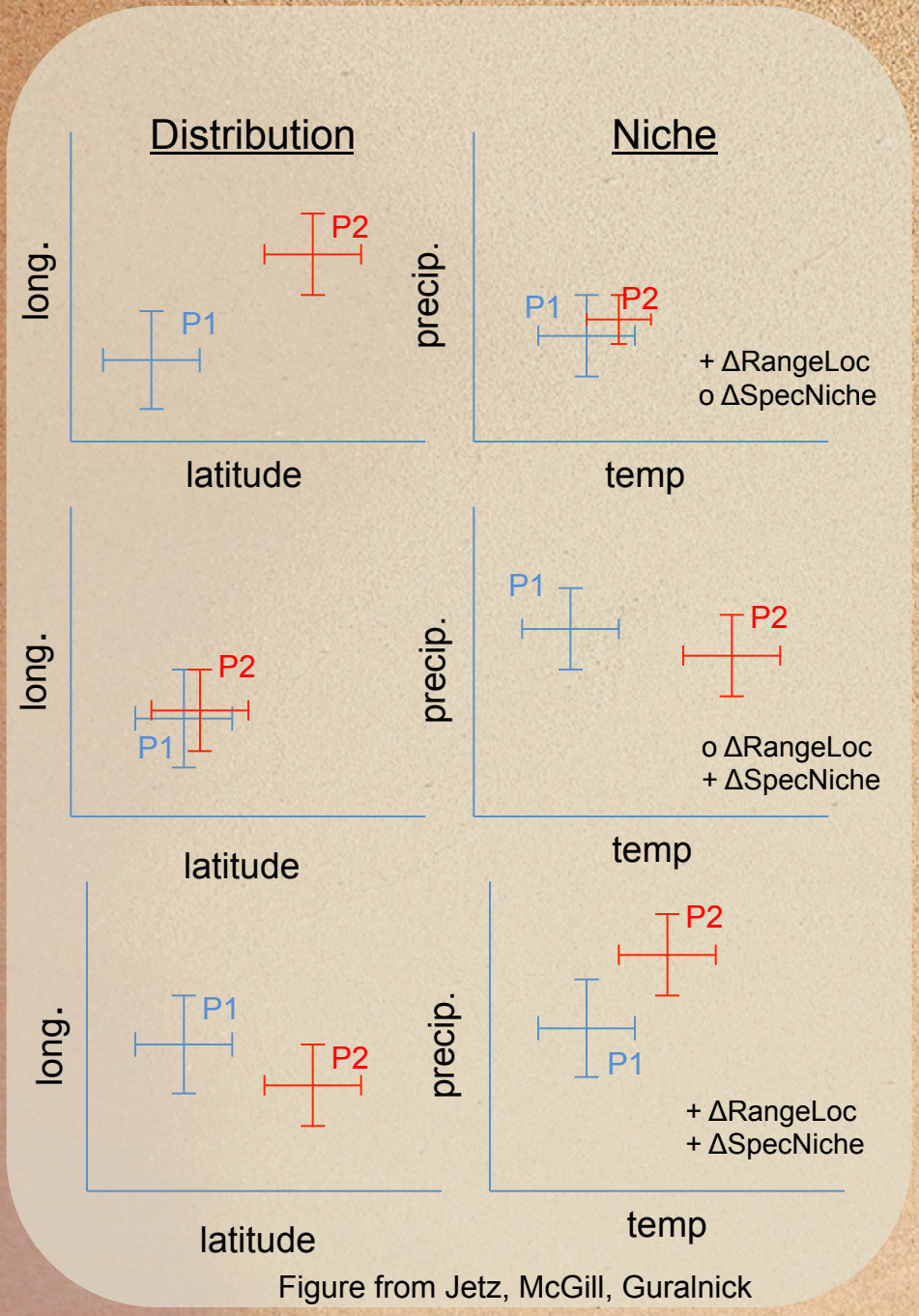


Figure from Jetz, McGill, Guralnick

Climate tracking

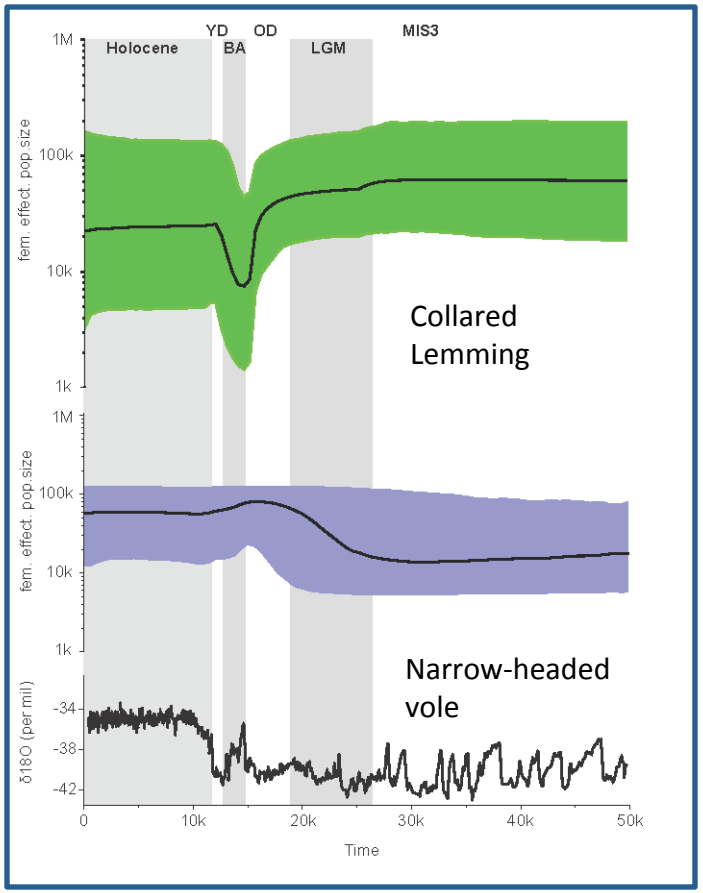
Niche shift

Mix of patterns

Ecological Tuned Environmental Data Layers

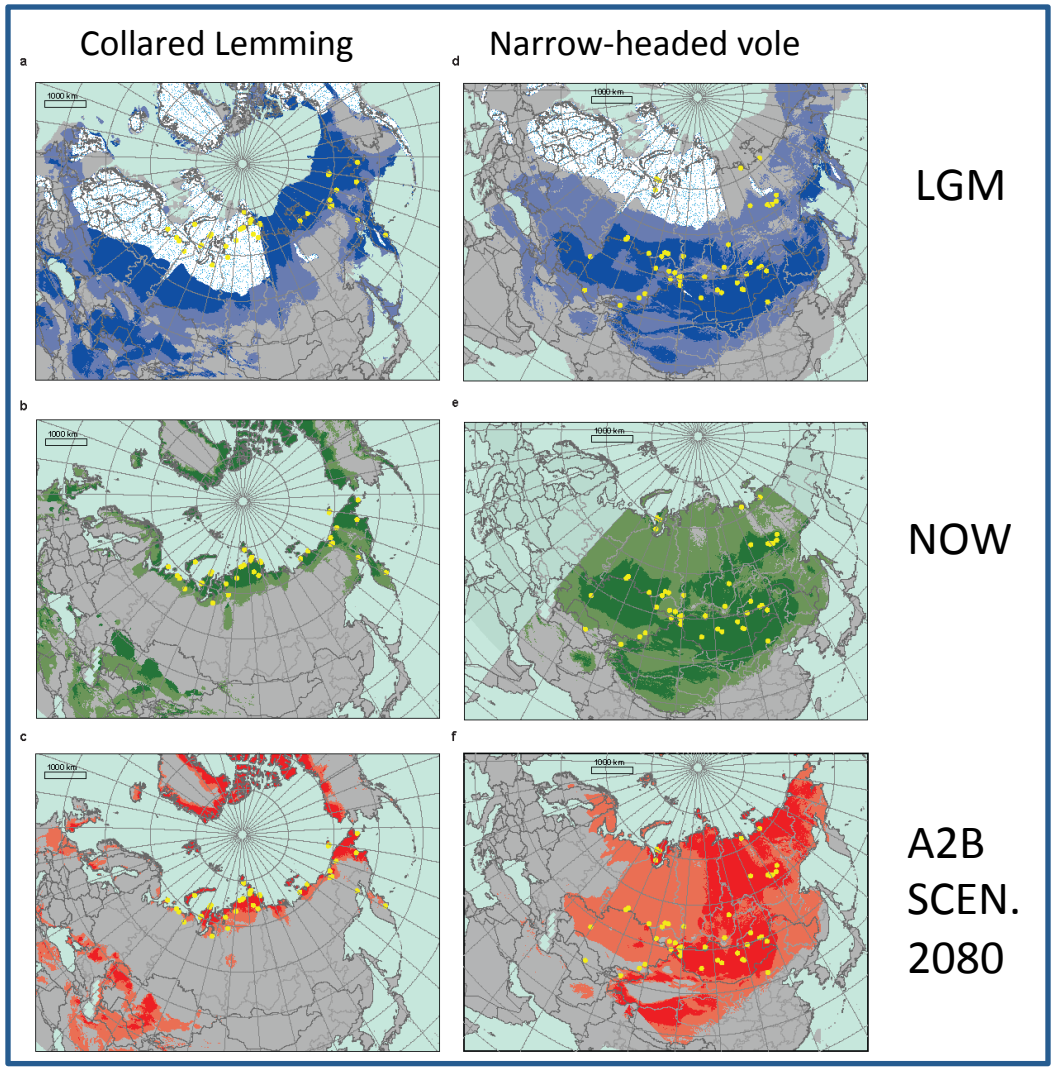
- **Fused climate 1km satellite & station for >20 yrs.**
  - **Bio-Ag variables** based on this dataset
    - Growing degree days, drought indices, etc.
- **Topography and hydrology 90m resolution**
- **Consensus landcover data layers over >20 yrs**
- **Layers coming online now**

METHODS CAN WORK OVER LONGER TIME SCALE AND INCLUDE OTHER DATA SUCH AS ANCIENT DNA: CONCERTED, SIGNIFICANT DEMOGRAPHIC BOTTLENECKS & DISTRIBUTION LOSS THAT MAY CALIBRATE FUTURE CHANGE FORECASTS.



Bayesian Skyline Plot (BSP) showing demographic histories of collared lemming (green) and narrow-skulled vole (blue) . Highest posterior densities (HPDs) shown in color; mean shown as a black line) and ^{18}Ocdeviations - proxied for temp on % (y-axis)

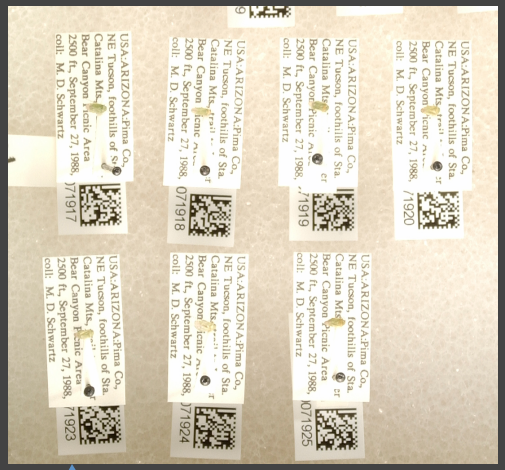
* Based on frags of CR and/or cytb for 33 NHV and 77 CL.



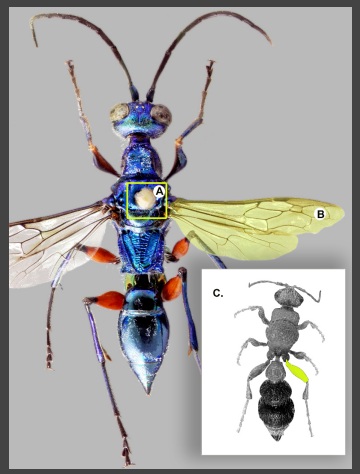
*lighter and darker shades represent more stringent and relaxed thresholds.

Prost, Guralnick, et al. 2013

Digitized specimen labels



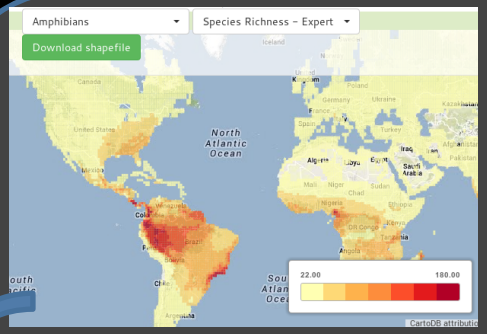
Images plus annotations (phenomics)



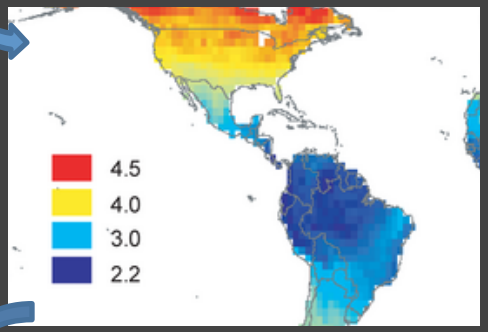
Direct species interactions

Genomics/transcriptomics

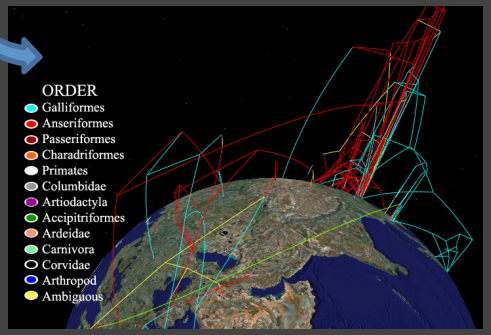
LINKED DATA FROM SPECIMENS



Richness patterns



Traits e.g. clutch Size (Jetz 2008)



avian flu evolution/ movement and hosts

GLOBAL INTEGRATED MAPS (in motion)

LINKED DATA FEEDING INTO LARGE SCALE ANALYSES:
WE ARE JUST STARTING DOWN THIS ROAD

It will be richer, more interesting, more integrative