SatRday
Chicago
2019

MISADVENTURES *IN* BIODIVERSITY DATA

Kate Webbink
Technology Department
Information Systems Specialist

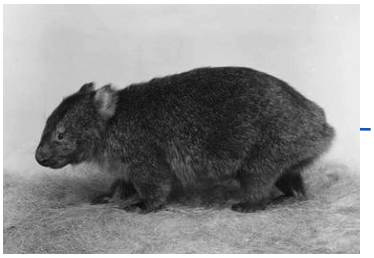
Want to share a few tools in the works at FMNH in the context of a large data cleanup project in preparation for some major database changes

---
- OSF / api / open data...?
- media file validation (needs work)

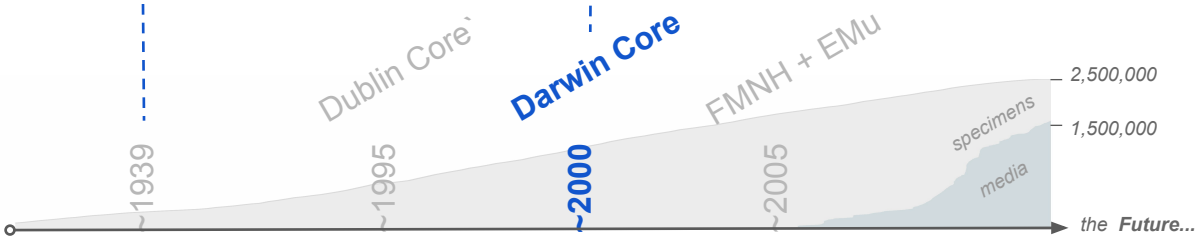
Shiny apps for: --

- Visualizing data cleanup / summaries
- admin tools / shiny app
- (registry & schema visualization w/ r2d3)
- setup/run shiny from github repo
- multimedia summaries…
- validity checks

Biodiversity Data ...Standards




```
recordedBy: "J. W. Woodhead",
country: "Australia",
decimalLatitude: "-42",
decimalLongitude: "147",
scientificName: "Vombatus ursinus hirsutus",
basisOfRecord: "PreservedSpecimen",
occurrenceID: "8864ef77-15ff-4ea8-bd97-f8bc396d7c8f",
catalogNumber: "49085",
collectionID: "Mammals",
year: "1939"
```

We have lots of data about NH collections stored in a system called “EMu” -- relational database

For a catalog record of a given specimen [say wombat collected in 1939), we attach
- A “Party” (person) record for who collected it
- A “Site” record for where collected
- A “Taxonomy record for what it is

Around 2000 - Darwin Core data standard was proposed, & lots of collections started publishing their data in this format

FMNH has about a century’s worth of records to deal with -- 2.5mil specimen records (of who collected what/where/when)
increasing amounts of multimedia now, too

While our biodiversity data primarily [ideally] takes the form of occurrence records...

- we have a lot of messes to clean up
  - just being able to see data/set/base structure helps a lot;
    - seeing 'where messes & missing pieces are (visdat)

Biodiversity data
- “Occurrence” records of
- published/maintained by museums/institutions/researchers with natural history

- collections/observations

An ideal occurrence record

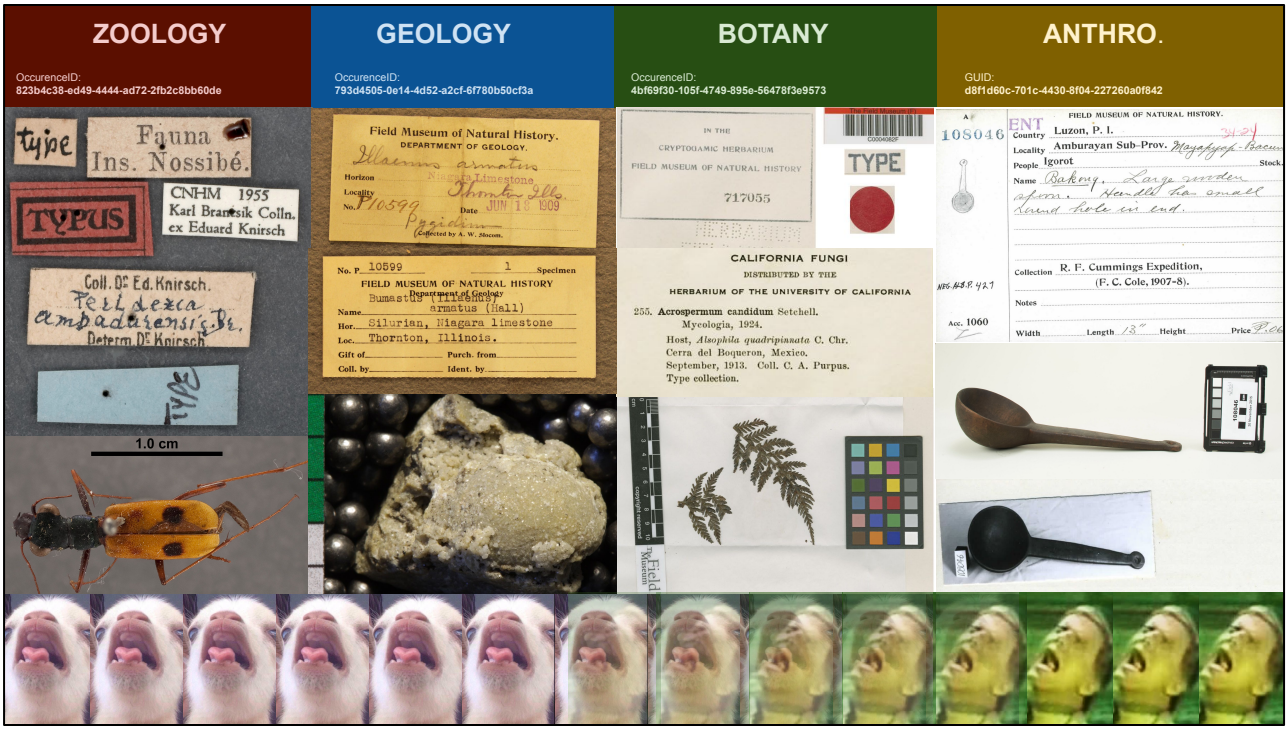
We & other museums are going through cleanup-crises lately -- not many institutions have high quality clean data (as reported by gbif & idigbio)
- show gbif summaries?
- show fmnh data summaries (+ ipt link)

Biodiversity data standards are a response to 100s of years of messy data

TDWG, GBIF, iDigBio, etc

https://www.gbif.org/occurrence/665900392

Add EML, ABCD, FAIRSHARING...
http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language



...While our biodiversity data primarily [ideally] takes the form of occurrence records

Realistically those datasets are collected using a variety of different research traditions

- Many data structures
  - Many different names for similar structures
  - Over years, database ended up with some unnecessarily duplicated fields

Within a record (zoology-insects is already many collections), this is hard enough over time

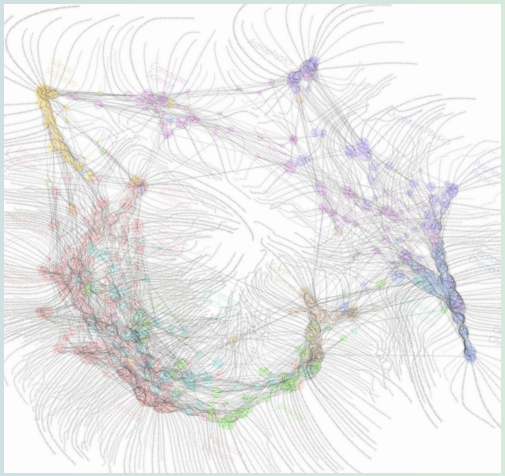
Within a whole collection...across collections...complicated

- What matters most differs across collections
  - e.g., Priorities in biology vs geology vs anthropology
  - Same names for different fields; different names for same fields
  - e.g., Priorities in cultural vs natural history [vs phys/geo sci]

A small chunk of chaos-nightmare
A cry for help
It's peeople

Diversity...

...is great!



© W B Paley - Relationships among Scientific Paradigms

But maybe not for data-structure.

HomeStarRunner.com / sbemail84.html

...

How to

build + maintain

ONE SYSTEM

for

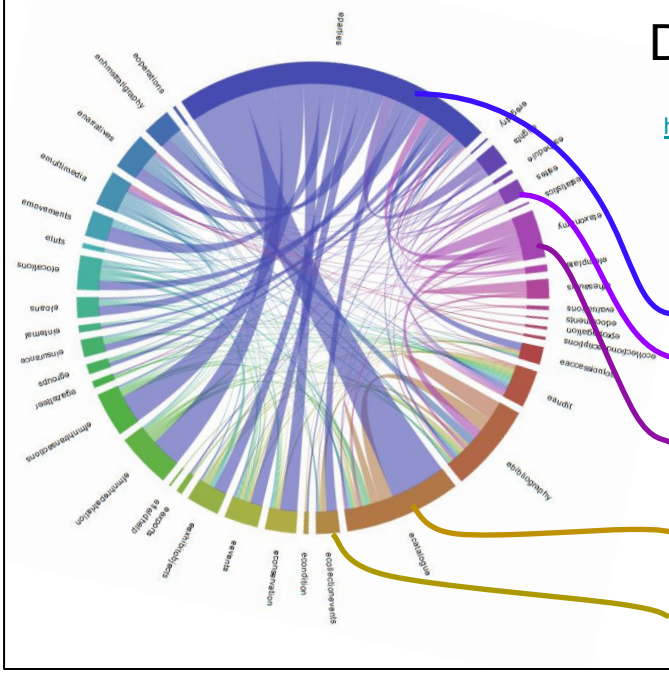
DIVERSE DATA?

How monitor biodiversity database in the throes of standardization?

- Changes happening to the data
- Changes happening to the database structure
    - Where can fields be deduplicated?
    - To more efficiently accommodate different priorities [for different users/producers of data]

Diversity is beautiful!
BUT this kind can be messy

- wheels get reinvented
- info goes missing between systems/migrations/fires
- SO...we need translators, migrators,



Database structure

https://github.com/**fieldmuseum/Collections-Scripts/**

+ **ShinySchemaVis**
+ **EMu database schema**

```
recordedBy: "J. W. Woodhead",

country: "Australia",
decimalLatitude: "-42",
decimalLongitude: "147",

scientificName: "Vombatus ursinus hirsutus",

basisOfRecord: "PreservedSpecimen",
occurrenceID: "8864ef77-15ff-4ea8-bd97-f8bc39",
catalogNumber: "49085",
collectionID: "Mammals",

year: "1939"
```

How can a database **accommodate for variety** without harboring chaos?
What are good tools for showing/assessing this?

How is/should be a “standardized” natural history record actually shaped?
(How can/should it accommodate for cultural history?)

Are these sunburst diagrams light at the end of the tunnel/rabbit hole?

- No. They are indeed oncoming trains
- But they at least help us see...until we’re dead/pancakes...

networkD3::chordNetwork()

EMu Schema
text file

prep-script
tidyr, zoo

matrix
igraph::
graph.data.frame()

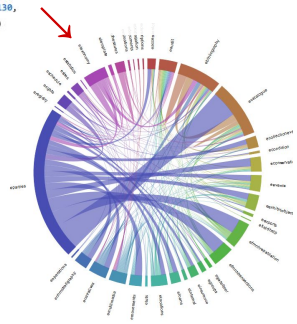
networkD3::**chordNetwork()**

```
# Setup Link-table of Source/Target nodes
schemaB <- schema[!schema$type=="table", c("source", "target")]
schemaB <- schemaB[order(schemaB$source, schemaB$target),]

schemaBlevels <- unique(append(schemaB$source, schemaB$target))
schemaBlevels <- as.factor(schemaBlevels[order(schemaBlevels)])

schemaB$sourceNum <- factor(schemaB$source,
                            levels = schemaBlevels)
schemaB$targetNum <- factor(schemaB$target,
                            levels = schemaBlevels)

schemaB$sourceNum <- as.numeric(schemaB$sourceNum) - 1
schemaB$targetNum <- as.numeric(schemaB$targetNum) - 1


# Show # of attachments b/t tables as line width
schemaBlinks <- dplyr::count(schemaB, sourceNum, targetNum)
schemaBnodes <- data.frame("id" = as.numeric(schemaBlevels) - 1,
                           "name" = as.character(schemaBlevels),
                           stringsAsFactors = FALSE)

schemaB2 <- dplyr::count(schemaB, source, target)
g <- graph.data.frame(schemaB2, directed = FALSE)
schemaMatrix <- get.adjacency(g, attr = "n", sparse = FALSE)
```

```
schemaB2 <- dplyr::count(schemaB,
                         source,
                         target)

g <- graph.data.frame(schemaB2,
                      directed = FALSE)

schemaMatrix <- get.adjacency(g,
                              attr = "n",
                              sparse = FALSE)
```

	eaccessionlots	eaudit	ebibl
eaccessionlots	0	1	
eaudit	1	1	
ebibliography	0	1	
ecatalogue	1	1	
ecollectionevents	1	1	
econdition	0	1	
econservation	0	1	
eevents	0	1	
eexhibitobjects	1	1	
eexports	0	1	
efieldhelp	0	1	

```
chordNetwork(Data = schemaMatrix,
             height = 850, width = 800,
             initialOpacity = 0.6,
             colourScale = col2hex(palette(rainbow(NROW(colnames(sch
             # colourScale = input$colorSchema,  # shemaRainbow,
             fontSize = 10,
             padding = 0.035,
             labels = colnames(schemaMatrix),
             labelDistance = 130,
             pdf(file = NULL))
```



1- Schema gives list of defined fields. (Want to butcher it for a list of attachment fields)
2 - count # reference fields
3 - convert data.frame to matrix
4 - plug into chordNetwoek()

Registry / Admin Patterns - `sunburstR::sunburst()`

EMu Registry
CSV

→

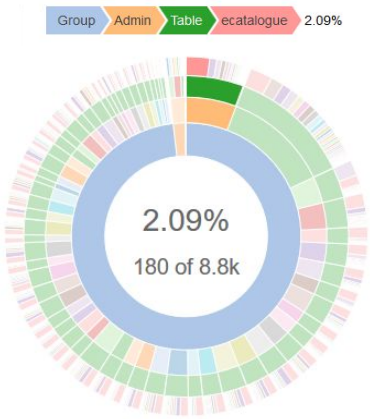
prep-script
dplyr

→

```
sunburst(regD3, count = TRUE)
```

↓

Key1	Key2	Key3	Key4	Key5
Group	Default	Table	Default	Operations
Group	Default	Table	ecatalog	Tabs
Group	Admin	Table	eevents	Tabs
Group	IZMgr	Table	esites	Report
User	emu	Table	esites	Tabs



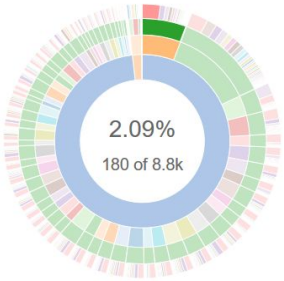
How do users actually enter/create natural history records? (Does it follow standards?)

(How can/should it accommodate for cultural history?)

Are these sunburst diagrams light at the end of the tunnel/rabbit hole?

- No. They are indeed oncoming trains
- But they at least help us see...until we're dead/pancakes...

Light at the end of a tunnel…



...Or an oncoming train?

How is a “standardized” natural history record actually shaped?
(How should it be?)
(How can/should it accommodate for cultural history?)

Are these sunburst diagrams light at the end of the tunnel/rabbit hole?

- No. They are indeed oncoming trains
- But they at least help us see...until we’re dead/pancakes...

https://wirralleaks.wordpress.com/2013/12/03/wirral-leaks-advent-calendar-day-3-light-at-the-end-of-the-tunnel/
https://wirralleaks.files.wordpress.com/2013/12/oncoming_train_shutterstock_87110158.jpg

http://www.autismafter16.com/sites/default/files/imagecache/article_large/article-images/Train%20Tunnel.jpg

Cited Things:

- **networkD3**
- **sunburstR**

- github.com/**fieldmuseum**/Collections-Scripts
  - Shiny**Schema**Vis
  - Shiny**Reg**Vis

- **Darwin Core**
  TDWG - https://dwc.tdwg.org/

- Science Map / **W. B. Paley**
- Children's book / **Homestar Runner**
- Screamy lamb / twitter.com/screaminglxmb
- Charlton Heston / Soylent Green

Acknowledgements:

FMNH
Technology Department

Sharon Grant
Janeen Jones
Pete Herbst
Rob Zschernitz

Action Department

Tomomi Suwa

sunburstR::sunburst()

EMu Registry
CSV

```
Key1    Key2    Key3    Key4    Key5
Group  Default Table  Default  Operations
Group  Default Table  ecatalog  Tabs
Group   Admin  Table  eevents   Tabs
```

→

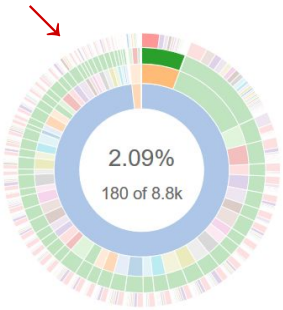
prep-script
dplyr

```
registry$Keys <- paste(registry$Key1,
                       registry$Key2,
                       registry$Key3,
                       registry$Key4,
                       sep = "-")

regD3 <- dplyr::count(registry, Keys)

regD3 <- regD3[order(regD3$n, decreasing = TRUE),]

add_shiny(sunburst(regD3, count = TRUE))
```

→

sunburstR::**sunburst()**

```
sunburst(regD3, count = TRUE)
```



Creating a matrix

Code & stuff & fun with chordNetwork()

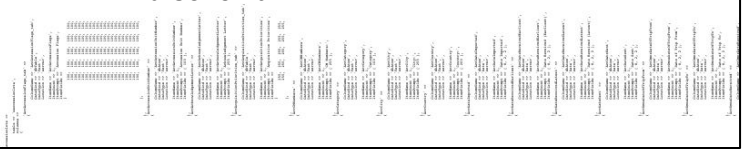
```
# shift & fill rows for main table names
schema$source <- NA
schema$source[schema$type=="table"] <- schema$target[schema$type=="table"]

schema$source <- gsub("^\\s+|\\s+$", "", schema$source)
schema$target <- gsub("^\\s+|\\s+$", "", schema$target)

schema$source <- na.locf(schema$source)


# Setup Link-table of Source/Target nodes
schemaB <- schema[!schema$type=="table", c("source", "target")]
schemaB <- schemaB[order(schemaB$source, schemaB$target),]

schemaBlevels <- unique(append(schemaB$source, schemaB$target))
schemaBlevels <- as.factor(schemaBlevels[order(schemaBlevels)])

schemaB$sourceNum <- factor(schemaB$source,
                            levels = schemaBlevels)
schemaB$targetNum <- factor(schemaB$target,
                            levels = schemaBlevels)

schemaB$sourceNum <- as.numeric(schemaB$sourceNum) - 1
schemaB$targetNum <- as.numeric(schemaB$targetNum) - 1
```

```
chordNetwork(Data = schemaMatrix,
height = 850, width = 800,
initialOpacity = 0.6,
colourScale = col2hex(palette(rainbow(NROW(colnames(schemaMatrix)), s = 0.6, v =
# colourScale = input$colorSchema,  # shemaRainbow,
fontSize = 10,
padding = 0.035,
labels = colnames(schemaMatrix),
labelDistance = 130,
pdf(file = NULL))
```

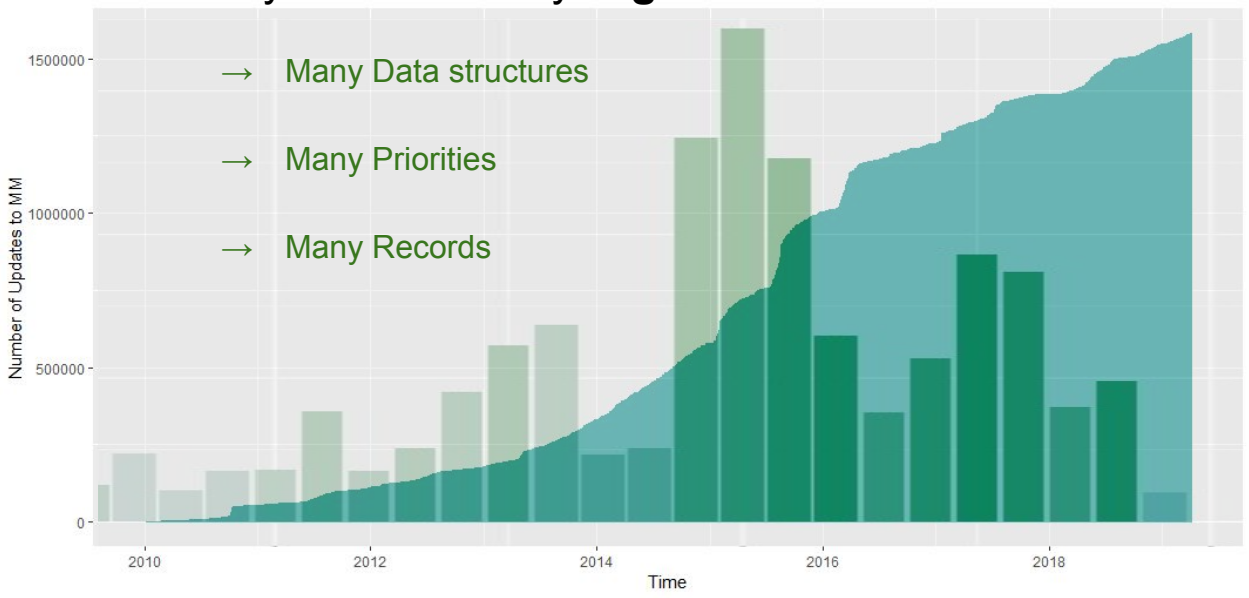
https://github.com/**fieldmuseum/Collections-Scripts/**

- **ShinySchemaVis**

\+ **EMu schema…**



Brief history of Biodiversity **Digitization**



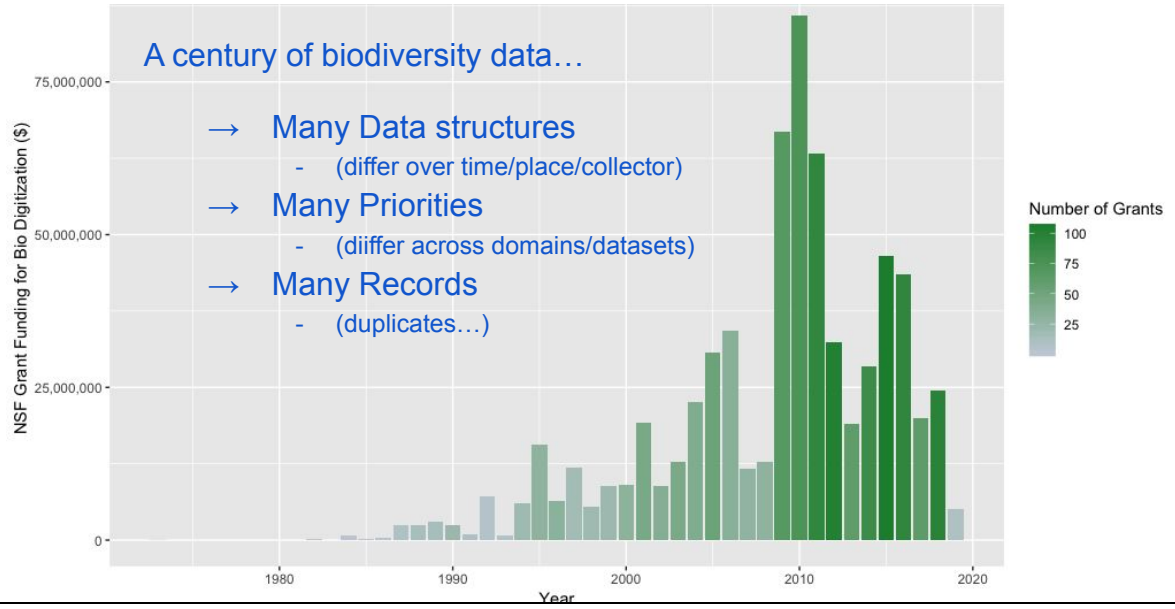
A few **decades** of biodiversity **digitization**

→ Many Data structures
- (differ over time/place/collector)

→ Many Priorities
- (differ across domains/datasets)

→ Many Records
- (duplicates…)

1895[ish] through 2019…



Brief history of **Biodiversity** data & the **FMNH** data-verse

~1995 - **Dublin Core** data schema *(for literary/published info - books, images, digital stuff)*

[some happy lambs → some screaming lambs …]

~2000 - **Darwin Core** data schema *(for biodiversity info - species/specimen 'occurrences')*

[some screaming lambs → more screaming lambs …]

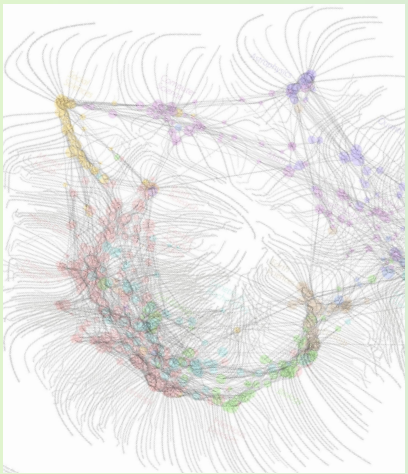
~2005 - FMNH starts merging collections data *(botany, zoology, geology, anthropology)*

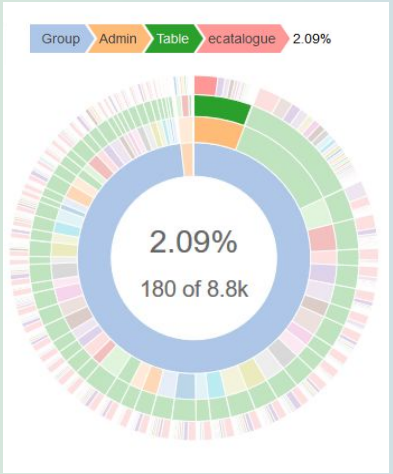
[more screaming lambs → moooaar screaming lambs …]

~2019 - Here we are today...

Priorities

Diversity is great! ...BUT... Maybe not for data-structure... SO...How to accommodate?





© W B Paley - Relationships among Scientific Paradigms

HomeStarRunner.com / sbemail84.html

How monitor biodiversity database in the throes of standardization?

- Changes happening to the data
- Changes happening to the database structure

Different priorities [for different users/producers of data]

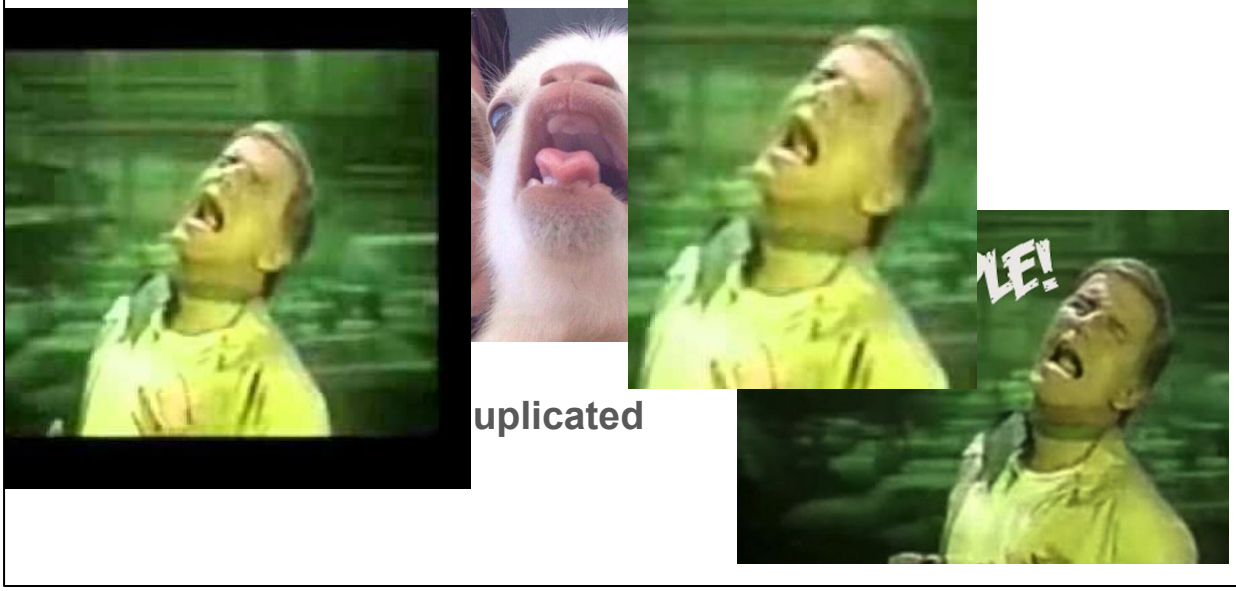
- What matters… many standards
  - e.g., Priorities in biology vs geology vs anthropology
  - sometimes info goes missing between systems/migrations/fires
- What matters… many standards
- e.g., Priorities in cultural vs natural history [vs phys/geo sci]

Diversity is beautiful!

BUT this kind can be messy

- wheels get reinvented
  - info goes missing between systems/migrations/fires
- SO...we need translators, migrators,

Records themselves...



Records themselves...

- ...go missing
  - in the wake of disaster & into the mystery-void between systems
  - but without cross-checking between systems…
- ...get re-re-dup-duplicated
  - We don’t always look before we leap
  - Workflows get in the way
  - (e.g., checking for dups by name...after renaming dups)