Towards user-definable, semi-automated workflows for curating biodiversity data

P.J. Morris R.A. Morris B. Ludäscher
D. Lowery J.A. Macklin
T. Song T. McPhillips J. Hanken

MUSEUM OF COMPARATIVE ZOOLOGY
HARVARD UNIVERSITY

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

UCDAVIS
UNIVERSITY OF CALIFORNIA

Agriculture and Agri-Food Canada

Webinar Outline

- Overview (Bertram)
- **Part 1: Presentation** [45 min]
  - 1.1 Example QC Spreadsheet & FilteredPush (Bob)
  - 1.2 Data Cleaning for Natural History Collections (Paul)
  - 1.3 Intro to current FP-Akka/post-proc. tools (Paul)
  - 1.4 Scientific Workflow Automation (Bertram)
- Q&A/Transition/Software check (David, Bob, Paul)
- **Part 2: Demo & Hands-on** (optional) [55 min]
  - Run FP-Akka workflow –w DwCa –a COL
  - …

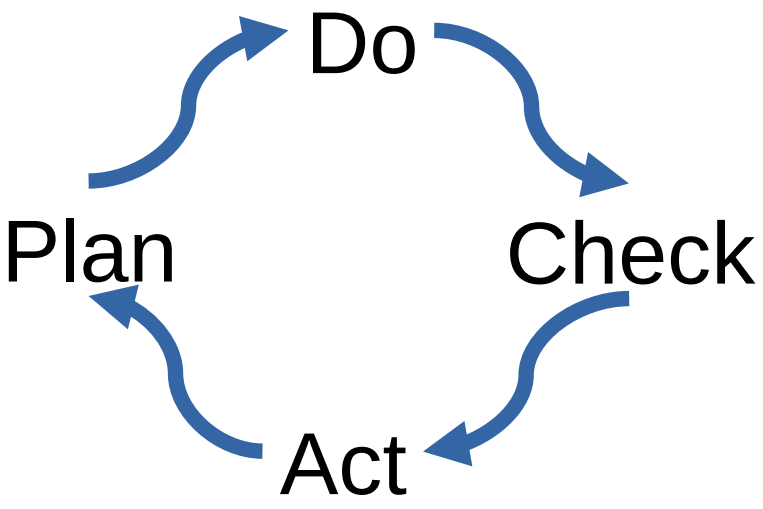
1.1 Example QC Spreadsheet

Bob (~ 10 min)

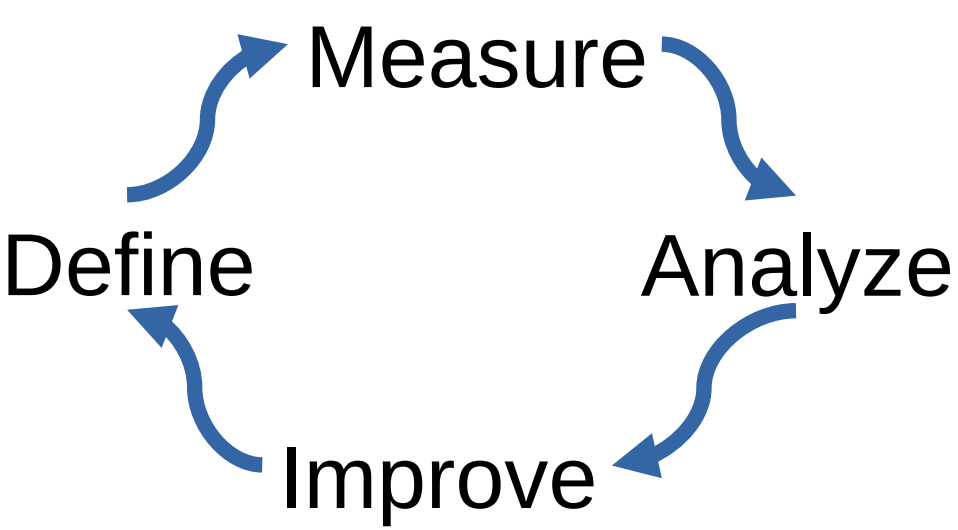
1.2 Data Cleaning for Natural History Collections

Paul (~ 15min)

Continuous Quality Control



Classical QC
Shewhart, 1939



Total Data Quality
Management
Wang, 1998

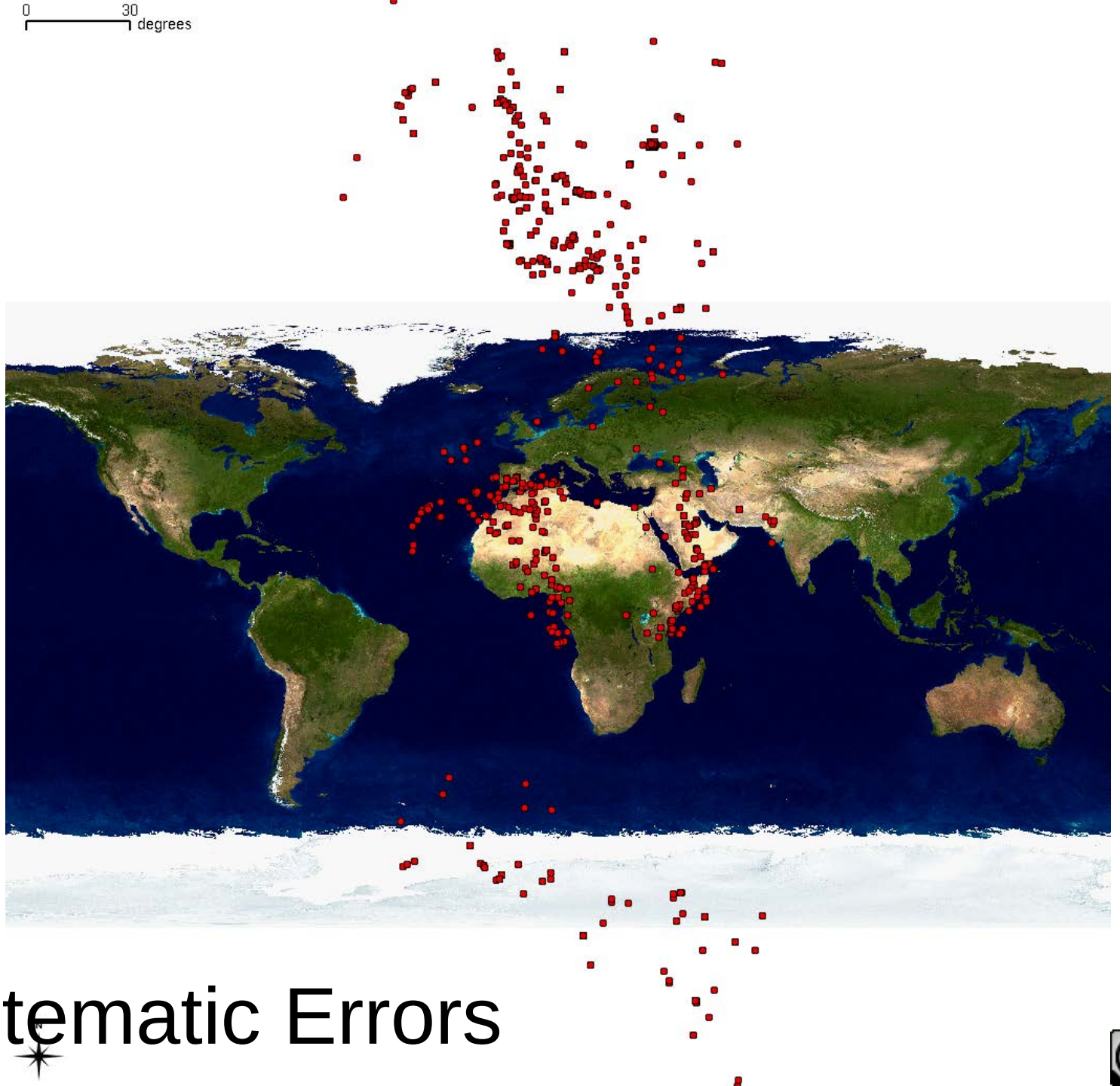
With tests for Fitness for Purpose

CC BY SA

Natural Science Collections Data
Fit for what purpose?

- Classical Use by Taxonomists visiting collections, and collections managers.
- Electronic Distribution: Many new uses
  - Species range modeling
  - Modeling effects of climate change
  - **Many new uses depend on**
    - **(1) Good georeferences (where).**
    - **(2) Good identifications (what taxon)**
    - **(3) Good collecting event dates (when)**

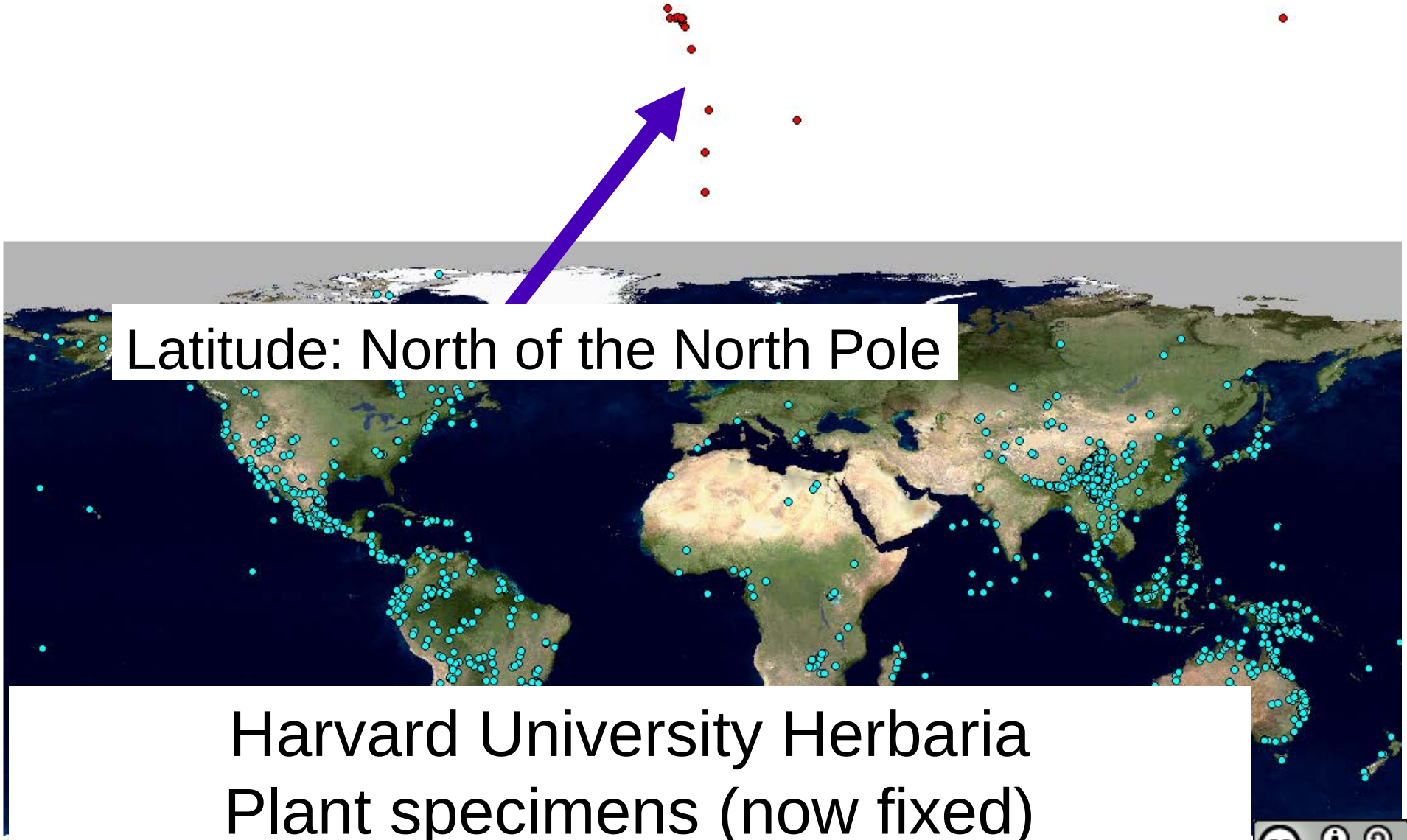
CC BY SA



Systematic Errors

cc BY SA

Illogical Values: Out of Range

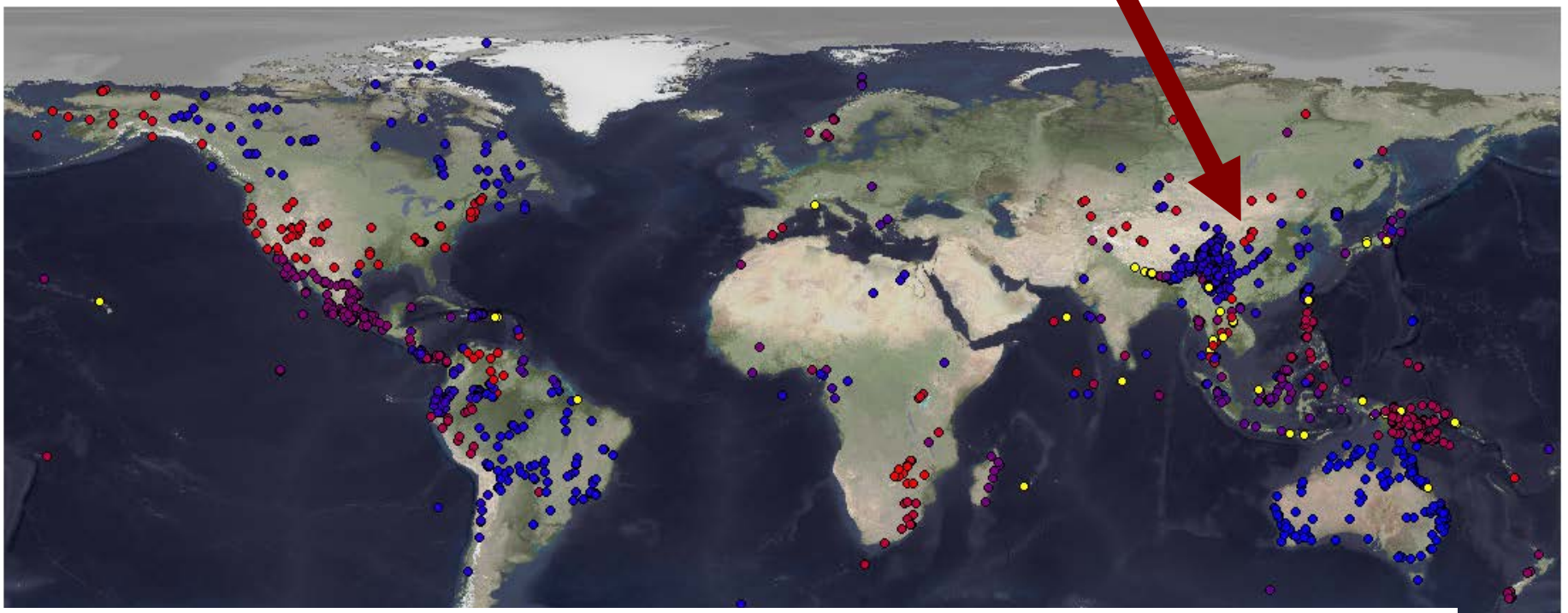


Harvard University Herbaria
Plant specimens (now fixed)

cc BY SA

Internal Inconsistency

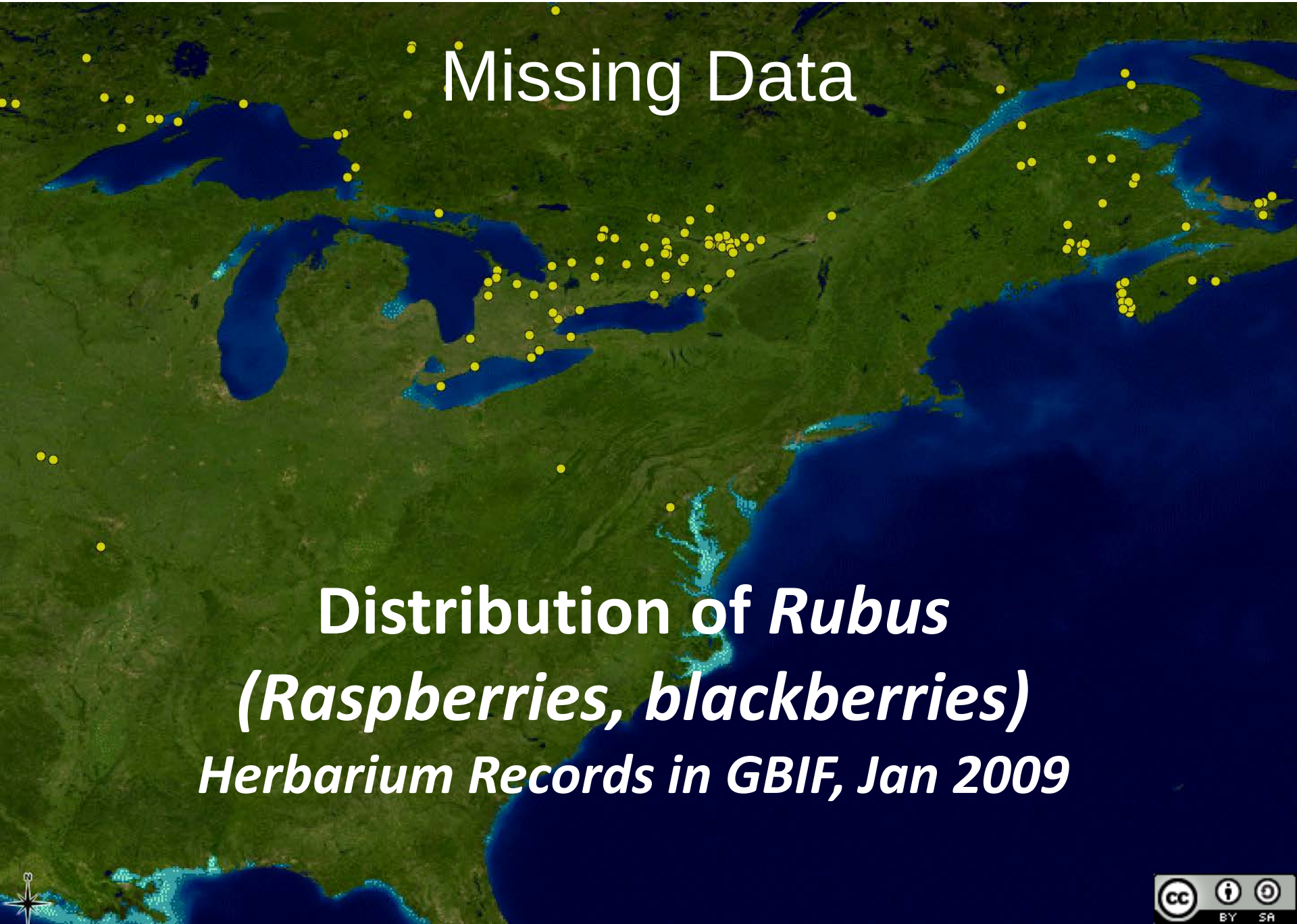
Latitude and Longitude: Somewhere in China
Country: United States of America



Harvard University Herbaria
Plant specimens, color coded by country

cc BY SA

Missing Data



Minimal Data Capture
- Current ID
- Locality
- Collector
- Date Collected

Georeference

Plan

Do

Check

Act

Internal
- Sample
- Look for Outliers
- Completeness

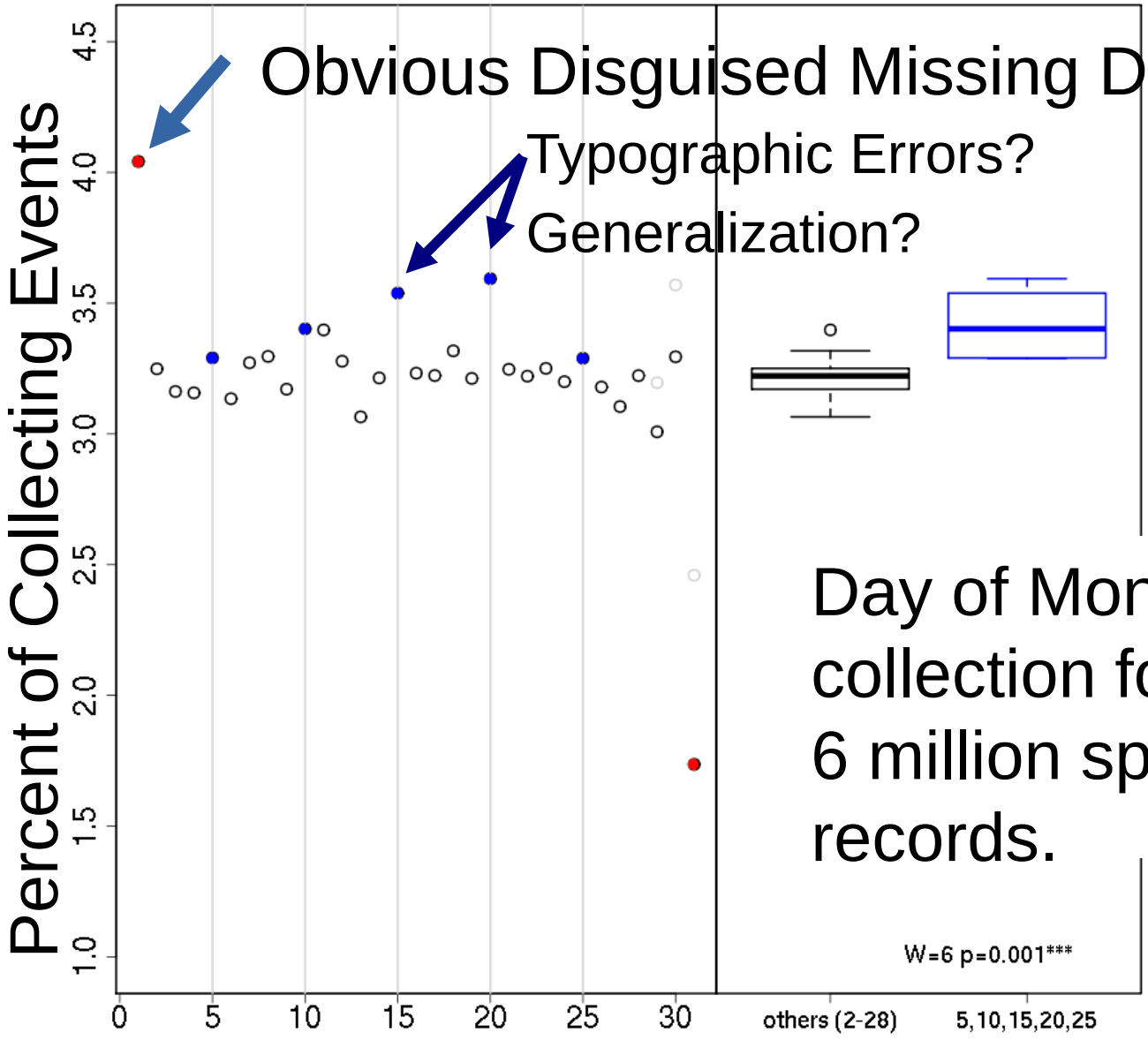
Clean Data With Data
- External Authorities
- Patterns
- Outliers

Measure

Fix issues found in data

Improve Process
- Training
- Improved Authority Files
- Improved User Interfaces

cc BY SA



Day of Month of collection for about 6 million specimen records.

Controlled Data Capture

Id: Formicidae:Myrmicinae:Pheidole

Label data: Verbatim Locality Text

Verbatim Locality: Causse de Kelifely; Madagascar

Elevation

Country: **Georeference**

State/Province:

Higher Geography

Specific Locality: **No Data**

Add another collector **Collector Name**

Verbatim Date:

ISO Date yyyy/mm/dd:

Date Collected:

Collection:

CollectingMethod:

Other Numbers

Notes

Specimen Notes: specimennotes

Features:

Habitat:

Microhabitat: microhabitat

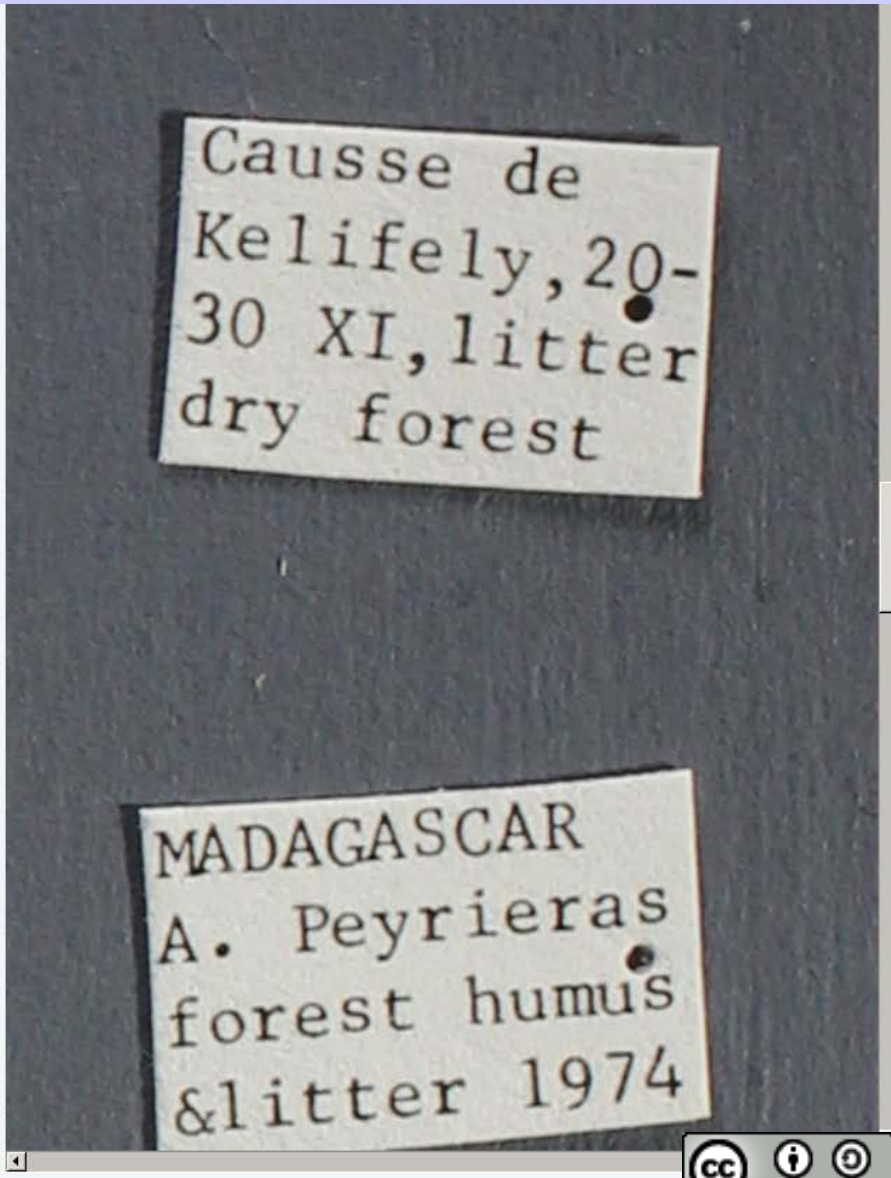
AssociatedTaxon:

ValidDistributionFlag: ☑

Part	Preparation	Count	Modifier
whole animal	pinned	1	

Attribute	Value	Remarks	Add Attributes
caste	worker		Edit

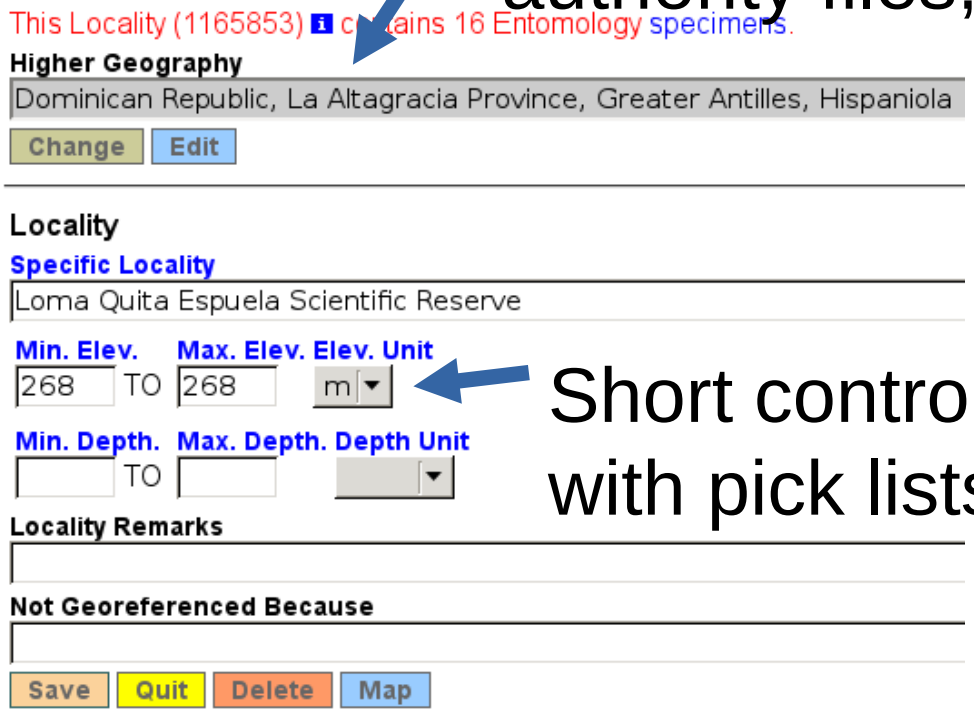
Add A Specimen Part



BY SA

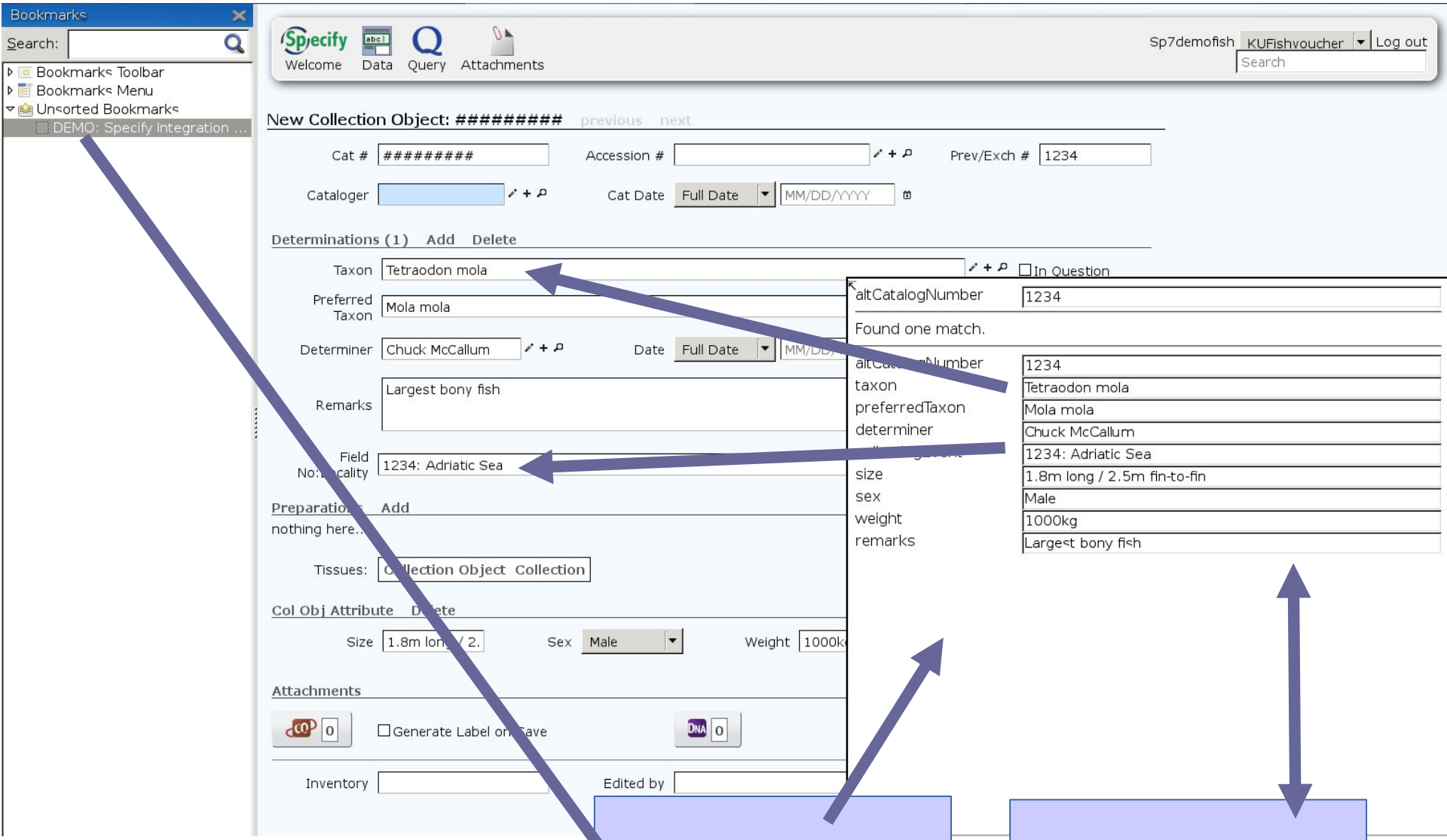
Controlled Data Capture

Large controlled vocabularies: Geographic, Taxonomic, etc. authority files, with lookups.

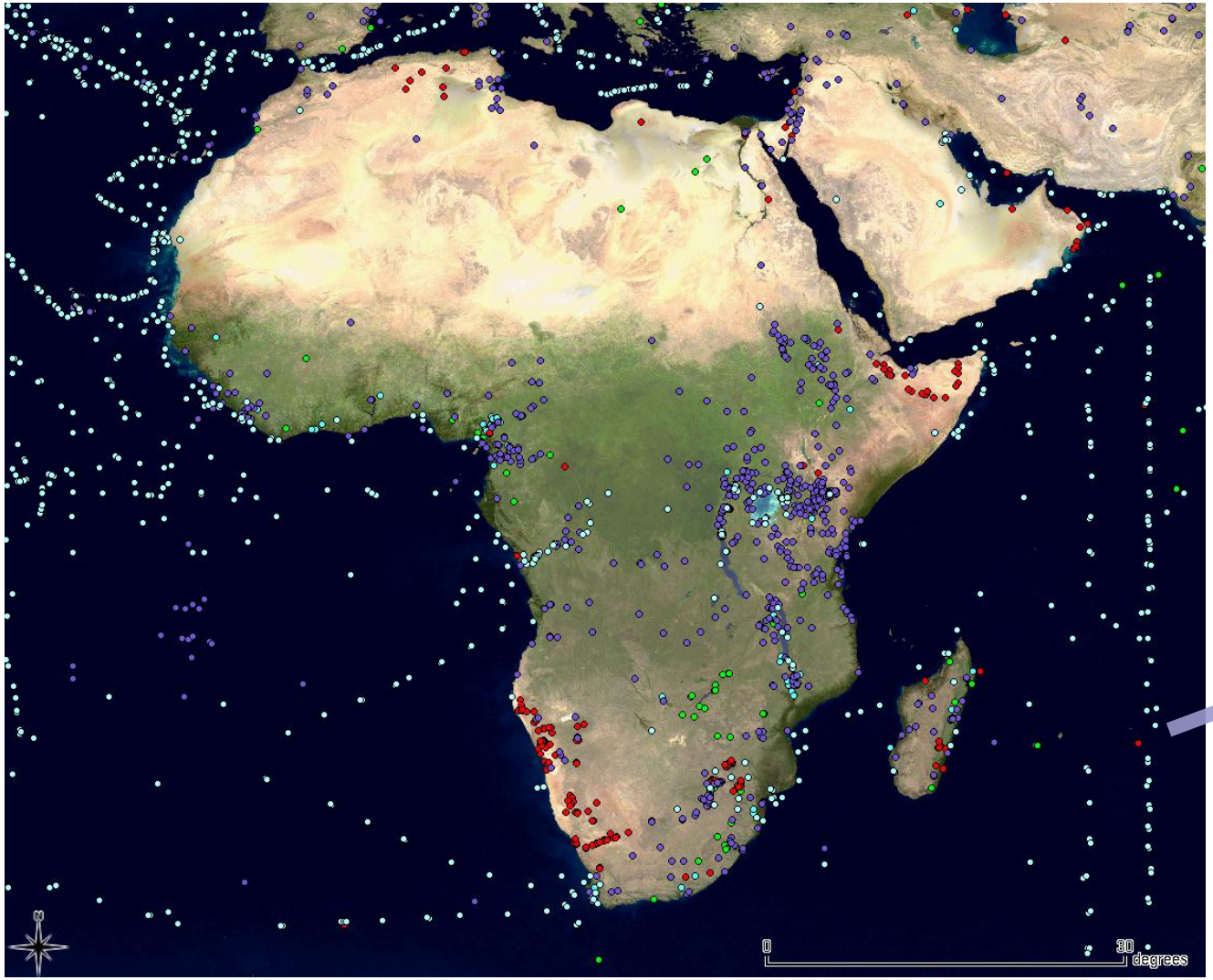


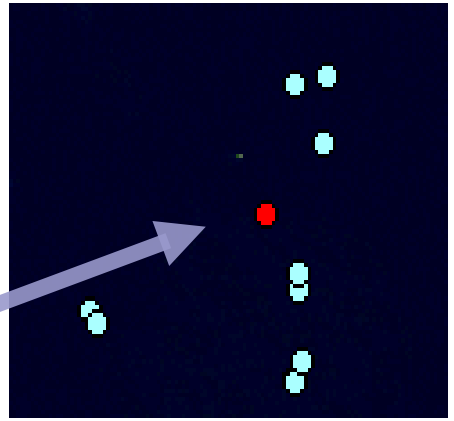
Short controlled vocabularies, with pick lists.

CC BY SA


Bookmarks
Search:
Bookmarks Toolbar
Bookmarks Menu
Unsorted Bookmarks
DEMO: Specify Integration ...
Specify
Welcome
Data
Query
Attachments
Sp7demofish
KUFishvoucher
Log out
Search
New Collection Object: #########
previous
next
Cat #
Accession #
Prev/Exch #
1234
Cataloger
Cat Date
Full Date
MM/DD/YYYY
Determinations (1)
Add
Delete
Taxon
Tetraodon mola
In Question
Preferred Taxon
Mola mola
Determiner
Chuck McCallum
Date
Remarks
Largest bony fish
Field No:
1234: Adriatic Sea
Preparations
Add
nothing here...
Tissues:
Collection Object
Collection
Col Obj Attribute
Delete
Size
Sex
Male
Weight
Attachments
Generate Label on Save
Inventory
Edited by
altCatalogNumber
1234
Found one match.
taxon
Tetraodon mola
preferredTaxon
Mola mola
determiner
Chuck McCallum
size
1.8m long / 2.5m fin-to-fin
sex
Male
weight
1000kg
remarks
Largest bony fish
FP-DataEntry
UI Service
FP-DataEntry
Query Service

Visualize



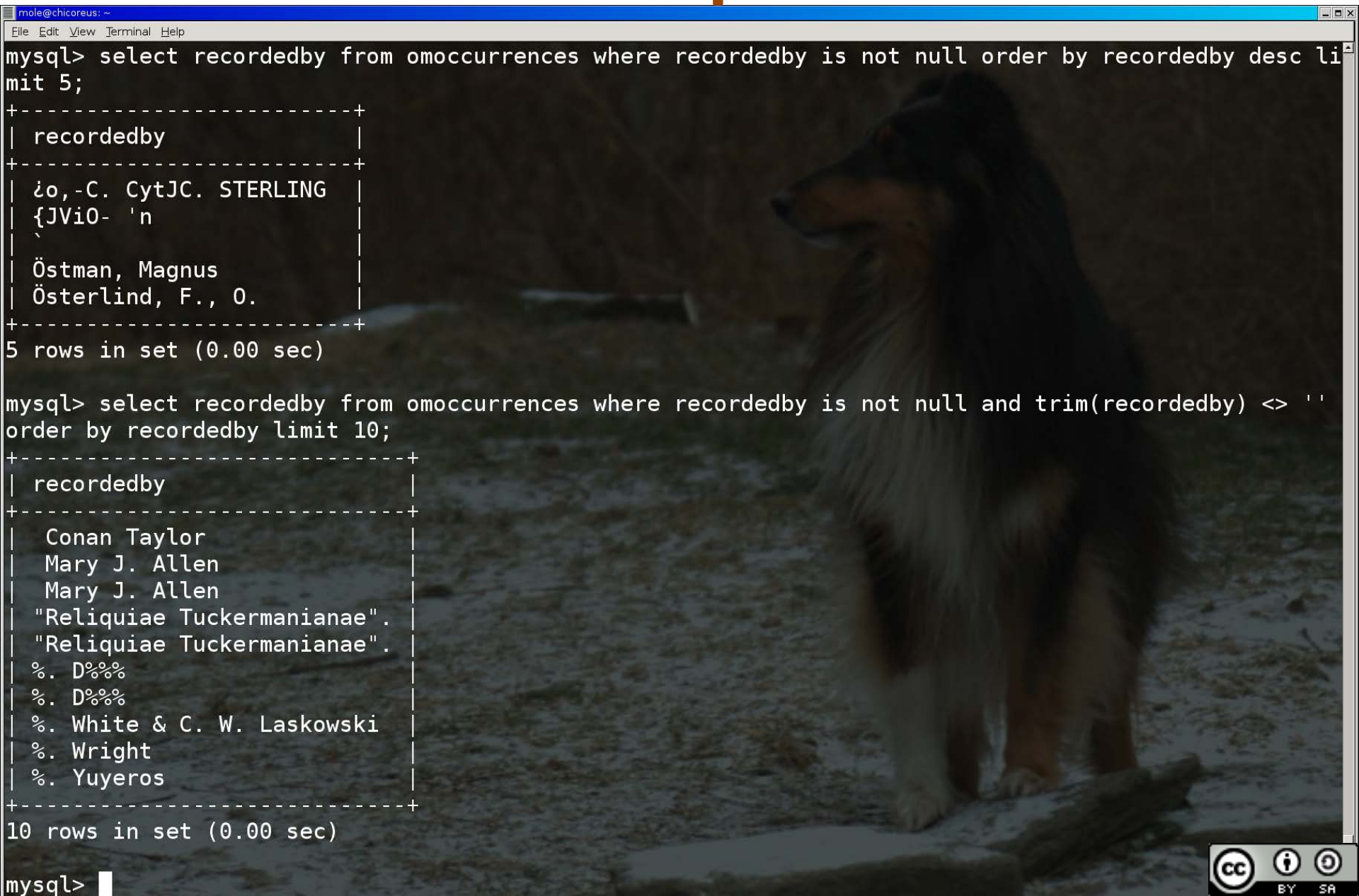


CC BY SA

1.3 Introduction: Current FP-Akka Workflows & Post-processing software

Paul (~ 5min)

Sort: Look at top and bottom



Examine rare values

```
•select count(*), country from omoccurrences group by country
order by count(*), country asc limit 10;
•+----------+-----------------------------------------+
•| count(*) | country                                 |
•+----------+-----------------------------------------+
•|        1 | 0                                       |
•|        1 | 1998                                    |
•|        1 | a                                       |
•|        1 | ABKHAZIA (GEORGIA)                      |
•|        1 | ABYSSINIA                               |
•|        1 | Africa Occidentalis [SÜo Tom_ and PrÕncipe] |
•|        1 | Alaska                                  |
•|        1 | ANDORRA                                 |
•|        1 | Arctic                                  |
•|        1 | AUSTIRA                                 |
•+----------+-----------------------------------------+
•10 rows in set (1.01 sec)
```

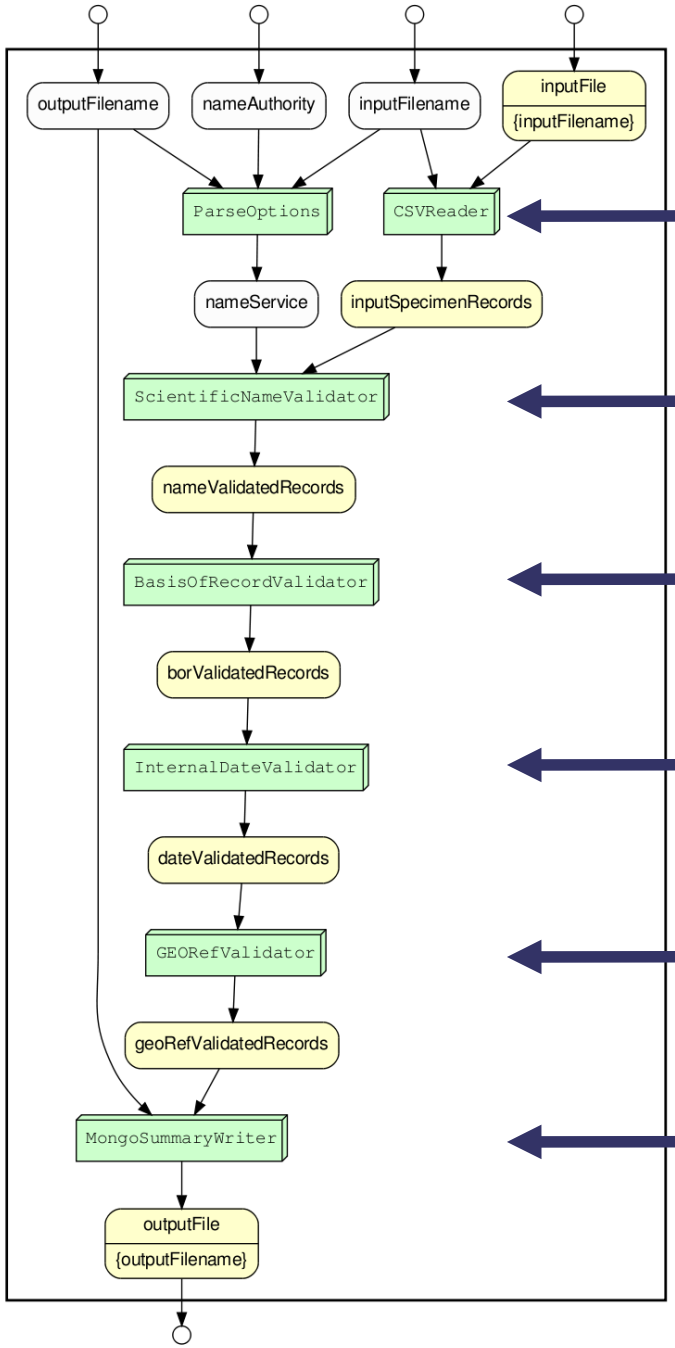
Look for patterns (in aggregated data)

recordedBy	fieldNumber	eventDate
S. Hammer	7540	1999-01-00
S. Hammer	7476	1999-01-00
S. Hammer	7279	1999-01-00
S. Hammer	8237	2000-01-00
S. Hammer	8321	2000-01-00
S. Hammer	8204	2000-01-00
S. Hammer	**7843**	2000-07-00
S. Hammer	**7851**	2000-07-00
S. Hammer	**7853**	2000-07-00
S. Hammer	**7849**	2000-07-00
S. Hammer	**7904**	2000-08-09
S. Hammer	**7909**	2000-08-09
S. Hammer	8002	2000-12-00
S. Hammer	8028	2000-12-00

Integer field number that gets larger over the life of the collector got smaller here.

CC BY SA

Data cleaning workflow



Load Data

IPNI, IF, WoRMS COL,GBIF

Check scientific name

GNI

GBIF

Check basisOfRecord

Check date collected

HUH Botanists

SCAN Entomo.

Check lat/long

Geolocate

Write out results

| sign changed coordinates are on the Earth's surface. | Coordinates not inside country. | transposed/sign changed coordinates to place inside the provided Country UNITED STATES | Transposed/sign changed coordinates are near (within 200.0 km) georeference of locality from the Geolocate service.

CC BY SA

Check names against nomenclators and taxonomic authority sources

Placopecten magellanicus WAS: Gmelin, 1791; CHANGED TO: (Gmelin, 1791)

| can't construct sciName from atomic fields | Found accepted name Placopecten magellanicus in Catalog of Life service. | Authorship: Differ only in Parentheses Similarity: 0.8333333333333334

CC BY SA

Minimal Data Capture
- Current ID
- Locality
- Collector
- Date Collected

Georeference

Plan

Do

Check

Act

Internal
- Sample
- Look for Outliers
- Completeness

Clean Data With Data
- External Authorities
- Patterns
- Outliers

Measure

Fix issues found in data

Improve Process
- Training
- Improved Authority Files
- Improved User Interfaces

cc BY SA

1.4 Scientific Workflow Automation:

- FilteredPush curation workflows
- Towards Kurator/P

Bertram (~ 10min)

What problems are we trying to solve?

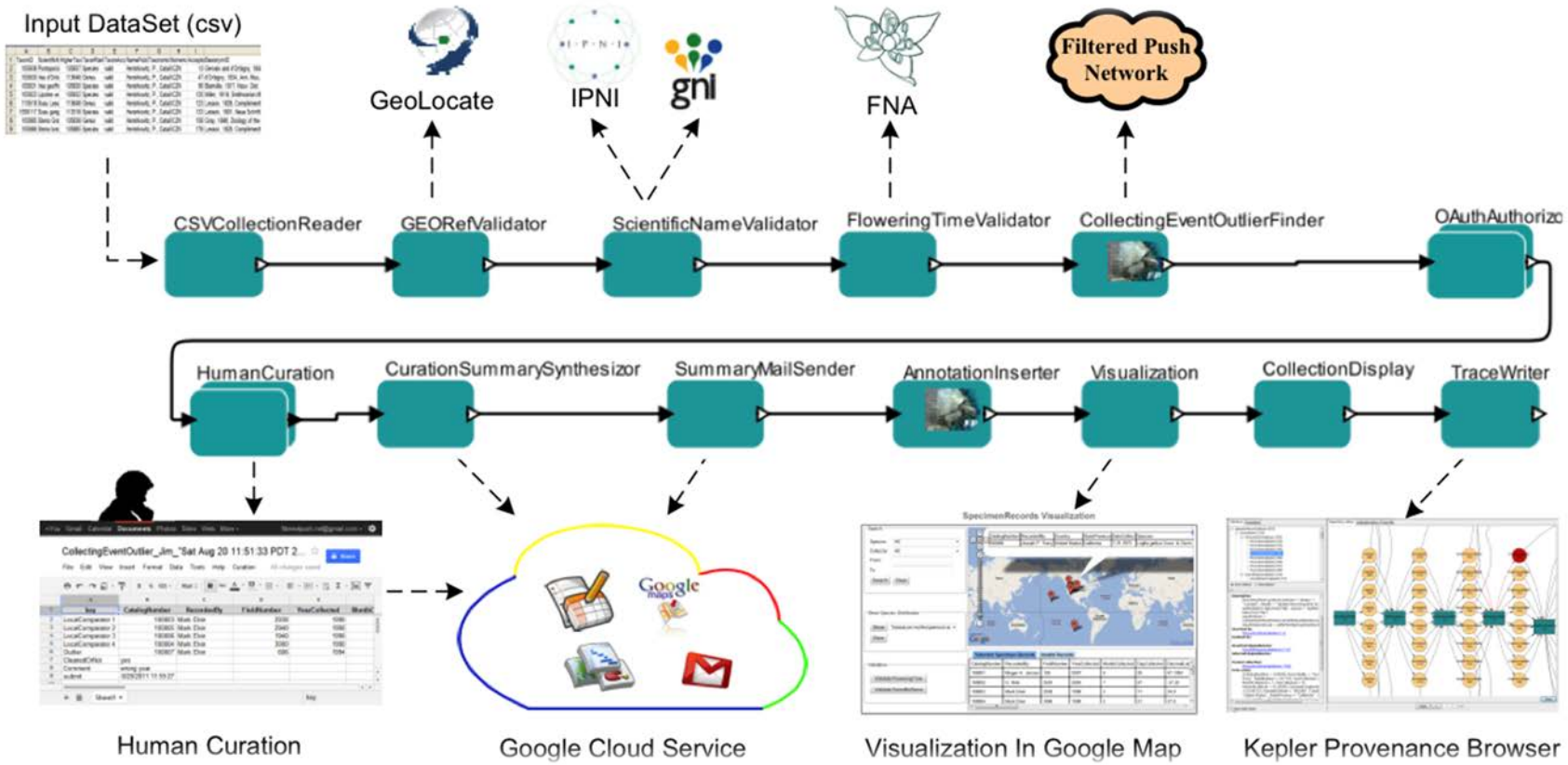
- **Detect** and **flag** **data quality issues**
- **Repair** if possible
  - … ask **human curators** as needed
- Keep track of **provenance**
  - (semi-)automatic repairs
  - human curators' edits

- Employ **workflow (semi-)automation**
  - **Scientific workflow systems:**
    - **Kepler**/COMAD, **Restflow**, **Galaxy**, **Biovel/Taverna**, **Argo**, **VisTrails**, …
  - **Related technologies**
    - **Akka** parallel execution platform
    - **Script-based** automation (e.g. **Python, R**), digital notebooks (**iPython**)

Curation Workflows Users

- **Collection Managers**
  - … who are managing the collections databases
  - Can run curation workflows periodically
    - … in the presence of new data and/or new curation services
- (Biodiversity) **Researchers**
  - To perform an analysis in the presence of (partially) dirty data, researchers need to
    - Clean or fix dirty data
    - Throw out unfixable data
  - Reporting back to the collection managers (.. *push*)

FilteredPush and Kepler Curation Workflows

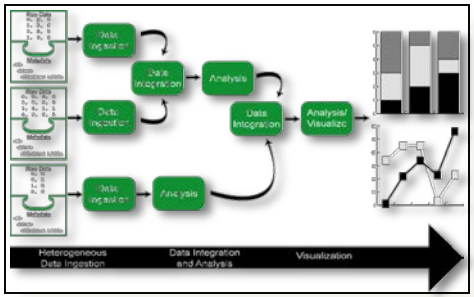


Dou, Lei., G. Cao, P.J. Morris, R.A. Morris, B. Ludäscher, J.A. Macklin, J. Hanken. 2012. **Kurator: A Kepler Package for Data Curation Workflows**, Procedia Computer Science, 9:1614-1619, doi:10.1016/j.procs.2012.04.177

- Today: FP-**Akka** workflow:
  - Validation of (1) SciName; (2) GeoReference; (3) CollectionDate
- Oh, and why workflows?? ASAP!

Scientific Workflows: ASAP!

- **Automation**
  - wfs to **automate** computational aspects of science



- **Scaling** (exploit and optimize ***machine*** *cycles*)
  - wfs should make use of **parallel compute resources**
  - wfs should be able handle **large data**

➔ ***Akka dataflow platform***

Kepler

swift

akka

Taverna

- **Abstraction, Evolution, Reuse** ***(human*** *cycles)*
  - wfs should be easy to **(re-)use, evolve, share**

TRIANA

VisTrails

- **Provenance**
  - wfs should capture **processing history**, **data lineage**

➔ traceable data- and wf-evolution

➔ **Reproducible Science**

ASKALON

Trident Workbench

So many choices …

- **Why not just use system X?**
  - Askalon, Kepler, Taverna, Trident, Triana, …
- Works well for:
  - custom libraries and
  - parameterized workflows

- But there are challenges:
  - … new actors/functionality (extensibility for mere mortals!)
    - powerful but also complex underlying MoCs (“PhD effect”)
  - … adopting system X = learning new language X
    - (tool makers already “speak” languages != X)
  - … sci-wf systems emphasize *process*
  - … but curation about ***data****!*

Scientific Workflows & Scripts

- **New Kurator Approach**
  - **Custom GUI** *(for tool users)* [Kurator Phase 2]
    - for custom workflows built from a small library
    - technology/system agnostic
    - include data viewer
  - **Kurator/P** *(for tool makers)* [Kurator Phase 1]
    - empower makers of curation tool
    - scientific workflow techniques for the rest of us
    - meeting (script-, batch-) programmers half way
    - easy integration with Custom GUI (Phase 2)
- **Spectrum of curation technologies:**

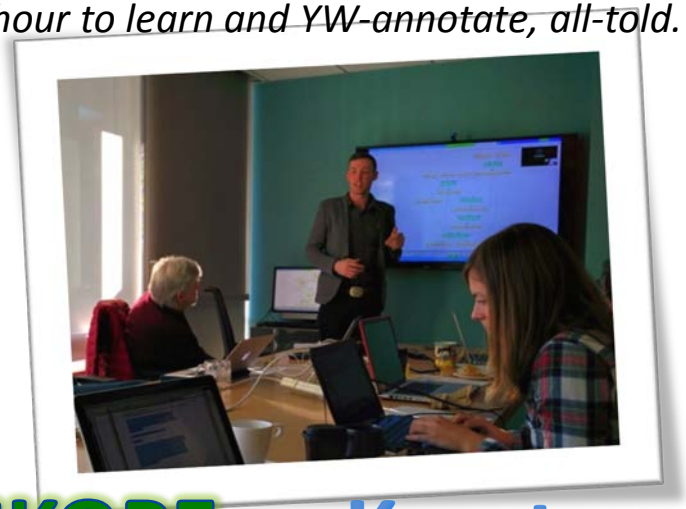
Scientific workflows … *YesWorkflow* … *anyScript*

YesWorkflow Example (EnviRecon.org)

- Python, MATLAB, Bash, … **R scripts** revealed as workflows with **YesWorkflow!**

Kyle B., (computational **R**-)archaeologist:

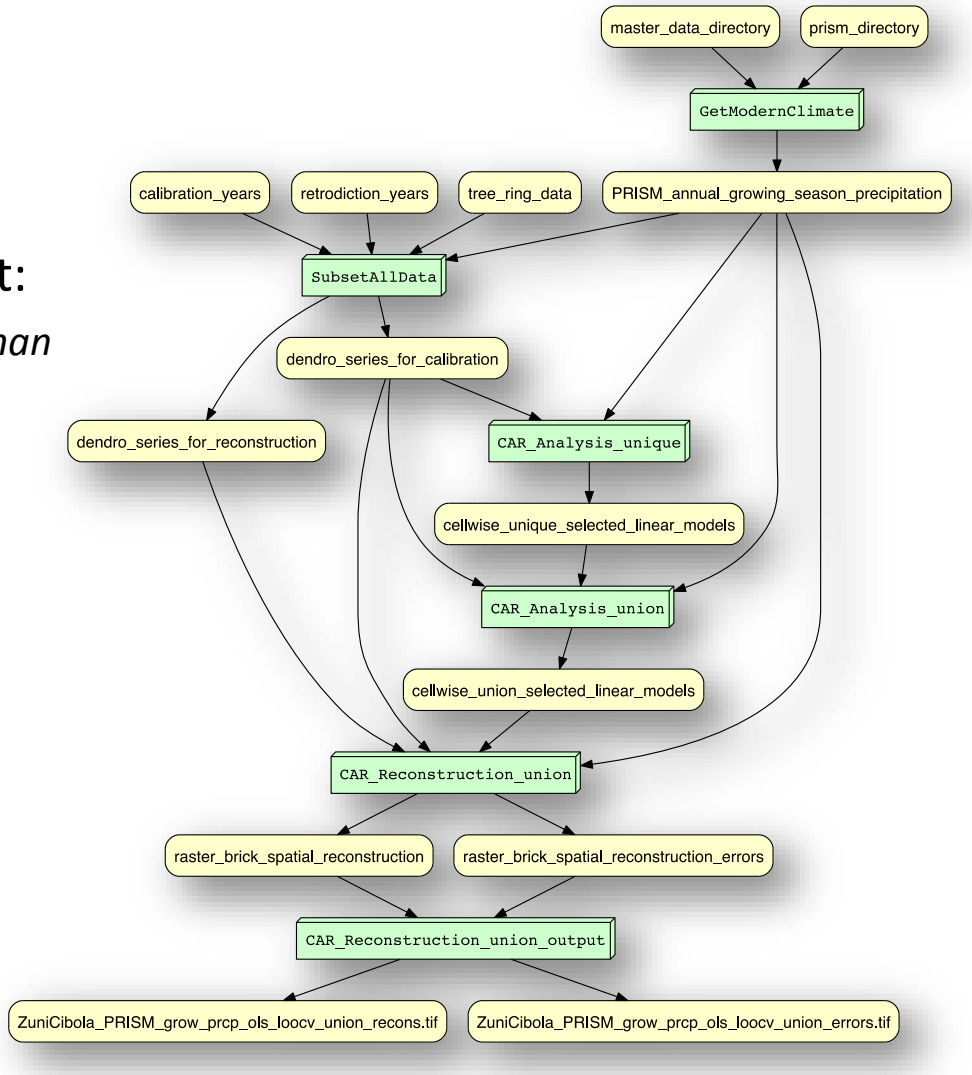
"It took me about 20 minutes to comment. Less than an hour to learn and YW-annotate, all-told."



SKOPE + Kurator

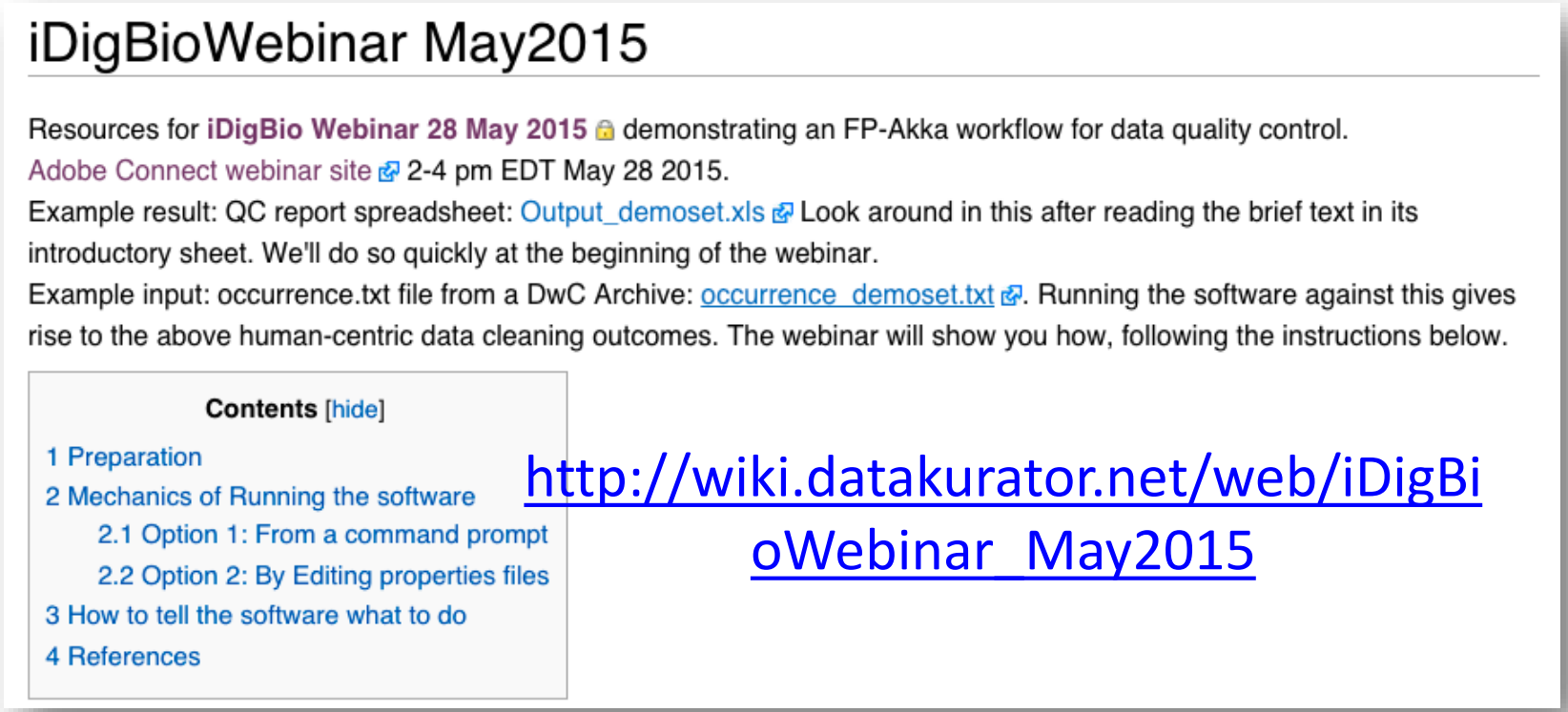
+ DataONE
Data Observation Network for Earth

=> YesWorkflow.org



2 FP-Akka Workflow
Demo & Hands-On Part (optional)

Paul (~ 55 min)



iDigBioWebinar May2015

Resources for **iDigBio Webinar 28 May 2015** demonstrating an FP-Akka workflow for data quality control.
Adobe Connect webinar site 2-4 pm EDT May 28 2015.
Example result: QC report spreadsheet: Output_demoset.xls Look around in this after reading the brief text in its introductory sheet. We'll do so quickly at the beginning of the webinar.
Example input: occurrence.txt file from a DwC Archive: occurrence_demoset.txt. Running the software against this gives rise to the above human-centric data cleaning outcomes. The webinar will show you how, following the instructions below.

Contents [hide]

1 Preparation
2 Mechanics of Running the software
  2.1 Option 1: From a command prompt
  2.2 Option 2: By Editing properties files
3 How to tell the software what to do
4 References

http://wiki.datakurator.net/web/iDigBioWebinar_May2015

Acknowledgments

- **NSF-DBI Filtered Push:** *Continuous Quality Control for Distributed Collections and Other Species-Occurrence Data* (ending)

- **NSF-DBI Kurator:** *A Provenance-enabled Workflow Platform and Toolkit to Curate Biodiversity Data* (started Aug. 2014)

Additional Material

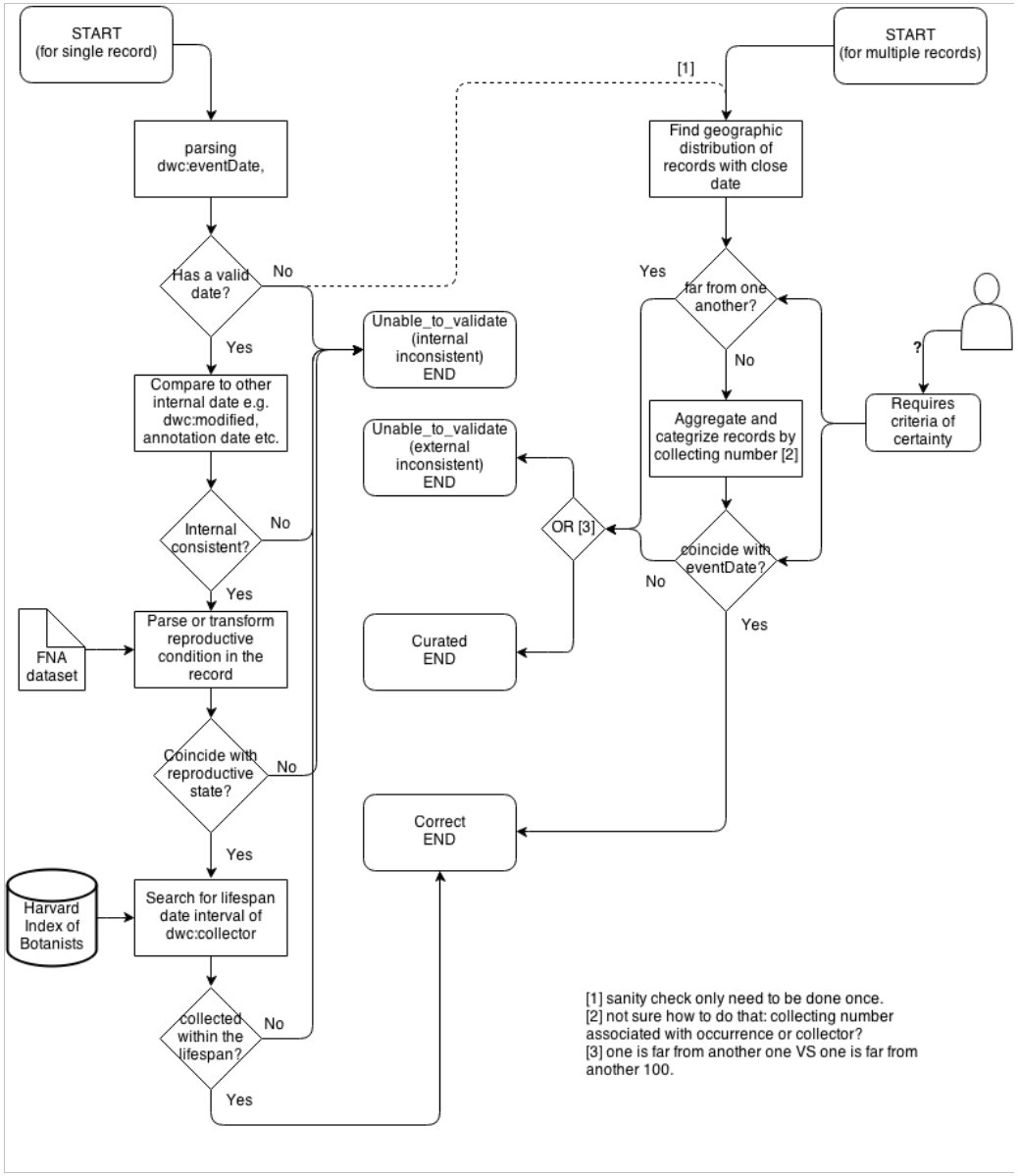
…not part of the presentation…

(for Q&A as needed)

Date Validation

- **Check**:
  - Collector's life span
  - .. vs. Date-Collected

- **Possible outcomes**:
  - **Valid**
  - **Corrected**
  - **Unable to validate**
    - Internal inconsistency
      - Contradicting dates
    - External inconsistency
      - Lack of date data

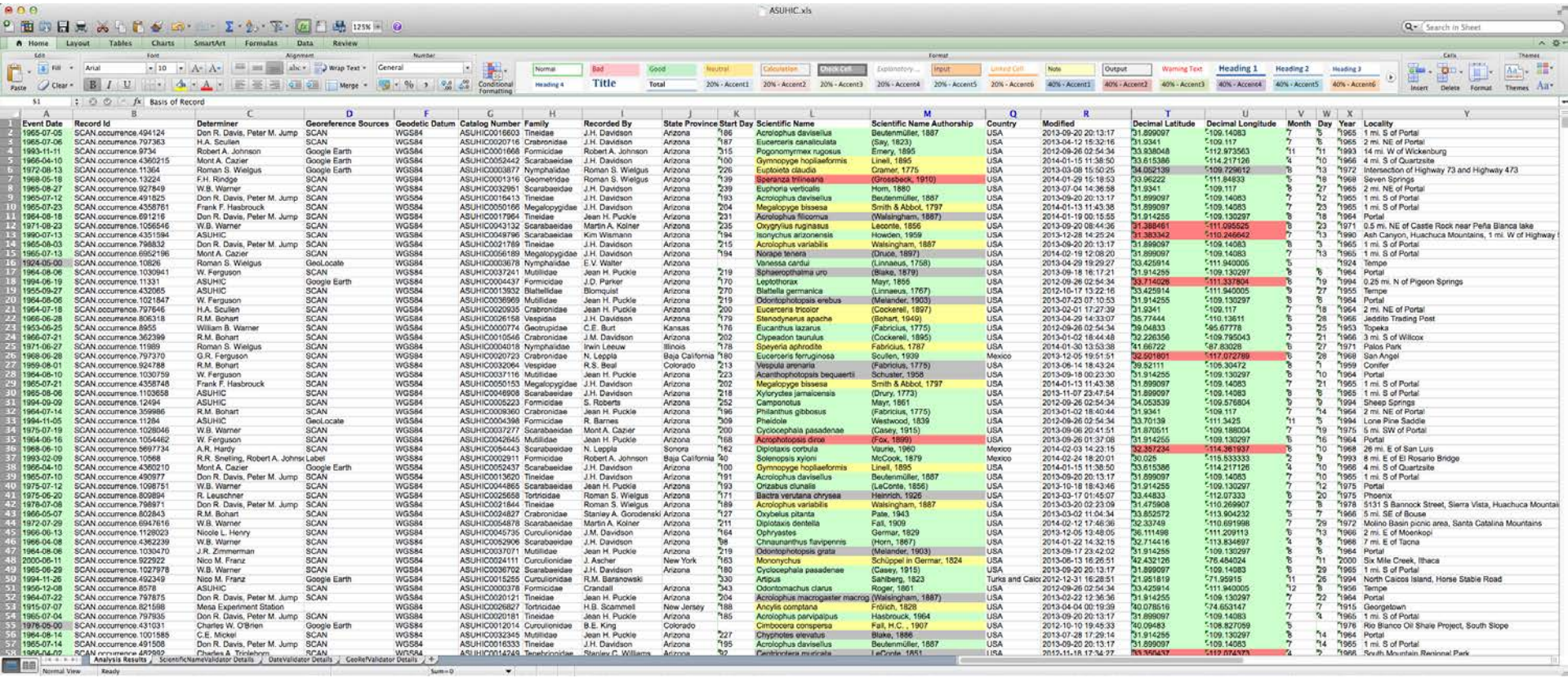


… Logic Behind Each Step (cont'd)

- **Scientific Name Validation**
  - Customer-dependent:
    - Collection Managers:
      - Nomenclature
    - Researchers:
      - Taxonomy (current names)
  - Several Remote services
    - IPNI, GNI, …

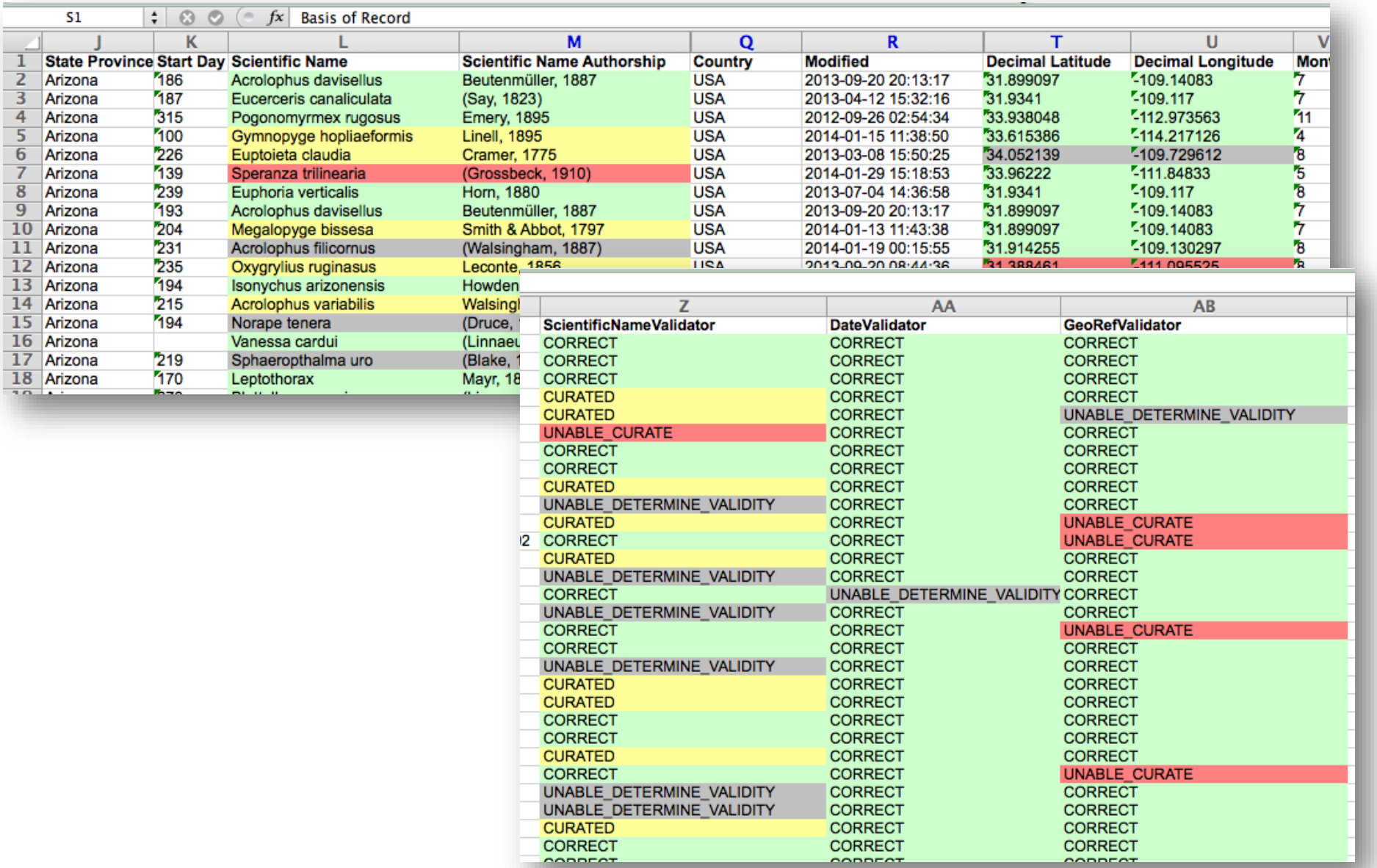
- ***…. <your logic here> …***

Example Output …



Event Date	Record Id	Determiner	Georeference Sources	Geodetic Datum	Catalog Number	Family	Recorded By	State Province	Start Day	Scientific Name	Scientific Name Authorship	Country	Modified	Decimal Latitude	Decimal Longitude	Month	Day	Year	Locality
1965-07-05	SCAN.occurrence.494124	Don R. Davis, Peter M. Jump	SCAN	WGS84	ASUHIC0016603	Tineidae	J.H. Davidson	Arizona	186	Acrolophus davisellus	Beutenmüller, 1887	USA	2013-09-20 20:13:17	31.899097	-109.14083	7	5	1965	1 mi. S of Portal
1965-07-06	SCAN.occurrence.797363	H.A. Scullen	SCAN	WGS84	ASUHIC0020716	Crabronidae	J.H. Davidson	Arizona	187	Eucerceris canaliculata	(Say, 1823)	USA	2013-04-12 15:32:16	31.9341	-109.117	7	6	1965	2 mi. NE of Portal
1993-11-11	SCAN.occurrence.9734	Robert A. Johnson	Google Earth	WGS84	ASUHIC0001668	Formicidae	Robert A. Johnson	Arizona	315	Pogonomyrmex rugosus	Emery, 1895	USA	2012-09-26 02:54:34	33.938048	-112.973563	11	11	1993	14 mi. W of Wickenburg
1966-04-10	SCAN.occurrence.4360215	Mont A. Cazier	Google Earth	WGS84	ASUHIC0052442	Scarabaeidae	J.H. Davidson	Arizona	100	Gymnopyge hoplieaformis	Linell, 1895	USA	2014-01-15 11:38:50	33.615386	-114.217126	4	10	1966	4 mi. S of Quartzsite
1972-08-13	SCAN.occurrence.11364	Roman S. Wielgus	Google Earth	WGS84	ASUHIC0003877	Nymphalidae	Roman S. Wielgus	Arizona	226	Euptoieta claudia	Cramer, 1775	USA	2013-03-08 15:50:25	34.052139	-109.729612	8	13	1972	Intersection of Highway 73 and Highway 473
1968-05-18	SCAN.occurrence.13224	F.H. Rindge	SCAN	WGS84	ASUHIC0001316	Geometridae	Roman S. Wielgus	Arizona	139	Speranza trilinearia	(Grossbeck, 1910)	USA	2014-01-29 15:18:53	33.96222	-111.84933	5	18	1968	Seven Springs
1965-08-27	SCAN.occurrence.927849	W.B. Warner	SCAN	WGS84	ASUHIC0032951	Scarabaeidae	J.H. Davidson	Arizona	239	Euphoria verticalis	Horn, 1880	USA	2013-07-04 14:36:58	31.9341	-109.117	8	27	1965	2 mi. NE of Portal
1965-07-12	SCAN.occurrence.491825	Don R. Davis, Peter M. Jump	SCAN	WGS84	ASUHIC0016413	Tineidae	J.H. Davidson	Arizona	193	Acrolophus davisellus	Beutenmüller, 1887	USA	2013-09-20 20:13:17	31.899097	-109.14083	7	12	1965	1 mi. S of Portal
1965-07-23	SCAN.occurrence.4358761	Frank F. Hasbrouck	SCAN	WGS84	ASUHIC0050166	Megalopygidae	J.H. Davidson	Arizona	204	Megalopyge bisaesa	Smith & Abbot, 1797	USA	2014-01-13 11:43:38	31.899097	-109.14083	7	23	1965	1 mi. S of Portal
1964-08-18	SCAN.occurrence.691216	Don R. Davis, Peter M. Jump	SCAN	WGS84	ASUHIC0017964	Tineidae	Jean H. Puckle	Arizona	231	Acrolophus filicornus	(Walsingham, 1887)	USA	2014-01-19 00:15:55	31.914255	-109.130297	8	18	1964	Portal
1971-08-23	SCAN.occurrence.1056546	W.B. Warner	SCAN	WGS84	ASUHIC0043132	Scarabaeidae	Martin A. Kolner	Arizona	235	Oxygrylius ruginasus	Leconte, 1856	USA	2013-09-20 08:44:36	31.388461	-111.095525	8	23	1971	0.5 mi. NE of Castle Rock near Peña Blanca lake
1990-07-13	SCAN.occurrence.4351594	ASUHIC	SCAN	WGS84	ASUHIC0049786	Scarabaeidae	Kim Wismann	Arizona	194	Isonychus arizonensis	Howden, 1959	USA	2013-12-28 14:25:24	31.383342	-110.246642	7	13	1990	Ash Canyon, Huachuca Mountains, 1 mi. W of Highway
1965-08-03	SCAN.occurrence.794832	Don R. Davis, Peter M. Jump	SCAN	WGS84	ASUHIC0021769	Tineidae	J.H. Davidson	Arizona	215	Acrolophus variabilis	Walsingham, 1887	USA	2013-09-20 20:13:17	31.899097	-109.14083	8	3	1965	1 mi. S of Portal
1965-07-13	SCAN.occurrence.6952196	Mont A. Cazier	SCAN	WGS84	ASUHIC0056189	Megalopygidae	J.H. Davidson	Arizona	194	Norape tehera	(Druce, 1897)	USA	2014-02-19 12:08:20	31.899097	-109.14083	7	13	1965	1 mi. S of Portal
1924-05-00	SCAN.occurrence.10826	Roman S. Wielgus	GeoLocate	WGS84	ASUHIC0003678	Nymphalidae	E.V. Walter	Arizona		Vanessa cardui	(Linnaeus, 1758)	USA	2013-04-29 19:29:27	33.425914	-111.940005	5		1924	Tempe
1964-08-06	SCAN.occurrence.1030941	W. Ferguson	SCAN	WGS84	ASUHIC0037241	Mutillidae	Jean H. Puckle	Arizona	219	Sphaeropthalma uro	(Blake, 1879)	USA	2013-09-18 16:17:21	31.914255	-109.130297	8	6	1964	Portal
1994-06-19	SCAN.occurrence.11331	ASUHIC	Google Earth	WGS84	ASUHIC0004437	Formicidae	J.D. Parker	Arizona	170	Leptothorax	Mayr, 1855	USA	2012-09-26 02:54:34	33.714028	-111.337804	6	19	1994	0.25 mi. N of Pigeon Springs
1955-09-27	SCAN.occurrence.432065	ASUHIC	SCAN	WGS84	ASUHIC0013932	Blattellidae	Blomquist	Arizona	270	Blattella germanica	(Linnaeus, 1767)	USA	2012-10-17 13:22:16	33.425914	-111.940005	9	27	1955	Tempe
1964-08-06	SCAN.occurrence.1021847	W. Ferguson	SCAN	WGS84	ASUHIC0036969	Mutillidae	Jean H. Puckle	Arizona	219	Odontophotopsis erebus	(Melander, 1903)	USA	2013-07-23 07:10:53	31.914255	-109.130297	8	6	1964	Portal
1964-07-18	SCAN.occurrence.797646	H.A. Scullen	SCAN	WGS84	ASUHIC0020935	Crabronidae	Jean H. Puckle	Arizona	200	Eucerceris tricolor	(Cockerell, 1897)	USA	2013-02-01 17:27:39	31.9341	-109.117	7	18	1964	2 mi. NE of Portal
1966-06-28	SCAN.occurrence.806318	R.M. Bohart	SCAN	WGS84	ASUHIC0026158	Vespidae	J.H. Davidson	Arizona	179	Stenodynerus apache	(Bohart, 1949)	USA	2013-04-29 14:33:07	35.77444	-110.13611	6	28	1966	Jeddito Trading Post
1953-06-25	SCAN.occurrence.8955	William B. Warner	SCAN	WGS84	ASUHIC0000774	Geotrupidae	C.E. Burt	Kansas	176	Eucanthus lazarus	(Fabricius, 1775)	USA	2012-09-26 02:54:34	39.04833	-95.67778	6	25	1953	Topeka
1966-07-21	SCAN.occurrence.362399	R.M. Bohart	SCAN	WGS84	ASUHIC0010546	Crabronidae	J.M. Davidson	Arizona	202	Clypeadon taurulus	(Cockerell, 1895)	USA	2013-01-02 18:44:48	32.226356	-109.795043	7	21	1966	3 mi. S of Willcox
1971-06-27	SCAN.occurrence.11989	Roman S. Wielgus	SCAN	WGS84	ASUHIC0004016	Nymphalidae	Irwin Leeuw	Illinois	178	Speyeria aphrodite	Fabricius, 1787	USA	2014-01-30 13:53:38	41.66722	-87.83028	6	27	1971	Palos Park
1968-06-28	SCAN.occurrence.797370	G.R. Ferguson	SCAN	WGS84	ASUHIC0020723	Crabronidae	N. Leppla	Baja California	180	Eucerceris ferruginosa	Scullen, 1939	Mexico	2013-12-05 19:51:51	32.501801	-117.072789	6	28	1968	San Angel
1959-08-01	SCAN.occurrence.924788	R.M. Bohart	SCAN	WGS84	ASUHIC0032064	Vespidae	R.S. Beal	Colorado	213	Vespula arenaria	(Fabricius, 1775)	USA	2013-06-14 18:43:24	39.52111	-105.30472	8	1	1959	Conifer
1964-08-10	SCAN.occurrence.1030759	W. Ferguson	SCAN	WGS84	ASUHIC0037116	Mutillidae	Jean H. Puckle	Arizona	223	Acanthophotopsis bequaerti	Schuster, 1958	USA	2013-09-18 00:23:30	31.914255	-109.130297	8	10	1964	Portal
1965-07-21	SCAN.occurrence.4358748	Frank F. Hasbrouck	SCAN	WGS84	ASUHIC0050153	Megalopygidae	J.H. Davidson	Arizona	202	Megalopyge bisaesa	Smith & Abbot, 1797	USA	2014-01-13 11:43:38	31.899097	-109.14083	7	21	1965	1 mi. S of Portal
1965-08-06	SCAN.occurrence.1103658	ASUHIC	SCAN	WGS84	ASUHIC0046908	Scarabaeidae	J.H. Davidson	Arizona	218	Xylorydes jamaicensis	(Drury, 1773)	USA	2013-11-07 23:47:54	31.899097	-109.14083	8	6	1965	1 mi. S of Portal
1994-09-09	SCAN.occurrence.12494	ASUHIC	SCAN	WGS84	ASUHIC0005223	Formicidae	S. Roberts	Arizona	252	Camponotus	Mayr, 1861	USA	2012-09-26 02:54:34	34.053539	-109.579804	9	9	1994	Sheep Springs
1964-07-14	SCAN.occurrence.359986	R.M. Bohart	SCAN	WGS84	ASUHIC0009360	Crabronidae	Jean H. Puckle	Arizona	196	Philanthus gibbosus	(Fabricius, 1775)	USA	2013-01-02 18:40:44	31.9341	-109.117	7	14	1964	2 mi. NE of Portal
1994-11-05	SCAN.occurrence.11284	ASUHIC	GeoLocate	WGS84	ASUHIC0004398	Formicidae	R. Barnes	Arizona	309	Pheidole	Westwood, 1839	USA	2012-09-26 02:54:34	33.70139	-111.3425	11	5	1994	Lone Pine Saddle
1975-07-19	SCAN.occurrence.1028046	W.B. Warner	SCAN	WGS84	ASUHIC0037277	Scarabaeidae	Mont A. Cazier	Arizona	200	Cyclocephala pasadenae	(Casey, 1915)	USA	2013-09-06 20:41:51	31.870511	-109.188004	7	19	1975	5 mi. SW of Portal
1964-06-16	SCAN.occurrence.1054462	W. Ferguson	SCAN	WGS84	ASUHIC0042645	Mutillidae	Jean H. Puckle	Arizona	168	Acrophotopsis dirce	(Fox, 1899)	USA	2013-09-26 01:37:08	31.914255	-109.130297	6	16	1964	Portal
1968-08-10	SCAN.occurrence.5697734	A.R. Hardy	SCAN	WGS84	ASUHIC0064443	Scarabaeidae	N. Leppla	Sonora	162	Diplotaxis cortula	Vaurie, 1960	Mexico	2014-02-03 14:23:15	32.357234	-114.361937	8	10	1968	26 mi. E of San Luis
1993-02-09	SCAN.occurrence.10568	R.R. Snelling, Robert A. Johns	Label	WGS84	ASUHIC0002911	Formicidae	Robert A. Johnson	Baja California	40	Solenopsis xyloni	McCook, 1879	Mexico	2014-02-24 18:20:01	30.025	-115.533333	2	9	1993	8 mi. E of El Rosario Bridge
1966-04-10	SCAN.occurrence.4360210	Mont A. Cazier	Google Earth	WGS84	ASUHIC0052437	Scarabaeidae	J.H. Davidson	Arizona	100	Gymnopyge hoplieaformis	Linell, 1895	USA	2014-01-15 11:38:50	33.615386	-114.217126	4	10	1966	4 mi. S of Quartzsite
1965-07-10	SCAN.occurrence.490977	Don R. Davis, Peter M. Jump	SCAN	WGS84	ASUHIC0013620	Tineidae	J.H. Davidson	Arizona	191	Acrolophus davisellus	Beutenmüller, 1887	USA	2013-09-20 20:13:17	31.899097	-109.14083	7	10	1965	1 mi. S of Portal
1975-07-12	SCAN.occurrence.1098751	W.B. Warner	SCAN	WGS84	ASUHIC0044865	Scarabaeidae	Jean H. Puckle	Arizona	193	Orizabus clunalis	(LeConte, 1856)	USA	2013-10-18 18:43:46	31.914255	-109.130297	7	12	1975	Portal
1975-06-20	SCAN.occurrence.809894	R. Leuschner	SCAN	WGS84	ASUHIC0025658	Tortricidae	Roman S. Wielgus	Arizona	171	Bactra verutana chrysea	Heinrich, 1926	USA	2013-03-17 01:45:07	33.44833	-112.07333	6	20	1975	Phoenix
1978-07-06	SCAN.occurrence.798971	Don R. Davis, Peter M. Jump	SCAN	WGS84	ASUHIC0021844	Tineidae	Roman S. Wielgus	Arizona	189	Acrolophus variabilis	Walsingham, 1887	USA	2013-03-20 02:23:09	31.475908	-110.269907	7	8	1978	5131 S Bannock Street, Sierra Vista, Huachuca Mountai
1966-05-07	SCAN.occurrence.802843	R.M. Bohart	SCAN	WGS84	ASUHIC0024827	Crabronidae	Stanley A. Gorodenski	Arizona	127	Oxybelus pitanta	Pate, 1943	USA	2013-03-02 11:04:34	33.852572	-113.904232	5	7	1966	5 mi. SE of Bouse
1972-07-29	SCAN.occurrence.6947616	W.B. Warner	SCAN	WGS84	ASUHIC0054878	Scarabaeidae	Martin A. Kolner	Arizona	211	Diplotaxis dentella	Fall, 1909	USA	2014-02-12 17:46:36	32.33749	-110.691998	7	29	1972	Molino Basin picnic area, Santa Catalina Mountains
1966-06-13	SCAN.occurrence.1128023	Nicole L. Henry	SCAN	WGS84	ASUHIC0045735	Curculionidae	J.M. Davidson	Arizona	164	Ophryastes	Germar, 1829	USA	2013-12-05 13:48:05	36.111498	-111.209113	6	13	1966	2 mi. E of Moenkopi
1966-04-08	SCAN.occurrence.4362239	W.B. Warner	SCAN	WGS84	ASUHIC0052906	Scarabaeidae	J.H. Davidson	Arizona	98	Chnaunanthus flavipennis	(Horn, 1867)	USA	2014-01-22 14:32:15	32.714416	-113.834697	4	8	1966	7 mi. E of Tacna
1964-08-06	SCAN.occurrence.1030470	J.R. Zimmerman	SCAN	WGS84	ASUHIC0037071	Mutillidae	Jean H. Puckle	Arizona	219	Odontophotopsis grata	(Melander, 1903)	USA	2013-09-17 23:42:02	31.914255	-109.130297	8	6	1964	Portal
2000-06-11	SCAN.occurrence.922922	Nico M. Franz	SCAN	WGS84	ASUHIC0024111	Curculionidae	J. Ascher	New York	163	Mononychus	Schüppel in Germar, 1824	USA	2013-06-13 16:26:51	42.432126	-76.484024	6	11	2000	Six Mile Creek, Ithaca
1965-06-29	SCAN.occurrence.1027978	W.B. Warner	SCAN	WGS84	ASUHIC0036702	Scarabaeidae	J.H. Davidson	Arizona	180	Cyclocephala pasadenae	(Casey, 1915)	USA	2013-09-20 20:13:17	31.899097	-109.14083	6	29	1965	1 mi. S of Portal
1994-11-26	SCAN.occurrence.492349	Nico M. Franz	Google Earth	WGS84	ASUHIC0015255	Curculionidae	R.M. Baranowski		330	Artipus	Sahlberg, 1823	Turks and Caic	2012-12-31 16:28:51	21.951819	-71.95915	11	26	1994	North Caicos Island, Horse Stable Road
1956-12-08	SCAN.occurrence.8578	ASUHIC	SCAN	WGS84	ASUHIC0000378	Formicidae	Crandall	Arizona	343	Odontomachus clarus	Roger, 1861	USA	2012-09-26 02:54:34	33.425914	-111.940005	12	8	1956	Tempe
1964-07-22	SCAN.occurrence.797875	Don R. Davis, Peter M. Jump	SCAN	WGS84	ASUHIC0020121	Tineidae	Jean H. Puckle	Arizona	204	Acrolophus macrogaster macrog	(Walsingham, 1887)	USA	2013-02-22 12:36:36	31.914255	-109.130297	7	22	1964	Portal
1915-07-07	SCAN.occurrence.821598	Mesa Experiment Station	SCAN	WGS84	ASUHIC0026827	Tortricidae	H.B. Scammell	New Jersey	188	Ancylis comptana	Frölich, 1828	USA	2013-04-04 00:19:39	40.078516	-74.653147	7	7	1915	Georgetown
1965-07-04	SCAN.occurrence.797935	Don R. Davis, Peter M. Jump	SCAN	WGS84	ASUHIC0020181	Tineidae	Jean H. Puckle	Arizona	185	Acrolophus parvipalpus	Hasbrouck, 1964	USA	2013-09-20 20:13:17	31.899097	-109.14083	7	4	1965	1 mi. S of Portal
1976-05-00	SCAN.occurrence.431031	Charles W. O'Brien	Google Earth	WGS84	ASUHIC0012014	Curculionidae	B.E. King	Colorado		Cimbocera conspersa	Fall, H.C. , 1907	USA	2012-10-10 19:45:33	40.09483	-103.8270	5		1976	Rio Blanco Oil Shale Project, South Slope
1964-08-14	SCAN.occurrence.1001585	C.E. Mickel	SCAN	WGS84	ASUHIC0032345	Mutillidae	Jean H. Puckle	Arizona	227	Chyphotes elevatus	Blake, 1886	USA	2013-07-28 17:29:14	31.914255	-109.130297	8	14	1964	Portal
1965-07-14	SCAN.occurrence.491508	Don R. Davis, Peter M. Jump	SCAN	WGS84	ASUHIC0016333	Tineidae	J.H. Davidson	Arizona	195	Acrolophus davisellus	Beutenmüller, 1887	USA	2013-09-20 20:13:17	31.899097	-109.14083	7	14	1965	1 mi. S of Portal
1966-04-02	SCAN.occurrence.447992	Charles A. Triplehorn	SCAN	WGS84	ASUHIC0014249	Tenebrionidae	Stanley C. Williams	Arizona	92	Centrioptera muricata	LeConte, 1851	USA	2012-11-18 17:34:27	33.305437	-112.074373	4	2	1966	South Mountain Regional Park

… close up …

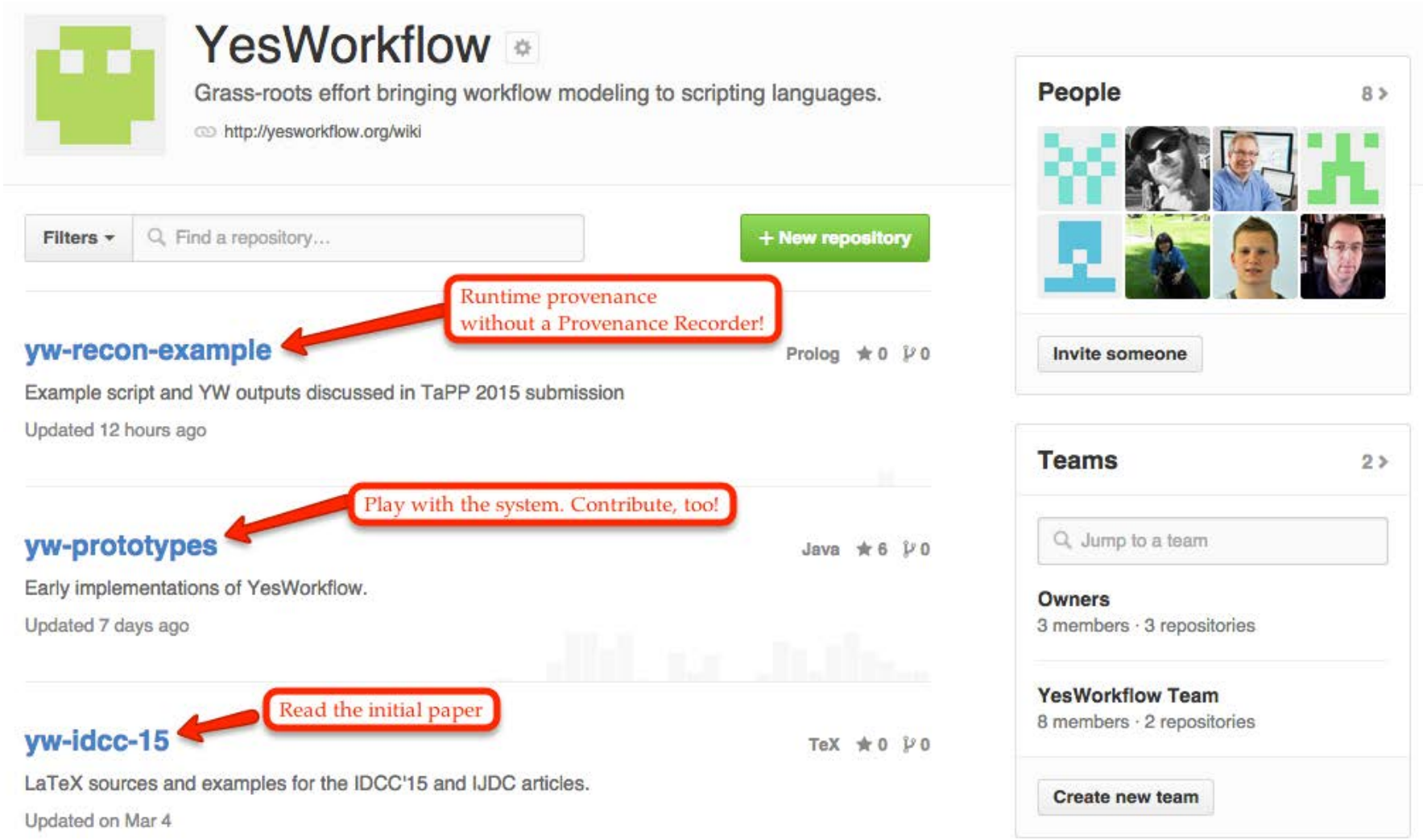

S1
Basis of Record
J State Province
K Start Day
L Scientific Name
M Scientific Name Authorship
Q Country
R Modified
T Decimal Latitude
U Decimal Longitude
Arizona 186 Acrolophus davisellus Beutenmüller, 1887 USA 2013-09-20 20:13:17 31.899097 -109.14083
Arizona 187 Eucerceris canaliculata (Say, 1823) USA 2013-04-12 15:32:16 31.9341 -109.117
Arizona 315 Pogonomyrmex rugosus Emery, 1895 USA 2012-09-26 02:54:34 33.938048 -112.973563
Arizona 100 Gymnopyge hopliaeformis Linell, 1895 USA 2014-01-15 11:38:50 33.615386 -114.217126
Arizona 226 Euptoieta claudia Cramer, 1775 USA 2013-03-08 15:50:25 34.052139 -109.729612
Arizona 139 Speranza trilinearia (Grossbeck, 1910) USA 2014-01-29 15:18:53 33.96222 -111.84833
Arizona 239 Euphoria verticalis Horn, 1880 USA 2013-07-04 14:36:58 31.9341 -109.117
Arizona 193 Acrolophus davisellus Beutenmüller, 1887 USA 2013-09-20 20:13:17 31.899097 -109.14083
Arizona 204 Megalopyge bissesa Smith & Abbot, 1797 USA 2014-01-13 11:43:38 31.899097 -109.14083
Arizona 231 Acrolophus filicornus (Walsingham, 1887) USA 2014-01-19 00:15:55 31.914255 -109.130297
Arizona 235 Oxygrylius ruginasus Leconte, 1856
Arizona 194 Isonychus arizonensis Howden
Arizona 215 Acrolophus variabilis Walsingl
Arizona 194 Norape tenera (Druce,
Arizona Vanessa cardui (Linnaeu
Arizona 219 Sphaeropthalma uro (Blake,
Arizona 170 Leptothorax Mayr, 18
Z ScientificNameValidator
AA DateValidator
AB GeoRefValidator
CORRECT CORRECT CORRECT
CORRECT CORRECT CORRECT
CORRECT CORRECT CORRECT
CURATED CORRECT CORRECT
CURATED CORRECT UNABLE_DETERMINE_VALIDITY
UNABLE_CURATE CORRECT CORRECT
CORRECT CORRECT CORRECT
CORRECT CORRECT CORRECT
CURATED CORRECT CORRECT
UNABLE_DETERMINE_VALIDITY CORRECT CORRECT
CURATED CORRECT UNABLE_CURATE
2 CORRECT CORRECT UNABLE_CURATE
CURATED CORRECT CORRECT
UNABLE_DETERMINE_VALIDITY CORRECT CORRECT
CORRECT UNABLE_DETERMINE_VALIDITY CORRECT
UNABLE_DETERMINE_VALIDITY CORRECT CORRECT
CORRECT CORRECT UNABLE_CURATE
CORRECT CORRECT CORRECT
UNABLE_DETERMINE_VALIDITY CORRECT CORRECT
CURATED CORRECT CORRECT
CURATED CORRECT CORRECT
CORRECT CORRECT CORRECT
CORRECT CORRECT CORRECT
CURATED CORRECT CORRECT
CORRECT CORRECT UNABLE_CURATE
UNABLE_DETERMINE_VALIDITY CORRECT CORRECT
UNABLE_DETERMINE_VALIDITY CORRECT CORRECT
CURATED CORRECT CORRECT
CORRECT CORRECT CORRECT

Kurator/P & YW: the road ahead …

- **YesWorkflow**:
  - … finishing support for **retrospective provenance** *without* using a runtime provenance recorder!
  - Key insight: scientists already leave provenance "bread crumbs" behind! (it's not an accident!)
  - Exploit that via annotations: URI-templates

- **Kurator**[/P]:
  - How far can we go towards ASAP via YW?

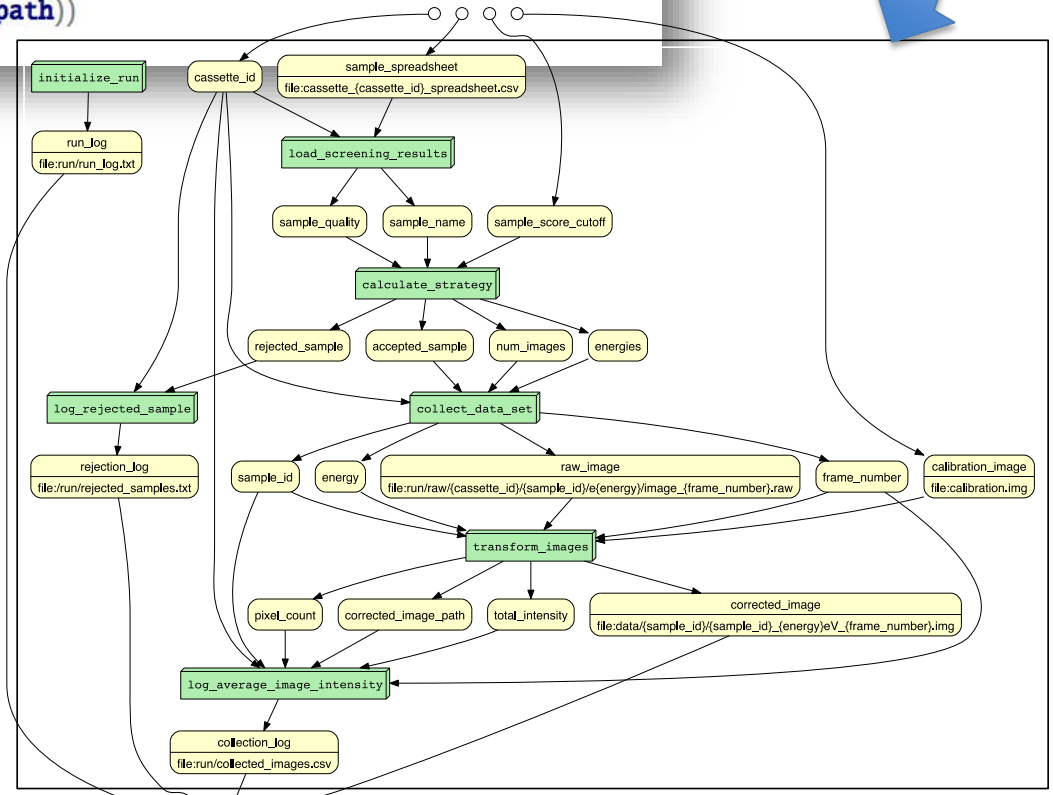
YesWorkflow.org



YW-RECON: Prospective & Retrospective Provenance … (almost) for free!

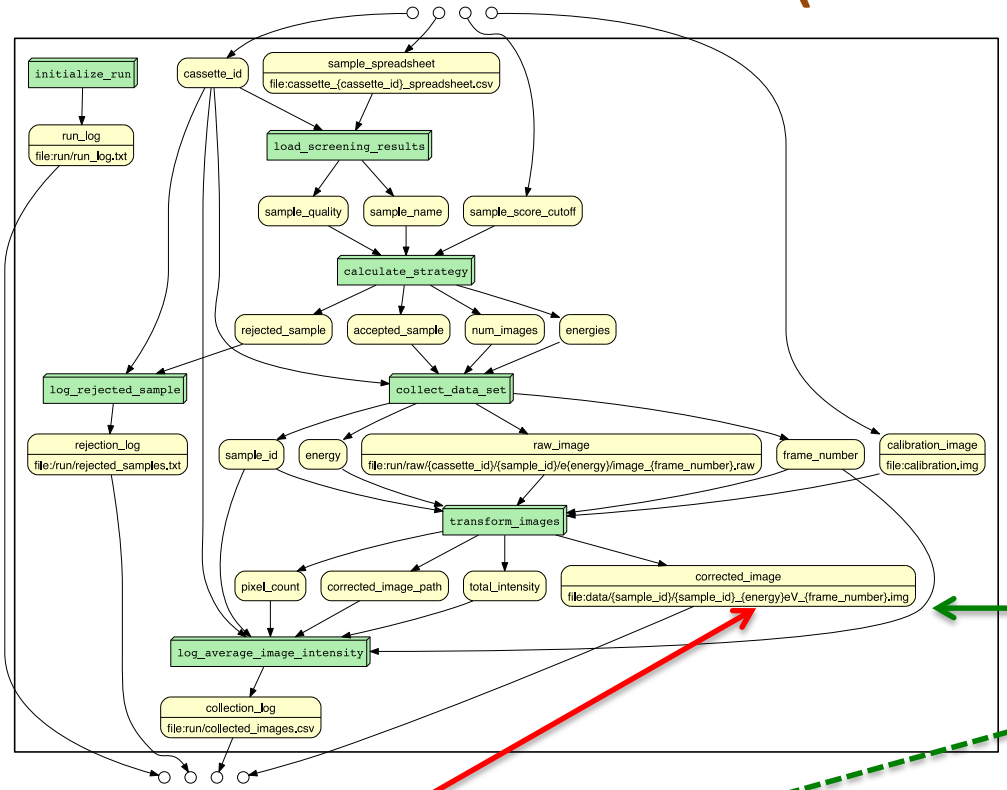
```
# @BEGIN collect_data_set
# @PARAM cassette_id @PARAM accepted_sample @PARAM num_images @PARAM energies
# @OUT sample_id @OUT energy @OUT frame_number
# @OUT raw_image_path @AS raw_image
# @URI file:run/raw/{cassette_id}/{sample_id}/e{energy}/image_{frame_number}.raw
run_log.write("Collecting data set for sample {0}".format(accepted_sample))
sample_id = accepted_sample
for energy, frame_number, intensity, raw_image_path in collect_next_image(
                        cassette_id, sample_id, num_images, energies,
                        "run/raw/{cassette_id}/{sample_id}/e{energy}/image_{frame_number:03d}.raw"):
    run_log.write("Collecting image {0}".format(raw_image_path))
# @END collect_data_set
```

YW



- YW annotations in the script (R, Python, Matlab) are used to **recreate the workflow view** from the script …

YW-RECON: Prospective & Retrospective Provenance … (almost) for free!




```
run/
├── raw
│   └── q55
│       ├── DRT240
│       │   ├── e10000
│       │   │   ├── image_001.raw
...    ...  ... ...
│       │   │   └── image_037.raw
│       │   └── e11000
│       │       ├── image_001.raw
...    ...    ...
│       │       └── image_037.raw
│       └── DRT322
│           ├── e10000
│           │   ├── image_001.raw
...        ... ...
│           │   └── image_030.raw
│           └── e11000
│               ├── image_001.raw
...            ...
│               └── image_030.raw
├── data
│   ├── DRT240
│   │   ├── DRT240_10000eV_001.img
... ... ...
│   │   └── DRT240_11000eV_037.img
│   └── DRT322
│       ├── DRT322_10000eV_001.img
...    ...
│       └── DRT322_11000eV_030.img
│
├── collected_images.csv
├── rejected_samples.txt
└── run_log.txt
```

- **URI-templates *link*** conceptual entities to **runtime provenance** "left behind" by the script author …
- … facilitating provenance reconstruction

Summary: Data Curation with Scientific Workflow Systems

Scientific Workflows

- [+] **A**utomation
- [+] **S**calability
- [+] **A**bstraction
- [+] **P**rovenance
- …
- [+/0] Easy to use
  - [0] learning a new paradigm
- [-] Teaching resources: learning a new language!
- [-] Special expertise needed for deep changes

  e.g. new Java actors, shims, …

Kurator/P: Scripts + YesWorkflow ++

Scripts: [+] **Automation**, [0] Scalability, [-] Abstraction, [0/-] Provenance

Now: Scripts + YesWorkflow Annotations

- **[+] Abstraction**
  - explain your methods to mere mortals

  => encourage (re-)use
- **[+] Provenance:**
  - **YesWorkflow** (***prospective*** and ***retrospective*** provenance)
- [+] Language independent (R, Matlab, Python, …)
- [+] **Empower tool makers** (script programmers): give them …
  - … some **immediate** benefits **(workflow views, retrospective provenance)**
  - … some **long term** benefits: think about your methods differently

  => dataflow programming => [+] Scalability