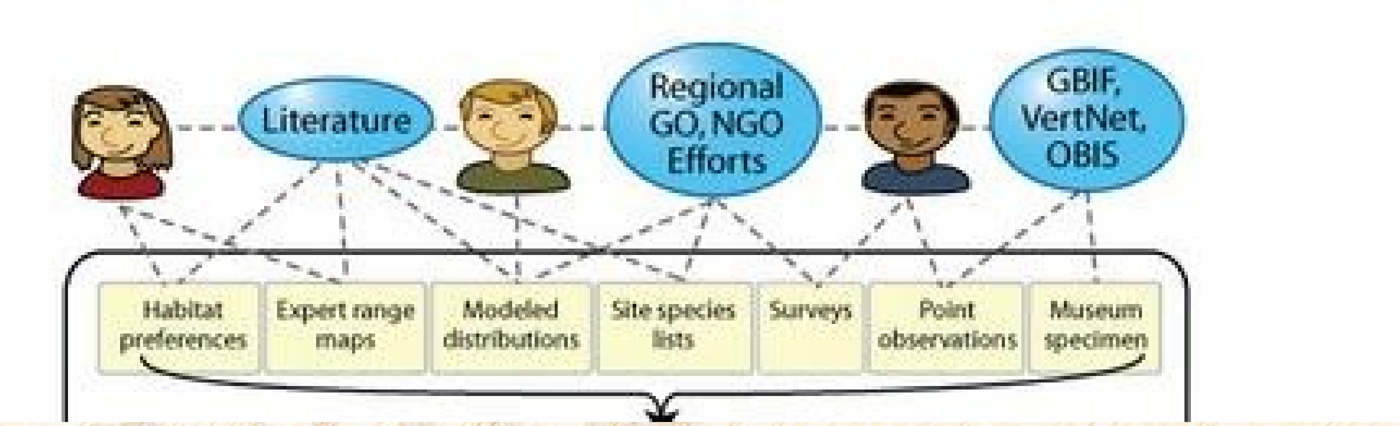
DATA and KNOWLEDGE INFRASTRUCTURE ↔ ANALYTICS, KNOWLEDGE DISSEMINATION

Herbarium of

ALS



Literature

Regional GO, NGO Efforts

GBIF, VertNet, OBIS

Habitat preferences

Expert range maps

Modeled distributions

Site species lists

Surveys

Point observations

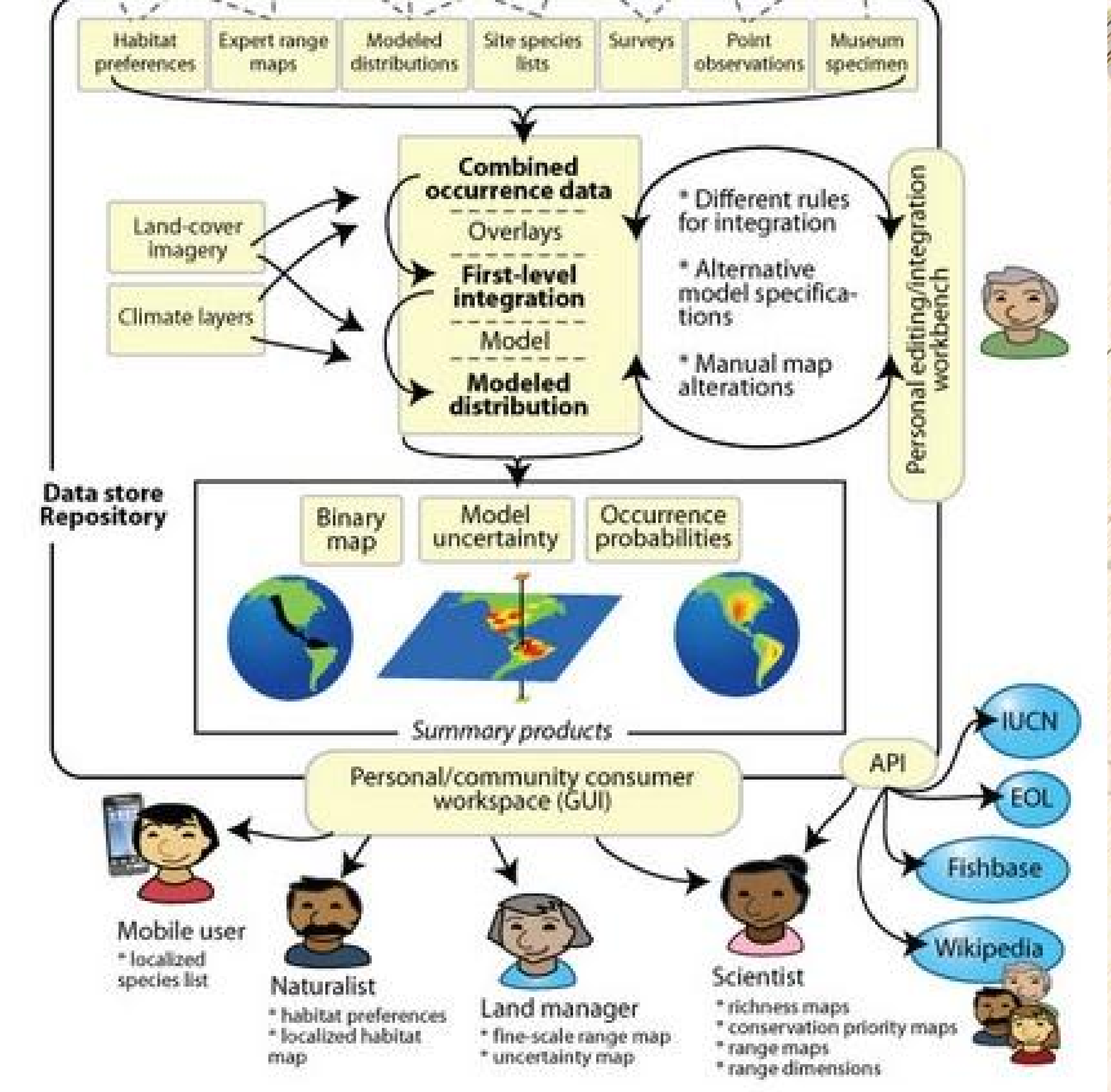
Museum specimen

Taxon: Andromeda glaucophylla

Notes: Yost Bog, Sec. 22, T38N,
Lagrange County, Indiana,

Contributors from multiple sectors add to Map of Life

legit: Anthony L. Swinehart Date:



Habitat preferences

Expert range maps

Modeled distributions

Site species lists

Surveys

Point observations

Museum specimen

Land-cover imagery

Climate layers

Combined occurrence data

Overlays

First-level integration

Model

Modeled distribution

* Different rules for integration

* Alternative model specifications

* Manual map alterations

Personal editing/integration workbench

Data store Repository

Binary map

Model uncertainty

Occurrence probabilities

Summary products

Personal/community consumer workspace (GUI)

API

IUCN

EOL

Fishbase

Wikipedia

Mobile user

* localized species list

Naturalist

* habitat preferences
* localized habitat map

Land manager

* fine-scale range map
* uncertainty map

Scientist

* richness maps
* conservation priority maps
* range maps
* range dimensions

Consumers use Map of Life knowledge for societal needs

BUT TO PROPERLY USE THESE DATA WE NEED TO UNDERSTAND DIFFERENT SOURCES OF DATA AND HOW TO DESCRIBE THEM

Description	Example	Contribut., Quality	Protocols	Effort Report	Source	Raw data	Temporal Scope	Geographic scope	Reporting basis	Complete &suited absence inference	Suited input occpy. Model
Summary inventory (limited protocol & effort report)	Protected area species list	many, heterogen.	multi, unclear	Very limited	Literature	no	long (years)	Clear (often >1km)	Multi (observation, Photo, Lit,)	Yes	No
Summary inventory (some protocol and effort reporting)	Standardized area survey (e.g. atlas grid cell)	many, heterogen.	multiple single, clear	Possible	Literature, Project reporting	no	long (months, years)	Clear (often >1km)	Single (e.g. Observation)	yes	no
Single person/group inventory: observation	Standardized area survey (e.g. transect count)	single, high & vetted	single, clear	Yes	Project data	yes	short (hours, days)	Clear (m to 1km)	Single (e.g. Observation)	no	Potent-ially
Inventory following protocol: stationary trapping	Camera traps & more typical trappings	single, high & vetted	singe, somewhat clear	Yes	Project data	yes	short (hours, days)	Very small (meters)	Single (e.g. Observation)	somewhat (over very small extent)	Yes
Inventory following protocol: active sampling campaigns	fish, zooplancton netting, algal sampling	single, high & vetted	single, clear	Yes	Project data	yes	usually short	Small (e.g. meters)	Single (e.g. Observation)	no	Yes
Full inventory following very defined protocol	CTFS forest plot, Revelle plots	single, high & vetted	single, clear	Perfect: full coverage	Project data	yes	short	V. small (e.g. meters)	Single (e.g. Observation)	yes (over very small extent)	NA
Inventory following loose protocol: citizen science observation	transect	single, heterogen., unvetted	single, clear	Yes	Project data	yes	short (hours, days)	Clear (m to 1km)	Single (e.g. Observation)	no	Potent-ially

A metadata schema for collating data from inventories

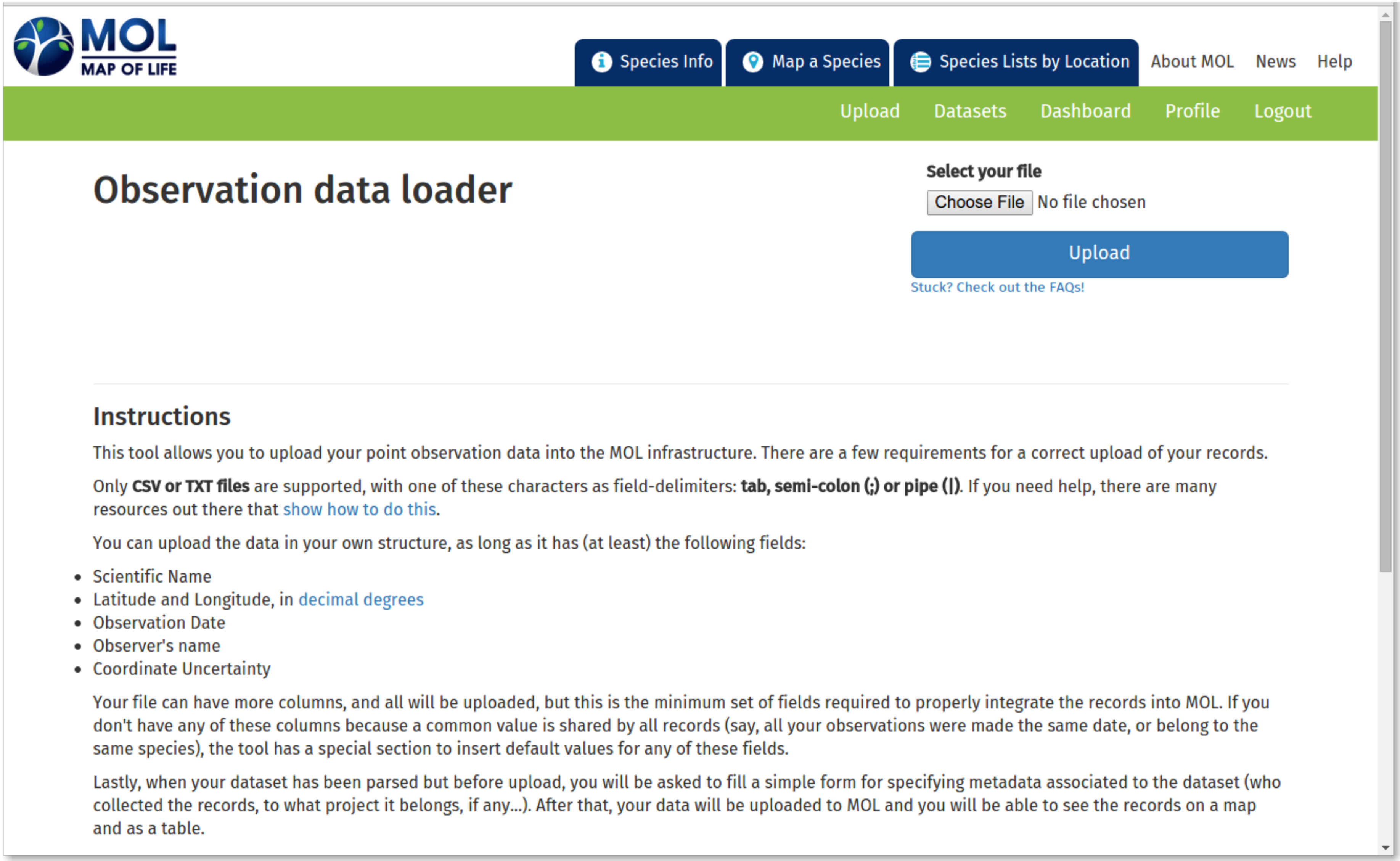
Humboldt Core Version 1	Area species checklists	Geographically restricted surveys
	Gridded Atlas survey Protected area species list	Transect count Trapping and netting CTFS forest, Revelle plots
General dataset & identification terms	inventory perfomed by; dataset name, identifier, publisher, licensing, rights holders; metadata recorded by; citation reference and id; taxa identifier by; identification quality; cited taxonomic authority	
Geospatial & Habitat Scope Terms	Geospatial scope; areal extent; total area inventoried; number of sites; site names and details; lat/lon by site; elevation range and units; habitats included and excluded.	
Temporal Scope Terms	Survey time blocks; start and end year, month day; time units spent In blocks; daily start, end time; study diurnality, study season.	
Taxonomic Scope Terms	Prospective taxonomic scope inclusion and exclusion; distribution status included and excluded; developmental stage included and excluded, size classes included and excluded.	
Methodology Description Terms	Inventory type; Compiled data Y/N & type; abundances and/or absences reported? Absence list	Inventory type, protocol name, detail, citation , reference, abundances reported Y/N &cap; absences reported?
Completeness & effort terms	Completeness reported and how; Inferred taxonomic completeness Upper/lower bound and how.	Effort reporting & lower/upper bound and granular breakdown; effort method; Vouchers or samples taken and how?

Putting it into practice – Assembling area species checklist data and metadata from the literature

Team of Boulder and Yale developers and students assembled metadata for (so far) 143 area checklists and collated information about area checklists characteristics

Humbolt Core Term	Possible Values	Percents
Compilation effort	Low, medium, high, na	Low – 38.7%, Medium – 6.2%, High – 8.5%, n.a – 45.7%
Abundances reported	Yes/No	Yes-57.5% No-42.6
Absences reported	Yes/No	Yes-41.8 , No-58.2%
Completeness assessment	Scale in 25 percent increments from 0-100	50-100% complete – 30.3%, 75-100% - 27.9%, other- 41.8%

A tool for the long-tail data on YOUR computer



Observation data loader

Select your file

Stuck? Check out the FAQs!

Instructions

This tool allows you to upload your point observation data into the MOL infrastructure. There are a few requirements for a correct upload of your records.

Only **CSV or TXT files** are supported, with one of these characters as field-delimiters: **tab, semi-colon (;) or pipe (|)**. If you need help, there are many resources out there that show how to do this.

You can upload the data in your own structure, as long as it has (at least) the following fields:

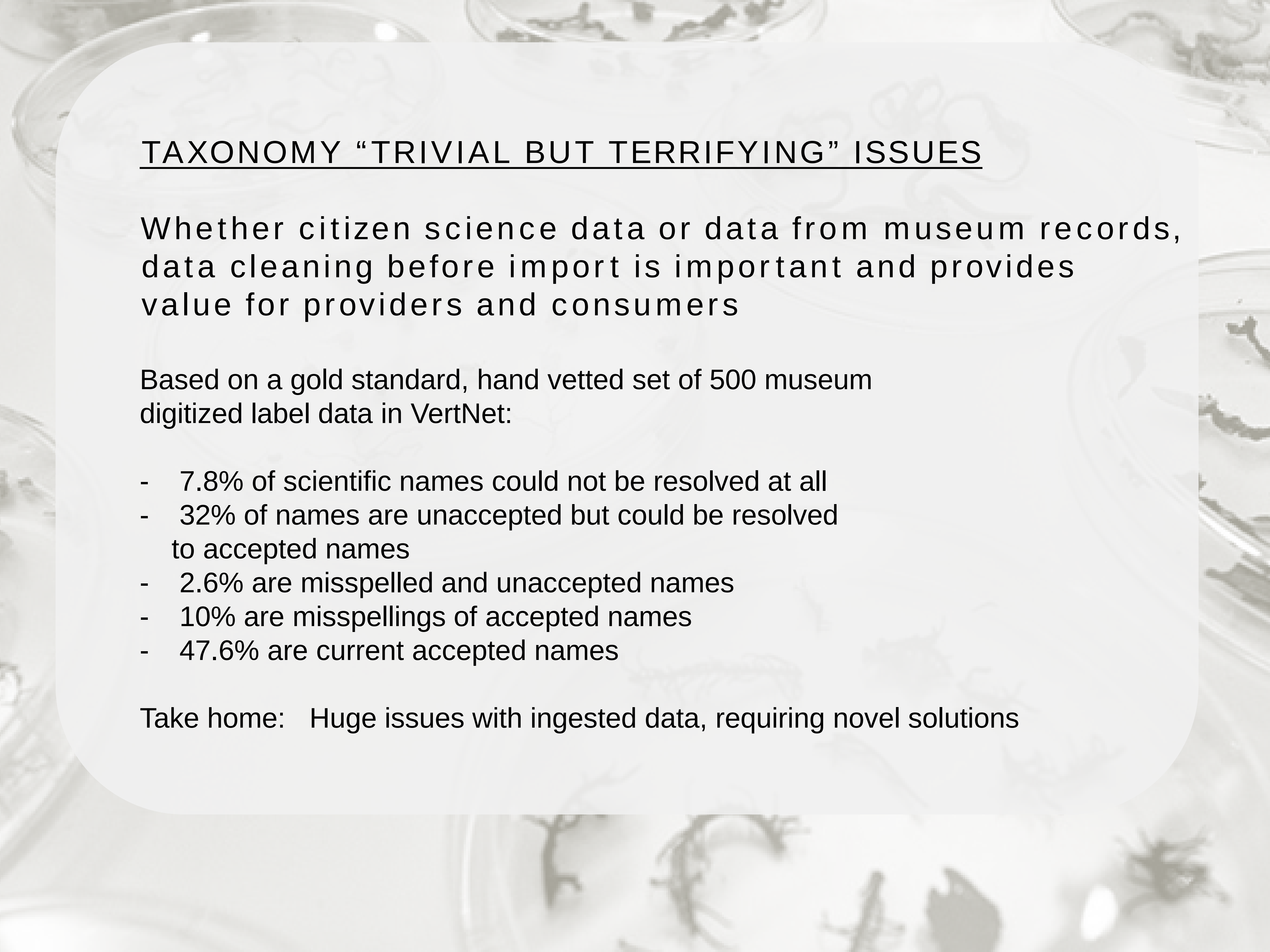
- Scientific Name
- Latitude and Longitude, in decimal degrees
- Observation Date
- Observer's name
- Coordinate Uncertainty

Your file can have more columns, and all will be uploaded, but this is the minimum set of fields required to properly integrate the records into MOL. If you don't have any of these columns because a common value is shared by all records (say, all your observations were made the same date, or belong to the same species), the tool has a special section to insert default values for any of these fields.

Lastly, when your dataset has been parsed but before upload, you will be asked to fill a simple form for specifying metadata associated to the dataset (who collected the records, to what project it belongs, if any...). After that, your data will be uploaded to MOL and you will be able to see the records on a map and as a table.

Point (and list) Uploader Key Points

- The point uploader is the first of many upload tools
- Metadata about datasets provides critical ownership and rights data
- As well as essential content for better use of the datasets e.g. probabilistic assessment of absences
- Data provided may be kept private for use
- Or made available for broader use, curation, improvement
- Metadata also provides value for downstream modeling (more expert versus novice data)



TAXONOMY “TRIVIAL BUT TERRIFYING” ISSUES

Whether citizen science data or data from museum records, data cleaning before import is important and provides value for providers and consumers

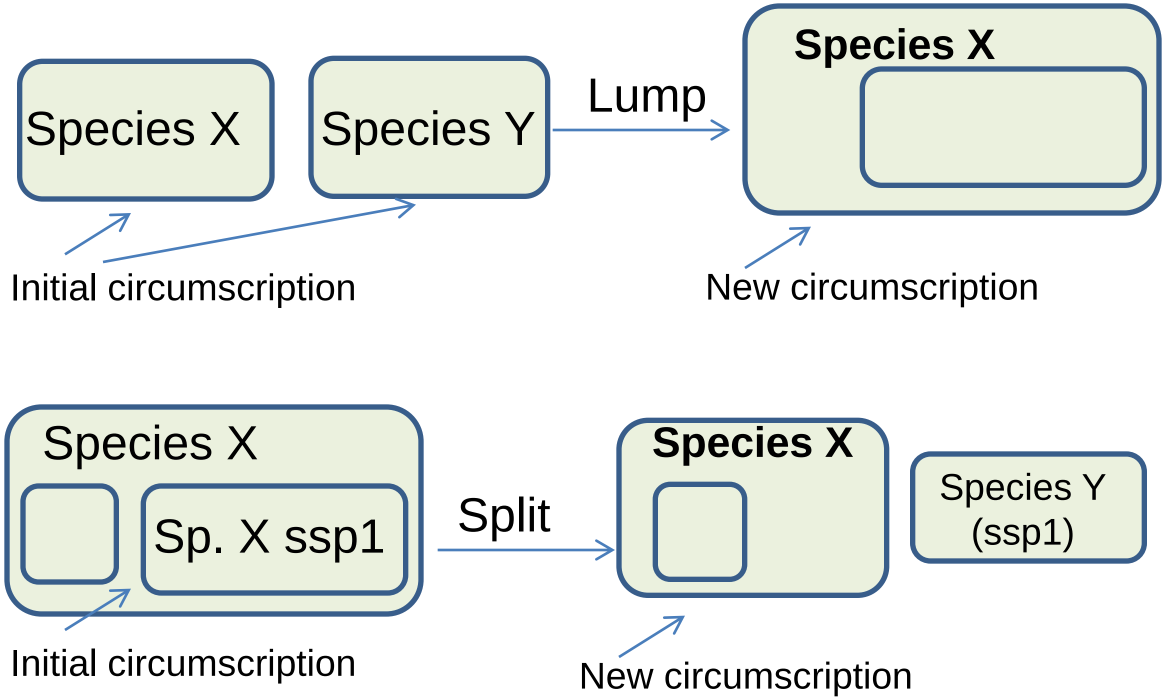
Based on a gold standard, hand vetted set of 500 museum digitized label data in VertNet:

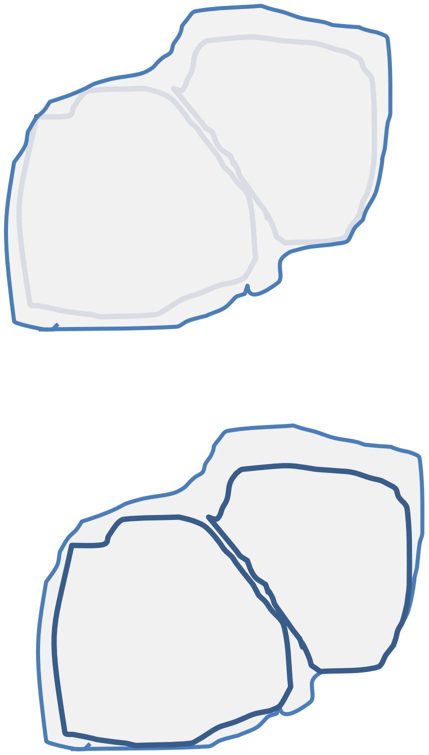
- 7.8% of scientific names could not be resolved at all
- 32% of names are unaccepted but could be resolved to accepted names
- 2.6% are misspelled and unaccepted names
- 10% are misspellings of accepted names
- 47.6% are current accepted names

Take home: Huge issues with ingested data, requiring novel solutions

The non-trival problem-

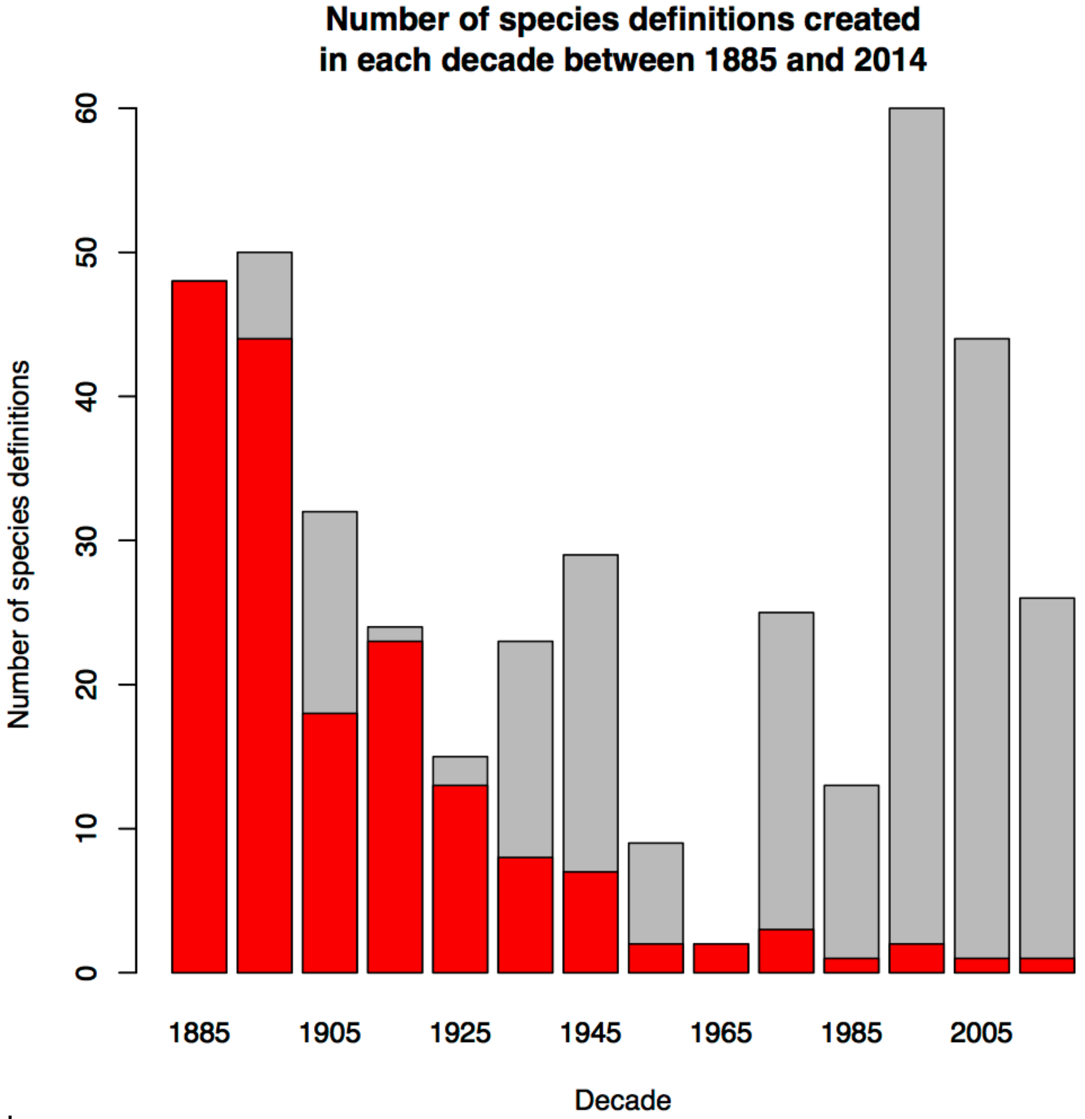
tracking how naming **meanings** change





Geographic range outcome

This is not a problem of “old records” or just “some groups”



Taxonomic effort
In **North American birds** 1885-2015

Primary descriptions

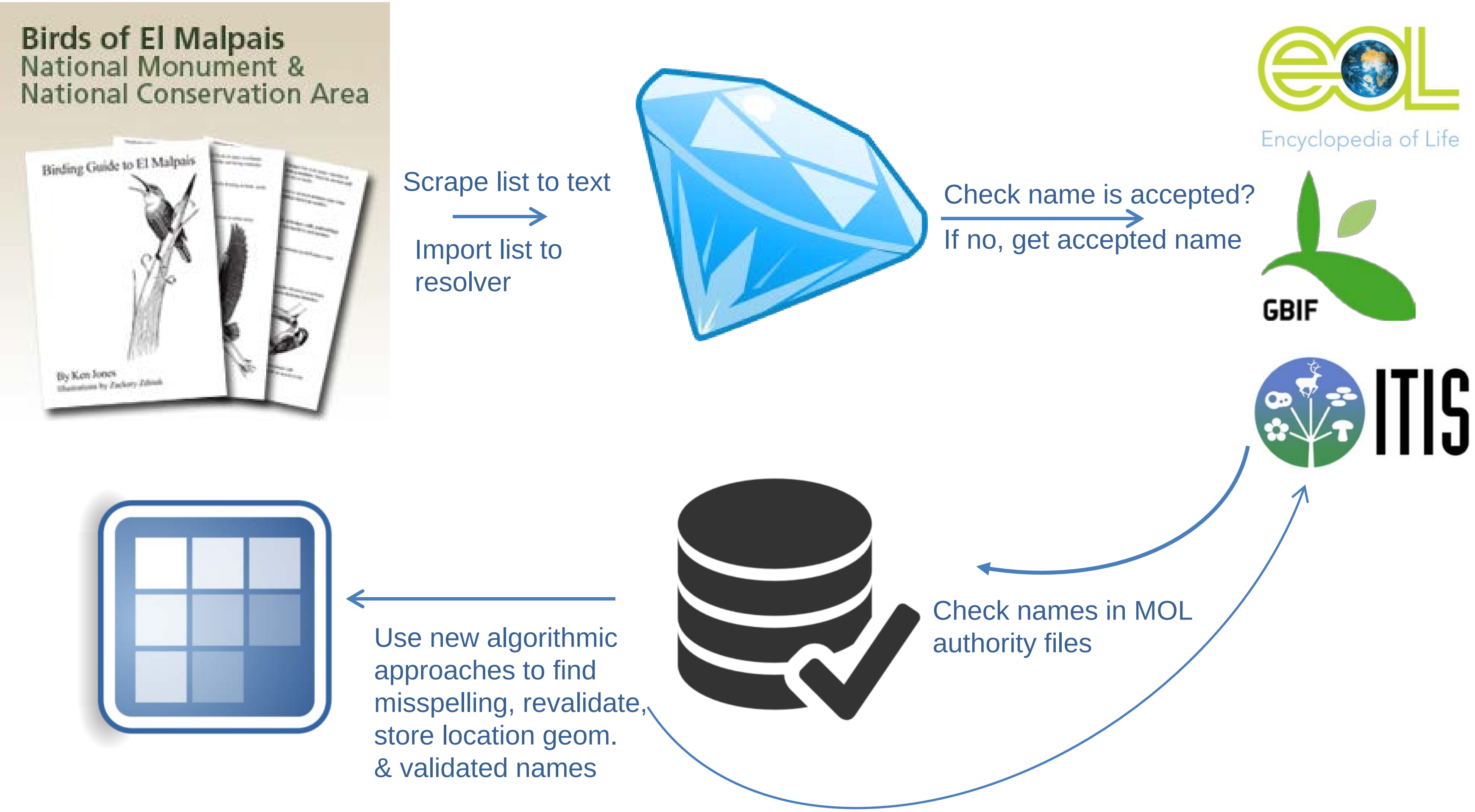
Redescriptions that change taxon concepts

* Based on AOU checklist (a conservative assessment)

Big Challenges working with Names –

Reconciling Names on Ingest

(for those resources where name validation has been inconsistent/problematic)



Scrape list to text

Import list to resolver

Check name is accepted?

If no, get accepted name

Check names in MOL authority files

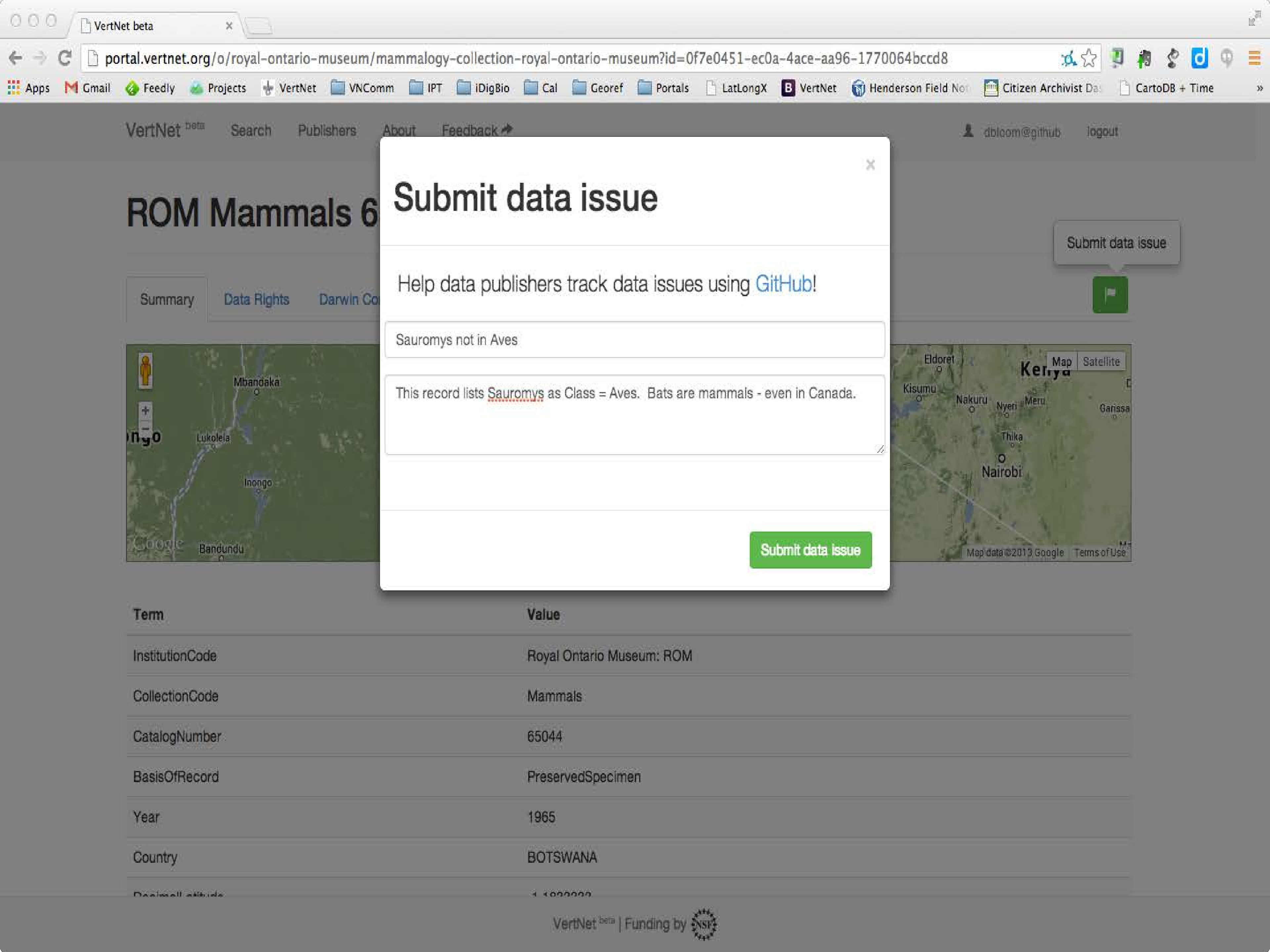
Use new algorithmic approaches to find misspelling, revalidate, store location geom. & validated names

Feature	Total
Total records	**200183168 (100.00%)**
(A) Geospatial issues	
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 issues	
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%)

A global assessment of terrestrial vertebrates using GBIF records. Data from Otegui and Guralnick

CLEANING IS NOT A ONE STEP PROCESS

... It is a constant process of further refining ...



ROM Mammals 6

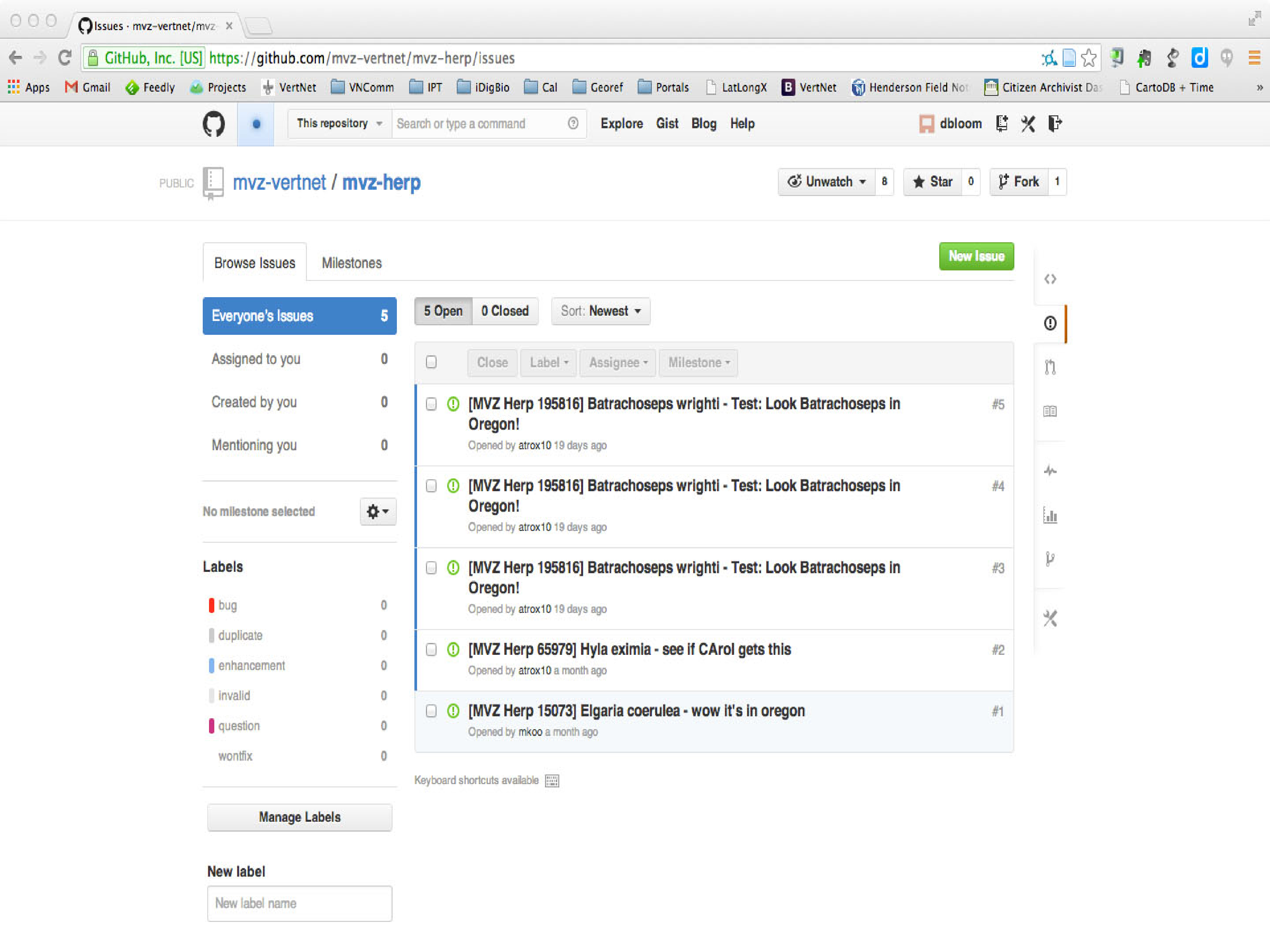
Submit data issue

Help data publishers track data issues using GitHub!

Sauromys not in Aves

This record lists Sauromys as Class = Aves. Bats are mammals - even in Canada.

Term	Value
InstitutionCode	Royal Ontario Museum: ROM
CollectionCode	Mammals
CatalogNumber	65044
BasisOfRecord	PreservedSpecimen
Year	1965
Country	BOTSWANA
DecimalLatitude	-1.1833333



mvz-vertnet / mvz-herp

Everyone's Issues	5
Assigned to you	0
Created by you	0
Mentioning you	0

No milestone selected

Labels

bug	0
duplicate	0
enhancement	0
invalid	0
question	0
wontfix	0

5 Open 0 Closed Sort: Newest

- □ [MVZ Herp 195816] Batrachoseps wrighti - Test: Look Batrachoseps in Oregon! #5
  Opened by atrox10 19 days ago
- □ [MVZ Herp 195816] Batrachoseps wrighti - Test: Look Batrachoseps in Oregon! #4
  Opened by atrox10 19 days ago
- □ [MVZ Herp 195816] Batrachoseps wrighti - Test: Look Batrachoseps in Oregon! #3
  Opened by atrox10 19 days ago
- □ [MVZ Herp 65979] Hyla eximia - see if CArol gets this #2
  Opened by atrox10 a month ago
- □ [MVZ Herp 15073] Elgaria coerulea - wow it's in oregon #1
  Opened by mkoo a month ago

WHY IT ALL MATTERS

- Reintegration of disparate data critical

  (but so is improving those data)
- The data and communities assembling data are *highly* heterogeneous and disconnected
- The data sciences components are *not trivial*.
- Map of Life provides tools for **ALL** to provision data, metadata and provide innovative tools to help curate & improve it
- To better serve needs for monitoring and assessment