DATA and KNOWLEDGE INFRASTRUCTURE

ANALYTICS, KNOWLEDGE DISSEMINATION
Taxon: Andromeda glaucophylla
Notes: Lost Bog, Sec. 22, T38N, Lagrange County, Indiana,

Contributors from multiple sectors add to Map of Life
Consumers use Map of Life knowledge for societal needs.
<table>
<thead>
<tr>
<th>Description</th>
<th>Example</th>
<th>Contribut., Quality</th>
<th>Protocols</th>
<th>Effort Report</th>
<th>Source</th>
<th>Raw data</th>
<th>Temporal Scope</th>
<th>Geographical scope</th>
<th>Reporting basis</th>
<th>Complete &amp; suited absence inferences</th>
<th>Suited input occupy.</th>
<th>Suited Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary inventory (limited protocol &amp; effort report)</td>
<td>Protected area species list</td>
<td>many, heterogen.</td>
<td>multi, unclear</td>
<td>Very limited</td>
<td>Literature</td>
<td>no</td>
<td>long (years)</td>
<td>Clear (often &gt;1km)</td>
<td>Multi (observation, Photo, Lit.)</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Summary inventory (some protocol and effort reporting)</td>
<td>Standardized area survey (e.g. atlas grid cell)</td>
<td>many, heterogen.</td>
<td>multiple single, clear</td>
<td>Possible</td>
<td>Literature, Project reporting</td>
<td>no</td>
<td>long (months, years)</td>
<td>Clear (often &gt;1km)</td>
<td>Single (e.g. Observation)</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Single person/group inventory: observation</td>
<td>Standardized area survey (e.g. transect count)</td>
<td>single, high &amp; vetted</td>
<td>single, clear</td>
<td>Yes</td>
<td>Project data</td>
<td>yes</td>
<td>short (hours, days)</td>
<td>Clear (to 1km)</td>
<td>Single (e.g. Observation)</td>
<td>no</td>
<td>Potentially</td>
<td></td>
</tr>
<tr>
<td>Inventory following protocol: stationary trapping</td>
<td>Camera traps &amp; more typical trappings</td>
<td>single, high &amp; vetted</td>
<td>single, somewhat clear</td>
<td>Yes</td>
<td>Project data</td>
<td>yes</td>
<td>short (hours, days)</td>
<td>Very small (meters)</td>
<td>Single (e.g. Observation)</td>
<td>somewhat (over very small extent)</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Inventory following protocol: active sampling campaigns</td>
<td>fish, zooplankton netting, algal sampling</td>
<td>single, high &amp; vetted</td>
<td>single, clear</td>
<td>Yes</td>
<td>Project data</td>
<td>yes</td>
<td>usually short</td>
<td>Small (e.g. meters)</td>
<td>Single (e.g. Observation)</td>
<td>no</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Full inventory following very defined protocol</td>
<td>CTFS forest plot, Revelle plots</td>
<td>single, high &amp; vetted</td>
<td>single, clear</td>
<td>Perfect: full coverage</td>
<td>Project data</td>
<td>yes</td>
<td>short</td>
<td>V. small (e.g. meters)</td>
<td>Single (e.g. Observation)</td>
<td>yes (over very small extent)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Inventory following loose protocol: citizen science observation</td>
<td>transect</td>
<td>single, heterogen., unvetted</td>
<td>single, clear</td>
<td>Yes</td>
<td>Project data</td>
<td>yes</td>
<td>short (hours, days)</td>
<td>Clear (to 1km)</td>
<td>Single (e.g. Observation)</td>
<td>no</td>
<td>Potentially</td>
<td></td>
</tr>
</tbody>
</table>
## A metadata schema for collating data from inventories

<table>
<thead>
<tr>
<th>Humboldt Core Version 1</th>
<th>Area species checklists</th>
<th>Geographically restricted surveys</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gridded Atlas survey</td>
<td>Transect count</td>
</tr>
<tr>
<td></td>
<td>Protected area species list</td>
<td>Trapping and netting</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CTFS forest, Revelle plots</td>
</tr>
</tbody>
</table>

### General dataset & identification terms
- Inventory performed 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

<table>
<thead>
<tr>
<th>Humbolt Core Term</th>
<th>Possible Values</th>
<th>Percents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compilation effort</td>
<td>Low, medium, high, na</td>
<td>Low – 38.7%, Medium – 6.2%, High – 8.5%, n.a – 45.7%</td>
</tr>
<tr>
<td>Abundances reported</td>
<td>Yes/No</td>
<td>Yes-57.5% No-42.6</td>
</tr>
<tr>
<td>Absences reported</td>
<td>Yes/No</td>
<td>Yes-41.8 , No-58.2%</td>
</tr>
<tr>
<td>Completeness assessment</td>
<td>Scale in 25 percent increments from 0-100</td>
<td>50-100% complete – 30.3%, 75-100% - 27.9%, other- 41.8%</td>
</tr>
</tbody>
</table>
A tool for the long-tail data on YOUR computer

Observation data loader

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.
• 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-trivial problem—tracking how naming meanings change

Species X  Species Y
Initial circumscription

Lump

Species X
New circumscription

Species X
Species Y (ssp1)

Split

Species X
New circumscription

Initial circumscription

Geographic range outcome
This is not a problem of “old records” or just “some groups”.

Number of species definitions created in each decade between 1885 and 2014

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

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 ...
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 | 1955
Country | BOTSWANA
[MVZ Herp 195816] Batrachoseps wrightii - Test: Look Batrachoseps in Oregon!
Opened by atrox 19 days ago

[ MVZ Herp 195816 ] Batrachoseps wrightii - Test: Look Batrachoseps in Oregon!
Opened by atrox 19 days ago

[ MVZ Herp 195816 ] Batrachoseps wrightii - Test: Look Batrachoseps in Oregon!
Opened by atrox 19 days ago

[MVZ Herp 65979] Hyla eximia - see if CARol gets this
Opened by atrox a month ago

[MVZ Herp 15073] Elgaria coerula - wow it's in oregon
Opened by mico a month ago
- 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