Better quality, less work: How to improve collections data with the efficient use of resources provided by aggregators and consortia

Erica R. Krimmel
The Chicago Academy of Sciences / Peggy Notebaert Nature Museum

Identifying opportunities

Digitizing specimen data both improves accessibility of biodiversity collections to external audiences, and highlights opportunities for improving internal data quality. Here we demonstrate the latter through the data management experience of The Chicago Academy of Sciences / Peggy Notebaert Nature Museum (CAS/PNNM) as it has collaborated with various collections data consortia and with biodiversity data aggregators (see Figure 1).

Consortia provide a broad base for comparing and cleaning collections data. Consortia also facilitate discussion over data standards and representation at a level that many collection managers and curators may feel more comfortable entering, as opposed to, e.g., similar discussions happening at the level of data standards governing bodies. CAS/PNNM benefits from being a member of biodiversity consortia including Arctos (*www.arctosdb.org*), and the InvertEBase Thematic Collections Network (*www.invertebase.org*). Biodiversity data aggregators, e.g. iDigBio (*www.idigbio.org*) or VertNet (*www.vertnet.org*), provide a variety of tools for mobilizing specimen data, and for improving the quality of data once published.

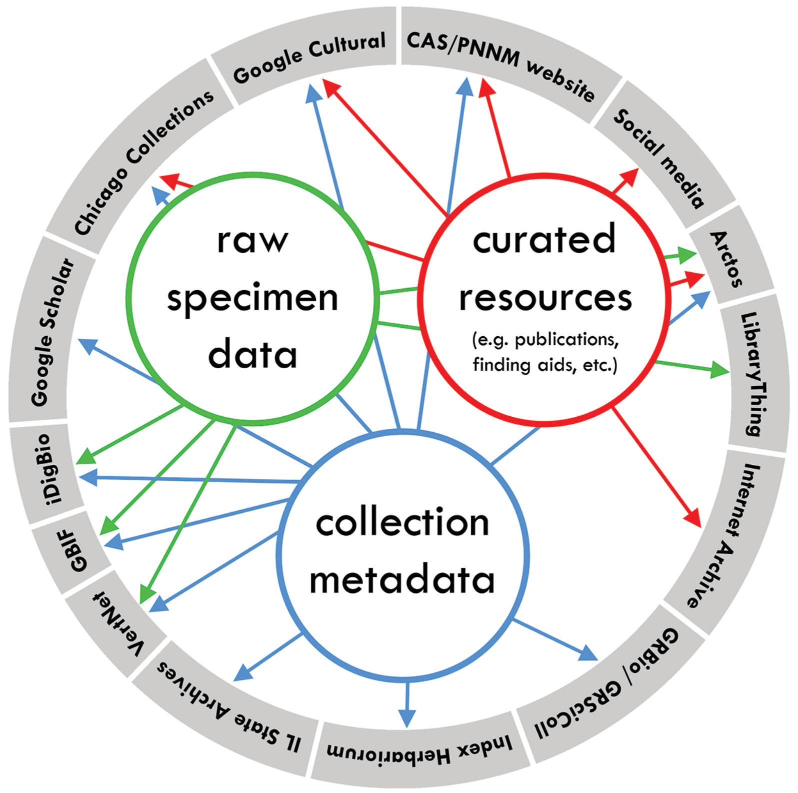

Figure 1. Map of CAS/PNNM collections collaborations and presence online, not yet updated to include collections metadata and raw specimen data on the InvertEBase Symbiota portal. In thinking about how to effectively use the resources available to us through consortia and aggregators we make every effort to centralize our digital resource management in Arctos.

Identifying priorities

Digitization at CAS/PNNM began in earnest from 2008-2012 when we digitally inventoried data from 16 biological and cultural collection disciplines into 50+ Excel spreadsheets. We began migrating this data to the Arctos collection management system in 2015, and in the process have been focusing on making specimens and their data more **complete**, more **accurate**, more **discoverable**, and more **research-ready**.

We use OpenRefine as our primary tool for cleaning data because it is designed to view and transform large amounts of data via reproducible operations. Additionally, OpenRefine simplifies the use of web services provided by online databases, e.g. it can look up the county for thousands of municipalities with just a few lines of code. For common in-house operations, such as converting dates to the ISO8601 format, we use OpenRefine to create and run JSON scripts. We also use OpenRefine to reconcile data against our own internal lookups, an approach similarly used by toolkits such as Kurator (*wiki.datakurator.org*).

Here we focus on resources we use in combination with OpenRefine that are relevant to improving the quality of data about people, places, and taxonomy—three common entry points for biodiversity data users.

CAS/PNNM OpenRefine scripts and workflow documentation are available online: *https://github.com/ChicagoAcademyofSciences/data-cleaning*

Data about People

Collectors, identifiers, and other people associated with specimens provide a link to knowledge previously generated from the specimen, typically via research publications. Agent information is often only marginally standardized as it may be difficult or impossible to know who incomplete verbatim names refer to. CAS/PNNM, however, is a small collection with rich institutional history, and we are often able to augment agent data that would appear skeletal out of context (e.g. in our ornithology collection "Deane" always refers to Ruthven V. Deane). To streamline and standardize this augmentation process, we have created a master agent lookup table for people whose names appear anywhere in our collection data.

RESOURCES:

→ Tools from Arctos help us to format names and compare them with existing Arctos agents. Because CAS/PNNM formerly exchanged specimens with institutions across the country, many of our agent names already exist in Arctos and we are able to connect our specimens to the existing agent as well as all of its associated metadata, relationships, and media (see example in Figure 2).

Figure 2. Images from a 1940 Chicago Academy of Sciences collecting expedition to Arizona, led by Howar K. Gloyd. Gloyd is an example of an agent who occurs across institutions in Arctos as a collector, identifier, donor, and research publication author.

Data about Taxonomy

Taxonomy is possibly the most common entry point for data users, and also the most challenging for CAS/PNNM to make discoverable. As a small collection with limited taxonomic expertise in house, we rely heavily on vetted taxon authorities, such as the Integrated Taxonomic Information Service (ITIS, *www.itis.gov*), the World Register of Marine Species (WoRMS, *www.marinespecies.org*), the International Plant Names Index (IPNI, *www.ipni.org*), etc. Our goal is to map our taxonomic data to taxonomic concepts supported by such online databases that have web services capable of (semi-)automating nomenclature updates for us in the future.

Despite the plethora of taxonomic resolution services offered by aggregators and consortia, accomplishing the goal above is easier said than done, as illustrated by Figure 3. In collections less recently curated, CAS/PNNM specimens are frequently identified with century-old nomenclature. These names are difficult to synonymize with more recent taxonomy due to alternate spellings and short-lived species concepts.

The community as a whole also faces significant challenges to making taxonomy a discoverable data entry point because taxonomic names are a reflection of current knowledge rather than a statement of standardized fact, as spatial data or collector names generally are.

RESOURCES:

→ Web services from aggregators and taxon authorities to resolve invalid taxonomic names.

→ Web services from aggregators and taxon authorities to fill in higher taxonomy and other associated data for taxa that have been resolved.

→ Access to taxonomic experts via InvertEBase, for advice on the quality of various online taxon authorities, as well as on specific taxonomic issues in our data.

→ Robust support for multiple identifications in Arctos, so that we can record and track provenance of all taxonomic names whether they are valid or not.

→ Data quality improvement to taxonomic identifications via annotation features in VertNet and in Arctos.

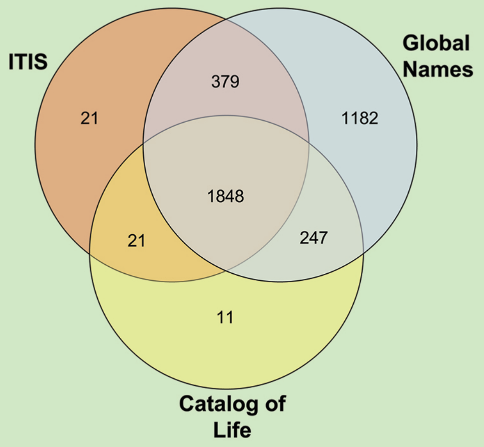

Figure 3. Disparities in taxonomic resolution between the Integrated Taxonomic Information Service (ITIS), Global Names, and Catalog of Life. Out of 5,616 unique taxon names in our non-Lepidopteran entomology collection, only 66% (3709 names) could be resolved semi-automatically via web services from these three sources. Even within this resolvable portion, we found significant differences in which source/s recognize which names.

Data about Places

Spatial data connects specimens over time and across taxa, and is critical for many areas of biodiversity research. Standardizing descriptive spatial data is an important step in our data migration, and sets the stage for efficient georeferencing in the future.

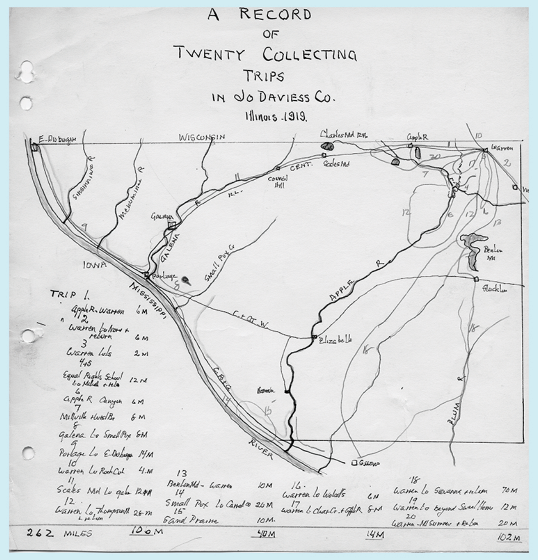
A Record of Twenty Collecting Trips in Jo Daviess Co. Illinois 1919.
Wisconsin
Iowa
Mississippi River
Galena
Warren
Elizabeth
Trip 1.
262 Miles

Figure 4. Example of ancilliary data from the CAS/PNNM archives detailing exact collecting localities of specimens. These field notes were recorded by Herman Silas Pepoon, "A record of twenty collecting trips in Jo Daviess Co., Illinois 1919."

In order to standardize descriptive spatial data without losing the granularity and authority that verbatim data provides, we have created an internal lookup for CAS/PNNM geography across all collections. This is particularly useful for information below the county level, including municipalities, neighborhoods, natural features, etc.

Georeferencing is arguably one of the most useful additions an institution can make to increase the research-readiness of its data. However, quality of georeferencing can be highly variable due to the georeferencing protocol used, the familiarity of the person doing the work with the area being georeferenced, etc. Many of the collecting localities in CAS/PNNM data are local to the Chicago region, supported by ancillary materials in our archives, or otherwise related to our institutional history (see Figure 4). Because of our familiarity with this context, we are able to provide higher quality georeference data than any automatic or semi-automatic service could do for such specimens.

RESOURCES:

→ Web services from the Getty Thesaurus of Geographic Names, Google's Geocoding API, etc. to resolve higher geography (county level and above).

→ Arctos lookups to format and standardize higher geography, as well as more specific localities.

→ Collaborative GeoLocate georeferencing tools via the InvertEBase Symbiota portal. Because collectors within a discipline often visit the same areas, pooling localities between institutions and collaboratively georeferencing them can greatly increase our efficiency.

→ Arctos data model, which shares localities across all institutions in the consortia, thereby ensuring that if one specimen is georeferenced, that georeference data exists for all other specimens using the same locality (see Figure 5).

→ Data quality flags from aggregators, e.g. iDigBio's "geocode_mismatch" flag alerts us to specimens with georeferenced lat/lon coordinates that do not match the country data provided.

→ Data quality improvement to spatial information via annotation features in VertNet and in Arctos.

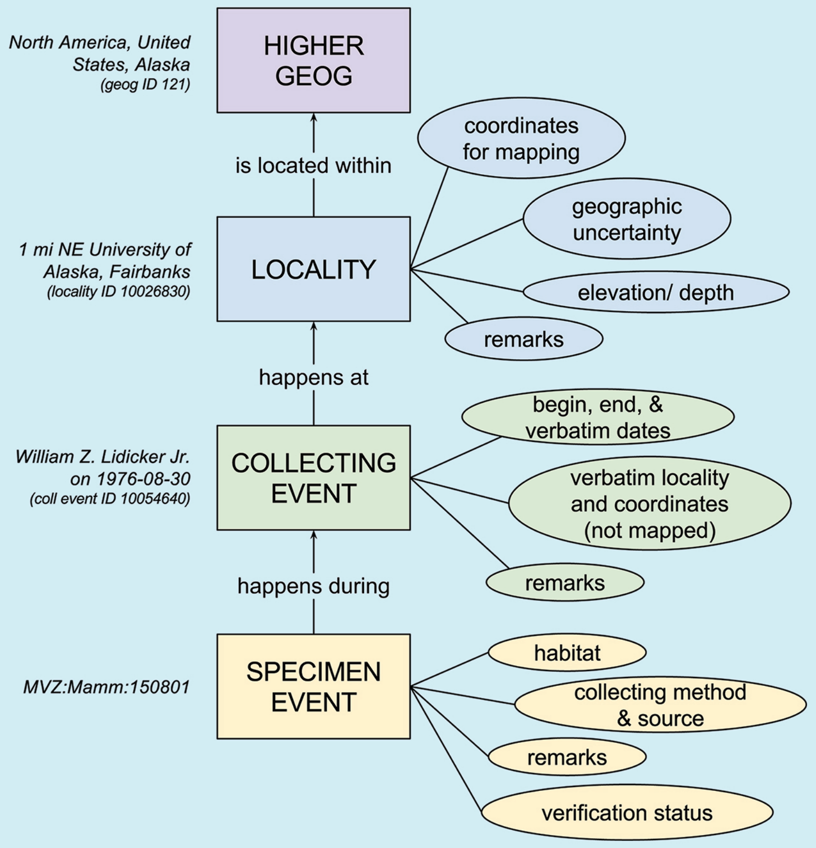
North America, United States, Alaska (geog ID 121)
HIGHER GEOG
is located within
coordinates for mapping
geographic uncertainty
elevation/ depth
remarks
1 mi NE University of Alaska, Fairbanks (locality ID 10026830)
LOCALITY
happens at
begin, end, & verbatim dates
verbatim locality and coordinates (not mapped)
remarks
William Z. Lidicker Jr. on 1976-08-30 (coll event ID 10054640)
COLLECTING EVENT
happens during
habitat
collecting method & source
remarks
verification status
MVZ:Mamm:150801
SPECIMEN EVENT

Figure 5. Diagram visualizing how the Arctos database captures locality data. Courtesy of *http://handbook.arctosdb.org/how_to/How-to-understand-the-Arctos-Locality-Model.html.*

Much thanks to Dawn Roberts and Samuel Howes, also at CAS/PNNM, for their collaboration in our data cleaning and migration process. Additional thanks to the consortia and aggregators mentioned here. This material is based in part on work supported by the National Science Foundation under Grant No. DBI 1601700. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.