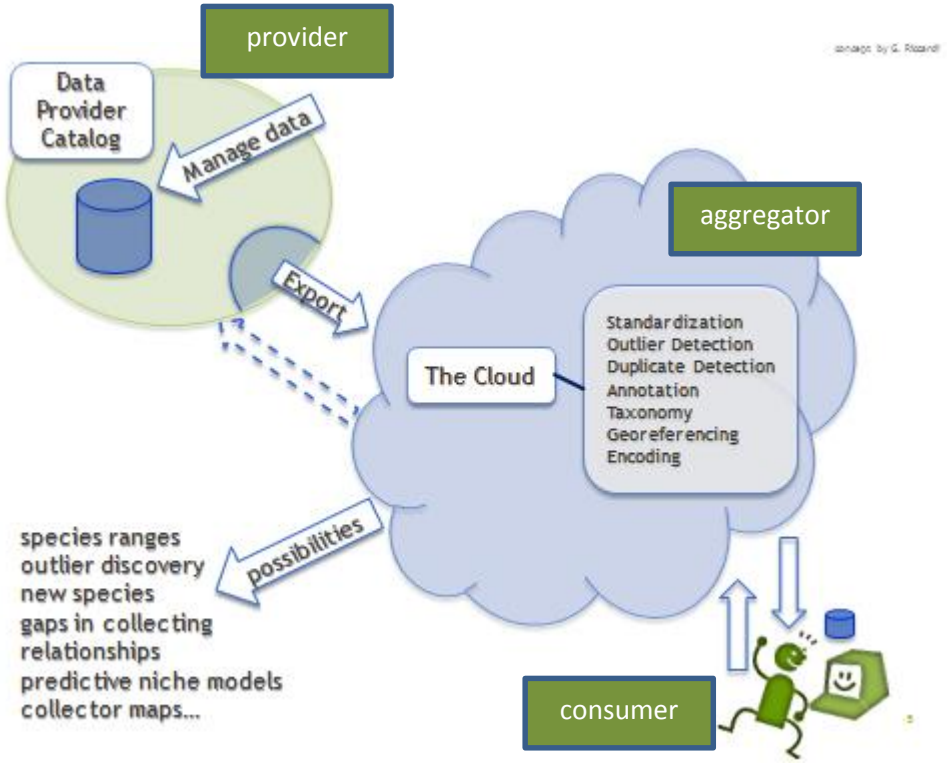
Data Requirements Workshop Report
Part 1: Needs of Providers and Consumers

Greg Riccardi
Florida State University
griccardi@fsu.edu

An iDigBio workshop in March 2014 considered data and service requirements for aggregators of biodiversity information. The workshop attendees were tasked with identifying needs of providers and consumers of biodiversity information. A subset of attendees (the writing group) was tasked with interpreting these requirements as they relate to aggregators. This document is a result of the discussions.

The initial session of the workshop brought the writing group together to prepare a list of topics for discussion. Two plenary sessions followed: the first discussed these and other topics from the perspective of providers and the second from the perspective of consumers. The final session allowed the writing group to discuss what aggregators need to do to meet the needs of providers and consumers.

This report is based on a model of interaction between providers of specimen data who manage primary data, consumers of specimen data who discover, acquire, filter and analyze data, and aggregators who acquire data from many sources, provide discovery and download services, and serve as the agent for feedback among users.



iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). 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.

Workshop attendees

Writing group

Greg Riccardi (iDigBio, Florida State U.)
Reed Beaman (iDigBio, U. of Florida)
Donald Hobern (GBIF)
Rich Pyle (Global Names Architecture, Bishop Museum)
Robert Whitton (Global Names Architecture, Bishop Museum)
Paul Flemons (Australian Museum, Sydney)
James Macklin (Agri-Food Canada)

Plenary Group

Joanna McCaffrey (iDigBio, U. of Florida)
Deb Paul (iDigBio, Florida State U.)
Neil Evenhuis (Global Names Architecture, Bishop)
Shelley James (Macro Algae TCN, Bishop Museum)
Michael Thomas (U. Hawaii)
Chris Neefus (Macro Algae TCN, U. of New Hampshire)
Matt Goodale (National Tropical Botanical Garden)
Tom Schils (Biodiversity Informatics Manager, U. of Guam)
Aubrey Moore (Entomology, U. of Guam)
Ryan Caesar (Entomology, U. Hawaii)

Needs of providers

Consistency and transparency of data

Providers

Attribution for data use

"Attribution" refers to information about the origin of something. Attribution for the use of data has great benefits to the providers of that data. Providers are particularly concerned about getting credit for the use of their data. Publication of results whether online or in print that carries links to the sources of the data used gives credibility to the provider. Giving credit to providers is only possible if the data carries information about its source. In addition, data analysis tools need to preserve this source information.

Providers need additional help in reading, processing and tracking attribution. Specimen database systems could be modified to ingest and manage attribution information that is available from aggregators and consumers. No end-to-end implementation of attribution is currently available. [Dryad reference]

Help with managing taxonomy

Many providers are unsure of the best current taxonomic names and concepts for their specimen determinations. They would greatly benefit from timely and accurate sources of scientific names and

classifications. They also need tools to migrate data to the current taxonomy and to import taxonomy into their specimen databases.

Feedback from determination and data cleaning

Consumers of data analyze specimen information to determine fitness for their use. That analysis often involves evaluation of taxonomic determinations, assessment of the accuracy of locality description, and inspection of images of specimens to determine characteristics like leaf shape, length of femur, etc. These analyses produce important new data that should be made available as feedback to providers and other consumers.

Systems like Filtered Push [REF] provide mechanisms for collecting and distributing this new data. The information can be represented in a standard form and stored in a repository. Interested parties can subscribe to information feeds tailored to their interests. The feeds provide a mechanism for downloading the feedback.

The effort required to process feedback will be considerable. Most specimen database tools are not equipped to download this type of data and help curators incorporate the data. New tools and enhancement of existing tools are required to provide end-to-end feedback mechanisms.

Help with identifiers

Globally unique identifiers for specimens, images and other objects are necessary for the distribution of specimen information and the proper functioning of attribution and feedback systems. Providers should attach suitable identifiers to data objects as part of data curation and include those identifiers in every data export operation.

Many providers are mystified by the whys and how of identifier management. They want to choose appropriate, sustainable technology for identifiers and partner with aggregators to make sure that an identifier can be used to acquire the specimen information. They need explicit, prescriptive instructions for how to create and implement an identifier policy.

Registries for people and localities

Lists of people who collected specimens and the places they were collected are valuable across collections. The systematic sharing of these lists with aggregators will allow providers to acquire and use them. These lists are especially important for transcription of labels and data cleaning.

Needs of Consumers

Good global information discovery services

Consumers of data want to find the most accurate and complete dataset possible. The acquisition of datasets from aggregators should include data from the providers as well as information that has been created by consumers and aggregators through assessments and analysis of primary information.

Taxonomic search

Consumers need to search by taxonomic terms and find all specimens that match. The search should be informed by synonymy, spelling, variant forms and classification. A search by higher taxon (e.g. family) should

find all objects in that subtree. The search should not be restricted to the exact scientific name given by the provider.

Assessments of data quality, per record or dataset

Each consumer of information needs assessments of the fitness for the intended use. This includes whatever is known about the quality of records and datasets as well as examples of prior use of the data. The attribution and feedback described above are just as useful for consumers as for providers.

As with attribution, the end-to-end solutions that record assessments in standard forms and make them available to all interested parties are not currently available.

Data cleaning services

Data cleaning includes finding misspellings and synonyms for scientific names, establishing geolocation from locality data, and finding errors in geolocation.

Good tools for data cleaning have been developed for GBIF, Vertnet and other systems. Consumers would benefit from access to services or installable libraries for data cleaning.

Tools to find related data, e.g. sequences

Consumers need to be able to link specimen data acquired from aggregators to data in other repositories, such as GenBank. It is typical to link information by scientific name and many repositories (EOL, Discover Life) use scientific name for their primary organization. Specimen-based research requires linking information by specimen, a capability that relies on good identifiers and identifier practices.

Tools to aid in integration of data from multiple sources

When consumers get data from multiple sources using different formats and data standards, they have trouble with integration. GBIF and iDigBio provide primary data in Darwin Core Archive format, which makes integration of the Darwin Core fields simpler. There continues to be a lack of tools in the community to find related data and to integrate heterogeneous data sets.