

Data Workshop

Honolulu March 2014

Greg Riccardi
Florida State University
iDigBio
griccardi@fsu.edu



This material is based upon work supported by the National Science Foundation under 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.



Goals of workshop

- Draft a requirements document for aggregators that describes information and services that are crucial to the success of biodiversity informatics.
- Writing group sessions
 - Discussion of issues for plenary
- Plenary session
 - Discussing issues from the perspective of
 - Providers
 - Users

Draft list of issues for discussion

- Full record-level information discovery and delivery
- Metadata harvesting protocols
- GUID per record with persistence
- Attribution metadata with all data records
- Media information ala Audubon Core
- Bi-directional portal
- Feedback from data users to providers (e.g. data quality)
- Usage analytics
- Attribution to providers from analysis
- Annotation management
- Active repository technology (incremental updates)

Plenary Discussion Summary

- Providers need
 - Attribution for data use
 - Help with managing taxonomy
 - Feedback from determination and data cleaning
 - The effort required to process feedback will be considerable
 - Tools are needed to help providers with feedback
 - Details about determinations and geolocations
 - Help with identifiers
 - Registries for people and localities

Plenary Discussion Summary

- Users need
 - Good global information discovery services
 - Assessments of data quality, per record or dataset
 - Data cleaning services
 - Feedback for data cleaning so that improvements are made by providers
 - Tools to find related data, e.g. sequences
 - Tools to aid in integration of data from multiple sources

Information Integrity and Attribution

- Provenance tracking
 - Keep track of source of information
 - Ensure information is not changed
 - Control versions
- Attribution
 - Keep track of delivery of information
 - Make attribution information available to providers
 - Provide mechanism for users to report
 - Publications
 - Derived data
 - Evaluation and corrections

Identifier and identifier services

- Encourage identifiers for objects
 - Require stable identifiers
 - Provider must commit to consistent use of identifiers
 - Strongly suggest that providers maintain GUIDs
 - Add GUIDs as necessary
- Identifier services
 - Return metadata document upon request
 - Discover and maintain relationships among identifiers
 - E.g. If a provider changes the identifier, the aggregator must record that the old and new identifiers are equivalent

Search and Discovery

- Search by common properties
- Discover across object types
 - E.g. Find image by scientific name or geography of specimen
- Provide for download in common formats
- Provide APIs for search and download

Taxonomic services

- Assumption
 - Provider sends scientific name and possibly higher classification
- Externalize taxonomic names and classifications
 - Participate in shared services
- Allow discovery beyond name string
 - Synonyms
 - Common names
 - Higher taxa

Dealing with extended schema

- Assumption:
 - Providers will have important information that is not Darwin Core
- Properties
 - Keep track of properties and evaluate new data sets for new properties
 - Allow both literal- and resource-valued (relationship)
- Transformations and normal forms
 - Maintain information content when changing formats
 - Transform or coalesce properties according to community standards
- extending schemas
 - traits, measurements, interactions
- property similarity with respect to discovery

Thanks to all participants

- Writing group
 - Greg Riccardi (iDigBio)
 - Reed Beaman (iDigBio)
 - Donald Hobern (GBIF)
 - Rich Pyle (GNA, Bishop)
 - Robert Whitton (GNA, Bishop)
 - Paul Flemons (Biodiversity Info Manager, Au)
 - James Macklin (Biodiversity Info Manager, Can)
- Plenary Group
 - Joanna McCaffrey (iDigBio)
 - Deb Paul (iDigBio)
 - Neil Evenhuis (GNA, Bishop)
 - Shelley James (Bishop, Macro Algae)
 - Michael Thomas (Biodiversity Informatics Manager, U. Hawaii)
 - Chris Neefus (Macro Algae)
 - Matt Goodale (data management/IT supervisor, NTBG)
 - Tom Schils (Biodiversity Informatics Manager, Botany, phycology, UOG)
 - Aubrey Moore (Biodiversity Informatics Manager, Entomology, UOG)
 - Ryan Caesar (IT Manager/programmer, Entomology, U. Hawaii)
- Others?