Version: May 2013

This guide is intended to describe the formats and requirements for ingesting data into iDigBio. If you have any questions about this guide, or to contact iDigBio to start the process of getting your data into iDigBio, please email info@idigbio.org and an iDigBio staff member will get back to you shortly.

iDigBio Data Ingestion Requirements and Guidelines

Supported File Formats

iDigBio strives to make data ingestion into our infrastructure as easy as possible. To achieve this, we have identified two lowest common denominator export file formats that we will initially support for dataset ingestion.

The first of these formats is CSV, which is available as an export format requiring very little work from most databasing software. When utilizing CSV files, care must be taken with free-form text fields to ensure that all line breaks and quotes are escaped, and all commas within fields are enclosed inside a quoted field. Give a visual inspection to the exported file to verify that any diacritics in the data have been preserved, preferably by encoding to UTF-8.

The second format is the Darwin Core Archive file.

Supported Term Names

Data submitted to iDigBio can be in a variety of formats and can contain virtually any set of fields needed to represent the information present in collections. That said, in order to maximize the reusability of the data, iDigBio does constrain the basic vocabulary used to describe core collections concepts.

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (#EF1115210)

Since iDigBio deals with many different potential types of data, and many standards may apply, we ask that all data element names be formatted using the xml-derived abbreviation syntax:

namespace:termName

where namespace is a known set of defined terms, and termName is the camel case, case-sensitive name of the term within the namespace. For example:

dwc:scientificName is the Darwin Core term for scientific name [1]
ac:caption is the Audubon Core term for caption [2]
dcterms:dateModified is the Dublin Core term for date modified [3]

As a general rule, iDigBio defers to the standards documents themselves for guidance, and fully supports all terms included within the standards. Links to all of the mentioned standards documents can be found at the end of the document.

At its core, iDigBio seeks to capture information about two types of objects, collection object data, and media data relating to those collection object. To this aim, iDigBio relies on two community-developed standards to represent the metadata elements around the data types. For the representation of collection object data, iDigBio relies heavily on Darwin Core [1] to provide field names. For the representation of media metadata, iDigBio relies heavily on Audubon Core [2]. iDigBio has also convened a Minimum Information for Scientific Collections working group [4] that has examined the existing standards and identified gaps in which the needs of the collections community are not met. Until those gaps are filled, iDigBio will also use the work of the MISC working group to support these additional terms within the iDigBio namespace [5].

Data Export Process

Local collections databases have a very different set of requirements than data aggregators such as iDigBio, and as a consequence often have internal structures and field names that do not map directly to those expected by external parties. There are a wide variety of solutions to this problem, both developed within the community and from outside, but the ones that iDigBio is choosing to focus on are those that provide structured, repeatable, automated transformations between source data and export data. These processes help to maintain the consistency and correctness of the data, ensure that data export updates are easy to perform, and minimize the workload on collections managers and curators after the initial setup.

As a long-term goal, iDigBio is working with collections management software providers to integrate standards-compatible automated export tools into their software packages. This level

of integration will provide the most accurate transformations of the data, and free collections managers and curators from having to learn and maintain a separate piece of software to export their data.

In the short term, iDigBio has the following recommendations for automating the data export process.

For collection management staff who use software that depends on a relational database, and who are able to set up and maintain a separate software package and connection to that database, GBIF's IPT [6] is the preferred method of exporting data to iDigBio. There is a wide body of existing knowledge on how best to map various collection databases into IPT, so if you need guidance iDigBio can point you to existing resources.

For collections using less complicated data storage methods, or whose databases are only capable of exporting CSV files in specific formats, the OpenRefine project [7] (formerly Google Refine) is an excellent way to provide a documented, repeatable, semi-automated transformation between your local CSV format and field names and the format and field names required by iDigBio. We hope to generate some example conversions that you could base your own workflows on in the near future.

Getting Your Data to iDigBio

If you are using IPT to export your data, and your IPT instance is internet accessible, iDigBio can use IPT's built-in RSS feed generator to detect and download updates to your dataset as they become available. For those using other export methods, including CSV, who can host files on a web server, iDigBio can provide guidance on how to configure a simple RSS feed generator to provide these files to iDigBio in much the same way as IPT.

For those with small numbers of exported data files, iDigBio can also accept direct links to files for downloading, which will be checked periodically for updated versions.

For those without the ability to host files on a web server, updates can be emailed to iDigBio for the time being. In the future, we will replace email with an interface for you to upload new versions of your dataset directly to iDigBio for processing.

Endnotes

1. Darwin Core: http://rs.tdwg.org/dwc/terms/index.htm
2. Audubon Core: http://terms.gbif.org/wiki/Audubon_Core_Term_List_(1.0_normative)
3. Dublin Core: http://dublincore.org/documents/dcmi-terms/

4. MISC Wiki: https://www.idigbio.org/wiki/index.php/MISC/Authority-File-Working-Group
5. iDigBio Namepsace: http://portal.idigbio.org/terms/
6. GBIF IPT: http://code.google.com/p/gbif-providertoolkit/
7. OpenRefine: http://openrefine.org/

Appendices

Appendix A - Darwin Core Archive Processing

When loading records from a Darwin Core Archive, iDigBio's data ingestion process obeys the following rules and procedures. Providers using IPT or another method to generate Darwin Core Archives should be aware of these rules and map their fields accordingly.

- The non-Darwin Core field idigbio:recordID will always be used as the primary record identifier of the data record if it is present.
- In the absence of the idigbio:recordID field, the following fields will be checked for a unique identifier, and it will be used as the idigbio:recordID of the object.
    - ac:providerManagedID in Audubon Core records
    - Record IDs can also be provided in darwin core archives by using the Resource Relationship extension. For iDigBio to recognize a resource relationship row as a valid record ID, it must have the dwc:relationshipOfResource set to "representedIn", and the value in dwc:relatedResourceID must consist solely of a URI or bare UUID. Use of this method for transmitting record IDs requires that the value in the ID column of the core file also be unique within the file.
- All fields whose names end in ID or Id, and which contain URIs or bare UUIDs will be harvested into iDigBio's identifier system, and will be associated with the collection object record to the best of our ability. For identifier harvesting, the id field of the core file will be assumed to the the primary identifier of the core type (for dwc:Occurrence, this is dwc:occurrenceID). The dcterms:identifier field is also harvested if present.
- At this time, iDigBio currently processes records out of files of type dwc:Occurrence and ac:Multimedia. All other extension files will be appended to the core record during processing, under the field name equal to the short name of their type (ex., dwc:ResourceRelationship, dwc:MeasurementOrFact).
- Special relationship processing, (e.g., media records):
    - Media in Audubon Core extension files are automatically associated with the record provided by the coreid they include.
    - ac:relatedResourceID is assumed to be a link to a related object if the identifier matches a record identifier in the iDigBio system.

- ac:associatedSpecimenReference is assumed to be the identifier of a related collection object, if the identifier matches a collection object in the iDigBio system.
- dwc:associatedMedia is assumed to be either the identifier of a media record in the iDigBio system, or an Audubon Core record's dcterms:identifier field if one was provided.
- if the resource relationship extension is present, iDigBio will also attempt to use those relationships to relate to known objects within the iDigBio system.

Appendix B - CSV Processing

When loading records from a CSV file, iDigBio's data ingestion process obeys the following rules and procedures.

- The non-Darwin Core field idigbio:recordID must be present in all CSV files, contain a URI or bare UUID, and must be globally unique.
- Any fields whose name ends in ID or Id, and which contains a URI or UUID will be harvested into iDigBio's identifier system, and will be associated with the collection object record to the best of our ability. The dcterms:identifier field is also harvested if present.
- At this time, iDigBio currently processes records out of files of type dwc:Occurrence, dwc:ResourceRelationship, and ac:Multimedia.
- Special relationship processing:
  - ac:relatedResourceID is assumed to be a related collection object, if the identifier matches a record identifier in the iDigBio system.
  - ac:associatedSpecimenReference is assumed to be the identifier of a related collection object, if the identifier matches a collection object in the iDigBio system.
  - dwc:associatedMedia is assumed to be either the identifier of a media record in the iDigBio system, or an Audubon Core record's dcterms:identifier field if one was provided.
  - If the provided CSV is a dwc:resourceRelationship file, iDigBio will attempt to use the stated relationships to relate to known objects within the iDigBio system.

Appendix C - Helpful Links for working with IPT and Darwin Core Archives

The GBIF Integrated Publishing Toolkit User Manual:
http://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes

http://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes
The gbif-providertoolkit google code repository:
http://code.google.com/p/gbif-providertoolkit/
http://code.google.com/p/gbif-providertoolkit/
Darwin Core Archive Assistant:
http://tools.gbif.org/dwca-assistant/
http://tools.gbif.org/dwca-assistant/
Darwin Core Archive Assistant, User Guide:
http://www.gbif.org/orc/?doc_id=2817&l=en
or search for pdf: gbif_dwc-a_asst_en_v1.1-1.pdf

Setting up IPT for the first time:
http://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes#Set_up_the_IPT_for_the_first_time
http://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes - Set_up_the_IPT_for_the_first_time
Darwin Core Archive Validator:
http://tools.gbif.org/dwca-validator/