GUID Guide for Data Providers

2013-06-26

Preface

A **Globally Unique Identifier** (**GUID**) is a unique reference number used as an identifier. Complexities associated with specimens and associated, dynamic data in natural history collections have led to differences in opinions about how to create GUIDs and where they are required vs. recommended. This guide is intended to give data providers the information necessary to assign GUIDs that will allow iDigBio to ingest and share data. Suggestions for improvements to this guide are welcome, and can be inserted as comments on https://www.idigbio.org/content/guid-guide. A more general discussion on the use of identifiers in natural history collections is appropriate in iDigBio's community forums: https://www.idigbio.org/forums/identifiers-natural-history-collections.

GUIDs for iDigBio

GUIDs are needed by data *providers*, intermediate data *aggregators*, and data *consumers*. A data *provider* is a person who sends biodiversity collections data to iDigBio to be ingested on behalf of the institution that owns the data. A data *aggregator* is an entity like iDigBio, Morphbank, and VertNet, which ingests data from *providers* or other *aggregators* and serves it to *consumers*. A *consumer* is a person who extracts data from a portal such as that provided by iDigBio for use in research, education and outreach activities.

GUIDS are necessary to identify unequivocally specimen-based records so that uses of the data, and recommended annotations (such as corrected species identifications or addition of georeference data) from aggregators or consumers, can be supplied back to the *provider*. To enable this, a **digital record for the specimen**[1] and **digital records for related objects** (e.g., image or audio recording of the specimen) must have their own GUIDs.

1 A "specimen" may be defined differently among various types of taxonomic collections. It may refer, for example, to (a) an individual (a mammal skin), (b) a 'lot' such as a vial containing several

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

A **digital record for the specimen** includes information usually contained in a collection database: a taxonomic name, a catalog number and/or barcode, locality, date of collection, etc. For this information to be submitted to iDigBio, the record should include, *as a minimum,* a GUID and a taxonomic name (usually a scientific name, but it can be a genus, family, or other taxonomic name).

Note: iDigBio will accept data without GUIDs assigned by the provider. However, without GUIDs, one will not be able to update records (i.e., they can only be read) making it impossible to create and track history, generate usage statistics or relate records to other records or data. In addition, recommended annotations to the dataset and information on uses of data cannot be easily captured and reliably returned to the provider. Please see Appendix 1.

I. How to generate GUIDs

iDigBio recommends generating GUIDs as **URIs** (Universal Resource Identifiers), which are strings that begin with a scheme name (or protocol). Registered scheme names that may be used to develop URIs include http, doi, and urn.

iDigBio further recommends URI schemes that rely on **UUIDs** (Universally Unique Identifiers). UUIDs are generated using web services or standard software packages and the corresponding URIs are globally unique.

1. Examples of UUIDs are:

   A simple UUID: f47ac10b-58cc-4372-a567-0e02b2c3d479
   A UUID using URI syntax: urn:uuid:f47ac10b-58cc-4372-a567-0e02b2c3d479

2. Another example of an acceptable URI scheme is Archival Resource Key (ARK), which originates from the library, archive and museum community. The EZID project of the California Digital Library (http://www.cdlib.org/services/uc3/ezid/) provides services in support of ARK UUIDs. An ARK-generated UUID has the following form, where the embedded "87286" is the Name Assigning Authority for the biodiversity community.

specimens from the same locality and date, (c) all of the various parts of a plant on an herbarium sheet, or (4) a group of fossils in a single concretion.

ark:/ 87286/f47ac10b-58cc-4372-a567-0e02b2c3d479

3. Other examples of GUID-generating resources are given in Appendix 2.

Remember: Digital records for the specimen and each related digital object must have their own GUIDs in order for annotations and other usage data to be sent to the provider.

*A **digital record for the specimen** might have this URN-generated UUID:*
urn:uuid:f47ac10b-58cc-4372-a567-0e02b2c3d479
*a record for an **image** of that specimen would have a separately generated URN UUID:*
*urn:uuid:*6796b640-4c50-4714-aad6-818346008282
*and **a record for a video recording** of that specimen would have a third URN UUID.*

II. How to add GUIDs to a database for export to iDigBio

Step 1. GUIDs from providers should be entered in a field named 'idigbio:recordID' (Fig. 1)

These identifiers do not replace catalog numbers, bar codes, or other collection management strategies that are used to connect database records with physical specimens. It is not necessary for GUIDs to be on labels or otherwise attached to physical specimens.

Step 2. Use of the 'ac:relatedResourceID field (Fig. 1)

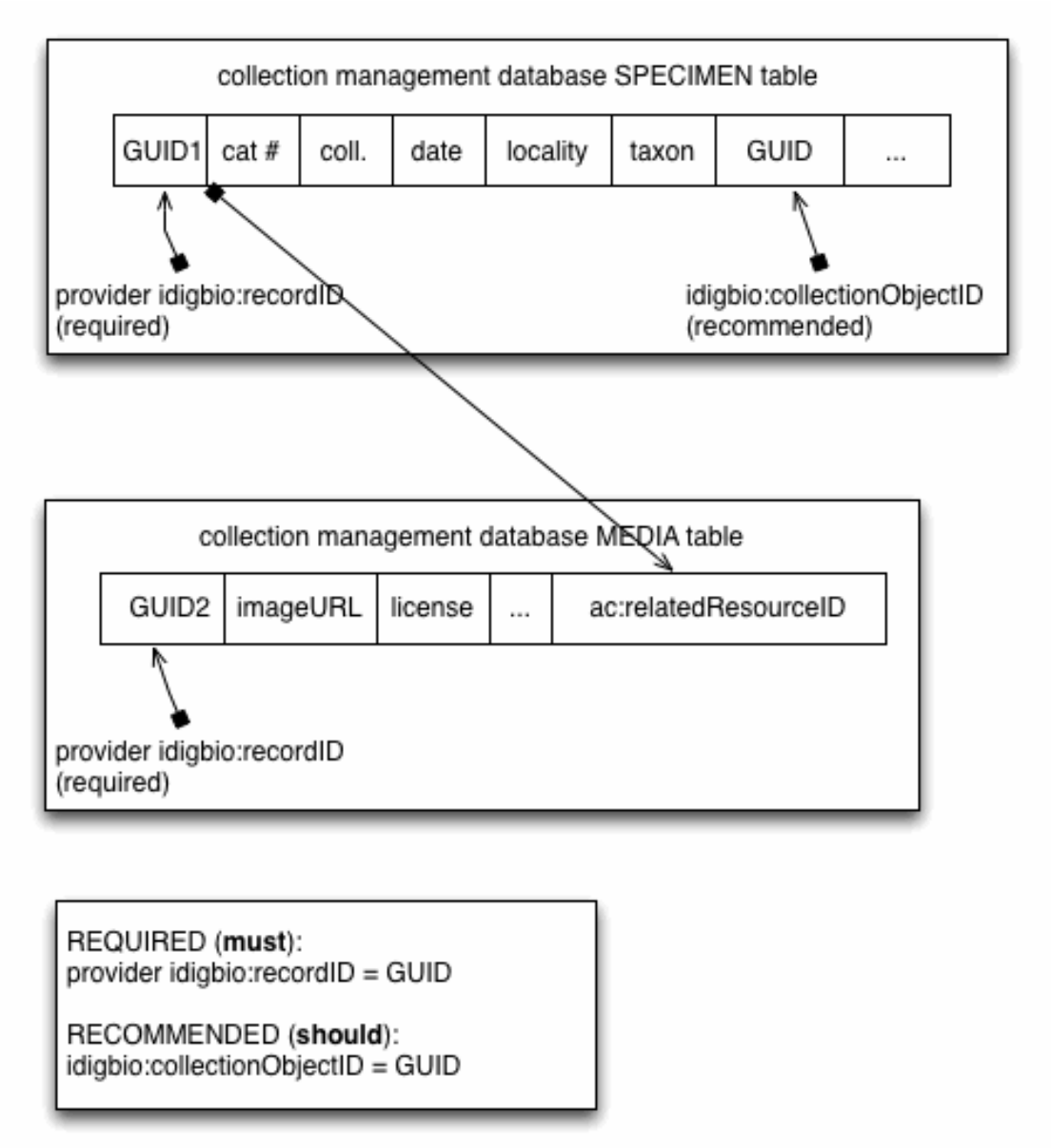
To clarify the relationship between a digital record for the specimen and another digital object record, e.g., for an image, put the GUID for the digital record for the specimen in a field named ac:relatedResourceID when the image record is submitted. (If an image or other related object record is submitted before the specimen record, put the GUID for the related digital object in the ac:relatedResourceID field when the digital record for the specimen is submitted.)

Step 3. Use of the 'idigbio:collectionObjectID field (Fig. 1)

To maintain a global reference to all records related to a physical specimen, a second GUID can be added to a field named 'idigbio:collectionObjectID. This could be used, for example, to

distinguish among several specimens on a slide or in a vial, or among fossils in a single concretion (in these instances, the GUID in the idigbio:recordID field refers to the slide, the vial, or the concretion, the encompassing cataloged collection object).

Figure 1. Relationships among GUIDs and names of fields for export to iDigBio



III. How to export record identifiers to iDigBio

a) If using the GBIF IPT tool, add the related resource extension and create the field mapping like this:

1. dwc:relatedResourceID = GUID entered in 'idigbio:recordID'
2. dwc:relationshipOfResource = "representedIn"

This additional mapping is necessary because there is no recordID field in the GBIF IPT tool.

b) If submitting records in a spreadsheet:

1. Create a field called "idigbio:recordID" and populate that field with the digital (specimen or media) record GUIDs from the collections database.
2. If a collectionObjectID is used, create another field called idigbio:collectionObjectID and populate it with a GUID as discussed above. (Note: For providers using Audubon Core fields for ingesting media records, this means adding the dcterms:identifier field and filling it with the primary identifier of the image record.)

In summary:

A GUID should be assigned to each digital record of a specimen, and a different GUID assigned to each related digital object submitted to iDigBio. These GUIDs are to be submitted in a field named 'idigbio:recordID' (Fig. 1).

It is recommended that providers place the GUID of the digital record for the specimen in a field named ac:relatedResourceID (Fig. 1) when a record for a related digital object (image) is submitted.

The GUID in the idigbio:collectionObjectID field can be used together with the 'idigbio:recordID' to maintain the global reference back to a particular physical specimen (e.g., in cases where it is necessary to relate different specimens to the same 'lot').

Appendix 1 – Why Use a GUID?

It is critical to be able to identify unequivocally a piece of information about a digital collection object, including who the *provider* is, regardless of the route it may have taken before arriving at the iDigBio portal; e.g., from one or many data aggregators or directly from the source (the *provider*). Ideally, every provider would add a globally unique identifier (in a field called recordID), to each digital collection object record. When this happens, different kinds of valuable services are available to the *provider* and to the *consumer*:

1) Write-back - This is the process where an annotation to a record by a *consumer* can be pushed back to the *provider* for presentation and possible update. Additionally, any other place where the specimen record appears in the global data universe, can be also be updated, as long as its identifier is persistent, i.e., no intervening aggregator has subverted, concealed, or reused for a different purpose the *provider's* GUID.

2) Impact tracking - this is the process where a *provider* can trace via the universe of globally linked data where their specimen records have been used; e.g., in research publications or environmental impact statements.

3) Validation - when the portal is able to run data validation reports against authority files, e.g., place names, taxonomy, collector names, it will be important to get the suggested corrections back to the source.

4) Data quality/research value/transparency - many *consumers* of the data are researchers who depend on the quality of the data, which includes proper accounting for statistical models. A GUID assigned at the source by the provider improves the quality of the data in general; e.g., alleviating issues caused by the number of duplicated collection object records.

- *For reasons given above, it is highly desirable that GUIDs be assigned by the data provider.*

- An internal iDigBio GUID is assigned to each record received for use by iDigBio, regardless of whether the provider has provided a GUID. These iDigBio GUIDs are visible too, and can be used by providers; however, that requires that they be returned to the

provider and entered into the original dataset by the provider at the institution, which is likely to be more difficult than adding GUIDs initially.

Consequences of Submitting Data to iDigBio without a GUID

In an effort to accommodate the broadest range of available data sources, iDigBio is relaxing its initial requirement of every record coming into the system with a GUID. iDigBio will make an effort to integrate records that have no identifiers into our system, but many of the features iDigBio has or intends to build may be disabled for these data. Specifically, annotations and systems which depend on annotation such as functionality (validation/quality assessments, impact reporting at the record level) will likely be wholly or partly disabled for datasets without identifiers.

For datasets that are unable to store or export even stable locally unique identifiers (CSVs with unstable row ordering, databases with duplicate catalog numbers and no primary keys, databases with fixed export formats that do not include identifiers), iDigBio may have to disable core system features such as versioning and relationships in order to incorporate them. Providers who find themselves in this category are advised to contact iDigBio for advice on modifying their existing system or migrating to a different cataloging system entirely, as systems in this category are likely to be increasingly isolated from the rest of the data universe.

Appendix 2 - Resources

dwc: Refers to 'Darwin Core,' a body of standards that "includes a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. The Darwin Core is primarily based on taxa, their occurrence in nature as documented by observations, specimens, and samples, and related information. Included are documents describing how these terms are managed, how the set of terms can be extended for new purposes, and how the terms can be used" (http://www.tdwg.org).

ac: "The Audubon Core metadata schema ("AC") is a representation-neutral metadata vocabulary for describing biodiversity-related multimedia resources and collections. Multimedia Resources are digital or physical artifacts which normally comprise more than text. These include pictures, artwork, drawings, photographs, sound, video, animations, presentation materials, interactive online media including identification tool packages involving text and other media. A multimedia collection is an assemblage of such objects whether curated or not and whether electronically accessible or not. For the purposes of

this schema we regard a collection of multimedia resources itself as a 'multimedia resource'"(http://www.tdwg.org).

UUID

- http://tools.ietf.org/pdf/rfc4122.pdf - canonical technical definition of UUID
- http://en.wikipedia.org/wiki/Globally_unique_identifier - the Wikipedia definition
- Online GUID generator: http://www.guidgenerator.com/online-guid-generator.aspx
- http://henbo.wordpress.com/2007/12/02/the-mystery-of-upper-case-and-lower-case-guid-values/
    - NOTE: 6.5.4 Software generating the hexadecimal representation of a UUID shall not use upper case letters. It is recommended that the hexadecimal representation used in all human-readable formats be restricted to lower-case letters. Software processing this representation is, however, required to accept both upper and lower case letters as specified in 6.5.2.
- MSSQL, PostgreSQL, and MySQL all provide native UUID datatypes.

GBIF

- Best place to start: http://www.gbif.org/communications/news-and-events/showsingle/article/a-beginners-guide-to-persistent-identifiers-published/
- http://rs.tdwg.org/dwc/terms/
- http://terms.gbif.org/wiki/Audubon_Core_Term_List_(1.0_normative) Audubon Core term list

GUID

- http://en.wikipedia.org/wiki/Globally_unique_identifier
- http://links.gbif.org/persistent_identifiers_guide_en_v1.pdf
- http://imsgbif.gbif.org/CMS_NEW/get_file.php?FILE=24d1fe88b849e4225f3117fac03d6c

LSID

- http://en.wikipedia.org/wiki/LSID

DOI

- http://www.doi.org
- http://www.handle.net

EZID

- http://www.cdlib.org/services/uc3/ezid/