GUID Guide for Data Providers

VERSION: 1 dated: 2012-12-14

Contents

Definition of Roles

When it comes to GUIDs (Globally Unique IDentifiers), one might well ask, 'what are they for?', and the answer depends on who you are: The data *provider*, the intermediate data *aggregator*, or the data *consumer*. The definitions of these roles are as follows:

1. The data *provider* is the person who brings collections data to iDigBio to be ingested on behalf of the institution who owns the data.

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

2. The intermediate data *aggregator* is the entity like iDigBio, Morphbank, VertNet that does not own the data but ingests it from *providers* and serves it to the *consumer*.
3. The *consumer* is the person who uses the iDigBio specimen data portal.

The aim of this entry level guide is to give the data *provider* enough explanation and context to understand GUIDs and how to implement them with data sets they provide to iDigBio. For the record, it is anticipated that four collection management tool providers, Symbiota, Specify, Arthropod Easy Capture and EMu, will all supply UUID-style *provider* GUIDs in their users' catalog records.

Why use a GUID?

The reason iDigBio is interested in GUIDs is because in a global data environment, it is critical to be able to identify unequivocally a piece of information (especially digital collection objects), 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 put a globally unique identifier on each digital collection object record, each media record, and so on. When this happens, different kinds of valuable services are available that are of interest to the *provider,* and also 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, in research, for instance.

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 by the source improves the quality of the data in general, e.g., alleviating issues caused by the number of duplicated collection object records.

It has been agreed by the iDigBio community that the identifier represents the digital record (database record) of the specimen not the specimen itself. Unlike the barcode that would be on the physical specimen, for instance, the GUID uniquely represents the digital record only.

UUID GUIDs Are Preferred

GUID (rhymes with 'squid') is a generic term for a globally unique identifier. There are several kinds of GUIDs, some with more distinguishing and desirable features than others. It is preferable for the GUID to be assigned by the *provider* so that there is never any question about its provenance. The 3 most important features of interest in using a GUID to iDigBio are:

1) persistency – ['referential consistency'] It is critical for each specimen record to keep its GUID for as long as the digital record exists. If for some reason the record needs to be split, new part records should be created by the data owner and given new GUIDs, and the original one retired (never reused for another purpose).

2) opacity - This is a controversial feature that some data owners favor for its utterly simplicity as 'just a string', no additional information can be inferred from looking at it. It may be combined with another protocol that is actionable.

3) actionable – [resolvable] This feature may or may not be desirable to the GUID creator. The actionable (or addressable) identifier has lexical meaning, and as such may

be presented to a web browser for Internet resolution services for more information. Using the 'URL' scheme as an identifier is deceptively actionable, i.e., the identifier looks like words in a string, a valid URL, but it may not have the actionable meaning it conveys. See http://www.w3.org/TR/webarch/#identification for more details.

There are several competent GUID protocols and management services that are available to the reader that are not covered in this document. Instead, only three will be discussed:

1) URI – this is one early solution proposed by iDigBio, it uses the W3C URI protocol, which might involve urn, file, or http, see

- https://www.idigbio.org/sites/default/files/iDigBio-GUID-Statement20MAR2012.pdf

2) the Darwin Core Triple – this solution is discussed briefly here only because it is the one more providers who have previously provided data to GBIF will be familiar with.

<institution code> + <collection code> + <catalog number>

The values of each of the parts of this identifier can vary depending on interpretation, and can lack global uniqueness. In the case where a specimen might change hands, the permanent institution code part of the identifier is no longer correct, and the syntax does not naturally allow for amendment to include the new owner (thereby possibly losing persistency in the bargain). Additionally, the 'catalog number' part of the identifier is ambiguous in many collections, for instance, it can be an accession number, a catalog number, a barcode. This triple has been used for a long time in the community, but it lacks some of the modern features that are required in a universe of data sources.

Examples:

'INHS' + 'Insect Collection' + '293937'

'FLAS' + 'vascular plants' + 'FLAS 100000'

'F' + Botany‘ + 'V0023234F'

3) Universally Unique IDentifier (UUID) URN Namespace (UUID)

These are 128 bit numbers in the style found here:

http://en.wikipedia.org/wiki/Universally_unique_identifier

Example:

020048f7-108a-437a-acad-23a73905dd28

These are the GUIDs that iDigBio prefers in that once created they persist and do not suffer from too much lexical friendliness.

While not immediately resolvable, UUIDs can be resolvable via a resolution service.

This UUID field should be included in the *provider* database as a 36 character field. Modern database applications provide native UUID datatypes that have significant advantages, especially when indexing those fields. The value is expressed in lowercase and although they are typically displayed in curly brackets ({}), it is not necessary to store those. There are various functions available in most programming languages for generating these UUIDs.

It is also an acceptable practice to combine a UUID with one of the URI/URN formats mentioned above.

Example: (URI with UUID)

http://www.example.org/specimen/020048f7-108a-437a-acad-23a73905dd28

GUID Resources

iDigBio

- https://www.idigbio.org/wiki/index.php/Globally_Unique_IDs_(GUID)
- https://www.idigbio.org/content/idigbio-guid-statement

GBIF

- Best place to start: http://www.gbif.org/communications/news-and-events/showsingle/article/a-beginners-guide-to-persistent-identifiers-published/
- http://links.gbif.org/persistent_identifiers_guide_en_v1.pdf
- http://imsgbif.gbif.org/CMS_NEW/get_file.php?FILE=24d1fe88b849e4225f3117fac03d6c

GUID

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

UUID

- http://tools.ietf.org/pdf/rfc4122.pdf
- http://en.wikipedia.org/wiki/Globally_unique_identifier
- 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 that have significant advantages, especially when indexing those fields for reference as a primary key. Microsoft programs typically utilize the Win32 API function CoCreateGuid and string manipulation functions.

LSID

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

DOI

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

EZID

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

Need Help?

Providers who need assistance with generating UUID-style GUIDs for their records should contact Alex Thompson (godfoder@acis.ufl.edu).