Stable identifiers for GBIF mediated data

Proposed action

1. Introduction

In a global network such as GBIF's, where data from numerous dispersed sources are integrated and made discoverable and accessible for re-use, it is essential that every data set and every record should have a stable identifier that is globally unique, persistent and resolvable.

2. Why do we need stable identifiers?

There are three main use cases for implementing resolvable, persistent identifiers for GBIF mediated data:

i. to support linking of biodiversity resources, e.g., linking a tissue sample or genetic sequence to a specimen
ii. to support citation of datasets, including "citation of citations" in which an identifier is issued for citing a collection of datasets (each with their own citable identifier) used in some particular study
iii. to provide information about a resource through a resolution service

2.1 Current situation

Currently, the GBIF registry issues identifiers for each dataset and record indexed. The former receive a UUID, e.g., **846d0710-f762-11e1-a439-00145eb45e9a** for the *Royal Saskatchewan Museum Collection*. Individual records within the collection receive a GBIF integer identifier based on the Darwin Core catalogue number provided, e.g., **88475251** for the lepidopteran *Boloria frigga* record with original catalogue number **425110**. GBIF can guarantee the stability and persistence of the dataset UUID. However, it cannot do so for all individual record identifiers as catalogue numbers in the original dataset sometimes become altered or re-assigned and so, as a consequence, when re-indexing, it is not possible to match the GBIF record identifier back to its original record. While the GBIF integer and UUID identifiers are resolvable (e.g., http://data.gbif.org/occurrences/88475251/; http://gbrds.gbif.org/browse/agent?uuid=846d0710-f762-11e1-a439-00145eb45e9a), these are not intended currently to be persistent URIs. Instead, persistent URIs for datasets based on the UUIDs will be resolvable through the new portal currently under test (e.g., http://uat.gbif.org/dataset/846d0710-f762-11e1-a439-00145eb45e9a). Persistent URIs for individual records in the form of http://gbif.org/occurrence/RecordNumber are technically trivial but require a commitment on the part of the data publisher to ensure the original record identifiers remain stable.

2.2 How can we improve this?

GBIF has already taken steps to ensure that all datasets in its network are provided with globally unique identifiers in the form of UUIDs which are also associated with stable HTTP URIs. The GBIF registry also allows other identifiers of a dataset, when known, to be mapped to the GBIF UUID. For GBIF to issue resolvable, globally unique and persistent identifiers for individual records within a dataset, the data publisher must guarantee the stability of their local record identifiers. Without this cooperation, implementing stable record level identifiers for the GBIF network becomes a very difficult task.

GBIF Secretariat • Universitetsparken 15 • DK-2100 Copenhagen Ø • Denmark
Tel: +45 35 32 14 70 • Fax: +45 35 32 14 80 • E-mail: gbif@gbif.org

HTTP URIs support two of the uses cases outlined above by offering a mechanism for both resolution and linking of biodiversity resources. Technically, the URI could also be used as an identifier for citation purposes. However, we propose to adopt the Digital Object Identifier (DOI)[1] system and, in particular, DataCite[2] with its support for "citable datasets" because it aligns the publishing of datasets, perhaps with an associated data paper[3], with traditional scholarly publications.

3. Digital Object Identifiers

DOIs are part of a system consisting of four main components[4]:

1. a naming/numbering syntax for DOIs
2. a resolution service for DOIs
3. a data model for the metadata/information registered about DOIs
4. a particular implementation (e.g., DataCite, CrossRef) with community infrastructure, regulations, etc., for issuing and registration of DOI names

The DOI is an opaque string with two parts, the prefix and suffix, separated by a "/". The prefix is allocated by a DOI registration agency (e.g., DataCite) and denotes a unique naming authority. The suffix, a unique string within the naming authority, is chosen by the registrant, and may be an already existing identifier and of any length. For example, the DOI for one particular dataset published by Canadensys is: doi:10.5886/txsd3at3. The suffix can have several parts and need not be so opaque, e,g., it could include institute, dataset name, etc.

Minting a DataCite DOI may be done through a web form interface or via an API (for automated machine processing). The requester is required to submit key information (metadata) that accurately identifies the data for citation and also the URL of a "landing page", i.e., the web page that the DOI resolves to. The DataCite metadata schema[5] defines 5 mandatory and 12 additional properties. The mandatory properties are DOI identifier, Creator, Title, Publisher, Publication Year. It is recommended that the DOI resolve to a rich landing page that provides much more information/metadata.

4. Proposed process for GBIF

In deploying DOIs on its network, GBIF will use the following principles:

1. promote use of DataCite DOIs as the preferred persistent identifiers for biodiversity datasets
2. mint DOIs when the data publisher does not have that capacity
3. never knowingly duplicate a DOI for a dataset that has already been issued with one
4. when duplicate DOIs exist, the preferred DOI to use will be clearly recognisable

GBIF will undertake the following tasks:

1. identify the most-appropriate technical options for GBIF to establish itself as an issuer of DOIs
2. document service requirements and best practices for data publishers to make use of DOIs and to adopt GBIF as their DOI issuing agency
3. undertake a test implementation of DOIs with the understanding that identifiers may change prior to a final implementation
4. generate and store a citation file for each data download, and issue a DOI for the citation file
5. assess the value of issuing DOIs for specimen records (in preference to HTTP-resolvable identifiers)
6. document requirements and assumptions for delivering consistent HTTP-resolvable globally-unique

[1] http://www.doi.org/
[2] http://datacite.org
[3] Chavan V, Penev L. 2011. The data paper: A mechanism to incentivise data publishing in biodiversity science. BMC Bioinformatics 12 (suppl. 15): S2
[4] http://www.doi.org/doi_handbook/1_Introduction.html
[5] http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

identifiers for all records in datasets

7. trial the use of consistent HTTP-resolvable globally-unique identifiers for each record in a data set (with the understanding that identifiers may change prior to any final implementation);
8. flag as unstable those datasets whose local record identifiers are not stable

5. Enabling citation of GBIF mediated data

The following illustrates the envisaged workflow for citation through the GBIF network where data download and re-use typically involve records derived from multiple datasets.

- all datasets are associated with a DOI
- on data download, a citation file is created which references the source datasets by their DOIs
- users of the downloaded data cite the data using the DOI for the citation file

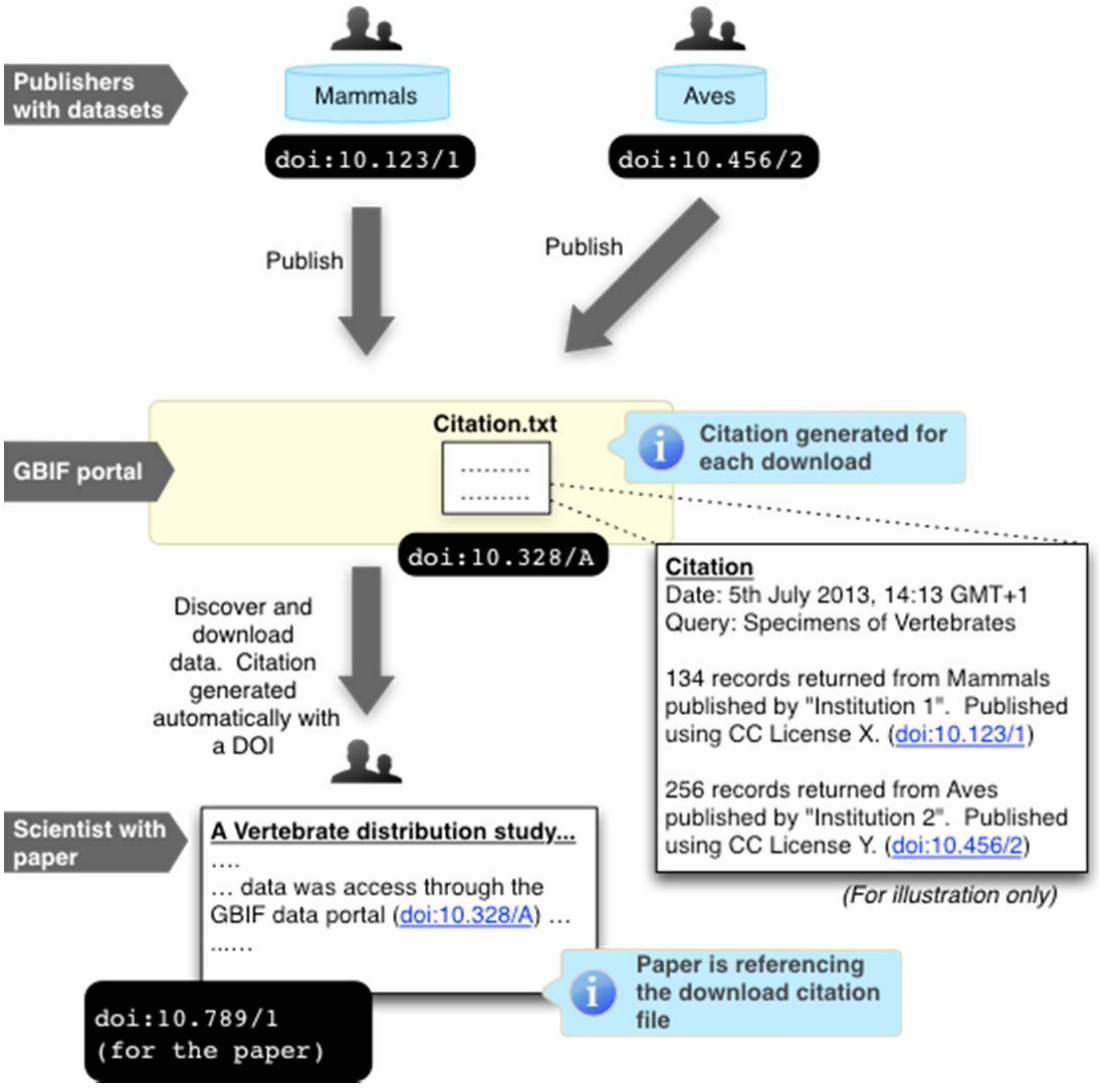

Figure 1. The envisaged workflow for issuing a citation for data downloaded from GBIF. Reference is made to three types of DOI: i) for the individual datasets published to the GBIF network, ii) the DOI for a scientific publication, and iii) a DOI for the citation file that contains the DOIs of multiple datasets downloaded from GBIF for the analysis reported in the research paper.

It should be noted in such an approach papers written using content accessed through GBIF will refer to the GBIF DOI for the citation file, and not the individual DOIs of the constituent datasets. There are pros and cons to this approach:

Cons:

- Currently, journals and DOI networks do not make it easy for people to view the citation graph, e.g., it is difficult to navigate beyond direct citations.

Pros:

- Data publishers can, through the DOI mechanism, see how many times their data has been accessed through GBIF
- GBIF could potentially index key journals and notify data publishers when papers are written using their data
- Users of GBIF will have a single clear mechanism to cite data publishers without having to cite potentially thousands of sources

Please send observations and responses, by 1 September, to:
identifiers@gbif.org

For any inquiries or clarifications about this briefing, contact:
Éamonn Ó Tuama
Senior Programme Officer for Interoperability
GBIF Secretariat
eotuama@gbif.org