Responses to the call for comments on the consultation document *Stable Identifiers for GBIF Mediated Data*

# Introduction

A draft document was first shared with Rod Page as he had expressed concern in the past about the lack of stable identifiers for GBIF records. He thought it looked good, suggested a couple of minor edits and was really pleased to see this work going ahead. The call for comments was issued on July 9th to the Nodes Interest Group mailing list with a request for input by September 1st.

# Respondents

Five responses were received.

1. Gregor Hagedorn
2. Francisco Pando (GBIF Spain)
3. Isabel Calabuig (DanBIF)
4. Rui Figueira (GBIF Portugal)
5. Anton Gunsch (BGBM)

# Responses

The text of the full responses are provided in the following sections (Response No. 1, Response No. 2, etc …). The main issues are outlined here.

1. All respondents were in favour of a policy of issuing DOIs for datasets
2. There is a need to be explicit about catalogue numbers [Gregor Hagedorn]:
   1. Will this identifier be stable if the data changes, but the catalogue number remains the same?
   2. When the catalogue number contains duplicates, because different subcollections of an organisations have duplicate catalogue numbers?
3. URIs for records [GH]
4. DOIs for versions of datasets [GH]
5. Identifiers not just for records but for specimens themselves [GH]
6. DOI for “citation of citations” needs to be generated just for the datasets that are actually used in a subsequent paper (and not on all data downloaded) [GBIF Spain; GBIF Portugal]
7. Cost of issuing DOIs [GBIF Spain]
8. How to handle records with duplicate identifiers when merging databases [DanBIF]
9. Recognising that the Darwin Core triple may not stable (e.g., because collections may merge) [DanBIF]
10. Tracking relationships of identifiers [DanBIF]
11. Need to follow the citation graph, i.e., if the DOI for the citation of citations alone is used, then only GBI gets credit; the DOIs within the GBIF citation must be indexed as well [GBIF Portugal]
12. Preference for http-URI based approach with identifiers issued at provider level (and not at the aggregator level) as exemplified by the http URIs for collection objects in the context of CETAF/ISTC and pro-iBiosphere in conjunction with BGBM, Paris-MNHN, Kew, RBGE Edinburgh, and MfN Berlin [BGBM].

# Response No.1: Gregor Hagedorn

Received: 9 July 2013

Comments on "Stable identifiers for GBIF mediated data"  
  
I find the paper in its present state somewhat confusing with respect to the following:  
  
1. The current situation with respect to record identifiers. The document says "receive a GBIF integer identifier based on the Darwin Core catalogue number  
provided". Here the "based on" is not explained. Will this identifier be stable if the data changes, but hte catalogue number remains the same? Is it stable where the catalogue number contains duplicates, because different subcollections of an organisations have duplicate catalogue numbers?  
  
2." 2.2 How can we improve this?" talks about dataset and record level, saying about record IDs: "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." (which I agree with). It continues with "HTTP URIs support two of the uses cases outlined above by offering a mechanism for both resolution and linking of biodiversity resources." -- to me it is not clear which are the two. Dataset and record IDs?  
  
3. It concludes, that instead of URIs, DOIs should be used: "However, we propose to adopt the Digital Object Identifier (DOI)1 system and, in particular, DataCite2 with its  
support for “citable datasets” because it aligns the publishing of datasets, perhaps with an associated data paper3, with traditional scholarly publications." I agree with this for the dataset, but after talking about dataset and records above, suddenly the record level is missing and URIs for this, which to me seem a very desirable solution, are no longer discussed.  
  
The record level is mentioned below as "to be assessed", but I believe it needs discussion.  
  
4. The most critical issue I consider missing is that each new version of a dataset requires a new DOI, because with DOIs we want to achieve citability. We could disregard citability in a new model of DOIs, but I believe not in an established one like DataCite.  
  
5. First and foremost, I believe it is mandatory that GBIF starts talking about Identifiers to make it possible no longer to talk only about records, but about the specimens itself. This is the main concern of the recent efforts to provide stable identifiers in the context of the semantic web.  
  
Whether GBIF issues record identifiers or not is a separte concern, but I believe what is needed is a stable identifier for the specimen, one that can be used inside and outside of GBIF, to start reasoning and seeing connections (e.g. to NCBI or BOL databases).  
  
Gregor

# Response No.2: GBIF Spain

Received: 21 August 2013

Dear identifiers,

We have discussed your paper at the GBIF Spanish node. We agree with the overall idea and proposed way forward.

We also note that it is the normal  fact that between the data  a researcher  downloads from GBIF, and the data finally considered for analyses and later publication, a lot of records are usually disregarded, so –as we see it-- the usefulness of the proposed "citation DOIs " is limited.

To overcome this limitation, we suggest an additional step in the envisaged workflow., i.e.,   to publish the actual "analysis datasets" in a DOI resolvable manner. Besides,  to facilitate tracking , an additional column may be added to the data files downloaded from GBIF  to include for each record its citation DOI (in addition to dataset DOIs, of course).

An additional concern is about the costs involving in implementing DOIs. We are sure you have made some calculations. How much would  the costs of implementing DOIs be for GBIF at the current and foreseen scale of operations?

With all best wishes,

Paco

Francisco Pando

Responsable

GBIF.ES, Unidad de Coordinación

# Response No.3: DanBIF

Received: 29 August 2013

Dear GBIF-secretariat,

Thank you for this draft, the Danish participant node supports the proposed process for GBIF on stable identifiers, it is a very important topic to plan and implement such a feature.

In the draft, it might be useful to specify further what is meant by ”the data publisher must guarantee the stability of their local record identifiers.”

Situations we could think of, with the current phrasing of the draft, where we would come in doubt on how to handle data

1) It is inevitable that records are deleted and hence a specific record identifier ceases to exist in a given dataset.

– e.g. imagine a huge dataset (atlas of Danish fungi) where we know there exists “doublets” of the same collecting instance because the atlas consists of the merging of at least three originally separate databases (a given occurrence originally appeared both in a “red-list-data-set”, a “curators personal colleting event dataset”, and an official country-wide “atlas survey dataset”).

These record “synonymies” will not be resolved in any foreseeable future, but we are eager to publish the dataset. At some point, hopefully, two records will be weeded out and a third will remain. How do we handle that?

2) The Darwin Core triplet identifier of a given record changes, e.g. if the owning institution of a dataset changes

– e.g. in the case of the Natural History Museum of Denmark, several datasets, previously “owned” by individual departments get a new common institution code. What should we do?

Some additional thoughts on the matter…: So we probably need some kind of system to keep track of record synonymy because e.g. changes in triplet codes and deletion of records are inevitable. If the new GBIF-system cannot handle deletion of a record, would we then need a system that links/maps an old identifier to a new one? But that of course poses problems as well, because mapping of an old to a new identifier probably must happen on the provider side.

Best regards,

The DanBIF secretariat,

/Isabel

# Response No 4: GBIF Portugal

Received: 30 August 2013

Stables identifiers for GBIF mediated data

Comments on the proposed action

**General comments**

The adoption of stable identifiers for datasets and records is essential to improve the quality of use of GBIF mediated data by users. Its implementation will expand the potential uses of biodiversity data to other domains, as ecology or phylogenetics. Presently, the data user is the only agent that can assert that no duplicate occurrence data is being used in his analysis. This will continue to be the case, but stable identifiers will enable him with more tools and control for this quality assessment.

Another important improvement in the use of stable identifiers is that it could boost the interest in the publication of context-based data papers, by allowing that data from different datasets or subsets are conveniently described under a coherent taxonomic, biogeographic or ecological context.

While the proposed action, and the workflow therein presented, addresses a critical issue that needs to be resolved, it has results in some consequences that need to be addressed.

*Consequences of the creation of the citation.txt file*

A new entity is created by this workflow, the citation file “citation.txt”. Each of these files needs, *a priori*, to be stored and become persistent. This file will be generated in every data download, whenever the download was made for exploratory analysis by the user, and could be discarded, or used in in a final study. This means that probably only a small fraction of the citation.txt files created would really need to be kept. The onus for the future for this storage obligation and how GBIFS will deal with it needs to be addressed.

Another aspect is is that, many times, the user only uses some records of the downloaded dataset, and not the whole file. It means that the citation file will not be representative of the data records that really need to by cited in the paper.

*Disincentive to data paper publication*

Concerning data papers, there is no obvious connection between the proposed action and the promotion of this form of publication. GBIF promotes the effort of publication of data papers by providers to i) promote complete meta-description of datasets; ii) facilitates attribution to providers by citation in papers using data. The proposed action makes it very unlikely that data papers get citations by users, because they will have a convenient citation file to cite. Providers can trace uses of data by crossing DOI of its datasets with the citation file, but will not receive direct citations in the final papers or documents published by users. This can result in a lack of incentive to providers to publish data papers, and to journal editors, to accept them.

Rui Figueira

Node Manager, Portugal

# Response No.5: BGBM

Received: 2 Sept 2013-09-16

Dear GBIF,

This is to let you know that the BGBM BDI group discussed the "Stable identifiers for GBIF mediated data" proposal report and concluded that deploying DOIs as a stable identifier system at data sets level seems to be the right step forward.

At individual record level we would however prefer an http-URI based approach with identifiers issued at provider level (and not at the aggregator level). We have started the implementation of http URIs for collection objects in the context of CETAF/ISTC and pro-iBiosphere together with BGBM, Paris-MNHN, Kew, RBGE Edinburgh, and MfN Berlin (see blog post at <http://stories.rbge.org.uk/archives/3846>) and we are quite optimistic that other institutions will follow. The system is cheap, easy to implement, does not prescribe a particular syntax for the structure of identifiers, makes a clear distinction between objects and data about objects, and it is linked open data compliant.

In early October, we will have a next workshop where we will look at and review the different implementations at CETAF partner institutions. We will also discuss the application of the system on objects beyond collections/observations. Markus Döring and Éamonn Ó Tuama did already get an invitation for the workshop.

Many thanks and best wishes,

Anton

--

Anton Güntsch

Freie Universität Berlin - Botanic Garden and Botanical Museum Berlin-Dahlem