Permanent IDentifiers on OpenBioMaps (an open conversation and research data management system)
Miklós Bán

THOR bootcamp Budapest – 28/09/2017
Presentation: http://openaccess.mtak.hu/dokumentumok/bootcamp_prez/BanM_THOR.pdf

Good morning/afternoon ladies and gentlemen (everybody), I'm very happy to be here today.
I'd like to begin my talk with highlighting the roles of open data and citizen science in Nature Conservation especially related to biodiversity loss.

[slide 1] Biodiversity loss is a global problem, Perhaps you agree with me, that it is the **biggest challenge** in human history ever - to be solved. This problem is bigger than anybody could solve alone, bigger than any country, any government could solve alone, bigger than science or industry could solve alone. We need collaboration on a massive scale to solve it and we should find the best way! It is a global problem, but trying to find global solutions for it is the wrong approach as it has no global origin but many local ones. **We need to solve** many local problems while keeping the global goal continuously in our minds. We must collaborate and communicate on different levels. So we should find very effective ways of communications and we need to find new and very effective ways to involve people - (if it is possible) everybody who can contribute in any way, to help solve local problems. It is an area where open research is becoming an essential (very important) part of the solution due to its technical and human infrastructure.

Even though it is obvious to many people that their contributions would be important, yet they do not contribute to conservation efforts (this is typical in Hungary). Meanwhile, the same people participate in several online communities, they build and improve collaborations for fun or for reputation. It is a key information: People have the capacity to build new online societies / communities, they can cooperate very effectively if they want to do something together. So, if we need their help in conservation, we have to try to involve them by tapping into their different motivations instead of convincing them to change their motivations or habits. To facilitate this, we should study how successful online communities operate, how they solve technical and human infrastructure problems, e.g. how they deal with permanency persistency in a rapidly changing world.

These are the thoughts which motivated us (some biologist researcher and conservationist) here in Hungary to start thinking about developing a good online data tool for scientists and conservationists. We wanted to create a swiss-army knife in data management and give it freely to people so they can create a possible interface for communication and data sharing. To reach this aim we have considered several known standards and data management habits and we are applying many of those. We have been developing an open tool which is compatible with several other open tools to help researchers and conservationists in their everyday work in data management.

[slide 2] Biologist have a long term relationship with permanent identifiers. The **scientific nomenclature** (Binomial nomenclature) is around 400 years old. The logic of this to give unambiguous names for species. Even though it is an old system, it still has some problems. The species are not permanent and our knowledge about them is also changing by time. Moreover, there is a global problem with the national names of species which used in several data sources.

Every large biological database has problems with the species names. Some of them are rigorous but it is an impossibly hard work to keep up-to-date them, some others are not strict but some algorithms needed to find the typos in the names and follow somehow the trivial mistakes, and almost none of them are able to follow the development of taxonomy (which realized in this level as name changes). To help solving this problem several online databases were established

specifically to collect and index species names and give unique and permanent identifiers to them. E.g GNI (Global Name Index), CoL (Catalogue of Life)... These can help a lot, but cannot solve every vagueness around the species names.

Data identification, which can also be complicated, should be considered when dealing with biodiversity data. In conservation, data point usually means an observation of species occurrence, which piece of information has at least four attributes: date-time, number of individuals, species name, and geographical location. In many cases, several other attributes are attached. Are these unique data points? Maybe, hopefully, mostly... Different observers can observe more or less the same event/creature and can create more or less same data about them. While many researchers work with firsthand data, it is more and more common to reuse data received from public data sources. In the latter case it is very important to give persistent identifiers to the data. Maybe the best is to give a DOI with some important metadata attributes. Probably one could be a field which contains those DOI-s which pointing to the same data - but was created independently! There is a huge global database which collects and shares these kinds of simple observation data. The GBIF, the world largest biodiversity database. In the GBIF there are ~850 million points of data. Theoretically this database contains individual data points which can be located with Permanent Identifiers.

The most straightforward area is **Identifying the data sources**. Thanks to the global effort around DOI, many of the scientific articles easily identifiable but not all. E.g. the old journals and books or private databases. Recently the scientific journals developed their own data repositories. Unfortunately (in many cases) the data coming from these have no unique identifiers but these are citable through using the published articles' identifiers or the parent database identifiers. An other way to put in or read data from a general data repository where the queried data has an identifier could be using a DOI.

Researcher and observer identifiers. In many databases and data sources the data collector (typically field biologist), the observer, (határozó) identifier and owner are listed by using their name and institutes. In most cases, these values can help to unambiguously identify these people but not always and it is difficult to handle these automatically. One possible solution would be a global researcher ID e.g. the ORCID.

In summary - conservation and biodiversity (including animal behaviour sciences) are very PID intensive fields. We generate lots of information which need permanent identifiers and we use lots of information where we have to use permanent identifiers.

[slide 3]

After this overview I would like to show how we are using Permanent IDentifiers in OpenBioMaps.

The OpenBioMaps as I mentioned previously is an open tool. This tool is a database management framework. It has few running nodes (server nodes) in some European countries. Most of these OpenBioMaps servers contain few databases for a special community such as National Parks and there are some which are dedicated for hosting small projects (openbiomaps.org).

The OpenBioMaps has some general feature: web interface to upload, query, show and share data. API interface. Permanent identifiers! The OpenBioMaps has some PID related services.
[slide 4] A list of our PID related services which will be paraphrased later:
PIDs for databases.
PIDs for datasets.
Using ORCID, DOI, DOI APIs ...

One of our first missions was to give **Permanent Identifiers for our databases**. Because the databases can move from one server to an other, Sometimes people buy domains which expire. So the database web addresses are not persistent. The databases itself also not immutable: in the most simple case more and more data coming into and their structure is also changing, developing.
The OpenBioMaps Consortium have a contract with this library to give DataCite DOIs for those OpenBioMaps databases which ask DOI and which can fill a DOI request FORM. This form includes some fields such as who is the founder, owner, what is the name of the database, description of it, web url, who are the managers and so on. This form produces automatically an online DOI metadata page, including all these information and the autogenerated future DOI. If you visit this example page [slide 5] (explanation) you will see the following information [slide 6-8] which will be included in the DOI metadata and we just send this URL to our DataCITE DOI manager to register a DOI for this database.
These DOIs help to cite the whole database, but sometimes we need more specific citations. A citation for a specified dataset within a database.

[slide 9] Users can **save queries and assign a (human readable) label and a PID for them**. These PIDs are located and accessible only on a specific OpenBioMaps server related to a specific database - so we do not try to reinvent the DOI. BUT, these PIDs are unique web links and people possessing these links can get the results of the original query independently the current state of the affected data in the database. However these PIDs only have local authority, these are unique web links - so these are digital objects. Therefore we can assign(?) DOI to them. Although it is one of our very first services, it has never been used until now. We just put together the first example this OBM service option at the last few days. I will explain this process shortly.
The Hungarian National Park Directorates operate independently from each other. There are 10 National Parks with 10 different database structures and database solutions. For example, in one of the National Parks there is a project to put together a book about the dragonflies of Hungary, therefore they asked for data from the other National Parks. The Duna-Ipoly National Park uses the OpenBioMaps. The ecology curator (an ecologist?) at this NP created a big SQL query [slide 10] for assorting the dragonflies data among the million animal and plant observation data. I put this query as a stored query in OpenBioMaps. I performed this stored query and I saved the result with a label attached to it. When I saved it, I got a LOCAL persistent identifier URL for referring to this query result. This web link has a unique string part [slide 11].We assign a DOI to this URL as a Digital Object to share data with the other National Park. In the book there will be two DOIs in the list of references, one referring to the database and the other to the stored results. SHOW WEB PAGE

We are also working on some ORCID integration. Earlier we used the ORCID API in the LogIn process but it was meaningless, nobody used it. Now we are working on an OBM module which helps to connect OBM people with ORCID and assigns data with the corresponding ORCID where the observers and owners of the data are available and relevant...

I mentioned the species names as a problematic field, previously. This is where we using our Local automatically generated species databases which is combined with automated typo handling and we are not connecting it to the global species name indexes and we try to avoid persistency here….

The last DOI related example in OpenBioMaps is a DOI usage.

There is an OBM project which is a closed scientific research project, where there are several (maybe 50 or more) researchers around the World who collect biometric data of animals from the literature to analyze them according to some questions. These data have some common attributes:
Species name,
Number of Individuals - sample size,

Date,
Location,
Reference,
We created a two dimensional database structure for handling ~200 variables with these 5 meta data variables. One of these, namely the reference is interesting here. Previously the researchers in this project put full text citations into the databases and later they processed it. It is a long and boring work. To make this easier, I wrote a small module to resolve DOI-s, that way a Citation List can be generated when needed. While putting in new data, researchers only have to copy the doi-s of the references into the input field instead of copy-paste full text references, although the latter is still possible using the same field!!!
Module usage: It has three possible ways:

1) Using PostgreSQL's notify service ability to resolve DOI-s as a background process soon after the data were deposited into the database. With the notify service we can fork processes into the background which can be important for actions with uncertain running time - which depend on unknown factors. So, users will see the full text citations after a while in the database in bibtex format. Disadvantage: complicated and not possible to solve the errors automatically.
2) Resolve DOI only when users need the citation list. The full text citations will appear in the database when users work with it for the first time. It can be achieved by using asynchronous Web calls. Disadvantage: slow.
3) Manually doing DOI resolution: It means that users create the citation list from DOIS and send the DOI list to the resolver, and the full text citation will be not stored in the database.

I don't know which solution will be realized.

Final thoughts

We need stable points in the Era of data deluge. There are two pivotal thoughts which should be considered: Far more open data are generated than ever will be used. Much more open data would be needed to create a clear and reliable picture about the current state of biodiversity and its problems. Why does this gap exist? Data quality problems? Reliability? Technical or human infrastructure problems? Maybe all together. But one thing is clear: Transparency is an essential ingredient of reliable biodiversity research and there is no transparency without persistent identifiers. Thank you.