Introduction to DataCite

Erika Bilicsi

THOR bootcamp Budapest – 28/09/2017

Presentation: https://prezi.com/0po_qdh11w_q/introduction-to-datacite/

I'm sure that all of you have seen this situation where you click on a link from a website and then it doesn't work, because the site was not available. For example DOIs are solving that problem, because the DOI name is unique, stable, persistent, and doesn't change with time.

Let's see what we need to operate this system: of course technical infrastructure and services, and we have to make them useful and useable for the researchers. We need statistcs and it is very important that we share the best practices.

Our colleagues realized these needs in 2009 or a bit earlier I think and established DataCite

DataCite is an international non-profit organisation which aims to improve data citation in three main ways:

- support data archiving that will allow results to be verified and reuse for future research
- support easier access to research data on the Internet
- facilitate acceptance of research data as legitimate, citable contributions

Let's see a little history! DataCite was created in London on 1 December 2009 as a German non-profit organisation by other organisations from 6 countries: the German National Library of Science and Technology (TIB) which was the first DOI registration agency for research data in the world since 2005; the British Library; the Technical Information Center of Denmark (DTIC); the TU Delft Library from the Netherlands; the National Research Council's Canada Institute for Scientific and Technical Information (NRC-CISTI); the California Digital Library (University of California Curation Center); and the Purdue University (USA).

Today DataCite has 47 members and 5 staff members from all over the world. According to June data it means 1249 datacenters and ca. 8 million DOIs. These datacenters create a very active community which meet online in every month in courses of Open Hours, staff of DataCites often organizes webinars and we have a mailing list too on which we can talk about everyday issues.

As I was saying DataCite is a German non-profit organisation but it has an international board. The board altogether consists of 9 members.

And the staff members who help us everyday to solve our daily businesses. They work in the German National Library of Science and Technology.

DataCite is a community driven organisation, it means that it operates under the control of 3 Steering Groups: Sustainability and Business; Services and Technology; and Community

and Engagement. The Steering Groups provide a venue for open participation by interested community members.

There are smaller, more specified working groups which are involved in the work. For example:

- the Metadata Working Group which determines and maintains DataCite's metadata schema and coordinates with community standards, such as ORCID
- and the Re3data Working Group which supports and develops the global registry of research data repositories, covering different academic disciplines.

DataCite is a DOI registration agency, so the main service is DOI registration.

Now all in all 10 registration agencies exist, which serve for publishers in several special field all over the world. These agencies have no difficulties to work with publishers from different disciplines or if they use special characters for example Cyrillic or Chinese letters.

The members register DOIs for datasets, gray literature and other non-textual materials.

Some DataCite's members register DOIs for journal articles, books and book chapters too, as we did, but the staff disapprove this practice - and we know that this isn't a good solution, because DataCite's metadata schema isn't suitable to make bibliographic records of these types of publications. The Library of HAS joined to DataCite to register DOI in 2014 for PhD dissertations but a little bit later we started to register DOIs for the articles of small journals. We are members of another regisrtation agency, CrossRef - which provide services especially for journals. The Hungarian journals have two main problems with the services of CrossRef: our clients have difficulties paying for the registration fee and to fullfill the cross linking obligations. The Hungarian Academy of Sciences supports our library's membership of DataCite so we can registrate DOIs for our clients for free. And the reference linking obligation of CrossRef is too much work for our clients. It means they should hyperlink to Crossref DOIs when they create the article's citation lists. This makes it possible for readers to jump with help of DOI link from the reference list to the cited work's fulltext and it enhances the scholarly communication and citations between articles.

But let's go back to DataCite: DataCite's definition of ''dataset'' is everything that can motivate new research. Therefore in this context ''dataset'' can be research data, images, videos, softwares and so on.

How we do this, how can we create a DOI? A DOI consists of 2 parts: the prefix and the suffix. The datacenter gets the prefix from DataCite but the suffix is freely manageable.

The core element of the service is the DataCite Metadata Store (MDS) which archives the metadata of all registered objects in a database.

To register an object, the metadata must be uploaded in XML format via browser interface or API to the MDS using the DataCite Metadata Schema.

When we register a DOI in MDS, we upload the DOI name and the URL where the object is available and the XML which contains the medatada fitting with the schema. as I was saying the DOI name is unique, stable, persistent, and doesn't change with time. The datacenter

has to manage the URLs in MDS, if they have been changed, they have to modify in MDS and the DOI will work again. But the DOI name isn't modifiable.

A little bit about this schema: it contains 6 mandatory fields, 6 recommended and 7 optional fields. Let me call attention to the rights or alternate identifiers fields in which we can add special information to facilitates data exchange and reuse.

In addition to the MDS, DataCite provides a search tool, the DataCite Search which contains metadata for DOI names that were registered through DataCite. We can search for researcher, for work and filter by publication year or by resource types, for data centers or for members.

Metadata are made available through DataCite's Open Archives Initiative Protocol for Metadata Harvesting service which provides an API for metadata harvesting which facilitates data-exchange between several databases - like MTMT for example. Furthermore, DataCite offers a detailed statistic portal too, which is very useful for us and for our clients when we need information about DOI retrieving.

But a DOI not worth anything if we don't use it! How can we facilitate this:

We can increase utility of DOIs if we link as much information as possible. For example we use ORCID to identify researchers as we have heard. In co-operation with CrossRef DataCite's team is working on linking institutions and funding information too.

Due to these cooperations there are a lot of efficient services in DataCite Search for researchers:

- like citation formatter, which provides the citation of DOI names in various formatting styles;
- or the ORCID button with which we can add works to our ORCID record manually but we can use the ORCID Auto-Update service too.

Why are these useful? Because these services and tools help to improve the scholarly infrastructure around data and other non-textual information.

The first step for us, here in Hungary to publish research data and archive them in a repository. We have 71 clients who register DOIs with our help but just 6 out of them mint DOIs to research data, and at this moment just 1 out of them archives data into a repository. It is a huge challenge for us here in the library too, to establish a data repository. We are working on it and I hope it will realize in this year. But we have a lot of questions: we don't know which type of records we need, how can we describe them, which metadata fields are necessary and will be the obligatory or which we need as minimum requirements, how can we guarantee the long term preservation and which file formats should we collect etc.

DataCite can help us realize this goal too. It operates the re3data.org - the Registry of Research Data Repositories service - which collects and makes more visible research data repositories from all over the world. This can help researchers to identify a suitable repository for their data, but it helps us to recognize best practices. There is a minimum requirements of

inclusion in this database, which is a guide for repository providers to evolve the suitable operation for research data archiving.

Last but not least it is very useful for us to archive research data in repository, because the popular repository softwares are integrable with DataCite services and we don't need to mint DOIs manually., Examples of these softwares are Eprints we use or Dspace which is very popular in Hungary - but OJS too, the Open Journal Systems which support to publish online, scientific, open access journals -

I think ensuring long-term accessibility will encourage researchers to link articles and underlying data. Don't forget that more and more founders recommend archiving and cite data in research articles. In one hand this confirms the results of the research, and in the other hand this makes possible to reuse the data.

DataCite community has a lot of challenges, I mention just two of which we are very interested in:

- DataCite, California Digital Library and DataONE have a project, the Make Data Count project. One of the main goal of this, is to develop standards for measuring and reporting of data usage statistics. This project started in May of 2017, we are interested in the results.
- we are expecting the new metadata schema which will contain relation types fieds to follow data versioning for the end of this year.