CASE STUDY

A Digital Curation Centre Case Study
April 2014

Assigning Digital Object Identifiers to Research Data at the University of Bristol

Section of the How to guide that this supports
Data Repositories and Data Catalogues

Stephen Gray, (University of Bristol) and **Monica Duke,** (DCC)

Introduction

The University of Bristol runs a dedicated research data repository as part of their Research Data Service. They are using the DataCite service at the British Library to assign digital object identifiers (DOIs) to research datasets in order to provide unique and perpetual identifiers for data, to allow easy citation and discoverability. The repository hosts data that underpins research outputs and provides a home for data with immediate and straightforward access. The Bristol Research Data Service provides guidance on how to use the identifiers to cite data and is developing appropriate policies to monitor usage.

Background context

The data.bris repository is run by the Research Data Service at the University of Bristol. The current service is a Library-led collaboration working closely with IT Services, Research and Enterprise Development and other Professional Services across the University.

Assigning identifiers to research data is part of the wider research data management strategy at the University of Bristol. Identifiers for digital objects serve several purposes, including helping to identify the object uniquely so that a reference to the object can be unambiguous. Some digital identifiers also provide mechanisms to locate the object to facilitate access. The identifier can be used in citations.

DataCite is an international organisation that specialises in services for assigning DOIs to research data and the British Library (BL) is the representative of DataCite in the UK. The BL team works with repositories and archives to assign DOIs to datasets to help make the data discoverable, accessible and citable.

The University of Bristol established its data.bris repository to curate datasets created by their researchers and through working together with DataCite, has put in place a system which automatically obtains a DOI for any dataset published by the research data repository.

What has been developed ? How can it be used?

As part of the data deposit process, the data.bris repository obtains a DOI for the deposited dataset from the DataCite service. Details of the workflow at Bristol are described in the next section. Each dataset is made available with a set of 'essential' metadata (DataCite Mandatory Properties). Currently, depositors are also encouraged to include rich, subject-specific metadata within the deposit itself, for example, as spreadsheets, user guides or simple text files.

The DataCite service is open to research institutions that wish to assign identifiers to the data they manage, but there are some set criteria that need to be met by organisations wishing to use it:

- the organisation must have the authority to assign DOIs to data
- a landing page, mandatory metadata and a URL that links to the data need to be provided
- mandatory and additional metadata must be made freely available for discovery
- a clear and public indication to make the data available over the long-term should be stated.

The aim of these obligations is to make sure that the data to which DOIs are assigned are trustworthy and persistent. There is a cost to using the DataCite service, based on an annual subscription, and a short trial period is used before a commitment to the service is made.

Practical example(s)

Implementing the DOI assignment process at the University of Bristol involved working at both a strategic level to acquire institutional policy and support, and at a practical level to establish process and workflow.

In joining the DataCite scheme, the University of Bristol was required to make a long-term commitment to support published datasets, therefore approval at the highest level was required. The decision was put before and approved by the Data Service steering group, chaired by Professor Guy Orpen (Pro Vice-Chancellor, Research). The University of Bristol team communicated regularly with the helpful BL staff via email, telephone, and in person at research data management events held at the BL itself. The cost of the University's DataCite licence is currently funded by the University until July 2015.

The platform used by Bristol for searching and previewing datasets is CKAN (from the Open Knowledge Foundation) which provides both human and machine (i.e. API) access to deposits. The following workflow runs from research project start to assignation of a DataCite DOI:

- Each new research activity is allocated a unique project identifier via the University's Finance System. The information collected at this point includes the names of staff and faculties involved.

- Early in the project a 'Data Steward' (typically the PI) is nominated, who takes responsibility for the project's data. The Data Steward registers online, agrees to the Data Storage Policy and is allocated 5TB of free storage space within the University's Research Data Storage Facility (RDSF) which also contains an empty, pre-made folder called 'data-bris'.

- Allocated RDSF storage space is then used for storing 'live' data throughout the project. At this stage access is limited to the research team.

- When the data is ready to be published (usually towards the end of the project) it is copied by a member of the research team to the 'data-bris' folder.

- The Data Steward then logs into an online depositing system, agrees to a Depositing License and completes a Deposit Form for each dataset. Any data copied into the 'data-bris' folder can be associated with one or more completed Deposit Forms. Much of the information required at this stage is automatically harvested from the the University's Finance System and also Pure, the University's Research Information System (RIS).

- When this process is complete, Research Data Service staff are made aware and validate each deposit before requesting a DOI from DataCite. Once a DOI is assigned, the dataset is published and the Data Steward notified.

All DOIs issued to Bristol are acquired via the depositing process outlined here. The DOI minting service provided by DataCite is only used in connection with the publication of University of Bristol data and not any other type of material or research output. The Research Data Storage Facility (RDSF) and Research Information System (RIS) provide some of the metadata describing published datasets. Specifically, the RDSF provides a starting point for the creators of the data, and (of course) the data itself. The RIS can be used to augment the list of creators and contributors, as well as related publications.

Policy

Several different policies must be adhered to if a dataset is to be assigned a DOI and ultimately published. The University's Data Storage Policy applies up to the point of publication and includes stipulations on data protection, encryption and sensitive data. At the point of publication the Repository's Depositing Licence applies and agreement is mandatory. The License is in place to ensure that deposited data is suitable for open publication as the repository does not currently offer controlled access. While the Deposit Licence does not permit publication of data where publication would be either illegal, a breach of contract or against ethical guidelines, no other 'rules' exist as to what type of data can be published. This choice is instead left entirely up to the Data Steward. For example, the deposit of raw data, creative works, records relating to physical objects and single file datasets are all accepted.

Issues addressed

Granularity

Granularity is dealt with by the depositor - deposits of a single file or deposits consisting of many thousands of files are acceptable, however, depositors are encouraged, via guidance documents, to carefully consider the logical structure of data before it is deposited.

An important factor for depositors to remember is that a DOI relates to a completed Deposit Form. This is a one-to-one relationship. However, each Deposit Form may be associated with none or many digital files.

Depositors are particularly encouraged to answer two questions:

1. What is the smallest unit of data which you are likely to cite? (e.g. for a historian this may be an individually transcribed manuscript, while a social scientist may want to cite the raw data from an entire longitudinal study).

2. What do you expect the needs of secondary data users to be and how can you best structure the data to support those needs? The file/folder structure at the time of deposit is preserved and made visible to secondary data users and is therefore a powerful organisational tool.

Versioning

The metadata schema supports the notion of one dataset 'superseding' another - the original dataset is not 'replaced'. Instead, users are made aware that a deposit is no longer current. A depositing Data Steward who wishes to 'version' a dataset does so by asking a member of Research Data Service staff to carry out this process, though this step is under review.

Choice of identifier scheme

"First and foremost the DataCite service enables the minting of persistent identifiers (DOIs) for datasets. Bristol considered that it was worth buying into the DataCite infrastructure as over the longer term it was felt to be a useful source of advice and support and that the community was beginning to grow. The team also felt that when talking to researchers and academics about publishing their data, the ability to refer to a DOI, DataCite, and the British Library would help in the 'conversation' given that these are well known, authoritative reference points."

The data.bris project also considered registering directly with the Handle System (http://www.handle.net/) as a source of persistent identifiers.

Acquiring metadata

From the earliest stages of setting up the University of Bristol repository and the Research Data Service, a joined-up and logical approach to the technical challenges involved was a priority. The aim was to minimise the effort required on the part of the depositing researcher. Therefore, if information exists elsewhere within the University a researcher is not required to re-enter it. In theory, only two pieces of information are required at the time of data deposit: agreement to the Depositing Licence and a date on which to publish the dataset (this allows an embargo period to be specified). All other fields are pre-populated using information harvested form other University systems, although researchers are free to overwrite any pre-populated information.

Testing

Much of the deposit and publication process is bespoke and a good part of the 18 month Jisc-funded project was dedicated to developing the deposit interface. From a technical point of view this process is complete, but development continues from a user experience perspective. Once published, a check is made on each URL to confirm it is resolving correctly.

Examples of other users of DataCite and DOIs

- The University of Nottingham, via the ADMIRE project, stated their intention to assign DOIs on request [1]

- In April 2013, the Univeristy of Oxford reached an agreement to use DataCite as part of the strategy for their emergent Research Data Management Infrastructure (described in [2]). Oxford has committed to using the DataCite metadata set as a minimum in its catalogue for research data.

- The Australian National Data Service (ANDS) offers a national service [3] to publically funded Australian research organisations for minting DOIs through its membership of DataCite. The use of this service by Griffith University is described in [4].

Alternative identifier systems and services to choose from

Besides DOIs through DataCite, there are some alternative identifier and resolution systems and services available that can be used for data.

- The Handle System includes an open set of protocols, a namespace, and a reference implementation of the protocols. The protocols enable a distributed computer system to store identifiers, known as handles, of arbitrary resources and resolve those handles into the information necessary to locate, access, contact, authenticate, or otherwise make use of the resources. It is the underlying system used to resolve DOIs.

- Persistent URLS (PURLs) are URLs that offer a simple indirection service, with the main advantage of being very simple to use. A free service to support the registration and resolution of PURLs is offered by OCLC. See http://www.purl.org

- Archival Resource Key (ARK)s are actionable identifiers that can connect to three things: the object itself, a metadata record, and a commitment statement. They are championed by the California Digital Library and there are no fees for assigning or using ARKs. See https://wiki.ucop.edu/display/Curation/ARK

- EZID is a service that allows you to choose from a variety of persistent identifiers, including ARKs and DataCite DOIs, to create identifiers and store citation metadata and update URL locations. A demo is available as well as a programming interface. See http://n2t.net/ezid

Where to get more info

- DataCite in the UK http://www.bl.uk/datasets email: datasets@bl.uk

- DataCite information for potential UK clients: http://www.bl.uk/aboutus/stratpolprog/digi/datasets/datacitefaq/faqhome.html

- The data.bris research data repository http://data.bris.ac.uk/data/

- Guidance on citing research datasets from the University of Bristol http://data.bris.ac.uk/files/2014/02/Citing-research-data.pdf

- A webinar from the UKDA on how they have worked with DataCite to implement their DOI system, addressing persistence and versioning. http://www.jisc.ac.uk/events/data-identifiers-how-to-ensure-your-data-is-properly-cited-11-apr-2012

- Working with the British Library and DataCite Institutional Case Studies http://www.bl.uk/aboutus/stratpolprog/digi/datasets/DataCiteCaseStudies_2013.pdf

References

[1] Alex Ball. “Making Citation Work: A British Library DataCite Workshop”. July 2013, Ariadne Issue 71 http://www.ariadne.ac.uk/issue71/datacite-2013-rpt

[2] Sally Rumsey, Neil Jefferies. “DataFinder: A Research Data Catalogue for Oxford”. July 2013, Ariadne Issue 71 http://www.ariadne.ac.uk/issue71/rumsey-jefferies

[3] ANDS Cite My Data service http://www.ands.org.au/services/cite-my-data.html

[4] Natasha Simons “Implementing DOIs for ResearchData” D-Lib Magazine, Vol 18, No 5/6, May/June 2012

Acknowledgements

The input of the datsets team at the British Library is gratefully acknowledged.