Towards systemic solutions for preservation and archiving, citation and referencing systems for data and web-based 'non-published' outputs of research

DataCite

Data sharing and re-use are becoming increasingly central to the research process, meaning that it is vital to have effective tools to find, access and use that data. DataCite is a global network of national libraries, data centres and other research organisations that work to increase the recognition of data as a legitimate, citable contribution to scholarly record. DataCite provides Digital Object Identifiers (DOIs) for data sets and other non-traditional research outputs. DOI assignment helps to make data persistently identifiable and citable.

DataCite chooses to work with DOIs because they have number of features that make them more suitable than other persistent identifiers[1]. For example, they are already well-established for research publications, recognised as an ISO standard and are centrally managed by the International DOI Foundation (IDF).

The British Library is one of the founding members of DataCite and provides UK-based organisations with the ability to mint DOIs for their data. The British Library provides a web interface for DOI creation and the Metadata Store, which is a centralised database that makes data easy to find and access.

The infrastructure and governance provided by the IDF and DataCite means that there are standards, policies and guidelines that allow for data citation to take place.

However, this technical infrastructure cannot function in isolation. The participating universities and research organisations have to have their own repository infrastructure, senior level buy-in, policies and skilled staff in place to ensure that they can meet the requirements for providing persistent access to data.

This means that in addition to providing the technical infrastructure for DataCite, the British Library is actively engaged in advocacy work with all relevant communities to enhance overall knowledge and capability of the sector to deal with data. It is in this area, in particular, where we require further support to ensure that this service is adopted by universities and research institutions, especially as further recommendations for the UK approach to open data are developed.

DataCite currently has 18 UK members:

Eight universities: Bristol, Edinburgh, Glasgow, Imperial College, Leeds, Oxford, Southampton and University of East London.

Ten other UK research institutes and data services: Archaeological Data Service, Cambridge Crystallographic Data Centre, CEFAS - Centre for Environment, Fisheries and Aquaculture Science, F1000 Research, Marine Science Scotland, NERC data centres, STFC, Sir Alistair Hardy Foundation for Ocean Science, UK Atomic Energy Authority and UK Data Archive.

As well as leading on the UK work, the British Library is also one of the leading organisations within the international DataCite network, currently holding presidency of DataCite, contributing to the continuing international work on developing metadata standards and conducting research that will ensure interoperability between DataCite and ORCID research identifiers[2].

The British Library also provides services for international territories without their own DataCite network (currently Beijing Genomics Institute and Digital Repository of Ireland).

It is important to us that we develop capability that ensures that UK is at the leading edge of international collaboration to enable greater use and re-use of research data.

The Research Sector Transparency Board could help with this work by:

- Ensuring that there is a recognition of DataCite as an integral part of the UK research policy and landscape for open data;

[1] Examples of other persistent identifiers: URLs, Persistent Uniform Resource Locator (PURL), Global Unique Identifiers (GUILDs) etc.

[2] The project known as ODIN – *ORCID and DataCite Interoperability Network* - is a two-year project which started in September 2012, funded by the European Commission's 'Coordination and Support Action' under the FP7 programme. Partners in ODIN are CERN, the British Library, ORCID, DataCite, Dryad, arXiv and the Australian National Data Service.

- Making further links and endorsement from the UK research funders when considering either open data mandates, or when working to enable further capacity building and policy development in this area;
- Helping us to put in place a broader advocacy and engagement programme which could reach and build bridges across different communities, whose future engagement with this programme is essential – most of all research communities, but also university decision makers, policy networks, publishers etc.

UK Web Archiving and 'Link Rot'

Under Non-Print Legal Deposit regulations, the British Library together with other five legal deposit libraries, has started to archive the UK web domain once a year. The first crawl of the .uk domain took place between April and June 2013, capturing 4.86 million individual domains, containing 1.38 billion URLs, at a total data volume of 31.6 GB (the equivalent of nearly 11 million e-books). Once there is a scaleable means of determining which .com, .org and other non-.uk domains are in scope, this will increase considerably. This work is currently taking place.

It is planned to make this first domain collection available to researchers in December 2013. The user would need to visit a legal deposit library to see the archived copy.

Before the advent of Non-Print Legal Deposit in April 2013, the British Library and its partners operated a permissions-based approach to web archiving, by gaining direct permission from site owners to archive their sites and make those archived copies publicly available through the Open UK Web Archive - webarchive.org.uk. This approach is still in operation, over and above legal deposit, for content of particular importance. The public archive, available on-line, now contains 13,500 archived sites, including a great many of scholarly and research value.

There is a significant level of web archiving happening worldwide. The most significant resource is the Internet Archive - archive.org – founded in 1996 and based in San Francisco.

The Internet Archive model is radically different to what we are implementing in the UK – it is a not-for-profit enterprise funded by commercial revenue and philanthropic donations, exercising 'right to remember', but not based on specific legislative underpinning. This provides for certain flexibility in approach, but also creates a number of legal challenges.

The majority of European national libraries have some level of web archiving capability and some of them operate on a similar scale as the British Library.

In the UK, another significant web archiving institution is the National Archives, which archives UK government websites.

Web archives constitute a new class of scholarly resource, and the requirements that researchers have of web archives are still in a state of flux and imperfectly expressed. This is particularly the case with citation methods, which are not systematically addressed anywhere even though a number of projects have attempted it - notably at the Harvard University web archive and the Webcite project - webcite.org.

With the advent of a greater amount of scholarly content being made available on the web, including via different forms of open access publications, the 'link rot' is a problem that is likely to grow. The need for a stable and acceptable method of citing archived web sites is one that researchers have expressed to the British Library in a number of contexts. The subject is ripe for investigation by researchers and archivists in collaboration.

We believe that the most satisfactory way of dealing with this issue would be to link a citation system with archived websites, which means that a suitable permanent identifier would point to already preserved content.

The next steps for this work could include:

- The British Library would like to explore with the UK research funders a potential for archiving any web-based research outputs created in the course of publicly funded research projects, which could include any transitory websites, blogs, research project sites and similar outputs. This would enable the British Library to preserve this material in a more accessible way than what is possible through a general web crawl under Non-Print Legal Deposit, which does not allow us to make this material accessible on-line as a part of the Open UK Web Archive.
- The British Library would be interested to facilitate and undertake further collaborative work regarding web citation to deal with the issue of 'link rot'. In our opinion the first step should include a consultative process with research community to determine what such referencing system should look like and how scholarly use of web citation should operate. We believe that any technical solution for 'link rot' (possibly similar to what we are implementing for data) should be preceded with some work that would improve our understanding of how web outputs and citation are perceived by researchers.