Implementing DOIs for research data

Natasha Simons, Griffith University, Australia

Abstract
Research is increasingly collaborative and global in nature and efforts to manage the vast amounts of research data generated daily require global solutions. The Digital Object Identifier (DOI) system provides a means of persistent identification of research data collections and datasets that is global, standardised and widely used. The Australian National Data Service (ANDS) partnered with DataCite to offer a DOI minting service. At Griffith University, implementing DOIs raised governance questions common to other institutions that encouraged discussion and collaboration.

Introduction
Huge volumes of research data, largely born digital and enabled by vast advances in computing power, are being generated worldwide. The world’s most powerful telescope, the Square Kilometre Array, will generate computer data in a single day that is equivalent to the amount generated by the whole world in a single year.¹ CERN’s Large Hadron Collider, built to help scientists answer key unresolved questions in particle physics, will produce roughly 15 petabytes of data annually which is enough to fill more than 1.7 million dual-layer DVDs a year.² It is predicted “The volume of data generated in research and by scientific instruments will soon dwarf all the technical and scientific data collected in the history of research”³. ‘Big data’ and the ‘data deluge’ affects all academic disciplines from science to the humanities and brings enormous opportunities for research along with difficult challenges in data storage and management.

Research institutions are faced with the immensely difficult task of finding ways to store and manage data in a format that facilitates discoverability, accessibility, and re-use. Data that is richly described, organised, integrated and connected allows the data to be more easily discovered by other researchers who may pose new questions to be answered, raise larger issues to be investigated, and identify data landscapes to be explored⁴. “Access to data enables system-level science, expands the instruments and products of research to new communities, and advances solutions to complex human problems”⁵.

Growing culture of data citation
Traditionally, knowledge derived from research is shared in the form of a publication such as a journal article. However the data used to produce the research publication is effectively lost to poor archival practice and only a very small proportion of the original data is made available in conventional journals. The result is unnecessary duplication of effort through re-creation of existing data, and an inability to verify results or re-purpose the data⁶. The vision of open access to data is to make it accessible and usable to anyone, anytime, anywhere, and for any purpose. Publicly available data is associated with significant increase in citations irrespective of journal impact factor, date of publication, and country of origin⁷.
As part of a global effort to improve access to research data, there is growing impetus for an international culture of data citation using the Digital Object Identifier (DOI) system. “Data citation refers to the practice of providing a reference to data in the same way as researchers routinely provide a bibliographic reference to printed resources.” DataCite is a global not-for-profit organisation formed in London in December 2009 that is facilitating the growing culture of data citation for scientific content. A key aim of DataCite is to “increase acceptance of research data as legitimate, citable contributions to the scholarly record” and they work with organisations that hold data to provide persistent identifiers in the form of DOIs.

**The DOI system**

The DOI system provides a framework for persistent identification, managing intellectual content, managing metadata, facilitating commerce and linking customers with content. DOIs are an implementation of the Handle System for persistent identifiers and seamlessly transport the user from one interface to another without requiring specific software. Information about a digital object may change over time, including where to find it and who owns it, but its DOI will not change.

A DOI is made up of alphanumeric characters and must be unique. It consists of a prefix and a suffix separated with a forward slash. The prefix always begins with ‘10’ as this distinguishes it from other implementations of the Handle System and then states the registrant code designating the creating organisation or publisher that is registering the DOI. The suffix identifies the individual work and is also known as the ‘item id’. It is assigned by the publisher/owner of the DOI.

The system evolved from the publishing industry and solidified in 1998 with the founding of the International DOI Foundation (IDF), an open membership consortium. A DOI Registration Agency infrastructure provides ongoing support and maintains quality and accuracy of DOI names. Agencies are appointed by the IDF to provide service, quality assurance and overall integrity of the DOI system. CrossRef, a consortium of around 3000 publishers, is one of many IDF Registration Agencies.

In addition to the International DOI Foundation (IDF) and Registration Agencies structure, DOIs differ from other persistent identifiers in that they require a minimal amount of metadata to be provided at the point of obtaining each DOI. They also require a commitment from the provider to maintain the URL associated with the DOI. In the context of research data, this requires the provider of the data to maintain access to, and preservation of, the data persistently over time.

The “DOI has emerged as the most widely used standard for digital resources in the publication world.” The DOI is an ISO International Standard and over 55 million DOI names have been assigned by DOI System Registration Agencies in the US, Australasia, and Europe. In December 2010, only one year after its formation, DataCite had registered over one million DOI names for research data material.
ANDS “Cite My Data” service

As part of a broad strategy to achieve its goals, the Australian National Data Service (ANDS) joined DataCite and in 2011 launched a Cite My Data service to offer minting of DOIs to Australian research institutions. ANDS is leading the creation of the ‘Australian Research Data Commons’, a cohesive Australian collection of research resources. The Commons will produce a richer data environment that will make better use of Australia’s research outputs and enable researchers to easily publish, discover, access and use data. The Cite My Data machine-to-machine service is offered free of charge to ANDS-partners and facilitates minting of DOIs for research data collections and datasets using DataCite as the DOI Registration Agency.

As an ANDS-partner institution, Griffith University ranks among Australia’s top-10 research universities and is placed in the top 500 Academic Ranking of World Universities. Funding from ANDS and internal sources has assisted Griffith to facilitate capture, discovery and re-use of the data produced by its researchers. The process has involved infrastructure development including a research data repository and the Griffith Hub, a metadata store solution with a semantic web implementation and a discovery layer. In addition to local discovery portals, Griffith’s research data collections and datasets are exposed via the ANDS Research Data Australia service and the National Library of Australia’s Trove service.

Implementing DOIs

The ANDS Cite My Data service provided Griffith University with a first-time opportunity to mint DOIs. The eResearch Services team within Scholarly Information and Research, Division of Information Services, are responsible for management of research data collections at the institution and volunteered to participate in the trial and production implementation. The scope included investigating the use of DOIs, their application to the Griffith collection, the benefits, implementation issues and workflows.

It became clear early in the trial of the Cite My Data service that implementing DOIs would be technically simple but required the resolution of a number of difficult governance questions. These included: What type of material should be assigned a DOI? Is the DOI a replacement or a compliment to other persistent identifiers? At what level of granularity should a DOI be applied? Should the DOI link to a landing page or to the actual digital dataset? What process should be followed regarding the DOI when the content of the object it is assigned to changes and a different version results? Who should mint and maintain the DOI for a data collection produced through research collaboration between different institutions? Should researchers be provided with direct or mediated access to DOI minting for their collections or should research data administrators facilitate this process? How will Griffith sustain support for DOIs in the long term? These questions are relevant to other institutions and a broad, open and collaborative approach is proving beneficial for resolution.
In late 2011 ANDS organised a Data Citation workshop featuring a range of presenters including the DOI expert Jan Brase, chair of the International DOI Foundation and Manager of DataCite. The DOI governance questions raised at Griffith were discussed and ANDS made responses available via their website. Answers to governance questions have been further progressed through collaborative, open discussion with other institutions implementing DOIs and a Griffith University Guide to Minting DOIs has been drafted.

**DOIs and collaborative research**

It is increasingly the case that research is the result of collaboration between researchers across institutions and across state and international borders. Therefore it’s necessary to provide guidelines of who, in a collaborative research project, mints the DOI for a data collection or dataset. While it is technically feasible to mint different DOIs for one data collection shared between institutions, each with a different landing page, this runs against the grain of the DOI system. Where more than one institution stores the same data collection, and each have their own landing page to that collection, only one DOI should be minted for the collection. Each institution then refers, in the metadata record about the collection, to the same DOI.

However there is no formal agreement between Australian research institutions that can provide guidance on this question. The solution may need to be resolved between researchers working on the collaborative project or between research data administrators at respective institutions. Considerations might include: Who is the primary researcher or author? Who is the lead institution? Who provides access to the material? Who is going to maintain access to the material in the long-term? Which institution has the capacity to mint a DOI?

**Versioning**

Material such as data collections and datasets are subject to change in version, scope and content over time. If a DOI is minted and assigned to a research data collection that is later changed (updated, revised, expanded) then a new DOI will need to be issued for the later version of the material. Otherwise a data citation may point to an incorrect version of the research data. This raises a technical issue. The relatedIdentifier metadata element within the DataCite Metadata Schema can be used to link versions to each other. However, there are only five mandatory metadata elements required to mint a DOI through DataCite. Implementing technical support for the full DataCite Schema will enable the provision of richer metadata such as versioning. The Schema is also subject to further changes via the activities of the DataCite Metadata Working Group, of which ANDS is a member. Additionally, consideration will need to be given to alerting users to the existence of different versions of the data collection, the differences between each version and their respective DOI.

**DOIs and other persistent identifiers**

Persistent identifiers are not a new concept and Griffith supports a range of these including Handles, ANDS-issued persistent identifiers, National Library of Australia identifiers and local institutional identifiers. These identifiers apply to different types of material including research publications, research data and
records about researchers themselves. In this mesh of persistent identifiers, what type of material do DOIs apply to and what is the advantage of using a DOI over another persistent identifier such as a Handle?

To assist institutions in deciding when to apply a DOI, a ‘Persistent Identifiers Decision Tree’ was produced by ANDS. The Tree suggests a DOI should be applied when: the data will be exposed and forms part of the scholarly record; the data can be kept persistent; and the minimum DataCite metadata schema requirements can be met. In deciding whether to assign a DOI to research data, additional questions may include: does the institution have the authority to expose and manage the research data? What happens to the data, and the DOI, when a researcher moves from one institution to another? What ‘grey literature’ such as theses and unpublished discussion papers should be issued with a DOI?

DOIs are intended to be persistent and to ensure research data becomes a citeable part of the scholarly record a DOI issued for a data collection requires persistent support of the data. However guaranteeing access to data over an indefinite period of time is problematic for research institutions. Digital storage requirements are ever-increasing, exponentially in some academic disciplines, along with associated storage costs. Research institutions are faced with the task of finding secure storage for large volumes of new data while maintaining storage of older digital data.

**Granularity**
One of the benefits of the DOI system is that DOIs can be assigned at any level of granularity. For example, a DOI can be assigned to a data collection and also to each item within the data collection e.g. a collection of digital film and each film within the collection. In deciding what level of granularity to apply a DOI to, consideration may be given to the expectations of data users. Is the material at a more granular level within each data collection likely to be cited? Does the material at an item level meet basic DOI requirements in terms of access, persistence and metadata?

**Storage and display**
Rather than pointing directly to a digital dataset, the DOIs minted by Griffith will point to a landing page in the Griffith Research Hub. The landing page contains relevant metadata that describes the digital object it links to and includes information about access rights and restrictions. In future it will also contain the preferred citation. The DOI will be stored in Griffith systems and will be displayed in the Research Hub and other discovery portals such as the Griffith Data Repository and Griffith Research Online. Importantly, the DOIs minted by Griffith will be included in the feed to collection records provided to the ANDS Research Data Australia service, the premier discovery service for Australian research data collections.

**Approaches and workflows**
There are a number of research data projects in 2012 at Australian institutions that involve the implementation of DOIs and a collaborative, open discussion between people working on some of these projects has begun. Two distinct
approaches and subsequent workflows have emerged from these discussions. One approach is to develop an interface for researchers that allows them to mint or place a mediated request for a DOI name for their data collection. Another approach is to mint DOIs for data collections that have already been curated and are maintained by research data administrators. The latter doesn’t require liaison with researchers though it will be important to communicate the DOI concept and its use in data citation. The two different approaches require different workflows and communication strategies though it is likely that as DOI implementation develops, an institution will support both workflows. Collaboration and information sharing between institutions implementing DOIs is expected to continue and there has been an informal agreement to make technical scripts available as open source.

Citation tracking
A method of tracking citation metrics will be required to evaluate the success of DOIs as tool for research data citation. DataCite have created a central metadata repository and allow Thomson Reuters to crawl the repository so that the metadata appears in the *Web of Science*. DataCite and CrossRef are also exploring the possibility for publishers to search the metadata repository and identify relations from datasets to their articles\(^2\). In 2011 BGI launched the GigaScience Journal\(^2\) as a 'big data' open-access open-data database\(^2\). A DOI can be assigned upon data submission to the journal to allow discovery, citing and citation tracking.

Conclusion
At Griffith University, implementing DOIs has raised governance questions common to other institutions that encouraged discussion and collaboration. In so doing, it has resulted in a new approach to a complex area. Further open and collaborative discussion will be important in the wider adoption of DOIs for research data collections by Australian research institutions.

---

2 European Organization for Nuclear Research (CERN), *The Large Hadron Collider.*
3 Joint Information Systems Committee (JISC), *The Data Deluge.*
4 Australian National Data Service (ANDS), *About ANDS.*
6 Brase, J. (2009). DataCite – A global registration agency for research data. *COINFO '09 Proceedings of the 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology:* 257-261. [http://dx.doi.org/10.1109/COINFO.2009.66](http://dx.doi.org/10.1109/COINFO.2009.66)
8 Australian National Data Service (ANDS), *Data Citation Resources*.
9 DataCite
10 DataCite, *What is DataCite*.
12 CrossRef
13 Brase, J. (2009). DataCite – A global registration agency for research data. *COINFO '09 Proceedings of the 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*: 258. [http://dx.doi.org/10.1109/COINFO.2009.66](http://dx.doi.org/10.1109/COINFO.2009.66)
14 The International DOI Foundation, *The DOI System*
15 DataCite, *Blog*.
16 Australian National Data Service.
17 Australian National Data Service (ANDS), *Cite My Data Service*.
18 Griffith University, *Griffith News*.
20 Australian National Data Service (ANDS), *Research Data Australia*.
22 Australian National Data Service (ANDS), *Data Citation Workshop*.
23 Australian National Data Service (ANDS), *DOI Questions and Answers*.
25 ANDS, *Data Citation Resources*.
26 DataCite, *Metadata Schema Repository*.
27 DataCite, *Blog*.
28 GigaScience Journal