Creating Citable Data Identifiers

Ryan Scherle (ryan@scherle.org), Dryad Digital Repository
Mark Diggory (mdiggory@atmire.com), @mire, Inc.

Data archiving is becoming increasingly important for science. As scientists become more aware of the need to archive data, they impose new requirements on repositories. One requirement of particular importance is support for data citation. For data citations to be most effective, a repository's identifier scheme must align with scientists' expected citation practices. We clarify the requirements necessary for repository platforms to support data citation and describe how well predominant repository platforms support data citation in the scientific publication process. Finally, we discuss the support for data citation in the Dryad repository via DOI identification and versioning of data products.

Principles of Citable Identifiers

Science depends on researchers' ability to build on previous work. To accurately replicate research, scientists need access to both publications and the data underlying those publications. For hundreds of years, scientists have been using citations to locate publications of interest and to guide future researchers to relevant publications. As data archiving becomes more common, scientists are starting to create citations cite specific to data objects. However, technical support for data citations varies greatly between individual repositories. If repositories are to become respected archives for scientific data, they must assign identifiers in ways that align with citation practices.

Previous work has determined that data citations should include information similar to publication citations, with minor additions. For example, Altman and King [1] argued that data citations should include the following: author(s), publication date, title, persistent identifier, and numeric fingerprint. Other proposals include slightly more or less information, but they are remarkably similar (e.g., [2][3]).

Researchers interact with publications in different ways than they interact with data. Data may be modified or supplemented after it has been published. Data may be used in machine-readable files, allowing for automatic replication of experiments. Data from multiple sources may be combined, enabling "synthetic science". Although these differences have little effect on most elements in a citation, they lead to more demanding requirements for data identifiers.

We were keenly aware of these issues while developing the identifier system for Dryad [4], a repository of data underlying peer-reviewed articles in the basic and applied biosciences. To ensure that Dryad identifiers met the needs of the science community, the Dryad development team sought feedback from a variety of scientists and journal editors. After much discussion, the

community agreed on basic properties of citable identifiers. Although these initial requirements come from the bioscience community, informal conversations with other scientists indicate that similar requirements exist in other science disciplines. The basic principles of Dryad's data identifiers are as follows:

1. **Data must be identified using DOIs**. Over the past decade, scientists have become accustomed to DOIs. Scientists are comfortable citing and resolving DOIs, even when they have no knowledge of other types of persistent identifiers. Many tools are available for managing DOIs and their associated metadata. Most importantly, to a scientist, a DOI conveys a certain amount of gravitas, indicating that a resource is worthy of citation.
2. **Identifiers should be as simple as possible.** Even though most identifiers are used in an online environment, there are still many cases where humans manually interact with identifiers. A scientist may locate an identifier in a print publication and want to view the associated data. A scientist may visually inspect two identifiers to determine whether they refer to the same object. Identifiers should be relatively short and simple to make these human-mediated activities as error-resistant as possible.
3. **Identifier syntax should clearly illustrate immutable relationships.** In the repository community, conventional wisdom has been to minimize the amount of semantics embedded in identifiers. Semantic information can make identifiers "brittle" and prone to failure. However, as described above, there are many circumstances in which identifiers are directly handled by humans. In these cases, scientists want critical information to be clearly visible within identifiers, including hierarchical and versioning relationships.
4. **When the content of a data file changes, it should receive a new identifier.** To support scientific review and analysis of research data, repositories must retain citable links to all published revisions of a data asset. Future researchers need the ability to view the data files that were originally associated with a publication, as well as any subsequent modifications. It is not the place of repositories to determine validity after data is published; however, it is the place of repositories to document corrections that may be necessary over the data's lifecycle.
5. **Metadata must be editable without creating a new identifier**. Edits to metadata frequently involve simple changes like corrections of typographical errors or addition of new subject keywords. Scientists typically do not want to change their citations when the actual content of data files has not changed. However, some metadata modifications have a greater effect on the interpretation of the data. In these cases, it should be possible to create a new version of the data object.

Repository Support for the Principles

Repository software typically provides support for persistent identifiers. But most repositories were not designed with a focus on scientific citation. Below, we summarize the identifier capabilities of various repository systems, with a particular focus on the capabilities of repositories for managing versioned content.

Fedora Commons

Fedora Commons has an internal identifier system that is accessible to external systems. Repository developers may choose to use this internal system or layer a new system on top of it. Thus, Fedora allows for an arbitrary identifier scheme, including schemes that follow the principles outlined above. Fedora stores a version history of every change made to a datastream. While this approach is desirable for maintaining the history of a repository, it is too granular for the purposes of citation. Given the requirements for identifiers to remain the same for metadata changes while changing for underlying bitstream modifications, a new identifier system must be applied to the repository for citation purposes.

EPrints

EPrints identifiers are based on the domain of the repository system. External identifier systems are not supported. EPrints has rich support for versioning. In EPrints, previously unrelated items may be grouped together and related by associating them in a version history. The version history need not be linear; it may include a hierarchy of versions (e.g., see http://files.eprints.org/319/). While different versions can be referenced by their identifier, the identifiers do not contain semantic version information.

DSpace

At the present time, DSpace only supports assignment of CNRI handles as external persistent identifiers. No support exists in DSpace to manage assigning of external identifiers such as DOI, ARK or PURL. Likewise, no versioning support exists on DSpace item or bitstream contents. In 2006, the DSpace community formed an Architectural Advisory group that clarified requirements for a number of important features that DSpace should support [5]. These features include the following: allowing use of more than one identification system; versioning of individual DSpace items; retaining both metadata and content-specific changes to revisions while allowing for metadata alteration independent of revision; and enabling version-specific identification in identifiers. While several prototype projects were initiated within the DSpace Community to address these requirements, none were formally contributed into recent DSpace releases.

Identifiers in Dryad

Dryad is built on top of DSpace. At the time we started working on version control, there was no support in the DSpace system for non-Handle identifiers or versioning. Rather than view this as a deficiency, we took this as an opportunity support the principles stated above by designing the required features from the ground up. These features were implemented as two new add-on services for DSpace; Identifier Service and Versioning Service.

Dryad's extensible Identifier Service supports one or more providers of external identifiers. For Dryad requirements, we implemented a DOI-based provider capable of reserving and registering DOIs via the California Digital Library EZID services [6]. Dryad's Versioning Service

provides the ability to generate a new version of an Item when desired by a scientist or a data curator. When a user selects the option to create a new version, the system generates a duplicate of the original item with a new revision identifier. The user may modify the new item in any way. Modifications occur in the user's private workspace. Other users of the repository will not see changes to the original item until the new version is complete and has been approved by a curator. Dryad Versioning and Identifier services work in concert to assure that proper new identifiers are reserved whenever a new version of an Item is created. Identifiers are not registered externally until after the data has been approved by a curator and made public. More details on Dryad's identifier system are available in [7].

While the Dryad source code is available to the public [8], the intention is that the Versioning and Identifier Services be contributed back to DSpace community for proper inclusion into a future DSpace release. The contribution of these features will allow all other DSpace repositories to support the versioning and identification requirements necessary for data citation.

Conclusion

Community practices for data citation are still evolving. Sometimes these practices arise from how the content is used. Too often, the practices are dictated by the technological capabilities of repository systems. Although we realize that the principles outlined in this paper will not meet the needs of every community, we hope that other repositories will expand their capabilities to accommodate these principles, allowing scientists to use a broader array of repositories for storing and citing scientific data.

References

[1] Altman, M. King, G. (2007). A Proposed Standard for the Scholarly Citation of Quantitative Data. D-Lib Magazine. 13(3/4). http://www.dlib.org/dlib/march07/altman/03altman.html
[2] Green, T (2009), "We Need Publishing Standards for Datasets and Data Tables", OECD Publishing White Paper, OECD Publishing. http://dx.doi.org/10.1787/603233448430
[3] How to Cite a Data Set. International Polar Year Data and Information Service. http://ipydis.org/data/citations.html
[4] Dryad Digital Repository. http://datadryad.org
[5] DSpace architecture review group. (2007). Toward the next generation: Recommendations for the next DSpace Architecture http://bit.ly/yRAzI5
[6] California Digital Library EZID services http://www.cdlib.org/services/uc3/ezid/
[7] Michener et al. (2011) DataONE: Data Observation Network for Earth — Preserving Data and Enabling Innovation in the Biological and Environmental Sciences, D-Lib Magazine 17(1/2) http://dx.doi.org/10.1045/january2011-michener
[8] Dryad Source Code. http://dryad.googlecode.com