UC Office of the President
CDL Staff Publications

Title
isCitedBy: A Metadata Scheme for DataCite

Permalink
https://escholarship.org/uc/item/6r03h784

Authors
Starr, Joan
Gastl, Angela

Publication Date
2011-01-24

eScholarship.org Powered by the California Digital Library
University of California

PRINTER-FRIENDLY FORMAT Return to Article

D-Lib Magazine

January/February 2011
Volume 17, Number 1/2

isCitedBy: A Metadata Scheme for DataCite

Joan Starr
California Digital Library
joan.starr@ucop.edu

Angela Gastl
ETH Zürich Library
angela.gastl@library.ethz.ch

doi:10.1045/january2011-starr

Abstract

The DataCite Metadata Scheme is being designed to support dataset citation and discovery. It features a small set of mandatory properties, and an additional set of optional properties for more detailed description. Among these is a powerful mechanism for describing relationships between the registered dataset and other objects. The Scheme is supported organizationally and will allow for community input on an ongoing basis.

Keywords: metadata, data citation, data description, data relationships

Introduction

Scholarly research across all disciplines is producing dramatically increasing amounts of data [1]. The knitting together of published research articles and the research data that substantiate their findings is of increasing importance as more disciplines take advantage of data-driven approaches to knowledge acquisition. One measure of the growth in these numbers is the experience of a single scientific journal, as described in the final report of NISO's Roundtable on Best Practices for Supplemental Journal Article Materials [2]. Between the years 1999 and 2009, the Journal of Clinical Investigation saw the portion of published research articles with supplemental data grow from zero to 87 percent, with 2010 trending toward 100 percent.

The result is a scholarly communication environment that is rapidly growing in both complexity and diversity of content. In this context, what has been missing until recently, is a *persistent* approach to access, identification, sharing, and re-use of datasets. Persistence is built on a two-part foundation. The first part is the trusted connection between an opaque identifier and an object. The second is the long-term maintenance of metadata about the object. In support of this critical piece, there must be a working infrastructure in place that meets the needs of the key constituents; in this case, academic researchers. A well-formed and right-sized metadata scheme is an important element of that infrastructure.

When the DataCite Consortium was founded in 2009, the development of a DataCite metadata scheme was an early priority. The Metadata Working Group was one of the four working groups to be initiated in the earliest meetings of the Consortium. The first two drafts of the DataCite metadata scheme emerged as a result of some of the Consortium's first discussions of the basic metadata schema for data used by the German National Library of Science and Technology (TIB), the first DOI Registration Agency for data from 2005 to 2009, and one of the founders of DataCite.

At the present time, DataCite is working with Digital Object Identifiers (DOIs) [3], which are one of several existing identifier schemes. DOIs are administered by the International DOI Foundation. Prior to DataCite's founding, DOIs had been used primarily for scholarly articles, and were identified fairly strongly with that model. In asserting that DOIs can be used equally effectively for datasets, DataCite must face the particular challenges of persistently identifying scientific data. Specifically, these include the need to link, at a very granular level, to components of a dataset, and to clearly identify relationships between components of one or more datasets.[4] Equally, there is a need to accommodate versions of datasets, which frequently go through many more iterations than typical scholarly publications.

In this paper we will discuss the development of, and next steps for, the Metadata Working Group's metadata scheme as an important way to address these challenges.

Objectives of the scheme

The scheme is designed to support DataCite's goals to "establish easier access to scientific research data on the Internet, increase acceptance of research data as legitimate, citable contributions to the scientific record, (and) support data archiving that will permit results to be verified and re-purposed for future study" (http://datacite.org/whatisdc.html). It should perform certain functions and provide a foundation for other functionality.

More specifically, the objectives of the scheme are to:

- recommend a standard citation format for datasets, based on a small number of properties required for identifier registration;
- provide the basis for interoperability with other data management schemas, many of which are domain specific;
- promote dataset discovery with optional elements allowing for flexible description of the resource, including its relationship to other resources, and other versions of the same resource; and
- lay the groundwork for future services (e.g., discovery) through the use of controlled terms from both a DataCite vocabulary and external vocabularies as applicable.

To achieve these objectives, the DataCite Metadata Working Group has endeavored to produce both a working version of the scheme and to establish procedures for its ongoing maintenance and support.

The core for citation

The metadata scheme's core is composed of a discreet number of required properties. It was determined that this set would be restricted to the information necessary to compose a citation. Table 1 shows these properties.

Table 1: DataCite Metadata Kernel: the mandatory properties

ID	*Property*	*Description and instructions*
1	Identifier	The Identifier is a unique string that identifies a resource. At present, the only allowed value is a DOI.
2	Creator	The main researchers involved in producing the data, or the authors of the publication, in priority order.
3	Title	A name or title by which a resource is known.
4	Publisher	A holder of the data (including archives as appropriate) or institution which submitted the work. Any others may be listed as contributors. This property will be used to formulate the citation, so consider the prominence of the role. In the case of datasets, "publish" is understood to mean making the data available to the community of researchers.
5	PublicationYear	The year when the data was or will be made publicly available. If an embargo period has been in effect, use the date when the embargo period ends.

In addition to these required properties, there are additional but optional properties that may be used in the citation when present and as appropriate, for example, version and resource type (e.g., dataset). Because many users of the scheme will be members of a variety of academic disciplines, and because DataCite must remain discipline-agnostic, DataCite recommends rather than requires a particular citation format, namely:

> Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier

For example,

> Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2. Geological Institute, University of Tokyo. Dataset. doi:10.1594/PANGAEA.726855. http://dx.doi.org/10.1594/PANGAEA.726855

Note that the Identifier may be expressed in the "native" format and/or in HTTP format, depending upon the requirements of the style sheet guidelines governing the publication.

Options for greater description

The metadata scheme provides a set of optional descriptive properties for users wanting to provide more detail about registered resources and, as desired, their relationship to other resources, including, for example, component pieces, prior versions, and referential material. The data centers and others who are registering DOIs with DataCite are free to store additional metadata fields in their own system catalogues. The DataCite scheme is the common denominator for metadata exchange. Table 2 shows the optional properties.

Table 2: DataCite Metadata Kernel: the optional properties

ID	*Property*	*Description and instructions*
6	Subject (with scheme attribute)	Subject, keywords, classification codes, or key phrases describing the resource.
7	Contributor (with type attribute)	The institution or person responsible for collecting or otherwise contributing to the development of the dataset.
8	Date (with type attribute)	Different dates relevant to the work. May be repeated to indicate a date range.
9	Language	Primary language of the resource.
10	ResourceType	The general type of a resource.
11	AlternateIdentifier (with type attribute)	An identifier other than the DOI applied to a resource. Any alphanumeric string which is unique within its domain of issue.
12	RelatedIdentifier (with type and relation type attributes)	Identifiers of related resources. Use this element to indicate subsets of elements, as appropriate.
13	Size	Size information about the resource; unstructured.
14	Format	Technical format of the resource.
15	Version	Version number of the resource. If the primary resource has changed the version number increases.
16	Rights	Any rights information for this resource. Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information. This could be at least a standard statement or include embargo information.
17	Description (with type attribute)	All additional information that does not fit in any of the other categories. Can be used for an abstract.

Two of the biggest potential challenges for DOIs in terms of their suitability for handling datasets, as noted earlier, pertain to the description of multiple subsets and versions of datasets. Properties 11, 12, and 15 (*AlternateIdentifier*, *RelatedIdentifier*, and *Version*) are designed to describe a complex set of relationships between objects and components of an object.

Both *AlternateIdentifier* and *Version* allow users to describe the object itself in more detail. That is, with *AlternateIdentifier*, it is possible to enter alternative unique identifiers that are associated with objects, so that they can be recognized as part of a particular context. Equally, if the object is registered with an identifier early in its existence, prior to deposit in a repository, then the early identifier can be stored in this field. In this way, the scheme provides a kind of life cycle support for the object.

The *Version* property, as noted earlier, may be used as part of the citation. It can be coupled, in effect, with the *RelatedIdentifier* property when describing the relationship between two identifiers that are versions of one another. DataCite does not enforce any validation rule that a resource ought to be re-registered each time it undergoes a version change. However, this is considered a recommended best practice for resource citation.

The glue that holds these pieces together is the *RelatedIdentifier* property. This is the place where references to other objects can be made. Importantly, the scheme provides a detailed and controlled list of relationship types in pairs, as shown in Table 3.

Table 3: These Related Identifier Values precisely describe relationships to other digital objects.

Allowed values	***Description and instructions***
IsCitedBy	Use to indicate the relation to the work that cites/quotes this data.
Cites	Use to indicate the relation to the work which the resource is citing/quoting.
IsSupplementTo	Use to indicate the relation to the work to which the resource is a supplement.
IsSupplementedBy	Use to indicate the relation to the work(s) which are supplements of the resource.
IsContinuedBy	Use to indicate the resource is continued by the work referenced by the related identifier.
Continues	Use to indicate the resource is a continuation of the work referenced by the related identifier.
IsNewVersionOf	Use to indicate the resource is a new edition of an old resource, where the new edition has been modified or updated.
IsPreviousVersionOf	Use to indicate the resource is a previous edition of a newer resource.
IsPartOf	Use to indicate the resource is a portion of another resource.
HasPart	Use to indicate the resource is a container of another resource.
IsReferencedBy	Use to indicate the resource is used as a source of information by another resource.
References	Use to indicate the relation to the work which is used as a source of information of the resource.
IsDocumentedBy	Use to indicate the work is documentation about/explaining the resource referenced by the related identifier.
Documents	Use to indicate the relation to the work which is documentation.
IsCompiledBy	Use to indicate the resource or data is compiled/created by using another resource or dataset.
Compiles	Use to indicate the resource is used for creating another resource or dataset.
isVariantFormOf	Use to indicate the resource is a variant or different form of another resource, e.g. calculated or calibrated form or different packaging.
isOriginalFormOf	Use to indicate the relation to the works which are variant or different forms of the resource.

With use of the *RelatedIdentifier* property with the relation type attribute, all of the following scenarios (and more) can be described:

- the connection between a scholarly article and the dataset upon which the research is based;
- the relationship between one version, or variant form, of an object and another; and
- the association between the whole and its parts, or the main body and its supplements.

These properties allow for repeating occurrences and can depict a wide range of nuanced relationships thereby providing a great degree of descriptive power. With these and the rest of the property set, the scheme fulfills the vision described recently by Helliwell and McMahon for a standard scheme that would permit "describing composite documents (including research article, component figures and tables, associated data sets and other supplementary materials) that allows individual components to be hosted on different platforms." [5]

Organizational support

The Metadata Working Group has always assumed that there would be an ongoing need for updates and inputs to the metadata scheme. One reason for the clarity on this issue is the nature of the group itself. It is a highly collaborative body, composed of members from eleven libraries and research organizations spread across ten countries and three continents. The working arrangements of the various DataCite member institutions vary, which means that representatives from each organization have been able to articulate different use cases and requirements for the metadata scheme.

Likewise, over time, changes to the metadata scheme will come from a number of sources. The most immediate are the direct requests from the data centers, universities, and researchers served by DataCite members. In addition, DataCite and DataCite members are active in the scientific data and data publishing communities, and discussions and information exchanges in these groups may surface new metadata requirements. Lastly, there are other task forces and working groups within the DataCite organization that may have interdependencies with the Metadata Working Group, which could lead to requests for changes to the scheme.

To meet the need for organizational support for the metadata scheme, including providing a mechanism for community input and scheme versioning, DataCite has named a Metadata Supervisor. This is a regular staff position at the German National Library of Science and Technology, the TIB, which is the hosting institution for the managing agency of DataCite. The Metadata Supervisor's exact workflow and procedures have yet to be precisely determined, for example, in terms of how often the scheme will be updated and how community members will be able to provide input. The Working Group may serve in an advisory capacity.

Next Steps

At the time of writing, the Working Group is completing the revision tasks following the community feedback period. Some aspects of the scheme will develop over time, including the acceptance of primary identifiers other than, or in addition to, a DOI. This change would affect the mandatory set (the allowed values for the Identifier property), and it would be a change that opens up the scheme considerably, increasing its utility for a broader community of academic researchers.

Once the scheme is in a final version, it will be converted to an XML schema format and published on the DataCite website for implementation by all DataCite members. The Metadata Supervisor will also put into place procedures for maintenance and make publicly available the schedule of updates, and mechanisms for community input.

In his well known call for publishing standards for datasets, Toby Green imagines a world in which everything a scholar creates is "compatible with and discoverable from all scholarly publishing and discovery systems," [6] and easy for publishers, librarians, and most of all, readers to find and use. DataCite is working to build toward this vision, and the DataCite Metadata Scheme is one of the foundational components.

Acknowledgements

The authors would like to give special thanks to: Jan Ashton (British Library), Patricia Cruse (California Digital Library), Alfred Heller (DTU Library), John Kunze (California Digital Library), Lynne McAvoy (CISTI), Elizabeth Newbold (British Library), Madeleine de Smaele (TU Delft), Anja Wilde (GESIS), and Wolfgang Zenk-Möltgen (GESIS).

We also acknowledge the contributions the other members of the metadata working group: Jan Brase (TIB / DataCite), Paul Bracke (Purdue University), Jacqueline Gillet (Inist), Birthe Krog (DTU Library), Karen Morgenroth (CISTI), and Scott Yeadon (ANDS), as well as the many community members who reviewed the Metadata Kernel Version 1.0.

Notes

[1] For an example of the explosive growth of data, see Figure 1 in Southan, Graham (2009), p. 118.

[2] See the information Scott Dineen gave about the "Interactive Science Publishing (ISP) Initiative" in NISO, NFAIS 2010, p. 5.

[3] There is a discussion underway within the DataCite organization regarding the complementary use of globally unique identifiers other than DOIs. Considerable interest was expressed on this topic by community members who participated in evaluating the Metadata Scheme.

[4] Consider this interview from Jon Udell (2007) with Tony Hammond (from Nature Magazine) about DOIs. Speaking about supplementary materials, Hammond says, "We've identified something like 25 million singletons out there and we need to set up some kind of dating service." (start at 21:20 mins.)

[5] Helliwell, McMahon (2010), p. 33.

[6] Green (2009), p.7.

References

Brase, Jan (2004): "Using digital library techniques — Registration of scientific primary data", in "Research and advanced technology

for digital libraries" Springer LNCS 3232. doi:10.1007/978-3-540-30230-8_44.

Green, Toby (2009): "We Need Publishing Standards for Datasets and Data Tables", OECD Publishing White Paper, OECD Publishing. doi:10.1787/603233448430.

Helliwell, John R., McMahon, Brian (2010): The record of experimental science. Archiving data with literature. In: Information Service & Use, Vol. 30 No. 1-2, pp. 31-37. doi:10.3233/ISU-2010-0609.

National Information Standards Organization (NISO), National Federation of Advanced Information Services (NFAIS) (2010): Roundtable on Best Practices for Supplemental Journal Article Materials. http://www.niso.org/apps/group_public/document.php?document_id=3708&wg_abbrev=ccm.

Southan, Christopher, Graham, Cameron (2009). Beyond the Tsunami: Developing the Infrastructure to Deal with Life Science Data. In: Tony Hey et al. (ed.): The Fourth Paradigm. Data-Intensive scientific Discovery. Microsoft Research, Washington, p. 117-123. http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx.

Udell, John (2007): Interview with Tony Hammond (from Nature Magazine) about DOIs. http://jonudell.net/podcast/ju_hammond.mp3.

About the Authors

Joan Starr is chair of the DataCite Metadata Working Group. At California Digital Library, she manages EZID, a new service that makes it easy to create and manage unique, persistent identifiers, including DataCite DOIs and more. Joan is also Manager of Strategic and Project Planning at CDL. She works closely with the Executive Director and CDL senior managers on strategic planning for CDL as a whole and for Program-specific planning, as appropriate. In addition, she provides oversight for the University of California Systemwide Library Planning function. Follow her at http://twitter.com/joan_starr.

Angela Gastl is member of different project teams at the library of ETH Zurich, Switzerland. In the last two years, the main focus of her work has been the setup of a DOI Registration Agency for primary and secondary data of Swiss universities. She recently joined a project-team that aims for the long-term preservation of digital data at ETH Zurich. Previously, having a Master of Arts in history, she worked at the corporate archives of a major Swiss bank (1996-2002) and was the manager of the historical archives at ETH Zurich (2003 to 2005).

Copyright © 2011 Joan Starr and Angela Gastl

PRINTER-FRIENDLY FORMAT Return to Article