Pre-Metadata Counseling: Putting the DataCite relationType Attribute into Action

Elise Dunham, Elizabeth Wickes, Ayla Stein, Colleen Fallaw, and Heidi J. Imker[*]

The library at the University of Illinois at Urbana-Champaign has launched an institutional data repository called the Illinois Data Bank. The Illinois Data Bank provides University of Illinois researchers with a library-based repository for research data that facilitates data sharing and ensures reliable stewardship of published data. Throughout development of the Illinois Data Bank self-deposit form, curation interface, and underlying infrastructure, we used the DataCite Metadata Schema Version 3.1 documentation as a guide for our descriptive metadata[21] so that when we mint a DataCite DOI at the end of the self-deposit process, we can register a robust set of metadata with the DataCite Metadata Store. Based on existing repository services, we anticipate that user discovery of data sets will frequently happen via harvesters and links from published papers and not directly within the Illinois Data Bank.[22] Therefore, we are committed to enhancing external discoverability and access to the data sets that we steward. Utilizing standards for data set description, like the DataCite Metadata Schema V 3.1, fulfills our commitment to following community-developed best practices for data publication and ensures that data sets are represented accurately and consistently within external discovery systems.

Our descriptive metadata schema for data sets defines relationships between data sets and other object types. The goal was to contextualize deposited data sets with related scholarly outputs, such as published papers, code, and other data sets. The functional requirements of the Illinois Data Bank that inspired this project are illustrated in table 6.2.

TABLE 6.2
Metadata-related functional requirements for the Illinois Data Bank.

Category	Functional Requirement of the Illinois Data Bank
Discoverability	Provide opportunity for end users reading a published journal article to find datasets that support conclusions of the article. Provide opportunity for end users viewing a dataset to find articles whose conclusions it supports. Provide opportunity for end users to navigate between related resources according to linked data principles.[a]
Provenance	Contribute to the overall record of a project's scholarly outputs. Track the life of a dataset, including its subsets and versions, over time.
Metrics	Position the Illinois Data Bank to be able to maintain a record of the number of times a dataset in the Illinois Data Bank is formally cited for internal and external reporting. Position the Illinois Data Bank to be able to monitor potential correlation between citation metrics and access metrics.

a. See Christina Bizer, Tom Heath, and Tim Berners-Lee, "Linked Data: The Story So Far," *International Journal on Semantic Web and Information Systems* 5, no. 3 (July–September 2009), http://go.galegroup.com/ps/i.do?id=GALE%7CA209477051&v=2.1&u=uiuc_uc&it=r&p=AONE&sw=w&asid=fd7e52083504e596d9b86bfe8a273f7c.

Prior to implementing the DataCite Metadata Schema for the Illinois Data Bank, we reviewed the schema's `relationType` vocabulary to determine if it met our needs for linking resources. We discovered that our functional requirements and DataCite 3.1's approach were misaligned. For example, one type of relationship that we required was the connection between a data set and the article whose conclusions it supports. The DataCite `relationType` pairs that appeared to be most applicable to representing relationships between articles and data sets were

- `IsCitedBy` and `Cites`
- `IsReferencedBy` and `References`
- `IsSupplementTo` and `IsSupplementedBy`

The definitions of these relationships in DataCite 3.1 are not specific, can overlap, and do not provide guidance on their explicit differences or suggested use. There were many interpretations of these relationship definitions within our team. For example, some interpreted the IsSupplementTo/IsSupplementedBy in a literal sense to mean that a resource must be shared in a "Supplemental Materials" section upon publication of a journal article in order for this relationType to apply. Others interpreted this relationship in a more abstract way to mean that a resource supplements one's understanding of a particular set of research conclusions. Our discussions about the meaning of this relationship occurred in the shadow of a wider scholarly publishing environment that has not developed consensus around definitions related to data and citation, including supplemental materials.[23] Not only did our small team's varying interpretations demonstrate just how difficult it is to navigate metadata semantics, but it also became clear that we needed a way to consistently describe a data set supporting an article regardless of how any particular author or journal decided to represent the relationship.

The troubles that we encountered with interpreting the DataCite relationType vocabulary definitions are indicative of the fact that the research data curation and scholarly communications fields are currently in a state of maturation on the issue of data set citation and linking of scholarly outputs.[24] As a collaborative community of practice, we see a need for solutions for linking scholarly outputs. Currently, the focus tends to be on assigning links after publication. For example, the Research Data Alliance Publishing Data Services Working Group is working to develop a common framework for cross-referencing data sets and articles that have already been published.[25] In time these efforts may feed back into the community's metadata standards and best practices to enable linking resources at the point or points of publication.

In the meantime, we developed more specific definitions of IsSupplementTo, isCitedBy, and the version and aggregation pairs based on our own interpretations (see table 6.3). We determined that the other relationTypes elements available in the DataCite vocabulary whose definitions were unclear to us (e.g., IsReferencedBy/References and IsCompiledBy/Compiles) would not immediately help us to achieve our goals with resource linking and therefore were not used. Through application of our limited set of the relationType terms, we are meeting our functional requirements in the short term and are hopeful that our definitions are flexible enough to respond to anticipated community developments surrounding issues of data set citation and linking. By developing our own clear definitions for relationTypes we are able to serve up our metadata consistently and in a way that fits our local needs, and by deciding to register our relatedIdentifers with DataCite even when our definitions do not match, we are able to represent these crucial relationships outside of our system.

TABLE 6.3

The DataCite relationType definitions used by the Illinois Data Bank.

DataCite relationType	Definition	Illinois Data Bank Usage Note
IsSupplementTo	The resource being described supports the conclusions of the related written work.	Use two instances of <RelatedIdentifier> for the same identifier when the data set whose conclusions support the paper is also formally cited in the paper: <RelatedIdentifier relatedIdentifierType=”DOI” relationType=”IsSupplementTo”> 10.1234/5678</RelatedIdentifier> <RelatedIdentifier relatedIdentifierType=”DOI” relationType=”IsCitedBy”> 10.1234/5678</RelatedIdentifier>
IsCitedBy	The resource being described is formally attributed as a source in the related written work.	Use two instances of <RelatedIdentifier> for the same identifier when the data set whose conclusions support the paper is also formally cited in the paper: <RelatedIdentifier relatedIdentifierType=”DOI” relationType=”IsSupplementTo”> 10.1234/5678</RelatedIdentifier> <RelatedIdentifier relatedIdentifierType=”DOI” relationType=”IsCitedBy”> 10.1234/5678</RelatedIdentifier>
IsNewVersionOf	The resource being described is a new version of the related resource.	
IsPreviousVersionOf	The resource being described is a previous version of the related resource.	

TABLE 6.3 (continued)

DataCite relationType	Definition	Illinois Data Bank Usage Note
IsPartOf	The resource being described is a member of the related aggregation.	
HasPart	The aggregation being described has the related resource as a member.	

We are not fully satisfied with the definitions we have presented here because we believe that they are not yet specific enough to support granular, meaningful expressions of all potential relationships between articles and data sets. Even so, we have determined that it is better to move forward with a well-documented plan that enables us to, at minimum, expose links between objects and count citations than to try to solve the problem without sufficient knowledge of what we will be facing in the Illinois Data Bank. As we implement our relationship definitions for scholarly resources, we will put a system in place that will support curator efforts to capture more detail about the relationships between data sets and articles that come onto our radar in the Illinois Data Bank. A systematic approach for gathering data about resource relationships will enable future research into the many linking challenges we face in the data curation community. We will also look to the data citation community for feedback and input as we collectively move forward to address the challenges of linking digital objects within the scholarly communication landscape.

ACKNOWLEDGMENTS

Our thanks to the DataCite Metadata Working Group for being responsive to our questions and for its continued work on the DataCite Metadata Schema.