Investigating DataCite: Evaluating a Metadata Schema & Its Impact

Jeffrey Kersh

School of Library & Information Science

LI 804: Organization of Information

Dr. Widdersheim

May 9, 2021

The growth of digital information resources accessed on the web has created a need for systems of organization to make sense of that information and to make it searchable and therefore useful to web users. The metadata, or data about data, can be used to index digital files and make them rapidly searchable. However, not all data is the same or intended for the same audience or use. This creates the need for different metadata schemas to provide a structure for the compilation of metadata that improves the findability and usability of records of a particular type. The DataCite Metadata Schema is one instance of a metadata schema, which is used to create metadata records for scientific data from universities, research institutions, and government agencies (DataCite, 2021). The different elements of this schema all work towards creating metadata that can make research data a citable source of information. The elements of the DataCite Metadata Schema not only make research datasets searchable, but they also ensure better authorship records.

The development of the DataCite Metadata Schema began in the early 2000's, primarily in Germany, with the idea that scientific data should be citable outside of research articles. The idea was that it would allow credit to be given to data authors and it would improve the verification of results which has always been of great importance in the natural sciences. With the growth and development of the internet and web, the focus was placed on finding and citing digital datasets and other digital research objects (Brase & Sens, 2015). Eventually digital object identifiers (DOI's) were chosen as the persistent identifiers to be assigned to research objects since they would create a stable method of finding and citing the objects. It also made sense since most scientists were already familiar with DOI's at the time (Brase & S). The work with DOI's and citation of digital research objects continued throughout the early 2000's and in 2009 the non-profit DataCite Consortium was founded by seven members from various scientific research institutions. DataCite established three fundamental goals: easy access to research data online, acceptance of research data as a legitimate, citable source, and the archiving of research data (DataCite Metadata Working Group, 2021). This required more than just registering DOI's to particular datasets. The result was the development of the DataCite Metadata Schema, which was to provide a structure that made digital scientific research data more searchable.

There are a wide variety of scientific disciplines, and the DataCite Consortium wanted a citation system and metadata schema that could be applied to all of them. From the beginning, the schema was intended to describe digital scientific data, but due to the scope of scientific material it was intended to be capable of describing any kind of digital object - images, numerical data, charts, graphs, etc. DataCite also wanted to be able to create metadata records for digital representations of non-digital research objects so that the greatest amount of scientific research could become accessible online (DataCite Metadata Working Group, 2015).

The authoritative source of information on the Datacite Metadata Schema is located on DataCite's website under the Resources page. The page contains a PDF with documentation for the schema, a change log describing the different updates, as well as examples and a brief history of the DataCite Consortium and its values and mission (DataCite Metadata Working Group, 2021). There is also an XML file containing the schema itself in XML format to illustrate the technical framework. The schema documentation also covers the current contributors and authors of the schema, as well as the properties and sub-properties required for individual records. It also addresses the schema's focus on interoperability with other metadata schemas. Another excellent source of information is the DataCite Metadata Schema google discussion group, which allows users to reach out with questions and concerns about the schema, as well as DataCite's Best Practice Guide on GitHub (MSpenger, 2019). There are also several useful articles that cover the

DataCite Metadata Schema. Starr and Gastl cover the metadata properties of the schema in an easy to understand way, and Adamich examines both the properties and also the partnerships DataCite maintains with other organizations that make the schema highly interoperable (2011)(2016).

To understand the DataCite Metadata Schema, it is helpful to look at its metadata properties. These properties, which are essentially a list of the information that is ideally included in a metadata record, are broken up into three distinct levels. These consist of Mandatory (M), Recommended (R), and Optional (O) metadata properties (DataCite Metadata Working Group, 2021). The first level, Mandatory, consists of five properties that are required for any DataCite metadata submission (Table 1). The reason these required properties are so few is to allow the broadest possible use of the schema, both in scientific subject and data object type. However, DataCite very strongly suggests that submissions also include the Recommended properties since it greatly strengthens the ability of research objects to be found and used (DataCite Metadata Working Group, 2021). The Recommended level consists of six properties that help describe the data and distinguish it from and relate it to other datasets Table 1). Finally, the Optional level consists of an additional eight properties that provide technical information about the submission and information such as rights and funding sources (Table 1).

Table 1

DataCite Metadata Schema Properties

ID	***Property***	***Obligation***
1	Identifier (with mandatory type sub-property)	**M**
2	Creator (with optional given name, family name, name identifier and affiliation sub-properties)	**M**

3	Title (with optional type sub-properties)	**M**
4	Publisher	**M**
5	PublicationYear	**M**
6	Subject (with scheme sub-property)	**R**
7	Contributor (with optional given name, family name, name identifier, and affiliation sub-properties)	**R**
8	Date (with type sub-property)	**R**
9	Language	**O**
10	ResourceType (with mandatory general type description subproperty)	**M**
11	AlternateIdentifier (with type sub-property)	**O**
12	RelatedIdentifier (with type and relation type sub-properties)	**R**
13	Size	**O**
14	Format	**O**
15	Version	**O**
16	Rights	**O**
17	Description (with type sub-property	**R**
18	GeoLocation (with point, box, place, and polygon sub-properties)	**R**
19	FundingReference (with name, identifier, and award related subproperties)	**O**
20	RelatedItem (with identifier, creator, title, publication year, volume, issue, number, page, publisher, edition, and contributor sub-properties)	**O**

Note. This table was created from information contained in "DataCite metadata schema documentation for the publication and citation of research data and other research output." M represents a mandatory property, R represents recommended, and O represents optional.

This three level structure was chosen for this metadata schema for several reasons. First of all, it makes the schema far more approachable for those researchers who are not experts in producing metadata records or are too busy to spare much time to the process. Since there are only five required properties, more researchers and organizations are likely to take the short amount of time to provide information to make their research more useful. The other reason for splitting the schema properties into three categories is to make the schema highly adaptable to a variety of data types and subject matters. The presence of too many required properties would limit submissions to those that had all of the necessary properties, even though there might be quality research that simply lacked one or two pieces of information.

Broadly speaking there are six metadata types: descriptive, technical, preservation, rights, structural, and markup languages (Zeng & Qin, 2008). This schema is an XML-based schema, which is primarily descriptive. The metadata properties 2-9, 15, and 17-19 can be classified as descriptive metadata since they help users find and understand the research object by providing publication and authorship information. These encapsulate the majority of properties and make it clear that the metadata's primary function is the location and understanding of the objects, which is consistent with DataCite's mission. The schema also includes administrative metadata, consisting of technical, preservation, and rights metadata. These are composed of properties 13-15 and 10, 1 and 11, and 16, respectively (Table 1). Of these, the only required properties are those dealing with the identification (DOI) and resource type. This is also consistent with DataCite's mission to make datasets citable and to accommodate a wide variety of digital object types (DataCite Metadata Working Group, 2021). Finally, metadata properties 12 and 20 are structural metadata, only one of which is required (Table 1). This property is included so that the research datasets include meaningful relationships that increase findability and useability.

Many of the metadata properties such as creator, title, date, etc. help with retrieval by creating a list of details about the object similar to a bibliographic record. The identifier property, like a call number, makes each record unique and therefore distinguishable to users, preventing record ambiguity. Due to the digital nature of these records, record type and other technical information is encouraged. This can help users filter by what type of resource they are looking for. As a whole, the record provides a description and context for the scientific research data that can let users limit searches to relevant information and ideally understand what purpose that data has and how it relates to any other research.

The metadata schema's characteristics can best be described as flexible and interoperable. As has already been mentioned, the schema can support a variety of object types and disciplines within the field of scientific research. The schema itself has dedicated resources for interoperability, but the interoperability of specific objects is also affected by how many of the Recommended properties are provided by the contributor for a particular metadata record. DataCite supports openness and extensibility by working with the Dublin Core Metadata Initiative (DCMI) Science and Metadata Community (SAM) through a DataCite to Dublin Core crosswalk (DataCite Metadata Working Group, 2021). The interoperability also extends to citations, with a partnership between DataCite and ORCID. The ORCID number is of great importance for addressing the issue of name ambiguity and making the work of a particular researcher easy to find. DataCite also maintains a partnership with the Consortia Advancing Standards in Research Administration Information (CASRAI). CASRAI's extensible dictionary of terms was created to allow greater interoperability between research organizations by providing a list of common terms with both their definitions and a general structure of the terminology. In essence, the CASRAI dictionary is a controlled vocabulary for research institutions. Their partnership with DataCite was in no small part caused by an existing relationship between ORCID and CASRAI (Adamich, 2016). The result is that ORCID can now exchange metadata with DataCite since they both use CASRAI's controlled vocabulary for the Type/ResourceTypeGeneral properties.

The following metadata record in Figure 1 was generated using DHVLAB's DataCite Metadata Generator (n.d.). The generator contains all the property and sub-property fields covered earlier. As the user enters the relevant information, an XML file is generated with the proper formatting for a DataCite record. While it is not much more complicated to manually write the XML file, the generator makes the process smoother and saves time. For each property in the generator, there is a link to DataCite's Best Practice Guide on GitHub, which contains information and explanations for each property as well as XML examples for each (MSpenger, 2019). Overall, the generator makes it very easy to generate the record and also to verify that the record is consistent with DataCite's guidelines. This is consistent with DataCite's goal of encouraging researchers to provide open access to their research data by making the process easily accessible and straightforward. The example shown below is based on undergraduate thesis work, but publication information has been added to demonstrate how the properties look within a record. This record demonstrates what a completed DataCite Metadata Schema record looks like in its XML format. All of the mandatory properties have been described and several of the recommended and optional properties as well.

Figure 1

DataCite Metadata Schema Record

```
<?xml version="1.0" encoding="UTF-8"?>
```

```xml
<resource xmlns=https://schema.datacite.org/meta/kernel-4.3/
xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance xsi:schemaLocation="https://schema.datacite.org/meta/kernel-4.3/ https://schema.datacite.org/meta/kernel-4.3/metadata.xsd">
<identifier identifierType="DOI">10.1000/xyz123</identifier>
<titles>
        <title xml:lang="En">Correlating Enzymatic Activity of BcHMGR with its Oligomeric States</title>
</titles>
<creators>
        <creator>
                <creatorName nameType="Personal">Jeffrey J. Kersh</creatorName>
                <givenName>Jeffrey</givenName>
                <familyName>Kersh</familyName>
                <nameIdentifier nameIdentifierScheme="ORCID" schemeURI="http://orcid.org/">0000-0001-1998-2021</nameIdentifier>
                <affiliation xml:lang="En">Gonzaga University</affiliation>
        </creator>
</creators>
<publisher xml:lang="En">ChEMBL</publisher>
<publicationYear>2020</publicationYear>
<resourceType resourceTypeGeneral="Dataset">Dataset</resourceType>
<subjects>
        <subject subjectScheme="DDC" schemeURI="http://dewey.info/" xml:lang="En">547 Organic Chemistry</subject>
</subjects>
<contributors>
        <contributor contributorType="ProjectLeader">
                <contributorName>Jeff Watson</contributorName>
                <givenName>Jeffery</givenName>
                <familyName>Watson</familyName>
                <nameIdentifier nameIdentifierScheme="ORCID" schemeURI="http://orcid.org/">0000-0002-6939-0137</nameIdentifier>
                <affiliation xml:lang="En">Gonzaga University</affiliation>
        </contributor>
</contributors>
<dates>
        <date dateType="Submitted">05/09/2020</date>
</dates>
```

```
35. <descriptions>
36.                 <description xml:lang="En" descriptionType="Abstract">Burkholderia cenocepacia is an opportunistic lung pathogen that can be found in soil and water and is one of the highest causes of morbidity and mortality in cystic fibrosis patients. In immunocompromised individuals B. cenocepacia can colonize and is especially troublesome to treat since it is naturally resistant to a wide range of antibiotics. One interesting aspect of B. cenocepacia is its unique use of 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGR), which is crucial for isoprenoid biosynthesis in mammals, but serves a different function here. The unorthodox use of HMGR by this pathogen makes the enzyme a good potential drug target. As a morpheein, BcHMGR can exist in at least three different oligomeric states and interconvert between them by disassembling into subunits, which undergo conformational changes, and then reassemble into one of the oligomeric states, depending on the changes to the subunits. One factor that can affect the oligomeric state of the enzyme is the pH of its surroundings. In order to better understand BcHMGR’s morpheein behavior and how it can be exploited to develop treatments for B. cenocepacia infections in immunocompromised individuals, the enzymatic activity was observed over a pH range. The effect of pH on the enzymatic activity of the enzyme was evaluated and supported previous research that proposed the level of enzymatic activity was dependent on the oligomeric state of the enzyme. The results of this study also highlight the need for further characterization of the distribution of BcHMGR’s oligomeric states as a function of pH.</description>
37. </descriptions>
38. <language>En</language>
39. </resource>
```

Note. the line numbers were added for ease of reference and were not generated as part of the record. In addition, the generator has not been updated for Version 4.4, so the RelatedItem property had not been added. However, there was no related item to include for this example so it was left out.

The DataCite Metadata Schema uses a three tiered system of metadata properties: mandatory, recommended, and optional. The majority of these properties are descriptive, but there are also some administrative and technical metadata properties. The structure of the schema is highly flexible, allowing researchers to describe a wide variety of digital research objects from all scientific disciplines. It is also highly interoperable due to DataCite’s work with ORCID, CASRAI, and Dublin Core to create a schema that can be translated to other existing schemas. The DataCite Metadata Schema also places a strong emphasis on authorship through its work with ORCID to prevent the loss of search results due to name ambiguity. As the DataCite Consortium developed its metadata schema, it focused on finding ways to make scientific research data more easily searchable and citable with the intent of helping the scientific community improve verification of results. Digital research objects have grown immensely over the past twenty years, and much of this research is not published in traditional journal articles despite its validity. The DataCite Metadata Schema has provided a strong foundation to make this research a useful source of information for current scientific work.

References

Adamich, T., (2016). Accessing scholarly research datasets using DataCite, ORCID, and CASRAI. *Technicalities, 36*(3), 18-21.

Brase, J. & Sens, I. (2015). The tenth Anniversary of assigning DOI names to scientific data and a five year history of DataCite. *D-Lib magazine, 21*(1/2). https://doi.org/10.1045/january2015-brase

DataCite metadata working group. (2021). *DataCite metadata schema documentation for the publication and citation of research data and other research outputs*. DataCite metadata schema 4.4. https://doi.org/10.14454/3W3Z-SA82

DataCite. (2021). *Members*. DataCite. https://datacite.org/members.html

Digital teaching and research infrastructure for the humanities. (n.d.). *DataCite metadata generator - kernel 4.3*. https://dhvlab.gwi.uni-muenchen.de/datacite-generator/

MSpenger. (2019, May 11). DataCite Best Practice Guide. GitHub. https://github.com/UB-LMU/DataCite_BestPracticeGuide

Starr, J., & Gastl, A. (2011). isCitedBy: A metadata scheme for DataCite. *D-Lib magazine, 17*(1/2). https://doi-org.emporiastate.idm.oclc.org/10.1045/january2011-starr

Zeng, M. L., & Qin, J. (2008). *Metadata*. Neal-Schuman Publishers.