DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2014

Metadata for Research Data: Current Practices and Trends

Sharon Farnel
University of Alberta,
Canada
sharon.farnel@ualberta.ca

Ali Shiri
University of Alberta,
Canada
ali.shiri@ualberta.ca

Abstract

This paper reports a study that examined the metadata standards and formats used by a select number of research data services, namely Datacite, Dataverse Network, Dryad, and FigShare. These services make use of a broad range of metadata practices and elements. The specific objective of the study was to investigate the number and nature of metadata elements, metadata elements specific to research data, compliance with interoperability and preservation standards, the use of controlled vocabularies for subject description and access and the extent of support for unique identifiers as well as the common and different metadata elements across these services. The study found that there was a variety of metadata elements used by the research data services and that the use of controlled vocabularies was common across the services. It was found that preservation and unique identifiers are central components of the studied services. An interesting observation was the extent of research data specific metadata elements, with Dryad making use of a wider range of metadata elements specific to research data than other services.

Keywords: metadata; research data; research data services; standards

1. Data Repositories

"And yet, data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself" (Gold 2007). Although the nature of research data can vary widely depending on the discipline, its importance to the replication, refutation or validation of the findings or observations of a research project has never been in doubt.

Research data has recently been viewed as being part of a larger data landscape, namely big data. A number of researchers have referred to research data, linked data, the web of data and open data as constituting elements of the big data landscape (Hudson, 2012; Shiri, 2013). The *Report of the 2011 Canadian Research Data Summit* (Research Data Strategy Working Group, 2011) provides a specific categorization of digital data, namely research data, produced by academia, industry and government.

The sharing of research data has long been a practice among many research communities, often through informal means made increasingly easy with the advent of the internet and associated tools such as email, ftp sites, etc. Borgman (2007) provides four rationales for the sharing research data, namely "to (a) reproduce or verify research, (b) make results of publicly funded research available to the public, (c) enable others to ask new questions of extant data, and (d) advance the state of research and innovation". She also notes that common metadata formats, ontologies and data structures will support the integration of multiple data sources and services.

The rise of the open data[1] and open science data[2] movements, in conjunction with the increasing implementation of data management and sharing policies by funding bodies[3],

[1] http://en.wikipedia.org/wiki/Open_data

[2] http://en.wikipedia.org/wiki/Open_science_data

[3] http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

https://doi.org/10.23106/dcmi.952136534

governments[4] and journals[5], has led to an explosion in the number of research data services created to serve institutions, association members, and research communities. Databib[6] and re3data.org[7] maintain listings of research data services, and as of August 2014 combined list nearly one thousand. Many services enable the deposit of research data and associated metadata, while others focus on metadata describing research data that is housed in other repositories.

This proliferation of services offering a range of functionalities and designed to serve different communities with different needs poses many challenges to researchers, librarians and others within the research community working to create an interoperable research environment. Documenting the range of functionalities as well as defining means of comparing one service to another have been recognized as important activities and have begun to be addressed by Databib[8] and Dryad[9] respectively. Key to any overall comparison or evaluation is an understanding of the metadata practices within services.

2. Metadata in Data Repositories

Metadata is structured information that provides context for information objects of all kinds, including research data, and in doing so enables the use, preservation, and reuse of those objects. The importance of quality, standards based metadata has long been understood by those in the fields of librarianship and research data management; NISO's six principles of good metadata (NISO 2007) being an excellent and oft-cited expression of that understanding. The same, however, has not always been the case among research communities. A recent study (Tenopir et al., 2011) found that there is a "lack of awareness about the importance of metadata among the scientific community - at least in practice" and recommended that institutions and individuals within them who work with researchers can and should do more to help researchers prepare the metadata necessary to enable the discovery, preservation, and reuse of their data. In a scoping study, Ball (2009) explored the feasibility and desirability of a harmonized application profile to improve resource discovery and reuse of scientific and research data in the repository landscape. The two key findings of his study were that a) a comparison of data models and metadata schemes from a variety of disciplines suggested that a carefully generalized metadata profile could be constructed that is both widely applicable and yet still fulfils the requirements of the use cases and b) while the comparison of several different data models shows sufficient common ground for a relatively detailed data model on which to base a Scientific Data Application Profile, from an implementation perspective a simple model is preferred.

One of the main arguments for the identification and documentation of metadata practices and formats for research data services is to create a solid basis upon which subject and semantic interoperability can be ensured. Identifying useful metadata elements and practices will support various interoperability models reported in the literature (Nicholson and Shiri, 2003; Hafezi, et al., 2010). The same arguments that were made in the first generation of digital libraries, open archives and content management systems hold true for research data services - the variety of disciplines involved and the vastness of research data call for a more systematic and holistic approach to metadata. In their 2012 study, Willis et al. identified 11 fundamental metadata goals for metadata documenting research data and highlighted the need for further metadata-related research. An evidence-based approach to the study of emerging research data management systems allows us not only to study emerging trends but also to develop a basis for formulating

[4] http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[5] http://www.plosone.org/static/policies#sharing

[6] http://databib.org/index.php

[7] http://www.re3data.org/

[8] http://goo.gl/mQvy0F

[9] http://www.dcc.ac.uk/webfm_send/750

https://doi.org/10.23106/dcmi.952136534

best practices and policies for research data management. This study aims to take a step towards that goal.

3. Purpose

Given the confluence of increased requirements around data management and sharing with greater demand by researchers for services around metadata standards and applications, an examination and comparison of the metadata standards and practices of research data services would be both timely and beneficial. Given the emerging nature of research data repositories and the urgent need for evidence-based practices, it is important to study examples of the repositories that have been experimenting with how best to organize and manage research data. This is not only useful for the metadata community in conceptualizing metadata standards in a new and emerging context, it is particularly important for planners and practitioners who aim to embark on research data repository projects. The objective of this study is to examine the metadata standards and formats used by a select number of research data services to address several specific research questions. These research questions are concerned with both theoretical as well as practical aspects of organizing, managing and providing access to research data.

1. What is the number and nature of metadata elements available?
2. What research data specific metadata do these services provide in addition to common metadata elements?
3. To what extent do the research data management services adhere to widely recognized interoperability and preservation metadata standards?
4. Which research data repositories benefit from and promote controlled vocabularies for subject description and access?
5. How many of the services provide support for unique identifiers (e.g., DOIs)?
6. What kind of metadata assistance (documentation, etc.) is provided?
7. What metadata elements are common and different across these services?

4. Methodology and Analysis

The nature of this study is exploratory in the sense that it aims to gain an insight into the current metadata practices and trends in four research data services: Datacite,[10] Dataverse Network,[11] Dryad,[12] and FigShare.[13] The rationale for the selection of these services lies in the fact that these are widely popular and internationally used research data services that cover multiple disciplines. A significant number of research-intensive and academic institutions are already using these services and some are considering them in their research data management planning.

Table 1 provides an overview of the geographic distribution of these research data services, their subject areas as well as their main services.

TABLE 1: Research data services

[10] http://www.datacite.org

[11] http://thedata.org/

[12] http://datadryad.org/

[13] http://figshare.com/

https://doi.org/10.23106/dcmi.952136534

Service	Subject area	Main services	Location
Datacite	General	Metadata, DOI	UK
Dataverse Network	General	Cite, analyze, preserve	US
Dryad	General	data underlying scholarly publications discoverable, accessible, understandable, freely reusable, and citable	US
FigShare	General	figures, datasets, media, papers, posters, presentations and filesets, altmetrics	UK

The seven research questions above, which are informed by the NISO principles for good metadata (NISO 2007), provide the analytical framework for examination of research data services focusing on various aspects of metadata elements, formats, and standards. As was stated earlier, an evidence-based approach for this study was thought particularly useful, partly because of the emerging nature of research data management systems and partly because of the variety of disciplines and domains that current research data management services cover. To address the research questions, existing metadata records, metadata creation interfaces, and associated documentation will be examined. The following comparative table addresses the key research questions.

5. Findings

Table 2 provides an overview of our sample set of research data services with respect to research questions 1 through 6.

TABLE 2: Research data services comparison (research questions 1-6)

	Datacite	**Dataverse Network**	**Dryad**	**Figshare**
Number of metadata elements	41	100	52	12
Research specific metadata elements	No	Yes	Yes	No
Compliance with standards	Datacite Metadata Schema, which is an application profile of Dublin Core (DC), OAI	Data Documentation Initiative (DDI) Codebook, compliant with Dublin Core (DC) and Content Standard for Digital Geospatial Metadata (CSDGM), MARC LOCKSS, OAI	Dublin Core, Darwin Core, Bibliographic Ontology, METS/MODS OAI/DC OAI/ORE (Object Reuse and Exchange) RDF/DC CLOCKSS For now, OAI/DC is the recommended format.	CLOCKSS
Use of controlled	Includes controlled vocabularies for	Supports use of controlled vocabularies	Supports use of ontologies and	No formal controlled

https://doi.org/10.23106/dcmi.952136534

vocabularies	some elements, supports use of controlled vocabularies for other elements; MESH, OBI, NCBI		controlled vocabularies such as Open Biomedical Ontologies & Gene Ontology. A trial version of HIVE is provided to support subject description. LCSH, TGN, MESH, Integrated Taxonomic Information Systems (ITIS), National Biological Information Infrastructure Biocomplexity Thesaurus, LC Name Authorities file	vocabularies; only 14 high level categories
Support for DOI	Yes	Yes	Yes	Yes
Metadata assistance	full documentation of metadata schema, user guidelines, full api documentation	metadata documentation available via user guide, contextual help available for each element in metadata entry form	Dryad Wiki pages provide detailed documentation including Cataloguing guidelines	Partner with DataCite

In terms of metadata elements, the services range in number from 12 to 100. Of course, the number of elements is not a measure of success or performance of a system. The number of metadata elements may be dependent on a wide range of factors, including the simple or sophisticated approaches that the research data repositories adopt, the disciplines and domains that they cover as well as the applicability of the elements in terms of metadata creation and maintenance. The proportion of general metadata elements in comparison to research data specific elements ranges quite dramatically; Datacite has no research data specific metadata elements while Dryad has 35 (of 52 total). Dataverse and Dryad provide a more sophisticated set of metadata elements and standards. Figshare takes a minimalist approach and provides a very basic set of metadata elements to facilitate quick and easy deposit of research data.

Preservation appears to be one of the central components of research data services to ensure long term access to data. Most have adopted preservation strategies associated with LOCKSS[14] (Lots of Copies Keep Stuff Safe) and CLOCKSS[15] (Controlled LOCKSS) as widely used and common information and data preservation approaches. Given the importance of interoperability in research data management services, DataCite, Dataverse Network and Dryad support OAI-PMH[16] (Open Archives Initiative/ Protocol for Metadata Harvesting) to ensure the wider findability and discoverability of research data

Initial comparison of several of the sample research data services demonstrates that a variety of metadata standards are in use, although Dublin Core is used or supported across most of the services. Support for controlled vocabularies is common, although few incorporate them by default into their schema. For instance, while Dryad and DataCite adopt a more systematic

[14] http://www.lockss.org/

[15] http://www.clockss.org/clockss/Home

[16] http://www.openarchives.org/pmh/

https://doi.org/10.23106/dcmi.952136534

approach to the use of various controlled vocabularies for subject description and access, recommending various thesauri and knowledge organization systems, Figshare does not provide any specific provision for this feature; the only subject access mechanism in Figshare is the high level subject categories that appear when users click on the 'browse' option on the homepage.

An encouraging sign is the common support for DOIs which are seen as key to discovery, preservation and citation of research data. All of the services appear to have metadata documentation available to aid users.

Table 3 provides a detailed account of the common and unique metadata elements used by the four research data repository services.

TABLE 3: Research data services comparison (research question 7)[17]

	Datacite	**Dataverse Network**	**Dryad**	**Figshare**
Titles	title	- title - subtitle - document title	- article title - journal title - data package title	title
Creators, Contributors	- creator - contributor - publisher	- author - producer - funding agency - distributor - depositor - contact - data collector	- author - creator	- author - collaborators
Topical subject(s)	subject	- keyword - topic classification	- keyword - scientific name	- categories - tags
General description	description	abstract	- article abstract - description	description
Object type(s)	resource type	kind of data	type	type
Date(s)	- date - publication year	- production date - distribution date - deposit date - version date - date of collection-start - date of collection-end	- date of issuance - deposit date - date available - embargo date	- date created - date published
Rights, Access, Use	rights	- data access place - original archive - availability status - confidentiality declaration - special permissions - restrictions - conditions - provenance - document holdings - disclaimer	- rights statement - location of related content outside of Dryad	license
Object technical characteristics	- size - format	- software - software version - size of collection - study completion	- file format - file size - provenance	file size

[17] Note that table 3 does not reference attributes or attribute values and is not meant to be an element by element mapping

https://doi.org/10.23106/dcmi.952136534

Spatial subject(s)	- geo location	- country/nation - geographic coverage - geographic unit - geographic bounding box	- spatial coverage	
Identifiers	- identifier - alternate identifier - related identifier	- study global ID - other ID	- article identifier - associated Dryad data package identifier - data package identifier - identifier for related data in Dryad partner repository - associated Dryad publication record identifier - associated Dryad data file record identifier - data file identifier - issn - electronic issn	
Temporal subject(s)		- time period covered-start - time period covered-end	- temporal coverage	
Citation		- citation requirements - depositor requirements	- journal volume number - journal issue - article start page - article end page - article pages	
Versioning	version	version		
Methodology		- unit of analysis - universe - time method - frequency - sampling procedure - major deviations for sample design - collection mode - type of research instrument - data sources - origin of sources - characteristics of sources noted - documentation and access to sources - characteristics of data collection situation - actions to minimize losses - control operations - weighting - cleaning operations - study level error nores - response rate - estimates of sampling errors - other forms of data appraisal		

https://doi.org/10.23106/dcmi.952136534

Related resources		- series - series information - replication for - related publications - related material - related studies - other references		
Language(s)	language			
Status			- status - article publication status	
Production		- production place		
Additional grant information		- grant number - grant number agency		
Note(s)		notes		

Dryad, Dataverse and DataCite make use of Dublin Core as well as other metadata schemes and standards. It is not surprising to note that there are common metadata elements across these services. Dryad also utilizes Darwin Core, Bibliographic Ontology and its own repository specific elements. While Figshare makes limited use of metadata elements, at least seven out of eleven metadata elements are consistent with Dublin Core. Therefore, one can argue that there is a set of elements across these four services that allow for basic interoperability if a meta-service were to be created for cross-searching and cross-browsing

One of the key questions this study aimed to address was the inclusion or creation of metadata elements specifically for research data. Our comparative analysis of the above research data services shows that there are research data specific metadata elements being used. Dataverse Network and Dryad incorporate metadata elements in this area. For instance, Dataverse makes use of such metadata elements as *date of data collection, data collectors, depositor, deposit date, data specific file types such as raw data, processed data*. Dryad offers a number of metadata elements related to the data package and data files deposited into Dryad. Examples of these elements include: *Associated Dryad Data Package Identifier, Data Package Title, Data Package Identifier, Associated Dryad Data File Record Identifier, Data File Identifier, Deposit Date.*

6. Conclusions and Future Work

This study compared four different research data services in terms of metadata and research data management practices. The results of this study will improve understanding among researchers, librarians and research data managers of the application of metadata in research data services. These preliminary findings contribute to the development of a set of guidelines and best practices for developing and implementing metadata for research data services in order to pave the way for the development of an interoperable research data environment. Furthermore, the identification of metadata elements and formats in commonly used research data services will contribute to the creation of an interoperable research data environment. Future work will include expanding this analysis to additional research data services, both general and domain-focused, as well as comparing in detail the metadata elements common across and unique among the services. The development of a framework that takes into account such important components as preservation infrastructures, unique identifiers, interoperability architecture and the definition of a set of research data specific metadata should guide further research and development in this area.

https://doi.org/10.23106/dcmi.952136534

References

Alipour-Hafezi, Mehdi, Abbas Horri, Ali Shiri, and Amir Ghaebi. (2010). Interoperability Models in Digital Libraries: An Overview. The Electronic Library, 28(3), 438-452.

Ball, A. (2009). Scientific data application profile scoping study report. *June 3rd*. Retrieved August 5, 2014, from http://alexball.me.uk/docs/ball2009sda/.

Borgman, Christine L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059-1078.

Gold, Anna. (2007, September/October). Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians. D-Lib Magazine. 13/9/10. Retrieved, August 5, 2014 from http://www.dlib.org/dlib/september07/gold/09gold-pt1.html.

Hodson, Simon. (2012). JISC and Big Data. Eduserv Symposium 2012: Big Data, Big Deal? May 10, 2012, London, UK.

Nicholson, Dennis and Ali Shiri. (2003). Interoperability in Subject Searching and Browsing. OCLC Systems & Services, 19(2), 58 - 61.

NISO. (2007). A Framework of Guidance for Building Good Digital Collections: Metadata. Retrieved, August 5, 2014, from http://www.niso.org/publications/rp/framework3.pdf.

Research Data Strategy Working Group. (2011). Mapping the Data Landscape: Report of the 2011 Canadian Research Data Summit. Retrieved, August 5, 2014, from https://web.archive.org/web/20140312192321/http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/obj/doc/2011_data_summit-sommet_donnees/Data_Summit_Report.pdf.

Shiri, Ali. (2013). Linked Data Meets Big Data: A Knowledge Organization Systems Perspective. Advances in Classification Research Online, North America, 24(1). Retrieved, August 5, 2014, from http://journals.lib.washington.edu/index.php/acro/article/view/14672.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE, 6(6). Retrieved, August 5, 2014, from http://dx.plos.org/10.1371/journal.pone.0021101.

Wiley, Christie. (2014), Metadata Use in Research Data Management. Bulletin of the Association for Information Science and Technology, 40(6). Retrieved, August 5, 2014 from http://www.asis.org/Bulletin/Aug-14/AugSep14_Wiley.html.

Willis, Craig, Jane Greenberg, and Hollie White. (2012). Analysis and synthesis of metadata goals for scientific data. Journal of the American Society for Information Science and Technology, 63(8), 1505 - 1520.

https://doi.org/10.23106/dcmi.952136534