An analysis of characteristics and structures embedded in data papers: a preliminary study

Ya-Ning Chen, arthur9861@gmail.com
Department of Information and Library Science, Tamkang University, Taiwan

Libellarium, IX, 2 (2016): 145 – 156.
UDK: 025.4.036:65.012.4=111
DOI: http://dx.doi.org/10.15291/libellarium.v9i2.266
Research paper

Abstract

Research data or datasets can be regarded as a catalyst to inspire new research by repurposing or combining existing research data, and grant applicants have been requested by funding institutions to include a data management plan as part of research project proposals. In addition to the metadata approach, data papers may mirror the scientific publication model as an alternative means of description and management of research data. However, there is not a common standard for all data papers across various communities. This study aimed to build up a common structural framework to investigate the embedded characteristics and structures of the content of data papers by using a content analysis approach, and 26 data journals from 16 publishers were selected as subject in this study. This study has proposed a common framework and further embodied a concept map (Candela et al. 2015) into more concrete components for the structure of data papers.

KEYWORDS: data papers, research data, datasets, research data management.

Introduction

Traditionally, journal articles and books are the primary conduit of scholarly communication. In the information age, these resources are still important, but research data or datasets are also emerging as another important source of scholarly output. Data have been regarded as a catalyst to inspire new research by repurposing or combining existing research data. Many international institutions have advocated making research data readily available, including the International Council for Science, the Global Biodiversity Information Facility, the UK Research Councils, the US National Science Foundation, and the US National Institutes of Health. These institutions have requested that grant

applicants include a data management plan as part of research projects to manage data for future reuse. In order to achieve reuse and share research data, data documentation is a required component of research data management (RDM) for data discovery and curation. In addition to the metadata approach, data papers may mirror the scientific publication model as an alternative means of description and management of research data.

Literature review

In the process of knowledge inquiry, journal articles and books are regarded as "standing on the shoulders of giants" to support exploration and research. In recent years, research has been enhanced by extensive use of data to test and examine hypotheses. Raw research data or datasets have become an essential part of data-driven research such as eResearch, eScience and digital humanities. However, sharing research data is not the norm for researchers in science. Thus, data are easily packaged and locked in dark archives. These data often lack adequate documentation for discovery and management, and are in danger of being lost (Chavan and Penev 2011, 2). Furthermore, the cost of recollecting or reproducing data is much more than documenting data, although data documentation is time consuming and costly (Kratz and Strasser 2014, 4). Even more importantly, some data cannot be recollected or reproduced. Therefore, adequate documentation of research data is an essential part of RDM for future sharing and reuse (Atici et al. 2013, 670, Chavan and Penev 2011, 2, Costello 2009, 421, Kansa and Kansa 2013, 4, Kratz and Strasser 2014, 4, Niu and Hedstrom 2008, 4, Rees 2010).

Metadata is a fundamental component of digital libraries, and various existing standards or guidelines have been developed for different communities or data types; however, descriptive metadata may not be suitable for RDM, as "metadata may not be sufficient to enable them [researchers] to use the data" (Costello 2009, 4). The reason may be that metadata "may not provide sufficient documentation of the context in which data were collected" (Borgman et al. 2007, 275), such as a "research methods" description not being included in most metadata standards or guidelines (Chao 2015). This means that most descriptive elements of metadata schemas are not suitable for the description and discovery of research data. Therefore, "data papers" that mirror the scientific publication model (Akers 2013) have been proposed as an alternative solution to metadata for description and discovery of research data.

Data papers are data publications resembling traditional journal articles, only shorter (Candela et al. 2015, 1751, Gray 2015). Data papers can be published to make research data or datasets public (Breure 2014, Callaghan et al. 2012, 112, Chavan and Penev 2011, 3, Gray 2015). It is required that data papers and research data have a digital object identifier (Candela et al. 2015, 1754, Gray 2015). Researchers have proposed various categories for the structure of data

papers including general information such as authors, keywords and abstracts on the title page (Candela et al. 2015, 1754), and specific information such as data collection or production (Akers 2013, Breure 2014, Candela et al. 2015, 1754), data processing (Akers 2013, Rees 2010), data analyzing or analytical methods (Akers 2013, Atici et al. 2013, 670), provenance (Candela et al. 2015, 1754, Rees 2010), context or coverage (e.g., time and place) (Atici et al. 2013, 670, Candela et al. 2015, 1754), background (Breure 2014, Candela et al. 2015, 1754), competing interests (Candela et al. 2015, 1754), license (Candela et al. 2015, 1754), attribution (Candela et al. 2015, 1754), reuse (Candela et al. 2015, 1754), etc. On the other hand, some researchers have proposed the 5Ws (i.e., what, where, why, how and who) as a documentation basis for the description of research data (Challaghan et al. 2012, 12, Kennedy, Ascoli, and De Schutter 2011, 318-319). Although some data journals have also defined structural categories in their templates or guidelines for data papers, there is not a common standard for all data papers across various communities (Candela et al. 2015, 1753-4, Callaghan et al. 2014, Chavan and Penev 2011, 3, Smith 2009, 2). With reference to the previous research outlined above, this study poses the following research questions:

RQ: Is there a common framework for the structure of data papers to describe research data to facilitate discovery, sharing and reuse across various communities? In addition to the proposed framework and its categories, what components and their characteristics are embedded in existing data papers?

Methodology

In this study, we aimed to build up a common structural framework for the content of data papers using a content analysis approach. Data papers are referred to using various terms including database article, data paper, data note, data article, data descriptor, data in brief, data original article, database paper, dataset paper, and genome database (Candela et al. 2015). In addition to the aforementioned variants of data papers, software papers were also included in this study to expand the examination of the characteristics and structures of data papers. In this study, 73 data journals provided by Akers (2014) were extended to 94 journals through query results for "data journal" in Ulrichsweb. One third of the 94 data journals (i.e., 31) were randomly selected for this preliminary study. In order to include diverse disciplinary domains and their characteristics, 31 data journals were reduced to 27. Then The Data Science Journal was excluded because, to date, it has not published any data papers. As a result, 26 data journals from 16 publishers were selected as subject in this study, and the disciplinary domain of research subject covered science, social science, and humanities. The publishers, number of journals, and their disciplinary domains are shown in Table 1.

Table 1. Subjects of this study

Publisher (No. of Journals)	Journal Title (Journal Code)	Discipline
BMC(3)	BMC Medical Education (JRN01)	Medical Sciences
	BMC Research Notes (JRN02)	Biology, Medical Sciences
	BMC Psychiatry (JRN03)	Medical Sciences
Brill (1)	Research Data Journal of the Humanities and Social Sciences (JRN04)	Social Sciences
Copernicus (2)	Earth Systems Science Data (JRN05)	Earth Sciences
	Geoscientific model development (JRN06)	Earth Sciences
Earthquake Engineering Research Institute (1)	Earthquake Spectra (JRN07)	Earth Sciences
Ecological Society of America (1)	Ecological Archives (JRN08)	Ecology
Elsevier (2)	Data in Brief (JRN09)	Computers
	Genomics Data (JRN10)	Biology
Faculty of 1000 (1)	F1000Research (JRN11)	Biology, Medical Sciences
Hindawi (1)	Dataset Papers in Science (JRN12)	Medical Sciences
Nature (1)	Scientific Data (JRN13)	Sciences
Pensoft (1)	Biodiversity Data Journal (JRN14)	Biology
Procon (1)	Biomedical Data Journal (JRN15)	Medical Sciences
Sage (1)	International Journal of Robotics Research (JRN16)	Computing, Engineering
Springer (1)	GigaScience (JRN17)	Biology
Ubiquity (5)	Journal of Open Archaeology Data (JRN18)	Archaeology
	Journal of Open Humanities Data (JRN19)	Humanities
	Journal of Open Psychology Data (JRN20)	Psychology
	Journal of Open Research Software (JRN21)	Publishing and Book Trade, Computing
	Open Health Data (JRN22)	Medical Sciences
	Open Journal of Bioresources (JRN23)	Biology
University of York in UK (1)	Internet Archaeology (JRN24)	Archaeology
Wiley (2)	British Journal of Educational Technology (JRN25)	Education
	Geoscience Data Journal (JRN26)	Geography

Templates or guidelines offered by journals for data papers were downloaded or printed out for content analysis. Online websites were also cross-checked to examine the characteristics and structures embedded in data papers. Based on the analysis of characteristics and structures, a common framework was generalized to examine the embedded characteristics and structures of data papers. Furthermore, a crosswalk between the common frameworks is proposed herein and a data papers concept map (Candela et al. 2015) was also created to examine the similarities and differences for in-depth investigation of data papers.

Results

Type of data journal: pure or hybrid data journals

Thirteen journals were pure data journals focused on data papers. Ten journals were hybrid journals meaning that these journals do not only focus on traditional journal articles, but also regularly include data papers. Three journals used special issues to publish data papers (Table 2).

Table 2. Type of data journal

Type of Journal	Amount	Instance
Pure Data Journal	13	Biodiversity Data Journal, Data in Brief, Dataset Papers in Science, Earth System Science Data, Geoscience Data Journal, Geoscientific Model Development, Internet Archaeology, Journal of Open Archaeology Data, Journal of Open Psychology Data, Open Health Data, Open Journal of Bioresources, Research Data Journal for the Humanities and Social Sciences, Scientific Data
Hybrid Data Journal	10	Biomedical Data Journal, BMC Medical Education, BMC Research Notes, BMC Psychiatry, Ecological Archives, F1000Research, Genomics Data, GigaScience, Journal of Open Humanities Data, Journal of Open Research Software
Other	3	British Journal of Educational Technology, Earthquake Spectra, International Journal of Robotics Research,

Publication model: data paper, software paper, or overlay journal

There were four models for publication of research data. Twenty journals published data papers only, and two published software papers only (JRN16 and 21). Three journals covered both data papers and software papers (JRN01, 03 and 11). Interestingly, one used an overlay journal approach to publish data papers (JRN04).

A framework for data papers

The proposed framework for the structure of data papers can be generalized into three categories (title page, description of datasets, and relationship) and each category is composed of individual components used for description of various different objects, including a title page, a description of datasets for research data, and relationships among data papers, datasets, journal articles and data repositories. Within the framework, the proposed categories and their components and contents can not only be regarded as a two-level hierarchical metadata schemas, but also access points for discovery and contextual background information for sharing and reuse research data. Detailed information is shown below:

- Title page
  This category is composed of title, authors, author's affiliation, author's email address, abstract, keywords, identifiers, copyright, citation and date, and the aforementioned components are often regarded as basic information on descriptive metadata elements for data papers. The detailed analysis is as follows:
  - ✓ One journal (JRN04) does not offer authors' affiliations.
  - ✓ Four journals (JRN04, 07, 12 and 25) do not offer keywords.
  - ✓ One journal (JRN04) assigns specific identifiers using the journal data platform, rather than a DOI or URL.
  - ✓ Two journals (JRN04 and 16) do not have citation data for users.
  - ✓ Four journals (JRN04, 08, 16 and 25) do not indicate the date of data papers. Most data papers offered four kinds of dates (received, revised, accepted and published online) to illustrate the publishing process.
  - ✓ With the exception of JRN25 that adopts "all rights reserved" as copyright for data papers, the journals adopted the Creative Commons as licensing terms and conditions to release data papers openly for wide public use the same as open access. CC-BY is the most popular for licensing terms and conditions of data papers.
  - ✓ HTML (24/26) is the most popular format provided by data journals, followed by PDF (23/26). Following the HTML and PDF formats, XML was the third most popular format. Twelve of the 26 data journals offered two formats (HTML and PDF). Nine data journals offered three formats (HTML, PDF and XML) and one offered four formats (HTML, PDF, XML, and EPub). Three data journals offered HTML only, and one journal offered PDF only.

- Description of datasets
  This category consists of collection, description, coverage, identifier, competing interest, ethics approval, consent for publication, funding statement, copyright, reuse, availability, author's contribution, authors' information, references, and acknowledgements. These components

are used for description of research datasets with adequate understanding of the context within which they were collected or processed to answer specific research question(s). The information embedded in the aforementioned components is useful for reuse and interpretation for future studies and cannot be found in most descriptive metadata standards. The detailed analysis is as follows:

- ✓ Collection: this component focuses on how data is captured or created and reflects the significance of "research method" (Chao 2015). Most of the content of data papers provides information describing methodology through which data are collected or produced to answer specific research problems in a certain context. Other important information is included within the description of methodology such as background, ideas of the project, experimental design, factors, features and quality control.
- ✓ Description: this component is focused on the description of file information of data, including file format, versions, creation date and file creators. There are three approaches to this description of the dataset. The first is structured components similar to structured metadata elements (e.g., JRN18-20, 22-23 and 25), the second tends towards textual based description (JRN01-02, 04-05, 07, 09, 11-14, 16-17, 21 and 24), and the last is a hybrid approach of the first two with a structured category name with accompanying textual statements (e.g., JRN10) with basic information about the dataset. Most of the file formats of datasets described by data papers are dependent on the requirements of the data repositories in which they are deposited. Therefore, there is no common agreement on the file format of datasets.
- ✓ Coverage: data papers related to the disciplinary domains of medicine, biology, and archaeology are inclined to provide temporal and spatial keywords (e.g., JRN15, 18 and 22-24). Furthermore, some data papers indicate the spatial coverage by tagging with longitude and latitude (e.g., JRN9-10). In addition to temporal and spatial coverage, this component is also used to indicate the taxonomy in terms of the biological classification of species (e.g., JRN14-15).
- ✓ Competing interest: the majority of data papers provide this component that clarifies potential factors that might affect the results of the dataset (e.g., JRN01-03, 09-14, 17, 19-22, and 25).
- ✓ Ethical approval and consent for publication: data papers in which the target subjects are related to individual privacy or human and animal rights are required to indicate whether researchers have received approval from subjects for the public release and use of data (e.g., JRN01-03, 09-11, 1718, 20, 22-23, and 25).
- ✓ Funding statement: two approaches are used to indicate whether production of data was supported by a funding grant. One is described in this component, and the other is embedded in the ac-

knowledgment component.
- ✓ Copyright: the majority of subjects in this study tend towards the adoption of open licensing terms and conditions such as CC, CC0 and PDDL.
- ✓ Author's contribution: an interesting point that deserves note is that some data papers offer information to clarify the contribution of each author to the data paper (e.g., JRN01-03, 05, 11, 13, 17 and 23).

- Relationships: there are three types of relationships between data papers and their datasets (or data repositories); indication of the relationship between versions of data papers (e.g., JRN11), between datasets and their derived journal article (e.g., JRN08-09, 12 and 15), and between datasets and the data repository in which they are deposited (e.g., JRN01-06, 09-15, 17-22 and 24-26). The aforementioned types of relationships can be regarded as three components of this category.

A crosswalk between the proposed common framework of this study and a concept map for data papers

A mapping between the common framework proposed in this study and a concept map (CP) (Candela et al. 2015) revealed that these two are fully interoperable (Table 3). We further illustrate the components of data papers in a more concrete manner. The Global Biodiversity Information Facility (GBIF) has transferred the content of the data paper into GBIF Integrated Publishing Tool (IPT) Metadata Profile elements with clear instructions (Penev et al. 2015). Using the successful case proved by GBIF, we mapped the components of the proposed common framework into those of CP (Table 4). In addition to one-to-one, many-to-one and one-to-many, some components embedded in data papers, such as ethical approval, consent to publication and relationships, have not been provided and defined by CP (Candela et al. 2015).

Table 3. A crosswalk between components of CP and the proposed framework found in this study

Components of CP	Components of Proposed Framework
Identifier-Data Paper	Identifier component of title page category
Content-Data Paper	Category of description of dataset and its components
	Category of relationships and its components
Metadata-Data Paper	Title page category and its components with exclusion of identifier

Table 4. A crosswalk between CP category and subcategories in the proposed framework found in this study

Category of CP	Common Framework
Quality	Collection
Provenance	
Project	
	Funding statement
Coverage	Coverage
Reuse	Copyright
License	
Competing Interest	Competing Interest
Microcontribution	Authors' contribution
Availability	Identifier
Format	Description
Not available	Ethical approval
Not available	Consent to publication
Not available	Relationships

Discussion: Granularity described by data papers

In terms of metadata, the objects described can be separated into different levels, such as item, collection, series and so on. The one-to-one principles are widely adopted for objects described by metadata. This means that each kind of object owns its metadata record respectively, and the relationship between objects is described by a specific metadata element (e.g. Relation of Dublin Core Element Set) embedded in metadata records. In this study, the objects described by data papers are mainly focused on the item (i.e. a single dataset). Only a few focused on multiple datasets or collection of datasets (e.g. databases) such as JRN01, 02, 05-06 and 13. We also found that data papers are a compound object used to describe data paper, research datasets and the relationship between datasets, data paper, journal article and data repository altogether. Therefore, data papers have two identifiers at least: one is for the data paper and the other is for the dataset(s). Furthermore, if a journal article is derived from research datasets, then the relationship between datasets and journal article is linked by the third kind of identifier: that is the identifier of the journal article (e.g., the DOI provided by the journal publisher). As a result, the approach to describe the object's granularity of research data between metadata and data papers is fundamentally distinctive.

Conclusion

In this study, 26 data journals of 16 publishers were selected to examine the characteristics and structures embedded in data papers. CP of data papers (Candela et al. 2015) was embodied into more concrete and extended components, and new components for the structure of data papers were provided although this preliminary study only partially reveals the phenomena associated with data papers. Furthermore, the proposed categories and their components and descriptions can be regarded as a useful facilitator for discovery, sharing and reuse of research data, as well as a useful basis to develop a set of structured descriptive metadata elements for research data. In the near future, we intend to include more target subjects to examine the feasibility of a proposed common framework and investigate more implicit characteristics and structures embedded in data papers, such as core and optional characteristics and structures of proposed categories, and the comparison with a metadata approach in RDM.

References

Akers, Katherine. 2012. "Data journals: Incentivizing research data dissemination." CLIR Blog, 12 December. Accessed November 19, 2015. http://connect.clir.org/blogs/katherine-akers/2013/12/12/data-journals-incentivizing-research-data-dissemination

Akers, Katherine. 2014. "A growing list of data journals." Data@MLibrary Blog, 9 May. Accessed January 7, 2015. https://mlibrarydata.wordpress.com/2014/05/09/data-journals/

Atici, Levent, Sarah Whitcher Kansa, Justin Lev-Tov, and Eric C. Kansa. 2013. "Other people's data: A demonstration of the imperative of publishing primary data." Journal of Archaeological Method and Theory 20, 4: 663-681. https://doi.org/10.1007/s10816-012-9132-9

Borgman, Christine L., Jillian C. Wallis, Matthew S. Mayernik, and Alberto Pepe. 2007. "Drowning in data: Digital library architecture to support scientific use of embedded sensor networks." In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, 269-277. New York, NY: ACM. https://doi.org/10.1145/1255175.1255228

Breure, Leen. 2014. "Enhanced data journal: Next generation science." E-data %26 research, Special Issue. Accessed November 19, 2015. http://www.edata.nl/2014_special-issue/pdf/Enhanced_data_journal.pdf

Callaghan, Sarah, Jonathan Tedds, Rebecca Lawrence, Fiona Murphy, Timothy Roberts, and Will Wilcox. 2014. "Cross-linking between journal publications and data repositories: A selection of examples."

International Journal of Digital Curation 9, 1: 164-75. https://doi.org/10.2218/ijdc.v9i1.310

Callaghan, Sarah, Steve Donegan, Sam Pepler, Mark Thorley, Nathan Cunningham, Peter Kirsch, Linda Ault, Patrick Bell, Rod Bowie, Adam Leadbetter, Gwen Moncoiffé, Kate Harrison, Ben Smith-Haddon, Anita Weatherby, and D. Wright. 2012. "Making data a first class scientific output: Data citation and publication by NERCs Environmental Data Centres." International Journal of Digital Curation 7, 1: 107-13. doi:10.2218/ijdc.v7i1.218 https://doi.org/10.2218/ijdc.v7i1.218

Candela, Leonard, Donatella Castelli, Paolo Manghi, and Alice Tani. 2015. "Data journals: a survey." Journal of the Association for Information Science 66, 9:1747–62. https://doi.org/10.1002/asi.23358

Chao, Tiffany C. 2015. "Mapping methods metadata for research data." International Journal of Digital Curation 10, 1: 82-94. https://doi.org/10.2218/ijdc.v10i1.347

Chavan, Vishwas, and Lyubomir Penev. 2011. "The data paper: A mechanism to incentivize data publishing in biodiversity science." BMC Informatics 12, S15: S2. doi: 10.1186/1471-2105-12-S15-S2 https://doi.org/10.1186/1471-2105-12-S15-S2

Costello, Mark J. 2009. "Motivating online publication of data." BioScience 59, 5: 418-427. doi: 10.1525/bio.2009.59.5.9 https://doi.org/10.1525/bio.2009.59.5.9

De Schutter, Erik, Giorgio A. Ascoli, and David N. Kennedy. 2009. "Review of papers describing neuroinformatics software." Neuroinformatics 7, 4: 211-212. doi: 10.1007/s12021-009-9058-x https://doi.org/10.1007/s12021-009-9058-x

Gorgolewski, Krzysztof J., Daniel S. Margulies, and Michael P. Milham. 2013. "Making data sharing count: A publication-based solution." Frontiers in Neuroscience 7, Article 9. doi: http://dx.doi.org/10.3389/fnins.2013.00009 https://doi.org/10.3389/fnins.2013.00009

Gray, Stephen. 2015. "Case study: Publishing a data paper." Accessed December 28, 2015. https://data.bris.ac.uk/files/2015/05/Publishing-a-data-paper.pdf

Kansa, Eric C., and Sarah Whitcher Kansa. 2013. "We all know that a 14 is a sheep: data publication and professionalism in archaeological communication." Journal of Eastern Mediterranean Archaeology and Heritage Studies 1, no. 1(2013): 1-14. Accessed November 19, 2015. doi: http://dx.doi.org/10.1353/ema.2013.0007

Kennedy, David N., Giorgio A. Ascoli, and Erik De Schutter. 2011. "Next steps in data publishing." Neuroinform 9, 4: 317-320. https://doi.org/10.1007/s12021-011-9131-0

Kratz, John, and Strasser Carly. 2014. "Data publication consensus and controversies." Version 3, F1000 Research 3:94: 1-21. doi: 10.12688/f1000research.3979.3 https://doi.org/10.12688/f1000research.3979.3

Niu, Jinfang, and Margaret Hedstrom. 2008. "Documentation evaluation model for social science data." Proceedings of the American Society for Information Science and Technology 45, 1: 1-11. doi: 10.1002/meet.2008.1450450223 https://doi.org/10.1002/meet.2008.1450450223

Penev, Lyubomir, Daniel Mietchen, Vishwas Chavan, Gregor Hagedorn, David Remsen, Vincent Smith, and David Shotton. 2015. "Pensoft data publishing policies and guidelines for biodiversity data." Accessed October 06, 2015. http:// www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf

Rees, Jonathan. 2010. "Recommendations for independent scholarly publication of data sets." Accessed October 06, 2015. http://neurocommons.org/report/data-publication.pdf

Smith, Vincent S. 2009. "Data publication: Towards a database of everything." BMC Research Notes 2, 113: 1-3. doi: 10.1186/1756-0500-2-113 https://doi.org/10.1186/1756-0500-2-113

Strasser, Carly. 2015. "Research data management." Accessed December 25, 2015. http://www.niso.org/apps/group_public/download.php/15375/PrimerRDM-2015-0727.pdf