

Data description in data repositories: metadata schemas for social sciences

Li Yang^{1,*}, Margaret E.I. Kipp¹

¹ University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA

Abstract

This project studies how data is described in data repositories related to social science. Using a qualitative content analysis method, we analyzed 39 metadata schemas and categorized them into five categories. The result shows the connections, common features in data description, and application of the metadata schemas.

Keywords

Data description, metadata, metadata schemas, data repositories, social sciences

1. Introduction

Data volume in the social sciences has surged due to increases in data-driven studies, cross-disciplinary research, funding requirements, etc. A large number of data repositories and data archives at international, national, regional, and institutional levels have appeared to provide long-term data preservation, curation, and dissemination for data sharing and reuse. Data description plays a critical and foundational role in implementing the FAIR (Findable, Accessible, Interoperable, and Reusable) principle to optimize reusing data [1].

Data description is the process of describing data following certain schemas to identify, access, reuse, and analyze data. Accordingly, a data description has at least three aspects: describing data for identification and access, describing data for reuse, and describing data for analysis. The first aspect requires describing the data features, similar to descriptive metadata, including the title, creator, dates, and subject. The second aspect is about describing features that will help users decide how the data will be reused, including copyright information, licenses, user terms or agreements. The third aspect describes the data content at the variable level and the context of the collection.

This study explores how research data is described in data repositories in the social sciences domain, focusing on the metadata schemas used and the attributes covered.

2. Practice Status

Data description requirements can be found in data policies of data repositories, data management plans required by grant funding agencies at various levels from federal to local, and best practices in general or specific domains.

Data repositories, platforms, or consortiums have similar but different data description requirements. For example, the consortium of European Social Science Data Archives (CESSDA) uses project-level and data-level documentation to describe contextual information and metadata for dataset description [2]. The CESSDA Metadata Model (CMM), defines elements in 10 categories for data description: study, person(s), institute(s), dataset,

* Corresponding author.

 liyang@uwm.edu (L. Yang); kipp.uwm.edu (M.E.I. Kipp)

 [0000-0002-4793-7592](https://orcid.org/0000-0002-4793-7592) (L. Yang); [0000-0002-4202-6239](https://orcid.org/0000-0002-4202-6239) (M.E.I. Kipp)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

instrument, questions and responses, concepts, further documents, publication, group of studies, and document description [3].

Data management plans required by grant funding agencies such as the National Science Foundation (NSF), the National Institutes of Health (NIH) and the National Institutes of Justice (NIJ) in the US, and the Economic and Social Research Council (ESRC) in the UK ask the funding applicants to describe data files, data types, data formats and standards, metadata, documentation, terms for data access and sharing, physical samples or collections, related tools, software or code, etc. [4, 5, 6].

3. Methodology

This study examines metadata for data repositories based on practical rather than theoretical considerations. Starting from the practice of data description, we used a qualitative content analysis method to analyze metadata schemas. First, we performed metadata source tracing analysis. We found four existing metadata standards to serve as the top categories. Second, we grouped all the data according to the top categories. We created another category to hold the data which could not be categorized. Third, we analyzed the hierarchies and elements of each schema. Two researchers performed the analysis, and then reviewed the codes and reached a consensus. We adopted a purposive sampling strategy with all the metadata schemas collected being used by one or more data repositories between October 2023 and February 2024. We used a set of criteria to select the repositories: 1. The data repositories must be open to the public nationally or internationally, including open institutional data repositories. 2. Individuals can deposit, search, view, and download datasets and their metadata from the data repositories. 3. No government open data portals for governmental data. 4. The data repositories must contain multiple disciplines including social sciences or be specifically for social sciences. 5. The data repositories must be accessible at the time of data collection.

We collected 48 data repositories according to the criteria. After preliminary analysis, we found some data repositories did not have sufficient information about their metadata schema. We retained 39 metadata schemas from collected data repositories in Asia, Australia, Europe, North America, and South America.

4. Findings

4.1. Metadata schemas and categorization

Some repositories use existing standards, while some use application profiles or derivatives of the standards. Others are homegrown schemas, which might refer to more than one standard. All the metadata schemas fall into five groups: DataCite, Dataverse, DDI, Dublin Core (DC), and homegrown schemas. The DDI and Dataverse schemas are mostly used in social sciences research data management. Table 1 shows the categorization of the schemas.

Table 1: Metadata Schemas Categorization

Category	Metadata Schemas
DataCite: 4	Figshare Schema, home region in the UK RADAR (Research Data Repository) Schema, Germany da ra Metadata Schema, Germany Zenodo Schema, Switzerland
Dataverse: 12	ADA (Australian Data Archive) Schema, Australia AUSSDA (The Austrian Social Science Data Archive) Schema, Austrian Borealis Schema, Canada CORA.RDR (CORA. Repositori de Dades de Recerca) Schema, Spain

	CROSSDA (Croatian Social Science Data Archive) Schema, Croatia DANS (Data Archiving and Networked Services) Schema, Netherlands DataverseNO Schema, Norway Harvard Dataverse Schema, US Peking University Open Research Data Platform Schema, China REDU (Unicamp Research Data Repository) Schema, Brazil So.Da.Net (Greek research infrastructure for the social sciences) Schema, Greece SODHA (Social Sciences and Digital Humanities Archive) Schema, Belgium
DDI: 15	ADP (Social Science Data Archives) Schema, Slovenia APIS (The Portuguese Archive of Social Information) Schema, Portugal CESSDA (Consortium of European Social Science Data Archives) Schema, Europe CSDA (Czech Social Science Data Archive) Schema, Czech DATICE (The Icelandic Social Science Data Service) Schema, Iceland FSD (Finnish Social Science Data Archive) Schema, Finland GESIS – Leibniz Institute for the Social Sciences Data Services Schema, Germany ICPSR (Inter-University Consortium for Political and Social Research) Schema, US ICSSR (Indian Council of Social Science Research) Data Service Schema, India ISSDA (Irish Social Science Data Archive) Schema, Ireland Sikt (Norwegian Agency for Shared Services in Education and Research), Norway SND (Swedish National Data Service) Schema, Sweden SSJDA (Social Science Japan Data Archive) Schema, Japan UKDC (UK data service) Schema, UK UniData – Bicocca Data Archive Schema, Italy
DC: 3	DANS-EASY, Netherlands DRI (The Digital Repository of Ireland), Ireland SCIDB (Science Data Bank), China
Home-grown: 5	DataON Schema, Korea DNA (Danish National Archives) Schema, Denmark QDR (Qualitative Data Repository) Schema, US RIF-CS (The Registry Interchange Format: Collections and Services), Australia SWISSUbase Schema, Switzerland

4.2. Existing standards, use applications and derivatives

DC, including all DCMI metadata terms, is the most widely used metadata standard that can be used to describe various types of resources including data [7]. DC is general but sufficient for citation and discovery purposes. All three schemas in the DC group used DCMI terms with additional local and light hierarchies.

DataCite provides the Digital Object Identifier (DOIs) registration service. The DataCite metadata schema aims to support data citation and discovery. It has 20 properties and many sub-properties, closely mapping to DC [8]. All four schemas in the DataCite group are derivatives with some enrichments based on the DataCite Schema.

DDI (Data Documentation Initiative) is a set of international metadata standards describing various data types in social sciences. It has DDI-Lifecycle, the full version of the standard supporting the data documentation at different stages; DDI-Codebook, a light version with a six-level hierarchy; and DDI-CDI for cross-domain integration [9]. All the metadata schemas in the DDI group are application profiles or derivatives that use DDI in a stripped-down manner.

Dataverse provides a customizable metadata schema primarily based on the DDI Codebook. Dataverse metadata schema consists of citation metadata, domain-specific metadata and file-level metadata as metadata blocks [10]. For the schemas in the Dataverse group, the most used blocks are Citation Metadata, Geospatial Metadata, and Social Science and Humanities Metadata.

4.3 Homegrown schemas

DataON is a national research data platform in Korea. The DataON Schema is a five-level hierarchy and has 149 elements in four categories, Collection, Dataset, File, and Repository. It was created based on DCMI, DataCite Schema, and Metadata Schema for the Description of Research Data Repositories 3.0 by re3data.org (re3data Schema) [11].

DNA is the national archives platform in Denmark. DNA Schema has two layers for describing research data: 1. describes general archive information and context information; 2. describes data. For quantitative data, 9 elements describe the software, data file, variables, reference, code, etc. For qualitative data, there are 16 fields, including SourcePath with an additional 15 elements from Dublin Core [12].

QDR provides qualitative data services in social sciences and related disciplines. The QDR Schema has 102 elements with 72 optional fields and 28 required fields. It describes information about citation, funding, data, related publications, geospatial information, terms of use, and data availability. It closely maps to DDI Codebook 2.5, Datacite 3.1, and Dataverse metadata [13].

RIF-CS, created based on ISO 2146:2010, is an Australian metadata interchange standard. RIF-CS is hierarchical with seven levels. It defines four classes of objects including activity, collection, party, and service. The elements describe data and its contextual information, access and terms, instruments and tools, etc. [14].

The SWISSUbase Schema has a three-level structure with three top categories Study, Dataset, and Data File [15]. Elements in Study describe the project information, funding, and related publications. Elements in Dataset describe the data information, citation, and access. Data File describes title, type, and access. The SWISSUbase metadata schema does not describe variable-level information.

4.4 Relationships between the metadata schemas

Besides the use application and derivative relationships, the existing standards and homegrown schemas are interconnected with mapping relationships and reference relationships. Figure 1 shows the interconnections of the metadata schemas.

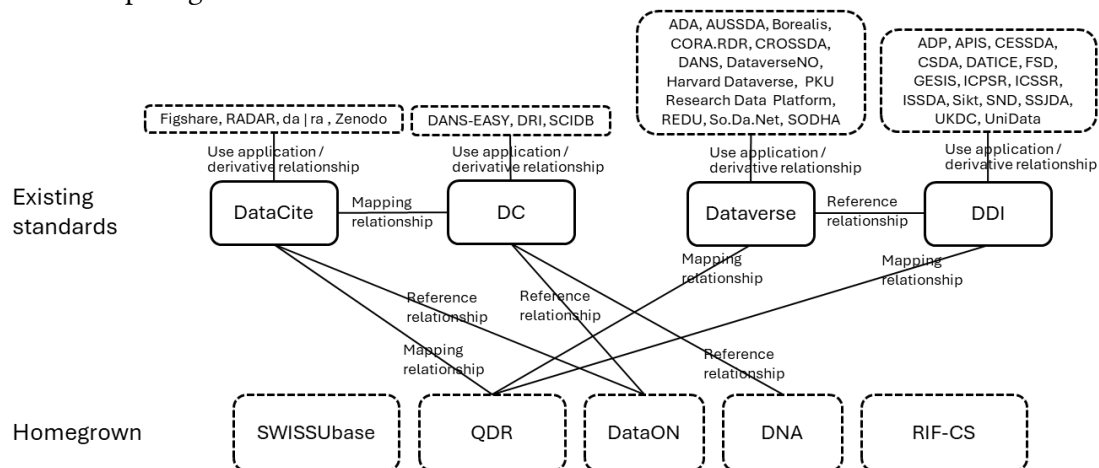


Figure 1: Interconnections with relationships between the metadata schemas

5. Summary

This is a study in progress. We collected 39 metadata schemas used by data repositories related to social sciences. We found that in practice, data description includes describing data at both descriptive and variable levels, describing the context of data in the form of

documentation, and describing the use of data in the form of a license or use agreement and the related auxiliary tools. In the next step, we will explore the schemas further, focusing on element analysis from describing data for identification and access, describing data for reuse, and describing data for analysis.

References

- [1] GO FAIR, FAIR Principles, 2016. URL: <https://www.go-fair.org/fair-principles/>.
- [2] CESSDA Training Team. CESSDA Data Management Expert Guide. CESSDA ERIC, 2020. URL: <https://doi.org/10.5281/zenodo.3820473>.
- [3] Borschewski K, Förster A, Friedrich T, Zenk-Möltgen W, Miranda P, Moura Ferreira P, et al., CMM CESSDA Metadata Model, 2019. URL: <https://doi.org/10.5281/zenodo.3543756>.
- [4] NIH, Final NIH Policy for Data Management and Sharing, 2020. URL: <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html>.
- [5] NIJ, Data Archiving, 2023. URL: <https://nij.ojp.gov/funding/data-archiving#data-archiving-plan>.
- [6] NSF. Proposal & Award Policies & Procedures Guide (PAPPG) (NSF 24-1), 2024. URL: https://nsf-gov-resources.nsf.gov/files/nsf24_1.pdf.
- [7] DCMi Usage Board, DCMi Metadata Terms, 2020. URL: <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>.
- [8] DataCite Metadata Working Group, DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.5. DataCite e.V, 2024. URL: <https://doi.org/10.14454/g8e5-6293>.
- [9] DDI, What is DDI?, 2024. URL: <https://ddialliance.org/learn/what-is-ddi>.
- [10] Dataverse Project, Dataset + File Management, 2023. URL: <https://guides.dataverse.org/en/latest/user/dataset-management.html>.
- [11] Korea Institute of Science and Technology Information. Research data management and utilization for government-funded research institutes in Korea, 2019. URL: <https://scienceon.kisti.re.kr/commons/util/originalView.do?cn=TRKO202200000068&dbt=TRKO&rn=>.
- [12] Rigsarkivet, Executive Order on Information Packages, 2022. URL: <https://en.rigsarkivet.dk/wp-content/uploads/2022/09/Executive-Order-on-Information-Packages128.pdf>.
- [13] Qualitative Data Repository, QDR Metadata Application Profile, 2024. URL: <https://qdr.syr.edu/content/qdr-metadata-application-profile>.
- [14] Research Data Australia, Registry Interchange Format - Collections and Services RIF-CS v1.6.5 Schema Guidelines, 2022. URL: <https://researchdata.edu.au/documentation/rifcs/guidelines/rif-cs.html>.
- [15] SWISSbase, Metadata Guide, 2022. URL: https://resources.swissubase.ch/wp-content/uploads/2023/06/SWISSUbase-Metadata-Guide_update_20230607_pw.pdf.