In: Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure, Edited by J.I. Lane, I. Mulvany, and P. Nathan, Sage Publishing, 2020.

4

METADATA FOR SOCIAL SCIENCE DATASETS

Robert B. Allen

Introduction 40
Data Elements and Datasets 40
Metadata Schemas and Catalogues 41
Linked Data 42
Richer Semantics 42
Data Repositories and Collections of Datasets 43
Repository Services 46
Infrastructure 49
Conclusion 49
Acknowledgements 50
References 50
Notes 51

Introduction

Evidence-based policy needs relevant data (Commission on Evidence-Based Policymaking, 2017; Lane, 2016). Such data is often difficult to find and use. The FAIR open access guidelines suggest that data should ideally be *f*indable, *a*ccessible, *i*nteroperable, and *r*eusable (FAIR).[1] Broad and consistent metadata can support these needs. Metadata and other knowledge structures could also supplement and ultimately even replace text.

This chapter surveys the state of the art of metadata for numeric datasets, focusing on metadata for administrative and social science records. Administrative records describe details about the state of the world as collected by organizations or agencies. They include governmental, hospital, educational, and business records. By comparison, social science data generally is collected for the purpose of developing or applying theory.

We start by considering data and datasets, and then the basic principles of metadata and their application to datasets. Modern metadata is often implemented with Resource Description Framework (RDF) linked data. Next, we introduce ontologies and other semantic approaches. We then move on to applications which use metadata. We examine repositories that hold and distribute collections of datasets. We then describe services and techniques associated with repositories. We conclude by briefly describing the computing infrastructure for repositories.

Data Elements and Datasets

While data may be incorporated in text, image or video, here we focus on numeric observations recorded and maintained in machine-readable form. Individual observations are rarely used in isolation. Rather, they are typically collected into datasets.

A dataset is defined in the W3C-DCAT (W3C Data Catalog Vocabulary)[2] as 'a collection of data, published or curated by a single agent'[3] such as a statistical agency. There are many different types of datasets; they differ in their structure, their source, and their use. A given data element may appear in many different datasets and may be numerically combined with other data to form derived data elements which then appear in still other datasets. In some cases, they are single vectors of data; in other cases, they comprise all the data associated with one study or across a group of related datasets. Reference datasets are generally collected and archived because they are of enduring value and can be used for answering many different types of questions. Other datasets, such as an individual's medical records, are associated with a relatively narrow set of applications.

There is wide variability in the organization and contents of datasets, as well as in the extent to which datasets are validated and curated. Potentially with frameworks such as the SDMX (Statistical Data and Metadata eXchange) Guidelines for the Design of Data Structure Definitions,[4] concise structured descriptions can be developed for how data elements are combined to form datasets.

Metadata Schemas and Catalogues

Many datasets are available; the DataCite repository alone contains over 5 million datasets. Metadata can support users in finding datasets and enable users to know what is in them. Metadata are short descriptors which refer to a digital object. However, there is tremendous variability in the types of metadata and how they are applied. One categorization of metadata identifies structural (or technical), administrative and descriptive metadata (Riley, 2017). Structural metadata includes the organization of the files. Administrative metadata describes the permissions, rights, preservation and usage relating to the data.[5] Descriptive metadata covers the contents.

A metadata element describes an attribute of a digital object. The simplest metadata (e.g., a digital object identifier (DOI) or ORCiD[6]) identifies the digital object or its creator.[7] Metadata elements are generally part of a schema or frame. DCAT[8] is a schema standard for datasets that is used by many repositories such as data.gov. Other structured frameworks for datasets include the DataCite[9] metadata schema and the Inter-university Consortium for Political and Social Research Data Documentation Initiative (ICPSR DDI; see below). ISO 19115–1:2014 establishes a schema for describing geographic information and services.[10]

The schema specifications provide a flexible framework. For instance, DCAT allows the inclusion of metadata elements drawn from domain schemas and ontologies. Some of these domain schemas are widely used resources which DCAT refers to as 'assets'. Figure 4.1 shows a fragment of properties (i.e. metadata elements) from an implementation of the Schema.org[11] dataset schema to describe gross domestic product (GDP).


```
{
    "name": "gdp",
    "title": "Country, Regional and World GDP (Gross Domestic Product)",
    "description": "Country, regional and world GDP in current US Dollars ($). Regional means collections of countries e.g. Europe & Central Asia. Data is sourced from the World Bank and turned into a standard normalized CSV.",
    "image": "http://assets.okfn.org/p/opendatahandbook/img/data-wrench.png",
    "readme": "Country, regional and world GDP in current US Dollars ($). Regional means\ncollections of countries e.g. Europe & Central Asia.\n\n## Data\n\nThe data is sourced from the World Bank (specifically [this dataset][current]) which\nin turn lists as sources: *World Bank national accounts data, and OECD National\nAccounts data files*.\n\nNote that there are a variety of different GDP indicators on offer from the\nWorld Bank including:\n\n* [GDP in current USD][current]\n* [GDP in constant USD (2000)][constant]\n* [GDP, PPP (constant 2005 international $)][ppp]\n* [GDP (constant LCU)][lcu]\n\n[constant]: http://data.worldbank.org/indicator/NY.GDP.MKTP.KD\n[current]: http://data.worldbank.org/indicator/NY.GDP.MKTP.CD\n[ppp]: http://data.worldbank.org/indicator/NY.GDP.MKTP.PP.KD\n[lcu]: http://data.worldbank.org/indicator/NY.GDP.MKTP.KN\n\n## Automation\n\nDatahub updates this dataset every year automatically.\n\n## License\n\nThis Data Package is made available under the Public Domain Dedication and License v1.0 whose full text can be found at: http://www.opendatacommons.org/licenses/pddl/1.0/",
    "keywords": [
        "GDP",
        "World",
        "Gross Domestic Product",
        "Time series"
    ],
    "last_updated": "2018-01-19",
    "licenses": [
    {
        "name": "ODC-PDDL-1.0",
        "path": "http://opendatacommons.org/licenses/pddl/",
```


Figure 4.1 Fragment of GDP properties described by the Schema.org dataset schema[12]

Metadata terms for an application are often assembled into namespaces from different metadata schemas. Metadata application profiles[13] provide constraints on the types of entities that can be included in the metadata for a given application. Moreover, application profiles can be used to validate standards. For instance, the DCAT application profile for data portals in Europe (DCAT-AP) supports the integration of data drawn from repositories in different jurisdictions in the EU.[14]

A collection of dataset schemas,[15] such as all the datasets in a repository, forms a catalogue. For data streams, there needs to be continuity but also the ability to update the records. In some cases, there may be relatively infrequent periodic updates. These could be given version numbers rather than an entirely new DOI.[16] However, collections of highly dynamic data streams present challenges; most of the data stay the same but some of the data and/or metadata (e.g. number of records) change.

Linked Data

RDF extends XML by requiring triples which assert a relationship (property) between two identifiers: 'identifier – property – identifier'. RDF Schema (RDFS) extends RDF by supporting subclass relationships. A graph is formed by linking triples.

Hierarchical classification systems are another knowledge structure with a long history. Indeed, Schema.org is based around a hierarchical ontology. Simple classification relationships are handled by the Simple Knowledge Organization System (SKOS). SKOS represents the hierarchical structure of traditional thesauri with RDFS. Collections of data organized by SKOS are often described as 'linked data'.

Depending on the rigour with which they are developed, these collections can support limited logical inference. Many administrative and social-science-related thesauri, such as EDGAR and those of the World Bank and the OECD, have now been implemented with SKOS. A knowledge base is, primarily, a SKOS graph that links real-world entities. For example, Wikidata[17] is an effort to develop a knowledge base based on structured data from Wikimedia projects, and VIVO[18] is a knowledge graph of scholarship.

But there are also many stand-alone classification schemes. The Extended Knowledge Organization System (XKOS)[19] was developed to allow classification systems to be incorporated into a SKOS framework.

Richer Semantics

Ontologies provide a coherent set of relationships between entities which cover a given domain. Well-constructed ontologies can support logical inference. Some vocabularies, such as Dublin Core, which is implemented in RDF, are said to have an ontology, but they are limited because relationships among the terms are not specified. FOAF (Friend of a Friend) provides a somewhat richer ontology which includes attributes associated with

people. Still more extensive ontologies often use OWL (Web Ontology Language) which can support stronger logical inference than RDFS.

One way to coordinate across terms is an upper ontology. Upper ontologies provide top-down structures for the types of entities allowed in domain and application ontologies. One of the best-known upper ontologies is the Basic Formal Ontology (BFO; Arp et al., 2015), which is a realist, Aristotelian approach. At the top-level, BFO distinguishes between continuants (endurants) and occurrents (perdurants) and also between universals and particulars (instances). Many biomedical ontologies based on BFO are collected in the Open Biomedical Ontology (OBO) Foundry.[20]

There are fewer rich ontologies dealing with social science content than for natural science. Social ontology, that is, developing rigorous definitions for social terms, is often a challenge. It is difficult to define precisely what is a family, a crime, or money. In most cases, an operational or approximate definition may suffice when formal definitions are difficult. However, those operational definitions often do not interoperate well across studies.

Data Repositories and Collections of Datasets

A data repository holds datasets and related digital objects. Ideally, it contains a stable collection selected according to a collection policy. It is organized by metadata and knowledge structures. It provides access to the datasets and typically supports search.

Version
Study Title
Alternate Title
PIs & Affliation
Funding Agencies
Summary
Subject Terms
Geographic Coverage Areas
Geographic Representation
Study Time Periods and Time Frames
Collection Notes
Study Purpose
Study Design
Description of Variables
Sampling: Sampling Procedure, Sampling Unit, Sampling Notes
Oversampled Group
Time Method
Data Source Type
Mode of Collection
Weight
Response Rates
Scales
Analysis Unit
Unit of Observation
Smallest Geographic Unit
Data Format
Restrictions
Version History

Figure 4.2 ICPSR DDI metadata elements

The Inter-university Consortium for Political and Social Research (ICPSR)

The ICPSR[21] is a major repository of public-use social science and administrative datasets derived mostly from questionnaires and surveys. We go into depth about it here because the ICPSR DDI[22] (e.g. Vardigan et al., 2008) is especially well crafted.[23] The DDI codebook saves the exact wording of all the questions and the ICPSR provides an index of all variable names. DDI-Lifecycle is an extension that describes the broader context in which the survey was administered as well as the details about the preservation of the file. DDI uses XKOS to provide linked data. Figure 4.2 shows the ICPSR DDI metadata schema.

The ICPSR metadata elements incorporate aspects of the implementation and design of research studies. However, many of the ICPSR metadata elements are not independent; potentially, they could be interlinked with terms such as organizations, locations, individuals, and research designs from other knowledge bases. Moreover, they could be linked with higher-level workflows and mechanisms.

Additional Examples of Repositories

Statistical data collection is a core function of government. Such collections often emphasize social data on employment, criminal justice, and public health, for example. They also include related indicators such as agricultural and industrial output and housing. Most countries have national statistical agencies such as Statistics New Zealand, and the Korean Social Science Data Archive. European datasets are maintained in the Consortium of European Social Science Data Archives[24] and the European Social Survey.[25] Australia has a broad data management initiative, the Australian National Data Service.[26] Many US federal governmental datasets are collected at data.gov. In addition, there are many other social survey repositories,[27] and many US states and cities have online statistics sites at varying levels of sophistication.

There are also many non-governmental and intergovernmental agencies such as the OECD, the World Bank, and the United Nations that manage datasets. Similarly, there are very large datasets from medical research such as from clinical trials and from clinical practice including electronic health records.

Many datasets are produced, curated, and used in the natural sciences such as astronomy and geosciences. Some of these datasets have highly automated data collection, elaborate archives, and established curation methods. Many repositories contain multiple datasets for which access is supported with portals or data cubes. For instance, massive amounts of geophysical data and related text documents are collected in the EarthCube[28] portal. The science.gov portal is maintained by the US Office of Science Technology and Policy. NASA supports approximately 25 different data portals. Each satellite in the Earth Observation System may provide hundreds of streams of data,[29] with much common metadata. Likewise, there are massive genomics and proteomics datasets which are accessible via portals such as UniProt[30] and the Protein Data Bank,[31] along with suites of tools for exploring them.

Repository Registries

There are a lot of different repositories, so it is useful to have a registry with a standard schema structure for describing them. The Registry of Research Data Repositories,[32] which is operated by DataCite, links to more than 2000 repositories, each of which holds many datasets. Each of those repositories is described by the re3data.org schema (Rücknagel et al., 2015).

Ecosystems of Texts and Datasets

Datasets are often associated with text reports, whether they describe the development of the datasets or their use. Ultimately, we would like to be able to move seamlessly from datasets to texts and other related materials. However, as demonstrated by several of the papers in this volume, it is often difficult to extract details about datasets from legacy publications.

Text associated with a dataset may be used to support searching for it. Indeed, Google Dataset Search uses texts marked up with Schema.org JavaScript Object Notation for Linked Data microdata to generate an index.

Going forward, great value can be achieved by persuading editors and authors to clearly cite and deposit datasets. In some cases, a separate data editor may be appointed. The Dryad Digital Repository[33] captures datasets from scholarly publications. It requires the deposit of data associated with scholarly papers accepted for publication. Such datasets are most often used to validate the conclusions of a research publication, but they may also be used more broadly.

Research datasets may be given DOIs[34] and cited in much the same way that research reports are cited. Formal citations can support tracing the origins of data used in analyses and help to acknowledge the work of the creators of the datasets.

Information Institutions and Organizations

The Open Archival Information System (OAIS) provides a reference model for the management of archives (Lee, 2010). A key part of the model is the inclusion of preservation planning and the requirement for stable administration over time. These attributes are part of all information institutions. Libraries, archives and museums have formal collection management strategies, metrics and policies.

In addition to traditional information institutions, there are now many other players. CrossRef[35] and DataCite are DOI registration agencies. CrossRef is a portal to metadata for scholarly articles, while DataCite provides metadata for digital objects associated with research. Schema.org's primary mission is to provide a structure that improves indexing by search engine companies. Still other organizations such as Health Level Seven International[36] and the Kyoto Encyclopedia of Genes and Genomes[37] manage controlled vocabularies and frameworks. These organizations are increasingly adopting best practices similar to those of traditional information organizations.

Repository Services

Administrative Metadata and Related Services

Administrative metadata is one of the three broad categories of metadata. Administrative metadata describes the permissions, rights, preservation, and usage of the data. While the focus of a traditional library is to support access and the focus of an archive is to ensure stability and quality, digital repositories must increasingly address both access and preservation.

Preservation and Trusted Datasets

Although data storage prices are declining dramatically, the cost of maintaining a trusted repository remains substantial, and we cannot save everything. These challenges are familiar from traditional archives; selection policies typical in archives could help in controlling the many poorly documented datasets in some repositories. Yet, prioritization of what to select is difficult (Whyte and Wilson, 2010).[38]

Lost data is often irreplaceable. Even if the data is not entirely lost, users need confidence that the validity of stored data has not been compromised. Indeed, some data may become the target of malicious attacks. Trust is a result of both technology and organizational procedures. Technology may include hash-based encoding of data. CLOCKSS (Controlled Lots of Copies Keep Stuff Safe)[39] is a distributed hash system for web-based scholarly literature. Blockchains provide hashed records of transactions and can be applied to data records.

The OAIS framework has been incorporated into the ICPSR DDI-Lifecycle model. The integrated Rule-Oriented Data System[40] is a policy-based archival management system[41] developed for large data stores. It implements a service-oriented architecture to support best practices established by archivists. Further, audits, such as by the Digital Repository Audit Method Based on Risk Assessment,[42] may be conducted to assess how well repositories implement trustworthy procedures.

Preservation and provenance metadata schemes such as PREMIS[43] and PROV-O[44] are state-based ontologies that include entities such as actors, events and digital objects. They record the history of transitions (e.g. changes in format) for digital objects.

Rights Metadata

For some data, there are many advantages to open publication. The rights for that data can be specified with a Creative Commons License. For other data, there can be strong justifications for limited access, such as privacy and economic factors.

For example, although survey results are generally aggregated across individuals, individual-level data is sometimes very useful. Some repositories of survey data include microdata, that is, data for the responses that individuals gave to survey questions.[45] However, analysis of such microdata raises privacy concerns and needs to be carefully managed; access should be limited to qualified researchers. Repositories of individual health records raise similar privacy concerns.

Usage Statistics

The number of visits and downloads for a dataset can give an indication to later users about the likely value of a given dataset. Such usage data are helpful for the managers and funders of repositories to evaluate their service. Citations are indicators for how a dataset is being used and its relationship to other work.

Analysis Platforms and Decision Support Systems

There is an increasingly rich set of analytic tools. Some of the earliest tools were statistical packages such as SPSS, R, SAS and STATA. These were gradually enhanced with data visualization and other analytic software. The current generation of tools such as Jupyter,[46] RSpace, and eLab notebooks (ELN) integrate annotations, workflows, raw data, data analysis, and annotations into one environment.

Virtual research environments are typically organized by research communities to coordinate datasets with search and analytic tools. For instance, the Virtual Astronomy Observatory uses Jupyter to provide users with a robust research environment. WissKI[47] is a platform for coordinating digital humanities datasets which are based on Drupal. Decision support systems are generally focused on finding optimal solutions in a parameter space. They often draw on data warehouses though recently they have begun to incorporate feeds from unstructured data (e.g. web searches).

Most repositories support search on metadata terms. In addition, some repositories have developed their own powerful data exploration tools such as ICPSR Colectica[48] for DDI and the GSS Data Explorer.[49] The Amundsen data discovery and metadata engine[50] uses metadata elements to provide a table explorer. Potentially, interactive visualization tools such as TableLens (Rao and Card, 1994) could also be employed.

Metadata Development, Standardization, and Management

Metadata, whether for texts or datasets, needs to be complete, consistent, standardized, machine processable, and timely (Park, 2009). Metadata registries provide clear definitions and promote standardization (ISO/IEC 11179). For instance, the Marine Metadata Interoperability Ontology Registry and Repository[51] records usage of different metadata terms. A registry may interoperate with editing tools for developers (Gonçalves et al., 2019). These tools may suggest candidate metadata terms. One of the keys to the development of good metadata is the involvement of a community that cares about the results.

Data Cubes, Data Warehouses, and Data Exchanges

An organization such as a large business often has many different databases. The data in the databases will likely have different formats and definitions and can be organized in a multidimensional cube. Some of cube's cells may be well populated with data that appears across many of the databases, but there will also be sparsely populated regions and cells. Online analytical processing users can generate different views of the data by

drilling down, rolling up, and slicing and dicing across cells. To facilitate retrieval, there can be a rich pre-coordinated index for common queries. Other queries can be implemented with slower methods such as hashing or B-trees.

While many organizations now have integrated enterprise data management systems, data cubes are still useful for warehousing data and for exchanging it across organizations. For instance, the W3C Data Cube[52] standard is applied in inter-organizational projects such as EarthCube[53]. SDMX[54] enables data exchange among statistical agencies in the EU.

Production Workflows, Research Workflows and Research Objects

Entities change over time, yet many knowledge representation frameworks do not model change. To represent change, models need to represent transitions, processes, and other sequential activities. Such modelling is closer to state machines, Petri nets, process ontologies, the Unified Modeling Language (UML) or even programming languages than to traditional knowledge representation.

One way to document a research project is by saving files developed during the study (Borycz and Carroll, 2018). Data files (e.g. Excel files) are just one type of artefact from a research programme; other research objects include workflows. Workflows are a natural fit for describing research methods and analyses (Austin et al., 2017). The Taverna[55] workflow tool has been used for the MyExperiment[56] project. It provides a framework for capturing and posting Taverna and other types of research workflows and incorporates simple ontologies such as FOAF. Workflows can also be used to specify and document statistical analyses; several of the analysis platforms support them. Sequential activities in the management of repositories are often tracked with workflows. For instance, the Generic Statistical Information Model (GSIM)[57] specifies workflows for the production of datasets by statistical agencies.

Semantic Modelling and Direct Representation

Semantic models attempt to represent entities. They could support unified descriptions of functionality, transitions of complex continuants, and sequential activities (Allen, 2018). Changes in semantic models are a form of qualitative simulation. While traditional knowledge representation is usually implemented with ontologies, models which allow transitions are more like programming languages. Although semantic modelling might be implemented by process ontologies, we have focused on the use of an object-oriented programming language which supports threads to allow parallel concurrent event streams and potentially to develop a 'unified temporal map'. Such semantic simulations may be useful for modelling historical events. For instance, a community described in a newspaper may be cast into a 'community model'. These go beyond social ontology to model social mechanisms (Ylikoski, 2017).

In addition, Allen (2015, 2018) has proposed rich semantic modelling of entire research reports and datasets. Structured evidence and argumentation about claims might then be applied for the evaluation of the models. Ultimately, such 'direct representation' may replace text as the primary representation for research and scholarship.

Infrastructure

Repository Servers

Semantic representations may be implemented with triplestores. Triplestores facilitate logical inference, but retrieval may be more efficient with relational databases. Many metadata catalogues are implemented with relational databases. Thus, they use SQL and are often characterized by UML class diagrams. Information models (e.g., National Information Exchange Model[58]) which could be used for metadata registries may be implemented as data dictionaries.

Some repositories are federated with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[59]. This allows the 'harvesting' of metadata from separate repositories. OAI-PMH is increasingly used as an API to allow external users to query and interact with the federated set of metadata.

Cloud Computing

We are well into the era of cloud computing (Foster and Gannon, 2017), allowing flexible allocation of computing, networking and storage resources, which facilitates software as a service. The compatibility of the versions of software packages needed for data management is often a challenge. Containers, such as those from Docker, allow compatible versions of software to be assembled and run on a virtual computer. A cloud-based virtual machine can hold datasets, workflows, and the programs used to analyse the data, which can be a complete digital preservation package.

Highly networked data centres facilitate the Internet of Things which generates massive and dynamic data. Increasingly, cloud computing is supporting edge computing and append-only stores which can capture streaming data. These technologies will provide the foundation for smart cities and have implications for the kinds of questions we may ask about social behaviour.

Conclusion

Many datasets, especially legacy datasets, are difficult to find and access. Some of the biggest issues for the retrieval of datasets concern information organization, which helps to provide context. Metadata supports the discovery and access to datasets.

More attention to metadata would also further support evidence-based policy. We need richer, more systematic, and more interoperable metadata standards. We need to improve the metadata associated with existing datasets. And we need to aggressively

upgrade the application of high-quality metadata and knowledge organization systems to datasets as they are created.

Acknowledgements

Julia Lane and members of NYU's Center for Urban Science and Progress provided useful advice and comments.

References

Allen, R. B. (2015) Repositories with direct representation. Preprint, arXiv:1512.09070.

Allen, R. B. (2018) Issues for using semantic modeling to represent mechanisms. Preprint, arXiv:1812.11431.

Allen, R. B. and Kim, Y. H. (2017/2018) Semantic modeling with foundries. Preprint, arXiv:1801.00725.

Arp, R., Smith, B. and Spear, A.D. (2015) *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: MIT Press.

Austin, C. C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V. K., Murphy, F., Nurnberger, A., et al. (2017). Key components of data publishing: Using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries*, 18(2) 77–92. doi: 10.1007/s00799–016–0178–2

Borycz, J. and Carroll, B. (2018) Managing digital research objects in an expanding science ecosystem: 2017 conference summary. *Data Science Journal*, 17. http://doi.org/10.5334/dsj–2018–016

Commission on Evidence-Based Policymaking (2017) *The Promise of Evidence-Based Policymaking*. https://www.cep.gov/cep-final-report.html

Foster, I. and Gannon, D. B. (2017) *Cloud Computing for Science and Engineering*. Cambridge, MA: MIT Press.

Gonçalves, R. S., O'Connor, M. J., Martínez-Romero, M., Egyedi, A. L., Willrett, D., Graybeal, J. and Musen, M. A. (2019) The CEDAR workbench: An ontology-assisted environment for authoring metadata that describe scientific experiments. Preprint, arXiv:1905.06480

InterPARES2 Project (2008) A framework of principles for the development of policies, strategies and standards for the long-term preservation of digital records.

Lane, J. (2016) Big data for public policy: The quadruple helix, *Journal of Policy Analysis and Management*, 35(3). doi: 10.1002/pam.21921

Lee, C.A. (2010) Open Archival Information System (OAIS) reference model. In M. J. Bates and M. N. Maack (eds), *Encyclopedia of Library and Information Sciences* (3rd edition). Boca Raton, FL: CRC Press.

Park, J.-R. (2009) Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47, 213–228. doi: 10.1080/01639370902737240

Rao, R. and Card, S. K. (1994) The Table Lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In *CHI '94: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM, pp. 318–322. doi: 10.1145/191666.191776

Riley, J. (2017) *Understanding Metadata: What Is Metadata, and What Is It For?: A Primer*. Bethesda, MD: NISO Press.

Rücknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D. et al. (2015) Metadata schema for the description of research data repositories: version 3.0. doi: 10.2312/re3.008

Vardigan, M., Heus, P. and Thomas, W. (2008) Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, 3(1). doi: 10.2218/ijdc.v3i1.45

Whyte, A. and Wilson, A. (2010) *How to Appraise and Select Research Data for Curation*. Edinburgh: Digital Curation Centre.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data*, 3, 160018. doi: 10.1038/sdata.2016.18

Ylikoski, P. (2017) Social mechanisms. In S. Glennan and P. Illari (eds), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge.

Notes

1 The FAIR guidelines have been extended from scholarly texts to datasets (Wilkinson et al., 2016).
2 https://www.ncbi.nlm.nih.gov/pmc/
3 https://www.aclweb.org/anthology/
4 https://github.com/allenai/science-parse
5 https://manpages.debian.org/testing/poppler-utils
6 https://github.com/explosion/spaCy
7 https://wiki.dbpedia.org/services-resources/ontology
8 https://en.wikipedia.org/wiki/Category:Statistical_methods
9 https://spacy.io/api/tokenizer
10 https://rasa.com/docs/nlu
11 https://nlp.stanford.edu/projects/glove
12 https://fasttext.cc/docs/en/crawl-vectors.html
13 https://www.w3.org/TR/shacl/
14 https://joinup.ec.europa.eu/release/dcat-ap/11
15 This differs from library or archival collections, which are usually thematically related, and for which the selection of items for inclusion is defined by an express collection policy.
16 The challenges of metadata for data streams are related to the cataloguing of different editions of a work and of serials in a text-based library.
17 https://wikidata.org/
18 https://duraspace.org/vivo/about/
19 https://ddialliance.org/Specification/RDF/XKOS
20 http://www.obofoundry.org/
21 https://www.icpsr.umich.edu/
22 http://ddialliance.org
23 DDI is also used for datasets from other organizations such as the National Opinion Research Center (NORC).
24 https://www.cessda.eu/
25 https://www.europeansocialsurvey.org/data/
26 https://www.ands.org.au/
27 There are additional collections at http://data.census.gov, http://gss.norc.org, http://electionstudies.org, http://psidonline.isr.umich.edu, and http://www.nlsinfo.org
28 https://www.earthcube.org/

29 https://pds.nasa.gov/
30 https://www.uniprot.org/
31 http://www.rcsb.org/
32 re3data.org
33 https://datadryad.org/
34 https://datacite.org/
35 https://www.crossref.org/
36 https://www.hl7.org/
37 https://www.genome.jp/kegg/
38 See also, for example, http://www.dcc.ac.uk/digital-curation/planning-preservation
39 https://clockss.org/
40 http://irods.org
41 The policies are based on the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) standard; see https://interparestrust.org/
42 http://www.dcc.ac.uk/sites/default/files/DRAMBORA_Interactive_Manual%5B1%5D.pdf; see also http://www.dcc.ac.uk/resources/repository-audit-and-assessment/drambora
43 http://www.loc.gov/standards/premis/ontology/
44 https://www.w3.org/TR/prov-o/
45 The term microdata is used in two distinct ways. In the context of HTML, it is associated with embedding Schema.org codes into web pages similar to micro-formats. In the context of survey data, it refers to individual-level data.
46 https://jupyter.org/
47 http://wiss-ki.eu
48 https://www.colectica.com/
49 https://gssdataexplorer.norc.org/
50 https://eng.lyft.com/amundsen-lyfts-data-discovery-metadata-engine–62d27254fbb9
51 https://mmisw.org/
52 https://www.w3.org/TR/vocab-data-cube/
53 https://www.earthcube.org/info/about
54 http://sdmx.org/
55 https://taverna.incubator.apache.org/
56 https://www.myexperiment.org/about
57 https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model. GSIM is coordinated with the Common Statistical Production Architecture; see https://unstats.un.org/unsd/nationalaccount/workshops/2015/gabon/BD/CSPA-ENG.pdf
58 https://www.niem.gov/about-niem
59 https://www.openarchives.org/pmh/