Metadata for Administrative and Social Science Data

Robert B Allen

[0000-0002-4059-2587]

rba@boballen.info

Data are valuable but finding the right data is often difficult. This chapter reviews current approaches and issues for metadata about numeric data and data sets that may facilitate the identification of relevant data. In addition, the chapter reviews how metadata support repositories, portals, and services. There are emerging metadata standards but they are applied unevenly so that there is no comprehensive approach. There has been greater emphasis on structural issues than on semantic descriptions.

# INTRODUCTION

Evidence-based policy needs relevant data (Commission on Evidence-Based Policy, 2018; Lane, 2016). Such data is valuable, but often difficult to find and/or replicate. The FAIR Open Access guidelines suggest that, ideally, data should be Findable, Accessible, Interoperable, and Reusable.[[1]](#footnote-1) Therefore, data curation and stewardship is needed.

While data may include text, image, or video, here we focus on numeric observations recorded and maintained in machine-readable form. There are many data sets of such observations available online; the DataCite[[2]](#footnote-2) repository alone contains over five million. There are many different types of data sets. Data sets differ in their structure, their source, and their use. 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 data sets. Following the approach of W3C-DCAT (World Wide Web Consortium-Data Catalog Vocabulary)[[3]](#footnote-3), a data set may be a collection of related observations which is developed and managed by a single entity such as a statistical agency. When stored as a unit online, the data set is a digital object.

Metadata consists of short descriptors which refer to a digital object. Metadata can support users in finding data sets, and enable users to know what is in them. 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. Structural metadata includes the organization of the files. Administrative metadata describes the permissions, rights, and usage. Descriptive metadata covers the contents.

This chapter surveys the state of the art of metadata for numeric data sets, 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 records, hospital records, educational records, and business records. By comparison, social science data generally is collected for the purpose of developing or applying theory.

Section 2 describes data, metadata, and digital objects. Section 3 discusses semantics. Section 4 considers repositories. Section 5 describes services. Section 6 describes the techniques for documenting the internal structure of data sets. Section 7 discusses cyberinfrastructure.

# DATA, METADATA, AND DIGITAL DATA OBJECTS

A metadata element describes some attribute of a digital object. The simplest metadata identifies the digital object.[[4]](#footnote-4) Individual metadata elements are generally part of a set which describes attributes of a data set. Such a set of metadata elements can be structured as a catalog, schema, or frame, and restrictions can be placed on the values allowed for the individual elements. A fragment of an example of the Schema.org[[5]](#footnote-5) dataset schema is shown in Figure 1. Note the distinct metadata elements in that fragment.

Figure 1: Fragment of schema.org dataset schema[[6]](#footnote-6).

{

"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\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/",

"title": "Open Data Commons Public Domain Dedication and License v1.0"

}

],

The W3C DCAT[[7]](#footnote-7) is a schema standard for data sets that is used by many repositories such as data.gov. Other structured frameworks for data sets include the DataCite[[8]](#footnote-8) metadata schema and the Inter-university Consortium for Political and Social Research Data Documentation Initiative (ICPSR DDI) discussed below (Section 4.1). The catalog specifications provide a flexible framework. For instance, DCAT allows the inclusion of metadata elements drawn from domain schema and ontologies. Some of these domain schemas are widely used resources which DCAT refers to as assets. For instance, spatial relationships are often modeled by the Federal Geographic Data Committee (FGDC) standard.[[9]](#footnote-9)

Many of the implementations for indexing collections of metadata schemas use relational databases. Thus, they use SQL and support tools such as data dictionaries. Moreover, they are often characterized by Unified Modeling Language (UML) Class Diagrams which are common for data modeling.

# SEMANTIC DESCRIPTIONS

Semantic data models have become widely explored. In particular, the Semantic Web utilizes nodes which are implemented with XML. RDF (Resource Description Framework) extends XML by requiring triples which assert a relationship (property) between two identifiers: “identifier”-“property”-“identifier”. By connecting triples, RDF can define a graph network of the relationships among a set of controlled vocabulary terms; this is the essence of linked data.

RDFS (RDF Schema) extends RDF by supporting class/subclass relationships. The classes allowed for identifiers in RDFS triples are controlled by domain and range parameters. Traditional thesauri have a simple hierarchical structure, the Simple Knowledge Organization System (SKOS) is an RDFS standard for a machine-processable representation of thesauri. 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. In addition, a knowledge graph is a model of a domain, sometimes including instances, which may be implemented in SKOS. For example, DBpedia[[10]](#footnote-10) is a knowledge graph based on Wikipedia.

Some frameworks for structural descriptions of data sets include aspects of ontologies. For example, less formal ontologies simply provide definitions and employ RDFS. Similarly, Schema.org schemas can be used with micro-formats which match schema elements with passages in an online text. Schema.org has a classification of topics and may incorporate other systems such as FOAF (Friend of a Friend) which includes attributes associated with people. Formal ontologies use OWL (Web Ontology Language) to add features to RDFS. These features lend themselves to logical inference provided that the entities and relationships are rigorously defined.

Upper ontologies provide top-down structures for the types of entities allowed in derivative domain and application ontologies. One of the best known upper ontologies is the Basic Formal Ontology (BFO) (Arp, Smith, & Spear, 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 biological ontologies have been developed based on BFO and are collected in the Open Biomedical Ontology (OBO) Foundry.

There are fewer rich ontologies dealing with social science content than there are for natural science. One challenge is “social ontology”, that is, developing definitions for social terms. It is difficult to define exactly what is a family, a crime, or money. In most cases, an operational definition or an approximate definition may suffice where structured documentation of the definitions are unavailable. Moreover, while social terms are especially difficult to define for vernacular speech, it seems possible to make clear, though perhaps cumbersome, definitions for scholarly applications.

# DATA REPOSITORIES

A data repository holds data sets and related digital objects. It provides access to the data sets and supports search. Metadata is integral to these services at several levels. In addition to item-level metadata for the data sets, there can also be study-level metadata or collection-level metadata.

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

ICPSR[[11]](#footnote-11) is a major repository of public-use social science and administrative data sets derived from questionnaires and surveys. The ICPSR DDI[[12]](#footnote-12) (e.g., Vardigan, Heus, & Thomas, 2009) defines a catalog code. A notable feature is a codebook which saves the exact wording of all the questions. In addition, the ICPSR provides an index of all variable names that are used in the data sets. DDI-Lifecycle is an extension of DDI that describes the broader context in which the survey was administered as well as the details about the preservation of the file (see Section 5.3).

## Repositories of Governmental and NGO Statistical Agencies

Statistical data collection is a core function of government. Most countries have national statistical agencies. While these statistical collections often emphasize social data, they also include related indicators such as agricultural and industrial output and housing, such as Statistics New Zealand, and the Korean Social Science Data Archive (KOSSDA). European data sets are maintained in the Consortium of European Social Science Data Archives (CESSDA)[[13]](#footnote-13) and the European Social Survey.[[14]](#footnote-14) Australia has a broad data management initiative, ANDS.[[15]](#footnote-15) Many U.S. governmental data sets are collected at data.gov.[[16]](#footnote-16) In addition, there are many non-governmental and inter-governmental agencies such as the OECD, the World Bank, and the United Nations, which host data sets.

## Other Data Repositories

Many data sets are produced, curated, and used in the natural sciences such as astronomy and geosciences. Some of these data sets have highly automated data collection, elaborate archives and established curation methods. Many of these repositories include multiple data sets for which access is supported with portals or data cubes (see Section 6.1). For instance, massive amounts of geophysical data and related text documents are collected in the EarthCube[[17]](#footnote-17) portal. The Science.gov portal is established by the U.S. Office of Science Technology and Policy. NASA supports approximately 25 different data portals. Each satellite in the Earth Observation System (EOS) may provide hundreds of streams of data,[[18]](#footnote-18) with much common metadata. This provides a context analogous to study-level metadata. Likewise, there are massive genomics and proteomics data sets which are accessible via portals such as UniProt[[19]](#footnote-19) and the Protein Data Bank[[20]](#footnote-20) along with suites of tools for exploring them. Similarly, there are very large data sets from medical research such as from clinical trials and from clinical practice including Electronic Health Records (EHRs).

## Ecosystem of Texts and Data Sets

Data sets are often associated with text reports. For example, the Dryad Digital Repository[[21]](#footnote-21) hosts data sets from scholarly publications which require the deposit of data associated with scholarly papers accepted for publication. In addition, data sets may be cited in much the same way that research reports are cited. Formal citation facilitates tracing the origins of data used in analyses and helps to acknowledge the work of the creators of the data sets. Standards have been developed for such citations (Martone, 2014; Silvello, 2017).

# SERVICES

The purpose of metadata and other aspects of information organization and management is to provide services to users. Indeed, “service science” is an approach in information technology which focuses on the design and delivery of services rather than on underlying technologies.

## Search

Searching for data sets differs from the familiar web-based text search because data repositories are generally hosted by either relational databases or semantic triplestores. Even where the data are stored on separate servers the metadata can be harvested and searched. This type of federated search is supported by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH);[[22]](#footnote-22) both data.gov and ICPSR use OAI-PMH.

## From Statistical Packages to Virtual Research Environments

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, RSpace, and eLab notebooks (ELN) integrate annotations, workflows, raw data, and data analysis into one platform. In addition, some repositories have developed their own powerful data exploration tools such as ICPSR Colectica[[23]](#footnote-23) for DDI and the GSS Data Explorer[[24]](#footnote-24). Virtual research environments (VREs) are typically organized by research communities to coordinate data sets with search and analytic tools. For instance, the Virtual Astronomy Observation (VAO) uses Jupyter to provide users with a robust research environment. WissKI[[25]](#footnote-25) is a platform for coordinating digital humanities data sets which are based on Drupal.

## Preservation

Lost data is often irreplaceable. Even if the data is not entirely lost, users need confidence that the quality of stored data has not been compromised. Moreover, although data storage prices are declining dramatically, we cannot save everything and the cost of maintaining a trusted repository remains substantial. Many of these challenges are familiar from traditional archives. For instance, selection policies typical in archives could help in controlling the many poorly documented data sets in some repositories. Yet, prioritization is difficult[[26]](#footnote-26) (Whyte & Wilson, 2010).

The Open Archival Information System (OAIS) provides a reference model for the management of archives (Lee, 2010). The OAIS framework has been incorporated into the ICPSR DDI-Lifecycle model. The Integrated Rule-Oriented Data System (iRODS)[[27]](#footnote-27) is a policy-based archival management system developed for large data stores. It implements a service-oriented architecture (SOA) to support best practices established by archivists.

Provenance records the history of an entity. This can help to ensure confidence in its authenticity. For data in a repository, provenance often means tracing the history of repository operations. The history of transitions is often recorded as event data, where the events are what happened to the data in the dataset. Typically, provenance ontologies include actors, events, and digital objects. Potentially, blockchains could provide an even greater level of trust in digital provenance.

## Metadata Quality and Metadata Management

Metadata, whether for texts or data sets, needs to be complete, consistent, standardized, machine processable, and timely (Park, 2009). A metadata editor supports the assignment of quality metadata (e.g., Gonclaves, O’Conner, et al., 2019). When collections or metadata standards change, the repository librarian must revise metadata (Tonkin, 2009). This might be particularly needed when updating metadata from data streams[[28]](#footnote-28) such as those from satellite downlinks or smart-city sensors.

Although survey results are generally aggregated across individuals, individual-level data is sometimes very useful. Some repositories of survey data include micro-data, data for the responses that individuals gave to survey questions.[[29]](#footnote-29) Currently, there are no distinct metadata tags for such data; they are embedded into repository data. Moreover, the individual level of analysis raises privacy concerns and needs to be carefully managed; at the least, access should be limited to qualified researchers.

Metadata registries, such as the Marine Metadata Interoperability Ontology Registry and Repository,[[30]](#footnote-30) record usage. The Registry of Research Data Repositories (re3data registry),[[31]](#footnote-31) which is operated by DataCite, links to more than 2000 different repositories each of which holds many data sets. Each of the repositories is described by the re3data.org schema for the description of research data repositories (Rücknagel, Vierkant, et al., 2015).

Metadata application profiles provide constraints on the types of entities that can be included in the metadata for a given application. For instance, DCAT Application Profiles (DCAT-AP) support standardized metadata exchange between repositories in different jurisdictions in the EU.[[32]](#footnote-32)

# DETAILS ABOUT THE DATA IN DATA SETS

## Data Cubes

Many notable data management techniques were originally developed for managing and processing business data.[[33]](#footnote-33) One such technique is data cubes, which support access to multidimensional data. They present data as if it filled cells of a high-dimensional cube, even if the data will probably not fill all of the cells. Users can generate different views of the data by drilling-down, rolling-up, and slicing-and-dicing across cells. For complex data sets, there will be many dimensions. 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.

Data cubes have been extended beyond business information processing to cubes such as the Statistical Data and Metadata eXchange (SDMX) used in the financial services industry and the W3C Data Cube[[34]](#footnote-34) standard that is applied in projects such as EarthCube.

## Sequential Activities and Modeling Research

Entities change over time, yet knowledge representation frameworks rarely model change. In order to represent changes, models need to represent transitions, processes, and other sequential activities. Such modeling is closer to the Unified Modeling Language (UML) or even programming languages than to ontologies. In fact, modeling change is routine for state machines, Petri nets, and other transition models. A “model-layer” that allows general statements to be made about sequential activities could incorporate both ontology and transitionals (Allen & Kim, 2018).

Models of sequential activities include workflows and mechanisms (e.g., Allen, 2018). A workflow is a structure for managing a sequence of activities and is a natural fit for describing research methods and analyses (Austin, Bloom, et al., 2017). The Taverna workflow tool has been widely used in the MyExperiment[[35]](#footnote-35) project, which provides a framework for capturing and posting research methods and incorporates ontologies such as FOAF.

Allen (2015, 2018) has proposed direct representation of entire research reports. This approach uses a programming language that blends upper ontologies with object-oriented programming to do semantic modeling. Beyond modeling events, it is also possible to use structured argumentation and assertions in scientific research reports. Potentially, social mechanisms (e.g., Hedstrom & Ylikoski, 2010) and community models could be implemented. Further, highly-structured evidence and claims might be applied to the evaluation of evidence-based social policy.

# CYBERINFRASTRUCTURE

## Information Institutions and Organizations

Libraries and archives (whether traditional or digital) have the mission, and often the resources, to develop standards and maintain information over the long-term. As noted earlier (Section 5.3), preservation is the fundamental concern for archival collections. Information institutions often have formal collection management strategies, metrics, and policies. These include Web and repository metrics and analytics, usage statistics such as reports of how many downloads were made from data sets, and procedures for updates and formatting standards.

In addition to traditional information institutions, there are now many other players. These new players have slightly different mandates. For example, Schema.org’s primary mission is to provide a structure that improves indexing by search engine companies. Nonetheless, these organizations often adopt best practices similar to those of traditional information organizations.

CrossRef[[36]](#footnote-36) 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. Increasingly, the two projects are coordinating. ORCID iDs[[37]](#footnote-37) are persistent digital identifiers assigned to authors. The emergence of structured identifiers such as DOIs and ORCID iDs has allowed the development of services such as VIVO[[38]](#footnote-38) and the Microsoft Academic Graph (MAG)[[39]](#footnote-39) which allow authors to be tracked across research reports and projects, and across publishers.

## Cloud Technologies, Software as a Service, and the Internet of Things

We are now well into the era of cloud computing (Foster & Gannon, 2017), allowing flexible allocation of computing, networking and storage resources, which facilitates Software as a Service (SaaS). 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 container could hold datasets, workflows, and the programs used to analyze the data, making the analyses readily replicable. Highly-networked data centers also facilitate the Internet of Things (IoT) which will generate massive data sets such as for “smart cities”.

# CONCLUSION

Data may not be useful when stand-alone without context. Some of the biggest issues for the retrieval of information from data sets concern information organization, which help provide context. Metadata supports the discovery of and access to data sets. We need richer, more systematic, and more interoperable metadata standards. Even more attention to metadata would further support evidence-based policy.

# ACKNOWLEDGMENTS

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, *Networked Knowledge Representation Systems,* arXiv: 1512.09070

Allen, R.B. (2018). *Issues for Using Semantic Modeling to Represent Mechanisms*, arXiv:1812.11431

Allen, R.B., & Kim, YH. (2017/2018). Semantic Modeling with Foundries, arXiv:1801.00725

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

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. *In*[*ternational Journal on Digital Libraries*](https://link.springer.com/journal/799)*, 18*(2) 77–92, doi: 10.1007/s00799-016-0178-2

Commission on Evidence-Based Policymaking. (2018). The Promise of Evidence-Based Policymaking, <https://www.cep.gov/cep-final-report.html>

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

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

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, *36* 49-67.

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

Lee, C.A. (2010). Open Archival Information System (OAIS) Reference Model. *Encyclopedia of Library and Information Sciences* (3rd Edition). CRC Press, doi: 10.1081/E-ELIS3-120044377

Martone M. (ed.) (2014). *Joint Declaration of Data Citation Principles*. San Diego CA: FORCE11, Data Citation Synthesis Group: <https://www.force11.org/group/joint-declaration-data-citation-principles-final>

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

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 (29), doi: 10.2312/re3.008

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology, 69*(1) 6-20. doi: 10.1002/asi.23917

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R.R., Duerr, R., et al. (2015). Achieving human and machine accessibility of cited data in scholarly publications, *PeerJ Computer Science* 1: e1, doi 10.7717/peerj-cs.1

Tonkin, E., (2009). MetRe [**supporting the metadata revision process.**](https://research-information.bristol.ac.uk/en/publications/metre-supporting-the-metadata-revision-process(b2fa9a79-e50b-4888-b510-336aacf73da5).html) *International Conference on Digital Libraries*,

Vardigan, M., Heus,P., & 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](https://doi.org/10.2218/ijdc.v3i1.45)

Whyte, A. & Wilson, A. (2010). How to appraise and select research data for curation. *DCC How-to Guides. Edinburgh*: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides

Wilkinson, M.D., [Dumontier](https://www.nature.com/articles/sdata201618#auth-2), 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

1. The FAIR Guidelines have been extended from scholarly texts to data sets (Wilkinson, [Dumontier](https://www.nature.com/articles/sdata201618#auth-2), et al., 2018). [↑](#footnote-ref-1)
2. <https://datacite.org/> [↑](#footnote-ref-2)
3. <https://www.w3.org/TR/vocab-dcat/> [↑](#footnote-ref-3)
4. Such operators need to be unambiguous. For example, Digital Object Identifiers (DOIs, <http://doi.org>) were developed for scholarly journals and are assigned by publishers, with a part of the DOI code being a unique publisher ID. While the DOIs may identify more than one data set, version numbers distinguish the data sets. For instance, the entire GSS (General Social Survey) has only one DOI, but it is possible to drill down through the different data sets by specifying version numbers. [↑](#footnote-ref-4)
5. Schema.org is a project of a consortium of search-engine companies. The Schema.org dataset schema (<https://schema.org/Dataset>) is used by the Google Data Search. [↑](#footnote-ref-5)
6. https://github.com/datasets/gdp [↑](#footnote-ref-6)
7. https://www.w3.org/TR/vocab-dcat/ [↑](#footnote-ref-7)
8. <https://schema.datacite.org/> [↑](#footnote-ref-8)
9. <https://www.fgdc.gov/> [↑](#footnote-ref-9)
10. <https://wiki.dbpedia.org/> [↑](#footnote-ref-10)
11. <https://www.icpsr.umich.edu/icpsrweb/> [↑](#footnote-ref-11)
12. <https://www.ddialliance.org/>. Note that the Data Documentation Initiative (DDI) is different from the Data Discovery Index (DDI) associated with DataMed. [↑](#footnote-ref-12)
13. https://www.cessda.eu/ [↑](#footnote-ref-13)
14. <https://www.europeansocialsurvey.org/data/> [↑](#footnote-ref-14)
15. The Australia National Data Service, <https://www.ands.org.au/> [↑](#footnote-ref-15)
16. 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>. [↑](#footnote-ref-16)
17. <https://www.earthcube.org/> [↑](#footnote-ref-17)
18. <https://pds.nasa.gov/> [↑](#footnote-ref-18)
19. <https://www.uniprot.org/> [↑](#footnote-ref-19)
20. <http://www.rcsb.org/> [↑](#footnote-ref-20)
21. <https://datadryad.org/> [↑](#footnote-ref-21)
22. <https://www.openarchives.org/pmh/> [↑](#footnote-ref-22)
23. <https://www.colectica.com/> [↑](#footnote-ref-23)
24. <https://gssdataexplorer.norc.org/> [↑](#footnote-ref-24)
25. <http://wiss-ki.eu> [↑](#footnote-ref-25)
26. <http://www.dcc.ac.uk/> [↑](#footnote-ref-26)
27. <http://irods.org> [↑](#footnote-ref-27)
28. <http://schema.org/dataset/datastreams> [↑](#footnote-ref-28)
29. The term micro-data 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. [↑](#footnote-ref-29)
30. <https://mmisw.org/> [↑](#footnote-ref-30)
31. <https://www.re3data.org/> [↑](#footnote-ref-31)
32. <https://joinup.ec.europa.eu/release/dcat-ap/11> [↑](#footnote-ref-32)
33. E.g., Online Analytical Processing (OLAP), Enterprise Data Warehouses (EDW), and Decision Support Systems (DSS). [↑](#footnote-ref-33)
34. <https://www.w3.org/TR/vocab-data-cube/> [↑](#footnote-ref-34)
35. <https://www.myexperiment.org/home> [↑](#footnote-ref-35)
36. <https://www.crossref.org/> [↑](#footnote-ref-36)
37. <https://orcid.org/> [↑](#footnote-ref-37)
38. <https://duraspace.org/vivo/> [↑](#footnote-ref-38)
39. <https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/> [↑](#footnote-ref-39)