ISSN 2161-3974
eScience in Action

Journal of eScience Librarianship 12(3): e758
DOI: https://doi.org/10.7191/jeslib.758

Journal of eScience Librarianship
putting the pieces together: theory and practice

Towards a Shared Framework: A Classificatory Matrix for Teaching Data Standards

Kelsey Badger, The Ohio State University, Columbus, OH, USA, badger.60@osu.edu

Abstract

Standards for research data can be a mystifying topic for both researchers and data professionals. A common source of confusion is that they are multipurpose: standards can (and should) be applied to both primary data and metadata, enabling a wide range of functions from the search features in a repository to the integration of disparate data sources. This paper reviews examples of classificatory approaches used by both librarians and researchers to describe data standards. This literature is synthesized into a classificatory matrix that can be used to map different types of standards. The matrix is constructed around two organizing principles: purpose (finding or using data) and type of information controlled (meaning or syntax). The objective of this classificatory exercise is to encourage further discussion about the misunderstandings between researchers and data support professionals and to spur further development of the educational resources needed to improve understanding and use of data standards.

Received: July 17, 2023 **Accepted**: November 14, 2023 **Published**: December 20, 2023

Introduction

Data standards play a pivotal role in research, enabling discoveries to be validated, replicated, and built upon by promoting consistency and shared norms. The phrase "data standard" is used to describe a vast array of normalizing practices, including the use of specific file formats, collection instruments, practices for identifying the provenance of materials, and approaches to documentation. While this range of applications is a clear indication of the importance of data standards, it is also a barrier to educational efforts. The topic quickly becomes overwhelming for researchers and support staff alike.

Information professionals understand better than most how classification provides scaffolding that makes complex domains of knowledge easier to traverse. Yet despite our collective experience in developing such systems for knowledge organization, there is not yet a widely accepted approach to classifying research data standards. This gap directly impacts the ability of researchers and data professionals to identify relevant standards and increase their use in research. This is especially problematic as federal requirements for data management and sharing increasingly ask researchers to list the data standards used (National Institutes of Health 2020; National Science Foundation, n.d.). While from a workflow perspective it may be desirable to make standards invisible to the user (Sansone et al. 2019), in an educational context the objective is inverted. It is beneficial during data management consultation to help researchers identify and build upon their own best practices.

This paper briefly introduces key concepts relevant to research data standards, reviews existing classificatory approaches used by librarians and researchers, and finally proposes a draft classificatory matrix that is intended to instigate further discussion about standards education in the broader community of research data professionals.

Data Standards: A Crash Course

To support data sharing and re-use, it is important to implement standards for both primary data and the documentation, or metadata, that is used to contextualize it (Sansone et al. 2019; Strasser 2015). Standardizing primary data and metadata often accomplishes different goals. For example, establishing common definitions for parameters and concepts in primary data—such as through the use of an existing data collection instrument or by mapping variables to an ontology—allows for easier integration of disparate datasets. These types of standards create semantic compatibility by ensuring data "use and speak the same language" (National Institutes of Health 2018, 6). On the other hand, a metadata schema that specifies how documentation should be standardized does not guarantee that the primary data will be semantically similar enough to integrate because there may not be any variables in common. Instead, standardized metadata is important for ensuring detailed context about the original conditions and methods of the research, which is also likely to be needed in secondary analysis (Lind et al. 2011). To oversimplify, primary data standards can make it possible to integrate data, whereas metadata standards can provide the context about whether you should.

But not all data and metadata standards serve the primary function of dataset integration. Many standards—especially those for metadata—support data discovery (Joudrey and Taylor 2017a). Repository search features, like filters for keywords or geographic coverage, only work if all of the datasets in the repository include metadata for those parameters. This can be accomplished by using standard element sets for metadata, also known as schema. These are often combined with other types of metadata standards, like content standards or thesauri, which ensure that the metadata values for all of the datasets follow the same rules (Joudrey and Taylor 2017b). For example, using singular nouns for keywords instead of plural, or better yet, using a controlled list of keyword values.

What We Talk About When We Talk About Standards

Miscommunication can occur in contexts when the phrase "data standard" is used without clearly distinguishing whether the goal is to support interoperability of primary data, data discovery through metadata records, or both. When terms are not well-defined, it is understandable that different audiences will fill in the blanks based on their own experiences and areas of expertise. For example, researchers may have greater familiarity with the primary data standards used in data collection or analysis and data support professionals with library backgrounds may be more familiar with the metadata standards used in information retrieval. This seems evident in many of the educational resources created by librarians, which often acknowledge the importance of applying standards to both primary data and metadata (for example, Strasser 2015), but which subsequently provide examples only for metadata schema. Conversely, researchers may be aware of metadata as a concept, but interviews indicate that its "purpose and importance in the process of sharing, accessing, and (re)using data [seems] elusive to many" (Bishop et al. 2019, 29). A first step to a shared understanding of the landscape of data standards is to examine the classifications used by each of these communities.

Library Classifications of Metadata Standards

Unlike the situation for data standards generally, there are well-established approaches to classifying metadata standards within the community of librarians and other information professionals. They are broadly consistent across many contexts, including textbooks used in library programs (Zeng & Qin 2022), research efforts of practicing librarians (Riley 2010), and the educational materials of standards developing organizations (Riley 2017). These classifications of metadata standards use two primary organizing principles, though sources may refer to them with slightly different labels.

The first organizing principle is labeled purpose (Riley 2010) or type (Riley 2017; Zeng & Qin 2022) and refers to the information content the metadata conveys. For example, descriptive metadata provides a characterization of a resource, whereas administrative metadata provides a record of how the resource has been and can be interacted with. If the resource described were a dataset, descriptive metadata might include a title, abstract, geographic coverage, and definitions of variables. The administrative metadata could include details about the data repository and curation process and the license terms specified by the researcher.

The second organizing principle categorizes metadata standards by their function (Riley 2010; Zeng & Qin 2022) or what it is they standardize (Riley 2017). For example, a schema or element set standardizes which categories of information need to be recorded, while a controlled vocabulary specifies the values that can be used to populate those elements. These functions are often discussed in terms of either syntax or semantics, where syntax describes format or structure and semantics describes meaning.

Researcher Classifications of Data Standards

Classifications of data standards by researcher communities are not nearly as consistent or well-defined. Standards that are developed in disciplinary siloes are prone to "overlaps in scope and arbitrary decisions on wording and substructuring" that complicate efforts to classify them (Taylor et al. 2008, 889). A notable exception to this disciplinary fragmentation is the FAIRSharing project, a registry for data sharing resources that catalogs data standards, repositories, and policies from across disciplines (Sansone et al. 2019). FAIRSharing grew out of the earlier MIBBI project, which sought to standardize reporting checklists used in research data documentation (Taylor et al. 2008). Both generations of this effort recognized the need to develop cross-disciplinary language for describing existing data standards as well as common practices for developing and registering new ones. The generalizability of these projects across research disciplines are a distinguishing characteristic.

Their classificatory groupings, shown in Table 1, have remained mostly consistent over the last decade, except for the addition to FAIRSharing of 'identifier schemata.'

Table 1: Cross-disciplinary classificatory groups for research data standards

MIBBI Project	FAIRSharing Registry
Minimum information checklists or guidelines	Reporting guidelines
Controlled vocabularies and ontologies (semantics)	Terminology artifacts
Formats (syntax)	Models/Formats
—	Identifier Schemata

Source: Taylor et al. 2008; Sansone et al. 2019

Unlike the examples drawn from libraries, these projects do not explicitly discuss the organizing principles that led to their groupings. Despite this, there are clues to how they were conceived. As in the examples from libraries, both classifications contain groups that distinguish the functions of regularizing syntax and semantics, with syntax referring to formats or models and semantics referring to terminology.

Both sources also provide short lists related to the purpose of standards, with MIBBI highlighting the "regularization of data capture, representation, annotation or reporting" (Taylor et al. 2008, 890) and FAIRSharing focused on the "identification, citation and reporting of data and metadata" (Sansone et al. 2019, 359). While reporting is an important commonality between the two initiatives, their stated purposes otherwise seem quite different. As its name suggests, the FAIRSharing registry's articulation of goals is influenced by the popularization of the FAIR Guiding Principles for research data (Wilkinson et al. 2016), which emphasize the use of standards for making data easier to both find and reuse. MIBBI, however, began nearly a decade prior to the popularization of FAIR and focuses on the use of standards for documenting the context of data, with only minimal references to the concept of repositories and data discovery.

Towards a Shared Framework

This exploration of classificatory approaches grew out of the desire for a teaching tool that would quickly convey the breadth of data standards while specifically addressing the need to differentiate standards that support data discovery and reuse. The library-derived and researcher-derived classifications both offer approaches for reducing the range of standards into concise groupings, but their value as instructional objects are undermined by the extensive use of jargon that may not be immediately understood. Moreover, neither domain fully articulates what the use of each type of standard enables. Despite frequent statements that standards should be applied to both data and metadata, they do not characterize what that looks like in practice—which standards are applied to each and to what end?

Determining a clear way to incorporate this concept into a teaching-friendly approach to classification required iteration. My first attempt, inspired by the educational materials of the U.S. Geological Survey (2021), was to present the classificatory groupings of standards as either dataset-level or parameter-level. I taught that a dataset-level standard applied to the entire digital object of the dataset, while a parameter-level standard applied to observations within that dataset. For example, this allowed for the introduction of dataset-level metadata that is used in information retrieval, but without the jargon and preconceived notions that may adhere to the word schema. However, it quickly became apparent to me that many standards could be applied at both the dataset-level and the parameter-level. For example, ISO 8601, an international standard for the display of dates, can be appropriate for formatting the values of dates in both a metadata record and in the observations of a primary data file.

This temporary setback resulted in two changes to my strategy. First, I realized that the distinction between the standards enabling data discovery and data reuse would be better represented as a continuous variable than as discrete categories. Second, I turned my focus from the classificatory groupings in the literature to the organizing principles that had produced them. The concept of function—the regularizing of syntax or semantics—is an organizing principle that is already used similarly within the library-derived and researcher-derived classifications. The concept of purpose is not so clearly defined by either community. However, the researcher-derived classifications seemed to hint at its suitability for addressing data discovery and data reuse as goals of standardizing research data. Drawing on these organizing principles, I drafted a classificatory matrix (Figure 1) to add context to the jargon-laden classificatory groups from the literature. To better distinguish between the meaning of purpose and function as organizing principles, function was relabeled as type of information controlled.

The y-axis of the figure represents type of control, ranging from meaning (or semantics) to syntax. The x-axis represents purpose, ranging from finding data to (re)using data. Plotted on the figure are examples of the many types of data standards, from common data elements to ontologies to ISO formatting standards. These were placed on the figure by reviewing examples within each category and approximating the extent to which they contributed to each of the four concepts. For example, my earlier observation that ISO 8601 could be used to format dates of both metadata records and primary data observations led to its placement at the midpoint of the find-use axis. It is placed at the syntax end of the meaning-syntax axis because a

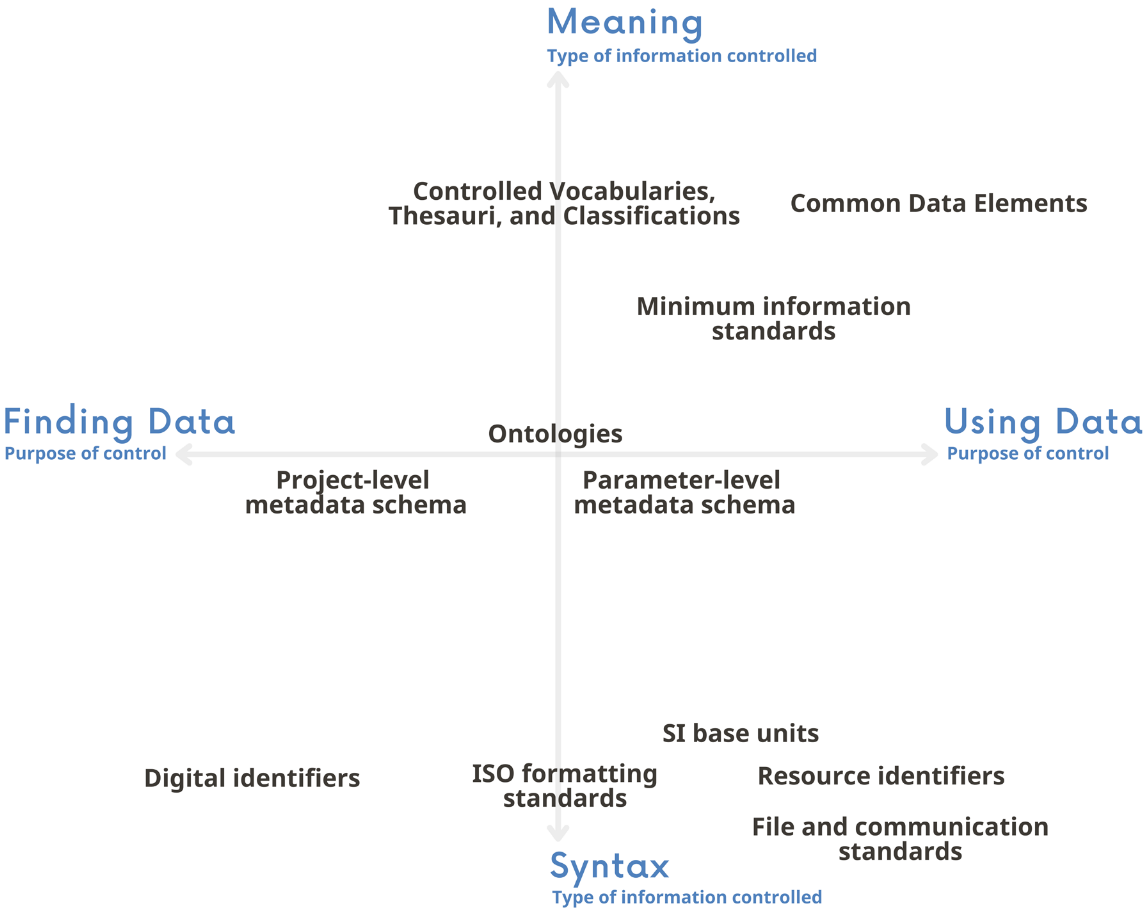
Meaning
Type of information controlled
Controlled Vocabularies, Thesauri, and Classifications
Common Data Elements
Minimum information standards
Finding Data
Purpose of control
Ontologies
Using Data
Purpose of control
Project-level metadata schema
Parameter-level metadata schema
SI base units
Digital identifiers
ISO formatting standards
Resource identifiers
File and communication standards
Syntax
Type of information controlled

Figure 1: Classificatory matrix for research data standards.

consistent format does not on its own provide any insight into the meaning or context of a date, such as whether it designates the start or the end of an observation.

While each standard's placement on the figure was carefully considered, the exact plotting of the points remains subjective. For now, the matrix serves less as a formal knowledge organization schema, and more as a way to encourage conversation around the instruction of data standards. Indeed—if the placement of a standard encourages debate, it will have fulfilled its purpose. It may also be useful as a lightweight scaffolding for designing instructional activities with researchers. For example, a librarian and a research team could use a blank matrix as an outline to brainstorm relevant community standards for a specific discipline.

Conclusion: A Call to Action

The classificatory matrix presented is undoubtedly only one of many ways to represent the range of research data standards. Moreover, it has been designed with the intent to address a particular point of confusion: the distinction between standards that facilitate data discovery and reuse. As research data infrastructure continues to mature and requirements for data sharing gradually become the norm across research disciplines, perhaps this point of emphasis will no longer be so crucial. For now, the matrix is intended as a call to action for research support professionals. When you teach data standards, what latent assumptions do you and your audience bring to the conversation? What frameworks, terminologies, or tools are needed for more comprehensive education in this area?

There is an inherent tension between theory and practice, between comprehensively describing the landscape of data standards and applying specific standards in practice. As the idiom goes, there is a risk that if we focus on the "forest," researchers will not see the "trees." However, it is no less of a gamble to send researchers to a catalog of data standards or a list of disciplinary metadata schema without contextualizing the information they will find there. As always, our theory and practice must plod ahead in tandem.

Acknowledgements

The content of this article is based on the poster entitled "Data Standards 101" originally presented at RDAP Summit 2023, available from Open Science Framework: https://osf.io/avj7b.

Competing Interests

The author declares that they have no competing interests.

References

Bishop, Bradley Wade, Carolyn Hank, Joel Webster, and Rebecca Howard. 2019. "Scientists' Data Discovery and Reuse Behavior: (Meta)Data Fitness for Use and the FAIR Data Principles." *Proceedings of the Association for Information Science and Technology* 56(1): 21–31. https://doi.org/10.1002/pra2.4.

Holdren, John P. 2013. “Increasing Access to the Results of Federally Funded Scientific Research.” Official Memorandum. Washington, DC: White House Office of Science and Technology Policy. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Joudrey, Daniel N., and Arlene G. Taylor. 2017a. “Introduction to Metadata.” In *The Organization of Information*. 4th ed. 181–247. Englewood, CO: Libraries Unlimited.

———. 2017b. “Systems for Vocabulary Control.” In *The Organization of Information*. 4th ed. 475–537. Englewood, CO: Libraries Unlimited.

Lind, Eric, Steve Aulenback, and Tom Burley. 2011. “Best Practice: Consider the compatibility of the data you are integrating.” DataONE Data Management Skillbuilding Hub. Accessed September 29, 2023. https://dataoneorg.github.io/Education/bestpractices/consider-the-compatibility.

National Institutes of Health. 2018. “NIH Strategic Plan for Data Science.” Office of Data Science Strategy. https://datascience.nih.gov/nih-strategic-plan-data-science.

———. 2020. “Supplemental Information to the NIH Policy for Data Management and Sharing: Elements of an NIH Data Management and Sharing Plan.” NOT-OD-21-014. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-014.html.

National Science Foundation. n.d. “Preparing Your Data Management Plan.” Funding at NSF. Accessed July 14, 2023. https://new.nsf.gov/funding/data-management-plan.

Riley, Jenn. 2010. *Seeing Standards: A Visualization of the Metadata Universe*. https://jennriley.com/metadatamap.

———. 2017. *Understanding Metadata: What Is Metadata, and What Is It For?* National Information Standards Organization. http://www.niso.org/publications/understanding-metadata-riley.

Sansone, Susanna-Assunta, Peter McQuilton, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Massimiliano Izzo, Allyson L. Lister, and Milo Thurston. 2019. “FAIRsharing as a Community Approach to Standards, Repositories and Policies.” *Nature Biotechnology* 37(4): 358–367. https://doi.org/10.1038/s41587-019-0080-8.

Strasser, Carly. 2015. Research Data Management: A Primer. National Information Standards Organization. https://www.niso.org/publications/primer-research-data-management.

Taylor, Chris F., Dawn Field, Susanna-Assunta Sansone, Jan Aerts, Rolf Apweiler, Michael Ashburner, Catherine A. Ball, et al. 2008. “Promoting Coherent Minimum Reporting Guidelines for Biological and Biomedical Investigations: The MIBBI Project.” *Nature Biotechnology* 26(8): 889–896. https://doi.org/10.1038/nbt.1411.

U.S. Geological Survey. 2021. “Data Standards.” USGS Data Management Website. Last updated May 3, 2021. https://doi.org/10.5066/F7MW2G15.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3(1): 160018. https://doi.org/10.1038/sdata.2016.18.

Zeng, Marcia Lei and Jian Qin. 2022. *Metadata*. 3rd ed. Chicago: ALA Neal-Schuman.