Towards an open-source model for data and metadata standards

Ariel Rokem Vani Mandava

1 Under construction

Please excuse our dust while we work on this report, which is currently under heavy construction.

2 Abstract

Recent progress in machine learning and artificial intelligence promises to advance research and understanding across a wide range of fields and activities. In tandem, increased awareness of the importance of open data for reproducibility and scientific transparency is making inroads in fields that have not traditionally produced large publicly available datasets. Data sharing requirements from publishers and funders, as well as from other stakeholders, have also created pressure to make datasets with research and/or public interest value available through digital repositories. However, to make the best use of existing data, and facilitate the creation of useful future datasets, robust, interoperable and usable standards need to evolve and adapt over time. The open-source development model provides significant potential benefits to the process of standard creation and adaptation. In particular, the development and adaptation of standards can use long-standing socio-technical processes that have been key to managing the development of software, and allow incorporating broad community input into the formulation of these standards. By adhering to open-source standards to formal descriptions (e.g., by implementing schemata for standard specification, and/or by implementing automated standard validation), processes such as automated testing and continuous integration, which have been important in the development of open-source software, can be adopted in defining data and metadata standards as well. Similarly, open-source governance provides a range of stakeholders a voice in the development of standards, potentially enabling use cases and concerns that would not be taken into account in a top-down model of standards development. On the other hand, open-source models carry unique risks that need to be incorporated into the process.

3 Introduction

Data-intensive discovery has become an important mode of knowledge production across many research fields and it is having a significant and broad impact across all of society. This is becoming increasingly salient as recent developments in machine learning and artificial intelligence (AI) promise to increase the value of large, multi-dimensional, heterogeneous data sources. Coupled with these new machine learning techniques, these datasets can help us understand everything from the cellular operations of the human body, through business transactions on the internet, to the structure and history of the universe. However, the development of new machine learning methods and data-intensive discovery more generally depends on Findability, Accessibility, Interoperability and Reusability (FAIR) of data (Wilkinson et al. 2016). One of the main mechanisms through which the FAIR principles are promoted is the development of *standards* for data and metadata. Standards can vary in the level of detail and scope, and encompass such things as *file formats* for the storage of certain data types, *schemas* for databases that organize data, *ontologies* to describe and organize metadata in a manner that connects it to field-specific meaning, as well as mechanisms to describe *provenance* of analysis products.

Community-driven development of robust, adaptable and useful standards draws significant inspiration from the development of open-source software (OSS) and has many parallels and overlaps with OSS development. OSS has a long history going back to the development of the Unix operating system in the late 1960s. Over the time since its inception, the large community of developers and users of OSS have developed a host of socio-technical mechanisms that support the development and use of OSS. For example, the Open Source Initiative (OSI), a non-profit organization that was founded in the 1990s developed a set of guidelines for licensing of OSS that is designed to protect the rights of developers and users. On the more technical side, tools such as the Git Source-code management system support open-source development workflows that can be adopted in the development of standards. Governance approaches have been honed to address the challenges of managing a range of stakeholder interests and to mediate between large numbers of weakly-connected individuals that contribute to OSS. When these social and technical innovations are put together they enable a host of positive defining features of OSS, such as transparency, collaboration, and decentralization. These features allow OSS to have a remarkable level of dynamism and productivity, while also retaining the ability of a variety of stakeholders to guide the evolution of the software to take their needs and interests into account.

The present report seeks to explore how OSS processes and tools have affected the development of data and metadata standards. The report will triangulate common features of a variety of use cases; it will identify some of the challenges and pitfalls of this mode of standards development; and it will make recommendations for future developments and policies that can help this mode of standards development thrive and reach its full potential.

4 Opportunities and risks for open-source standards

Data and metadata standards that adopt tools and practices of OSS ("open-source standards" henceforth) stand to reap many of the benefits that the OSS model has provided in the development of other technologies. At the same time, these tools and practices are associated with risks that need to be mitigated.

4.1 Flexibility vs. stability

One of the defining characteristics of OSS is its dynamism and its rapid evolution. Because OSS can be used by anyone and, in most cases, contributions can be made by anyone, innovations flow into OSS in a bottom-up fashion from user/developers. Pathways to contribution by members of the community are often well-defined: both from the technical perspective (e.g., through a pull request on GitHub, or other similar mechanisms), as well as from the social perspective (e.g., whether contributors need to accept certain licensing conditions through a contributor licensing agreement) and the socio-technical perspective (e.g., how many people need to review a contribution, what are the timelines for a contribution to be reviewed and accepted, what are the release cycles of the software that make the contribution available to a broader community of users, etc.). Similarly, open-source standards may also find themselves addressing use cases and solutions that were not originally envisioned through bottom-up contributions of members of a research community to which the standard pertains. However, while this dynamism provides an avenue for flexibility it also presents a source of tension. This is because data and metadata standards apply to already existing datasets, and changes may affect the compliance of these existing datasets.

4.2 Mismatches between standards developers and user communities

There is an inherent gap in both interest and ability to engage with the technical details undergirding standards and their development between the core developers of the standard and their users. In extreme cases, these interests may even be at odds, as developers implement sophisticated mechanisms to automate the creation of the standard or advocate for more technically advanced mechanisms for evolving the standard, leaving potential users sidelined in the development of the standard, and limiting their ability to provide feedback about the practical implications of changes to the standards.

4.3 Unclear pathways for standards success

Standards typically develop organically through sustained and persistent efforts from dedicated groups of data practitioneers. These include scientists and the broader ecosystem of data curators and users. However there is no playbook on the structure and components of a data

standard, or the pathway that moves a data implementation to a data standard. As a result, data standardization lacks formal avenues for research grants.

4.4 Cross domain funding gaps

Data standardization investment is justified if the standard is generalizable beyond any specific science domain. However while the use cases are domain sciences based, data standardization is seen as a data infrastructure and not a science investment. Moreover due to how science research funding works, scientists lack incentives to work across domains, or work on infrastructure problems.

4.5 Data instrumentation issues

Data for scientific observations are often generated by proprietary instrumentation due to commercialization or other profit driven incentives. There islack of regulatory oversight to adhere to available standards or evolve Significant data transformation is required to get data to a state that is amenable to standards, if available. If not available, there is lack of incentive to set aside investment or resources to invest in establishing data standards.

4.6 Sustainability

4.7 The importance of automated validation

5 Use cases

To understand how OSS development practices affect the development of data and metadata standards, it is informative to demonstrate this cross-fertilization through a few use cases. As we will see in these examples some fields, such as astronomy, high-energy physics and earth sciences have a relatively long history of shared data resources from organizations such as LSST and CERN, while other fields have only relatively recently become aware of the value of data sharing and its impact. These disparate histories inform how standards have evolved and how OSS practices have pervaded their development.

5.1 Astronomy

One prominent example of a community-driven standard is the FITS (Flexible Image Transport System) file format standard, which was developed in the late 1970s and early 1980s (Wells and Greisen 1979), and has been adopted worldwide for astronomy data preservation and exchange. Essentially every software platform used in astronomy reads and writes the FITS

format. It was developed by observatories in the 1980s to store image data in the visible and x-ray spectrum. It has been endorsed by IAU, as well as funding agencies. Though the format has evolved over time, "once FITS, always FITS". That is, the format cannot be evolved to introduce changes that break backwards compatibility. Among the features that make FITS so durable is that it was designed originally to have a very restricted metadata schema. That is, FITS records were designed to be the lowest common denominator of word lengths in computer systems at the time. However, while FITS is compact, its ability to encode the coordinate frame and pixels, means that data from different observational instruments can be stored in this format and relationships between data from different instruments can be related, rendering manual and error-prone procedures for conforming images obsolete.

5.2 High-energy physics (HEP)

Because data collection is centralized, standards to collect and store HEP data have been established and the adoption of these standards in data analysis has high penetration (Basaglia et al. 2023). A top-down approach is taken so that within every large collaboration standards are enforced, and this adoption is centrally managed. Access to raw data is essentially impossible, and making it publicly available is both technically very hard and potentially ill-advised. Therefore, analysis tools are tuned specifically to the standards. Incentives to use the standards are provided by funders that require data management plans that specify how the data is shared.

5.3 Neuroscience

In contrast to astronomy and HEP, Neuroscience has traditionally been a "cottage industry", where individual labs have generated experimental data designed to answer specific experimental questions. While this model still exists, the field has also seen the emergence of new modes of data production that focus on generating large shared datasets designed to answer many different questions, more akin to the data generated in large astronomy data collection efforts (Koch and Clay Reid 2012). This change has been brought on through a combination of technical advances in data acquisition techniques, which now generate large and very high-dimensional/information-rich datasets, cultural changes, which have ushered in new norms of transparency and reproducibility, and funding initiatives that have encouraged this kind of data collection. However, because these changes are recent relative to the other cases mentioned above, standards for data and metadata in neuroscience have been prone to adopt many elements of modern OSS development. Two salient examples in neuroscience are the Neurodata Without Borders file format for neurophysiology data (Rübel et al. 2022) and the Brain Imaging Data Structure (BIDS) standard for neuroimaging data (Gorgolewski et al. 2016). BIDS in particular owes some of its success to the adoption of OSS development mechanisms (Poldrack et al. 2024). For example, small changes to the standard are managed through the GitHub pull request mechanism; larger changes are managed through a a BIDS Enhancement

Proposal (BEP) process that is directly inspired by the Python programming language community's Python Enhancement Proposal procedure, which used to introduce new ideas into the language. Though the BEP mechanism takes a slightly different technical approach, it tries to emulate the open-ended and community-driven aspects of Python development to accept contributions from a wide range of stakeholders and tap a broad base of expertise.

5.4 Automated discovery

5.5 Citizen science

6 Cross-sector interactions

The importance of standards stems not only from discussions within research fields about how research can best be conducted to take advantage of existing and growing datasets, but also arises from interactions with other sectors. Several different kinds of cross-sector interactions can be defined as having important impact on the development of open-source standards.

6.1 Governmental policy-setting

The development of open practices in research has entailed an ongoing interaction and dialogue with various governmental bodies that set policies for research. For example, for research that is funded by the public, this entails an ongoing series of policy discussions that address the interactions between research communities and the general public. One way in which this manifests in the United States specifically is in memos issued by the directors of the White House Office of Science and Technology Policy (OSTP), James Holdren (in 1) and Alondra Nelson (in 2022). While these memos focused primarily on making peer-reviewed publications funded by the US Federal government available to the general public, they also lay an increasingly detailed path toward the publication and general availability of the data that is collected in research that is funded by the US government. The general guidance and overall spirit of these memos dovetail with more specific policy guidance related to data and metadata standards. For example, the importance of standards was underscored in a recent report by the Subcommittee on Open Science of the National Science and Technology Council on the "Desirable characteristics of data repositories for federally funded research" (The National Science and Technology Council 2022). The report explicitly called out the importance of "allow[ing] datasets and metadata to be accessed, downloaded, or exported from the repository in widely used, preferably non-proprietary, formats consistent with standards used in the disciplines the repository serves." This highlights the need for data and metadata standards across a variety of different kinds of data. In addition, a report from the National Institute of Standards and Technology on "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools" emphasized that – specifically for the case of AI – "U.S. government agencies should prioritize AI standards efforts that are [...]

Consensus-based, [...] Inclusive and accessible, [...] Multi-path, [...] Open and transparent, [...] and [that] result in globally relevant and non-discriminatory standards..." (National Institute of Standards and Technology 2019). The converging characteristics of standards that arise from these reports suggest that considerable thought needs to be given to how standards arise so that these goals are achieved.

A compelling road map towards implementation and adoption of community-developed standards is offered in a blog post authored by the Center for Open Science's Brian Nosek, entitled "Strategy for Culture Change" (Nosek, n.d.). The core idea is that affecting a turn toward open science requires an alignment of not only incentives and values, but also technical infrastructure and user experience. A sociotechnical bridge between these pieces, which make adoption of standards possible, and maybe even easy, and the policy goals, arises from a community of practice that makes adoption of standards normative. Once all of these pieces are in place, making adoption of open science standards required becomes more straightforward and less onerous.

6.2 Funding

While government-set policy is primarily directed towards research that is funded through governmental funding agencies, there are other ways in which funding relates to the development of open-source standards. One way is in funding the development of these standards. For example, the National Institutes of Health have provided some of the funding for the development of the Brain Imaging Data Structure standard in neuroscience.

7 Recommendations for open-source data and metadata standards

In conclusion of this report, we propose the following recommendations:

7.1 Funding or Grantmaking entities:

7.1.1 Fund Data Standards Development

While some funding agencies already support standards development as part of the development of informatics infrastructures, data standards development should be seen as integral to science innovation and earmarked for funding in research grants, not only in specialized contexts. Funding models should encourage the development and adoption of standards, and fund associated community efforts and tools for this. The OSS model is seen as a particularly promising avenue for an investment of resources, because it builds on previously-developed procedures and technical infrastructure and because it provides avenues for community input along the way. The clarity offered by procedures for enhancement proposals and semantic

versioning schemes adopted in standards development offer avenues for a range of stakeholders to propose to funding bodies well-defined contributions to large and field-wide standards efforts.

7.1.2 Invest in Data Stewards Recognize data stewards as a distinct role in

research and science investment. Set up programs for training for data stewards and invest in career paths that encourage this role. Initial proposals for the curriculum and scope of the role have already been proposed (e.g., in (Mons 2018))

7.1.3 Review Data Standards Pathways

Invest in programs that examine retrospective pathways for establishing data standards. Encourage publication of lifecycles for successful data standards. Lifecycle should include process, creators, affiliations, grants, and adoption journeys. Make this documentation step integral to the work of standards creators and granting agencies. Retrocactively document #3 for standards such as CF(climate science), NASA genelab (space omics), OpenGIS (geospatial), DICOM (medical imaging), GA4GH (genomics), FITS (astronomy), Zarr (domain agnostic n-dimensional arrays)… ?

7.1.4 Establish Governance

Establish governance for standards creation and adoption, especially for communities beyond a certain size that need to converge toward a new standard or rely on an existing standard. Review existing governance practices such as TheOpenSourceWay. Data management plans should promote the sharing of not only data, but also metadata and descriptions of how to use it.

7.1.5 Program Manage Cross Sector alliances

Encourage cross sector and cross domain alliances that can impact successful standards creation. Invest in robust program management of these alliances to align pace and create incentives (for instance via Open Source Program Office / OSPO efforts). Similar to program officers at funding agencies, standards evolution need sustained PM efforts. Multi company partnerships should include strategic initiatives for standard establishment e.g. Pistoiaalliance.

7.1.6 Curriculum Development

Stakeholder organizations should invest in training grants to establish curriculum for data and metadata standards education.

7.2 Science and Technology Communities:

7.2.1 User Driven Development

Standards should be needs-driven and developed in close collaboration with users. Changes and enhancements should be in response to community feedback.

7.2.2 Meta-Standards development

Develop meta-standards or standards-of-standards. These are descriptions of cross-cutting best-practices and can be used as a basis of the analysis or assessment of an existing standard, or as guidelines to develop new standards. For instance, barriers to adopting a data standard irrespective of team size and technological capabilities should be considered. Meta standards should include formalization for versioning of standards & interaction with related software. Naming of standards should aid marketing and adoption.

7.2.3 Ontology Development

Create ontology for standards process such as top down vs bottom up, minimum number of datasets, community size. Examine schema.org (w3c), PEP (Python), CDISC (FDA).

7.2.4 Formalization Guidelines

Amplify formalization/guidelines on how to create standards (example metadata schema specifications using LinkML.

7.2.5 Landscape and Failure Analysis

Before establishing a new standard, survey and document failure of current standards for a specific dataset / domain. Use resources such as Fairsharing or Digital Curation Center.

7.2.6 Machine Readability

Development of standards should be coupled with development of associated software. Make data standards machine readable, and software creation an integral part of establishing a standard's schema e.g. For identifiers for a person using CFF in citations, cffconvert software makes the CFF standard usable and useful. Additionally, standards evolution should maintain software compatibility, and ability to translate and migrate between standards.

8 Acknowledgements

This report was produced following a workshop held at NSF headquarters in Alexandria, VA on April 8th-9th, 2024. We would like to thank the speakers and participants in this workshop for the time and thought that they put into the workshop.

The workshop and this report were funded through NSF grant #2334483 from the NSF Pathways to Enable Open-Source Ecosystems (POSE) program.

References

Basaglia, T, M Bellis, J Blomer, J Boyd, C Bozzi, D Britzger, S Campana, et al. 2023. "Data Preservation in High Energy Physics." *The European Physical Journal C* 83 (9): 795.

Gorgolewski, Krzysztof J, Tibor Auer, Vince D Calhoun, R Cameron Craddock, Samir Das, Eugene P Duff, Guillaume Flandin, et al. 2016. "The Brain Imaging Data Structure, a Format for Organizing and Describing Outputs of Neuroimaging Experiments." *Sci Data* 3 (June): 160044. https://www.nature.com/articles/sdata201644.

Koch, Christof, and R Clay Reid. 2012. "Observatories of the Mind." http://dx.doi.org/10.1038/483397a.

Mons, Barend. 2018. *Data Stewardship for Open Science: Implementing FAIR Principles*. 1st ed. Vol. 1. Milton: CRC Press. https://doi.org/10.1201/9781315380711.

National Institute of Standards and Technology. 2019. "U.S. LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools."

Nosek, Brian. n.d. "Strategy for Culture Change." https://www.cos.io/blog/strategy-for-culture-change.

Poldrack, Russell A, Christopher J Markiewicz, Stefan Appelhoff, Yoni K Ashar, Tibor Auer, Sylvain Baillet, Shashank Bansal, et al. 2024. "The Past, Present, and Future of the Brain Imaging Data Structure (BIDS)." *ArXiv*, January.

Rübel, Oliver, Andrew Tritt, Ryan Ly, Benjamin K Dichter, Satrajit Ghosh, Lawrence Niu, Pamela Baker, et al. 2022. "The Neurodata Without Borders Ecosystem for Neurophysiological Data Science." *Elife* 11 (October).

The National Science and Technology Council. 2022. "Desirable Characteristics of Data Repositories for Federally Funded Research." *Executive Office of the President of the United States, Tech. Rep.*

Wells, Donald Carson, and Eric W Greisen. 1979. "FITS-a Flexible Image Transport System." In *Image Processing in Astronomy*, 445.

Wilkinson, Mark D, Michel Dumontier, I Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Sci Data* 3 (March): 160018.