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, an 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, 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 has had 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, rely heavily on the availability and usability of these large datasets. Data can be openly available but still not useful if it cannot be properly understood. In current conditions in which almost all of the relevant data is stored in digital formats, and many relevant datasets can be found through the communication networks of the world wide web, Findability, Accessibility, Interoperability and Reusability (FAIR) principles for data management and stewardship become critically important [**?**].

One of the main mechanisms through which these 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 storing of certain data types, *schemas* for databases that store a range of data types, *ontologies* to describe and organize metadata in a manner that connects it to field-specific meaning, as well as mechanisms to describe *provenance* of different data derivatives. The importance of standards was underscored in a recent report report by the Subcommittee on Open Science of the National Science and Technology Council on "Desirable characteristics of data repositories for federally funded research" [**?**]. 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..." [**?**]. The converging characteristics of standards that arise from these reports suggest that considerable thought needs to be given to the manner in which standards arise, so that these goals are achieved.

Standards for a specific domain can come about in various ways, but very broadly speaking two kinds of mechanisms can generate a standard for a specific type of data: (i) top-down: in this case a (usually) small group of people develop the standard and disseminate it to the communities of interest with very little input from these communities. An example of this mode of standards development can occur when an instrument is developed by a manufacturer and users of this instrument receive the data in a particular format that was developed in

tandem with the instrument; and (ii) bottom-up: in this case, standards are developed by a larger group of people that convene and reach consensus about the details of the standard in an attempt to cover a large range of use-cases. Most standards are developed through an interplay between these two modes, and understanding how to make the best of these modes is critical in advancing the development of data and metadata standards.

One source of inspiration for bottom-up development of robust, adaptable and useful standards comes from open-source software (OSS). 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 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 1990s has evolved a set of guidelines for licensing of OSS that is designed to protect the rights of developers and users. Technical tools to support the evolution of open-source software include software for distributed version control, such as the Git Source-code management system. 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.

A necessary complement to these technical tools and legal instruments have been a host of practices that define the social interactions *within* communities of OSS developers and users, and structures for governing these communities. While many OSS communities started as projects led by individual founders (so-called benevolent dictators for life, or BDFL; a title first bestowed on the originator of the Python programming language, Guido Van Rossum [**?**]), recent years have led to an increased understanding that minimal standards of democratic governance are required in order for OSS communities to develop and flourish. This has led to the adoption of codes of conduct that govern the standards of behavior and communication among project stakeholders. It has also led to the establishment of democratically elected steering councils/committees from among the members and stakeholders of an OSS project's community.

It was also within the Python community that an orderly process for community-guided evolution of an open-source software project emerged, through the Python Enhancement Proposal (PEP) mechanism [**?**], which lays out how major changes to the software should be proposed, advocated for, and eventually decided on. While these tools, ideas, and practices evolved in developing software, they are readily translated to other domains. For example, OSS notions surrounding IP have given rise to the Creative Commons movement that has expanded these notions to apply to a much wider range of human creative endeavours. Similarly OSS notions regarding collaborative structures have pervaded the current era of open science and team science [**?**, **?**].

4 Challenges for open source data and metadata standards, and some solutions

4.1 Too much flexibility, or too little

It's a story as old as time (or at least as old as standards): users fail to consider existing standards, or perceive an existing standard as not offering enough flexibility to cover some use case, and they embark on the development of a new standard [1].

Another failure is the mismatch between developers of the standard and users. There is an inherent gap in both interest and ability to engage with the technical details undergirding standards and their development between the developers of the standard and their users. In extreme cases, these interests may 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.

4.2 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.

4.3 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 infrastrucutre 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.

4.4 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

[1] So old in fact that an oft-cited XKCD comic has been devoted to it.

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.

4.5 Sustainability

4.6 The importance of automated validation

5 Recommendations

We make the following recommendations:

1. Training for data stewards and career paths that encourage this role.
2. Development of meta-standards or standards-of-standards. These are descriptions of cross-cutting best-practices. These can be used as a basis of the analysis or assessment of an existing standard, or as guidelines to develop new standards.
3. Recommend pathways or lifecycles for successful data standards. Include process, creators, affiliations, grants, and adoption journeys. Make this documentation step integral to the work of standards creators and granting agencies.
4. 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)… ?
5. 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).
6. Amplify formalization/guidelines on how to create standards (example metadata schema specifications using https://linkml.io).
7. Make data standards machine readable, and software creation an integral part of establishing a standard's schema e.g. identifiers for a person using CFF in citations. cffconvert software makes the CFF standard usable and useful.
8. Survey and document failure of current standards for a specific dataset / domain before establishing a new one. Use resources such as Fairsharing.org or Digital Curation Center https://www.dcc.ac.uk/guidance/standards.
9. Funding agencies and science communities need to establish governance for standards creation and adoption (cite https://www.theopensourceway.org/the_open_source_way-guidebook-2.0.html#_project_and_community_governance).
10. Cross sector alliances such as industry - academia need closer coordination and algnment of pace through strong program management (for instance via OSPO efforts).
11. Multi company partnerships should include strategic initiatives for standard establishment (example https://www.pistoiaalliance.org/news/press-release-pistoia-alliance-launches-idmp-1-0/).
12. Stakeholder organizations should invest in training grants to establish curriculum for data and metadata standards education.