Towards an open-source model for data and metadata standards

Ariel Rokem Vani Mandava Nicoleta Cristea

1 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.

2 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) as well as metadata (Musen 2022). 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 technical side, tools such as the Git Source-code management system support complex and distributed open-source workflows that accelerate, streamline, and robustify OSS development. 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.

Data and metadata standards that use tools and practices of OSS ("open-source standards" henceforth) reap many of the benefits that the OSS model has provided in the development of other technologies. The present report explores 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, with a particular focus on cross-sector interactions; and it will make recommendations for future developments and policies that can help this mode of standards development thrive and reach its full potential.

3 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, CERN, and NASA, 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. It also demonstrates field-specific limitations on the adoption of OSS tools and practices that exemplify some of the challenges, which we will explore subsequently.

3.1 Astronomy

An early 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 backward 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.

3.2 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 because of its large volume, and making it publicly available is both technically very hard and potentially ill-advised. Therefore, analysis tools are tuned specifically to the standards of the released data. Incentives to use the standards are provided by funders that require data management plans that specify how the data is shared (i.e., in a standards-compliant manner).

3.3 Earth sciences

The need for geospatial data exchange between different systems began to be recognized in the 1970s and 1980s, but proprietary formats still dominated. Coordinated standardization efforts brought the Open Geospatial Consortium (OGC) establishment in the 1990s, a critical step towards open standards for geospatial data. The 1990s have also seen the development of key standards such as the Network Common Data Form (NetCDF) developed by the University Corporation for Atmospheric Research (UCAR) and the Hierarchical Data Format (HDF), a set of file formats (HDF4, HDF5) that are widely used, particularly in climate research. The GeoTIFF format, which originated at NASA in the late 1990s, is extensively used to share image data. In the 1990s, open web mapping also began with MapServer (https://mapserver.org) and continued later with other projects such as OpenStreetMap (www.openstreetmap.org). The following two decades, the 2000s-2020s, brought an expansion of open standards and integration with web technologies developed by OGC, as well as other standards such as the Keyhole Markup Language (KML) for displaying geographic data in Earth browsers. Formats suitable for cloud computing also emerged, such as the Cloud Optimized GeoTIFF (COG), followed by Zarr and Apache Parquet for array and tabular data, respectively. In 2006, the Open Source Geospatial Foundation (OSGeo, https://www.osgeo.org) was established, demonstrating the community's commitment to the development of open-source geospatial technologies. While some standards have been developed in the industry (e.g., Keyhole Markup Language (KML) by Keyhole Inc., which Google later acquired), they later became international standards of the OGC, which now encompasses more than 450 commercial, governmental, nonprofit, and research organizations working together on the development and implementation of open standards (https://www.ogc.org).

3.4 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 BIDS Enhancement Proposal (BEP) process that is directly inspired by the Python programming language community's Python Enhancement Proposal procedure, which isused 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.

3.5 Community science

Another interesting use case for open-source standards is community/citizen science. This approach, which has grown in the last 20 years, has many benefits for both the research field that harnesses the energy of non-scientist members of the community to engage with scientific data, as well as to the community members themselves who can draw both knowledge and pride in their participation in the scientific endeavor. It is also recognized that unique broader benefits are accrued from this mode of scientific research, through the inclusion of perspectives and data that would not otherwise be included. To make data accessible to community scientists, and to make the data collected by community scientists accessible to professional scientists, it needs to be provided in a manner that can be created and accessed without specialized instruments or specialized knowledge. Here, standards are needed to facilitate interactions between an in-group of expert researchers who generate and curate data and a broader set of out-group enthusiasts who would like to make meaningful contributions to the science. This creates a particularly stringent constraint on transparency and simplicity of standards. Creating these standards in a manner that addresses these unique constraints can benefit from OSS tools, with the caveat that some of these tools require additional expertise. For example, if the standard is developed using git/GitHub for versioning, this would require learning the complex and obscure technical aspects of these system that are far from easy to adopt, even for many professional scientists.

4 Opportunities and risks for open-source standards

While open-source standards benefit from the technical and social aspects of OSS, 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. Similarly, analysis technology stacks that are developed based on an existing version of a standard have to adapt to the introduction of new ideas and changes into a standard. Dynamic changes of this sort therefore risk causing a loss of faith in the standard by a user community, and migration away from the standard. Similarly, if a standard evolves too rapidly, users may choose to stick to an outdated version of a standard for a long time, creating strains on the community of developers and maintainers of a standard who will need to accommodate long deprecation cycles.

4.2 Mismatches between standards developers and user communities

Open-source standards often entail 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 the users of the standard, which are members of the broader research field to which the standard pertains. This gap, in and of itself, creates friction on the path to broad adoption and best utilization of the standards. In extreme cases, the interests of researchers and standards developers may even seem at odds, as developers implement sophisticated mechanisms to automate the creation and validation of the standard or advocate for more technically advanced mechanisms for evolving the standard. These advanced capabilities offer more robust development practices and consistency in cases where the standards are complex and elaborate. They can also ease the maintenance burden of the standard. On the other hand, they may end up 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. One example of this (already mentioned above in Section 3) is the use of git/GitHub for versioning of standards documents. This sets a high bar for participation in standards development for researchers in fields of research in which git/GitHub have not yet had significant adoption as tools of day-to-day computational practice. At the same time, it provides clarity and robustness for standards developers communities that are well-versed in these tools.

4.3 Cross-domain gaps

There is much to be gained from the development of standards that apply in multiple different domains. For example, many research fields use images as data and array-based computing that is applicable to images in various research fields is at the core of many scientific computing codes. This means that practitioners within any given field should be motivated to draw on shared data standards and shared software implementations of operations that are common across fields. On the other hand, it is very hard to justify the investment in these common resources. On the one hand, data standardization investment is even more justified if the standard is generalizable beyond any specific science domain. On the other hand, while the use cases are domain sciences based, data standardization is seen as a data infrastructure and not a science investment, reducing the immediate incentives for researchers to engage with such efforts. This is exacerbated by science research funding schemes that eschew generalized cross-domain solutions, and that seek to generate tangible impact only with a specific domain.

4.4 Data instrumentation issues

Where there is commercial interest in the development of data analysis tools (e.g., in biomedical applications or applications were research funding can be directed towards commercial solutions) there is an incentive to create data formats and data analysis platforms that are proprietary. This may drive innovative applications of scientific measurements, but also creates sub-fields where scientific observations are generated by proprietary instrumentation, due to these commercialization or other profit-driven incentives. There is a lack of regulatory oversight to adhere to available standards or evolve common tools, limiting integration across different measurements. In cases where a significant amount of data is already stored in proprietary formats, significant data transformations may be required to get data to a state that is amenable to open-source standards. In these sub-fields there may also be a lack of incentive to set aside investment or resources to invest in establishing open-source data standards, leaving these sub-fields relatively siloed.

4.4.1 Harnessing new computing paradigms and technologies

Open-source standards development faces the challenges of adapting to new computing paradigms and technologies. Cloud computing provides a particularly stark set of opportunities and challenges. On the one hand, cloud computing offers practical solutions for many challenges of contemporary data-driven research. For example, the scalability of cloud resources addresses some of the challenges of the scale of data that is produced by instruments in many fields. The cloud also makes data access relatively straightforward, because of the ability to determine data access permissions in a granular fashion. On the other hand, cloud computing requires reinstrumenting many data formats. This is because cloud data access patterns are fundamentally different from the ones that are used in local

posix-style file-systems. Suspicion of cloud computing comes in two different flavors: the first by researchers and administrators who may be wary of costs associated with cloud computing, and especially with the difficulty of predicting these costs. Projects such as NSF's Cloud Bank seek to mitigate some of these concerns, by providing an additional layer of transparency into cloud costs (Norman et al. 2021). The other type of objection relates to the fact that cloud computing services, by their very nature, are closed ecosystems that resist portability and interoperability. Some aspects of the services are always going to remain hidden and privy only to the cloud computing service provider. In this respect, cloud computing runs afoul of some of the appealing aspects of OSS. That said, the development of "cloud native" standards can provide significant benefits in terms of the research that can be conducted. For example, NOAA plans to use cloud computing for integration across the multiple disparate datasets that it collects to build knowledge graphs that can be queried by researchers to answer questions that can only be answered through this integration. Putting all the data "in one place" should help with that. Adaptation to the cloud in terms of data standards has driven development of new file formats. A salient example is the ZARR format (Miles et al. 2024), which supports random access into array-based datasets stored in cloud object storage, facilitating scalable and parallelized computing on these data. Indeed, data standards such as NWB (neuroscience) and OME (microscopy) now use ZARR as a backend for cloud-based storage. In other cases, file formats that were once not straightforward to use in the cloud, such as HDF5 and TIFF have been adapted to cloud use (e.g., through the cloud-optimized geoTIFF format).

4.5 Unclear pathways for standards success and sustainability

The development of open-source standards faces similar sustainability challenges to those faced by open-source software that is developed for research. Standards typically develop organically through sustained and persistent efforts from dedicated groups of data practitioners. 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 the implementation of a specific data architecture (e.g., a particular file format) to become a data standard. As a result, data standardization lacks formal avenues for success and recognition, for example through dedicated research grants (and see Section 5). This hampers the long-term trajectory that is needed to inculcate a standard into the day-to-day practice of researchers.

4.6 The importance of automated validation

5 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 stakeholders in other sectors. Several different kinds of

cross-sector interactions can be defined as having an important impact on the development of open-source standards.

5.1 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 2013) 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. Importantly, open-source standards seem to well-match at least some of these characteristics.

The other side of policies is the implementation of these policies in practice by developers of open-source standards and by the communities to which the standards pertain. A compelling road map towards implementation and adoption of open science practices in general and open-source standards in particular is offered in a blog post authored by the Center for Open Science's co-founder and executive director, 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 makes the adoption of standards

possible, and maybe even easy, and the policy goals, arises from a community of practice that makes the adoption of standards *normative*. Once all of these pieces are in place, making adoption of open science standards *required* through policy becomes more straightforward and less onerous.

5.2 Funding

Government-set policy intersects with funding considerations. This is because it is primarily directed towards research that is funded through governmental funding agencies. For example, high-level policy guidance boils to practice in guidance for data management plans that are part of funded research. In response to the policy guidance, these have become increasingly more detailed and, for example, NSF- and NIH-funded researchers are now required to both formulate their plans with more clarity and increasingly also to share data using specified standards as a condition for funding.

However, there are other ways in which funding relates to the development of open-source standards. For example, through the BRAIN Initiative, the National Institutes of Health have provided key funding for the development of the Brain Imaging Data Structure standard in neuroscience. Where large governmental funding agencies may not have the resources or agility required to fund nascent or unconventional ways of formulating standards, funding by non-governmental philanthropies and other organizations can provide alternatives. One example (out of many) is the Chan-Zuckerberg Initiative program for Essential Open Source Software, which funds foundational open-source software projects that have an application in biomedical sciences. Distinct from NIH funding, however, some of this investment focuses on the development of OSS practices. For example, funding to the Arrow project that focuses on developing open-source software maintenance skills and practices, rather than a specific biomedical application.

5.3 Industry

Interactions of data and meta-data standards with commercial interests may provide specific sources of friction. This is because proprietary/closed formats of data can create difficulty at various transition points: from one instrument vendor to another, from data producer to downstream recipient/user, etc. On the other hand, in some cases, cross-sector collaborations with commercial entities may pave the way to robust and useful standards. One example is the DICOM standard, which is maintained by working groups that encompass commercial imaging device vendors and researchers.

6 Recommendations for open-source data and metadata standards

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

6.1 Policy-making and Funding entities:

6.1.1 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 the democratization of development processes and for community input along the way. The clarity offered by procedures for enhancement proposals and semantic versioning schemes adopted in standards development offers avenues for a range of stakeholders to propose well-defined contributions to large and field-wide standards efforts (e.g., (Pestilli et al. 2021)).

6.1.2 Invest in Data Stewards

Advancing the development and adoption of open-source standards requires the dissemination of knowledge to researchers in a variety of fields, but this dissemination itself may not be enough without the fostering of specialized expertise. Therefore, it is important to recognize the distinct role that *data stewards* play in contemporary research. As policy demands for openness become increasingly high, it is crucial to truly support experts whose role will be to develop, maintain, and facilitate the adoption and use of open-source standards. This support needs to manifest in all stages of the career of these individuals: it will be necessary to set up programs for training for data stewards, and to invest in the career paths of individuals that receive such training so that this crucial role is encouraged. Initial proposals for the curriculum and scope of the role have already been proposed (e.g., in (Mons 2018)), but we identify here also a need to connect these individuals directly to the practices that exemplify open-source standards. Thus, it will be important for these individuals to be facile in the methodology of OSS. This does not mean that they need to become software engineers – though for some of them there may be some overlap with the role of research software engineers (Connolly et al. 2023) – but rather that they need to become familiar with those parts of the OSS development life-cycle that are specifically useful for the development of open-source standards. For example, tools for version control, tools for versioning, and tools for creation and validation of compliant data and metadata.

6.1.3 Review Data Standards Pathways

Invest in programs that examine retrospective pathways for establishing data standards. Encourage publication of lifecycles for successful data standards. These lifecycles should include

the process, creators, affiliations, grants, and adoption journeys of open-source standards. To encourage sustainable development of open-source standards, and to build on prior experience, the documentation and dissemination of lifecycles should be seen as an integral step of the work of standards creators and granting agencies. In the meanwhile, it would be good to also retroactively document the lifecycle of existing standards that are seen as success stories. Research on the principles that underlie successful open-source standards development can be used to formulate new standards and iterate on existing ones.

6.1.4 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.

6.1.5 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.

6.1.6 Curriculum Development

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

6.2 Science and Technology Communities:

6.2.1 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.

6.2.2 Meta-Standards development

In surveying the landscape of existing standards, a readiness/maturity model can be developed that assesses the challenges and opportunities that a specific standard faces. This process in itself can be standardized to develop meta-standards or standards-of-standards. These are the succinct descriptions of cross-cutting best-practices that can be used as a basis for 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 and interactions with specific related software. Aspects of communication with potential user audiences (e.g., researchers in particular domains) should be taken into account as well. For example, in the quality of onboarding documentation and tools for ingestion or conversion into standards-compliant datasets. Relatedly, it would be good to create an ontology for standards process such as top down vs bottom up, minimum number of datasets, target community size and technical expertise typical of this community, etc. This ontology can help guide the standards-development process towards more effective adoption and use.

6.2.3 Formalization Guidelines

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

6.2.4 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.

6.2.5 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.

7 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.

Connolly, Andrew, Joseph Hellerstein, Naomi Alterman, David Beck, Rob Fatland, Ed Lazowska, Vani Mandava, and Sarah Stone. 2023. "Software Engineering Practices in Academia: Promoting the 3Rs—Readability, Resilience, and Reuse." *Harvard Data Science Review* 5 (2).

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.

Miles, Alistair, jakirkham, M Bussonnier, Josh Moore, Dimitri Papadopoulos Orfanos, Davis Bennett, David Stansby, et al. 2024. "Zarr-Developers/Zarr-Python: V3.0.0-Alpha." Zenodo. https://doi.org/10.5281/zenodo.11592827.

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

Musen, Mark A. 2022. "Without Appropriate Metadata, Data-Sharing Mandates Are Pointless." *Nature* 609 (7926): 222.

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

Norman, Michael, Vince Kellen, Shava Smallen, Brian DeMeulle, Shawn Strande, Ed Lazowska, Naomi Alterman, et al. 2021. "CloudBank: Managed Services to Simplify Cloud Access for Computer Science Research and Education." In *Practice and Experience in Advanced Research Computing*. PEARC '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3437359.3465586.

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

Pestilli, Franco, Russ Poldrack, Ariel Rokem, Theodore Satterthwaite, Franklin Feingold, Eugene Duff, Cyril Pernet, Robert Smith, Oscar Esteban, and Matt Cieslak. 2021. "A

Community-Driven Development of the Brain Imaging Data Standard (BIDS) to Describe Macroscopic Brain Connections." *OSF*.

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.