Towards an open-source model for data and metadata standards

Ariel Rokem[a,b,*], Vani Mandava[c,b], Nicoleta Cristea[d,b], Anshul Tambay[c,b], Andrew J. Connolly[e,b]

[a] *University of Washington, Department of Psychology, Seattle, USA,*

[b] *University of Washington, eScience Institute, Seattle, USA,*

[c] *University of Washington, Scientific Software Engineering Center, Seattle, USA,*

[d] *University of Washington, Department of Civil and Environmental Engineering, Seattle, USA,*

[e] *University of Washington, Department of Astronomy, Seattle, USA,*

Abstract

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, data and meta-data standards can use long-standing technical and socio-technical processes that have been key to managing the development of software, and which allow incorporating broad community input into the formulation of these standards. On the other hand, open-source models carry unique risks that need to be considered. This report surveys existing open-source standards development, addressing these benefits and risks. It outlines recommendations for standards developers, funders and other stakeholders on the path to robust, interoperable and usable open-source data and metadata standards.

*Corresponding author
Email address: arokem@uw.edu (Ariel Rokem)

October 17, 2024

1. 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 [1] as well as metadata [2]. 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 make OSS development more robust. 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 survey 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.

2. 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 SDSS, 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.

2.1. 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 [3], 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 the International Astronomical Union (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 a coordinate frame for pixels, means that data from different observational instruments can be stored in this format and relationships between data from different instruments can be defined, rendering manual and error-prone procedures for conforming images obsolete. Nevertheless, the stability has also raised some issues as the field continues to adapt to new measurement methods and the demands of ever-increasing data volumes and complex data analysis use-case, such as interchange with other data and the use of complex data bases to store and share data [4]. Another prominent example of the use of open-source processes to develop standards in Astronomy is in the tools and protocols developed by the International Virtual Observatory Alliance (IVOA) and its national implementations, e.g., in the US Virtual Astronomical Observatory[5]. The virtual observatories facilitate discovery and access across observatories around the world and underpin data discovery in astronomy. The IVOA took inspiration from the World-Wide Web Consortium (W3C) and adopted its process

for the development of its standards (i.e., Working drafts → Proposed Recommendations → Recommendations), with individual standards developed by inter-institutional and international working groups. One of the outcomes of the coordination effort is the development of an ecosystem of software tools both developed within the observatory teams and within the user community that interoperate with the standards that were adopted by the observatories.

2.2. 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 [6]. 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 would be technically very difficult. 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).

2.3. 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. 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).

2.4. Neuroscience

In contrast to the previously-mentioned fields, 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 [7]. 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 [8] and the Brain Imaging Data Structure (BIDS) standard for neuroimaging data [9]. BIDS in particular owes some of its success to the adoption of OSS development mechanisms [10]. 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 is 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.

2.5. Community science

Another interesting use case for open-source standards is community/citizen science. An early example of this approach is OpenStreetMap (https://www.openstreetmap.org), which allows users to contribute to the project development with code and data and freely use the maps and other related geospatial datasets. But this example is not unique. Overall, this approach has grown in the last 20 years and has been adopted in many different fields. It 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.

3. 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.

3.1. 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 users/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. These existing datasets may have a lifespan of decades, making continued compatibility crucial. 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. On the other hand, in cases in which some forms of dynamic change is prohibited – as in the case of the FITS file format, which prohibits changes that break backwards-compatibility – there is also a cost associated with the stability [4]: limiting adoption and combinations of new types of measurements, new analysis methods or new modes of data storage and data sharing.

3.2. Mismatches between standards developers and user communities

Open-source standards often entail an inherent gap between the core developers of the standard and the users of the standard. The former may be possess higher ability to engage with the technical details undergirding standards and their development, while the latter still have a high level of interest as 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 experimental researchers and data providers 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 2) 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.

Another layer of potential mismatches arises when a more complex set of stakeholders needs to be considered. For example, the Group on Earth Observations (GEO) is a network that aims to coordinate decision making around satellite missions and to standardize the data that results from these missions. Because this group involves a range of different stakeholders, including individuals who more closely understand potential legal issues and researchers who are better equipped to evaluate technical and domain questions, communication is slower and hindered. As the group aims to move forward by consensus, these communication difficulties can slow down progress. This is just an example, which exemplifies the many cases in which OSS process which strives for consensus can slow progress.

3.3. 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.

3.4. 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. FTIR Spectroscopy is one such example, wherein use of Bruker instrumentation necessitates downstream analysis of the resulting measurements using proprietary binary formats necessary for the OPUS Software. Another example is the proliferation of proprietary file formats in electrophysiological measurements of brain signals [11, @Hermes2023-aw]. And yet another one is proprietary application programming interfaces (APIs) used in electronic health records [12, @Adler-Milstein2017-id]. In most cases, 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, or where access is limited by proprietary APIs 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.

3.4.1. 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. This can particularly affect scenarios where long-term preservation is required. Projects such as NSF's Cloud Bank seek to mitigate

some of these concerns, by providing an additional layer of transparency into cloud costs [13]. 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 [14], 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).

3.5. 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 4). This hampers the long-term trajectory that is needed to inculcate a standard into the day-to-day practice of researchers.

4. 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.

4.1. 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" [15]. 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, [...] Multipath, [...] Open and transparent, [...] and [that] result in globally relevant and non-discriminatory standards..." [16]. 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" [17]. 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.

4.2. 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.

4.3. 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. For example, imaging measurements in human subjects (e.g., in brain imaging experiments) significantly interact with standards for medical imaging, and chiefly the Digital Imaging and Communications in Medicine (DICOM) standard, which is widely used in a range of medical imaging applications, including in clinical settings [18, @Mustra2008-xk]. The standard emerged from the demands of the clinical practice in the 1980s, as digital technologies were came into widespread use in medical imaging, through joint work of industry organizations: the American College of Radiology and the National Association of Electronic Manufacturers. One of the defining features of the DICOM standard is that it allows manufacturers of instruments to define "private fields" that are compliant with the standard, but which may include idiosyncratically organized data and/or meta-data. This provides significant flexibility, but can also easily lead to the loss of important information. Nevertheless, the human brain imaging case is exemplary of a case in which industry standards and research standards coexist and

need to communicate with each other effectively to advance research use-cases, while keeping up with the rapid development of the technologies.

5. Recommendations for open-source data and metadata standards

In conclusion of this report, we would like to propose a set of recommendations that distill the lessons learned from an examination of data and metadata standards through the lense of open-source software development practices. We divide this section into two parts: one aimed at the science and technology communities that develop and maintain open-source standards, and the other aimed at policy-making and funding agencies, who have an interest in fostering more efficient, more robust, and more transparent open-source standards.

5.1. Science and technology communities:

5.1.1. Establish standards governance based on OSS best practices

While best-practice governance principles are also relatively new in OSS communities, there is already a substantial set of prior art in this domain, on which the developers and maintainers of open-source data and metadata standards can rely. For example, it is now clear that governance principles and rules can mitigate some of the risks and challenges mentioned in Section 3, especially for communities beyond a certain size that need to converge toward a new standard or rely on an existing standard. Developers and maintainers should review existing governance practices such as those provided by The Open Source Way, (https://www.theopensourceway.org/).

5.1.2. Foster meta-standards development

One of the main conclusions that arise from our survey of the landscape of existing standards is that there is significant knowledge that exists across fields and domains and that informs the development of standards within each field, but that could be surfaced to the level where it may be adopted more widely in different domains and be more broadly useful. One approach to this is a comparative approach: in this approach, a readiness and/or maturity model can be developed that assesses the challenges and opportunities that a specific standard faces at its current phase of development. Developing such a maturity model, while it goes beyond the scope of the current report, could lead to the eventual development of a meta-standard or a standard-of-standards. This would facilitate a succinct description 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, specific barriers to adopting a data standard that take into account the size of the community and its specific technological capabilities should be considered.

More generally, meta-standards could include formalization for versioning of standards and interactions with specific related software. This includes amplifying formalization/guidelines on how to create standards (for example, metadata schema specifications using LinkML, https://linkml.io). However, 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.

An ontology for the standards-development process – for example top-down vs bottom-up, minimum number of datasets, target community size and technical expertise typical of this community, and so forth – could help guide the standards-development process towards more effective adoption and use. A set of meta-standards and high-level descriptions of the standards-development process – some of which is laid out in this report – could help standard developers avoid known pitfalls, such as the dreaded proliferation of standards, or complexity-impeded adoption. Surveying and documenting the success and failures of current standards for a specific dataset / domain can help disseminate knowledge about the standardization process. Resources such as Fairsharing (https://fairsharing.org/) or the Digital Curation Center (https://www.dcc.ac.uk/guidance/standards) can help guide this process.

5.1.3. Develop standards in tandem with standards-associated software

Development of standards should be coupled and tightly linked with development of associated software. This produces a virtuous cycle where the use-cases and technical issues that arise in software development informs the development of the standard and vice versa. One of the lessons learned across a variety of different standards is the importance of automated validation of the standard. Automated validation is broadly seen as a requirement for the adoption of a standard and a factor in managing change of the standard over time. To advance this virtuous cycle, we recommend to make data standards machine readable, and make software creation an integral part of establishing a standard's schema. Additionally, standards evolution should maintain software compatibility, and ability to translate and migrate between standards.

5.2. Policy-making and funding entities:

5.2.1. Fund the development of open-source standards

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. At the same time, there are significant challenges associated with incentives to engage, ranging from the dilution of credit to individual contributors, and ranging through the burnout of maintainers and developers. 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., [19]), and potentially helps alleviate some of these concerns by providing avenues for individual contributions to surface, as well as clarity of process, which can alleviate the risks of maintainer burnout.

5.2.2. 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 [20]), 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 conversant 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 [21] – 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. Stakeholder organizations should invest in training grants to establish curriculum for data and metadata standards education.

Ultimately, efficient use of data stewards and their knowledge will have to be applied. It is evident that not every project and every lab that produces data requires a full-time data steward. Instead, data stewardship could be centralized within organizations such as libraries, data science, or software engineering cores of larger research organizations. This would be akin to recent models for research software engineering that are becoming common in many research organization [22]. Efficiency considerations also suggest that the development of data standards would not have its intended purpose unless funds are also allocated to the implementation of the standard in practice. Mandating standards without appropriate funding for their implementation by data producers and data users could risk hampering science and could leading to researchers doing the bare minimum to make their data "open".

5.2.3. Review open-source 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, and to foster the awareness of these standards. In addition, fostering research projects on the principles that underlie successful open-source standards development will help formulate new standards and iterate on existing ones. In accordance, data management plans should promote the sharing of not only data, but also metadata and descriptions of how to use it.

5.2.4. 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 Offices at Universities or other research organizations). Similar to program officers at funding agencies, standards evolution need sustained PM efforts. Multi-party partnerships should include strategic initiatives for standard establishment such as the Pistoia Alliance (https://www.pistoiaalliance.org/).

6. 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. A list of workshop participants is provided as an appendix (Section 8).

The workshop and this report were funded through NSF grant #2334483 from the NSF Pathways to Enable Open-Source Ecosystems (POSE) program. The opinions expressed in this report do not necessarily reflect those of the National Science Foundation.

7. References

[1] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester,

P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR guiding principles for scientific data management and stewardship, Sci Data 3 (2016) 160018.

[2] M. A. Musen, Without appropriate metadata, data-sharing mandates are pointless, Nature 609 (7926) (2022) 222.

[3] D. C. Wells, E. W. Greisen, Fits-a flexible image transport system, in: Image processing in astronomy, 1979, p. 445.

[4] M. Scroggins, B. M. Boscoe, Once FITS, always FITS? astronomical infrastructure in transition, IEEE Ann. Hist. Comput. 42 (2) (2020) 42–54.

[5] R. J. Hanisch, G. B. Berriman, T. J. W. Lazio, S. Emery Bunn, J. Evans, T. A. McGlynn, R. Plante, The virtual astronomical observatory: Re-engineering access to astronomical data, Astron. Comput. 11 (2015) 190–209.

[6] T. Basaglia, M. Bellis, J. Blomer, J. Boyd, C. Bozzi, D. Britzger, S. Campana, C. Cartaro, G. Chen, B. Couturier, G. David, C. Diaconu, A. Dobrin, D. Duellmann, M. Ebert, P. Elmer, J. Fernandes, L. Fields, P. Fokianos, G. Ganis, A. Geiser, M. Gheata, J. B. G. Lopez, T. Hara, L. Heinrich, M. Hildreth, K. Herner, B. Jayatilaka, M. Kado, O. Keeble, A. Kohls, K. Naim, C. Lange, K. Lassila-Perini, S. Levonian, M. Maggi, Z. Marshall, P. M. Vila, A. Mečionis, A. Morris, S. Piano, M. Potekhin, M. Schröder, U. Schwickerath, E. Sexton-Kennedy, T. Šimko, T. Smith, D. South, A. Verbytskyi, M. Vidal, A. Vivace, L. Wang, G. Watt, T. Wenaus, DPHEP Collaboration, Data preservation in high energy physics, The European Physical Journal C 83 (9) (2023) 795.

[7] C. Koch, R. Clay Reid, Observatories of the mind, http://dx.doi.org/10.1038/483397a, accessed: 2024-6-17 (Mar. 2012).

[8] O. Rübel, A. Tritt, R. Ly, B. K. Dichter, S. Ghosh, L. Niu, P. Baker, I. Soltesz, L. Ng, K. Svoboda, L. Frank, K. E. Bouchard, The neurodata without borders ecosystem for neurophysiological data science, Elife 11 (Oct. 2022).

[9] K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, D. A. Handwerker, M. Hanke, D. Keator, X. Li, Z. Michael, C. Maumet, B. N. Nichols, T. E. Nichols, J. Pellman, J.-B. Poline, A. Rokem, G. Schaefer, V. Sochat, W. Triplett, J. A. Turner, G. Varoquaux, R. A. Poldrack, The Brain Imaging Data Structure, a format for organizing and describing outputs of neuroimaging experiments, Sci Data 3 (2016) 160044.
URL https://www.nature.com/articles/sdata201644

[10] R. A. Poldrack, C. J. Markiewicz, S. Appelhoff, Y. K. Ashar, T. Auer, S. Baillet, S. Bansal, L. Beltrachini, C. G. Benar, G. Bertazzoli,

S. Bhogawar, R. W. Blair, M. Bortoletto, M. Boudreau, T. L. Brooks, V. D. Calhoun, F. M. Castelli, P. Clement, A. L. Cohen, J. Cohen-Adad, S. D'Ambrosio, G. de Hollander, M. de la Iglesia-Vayá, A. de la Vega, A. Delorme, O. Devinsky, D. Draschkow, E. P. Duff, E. DuPre, E. Earl, O. Esteban, F. W. Feingold, G. Flandin, A. Galassi, G. Gallitto, M. Ganz, R. Gau, J. Gholam, S. S. Ghosh, A. Giacomel, A. G. Gillman, P. Gleeson, A. Gramfort, S. Guay, G. Guidali, Y. O. Halchenko, D. A. Handwerker, N. Hardcastle, P. Herholz, D. Hermes, C. J. Honey, R. B. Innis, H.-I. Ioanas, A. Jahn, A. Karakuzu, D. B. Keator, G. Kiar, B. Kincses, A. R. Laird, J. C. Lau, A. Lazari, J. H. Legarreta, A. Li, X. Li, B. C. Love, H. Lu, E. Marcantoni, C. Maumet, G. Mazzamuto, S. L. Meisler, M. Mikkelsen, H. Mutsaerts, T. E. Nichols, A. Nikolaidis, G. Nilsonne, G. Niso, M. Norgaard, T. W. Okell, R. Oostenveld, E. Ort, P. J. Park, M. Pawlik, C. R. Pernet, F. Pestilli, J. Petr, C. Phillips, J.-B. Poline, L. Pollonini, P. R. Raamana, P. Ritter, G. Rizzo, K. A. Robbins, A. P. Rockhill, C. Rogers, A. Rokem, C. Rorden, A. Routier, J. M. Saborit-Torres, T. Salo, M. Schirner, R. E. Smith, T. Spisak, J. Sprenger, N. C. Swann, M. Szinte, S. Takerkart, B. Thirion, A. G. Thomas, S. Torabian, G. Varoquaux, B. Voytek, J. Welzel, M. Wilson, T. Yarkoni, K. J. Gorgolewski, The past, present, and future of the brain imaging data structure (BIDS), ArXiv (Jan. 2024).

[11] C. J. Gillon, C. Baker, R. Ly, E. Balzani, B. W. Brunton, M. Schottdorf, S. Ghosh, N. Dehghani, ODIN: Open data in neurophysiology: Advancements, solutions & challenges, arXiv [q-bio.NC] (Jul. 2024).

[12] W. Barker, N. Maisel, C. E. Strawley, G. K. Israelit, J. Adler-Milstein, B. Rosner, A national survey of digital health company experiences with electronic health record application programming interfaces, J. Am. Med. Inform. Assoc. 31 (4) (2024) 866–874.

[13] M. Norman, V. Kellen, S. Smallen, B. DeMeulle, S. Strande, E. Lazowska, N. Alterman, R. Fatland, S. Stone, A. Tan, K. Yelick, E. Van Dusen, J. Mitchell, CloudBank: Managed Services to Simplify Cloud Access for Computer Science Research and Education, in: Practice and Experience in Advanced Research Computing, PEARC '21, Association for Computing Machinery, New York, NY, USA, 2021. doi:10.1145/3437359.3465586.
URL https://doi.org/10.1145/3437359.3465586

[14] A. Miles, jakirkham, M. Bussonnier, J. Moore, D. P. Orfanos, D. Bennett, D. Stansby, J. Hamman, J. Bourbeau, A. Fulton, G. Lee, R. Abernathey, N. Rzepka, Z. Patel, M. R. B. Kristensen, S. Verma, S. Chopra, M. Rocklin, A. B. AWA, M. Jones, M. Durant, E. S. de Andrade, V. Schut, raphael dussin, S. Chaudhary, C. Barnes, J. Nunez-Iglesias, shikharsg, zarr-developers/zarr-python: v3.0.0-alpha (Jun. 2024). doi:10.5281/zenodo.11592827.
URL https://doi.org/10.5281/zenodo.11592827

[15] The National Science and Technology Council, Desirable characteristics of data repositories for federally funded research, Executive Office of the President of the United States, Tech. Rep (2022).

[16] National Institute of Standards and Technology, U.S. LEADERSHIP IN AI: A plan for federal engagement in developing technical standards and related tools, Tech. rep. (2019).

[17] B. Nosek, Strategy for culture change, https://www.cos.io/blog/strategy-for-culture-change, accessed: 2024-6-19.

[18] M. Larobina, Thirty years of the DICOM standard, Tomography 9 (5) (2023) 1829–1838.

[19] F. Pestilli, R. Poldrack, A. Rokem, T. Satterthwaite, F. Feingold, E. Duff, C. Pernet, R. Smith, O. Esteban, M. Cieslak, A community-driven development of the brain imaging data standard (bids) to describe macroscopic brain connections, OSF (2021).

[20] B. Mons, Data Stewardship for Open Science: Implementing FAIR Principles, 1st Edition, Vol. 1, CRC Press, Milton, 2018. doi:10.1201/9781315380711.

[21] A. Connolly, J. Hellerstein, N. Alterman, D. Beck, R. Fatland, E. Lazowska, V. Mandava, S. Stone, Software Engineering Practices in Academia: Promoting the 3rs—Readability, Resilience, and Reuse, Harvard Data Science Review 5 (2), https://hdsr.mitpress.mit.edu/pub/f0f7h5cu (apr 27 2023).

[22] Hiring, managing, and retaining data scientists and research software engineers in academia: A career guidebook from ADSA and US-RSE (2023). doi:https://doi.org/10.5281/zenodo.8329337.
URL https://zenodo.org/records/8329337

8. Appendix: List of participants

Name	Affiliation
Alex D Wade	Digital Science
Alexander Szalay	Johns Hopkins University
Andrew Connolly	University of Washington
Anshul Tushar Tambay	University of Washington
Ariel Rokem	University of Washington
Carolina Lorena Berys	University of California, San Diego
Christine Kirkpatrick	San Diego Supercomputer Center
Fernando Seabra Chirigati	Nature Computational Science
Jessica Morgan	NOAA
John Relph	NOAA
Julia Ferraioli	Open Source Stories
Jurriaan Hein Spaaks	formerly Netherlands eScience Center
Justin (Jay) Hnilo	Department of Energy
Kalynn Elisabeth Kennon	Infectious Diseases Data Observatory
Kevin Christopher Booth	Radiant Earth
Kristofer E. Bouchard	Lawrence Berkeley National Labs
Lea A. Shanley	University of California, Berkeley
Michael Spannowsky	Durham University
Nicoleta C Cristea	University of Washington
Nina Amla	NSF
Oliver Ruebel	Lawrence Berkeley National Labs
Ray E. Habermann	Metadata Game Changers
Raymond (Ray) Plante	NIST
Robert Hanisch	NIST
Saskia de Vries	Allen Institute for Neural Dynamics
Steven Crawford	NASA
Vani Mandava	University of Washington
Yaroslav O. Halchenko	Dartmouth College
Ziheng Sun	George Mason University