INCF Standards Review Criteria v.2.0

Author: Standards and Best Practices Committee,
International Neuroinformatics Coordinating Facility

June 2023
Submitted for consideration: 2 September 2021
Initial review: 25 April 2023

Basic Metadata

Title: SPARC Data Structure (SDS)

Brief description: The Stimulating Peripheral Activity to Relieve Conditions (SPARC) Data Structure is a consistent file structure and naming convention, based on the Brain Imaging Data Structure (BIDS) to ensure that the diverse types of data in SPARC is organized in a similar manner. The current version is SDS 2.0 (released June 24, 2022) the first version is SDS 1.2.3 (released October 30, 2019).

URL: https://sparc.science/help/qvEcnv56c76V0JC0KvtSd

Steward: Tom Gillespie tgbugs@gmail.com

Relevant publication: https://doi.org/10.1101/2021.02.10.430563

Summary of Discussion

Overall, the members of the INCF Standards and Best Practices Committee could potentially meet the criteria for INCF endorsement. It is open, has strong documentation, and supports FAIR reasonably well with evidence of efforts to align it with BIDS and the DANDI metadata structure. Its use is currently imposed on the SPARC community of approximately 500 investigators with no evidence of use outside of this community. While SDS is inspired by BIDS, it was designed explicitly to accommodate data collection patterns that are fundamentally incompatible with BIDS 1.0 structures (the 20% that BIDS does not attempt to cover). Since SDS has been a consortium used standard, it currently lacks a formal governance structure; however, the submitters have indicated that once SDS is used by groups outside of the consortium that a formal governance structure will be established.

Recommendation

The INCF Standards and Best Practices Committee voted to put SDS forward for community review. The committee is particularly interested in receiving comments from the BIDS and DANDI metadata structure communities. In addition, the committee would also like for SDS, BIDS, and DANDI to draft a commentary on the relationship between the standards to better help the community in determining which standard to use.

Conflicts of Interest

No conflicts of interest declared.

Open criteria

It is essential that a FAIR supporting standard is open and allows free use by the community. Open development practices are also strongly encouraged to facilitate transparency and adoption. If questions do not apply, leave them blank or mark N/A.

1. **Is the SBP covered under an open license so that it is free to implement and reuse by all interested parties (including commercial)? (List of open source licenses)**
   a. Yes.
2. **What license is used?**
   a. The code is licensed under the MIT license. The paper is open access, and any supporting documentation is licensed under the MIT license or CC BY 4.0.
3. **Does the SBP follow open development practices?**
   a. Yes.
4. **Where and how are the code/documents managed?**
   a. The implementation is maintained as a git repository with a remote located at https://github.com/SciCrunch/sparc-curation/. This repository also houses the developer documentation for the reference implementation.
   b. User facing documentation can be found at:
      i. High level overview: https://docs.sparc.science/docs/overview-of-sparc-dataset-format
      ii. Template for the SDS 2.0 release can be downloaded at https://github.com/SciCrunch/sparc-curation/releases/tag/dataset-template-2.0.0
      iii. Webinar overview: https://youtu.be/_WE6BUY7Bbg
5. **Any additional comments on the openness of the SBP?**
   a. We welcome the use of this standard by anyone, it is open and we think the SDS is flexible enough to handle the structure of most experiments, models, or simulations. However, as the standard is integrated into the data shared on the

SPARC portal, this imposes certain technical constraints on standard modifications; as modifications have to be vetted and tested with all tools and resources that touch the standard on the SPARC portal before any changes can be integrated into the existing standard.

FAIR criteria:

Considers the SBP from the point of view of some (not all) of the FAIR criteria (Wilkinson et al. 2016). Is the SBP itself FAIR? Does it result in the production of FAIR research objects? Note that many of these may not apply. If so, leave blank or mark N/A.

1. **SBP uses/permits persistent identifiers where appropriate (F1).**
    a. Yes, though the system is designed to be agnostic to any particular system. The standard provides clear guidance for how to create good local identifiers within a dataset to make it possible to address content and metadata within a dataset easier when the dataset is identified by a persistent identifier (e.g. by a DOI).
2. **SBP allows addition of rich metadata to research objects (F2).**
    a. Yes. The SDS has a variety of ways that users can add arbitrary rich metadata for subjects, samples, specimens, digital resources, files, folders, protocols, and performances.
3. **SBP uses/permits addition of appropriate PIDs to metadata (F3).**
    a. Yes. Version 2 of the SDS includes a direct mapping to the datacite related identifiers schema to make inclusion of PIDs easier and more regular. SDS also explicitly recommends the use of RRIDs.
4. **The protocol allows for an authentication and authorization when required (A1.2).**
    a. Not relevant. The SDS is expected to be used in a variety of situations and is orthogonal to any auth/auth system.
5. **SBP uses or allows the use of vocabularies that follow the FAIR principles (I2).**
    a. Yes. We do not require the use of ontology identifiers in the standard but we make extensive use of them and encourage users to do so as well.
6. **SBP includes/allows qualified links to other identifiers (I3).**
    a. Yes. See response to F3 about datacite related identifiers.
7. **Does the standard interoperate with other relevant standards in the same domain? (I1).**
    a. SPARC accepts multiple different types of data derived from a variety of experimental approaches. Therefore a standard is required that can handle a wider variety of cases than those developed in many other domains. We have made attempts to align to standards such as BIDS, and DANDI's metadata structure, however the SDS is a superset of BIDS and thus not all datasets can be easily converted to BIDS.
8. **Does the SBP provide citation metadata so its use can be documented and tracked? (R1.2).**

    a. The standard itself has the paper that can be used as a reference, and we are in the process of registering an RRID.

9. **Does the SBP have a clear versioning scheme and appropriate documentation?**
    a. Yes. The dataset template follows a semantic versioning scheme. See links to user documentation above.
10. **Any additional comments on aspects of FAIR?**
    a. In addition to user documentation, there is also extensive documentation and examples for how to query for datasets that have been run through the SDS dataset export pipelines.
    b. https://github.com/SciCrunch/sparc-curation/blob/master/docs/queries.org
    c. https://doi.org/10.5281/zenodo.5337442 a integrated data release that contains queryable SDS metadata with links to how to run a docker image with examples.

Design, Testing, and Implementation

These may not all apply, if so, leave blank or mark N/A. Proper design, testing, and implementation, in addition to supporting tools greatly aid in adoption of a standard.

1. **What is the technical expertise level required to implement this? Even if it is quite difficult, should it be implemented anyway?**
    a. Implementing this standard does require some time and effort on the part of the data submitter. However, doing so puts their dataset into a FAIR format, allowing discovery on the SPARC portal (and theoretically beyond) and reusability of these datasets. There are multiple mechanisms in place to help users implement SDS during data submission. Using the standard (i.e. implementing it in an experimental laboratory) to impose file structure is fairly straightforward and can be done manually. Dozens of labs are actively using the SDS to submit data as part of SPARC. For those people, we have documentation and readily accessible curators to aid the process. In addition the SODA tool (referenced above) will help step people through the process and reorganize files as required.
    b. With regard to implementing a validator for the SDS there is a reference python implementation. The difficulty of implementing the validator depends on the balance between automated and human curation that will be used to enforce it. Anyone who can implement an ETL pipeline can probably implement the standard, however a simple ETL pipeline is not sufficient to ensure that the file system structure conforms to the standard and that it matches the metadata files.
2. **Does the SBP provide an architectural concept to understand its implementation and relationships to external entities?**
    a. Yes, we have an owl ontology that specifies all the major entities inside and outside a dataset.
3. **Does the SBP have a reference implementation?**
    a. Yes. https://github.com/SciCrunch/sparc-curation.

    b. A zipped Dataset template can be found at https://github.com/SciCrunch/sparc-curation/releases/tag/dataset-template-2.0.0. The development version is at https://github.com/SciCrunch/sparc-curation/tree/master/resources/DatasetTemplate

4. **What software artifacts (resources files/scripts/libraries/tools) are available to support the SBP?**
    a. The software tool SODA (Software to Organize Data Automatically) for SPARC uploads datasets into the SDS 2.0 format https://fairdataihub.org/sodaforsparc
    b. There is also a docker image that is used to run single dataset pipelines https://hub.docker.com/r/tgbugs/musl/tags?page=1&name=sparcur-user. The specification for building the image is here https://github.com/tgbugs/dockerfiles/blob/master/source.org#sparcur-user.
5. **Are the supporting software resources tools and implementations covered under an open source license?**
    a. Yes, open-source MIT License
    b. Are the supporting software resources well documented (documentation of I/O operations, programming interfaces, user interfaces, installation)?
        i. SODA: https://docs.sodaforsparc.io/docs/intro
        ii. https://github.com/SciCrunch/sparc-curation/blob/master/docs/developer-guide.org
        iii. https://github.com/SciCrunch/sparc-curation/blob/master/docs/setup.org
    c. Were the supporting software resources validated?
        i. Yes. We run regular unit and integration tests over all the parts of the pipelines, and they are run in production for the SPARC consortium.
    d. What is your assessment of the quality of the code/document?
        i. Speaking as the author, the current version does the job, but we are in the process of improving and regularizing ripping out and replacing some of the more … bespoke portions.
    e. Have the supporting software resources been deployed, is there any experience or references to their use by the community?
        i. Yes, we use the pipelines in production as mentioned. We are also in the process of getting the validator integrated into SODA, and plan to get an experience report from that team.
6. **Any additional comments on design, testing, and implementation?**
    a. If a group wanted to reuse the implementation on top of a separate data platform they would only need to add a new backend that would allow the pipelines to retrieve data (e.g. from S3 buckets directly, or from datalad, etc.). For example, an alternate ssh based backend already exists.

Governance

Ongoing governance is key to ensuring the transparency about how a standard was created, and ensuring the stewards are responsive to the needs of the community. Standards require transparent governance practices; however it is possible some of the following questions do not apply; if so, leave blank or mark N/A.

1. **Does the SBP have a clear description of how decisions regarding its development are made?**
    a. Not at this time
2. **Is the governing model document for maintenance and updates compatible with the INCF project governing model document and the open standards principles?**
    a. NA
3. **Is the SBP actively supported by the community? If so, what is the evidence?**
    a. At this time, the standard is imposed on the SPARC community, uptake beyond this community is unclear at this time. However, we will begin to accept some dataset submissions from outside SPARC and those will also be asked to use the SDS.
4. **Does the SBP provide tools for community feedback and support?**
    a. The latest version of the SDS (2.0) is just beginning to be used. The SPARC Curation team are actively interacting with the data contributors about their experience about the submission process. Once the data are offered on the Portal in this new format, data users will also be approached for feedback. In addition, there is an issue tracker on GitHUB https://github.com/SciCrunch/sparc-curation/issues where we will collect feedback and offer support. The SPARC portal also offers a mechanism to gather feedback.
5. **Any additional comments on governance?**
    a. At the moment we are awaiting outside interest before we commit to the overhead of spinning up a full governance structure. Once this happens we would start with the basic open source project governance using issues and pull requests to coordinate activity.

Adoption and Use

The standard must have substantive evidence of use outside of the group or individual that develops and maintains it. However, different levels of adoption and use will be taken into consideration depending on the purpose of the standard and the size/type of audience that might implement the standard. Because INCF represents organizations world-wide, evidence of international use is highly desirable.

1. **Is there evidence of community use beyond the group that developed the SBP?**

    a. The format is used by all labs in the SPARC consortium but they are required to do so, this covers 38 awards and about 500 investigators. To our knowledge it is not currently being used outside of this community.
2. **Please provide some concrete examples of use, e.g., publications where the use of the SBP is cited; databases or other projects that have adopted the SBP.**
3. **Is there evidence of international use?**
   N/A
4. **Any additional comments on use?**

Stability and Support

Standards need some sort of ongoing stability and support to ensure it will be useful in the future. However, given the nature of research projects, the level of acceptable stability and support is somewhat at the discretion of the SBP Committee and reviewers.

1. **Does the SBP have a clear description on who is maintaining the SBP?**
    a. We have multiple user support email addresses and the maintenance of the reference implementation is clear from the commit logs and activity on github.
2. **How is it currently supported?**
    a. NIH funded
3. **What is the plan for long term support?**
    a. This standard can be used by anyone as is beyond the SPARC project. Tools built on top of the standard would be useful to a certain extent as well.
4. **Are training and other supporting materials available?**
    a. Yes, refer to section Open Criteria #4
5. **Any additional comments on sustainability and support?**
    a. We had floated the idea of being able to merge some of the changes back into BIDS as part of BIDS 2.0, however that is far in the future, and at the moment the complexity of some of the experiments represented in SDS is beyond what BIDS can handle, making it difficult to merge those back into BIDS.

Extensibility

If it is possible to update or potentially apply the standard to other areas, that should be indicated. The ability for a standard to be extensible is highly desirable, but not required. This is an area where having this knowledge is valuable to INCF and the community in general.

1. **Can the SBP be extended to cover additional domains/use cases?**
    a. Yes. It was designed to handle a wide range of experiments, computation studies, simulations, and models. It is domain agnostic and could probably be used for fields as different from neuroscience as botany or geology if one were so inclined.
2. **If so, how is the process documented and managed?**

    a. At the moment, extensions to the standard are not expected to be needed for narrow use cases, the issue is more around how to use the existing standard to represent a particular experimental structure. If someone were to propose an extension, we would use the same process we used internally to extend it for simulations. This would be incorporated into a new version of the SDS.

3. **Any additional comments on extensibility?**
    a. General extensibility is provided by the fact that users can add whatever fields and columns they need to subjects/samples/manifests, etc. For example, more specific file type information was needed for certain use cases and a column was added. That column has been used informally for many months without issue, and we might at some point consider promoting it to a formally recognized field since it has proved to be useful. When that happens it would be incorporated into a new version of the standard.

Comparison

This information is important to INCF and the community. It may be important for evaluating one standard vs. another, or where INCF may help facilitate interoperability between complementary standards within a similar area.

1. **Are there other similar SBPs available?**
    a. Yes, BIDS.
2. **If yes, how do they compare on key INCF criteria?**
    a. We are missing governance, and we have a smaller user community with greater diversity of experiment types.
3. **What are the key advantages of the SBP when compared to other SBPs?**
    a. The key advantage of SDS compared to other SBPs is that it is designed from the bottom up to handle a wide diversity of types of experiments and it has been designed to meet investigators where they are with the tools they are used to using.
    b. Another advantage is that it can accommodate computational studies in addition to experimental studies.
4. **Any additional comments on comparison with other SBPs?**
    a. None.