# The SPARC Minimal Information Standard (MIS)

Authors: Tom Gillespie[1, 2], Bernard de Bono, [2, 3], Jyl Boline [2,4], Maryann E. Martone[1,2]

Affiliations:

[1] Dept of Neuroscience, University of California, San Diego
[2] SPARC Data and Resource Center
[3] Whitby et al., Inc.
[4] Informed Minds, Inc.

Acknowledgements: SPARC Data Standards Working group for initial recommendations

# Table of Contents

# Purpose:

This document provides a high level overview of the SPARC Minimal Information Standard (MIS) and its use to create the SPARC Knowledge Graph. The MIS is a semantic metadata model that specifies key metadata attributes and their relationships for SPARC datasets. This document provides the history, rationale, structure and content, and provides links to files and additional documentation. It is intended primarily as an internal document for developers or knowledge engineers who are part of the DRC or who will be working with the SPARC Knowledge Graph. Because some files contain both published and unpublished data, not all links are publicly available.

# Introduction

The SPARC Minimal Information Standard (MIS) is a semantic metadata model developed in OWL2 (https://www.w3.org/TR/owl2-overview/) to specify experimental metadata and the data set structure associated with SPARC data sets and other resources (Osanlouy et al. 2021; Bandrowski et al. 2021). Note that in various communications, it has also been referred to as the Minimal Information Specification, but Minimal Information Standard is the preferred name. The MIS is the basis of the SPARC Knowledge Graph which relates SPARC resources such as datasets, flat maps, and scaffolds (Osanlouy et al. 2021) to biological entities, provenance and experimental metadata. The initial MIS was developed through an iterative process by SPARC investigators, who formed working groups to determine a common set of metadata to be acquired for all datasets along with domain-specific metadata for certain data types. The initial data types considered were:

- Anatomical mapping (includes imaging)
- Functional mapping
- Transciptomics/informatics
- Computational (atlases, models)

In considering these domains, the working groups reviewed existing standards including the Minimum Information about a Cardiac Electrophysiology Experiment (MICEE) and a pre-release of the HBP-Open MINDS. The further development of the MIS was taken over by the SPARC Curation and Knowledge Management (K-Core) team when they joined the project in 2018. At that time, only the anatomical mapping model had been implemented.

# Accessing the MIS

The MIS is maintained in an open GitHub repository and is available at: https://github.com/SciCrunch/NIF-Ontology/blob/sparc/ttl/sparc-methods.ttl. The .ttl file can be viewed in Protege 5.5.0. This file provides the MIS data model in OWL2, including classes and predicates. However, objects appearing in the MIS may not have their class definitions in the MIS because they are defined in external community ontologies such as UBERON. These large ontologies can be loaded alongside the MIS in Protege but are not automatically included in the sparc-methods.ttl file to simplify viewing. A list of ontologies referenced in the MIS are provided in Table 1. These ontologies may be imported individually, or via an import of the full developmental version of the NIFSTD ontology (Bug et al. 2008) which imports all ontologies

currently in use with the exception of the FMA. The developmental version of NIFSTD is available at https://github.com/SciCrunch/NIF-Ontology/blob/dev/extra.ttl.

***Table 1: Ontologies used in MIS. Those marked with an * are included in the NIFSTD***

| Entity Type | Ontology | Comments |
|---|---|---|
| Information entities | Information Artifact Ontology (IAO)* | |
| Organisms | NCBI Taxonomy* | |
| Upper ontology | BFO* | |
| Qualities | PATO* | |
| Techniques | NIFSTD* | |
| Relations | Relations ontology (RO)* | |
| Anatomical terms | UBERON*, FMA, EMAPA | |
| Phenotypes | Human Phenotype Ontology | Imported for future use |
| MONDO | Diseases | Imported for future use |

***Table 1: List of external ontologies used by or potentially used by the MIS. Those marked by * in column 2 are automatically imported via the developmental version of NIFSTD.***

## SPARC Knowledge Graph

The MIS is the basis for the SPARC Knowledge Graph which contains SPARC datasets modeled using the MIS. The full SPARC Knowledge Graph comprises the following files

- MIS (sparc-methods.ttl): The MIS data model
- External ontologies (Table 1)
- Curation-export.ttl: Contains instances of SPARC datasets modeled according to the MIS. Available upon request.
- Protcur.ttl: Contains metadata details about protocols derived from the translation of the SPARC protocol curation pipeline. Available upon request

To support the current use of MIS in the SPARC curation workflow [see (Bandrowski et al. 2021) for details about the workflow], a developer guide is available.

The SPARC Knowledge Graph may be accessed in multiple ways;

1) **Protege**: Load the SPARC curation export .ttl file. This file is available on request because it includes both published and unpublished data. Additional ontologies must be loaded as per above.
2) **Triple store (Blazegraph)**: The SPARC Knowledge Graph is available as RDF via a SPARQL endpoint. In the future a public version will be made available that contains only the published data.

3) **SciGraph:** The full knowledge graph is available in a development version of SciGraph, a neo4j graph database, and is available on request. This version contains minimal external ontology support. A public version of the knowledge graph will be released in Sept 2021 which will also include all required external ontologies. Datasets will be available via a neo4J graph database.

Additional information on accessing the SPARC Knowledge Graph can be found at: https://github.com/SciCrunch/sparc-curation/blob/master/docs/queries.org#ontology-content

# Structure of MIS

A summary of the current structure of the MIS is shown in Fig 1 as of August 1, 2021 (MIS 2.0). The current MIS reflects the types of resources produced by SPARC and their organization. A more detailed explanation of SPARC datasets and other products is provided in (Osanlouy et al. 2021; Bandrowski et al. 2021).

The MIS contains a series of classes in a mostly flat hierarchy under OWL:Thing comprising entities associated with descriptions of datasets and other resources produced by SPARC, and experimental details for specific data types. Entities are imported from or mapped to existing ontologies where possible. A list of ontologies used in the MIS is given in Table 1.

The MIS is undergoing active evolution and so there may be slight deviations from the class and property structure noted here.

## Resource description

The main classes currently used by SPARC to describe a resource are given in Table 2:

| Class | Definition |
|---|---|
| Resource | Any digital artifact. Closely follows the RDF meaning of resources. Examples of subclasses are, file, folder, dataset, maps and scaffolds. |
| Dataset | Child of resource; "The atomic unit of publication of a nested collection of files and folders. Within SPARC this also implies that the dataset should follow the SPARC Data Structure (SDS). Metadata may be shared between datasets but all metadata relevant to the scientific content of the dataset must be contained inside of it, so that it is at least minimally self describing. For example a dataset as defined by the Pennsieve platform. |
| Protocol | An information content entity that contains instructions for performing a scientific experiment or procedure. Protocols are prior informational constraints on scientific processes. |

| | |
|---|---|
| Person | An agent or being of some variety, usually classified as such because they are participating in activities modeled by this ontology in some way. |
| Participant | Comprises experimental samples, subjects, specimens, and populations used within a study. This is probably more accurately “Material Entity,” but one that we have data about since it is appearing in some information artifact. Despite the fact that we are using the word participant in the identifier, this class is meant to abstract over atomic specimens and collective (but opaque) populations which participate in some process. If there were a better term that subsumed both atomic and collective beings that wasn't material entity (object + object aggregate), we could go with that. Normally this would be role in BFO and this classification would be inferred, because technically all continuants participate in a dual occurent that places them in time. See https://github.com/SciCrunch/sparc-curation/blob/master/docs/participants.org for details. |

***Table 2: Main classes used in the SPARC Knowledge Graph. The mappings of these classes to the IAO is given in the last column.***

SPARC datasets are represented in the SPARC Knowledge Graph as individuals of Class:Dataset and are related to other classes and metadata through a set of object, data type and annotation properties. A high level overview of how SPARC datasets are represented in the MIS is shown in Figure 1 while a specific example is shown in Figure 2. A human readable list of properties used in SPARC and their definitions is available here and a more complete documentation is available here. At this time, we make limited use of OWL axioms.

- Dataset
  - isAboutParticipant Participant isA Subject
  - hasResponsiblePrincipleInvestigator Person hasFirstName, hasLastName
  - hasContactPerson Person
  - hasURIHuman Dataset (label = Link for human consumption)
  - hasURIApi Dataset (label = Link for machine consumption)
  - hasAwardNumber Funded Research Project
  - hasProtocol Protocol
  - hasExpectedNumberofSubjects → Integer
  - hasExpectedNumberofSamples → Integer
- Subject
  - hasBiologicalSex PATO
  - hasDerivedInformationAsParticipant Dataset
  - animalSubjectHasWeight → Integer
  - animalSubjectIsOfStrain → String
  - stimulatorUtilized → String

- Protocol
  - hasURIHuman
  - hasDOI
  - protocolEmploysTechnique → Interlex URI
  - involvesAnatomicalRegion UBERON

***Figure 1: A subset of the MIS property graph. Entities are color coded as follows: Gold = OWL classes, Blue = object properties, Green = datatype properties.***

Datatype properties relate individuals to properties such as number of subjects or subject weight. Current data type properties used in SPARC are String, Date, DateTime, Double, and Integer. Double refers to 64bit floating point.

Annotation properties provide standard Dublin Core annotation metadata, supplemented by additional properties that provide details about the data set, e.g., wasCreatedAtTime, hasSizeInBytes. Annotation properties are also used as a short cut to relate key properties to datasets without having to go through the property chain. See the examples provided in the Protocol section below.

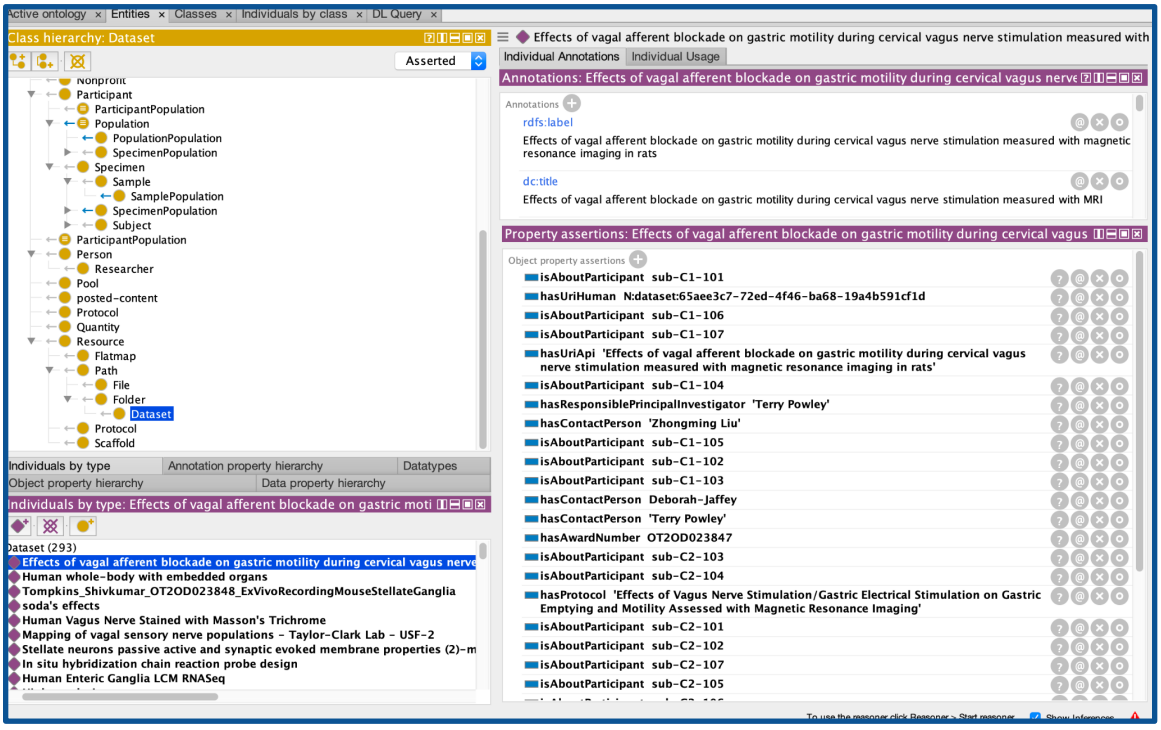


***Figure 2: An example of an individual dataset modeled according to the MIS viewed in Protege.***

## Duplicate properties

There are several cases where the same property is present across property types. MIS 1.0 made extensive use of datatype properties for capturing key metadata in the form of strings based on the original requirements established by SPARC working groups. As the MIS has

evolved, many of the original data type properties are being re-implemented as object properties where the object is now an ontology class. This migration results in the presence of the same property as a data type and object property. When such a duplication is encountered, the object property should be used.

As described above, an apparent duplication of properties also occurs across annotation and object properties. However, in this case, the annotation property version is usually used as a shortcut for a fully normalized object property chain, as illustrated by the example relating techniques to datasets shown in Figure 3.

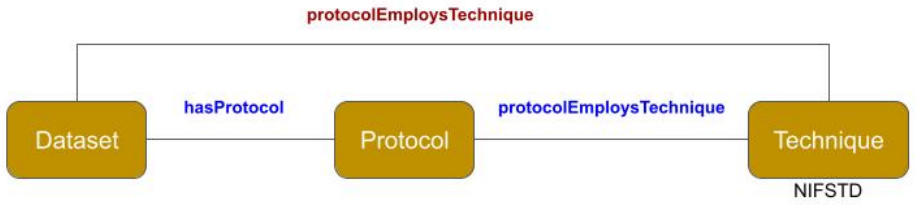


**Figure 3: Use of annotation property as a shortcut for a fully normalized object property chain. In this case, the technique used in a dataset can be derived from the object property chain (blue) by way of Protocol or directly from the annotation property (red).**

In other cases, the annotation property is used for raw information provided by the investigator while the corresponding object property is used for normalized or modified information. An example of this use is provided in the **wasExtractedFromAnatomicalRegion** example given the Anatomical Information section.

# Experimental details

## Participants

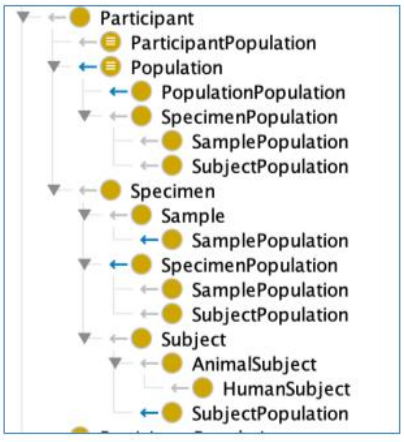


***Figure 4: Asserted hierarchy for participant***

In the MIS, **participants** are the things that are measured to derive a dataset. The asserted hierarchy for the participant class is shown in Fig 4. The two main branches are **Population** (equivalent to Participant Population) and **Specimen**. Populations are collective participants that may be opaque to their individual members but still have relations to generated data and sample/specimen derivation chains.

**Specimen** encompasses **Subject**, i.e., organism, and **Sample** derived from an organism, e.g., a tissue slice, a cell. All specimen types have a corresponding specimen population class which can be used to describe a collection of participants, e.g., an experimental group. The class **Pool** is used when the measurements are derived from multiple participants that cannot be distinguished as individual, e.g., a pooled sample.

For more information about the treatment of participants et al in MIS, see https://github.com/SciCrunch/sparc-curation/blob/master/docs/participants.org

### Protocols

Protocol is a primary resource type in the MIS related to the dataset via the **hasProtocol** object property. All SPARC investigators are required to provide a detailed experimental protocol deposited in Protocols.io for each dataset. Each protocol is therefore identified by a DOI, once published, or an internal protocols.io identifier prior to publication. Experimental techniques are related to datasets via the protocol, using the **protocolEmploysTechnique** *object* property. The **protocolEmploysTechnique** *annotation* property has been used to relate techniques directly to a dataset Techniques in SPARC are described through methods ontology in NIFSTD, and may be viewed through Interlex (https://interlex.org), a web-based ontology management system. Each Interlex term is given a unique URI, which is currently used as the object for this property.

Details about protocols are split across two .ttl files. The links between Protocol and protocol files via DOIs or other identifiers are contained in the main curation-export.ttl file. More granular details about the contents of the protocol are derived from a curation workflow involving tagging with the on-line annotation tool Hypothes.is. These tags, which capture entities like chemicals or technical details are contained in a separate file, protcur.ttl.

### Anatomical information

Anatomical information is related to datasets through three primary predicates:

**involvesAnatomicalRegion**, object property applied to a protocol that takes as its object an anatomical region from either UBERON, FMA or EMAPA.

**wasExtractedFromAnatomicalRegion**, an annotation property that takes a participant (mainly sample) as its subject and an anatomical region in free text as its object. We haven't been able to normalize these to ontology IDs because of the complexity of the anatomical details provided, e.g., "anatomical region 10 mm from esophageal junction". This free text is provided under the TEMPRAW and TEMP namespaces (see Internal curation), indicating that at some future time, these may be normalized to the appropriate ontology classes. TEMPRAW denotes unmodified data provided by the investigator such as the example provided above, while TEMP denotes text that has been mapped to an anatomical entity and will be normalized in the future using an object property.

**isAbout**: A catchall relationship for keywords and is used for anatomical entities that are not the primary subject but which are useful for search. Note that it may show up in Protege as **IAO:0000136.**

## Internal curation

All SPARC datasets that have been submitted to the Pennsieve platform (formerly Blackfynn) are ingested into the MIS using automated pipelines (see Supporting software), regardless of

their state of curation. Metadata from the MIS.ttl file is automatically extracted from the .ttl file and used to populate equivalent metadata fields within Pennsieve. The associated JSON representation of the .ttl file is used by K-Core to populate the Elasticsearch and Algolia index endpoints via Foundry, a message-oriented, horizontally scalable ETL system for scientific data integration and enhancement. The curation-export.ttl therefore contains both datasets that are under active curation and those that have been curated and published or still under embargo.

During curation, a set of automated scripts extract metadata and validate that file organization, naming and metadata adhere to the SPARC SDS. These scripts produce information that is used by the curation team, recorded in a set of annotation properties, e.g., **errorIndex**, **curationndex**, **milestoneCompletionDate**. These properties are not meant to be uploaded to Pennsieve and are identified with the predicate: ilxtr:curationInternal = true. Any url starting with TEMPRAW: -> http://uri.interlex.org/temp/uris/raw/ should be excluded from import as indicated by the presence of the triple `TEMPRAW:` ilxtr:curationInternal true.

As noted above, the TEMP RAW namespace is also used for data that has not yet been normalized, e.g., the **wasExtractedFromAnatomicalRegion**. These properties are always annotation properties. Normalized versions are under the TEMP versions in general.

The objective is to replace these relationships with versions that are not temporary, but as there are external dependencies involving the Pennsieve platform that use these temporary ID's, we continue to retain them so as not to break the import.

## Evolution of the MIS

The MIS has evolved as the SPARC project has evolved. The first version of the MIS was produced before the current organization of SPARC datasets and resources were produced, and therefore reflected a more general data model for experimental datasets. In 2018-2019, SPARC adopted a standard way for organizing and documenting datasets called the SPARC Dataset Structure (SDS; (Bandrowski et al. 2021). The MIS evolved to align with this structure and has also been adapted as additional standards have been implemented, e.g., the SPARC Imaging Metadata Standard (Tappan and Martone 2021). Aspects of the original MIS have also been refined to make it more efficient and to align it with best practices for ontology design. As noted above, the MIS contains a set of temporary classes and properties that are being replaced over time.

MIS 1.0 was released in Sept of 2018 ahead of the first round of data submissions by SPARC investigators. MIS 1.0 was used to model metadata associated with datasets submitted according to the SPARC Dataset Structure (SDS) as detailed in Bandrowski et al., (2020). MIS 1.0 established a common metadata model for describing datasets, and detailed metadata for anatomical mappings. All metadata were applied at the level of "Dataset" and not to individual components of a SPARC dataset.

MIS 2.0 was released in August 2020. V2.0 added classes and properties for additional resource types, mainly the 3D scaffolds used for spatial registration of SPARC data (Osanlouy et al. 2021) and also direct support for the file and folder structure required by the SDS:

sparc:Dataset, sparc:Folder, sparc:File, and sparc:Path. These classes lay the foundation for more granular search through SPARC datasets, by allowing us to attach metadata to individual folders and files. A folder in the SDS, for example, may be associated with a single experimental modality. SPARC datasets are automatically recurated to the 2.0 through the use of automated metadata extractors.

MIS 3.0: is scheduled for release in Fall 2021. The major additions will be: the SPARC Minimal Information Standards for Optical Microscopy; classes and relationships necessary for modeling simulations via Sim Core; deeper metadata for physiology, once the proposed metadata standard for physiology has been implemented.

## Supporting software

The MIS is populated almost exclusively through automated extraction tools via the metadata provided through the SDS by SPARC Investigators. A limited number of fields are supplied by curators (organ term, modality/approach and technique) via a Google Doc.

Populating the SPARC Knowledge Graph is supported by the SPARC curation pipelines, which include metadata extractors, and multiple levels of validation, integration and export to a variety of formats. Curation pipelines can be accessed https://github.com/SciCrunch/sparc-curation . Validation of data that is exported to MIS involves validation against several JSON schemas available in that same repository. We are in the process of documenting these schemas more fully.

## Helpful Links

MIS
https://github.com/SciCrunch/NIF-Ontology/blob/sparc/ttl/sparc-methods.ttl

NIFSTD developmental version
https://github.com/SciCrunch/NIF-Ontology/blob/dev/extra.ttl.

MIS releases
https://github.com/SciCrunch/NIF-Ontology/releases/tag/sparc-mis-2.0

MIS internal structure details
https://github.com/SciCrunch/sparc-curation/blob/master/docs/participants.org

MIS predicate reports
sparc automated pipeline reports

Curation pipeline overview
https://github.com/SciCrunch/sparc-curation/blob/master/docs/background.org

Curation pipeline details
https://github.com/SciCrunch/sparc-curation/blob/master/docs/developer-guide.org

https://github.com/SciCrunch/sparc-curation/blob/master/docs/images/sparc-curation-pipelines.png

Protcur
https://cassava.ucsd.edu/sparc/preview/exports/protcur.ttl

## References

Bandrowski, Anita, Jeffrey S. Grethe, Anna Pilko, Tom Gillespie, Gabi Pine, Bhavesh Patel, Monique Surles-Zeigler, and Maryann E. Martone. 2021. “SPARC Data Structure: Rationale and Design of a FAIR Standard for Biomedical Research Data.” *bioRxiv*. https://doi.org/10.1101/2021.02.10.430563.

Bug, William J., Giorgio A. Ascoli, Jeffrey S. Grethe, Amarnath Gupta, Christine Fennema-Notestine, Angela R. Laird, Stephen D. Larson, et al. 2008. “The NIFSTD and BIRNLex Vocabularies: Building Comprehensive Ontologies for Neuroscience.” *Neuroinformatics* 6 (3): 175–94.

Osanlouy, Mahyar, Anita Bandrowski, Bernard de Bono, David Brooks, Antonino M. Cassarà, Richard Christie, Nazanin Ebrahimi, et al. 2021. “The SPARC DRC: Building a Resource for the Autonomic Nervous System Community.” *Frontiers in Physiology* 12 (June): 693735.

Tappan, Susan, and Maryann Martone. 2021. *SPARC Optical Microscopy Imaging Data and Imaging Metadata Standard*. https://doi.org/10.5281/zenodo.5347993.