SPASE metadata as a building block of a heliophysics science-enabling framework

Shing F. Fung, Arnaud Masson, Lee F. Bargatze, Todd King, Rebecca Ringuette, Robert M. Candey, Chiu Wiegand, Lan K. Jian, Darren de Zeeuw, Karin Muglach, et al.

Available online at www.sciencedirect.com

ScienceDirect

Advances in Space Research 72 (2023) 5707–5752

ADVANCES IN SPACE RESEARCH (a COSPAR publication)

www.elsevier.com/locate/asr

SPASE metadata as a building block of a heliophysics science-enabling framework

Shing F. Fung [a,*], Arnaud Masson [b], Lee F. Bargatze [c,d], Todd King [d], Rebecca Ringuette [e,a], Robert M. Candey [a], Chiu Wiegand [g,a], Lan K. Jian [a], Darren De Zeeuw [h,m], Karin Muglach [h,a], Ryan M. McGranaghan [i,a], D. Aaron Roberts [a], Baptiste Cecconi [f], Nicolas André [j], V. Génot [j], Jon Vandegriff [k], Martin A. Reiss [l,m]

[a] *Heliophysics Science Division, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA*
[b] *European Space Agency (ESA), European Space Astronomy Centre, Camino Bajo del Castillo s/n 28692, Villafranca del Castillo, Madrid, Spain*
[c] *Department of Earth, Planetary, & Space Sciences, Univ. of California, Los Angeles, CA 90095, USA*
[d] *Institute of Geophysics and Planetary Physics, University of California, Los Angeles, CA 90095, USA*
[e] *ADNET Systems, Inc., Bethesda, MD 20817, USA*
[f] *Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique (LESIA), Observatoire de Paris-PSL, CNRS, Meudon 91290, France*
[g] *Instrument Systems and Technology Division, NASA GSFC, Greenbelt, MD 20771, USA*
[h] *Catholic University of America, Washington, DC 20064, USA*
[i] *Orion Space Solutions, Louisville, CO 80027, USA*
[j] *Institut de Recherche en Astrophysique et Planétologie, CNRS-UPS-CNES, 31028, Toulouse, France*
[k] *Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723, USA*
[l] *Austrian Space Weather Office, Zentralanstalt für Meteorologie und Geodynamik, Graz 8020, Austria*
[m] *Community Coordinated Modeling Center, NASA Goddard Space Flight Center, 8800 Greenbelt Rd., Greenbelt, MD 20771, USA*

Received 7 February 2022; received in revised form 26 September 2023; accepted 28 September 2023
Available online 4 October 2023

Abstract

Heliophysics and space weather research encompass the effects of solar output on practically the entire Solar System and are fundamentally cross-disciplinary. Cross-domain science investigations, such as in Sun-heliosphere interactions, solar wind-magnetosphere interactions, or magnetosphere-ionosphere coupling, often require the use of data, models, and other digital resources pertaining to different heliophysical domains: the Sun, the solar wind, the magnetosphere, the ionosphere, the thermosphere and the mesosphere. Due to differences in measurement platforms, techniques and instruments, heliophysics data obtained from different domains are diverse and complex, making the resource landscape difficult for untrained users to navigate. Without proper and adequate guidance from domain experts, it is often difficult for early-career scientists and non-domain experts to discover useful datasets and to know from where and how to obtain and understand the data they need to support their research. This paper describes the roles of metadata in providing the identification, location, access protocol, and detailed content description of a digital resource. More specifically, we point out that metadata written according to the Space Physics Archive Search and Extract (SPASE) metadata model are fully compatible with the FAIR principles so that digital resources described using the SPASE model can be uniformly Findable, Accessible, Interoperable, and Reusable. SPASE metadata can thus be the key element, the lingua franca so to speak, that enables unfettered information flow between data systems and services throughout the heliophysics data environment and lowers the understandability barrier of the resources to ensure their independent usability. After describing various components of the heliophysics data environment, their metadata requirements for effective operations, and some essential features of the SPASE metadata model, we then illustrate how metadata in SPASE can enable or

* Corresponding author.
E-mail address: shing.f.fung@nasa.gov (S.F. Fung).

https://doi.org/10.1016/j.asr.2023.09.066
0273-1177/Published by Elsevier B.V. on behalf of COSPAR.

facilitate the performance of different science tasks. The current status and future outlook of SPASE are also presented.

Keywords: Heliophysics; Space Weather; Metadata; FAIR principles; Informatics; Data archiving and access

1. Introduction

It is often said that because of the large variety of measurement techniques, instrument types, and product types, etc., involved in supporting heliophysics research, heliophysics data are generally complex and heterogeneous. Data products and other digital resources in different heliophysics domains or disciplines (solar, heliospheric, magnetospheric, and ionospheric) may also be different in the ways they are stored (location, format), served (access protocol), and used (domain knowledge, method of analysis). However, any research task involving the use of heliophysics data, either from observations or models, and other digital resources must begin by acquiring the relevant resources, such as the required datasets, analysis and visualization tools, and the documentation needed to understand the correct use of the resources. The complexity and difficulty of the research task then depend on how easy it is to address each of the following questions:

(1) What are the required resources?
(2) Where are the required resources located and stored?
(3) How can the identified resources be accessed and retrieved?
(4) Are the resources obtained understandable, so they are readily and independently usable?
(5) Can the newly obtained resources be used in conjunction with existing resources?

Cross-disciplinary research requiring resources from different disciplines, in particular, can be hampered if discipline resources are not findable and discoverable [questions (1) and (2) above], accessible [question (3)], interoperable [questions (2), (3) and (5)], and re-usable [questions (4) and (5)], i.e., not compliant with the FAIR principles (Wilkinson et al., 2016) across all heliophysics disciplines. The general question we would like to address in this paper then is: Given their complexity and heterogeneity, **can heliophysics digital resources be made FAIR-compliant by simply imposing a uniform description scheme, i.e., a standard metadata model that can describe resources with sufficient detail?** To address this general question, we first examine how each of the above questions relate to resource descriptions as indicated above, i.e., their metadata requirements**.** We then explore the capabilities of the Space Physics Archive Search and Extract (SPASE) metadata model (Roberts et al., 2018) and consider SPASE as a standard model for heliophysics metadata. Finally, we look at various types of science studies: event analysis, statistical studies, data-model comparisons (model testing and validation) and even data science analysis (e.g., machine learning) to see how SPASE metadata facilitates heliophysics and space weather research by removing or minimizing the stumbling blocks typically associated with the above questions.

It is important to note the difference in focus of the present paper from the SPASE model documentation (available from the SPASE website) and from the earlier paper by Roberts et al. (2018). The present paper primarily explores how the SPASE metadata model would actually facilitate various science tasks rather than just describing the metadata model. However, some basic description of the SPASE model is still needed to facilitate discussions in the paper. In doing so, we hope to also make the informatic description of the model more understandable and accessible to a broader scientific community.

A brief outline of the paper is as follows. We first describe in section 2 how metadata in general is required to support the operations and functionalities of the heliophysics data environment and supporting infrastructure to search, locate, access, and deliver digital resources to a user carrying out a given research task. Section 3 provides an overview of the SPASE metadata model and its specific capabilities for resource descriptions and referencing and gives a behind-the-scenes illustration of how the SPASE schema captures all the essential descriptive information on digital resources, ranging from observational data to simulation models and model data, along with their digital object identifier (DOI) references. It then illustrates how resource locations and access mechanisms are identified, and how information detailing their contents, format, and caveats, etc., would lead to a better understanding of the resources so users can use them independently. Next, we describe in section 4 how the SPASE model can actually be used as a standard to facilitate various scientific tasks, such as finding, accessing, and visualizing data, supporting data-model comparisons, and providing persistent references to the resources so that they can be found again, understood, and reused correctly and independently by the international research community. The subsections therein present a few examples of how different types of studies can be more easily carried out with the support of SPASE. Finally, sections 5 and 6 provide an outlook for the future of SPASE and some conclusions, particularly on how the FAIR principles (for Findability, Accessibility, Interoperability, and Reusability of digital assets) (Wilkinson et al., 2016) can be supported to enable reproducibility of research results. A few appendices are also

included at the end of the paper to provide a list of acronyms and websites referred to in the paper (Appendix A), a mapping of global and variable (parameter) attributes between Common Data Format (CDF) and SPASE metadata (Appendix B), and a list of systems that use SPASE compliant metadata (Appendix C).

2. Infrastructure supporting the heliophysics data environment

Questions (1)-(5) above and the subsequent general question raised in section 1 all pertain to enabling the acquisition and the ease of utilization of the resources needed to support a given science analysis. Except for Question (1), which is derived from science requirements, the answers to the other questions rest upon the effectiveness of the information architecture and supporting infrastructure in providing the required resources to the user, even if not previously known or expected. We describe in the following subsections the heliophysics data environment and its supporting infrastructure, and in particular the need for metadata to effectively enable various functionalities.

2.1. Heliophysics data environment

For some, "data environment" may seem like a nebulous term or concept. The word "environment" conjures up the idea of the "look and feel" of a room, a house, or a neighborhood. It could also refer to the scenic view of a field, landscape, or just nature. Sounds and other sensations, natural or otherwise, add to the environment. In a city, all the buildings, structures, roads, highways, and whatever else that make up the cityscape, and how they interact visually, physically and operationally, create the environment of the city. Architecture and infrastructure operations are thus important aspects to an environment. Fig. 1 shows the components of the information architecture of the heliophysics data environment and the information flows between them, represented by the arrowheads. In total, the information architecture provides the framework for developing the infrastructure that forms the heliophysics data environment. Operations in different parts of the heliophysics data environment however are different, depending on the information sources and services involved, and how resources are discovered, accessed, retrieved, and consumed. It is thus not straightforward to describe generally the operations throughout the data environment except for the general requirement that resources must be able to flow freely within the environment (as depicted by Fig. 1), or operations would stop. **How good or "pleasing" a data environment is depends on how effective the information architecture or infrastructure is in supporting the flow of information (i.e., resource) through different parts of the architecture and enabling consumption of the resource.** Finally, we should emphasize that the word "data" in "data environment", or in this paper for that matter, represents not just data (observational or otherwise), but any digital resource being transferred throughout the data environment and utilized by users.

Fig. 1 shows the different components of the heliophysics information architecture (blue boxes), information flows (thick blue arrows) and the interaction pathways (thin arrows) between the users (red oval) and the infrastructure components: data sources (data producers, large databases, and repositories), modeling centers, data visual-

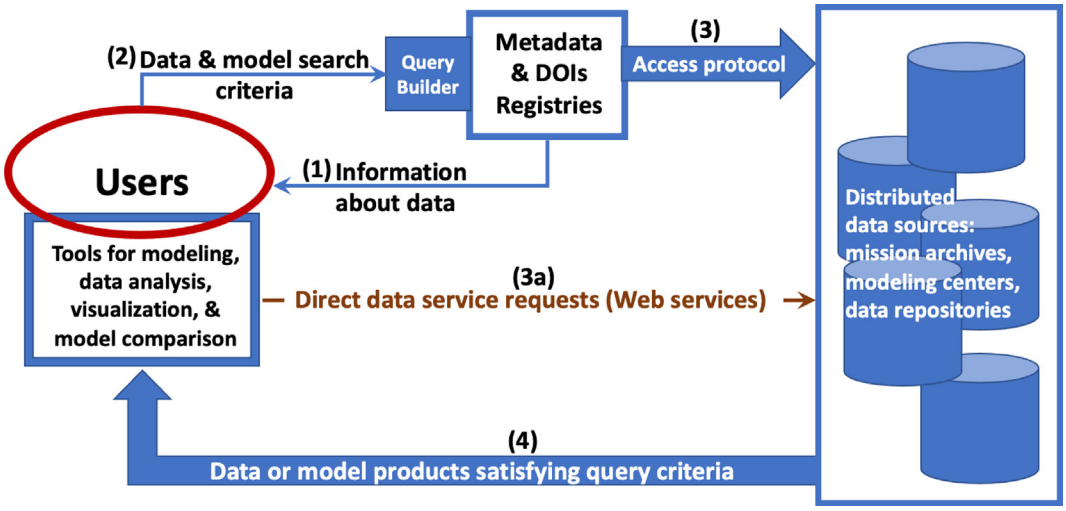

Fig. 1. A schematic of information (e.g., data) flows between different components of the heliophysics information architecture. The numbers associated with each segment of the information flow pathways indicate the operational order of information or data transfer within the heliophysics data environment. Users (red oval) would first obtain information (metadata) and gain understanding of registered resources along path (1), then discover, locate, and access their needed resources either by relevant data services through their interfaces along paths (2) and (3) or directly using Web services via APIs along path (3a). Required resource types and resources satisfying users' queries would then be returned along path (4) for consumptions in various science tasks captured collectively in the square box. More detailed illustration of information flows in support of various science tasks is shown later in Fig. 9.

ization and analysis tools, data services (e.g., processing, search, access protocols, and citation), as well as the underlying cyberinfrastructure on which a large amount of data may be stored and processed. **In terms of creating the effective supporting infrastructure, efforts should focus on setting up the appropriate services that would support unfettered flows of resources and effective understanding of the resources to ensure their proper utilization or consumption.** In a traditional data environment in which data was served primarily by a centralized data archive, the archive incorporated all the data querying, accessing and storage functionality into one entity so users could only obtain the resources stored at that archive. Searching and accessing resources at different archives would require the use of different interfaces. As we will elaborate in section 4.2, the information architecture depicted in Fig. 1 will be more effectively served by a middleware dedicated to formulating and executing resource querying and accessing, including resource discovery. The middleware approach that operates on metadata only would enable distributed and network-accessible resources to be more effectively used. But this approach would require proper and adequate metadata descriptions of the resources to enable such interoperability.

2.2. Supporting infrastructure

In this subsection, we consider each of the components of the heliophysics information architecture shown in Fig. 1 to understand how metadata in general supports the functionalities of various infrastructure components in the heliophysics data environment. Table 1 below lists the various components of the data environment infrastructure (left column) and their corresponding functionalities within the environment (middle column). The right column summarizes the significance of metadata, i.e., its role in enabling the functionalities of each of the data environment components. In subsections 2.2.1–2.2.5 below, we consider each case in more detail.

2.2.1. Data sources

Data sources are critical components of the information architecture. While users may request digital resources via segment (3) or (3a) in Fig. 1, the requested information is then supplied to the users or to analysis and modeling tools directly through segment (4) for consumption.

Observational heliophysics data (section 4.1.1) are produced originally from measurements by many spaceborne, sub-orbital, and ground-based experiments, including single spacecraft, multi-spacecraft constellations, small-

Table 1
Metadata has different roles and significance in supporting various functionalities in the heliophysics data environment. As noted in the Table, metadata is the key to enable finding, accessing and independent use of digital resources; standard metadata, however, is the basis for interoperability.

Heliophysics Data Environment Infrastructure Components	Functionalities	Significance of metadata
Data sources (e.g., instrument teams of space missions or ground instruments, researchers who produce higher level data products from research projects)	· Production or compilation of data resources · Generation of resource documentations · Distribution and archiving of data resources	· Origins of data resource descriptions (“F” in FAIR) · Source of expertise that can ensure metadata accuracy· Source of original information to ensure understability and correct usability of resources (“R” in FAIR)
Data repositories (e.g., SPDF, SDAC, ESAC, IUGONET)	· Archiving and long-term preservation of data resources · Distribution of resources for community utilization	· Proper identification, descriptions, storage, tracking, and maintenance of data resources (“F” in FAIR)· Enabling finding, accessing, and understanding of archived resources and independent reuse of the resources (“F,” “A,” and “R” in FAIR)
Modeling centers (e.g., CCMC, VSWMC)	· Support large-scale computer simulations and analysis of modeling results · Making accumulated modeling results available for post-analysis and reuse by the community	· Proper descriptions (with identifications), storage, tracking, and maintenance of models, model setups for different runs and corresponding modeling outputs· Enabling search, access, and understanding of modeling artifacts and ensuring independent use of the resources (“F,” “A,” and “R” in FAIR)
Data format (e.g., CDF, FITS) **and service** (e.g., HAPI, TAP) **standards**	· Uniform access to digital resources	· Interoperability (“I” in FAIR) is facilitated by the adoption of metadata standards
Data visualization and analysis tools (e.g., Autoplot, SPEDAS, PyHC packages)	· Browsing or inspection of digital resources for physical insight into the systems or processes under investigation · Plotting or graphical representation of data · Quantitative data analysis	· Need information on access protocol and data format for retrieving and manipulating resources· Need parameter descriptions such as coordinate systems, units, caveats, time cadences, extremum values, etc. for plotting and proper performance of analysis and interpretation of results (“R” in FAIR)

sats, rockets, balloons, aircraft, and various types of ground-based instruments. Data are also produced from simulation model runs (section 4.1.2), while higher-level products (section 4.1.3) can also be generated from various analysis and research projects including datasets built upon multiple sources (e.g., the OMNI datasets, https://omniweb.gsfc.nasa.gov/). Consequently, heliophysics data are diverse in data types and in disciplines, complex in terms of the differences in the characteristics of the instruments with which the data were obtained, and otherwise different in terms of their data formats, storage locations and methods of access.

Due to the diversity and complexity of heliophysics data, data products produced from one discipline may not be readily accessible or consumed by users in different disciplines without having some familiarity and understanding of the discipline data products. For a given data product to be discoverable, accessible, understandable, and thus usable, particularly by users trained in different disciplines, it is imperative for the product to be uniquely identified and adequately described by the resource producer, who presumably should have the most accurate knowledge and expertise on their data products. In time, the original resource description will become the best, if not the only, source of information on the product; so, it is important to preserve the original, most-detailed metadata and its association with the product, particularly if the management and storage of the product might be transferred or migrated over time.

As alluded to above, cross-disciplinary heliophysics research is often hindered by the complexity and diversity of data and models, making them less accessible and understandable by different discipline users. As we will discuss in section 2.2.4 below, metadata standards can lead to uniform descriptions of otherwise heterogeneous data and model resources, and thus can facilitate cross-disciplinary research by lowering the barriers that tend to hinder interoperability.

2.2.2. Data repositories

In the heliophysics domain, there are a number of long-term active archives that serve the community generally while some operating missions may also operate their own data services. Examples of significant infrastructure class and several mission-specific data repositories operating presently (below the black line) are listed in Table 2 below. All ESA heliophysics science mission archives are operated by the European Space Astronomy Centre (ESAC) Science Data Centre (ESDC).

Each of these archives and data services have their own way of identifying data products, implementing search capabilities and accessing methods. They thus have different metadata requirements for finding and accessing the data. To ensure usability of the data, a data repository also needs to capture from the data provider specific information about the data content, such as data format and parameter descriptions, such as, coordinate system used, physical units, temporal cadence, etc. For instance, the popular NASA Coordinated Data Analysis Web (CDAWeb) interface https://cdaweb.gsfc.nasa.gov/ allows selection by mission name and general instrument types, with datasets and variables described by text strings and time ranges, as provided by the ISTP metadata internal to the datasets; but the system works only with data stored in specific data formats, CDF and netCDF (see section 2.2.4.1). Some archives also provide direct access to data files in a directory hierarchy, where selection metadata are essentially the names of the directories and files, which requires custom parsing to find, for instance, the begin time of the data in a file (see section 2.2.4.4). This parsing can be described with standardized template descriptions, such as, https://github.com/hapi-server/uri-templates/wiki/Specification, which is more useful when datasets and filenames follow common recommendations https://spdf.gsfc.nasa.gov/guidelines/filenaming_recommendations.html.

Like the NASA CDAWeb, the ESA heliophysics archives offer the possibility to select multiple datasets by time range to either download or visualize them. Advanced search tools are also available such as the Cluster data mining tool, allowing science based searches over an entire archive (https://caa.esac.esa.int/data-mining/data-mining/). However, search capabilities could be improved to make use of the full potential of the detailed metadata included in the files. For instance, detailed metadata could be used to access and retrieve data related to direct current (DC) electric field (already stored in the Cluster Science Archive (CSA) metadata). Thanks to these metadata, users could retrieve not only measurements from classic DC double-probe sensors, but also high-level DC electric field data products derived from particle experiments, which are already available (e.g., see Paschmann et al., 1997; Torbert et al., 2016). More generally, some archives or overarching interfaces could offer more scientifically oriented search capabilities (e.g., the Heliophysics Data Portal (HDP); https://heliophysicsdata.gsfc.nasa.gov). It is clear that data repositories need to have accurate and adequate metadata to enable effective finding, accessing, and using (or reusing) the data, i.e., satisfying "F," "A," and "R" in FAIR. It would be very helpful if all data repositories could adopt the same standard metadata model so that they can support a standardized interface for data access by different access protocols (see section 2.2.4.4), and thereby become interoperable.

2.2.3. Modeling centers

Modeling centers play a special role in supporting heliophysics and space weather research. Typically representing significant strategic and programmatic investments by governmental agencies and large research organizations to support research, modeling centers provide the computational resources required to construct and execute simulation runs of complex or composite models that span different heliophysical domains. Modeling centers can be the proving grounds for validating systems-science and

Table 2
Examples of infrastructure-class and mission-specific (below thick black line) data repositories and services.

Data repository or service	Host Location	Heliophysics Data Type	URL	Note
The NASA Solar Data Analysis Center (SDAC)	NASA Goddard Space Flight Center, USA	Solar data	https://umbra.nascom.nasa.gov/	Part of NASA HDRL https://hdrl.gsfc.nasa.gov/
Space Physics Data Facility (SPDF)	NASA Goddard Space Flight Center, USA	Non-solar heliophysics data	https://spdf.gsfc.nasa.gov	Part of NASA HDRL https://hdrl.gsfc.nasa.gov/
The European Space Astronomy Centre (ESAC) Science Data Centre (ESDC)	European Space Agency (ESA)	Solar and space physics data	https://www.cosmos.esa.int/web/esdc Solar orbiter https://soar.esac.esa.int Cluster and Double star https://csa.esac.esa.int SOHO https://ssa.esac.esa.int/ Proba-2 http://p2sa.esac.esa.int/ Ulysses http://ufa.esac.esa.int/ufa/	Satellite-focused data service
The Centre des Données de la Physique des Plasmas (CDPP)	Centre National d'Études Spatiales (CNES)	All heliophysics data generated by CNES heliophysics missions and CNES-financed experiments	http://www.cdpp.eu	
The Data ARchives and Transmission System (DARTS)	Japan Aerospace Exploration Agency (JAXA)	Multi-disciplinary space science data archive	https://darts.isas.jaxa.jp/	
ESA space weather service network	ESA	Ground-based and space-based measurements, together with near Earth solar wind forecasts based on real time simulations	https://swe.ssa.esa.int/current-space-weather	
ESA Earth Observation Virtual environments for Earth Scientists (VirES)	ESA	Data, indices,value added products and services from ESA ionospheric missions (inc. SWARM) and other Earth's magnetic field related missions	https://vires.services/	
Open Madrigal Initiative	MIT Haystack Observatory / NSF	Geospace data	http://cedar.openmadrigal.org/openmadrigal	International collaboration
SuperMAG	Johns Hopkins University Applied Physics Lab, USA	Ground-based magnetometers network	https://supermag.jhuapl.edu/	International collaboration
International Real-time Magnetic Observatory Network (INTERMAGNET)	Various	Ground-based magnetometer data collections	http://intermagnet.org/	International collaboration led by an executive council
SuperDARN	Johns Hopkins University Applied Physics Lab, USA	Ionospheric coherent radar network	http://superdarn.jhuapl.edu/	International collaboration
BASS2000 Solar Survey Archive	Paris Observatory, France	Ground-based solar measurements	https://bass2000.obspm.fr/	
Inter-university Upper Atmosphere Global Observatory Network (IUGONET)	Japan	Ground-based and space-based solar and non-solar measurements	http://search.iugonet.org/list.jsp	Japanese collaborative project
Magnetospheric Multiscale (MMS) science data center	LASP at CU Boulder, USA	Measurements performed by the MMS satellite constellation	https://lasp.colorado.edu/mms/sdc/public/search/	Satellite-focused data service
THEMIS mission data service	UC Berkeley, USA	Measurements performed by the THEMIS mission	http://themis.ssl.berkeley.edu/overview_data.shtml	Satellite-focused data service
Proba-2 science center	Brussels, Belgium	Processed data for ESA's PROBA2 spacecraft	https://proba2.sidc.be/	Satellite-focused data service
Exploration of energization and Radiation in Geospace (ERG) science center	Nagoya, Japan	Ground-network observations, simulation/integrated analysis, and Arase data	https://ergsc.isee.nagoya-u.ac.jp/	Satellite-focused data service

space weather models. The Virtual Space Weather Modelling Centre (VSWMC; https://esa-vswmc.eu/, https://swe.ssa.esa.int/gen_mod) implemented within the ESA Space Weather Service Network (https://swe.ssa.esa.int/) and the Center for Heliospheric Science (CHS, https://chs.isee.nagoya-u.ac.jp/en/about/) of the Nagoya University are two examples of recently established modeling centers with significant simulation components.

Another long-standing modeling center example is the CCMC located at the NASA Goddard Space Flight Center (GSFC). The CCMC has been serving the international heliophysics science community since 2000. The main goals identified in the CCMC concept of operations are: 1) to facilitate research and model development to advance understanding; 2) to support the transition of research models to space weather operations. Fast-forwarding to the present day, the CCMC currently hosts the largest expanding collection of space science and space weather models developed by the heliophysics science community. As of 2023, the CCMC offers more than 80 models and model combinations for public use. Anyone can request a model run via CCMC's flagship Runs-On-Request service (ROR; https://ccmc.gsfc.nasa.gov/requests/requests.php). In addition, there is a subset of models that CCMC executes continuously using near real-time data as input feeding CCMC's integrated Space Weather Analysis system (iSWA; https://ccmc.gsfc.nasa.gov/iswa/). It is clear that the collection of modeling artifacts: models, model runs, and their output, represent a treasure trove of information that would benefit research if all of them can be made widely available and accessible to all researchers, just like data resources are available from data repositories.

Given the CCMC's continuously growing archive of simulation models, model runs, and products, it will be extremely useful, more efficient, and economical for researchers to simply search, obtain and understand existing modeling results relevant to their studies rather than to rerun the models and re-create earlier modeling studies. With that thought in mind, the CCMC has been researching on how to make a CCMC 'knowledgebase', which includes all CCMC hosted models, model runs, output, derived products, and services, easily searchable and accessible by any user. One essential building block to achieve such a goal is a common metadata standard that all CCMC services/tools would understand. One can think of such metadata as a common language that all CCMC developed services/tools can use to communicate, understand, and exchange information with. Currently, however, there is a limited set of standardized metadata describing the various models offered by the CCMC. With that realization, CCMC is actively working on adding metadata to all information stored in their 'knowledgebase'. The metadata standard that CCMC plans to follow is the SPASE metadata standard described in section 3 of this paper, with current progress detailed in 3.4.

A similar exercise is being pursued in Europe for all run-on-request simulations from a few tens of heliophysics simulations including the EUropean Heliospheric FORecasting Information Asset or EUFHORIA (Poedts et al., 2020) through the ESA VSWMC. More details about the simulation services of CCMC and the VSMC are presented in Masson et al. (2023).

2.2.4. Standards for data formats and services

General adoption of standards by systems is a key to enabling interoperability between those systems. Interoperability is the "I" in the FAIR principles, and so is a critical part of infrastructure capability. Information transfer between two components of the information architecture in Fig. 1 is hampered if the information from one subsystem (e.g., data stored in a data file) cannot be opened, clearly and completely mapped or translated into a downstream structure, operated, interpreted, or understood by other parts of the systems downstream, such as an analysis tool or a user. The standards that support information flow throughout the heliophysics data environment must therefore also be parts of the supporting infrastructure. If different parts of the environment employ different standards, then proper translation or mapping between the two standards will be required. Unfettered flow of data and information is thus enabled by two complementary measures: (1) the use of the same metadata standards as a common language for information flow, and (2) the development and implementation of versatile mapping and translation tools as needed. In the subsections below, we outline several key aspects of the information environment components in which the use of the same standards as a common language, i.e., a lingua franca, is particularly beneficial.

2.2.4.1. Data formats. A data format specifies how information or data content is digitized and organized (and stored) in a data file. It is easy to imagine even for a simple, time-series dataset written into an ASCII (American Standard Code for Information Interchange) table or a spreadsheet, correct use of the data is possible only if proper metadata descriptions of the data file (type and format) are available to the user before it can be independently used. Community users are often confronted with (1) not knowing the specific file and data formats in advance of acquiring the data, (2) having to deal with resources in different formats that they might not be familiar with, and (3) not knowing what tools are available and suitable for dealing with the data once they have acquired it. The lack of knowledge about file and data formats and related tools invariably hinders the effective use of the data resources by the wider Heliophysics and Space Weather community.

Due to their special capability of handling multi-dimensional data records and embedding of structured metadata, the so-called "self-describing and self-documenting" data formats, such as the Common Data Format (CDF, https://cdf.gsfc.nasa.gov/), the Flexible

Image Transport System (FITS, https://fits.gsfc.nasa.gov/), and the Network Common Data Form (netCDF, https://www.unidata.ucar.edu/software/netcdf/), have been the mainstay of data formats used for storing heliophysics resources. Since the availability of metadata is guaranteed for these file formats, they are often preferred to the more traditional formats (e.g., ASCII) that require separate metadata documentations, which could be lost easily without having the embedded metadata. In addition, the use of standard data formats with standard metadata specification agreed upon at the international level [such as the ISTP Guidelines (https://github.com/IHDE-Alliance/ISTP_metadata/) for CDF and netCDF files or the World Coordinate System (WCS) for FITS (https://fits.gsfc.nasa.gov/fits_wcs.html)] would ensure interoperability of the resources stored in those file formats. The drawback of these binary standard formats, however, is that the embedded metadata is not fully accessible without opening the data files, making the data less convenient to be searched and found (i.e., less supportive of the "F" in FAIR).

An intermediate solution was found in the ASCII-based Cluster Exchange Format (CEF) with associated structured metadata and a detailed metadata dictionary that was developed by ESA for the Cluster and the Double Star missions. It enables storing of multi-dimensional time series datasets and images. Structured metadata is embedded at the top of any file and read by various data analysis software. CEF metadata headers are also made available as separate files. A data converter from CEF to CDF ISTP was developed to ensure interoperability. Unlike CDF, however, the CEF is not a widely adopted standard format for heliophysics data.

While it is understandable that different standard data formats may be used to store different types of data particularly in different heliophysics disciplines (e.g., FITS for solar images and CDF for in situ time series data), we should still seek to have a uniform (standard) way to describe these data despite their format differences. The need for different data tools to deal with different data formats has the potential of setting up barriers to information flow and data exchange between disciplines, unless the information about and access to those tools are also available to the user somehow. On the other hand, general adoption of the same standards would facilitate the "I" and "R" in the FAIR principles. As we will see in the sections below, having metadata descriptions of resources electronically accessible and independently of the data, but still permanently associated with the data (resources) (see section 2.2.1), would greatly facilitate the discovery of the resources by a middleware (Fig. 1).

2.2.4.2. Metadata. Metadata is often described, in short, as "the data about data." That is to say that metadata are for describing data, giving information about the data (e.g. its identification and location, coordinate system, variable names, units, time cadence, etc.) and not necessarily revealing the information (the data records) it describes. For proper analysis of the data, however, the associated metadata must uniquely identify and describe the data completely in order to enable independent use of the data. Furthermore, the identifying attributes of the data can be used to search for and locate the specific data elements from among all the data stored in a data repository, so that the relevant data can be extracted and delivered to the users.

Once delivered, a user would need to open the data files and make use of the data in analysis tasks. But that cannot be done unless the user knows how (1) to open the data files, (2) to read the content of the data files correctly, (3) to properly manipulate the data, and (4) to understand the contents of the data. Steps (1)-(3) relates directly to the data formats discussed in the previous subsection. When the data are stored in *standard data formats*, tools developed by the community are generally available and accessible for handling the data files and their contents.

Providing clear, unambiguous meaning of the contents of the data is of utmost importance for supporting point (4) above and enabling independent use of the data. Given the diversity of measurement platforms, instruments, measurement types, and measured quantities in heliophysics, there is a great deal of complexity in heliophysics data. Similar measurement quantities appearing in different datasets may have different coordinate systems, units or measurement cadence. Without proper definitions and explanations, it would be hard for datasets to be utilized by users from different disciplines. This is akin to different people trying to communicate in different languages. The most effective way to enable independent use of the data by diverse users is the adoption of a metadata standard that can serve as a lingua franca between different disciplines, data services and systems.

We describe in the next subsection some general guidelines on how data contents at the measurement parameter level should be described in order to enable independent usability of the data.

2.2.4.3. Metadata for parameter descriptions. While global attributes of a digital resource are needed for identifying and locating the resource, they do not contain specific information about the contents of the resource. And, while the data file format specifies how resource contents are organized, parameter descriptions are needed for explaining the contents of the resource. Complete and unambiguous parameter descriptions are the key to understandability, thus making the resource **independently usable** at the end of path (4) in Fig. 1 and ensuring the "I" and "R" in the FAIR principles.

Metadata provides detailed descriptions of data formats, datasets, parameters, time conventions, and dataset and file naming conventions enabling effective data analysis and browsing with generic easy-to-use software and web services. Restricting metadata descriptions to standard, and thus uniform, representations would limit the number of equivalent possibilities which software must deal with

and would thus foster interoperability. Conventions help standardize ways to name things, represent relationships, locate data in space and time, and abstract the general data models to represent the data semantics. This enables the development of reusable applications with powerful data extraction, gridding, analysis, visualization, and processing capabilities. These standards embody the data provider's experience and capture the meaning in the data, making the data semantics accessible to humans as well as software tools. Higher-level abstractions such as coordinate systems and standard names for physical quantities further facilitate comparing different datasets and distinguishing between variables.

The ISTP Metadata Guidelines mentioned above (section 2.2.4.1) have been in use by the traditional space physics community over the last few decades mostly for time series data. It provides an internally and logically complete set of metadata to enable a dataset to be correctly and independently usable, especially by generic (non-dataset specific) automated processing, analysis, and display software. The guidelines define self-documenting (internal) metadata for data stored in CDF and netCDF files and include general file naming conventions. Data are time-ordered and time-identified, with standardized time formats. The ISTP Metadata Guidelines define a set of required and optional global (dataset-level) and variable (parameter-level) attributes. Variable attributes can point to other variables by name and carry arguments, and thus may convey information about relationships among variables. The ISTP guidelines may thus provide a good candidate for paving the way toward standardizing parameter descriptions for time series data.

2.2.4.4. Data access protocols. Data access protocols are used to access, retrieve, and deliver a specific element of digital content to the user. Before the execution of an access protocol, however, the specific element of digital content defined by, e.g., a dataset name, measurement parameters and time range, must first be identified and located, requiring the use of metadata (see paths (2) and (3), or (3a), and (4) in Fig. 1). This means that metadata is needed to first define the resource object for an access protocol to access, satisfying the "A" in the FAIR principles. If the same access protocol is used generally by different data sources (data providers or repositories in Fig. 1), then the protocol would become a de facto standard and the data sources would meet the interoperability requirement of the FAIR principles.

File-level delivery of data resources is traditionally accomplished by the File Transfer Protocol (FTP) or HyperText Transport Protocol (HTTP), or their secured versions. To promote and serve an interval of data requires the determination of the minimum set of files that should be collected and retrieved. As data volume and complexity increase and storage facility multiplies, more robust discovery (search and access) tools that make effective use of the data products' metadata will be needed. As depicted by paths (2) and (3), or more directly by path (3a) in Fig. 1, data access protocols need to dovetail with data search mechanisms so that the result from a data query can be fed into the access protocol of the appropriate data service or repository and retrieve the data for delivery to the user or analysis tool, as represented by path (4) in Fig. 1. Two protocols are used at the NASA SPDF to access time series data: HAPI (https://hapi-server.org/servers/, Weigel et al. 2021a) and the Coordinated Data Analysis System (CDAS) Web services. They and a few others currently in use for accessing heliophysics resources are described briefly below and summarized in Table 3.

The HAPI protocol is an API with endpoints that allow access to time series data values at the parameter level within one or more data collections. Since HAPI is focused on data access, the required HAPI-specific metadata is not intended for complex search and discovery. However, the metadata schema can provide information as to where more descriptive details for any dataset could be found. The HAPI API is based on REpresentational State Transfer (REST) principles (https://en.wikipedia.org/wiki/Representational_state_transfer), which emphasize that Uniform Resource Locators (URLs) are stable endpoints through which clients can request specific data elements. Because it is based on well-established HTTP request and response rules, a wide range of HTTP clients can be used to interact with HAPI servers. HAPI is also a COSPAR recommended access protocol standard (see COSPAR, 2021) for serving time series data.

CDAS is also a RESTful web service for querying data and metadata components from data sets in the NASA SPDF CDAWeb system. Queries can include requests for information regarding instruments, observatories, and the data inventory. It can support simultaneous multi-mission, multi-instrument selection and comparison of science data. The metadata associated with the digital resources is ISTP/SPDF compliant. There is also a Python library (https://pypi.org/project/cdasws/) which provides a simple python interface to the CDAWeb data and services.

Solar data collections at the SDAC are primarily delivered through the Virtual Solar Observatory (VSO). VSO provides access to distributed solar data collections stored around the world, all described with the VSO data model. Two popular tools in the solar community have been developed to access the VSO API: the IDL Solar SoftWare (SSW) package (SolarSoft) and the SunPy (Mumford, et al., 2021) unified Finding and Downloading object of sunpy.net called Fido (https://docs.sunpy.org/en/stable/generated/gallery/acquiring_data/searching_vso.html).

At ESA ESDC, apart from the Ulysses final archive (UFA; see Table 2) which uses a FTP-like method for file access, all heliophysics archives are accessible through the International Virtual Observatory Alliance (IVOA; https://ivoa.net/) Table Access Protocol (TAP, see Table 3). This data access protocol enables access to both data and metadata for time series and remote sensing data. The

Table 3
Data access protocols often used in heliophysics.

Data Access Protocol	Data type	Access Method(s)	URL	Linked Data Sources
Heliophysics API (HAPI)	time series data	RESTful API, HTTP	http://hapi-server.org/servers/	SPDF, SDAC, CCMC, and more
Coordinated Data Analysis System (CDAS) Web services	time series data	RESTful API	https://cdaweb.gsfc.nasa.gov/WebServices/	SPDF
Virtual Solar Observatory (VSO)	solar data	API	https://nso.edu/data/vso/	SDAC
International Virtual Observatory Alliance (IVOA)	time series and remote sensing data	Table Access Protocol (TAP)	https://www.ivoa.net/documents/TAP/, https://solarnet.oma.be/	ESA ESDC heliophysics archives, Solar Virtual Observatory (SOLAR VO), SOLARNET
Europlanet TAP (EPN-TAP)	observations, simulations, or experimental data	Extended version of TAP for planetary data	https://www.ivoa.net/documents/EPNTAP/, http://www.europlanet-vespa.eu/standards.shtml, https://vespa.obspm.fr/planetary/data/	Virtual European Solar and Planetary Access (VESPA)

Cluster Science Archive (CSA; see Table 2) and Solar Orbiter ARchive (SOAR; see Table 2) in particular contain more than two thousand datasets with detailed metadata, some even described in SPASE (see section 3.0), all accessible through TAP. The first HAPI server access interoperable with a TAP server (starting with CSA) is now implemented. Meanwhile, on-going collaboration between ESDC and SDAC has made the VSO interoperable with TAP servers, starting with Proba-2 data (now available on the VSO) and soon the SOAR TAP server. TAP-services are also employed by the Solar Virtual Observatory (SOLAR VO, see Table 3), which was built during the H2020 SOLARNET project.

Europlanet (EPN) TAP, also known as EPN-TAP (Table 3), is the TAP protocol with the inclusion of the EPNcore metadata dictionary (Erard et al, 2014). It describes tables with a common set of required metadata (parameters) in standard units which can be used to query all EPN-TAP services. A unique query can then be sent and be answered by multiple data services. EPNcore (https://www.ivoa.net/documents/EPNTAP/) defines the core groups of metadata (components) that are necessary to perform data discovery in science fields related to the Solar System and related fields. It includes keywords to describe data products coverage (temporal, spectral, spatial, illumination conditions), origin (instrument, facility), content (target, physical parameters), access, references, etc. These keywords are intended either as search parameters or as descriptive information. All keywords can be searched by value with the TAP. Mapping the SPASE dictionary with the EPNcore metadata to enable cross-resource searches is not straightforward for two reasons: (a) the SPASE metadata are dedicated to space physics (hence with implicit knowledge), whereas EPNcore includes all solar system science topics; (b) the SPASE metadata has originally been defined as a registry of resources, whereas EPNcore is dedicated to data discovery. A first attempt to map the SPASE metadata with the astronomy metadata has been conducted by Cecconi et al. (2014). The resulting mapping has been included in version 1.3 of the Unified Content Descriptors (UCDs) (Preite Martinez et al, 2018). The full mapping between the two metadata schemas is currently being developed.

EPN-TAP uses the notion of "granule" (inherited from the SPASE standard) to refer to the data service granularity, which is the smallest data unit that can be provided by a service. A "granule" can correspond to a data file, a set of scalar values, a call to a web service, a query to data service in a different protocol, etc. Each granule is described in an associated table, one granule per row, which is accessed and used by the EPN-TAP service. An EPN-TAP service can support observations, simulations, or experimental data.

2.2.5. Data visualization and analysis tools

Data visualization tools are used for browsing and screening the data for interesting events or for exploring and analyzing quantitatively the data content in support of a given study and are important and indispensable functionality for supporting science analysis. Proper graphical display of data requires precise knowledge of the parameters contained in a data product, including for example the number and identities of the parameters, and how they are organized in a data file (i.e., data format as described in section 2.2.4.1). Attributes of each parameter, such as dimensions and ranges, coordinate systems, sampling rates in time and/or space (cadence and resolution), units of measurements, normalizations (if any), and caveats, etc., must all be properly specified before they can be accurately represented and displayed. Our ability to display and analyze the data correctly is therefore intimately dependent on the availability of the metadata describing the parameters contained in the data files, as discussed in sections

2.2.4.1-2.2.4.3. Data visualization examples are provided in sections 4.2 and 4.3.

3. SPASE metadata model

Metadata - the data about the data - is a critical supporting infrastructure element (section 2.2.4.2) enabling the several important functionalities of the heliophysics data environment (Table 1). Without metadata, it would be impossible to find or discover, access and understand digital resources. However, the large variety of resource types and their current metadata descriptions mentioned thus far demand a uniform metadata standard to enable a truly simplistic and capable digital resource environment. Of the variety of metadata formats currently in use to describe those resources, the ISTP Metadata guidelines are so far the most proliferated, in particular for parameter-level descriptions of data stored in CDF and netCDF. However, the ISTP Metadata guidelines are also limited (sections 2.2.2, 2.2.4.1, 2.2.4.3) and can only support a portion of the digital resource environment. Heliophysics and space weather research require a broader, more comprehensive metadata standard that can be applied to all the digital resources in our current environment. We advocate that the SPASE metadata model is the most developed candidate for that metadata standard in heliophysics.

The Space Physics Archive Search and Extract (SPASE) metadata model has been designed, developed, and maintained over the past 20 + years largely by members of the international space physics (heliophysics) community, the SPASE Group (https://spase-group.org/about.html), specifically for recording the descriptive information, i.e., the metadata, of space physics or heliophysics data products (Roberts et al., 2018). Even though the SPASE information model and the SPASE Group got their start in the international community at the grassroots level, their development, maintenance, and operations have primarily been sponsored by NASA. The SPASE Group, however, is still an international heliophysics community group and remains a grassroots effort to ensure community involvement and as such is open to all interested parties (https://spase-group.org/connect.html). In addition, NASA has established a SPASE metadata working team (SMWT, https://hdrl.gsfc.nasa.gov/smwt_home/smwt_index.html) to assist the community in implementing the SPASE model and to produce and maintain SPASE metadata. Consequently, NASA currently maintains the largest, openly-accessible SPASE metadata registry on Github (https://github.com/hpde) with the SPASE metadata landing pages also posted at https://hpde.io/.

The SPASE metadata specifications are permanently referenceable with DOI urls: https://doi.org/10.48322/E72C-5Y75 for the base information model and https://doi.org/10.48322/TXCA-X050 for the simulation extensions. The DOI urls lead to the landing pages containing the SPASE descriptions of the model specifications. Upon accessing a referenced DOI landing page, the actual citation of the resource with the DOI URL is given near the top of the page with an access timestamp appended. By this citation scheme, the DOI URL will always point to the latest version of the data model. The timestamp, when compared against the resource revision history in the SPASE description, will then help identify the version of the resource referred to or used. We note here that all heliophysics digital resources (dataset, model, modeled data, etc.) can be cited similarly by using DOI URLs and SPASE description landing pages.

Due to its specific applicability to heliophysics data, SPASE has been adopted by the COSPAR Panel on Space Weather in 2018 as a recommended metadata standard for describing space physics and space weather resources (COSPAR Panel on Space Weather, 2021). The subsections below give an overview of the SPASE metadata schema and a description of how heliophysics digital resources can be described and referenced, enabling their archival, searchability, accessibility, understandability, and independent reusability.

3.1. SPASE model ontology

The SPASE information model is basically divided into four separate resource description domains: Data, Origination, Infrastructure, and Simulation Extensions, so SPASE can be used to describe any type of digital resources being used in the heliophysics data environment (section 2.1). Fig. 2 shows a schematic of the four metadata domains and their associative relationships, i.e., the SPASE ontology (since version 2.4.0). We should note here that the "infrastructure" depicted in Fig. 1 does not show services and cyberinfrastructure that are also implied in the broader SPASE ontology shown in Fig. 2 (SPASE Group, 2021). The Data domain (blue labels) contains the digital resource (e.g., document, catalog, numerical data, display data, annotation, or granule) to be described in SPASE. Origination (green labels) is the domain of all the entities from which a resource originates or is created. It is also the domain through which resource provenance can be traced. The Infrastructure domain (pink labels) consists of all the systems, services, software (tools), and facilities needed to support the operations of the heliophysics data environment described in section 2.1. Finally, the Simulation Extensions domain (yellow labels) is where simulation models and results are described. The meanings of the SPASE terms mentioned here and elsewhere in the paper can be found by consulting the SPASE metadata dictionary (https://spase-group.org/data/model/search/index.html).

This dictionary has been developed over the years by space scientists specifically for describing heliophysics resources and is thus quite extensive, yet it is by no means complete. New quantities and terms can be defined and incorporated into the SPASE metadata dictionary on an as-needed basis. Any requirement for a new SPASE term or a possible change in the SPASE schema should be dis-

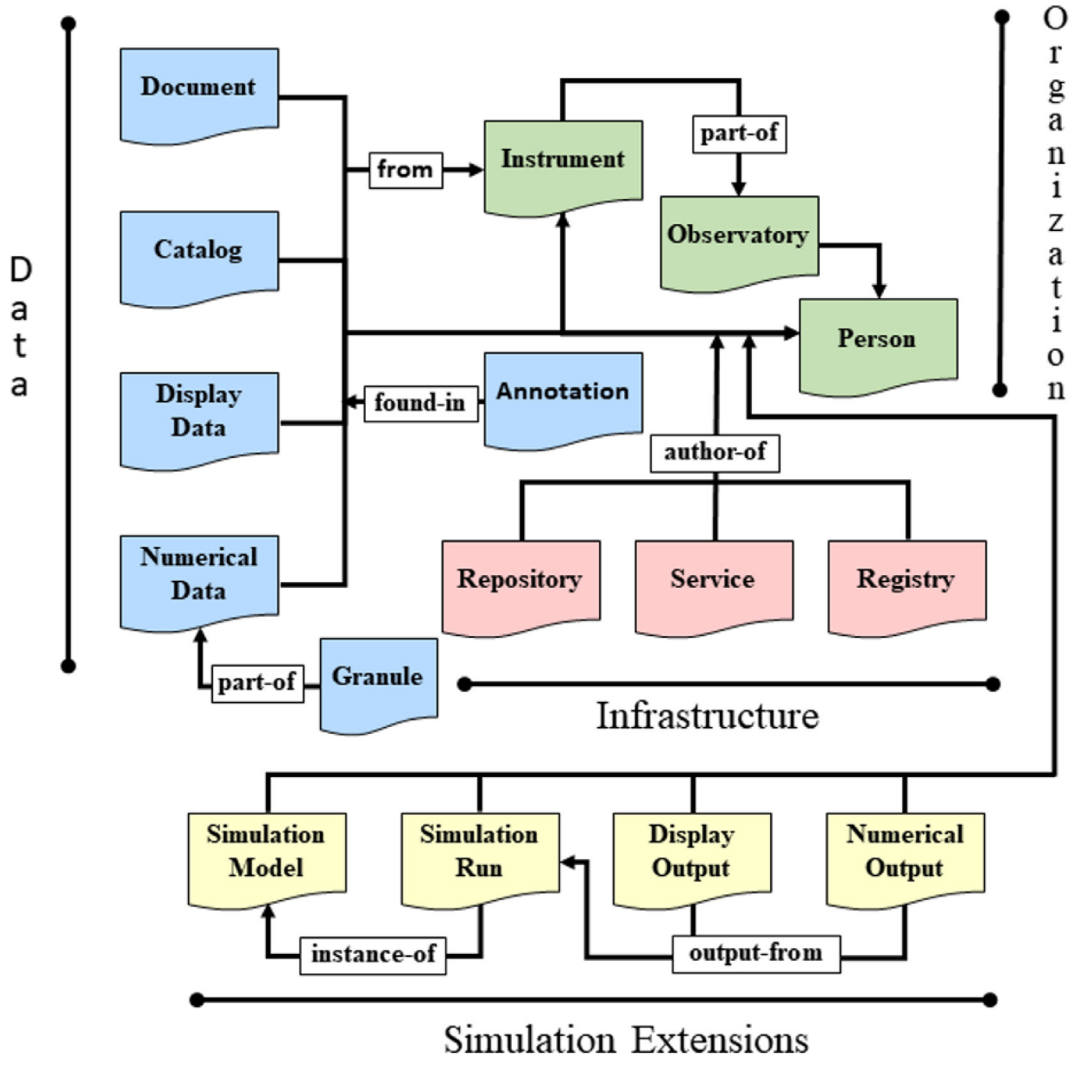

Fig. 2. A schematic of the SPASE ontology (since version 2.4.0), showing the interrelationships between the SPASE descriptive domains: data or resources (blue), origination (green), infrastructure (pink), and simulation extensions (yellow). Arrows in the schematic point in the direction of association. For example, the top of the figure shows that Instrument is "part of" Observatory, and the association is not reversible. (Note added in galley proof: Simulation extension has been incorporated into a recent release of the SPASE information model (version 2.6.0) with which description of empirical models is also enabled. An updated SPASE ontology figure can be found in the PDF version of the SPASE model documentation posted at https://spase-group.org/data/model/index.html. This shows that the SPASE information model is a living model that can be improved and developed upon as needed.)

cussed with the SPASE Group by emailing spase@groups.io for consideration. General queries about implementation of the SPASE information model can be sent to spase-support@groups.io. Since the SPASE Group meets regularly on a bi-weekly basis, simple adoption of a new term into the SPASE dictionary can occur over a two-week interval, from proposal to discussions and to final voting for adoption. The established procedure for updating, developing and maintaining the SPASE metadata model has over the years produced both major and minor releases of the dictionary and schema under strict version control. The SPASE model revision history, releases and specifications can be found on the SPASE model page at https://spase-group.org/data/model/index.html. Versioning of datasets, as provided by their producers, are similarly tracked by their DOI references and revision history in their SPASE descriptions.

3.2. Identifying datasets and tracking their origins

Minimizing barriers to the flow of data is the key to data sharing and enabling effective use of the resources by the international community. In an environment where resources are freely exchanged, however, it is easy to lose track of the origin or the evolutionary track of a resource. It is thus important to identify and track the origins of and changes to resources while sharing the resources in accordance with the FAIR principles (Wilkinson et al., 2016).

We describe here how the origin of a resource is built into the SPASE metadata model and used for tracking. We then describe in section 3.5 below how digital resources can be cited and referenced by using Digital Object Identifiers (DOIs).

It is important to note that the origin of an object, its source or history of ownership, is referred to as the provenance of the object in the English language [see for examples, the Merriam-Webster (https://www.merriam-webster.com/dictionary/provenance), the Cambridge (https://dictionary.cambridge.org/us/dictionary/english/provenance), and the Oxford Learner's (https://www.oxfordlearnersdictionaries.com/us/definition/english/provenance?q = provenance) dictionaries]. In informatics, however, the concept of provenance has been broadened or modified significantly; its definition may also depend on the context and the community involved. An often quoted or adopted "definition" has been promoted by the World Wide Web Consortium (W3C; see https://www.w3.org/TR/prov-overview/), which states that "provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness." This definition has also been adopted by the International Virtual Observatory Alliance (IVOA; see https://www.ivoa.net/documents/ProvenanceDM/).

Fig. 2 shows that all the information required for provenance in accordance with the *W3C* is spread throughout the SPASE metadata description of a resource. It would thus be hard to track provenance in the W3C sense using a single, convenient handle or tag. For reasons that will become clear, we refer to the *origin* of a resource in this paper simply to the "original ownership" or the *root source* of the resource. A digital resource described by SPASE is then uniquely identified by a SPASE Resource ID, typically shortened in CamelCase and italicized as *ResourceID*, which has a uniform resource identifier (URI) of the form (see Guidelines for Resource ID Formation, 2022):

spase://NamingAuthority/ResourceType/Project/Observatory/InstrumentType/Cadence.

where "*spase*" identifies SPASE as the *Namespace* (https://en.wikipedia.org/wiki/Namespace) or the declarative scheme being used to construct the *URI* for identifying an object, *i.e.*, a digital resource. The *NamingAuthority* then identifies the *root* or *original* ownership of the resource and is thus the top-level identifier of all the SPASE metadata resources that belong to the same Naming Authority. In practice, original ownership cannot be assigned to a person, but to an organization, so the most logical origination entity would be the organization or agency that commissioned the creation of the resource in the first place. Therefore, a *NamingAuthority* is most suitably assigned to the commissioning agency or group that has the overarching authority over the digital resource. By agreement or management arrangements, *NamingAuthority* can also be delegated.

While a Naming Authority is typically the funding agency that sponsored the creation of the resource, many projects, observatories, or even instruments, particularly in ground-based facilities, are sponsored or managed by multiple agencies. It is sometimes difficult to determine which agency holds the ultimate oversight responsibility. In that case, the SPASE Group has recommended a set of rules for establishing a *NamingAuthority* for both space-based and ground-based resources (see https://spase-group.org/services/naming-authority.html, where a number of currently defined Naming Authorities are also listed).

The rest of the *ResourceID* URI in the example for an observational dataset above then consists of item identifications in the Data domain (*ResourceTypes),* the Origination domain (*Project, Observatory,* and *InstrumentType*), and the specification of the most distinctive or representative characteristic of the resource (e.g., time cadence) to make up a *unique path* for identifying the described resource (see also section 4.1.3.1). *ResourceIDs* for other types of resources (see Fig. 2) can be constructed similarly to the above example. Since all digital resources are uniquely identified by their corresponding ResourceIDs, the SPASE URI structure thus provides a logical way to organize all the SPASE resources on the SPASE registry (see https://hpde.io for the SPASE registry landing pages).

We should note here that there is not yet a globally recognized general SPASE metadata registry. While the SPASE Group maintains the largest SPASE metadata registry (https://github.com/hpde/), individual data services such as the Automated Multi-Dataset Analysis tools (AMDA) of the CDPP, ESAC science data centre (ESDC), and EuroPlanet, have also set up their own metadata registries based on their adopted metadata models: SPASE, TAP, and EPN-TAP, respectively. Discussions in international forums on collaborations and coordination on the development and use of standards for data and tools, as described later in section 5.5, will help promote sharing of metadata and ensure interoperability of data services worldwide.

3.3. Digital resource description

As pointed out in section 2.2.4.2 (metadata), a user must be able to do four things in order to use the resource effectively and independently after retrieving a digital resource. As noted also in section 2.2.4.3 (parameter descriptions), independent usability of a dataset depends on the clarity, correctness, completeness, and self-sufficiency of the metadata describing the dataset. Fig. 3 shows an example of the hierarchy of SPASE fields and labels for identifying, locating, and understanding each of the parameters of a numerical data resource. The *NumericalData* resource schema has two important metadata containers that need careful attention when generating data product descriptions. The first is the *ResourceHeader*, which includes a *ResourceName* text box and a Description text box among other things (see

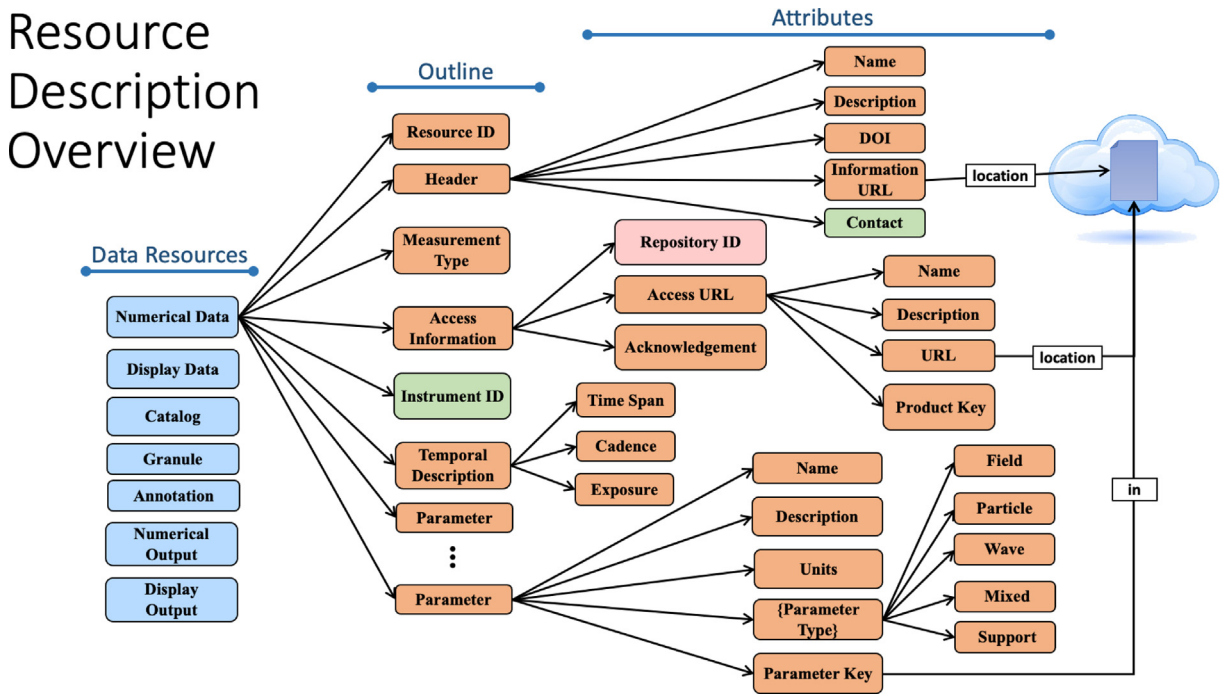

Fig. 3. An example of numerical data description based on the *SPASE* schema. The blue, green, and pink labels correspond respectively to the elements in the Data, Origination, and Infrastructure description domains shown in Fig. 1, whereas the orange labels correspond to the various descriptive fields used to identify and locate the numerical data resource and provide the information and explanations needed to understand every parameter contained in the resource.

Fig. 3). The *ResourceName* should list a unique dataset "title" along with the data processing level and time resolution. The Description text field allows one to provide a much more complete account of the dataset content. One can utilize other parts of the *ResourceHeader* schema to shorten the overall Description and reduce the work required to write SPASE descriptions. For instance, one can provide the DOI and *PublicationInfo* for an already published paper or other resource that describes the data (see section 3.5). *InformationURL* metadata can also be used to link to web sites that describe the data product, though it is important to remember that URLs are not guaranteed to have the persistence of DOIs. It is also important to provide a complete list of the people involved in all aspects of the data product including generation, access, and archival by using the Contact portion of the header schema.

Note that *NumericalData* allows one to list *InstrumentID* metadata records that refer to instruments also described in SPASE, which themselves allow contacts for the instrument to be specified. So, the *NumericalData* description needs to identify only the individuals associated directly with the generation of the data product, while other individuals can be associated by linking the relevant records. Finally, we wish to stress the importance of providing accurate and complete acknowledgements to persons involved in the project. Also, if a link is available that specifies the exact text to give credit for any aspect related to data product legacy, then it is good practice to provide that information via *InformationURLs*.

The second important container is *AccessInformation* (above the green 'Instrument ID' box in Fig. 3) that allows one to provide different methods for accessing data products via population of *AccessURL* Name, Description, and URL metadata with one or more *AccessURL* containers for each data repository. Multiple *AccessInformation* containers would be used if the same data are hosted by multiple institutions with potentially different access protocols (web portals, HAPI servers, or other web service). Finally, issues related to data set storage and archival can only be handled by those generating and providing data. If a data product is accessible and stored in a consistent fashion, then high quality SPASE descriptions can be written to provide all the information required to document the data for archival purposes.

Descriptions of other resource types (blue labels in Figs. 2 and 3) can be similarly represented. A more complete view of the SPASE metadata layout, including the simulation extensions, can be found in Fig. 1 of Roberts et al., (2018). To understand the meaning of each tag or term in Fig. 3 and the associated information pertaining to a dataset being described, the readers are encouraged to consult the SPASE metadata model (https://spase-group.org/data/model/index.html) (SPASE Group, 2021) and the SPASE dictionary (https://spase-group.org/data/model/search/index.html). Since not everyone is well versed in XML and the SPASE schema, the SMWT has

developed a web-based SPASE editor (https://xmleditor.spase-group.org/) for anyone to create and edit SPASE XML documents without needing to know XML and SPASE.

Under the auspices of the NASA HDRL, there is ongoing work by the SMWT to assist the community in creating and maintaining SPASE descriptions, and registering them at the SPASE registry mentioned earlier in section 3. One of the projects called ADAPT, which stands for Active Data Archive Product Tracking (Bargatze, 2018; Bargatze et al., 2022), leverages metadata embedded in data resources that are stored in self-documented data formats, such as CDF and netCDF, to automate the generation of SPASE data product descriptions. The ADAPT tool kit currently comprises a core set of IDL programs designed to create, populate, and write data product descriptions for each resource type defined by the SPASE data model. The ADAPT IDL routines are also supplemented by a set of support software built to harvest metadata from self-documented data products archived at and made available from data repositories and services (sections 2.2.2 and 2.2.4). The support software also scrapes textual information from other sources, including previously generated SPASE resource descriptions already registered and in use in the global Heliophysics data environment (Fig. 1). The Heliophysics data environment, ADAPT and its support software thus form an ecosystem in which the process of SPASE metadata generation and updating can be effectively automated by regularly harvesting metadata and re-running ADAPT over any new and updated metadata.

In the present context, ADAPT uses metadata present in ISTP compliant data files (section 2.2.4.3) to generate SPASE descriptors via mapping of CDF (operational) and netCDF (developmental) global and variable attribute content into the appropriate text field elements appearing in SPASE (see Appendix B). The ISTP-to-SPASE metadata mapping is relatively robust, but it cannot always be trusted to generate valid SPASE data product descriptions. Also, sometimes the ISTP attribute text needs standardization, correction for errors, or translation of an ISTP term into its SPASE equivalent cognate. The tables in Appendix B, described in more detail below, show the typical ISTP to SPASE metadata mappings separately for global and variable attribute metadata with some notes concerning implementation strategy and other issues. Finally, we note that the automation of SPASE document generation, as currently done for ISTP compliant CDF data products, speeds production and also minimizes error, human or otherwise, in populating metadata. Thus, the metadata generated by utilizing ADAPT software yields more precise, accurate, and complete data product descriptions. While full automation is ideal, there are some cases when harvested metadata require human editing.

Each of the tables in Appendix B has four columns that show the correspondence or mapping between the ISTP attributes and SPASE descriptions. The first column in both tables contains a list of CDF attributes/ISTP keywords with global attributes in Table B1 and variable attributes in Table B2. The second columns where the harvested CDF attribute metadata are mapped to within a SPASE *NumericalData* description. The third column, titled "Edit?", shows whether CDF metadata require hand editing in order for them to be used in populating the SPASE description while the fourth columns show how such metadata changes are handled by hand editing, text stream editing, etc. Stream edit commands, also allow one to automatically revise a file on a line-by-line basis. ADAPT mostly uses sequences of BASH sed commands for updating text (see the sed manual listing at: https://www.gnu.org/software/sed/manual/sed.html). The ADAPT sed tools are occasionally revised to stay up to date as new textual issues are encountered.

3.4. Model and simulation data description

Fig. 4 shows a more detailed ontology of the SPASE extensions for describing simulation models and their data (https://spase-group.org/data/simulation/). The readers are reminded that arrows in the schematic point in the direction of association, as in Fig. 2. Generally speaking, simulation models and model results are described separately although model run descriptions should also include the model identifiers and their associated setup for executing the model runs. These extensions were added to the SPASE model in parallel with the observational heliophysics data products beginning in version 2.2.4 in 2015.

The SPASE model simulation extensions were originally developed by members of the Integrated Medium for Planetary Exploration (IMPEx) project, which was funded by the European Union under the Seventh Framework Programme with the goal of bridging "the gap between observational databases and scientific modeling tools" (see https://impex-fp7.oeaw.ac.at). With the IMPEx work started in June 2011, the IMPEx team was able to take advantage of the basic framework of the SPASE model established several years earlier and develop the extensions for describing simulation products in a manner consistent with the SPASE metadata model (Khodachenko et al., 2011, Hess et al., 2012a,b; Génot et al., 2012, Modolo et al., 2018, Genot et al., 2021). The IMPEx simulation extensions were officially adopted by the SPASE consortium in May 2014. As a result, four new resource classes were added with the following names (Fig. 4): *SimulationModel*, *SimulationRun*, *NumericalOutput*, and *DisplayOutput*, which eponymously define their intended purpose in the context of the SPASE simulation extension schema.

All four simulation extension resource types include the *ResourceID* and *ResourceHeader* information that is present in most high level SPASE-based informational resource types. However, the simulation extensions also require the addition of attributes that are necessary for describing the models themselves, the runs executed, and the data output. For instance, the SPASE *SimulationModel*

Table B1
Comparison of global attributes for numerical data between the SPASE metadata model and ISTP Guidelines.

Metadata Mapping - CDF Global Attribute to SPASE Numerical Data			
CDF Global Attribute	**SPASE Numerical Data Mapping**	**Edit?**	**Non CDF Metadata Sources, Processing Programs, Misc. Notes**
LOGICAL_SOURCE	NumericalData/ResourceID		SMWG, Spacecraft lookup table: cdaweb_sc_list.tab, stream editing
LOGICAL_SOURCE	NumericalData/ResourceID, NumericalData/Parameter/Cadence	Yes	SMWG, cdaweb_spase_map_cadence.pro
LOGICAL_SOURCE_DESCRIPTION	NumericalData/ResourceHeader/ResourceName	Yes	Hand edits as required
Title	NumericalData/ResourceHeader/ResourceName	Yes	Hand edits as required
DESCRIPTOR	NumericalData/Parameter/Caveats	Yes	Hand edits as required
TEXT	NumericalData/Parameter/Caveats	Yes	Hand edits as required
PI_NAME	NumericalData/ResourceHeader/Contact/Name		Stream editing
PI_NAME	NumericalData/ResourceHeader/Contact/PersonID		Stream editing
ACKNOWLEDGEMENT	NumericalData/ResourceHeader/Contact/Acknowledgement	Yes	Stream editing, Hand edits as required
ACKNOWLEDGEMENT	NumericalData/AccessInformation/Acknowledgement	Yes	Stream editing, Hand edits as required
PI_NAME	NumericalData/AccessInformation/Acknowledgement	Yes	Stream editing, Hand edits as required
Not applicable	NumericalData/AccessInformation/RepositoryID		SMWG
Not applicable	NumericalData/AccessInformation/AccessURL/URL		URLs set to match the SPDF CDF directory tree structure
LINK_TITLE	NumericalData/ResourceHeader/InformationURL/Name	Yes	Hand edits as required
LINK_TEXT	NumericalData/ResourceHeader/InformationURL/Description	Yes	Hand edits as required
HTTP_LINK	NumericalData/ResourceHeader/InformationURL/URL	Yes	Hand edits as required
PI_AFFILIATION	NumericalData/ResourceHeader/InformationURL/Acknowledgement	Yes	Hand edits as required
LOGICAL_SOURCE	NumericalData/InstrumentID		SMWG
MISSION_GROUP	NumericalData/InstrumentID		SMWG
INSTRUMENT_TYPE	NumericalData/MeasurementType	Yes	Stream editing, Hand edits as required
Not applicable	NumericalData/TemporalDescription/TimeSpan/StartDate		Dates set by tracking of the CDAWeb data product CDF file content
Not applicable	NumericalData/TemporalDescription/TimeSpan/[Relative]StopDate		Dates set by tracking of the CDAWeb data product CDF file content
LOGICAL_SOURCE	NumericalData/TemporalDescription/Cadence	Yes	Hand edits as required, cdaweb_spase_map_cadence.pro
CAVEATS	NumericalData/Caveats		Stream editing
TITLE	NumericalData/Keyword		Stream editing
MISSION	NumericalData/Keyword		Stream editing
PROJECT	NumericalData/Keyword		Stream editing
DATA_VERSION	NumericalData/Keyword		Stream editing
DISCIPLINE	NumericalData/Keyword		Stream editing
DATA_TYPE	NumericalData/Keyword		Stream editing
ADID_REF	NumericalData/Keyword		Stream editing
GENERATION_DATE	NumericalData/Keyword		Stream editing
NSSDC_ID	NumericalData/Keyword		Stream editing
MODS	NumericalData/Keyword		Stream editing
SOFTWARE_VERSION	NumericalData/Keyword		Stream editing
GENERATED_BY	NumericalData/Keyword		Stream editing
RULES_OF_USE	NumericalData/Keyword		Stream editing
TEXT_SUPPLEMENT_1	NumericalData/Keyword		Stream editing
LOGICAL_FILE_ID	Used to cross check LOGICAL_SOURCE Metadata		Stream editing

Table B2
Comparison of variable (parameter) attributes for numerical data between the SPASE metadata model and ISTP Guidelines.

Metadata Mapping - CDF Variable Attribute to SPASE Parameter			
CDF Variable Attribute	**SPASE Numerical Data Parameter Mapping**	**Edit?**	**NonCDF Metadata Sources, Processing Program, etc.**
FIELDNAM	NumericalData/Parameter/Name	Yes	Often custom editing required
DEPEND_0	NumericalData/Parameter/Set		
CATDESC	NumericalData/Parameter/Set		
cdf_variable_info.name	NumericalData/Parameter/ParameterKey	No	Parameter Key populated without using Var. Attr.
VAR_NOTES	NumericalData/Parameter/Caveats	Yes	Hand edits as required
AVG_PTR_1	NumericalData/Parameter/Caveats		Stream editing
AVG_TYPE	NumericalData/Parameter/Caveats		Stream editing
VIRTUAL	NumericalData/Parameter/Caveats		Virtual Variable designation flag
FUNCT	NumericalData/Parameter/Caveats		Virtual Variable support metadata
FUNCTION	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_0	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_1	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_2	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_3	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_4	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_5	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_6	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_7	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_8	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_9	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_10	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_11	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_12	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_13	NumericalData/Parameter/Caveats		Virtual Variable support metadata
Component_14	NumericalData/Parameter/Caveats		Virtual Variable support metadata
MONOTON	NumericalData/Parameter/Caveats	Yes	Rarely needs editing
Not applicable	NumericalData/Parameter/Cadence	Yes	cdaweb_spase_map_cadence.pro via CDF Global Variable LOGICAL_SOURCE
TIME_RES	NumericalData/Parameter/Cadence		Stream editing
RESOLUTION	NumericalData/Parameter/Cadence		Stream editing
UNITS	NumericalData/Parameter/Units		Stream editing
SI_CONVERSION	NumericalData/Parameter/UnitsConversion		Set from UNITS value via stream editing
DICT_KEY	NumericalData/Parameter/CoordinateSystem/CoordinateSystemName		Stream editing
FRAME	NumericalData/Parameter/CoordinateSystem/CoordinateSystemName		Stream editing
COORDINATE_SYSTEM	NumericalData/Parameter/CoordinateSystem/CoordinateSystemName		Stream editing
DICT_KEY	NumericalData/Parameter/CoordinateSystem/CoordinateSystemRepresentation		Stream editing
FRAME	NumericalData/Parameter/CoordinateSystem/CoordinateSystemRepresentation		Stream editing
REPRESENTATION_1	NumericalData/Parameter/CoordinateSystem/CoordinateSystemRepresentation	Yes	Rarely needs editing
DISPLAY_TYPE	NumericalData/Parameter/RenderingHints/DisplayType		
LABLAXIS	NumericalData/Parameter/RenderingHints/AxisLabel		
LABLAXIS	NumericalData/Parameter/RenderingHints/RenderingAxis		
LABLAXIS	NumericalData/Parameter/RenderingHints/Index		
FORMAT	NumericalData/Parameter/RenderingHints/ValueFormat		

(*continued on next page*)

Table B2 (*continued*)

Metadata Mapping - CDF Variable Attribute to SPASE Parameter			
CDF Variable Attribute	**SPASE Numerical Data Parameter Mapping**	**Edit?**	**NonCDF Metadata Sources, Processing Program, etc.**
FORMAT_PTR	NumericalData/Parameter/RenderingHints/ValueFormat		
SCALEMIN	NumericalData/Parameter/RenderingHints/ScaleMin		
SCALEMAX	NumericalData/Parameter/RenderingHints/ScaleMax		
SCALETYP	NumericalData/Parameter/RenderingHints/ScaleType		
PROPERTY	NumericalData/Parameter/Structure/Size		
SCALEMIN	NumericalData/Parameter/Structure/Size		
DEPEND_1	NumericalData/Parameter/Structure/Element/Name		
DEPEND_1	NumericalData/Parameter/Structure/Element/Qualifier	Yes	CDF Dictionary Key mapping but not via ADAPT, AI mapping desired
SI_CONVERSION	NumericalData/Parameter/Structure/Element/Index		Automatically loop indexed
LABL_PTR_1	NumericalData/Parameter/Structure/Element/ParameterKey		
UNIT_PTR	NumericalData/Parameter/Structure/Element/Units		
Not applicable	NumericalData/Parameter/Structure/Element/UnitsConversion		Stream editing
VALIDMIN	NumericalData/Parameter/Structure/Element/ValidMin		
VALIDMAX	NumericalData/Parameter/Structure/Element/ValidMax		
FILLVAL	NumericalData/Parameter/Structure/Element/FillValue		
Not applicable	NumericalData/Parameter/Particle/ParticleType	Yes	CDF to SPASE Mapping not fully automated via ADAPT
FIELDNAM	NumericalData/Parameter/Support/SupportQuantity	Yes	CDF to SPASE Mapping not fully automated via ADAPT
FIELDNAM	NumericalData/Parameter/Support/SupportQuantity		Parse variable name to identify positional/orientational variables

resource class includes information concerning the model's *ResourceID*, *ResourceHeader*, *Versions* (of the model), *SimulationType*, *CodeLanguage*, *TemporalDependence*, *SpatialDescription*, *SimulatedRegion*, *InputProperties*, *OutputParameters*, and *ModelURL* (see Fig. 5). *InputParameters* and other settings that specify a given model run are described in *SimulationRun* (Fig. 6) while the model output would be described in *NumericalOutput* (Fig. 7). Note that the SPASE simulation extensions Granule and Particle classes override those defined in the base SPASE schema. A full description of the SPASE simulation extension data model can be downloaded via the SPASE website, specifically: https://spase-group.org/data/simulation/spase-sim-1.0.0.pdf. We should note at this point, however, that the simulation extension is in the process of being incorporated into the SPASE base model so a unified metadata model is to be released in the near future.

3.5. Resource citation and referencing

A Digital Object Identifier (DOI) provides a unique, stable reference to a piece of digital content including articles, observational datasets, model output, and software. The DOI overarching authority is the International DOI Foundation at https://doi.org/ Many members of this group, notably CrossRef (https://www.crossref.org/) mainly for scholarly articles and DataCite (https://datacite.org/) for datasets, give users the ability to "mint" (create) DOIs. There is considerable leeway in what qualifies as a data "digital object," and it can vary from a table of numbers to a petabyte of files or images, such as the nearly 3 petabytes of data collected by the Solar Dynamics Observatory during its five-year nominal mission.

The DOI is a handle to a "landing page" that provides not only a persistence reference to a resource but also the full metadata description of the resource, such as access URLs and methods, descriptions of content including variables, caveats for data use, and potentially other things (as shown in Fig. 3). Methods or protocols, such as web services (e.g., HAPI) or a web portal (e.g., CDAWeb), for accessing digital resources are provided in *AccessURL* under *AccessInformation* of the SPASE schema (as shown in Figs. 3 and 9). The provision of access protocols by SPASE enables users to directly obtain the resources they need to support their research.

SPASE uses DOIs to allow citation of the data resources, the same way as is done for research papers. The essential feature of DOIs is the persistence of the pointer and the assurance that it points to the expected digital object. Should the digital object change in any way, the landing page can be updated without minting a new DOI. If the digital object is retracted or definitively unavailable, the DOI is kept, but points to a "tombstone page" explaining the status of the removed digital object. This form of DOI referencing for digital resources is more versatile than the traditional use of DOIs for referencing publications.

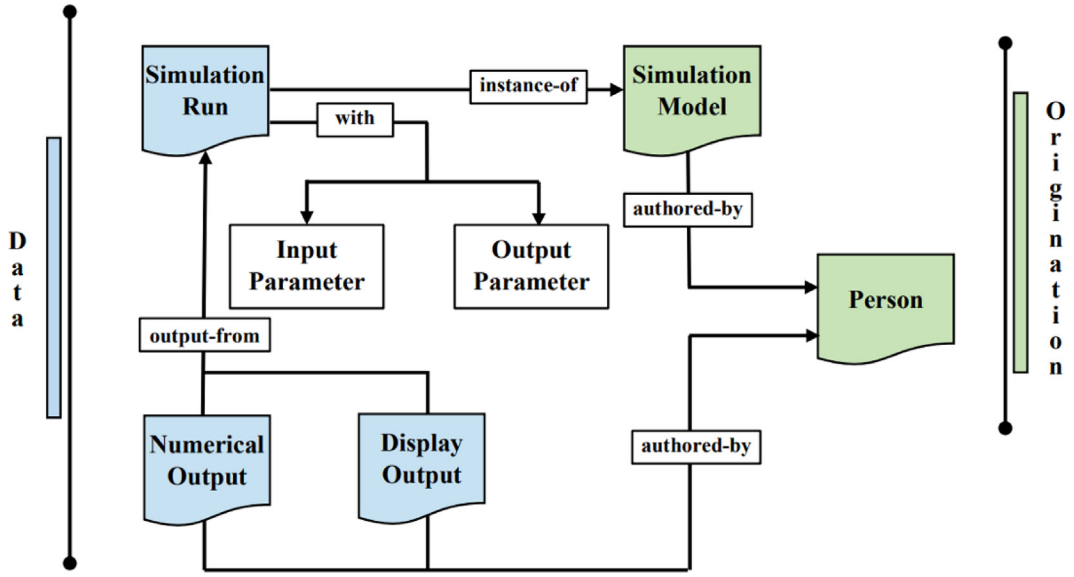

Fig. 4. The association between resources in the SPASE Simulation Extensions Information Model. All resource types available in the SPASE Base Information Model may also be used within the Simulation Extensions. Arrows point in the direction of association, as in Fig. 5.

There are various degrees of granularity that can be invoked in minting DOIs. The general practice for NASA Heliophysics is that any NASA dataset that has a SPASE description and is publicly accessible is assigned a DOI. For example, the 1-min resolution magnetometer data from the ISEE-2 spacecraft in various coordinate systems has been collected as a set of files that are registered as a SPASE dataset. It has been given a DOI (https://doi.org/10.21978/p8t923). Clicking on the DOI (note that the full URL for the DOI is the expected form of the DOI) brings up the landing page that includes an example of how to cite the dataset along with access methods (under *AccessInformation*; see Figs. 3 and 8) and a great deal of other metadata. All of the information on the landing page is essentially the SPASE description. Value-added data products, derived from the measurements of multiple spacecraft, like the very popular OMNI datasets have also been assigned DOIs (e.g., https://doi.org/10.48322/mj0k-fq60 for the OMNI version 2, 1-min resolution).

Although over 10,000 datasets, mostly under the NASA *NamingAuthority*, have already been described using SPASE, NASA alone still has some way to go to mint DOIs for all the SPASE-registered data products. A given SPASE description of a resource, with PublicationInfo provided under the ResourceHeader [see Fig. 2 in Roberts et al.(2018) or Fig. 3**],** can nevertheless contain all the information required to mint a DOI. Using simple software tools, all the pertinent information can be extracted to create a landing page that will become the permanent DOI reference page. In turn, the DOI can then be inserted back into the SPASE description to cement their linkage. **The utilization of DOIs within SPASE descriptions allows for precise and consistent citation of data resources, a clear benefit for the Heliophysics research community.**

In the case of the ESDC datasets, the ESA approach has been to first register collections of heliophysics data associated with each "experiment" (an instrument or a set of related instruments) for any ESA heliophysics mission (Masson et al., 2021). This provides users with the ability to cite, in a paper, any bunch of datasets produced by one experiment with one DOI, and acknowledge the experiment Principal Investigator. There are on-going efforts to also generate a DOI per dataset.

Another level of granularity is that of the "collection" (rather than a dataset from a single instrument as above). A collection can be a composite dataset, as described later in section 4.1.3.1, that is a heterogeneous grouping of digital objects united by some purpose, such as in support of a particular investigation reported in a journal article. In astrophysics it is common to collect the pointers to all the observations needed for the article and to bundle them with one DOI. ESA has generated thousands of DOIs related to data bundles associated with each selected observation proposal for observatory-type astrophysics missions (e.g., XMM-Newton) or data releases for survey-type astrophysics missions like Gaia. In the planetary domain, ESA has assigned a DOI to all planetary datasets which are, by definition, bundles in this field (see Masson et al., 2021 for more details). Eventually, ESA envisages allowing users to mint DOIs on the fly in its archives related to datasets, versions, and time periods used in a particular research study in order to include that such DOI in their paper. This will enable individual scientists to directly link their paper to the datasets used, enhancing the reproducibility of their results.

There might also be a DOI associated with data output from software, such as a simulation model. For these types of objects both the software and the output will have DOIs

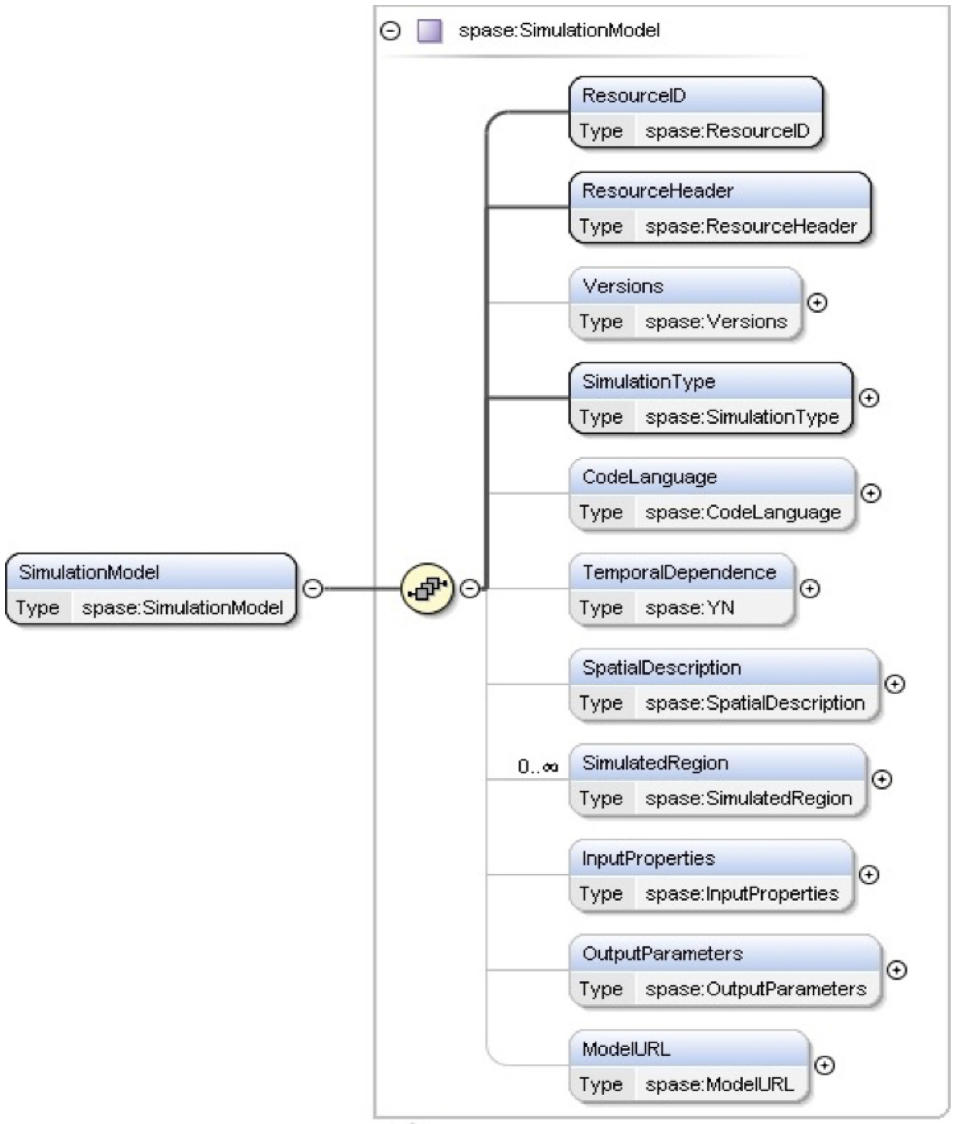

Fig. 5. SPASE description schema for a simulation model. Required description items are linked with dark connection lines, whereas those with light lines, though optional, are recommended to ensure full description. Furthermore, some attributes, "0" for required and "1" for optional, may have multiple entries, signified by "∞".

which will aid in the reproducibility and validation of results. This style of DOI usage is becoming more common in Heliophysics as Open Science (see section 5.1) gets more widely adopted. One could also imagine a study partially based on heliophysics datasets/software/simulation described in SPASE combined with datasets (e.g., planetary, astronomy) not described in SPASE.

In this context, one service worth highlighting is Zenodo (https://zenodo.org), which allows a user to submit any digital object and generate a DOI. It is indeed one of the general repositories (https://doi.org/10.5281/zenodo.3946720) listed in the Data and Software sharing guidance for authors submitting to AGU journals for instance (https://doi.org/10.5281/zenodo.5124741). However, it should be noted that this AGU sharing guidance recommends its authors to first look for a repository that specializes in the data of the authors' scientific domain (see https://data.agu.org/resources/useful-domain-repositories), like the SPDF and SDAC for heliophysics data, as this will maximize the probability that the deposited data will be FAIR-compliant.

4. Science-enabling data services and tools

We have so far examined how metadata is needed to support different infrastructure functionalities in the heliophysics data environment (sections 2) and how the SPASE metadata model is compliant with the FAIR principles and would provide uniform description of multi-disciplinary digital resources to support cross-disciplinary research (sections 3). Here we use a few science task examples to illustrate the use of SPASE metadata to facilitate unfettered flow of digital resources, data in particular, required by various heliophysics and space weather research activities

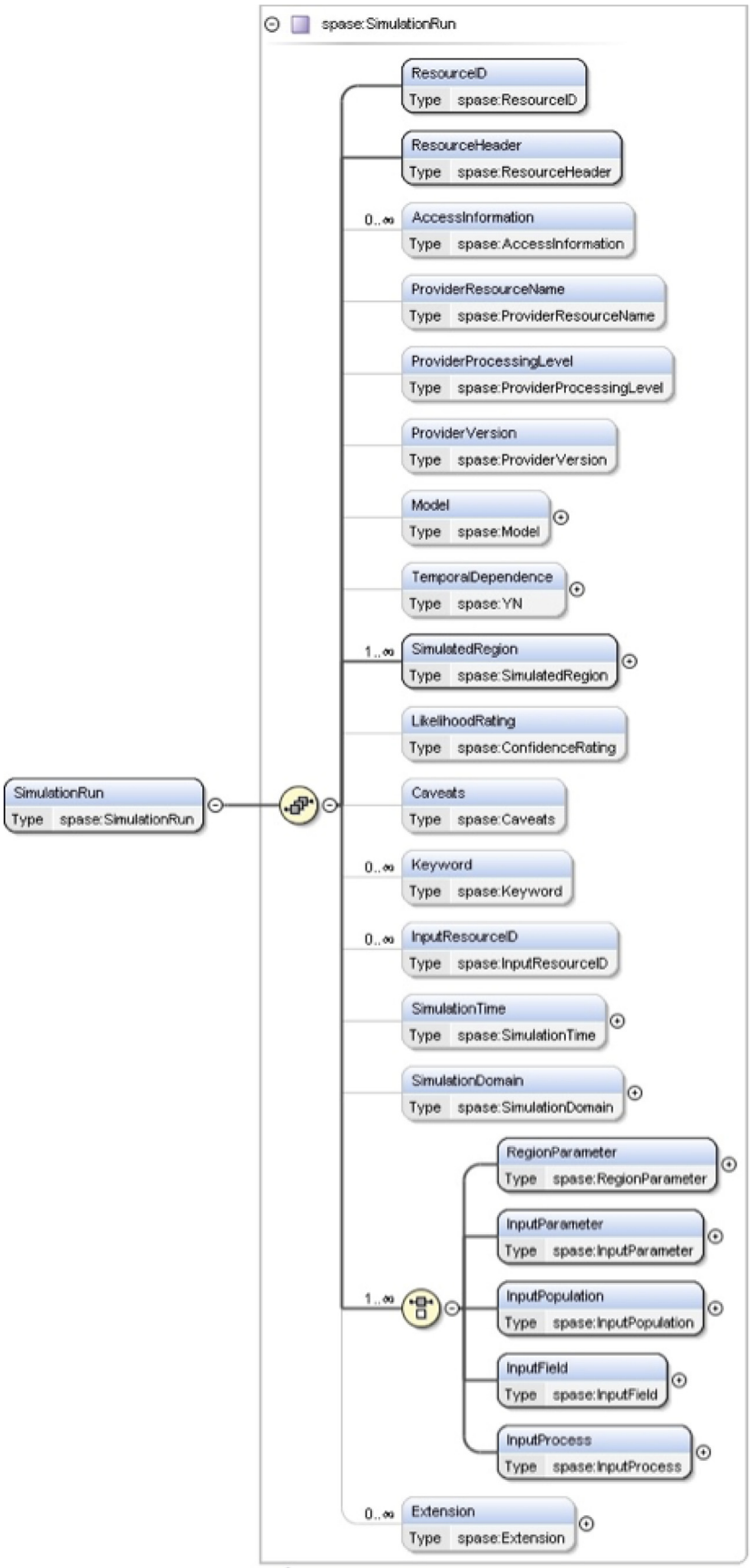

Fig. 6. SPASE description schema for a simulation run in the same format as Fig. 5.

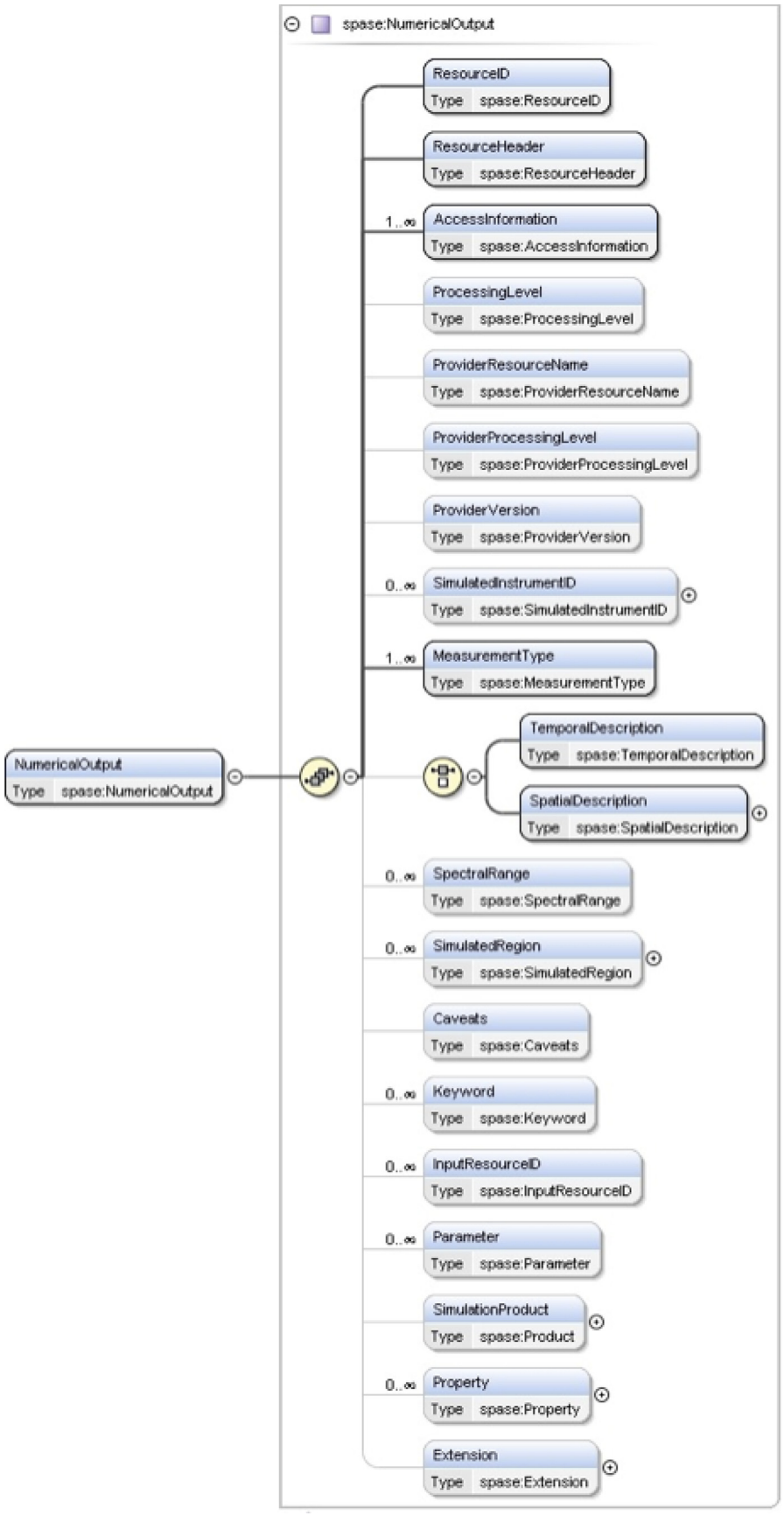

Fig. 7. SPASE description schema for simulation output in the same format as Fig. 5.

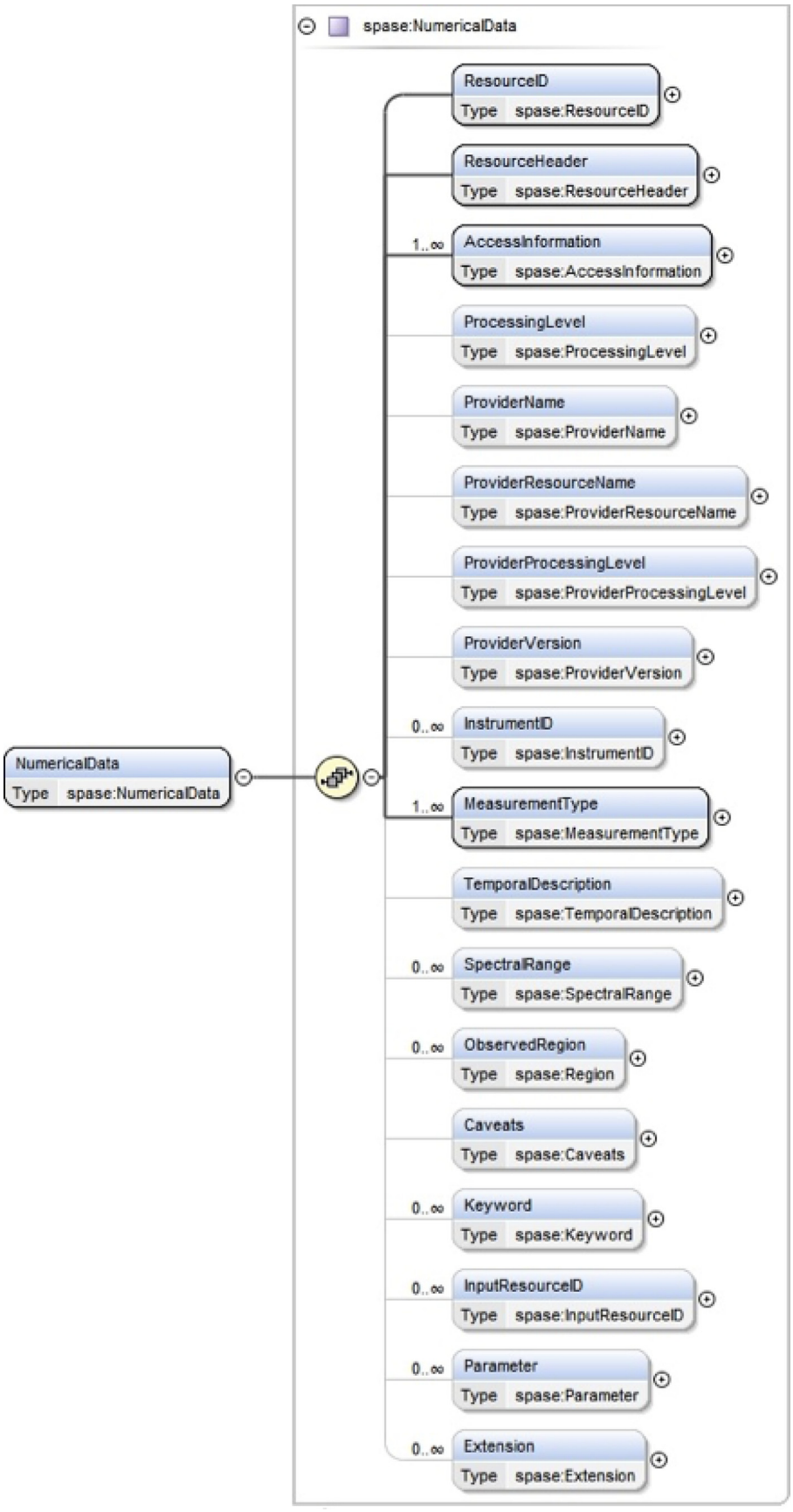

Fig. 8. SPASE schema for numerical data resource attributes in the same format as Fig. 5.

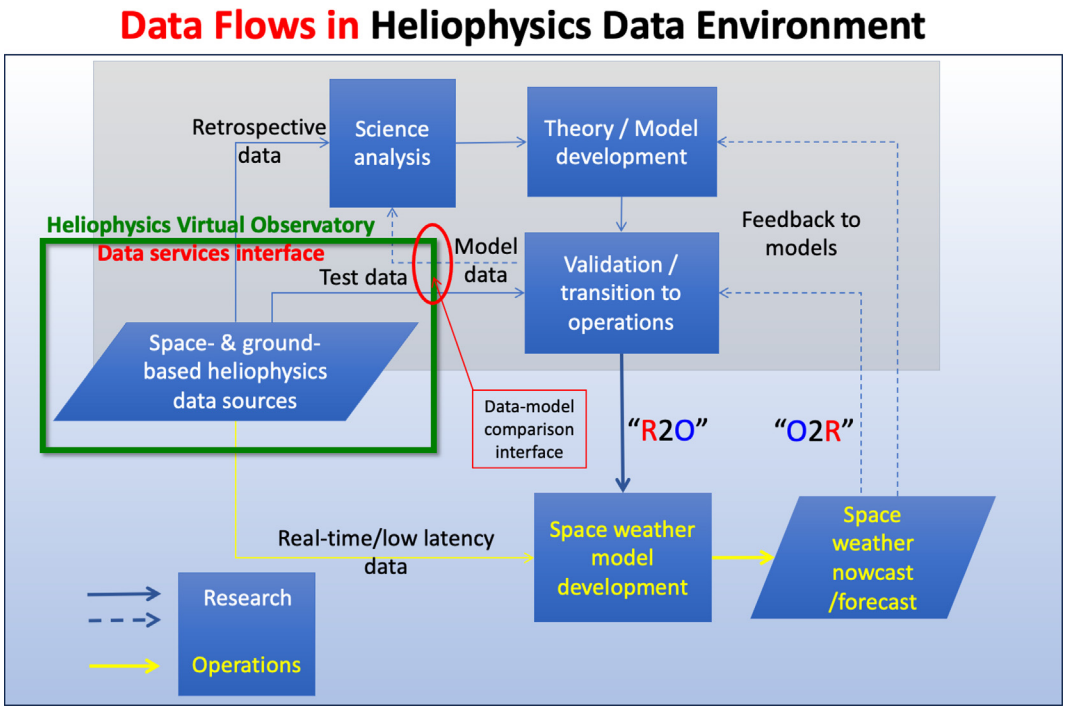

Fig. 9. A schematic of data flows throughout the heliophysics data environment, from different data sources or repositories, served through an appropriate user or application-programming interface (green boundary), to support different types of retrospective research tasks (in gray box) and space weather modeling and forecasting. The data system interface may also support data-model comparison (red oval region) as described in section 4.2.4. The "R2O" and "O2R" pathways between the research and space weather operations domains then form a feedback loop for improving model performance and physical understanding of space weather processes.

occurring within the heliophysics data environment (Fig. 1) as laid out in section 2.1.

Fig. 9 shows a schematic of data flows from different data sources or repositories, through interfaces represented schematically by the green boundary to various user locations where data would be used to support different types of science tasks, such as analysis and modeling. The blue tags connected by blue arrows within the gray box cover the domain of traditional research activities based on retrospective science analysis and modeling. Outside the gray box, the tags with yellow labels connected by yellow arrows represent the domain of space weather modeling and forecasting for which low-latency data (e.g., real-time data) are more typically used. The green interface to both archived and space weather data sources may also serve as an interface to support data-model comparison (red oval, see section 4.2.4).

Since Heliophysics is the underlying science of space weather, knowledge gained through science analyses are represented in models. Physical understanding is attained when models are tested and validated by independent observations. Through the research-to-operations ("R2O") transition, research models can be used to develop space weather nowcast/forecast models. Space weather forecast performance, fed back into the research domain via the operations-to-research ("O2R") pathway for error analysis, can in turn lead to model refinements and improvements in physical understanding.

4.1. Documenting data for archiving and distribution

The primary purpose of data archiving is to preserve the data for future use, especially by users other than the originator or producer of the data. Distribution of archived data to widely distributed users who may have different data requirements and expertise is thus an inherent function of a data archive as a key component of the heliophysics information architecture (Fig. 1) and data flow environment (Fig. 9). Upon archiving a resource (any of the blue tags in Fig. 1), the documentation of the resource should provide all the information needed to locate, retrieve, and understand the resource sufficiently so that it can be used independently. Section 4.1 first discusses briefly how the SPASE metadata model facilitates the documentation of different types of data products so that when a data source is being queried or searched, as depicted in Fig. 1, each product can be uniquely identified and retrieved to meet the query requirements. After obtaining the relevant resource from the data source, users would have sufficient understanding of the resource to be able to use the resource independently in various science tasks, as shown in Fig. 9.

4.1.1. Observational data

As depicted in Fig. 9, observations are the data sources at the beginning of many heliophysics studies of fundamental and operational significance. Understanding the con-

tents of the measurements is basic to analyzing and interpreting the analysis results. The SPASE metadata model (section 3) provides a framework and vocabulary for describing and conveying the meaning of the contents of various resource types shown in Fig. 2. Following the example in Fig. 3, we illustrate how the SPASE metadata model captures the essential information of a numerical data resource obtained from heliophysics observations.

Fig. 8 shows the SPASE schema and the essential attributes or categories of information for describing a numerical data resource. The four required attributes (indicated by the dark connecting lines) correspond to the top four items under "Outline" in Fig. 3. *ResourceID* is the unique identifier of the resource. *ResourceHeader* contains all the global attribute descriptions of the resource as a whole (see Fig. 2 in Roberts et al. (2018) for the full list of header information). *AccessInformation* provides information on how and from where to access the resource electronically, such as through a web interface or web services. This is also where data format information is provided so that users would be able to know how or what tools would be needed to access the content of the data. *MeasurementType* indicates the nature of the type of observational measurement, such as ion composition, field, or wave, etc.

The information required to describe a digital resource as a whole, such as an entire dataset, may be sufficient for searching and locating the resource; but the global-level description alone is insufficient knowledge for accessing and understanding the contents of the dataset. As discussed in section 2.2.4 above, descriptions of data format (2.2.4.1) and parameters (2.2.4.3) effectively provide the balance of the necessary information on accessing and understanding the data contents. Figs. 3 and 8 show that both *AccessInformation*, which includes the specification of data format, and parameter descriptions that are included in the SPASE information model are important information for enabling accessibility and usability of the data.

SPASE descriptions of observational datasets registered on the SPASE registry can be found by perusing the SPASE resource landing pages posted at https://hpde.io/. For example, the landing page of the SPASE description for the fast-mode (4.5 s) electron phase space distribution data obtained by the dual electron spectrometer (DES) of the fast plasma instrument (FPI) on the number 2 spacecraft of the Magnetospheric Multi-Scale (MMS) mission is posted at https://hpde.io/NASA/NumericalData/MMS/2/FastPlasmaInvestigation/DES/Fast/Level2/Distribution/PT4.5S.html and displayed in Fig. 10. From the landing page, one can find that the *ResourceID* of the dataset is simply given by the landing page url but with "https://hpde.io/" replaced by "spase://". As discussed in section 3.2, The URI of the *ResourceID*, or the landing page url, of a resource essentially reflects the unique path by which the resource is found on the SPASE registry. Using the *AccessInformation* provided in the SPASE descriptions, all SPASE-registered resources can be uniquely identified and found. The description, *InformationURLs*, and other metadata on the landing page help the user understand the dataset.

4.1.2. Model data

As mentioned in section 3.4, the SPASE metadata standard has adopted the simulation extension that was first developed by the IMPEx project. Any group can now choose to adopt SPASE and use it for describing models, model runs, and model output. CCMC has decided to do exactly that. The CCMC currently hosts more than 80 models and model combinations. The Runs-On-Request (ROR) Systems at the CCMC currently has over 22,500 simulation runs from those CCMC hosted models. Outputs of all runs can be visualized through the CCMC sophisticated web-based visualization and analysis system and requested for downloads through each simulation runs' webpage (specifically via the "Request output data as a single archive file" button, see Fig. 11).

Providing tailor search, discovery and easy access to those models and simulation runs has been one of the CCMC goals to better support the community. As such, CCMC has been adding metadata following the SPASE metadata standard (Fig. 4) for all CCMC hosted models (Fig. 5) and ROR runs (Figs. 6 and 7). All such metadata will be made available on the CCMC Metadata Registry (CMR), https://kauai.ccmc.gsfc.nasa.gov/CMR/view/metadata The current version of the ROR runs' metadata are viewable on each simulation output's landing page (e.g., the example displayed in Fig. 11). The CCMC plans to use such metadata as an important backend to provide information to other CCMC systems as well as external systems (e.g., other heliophysics archives that use SPASE as the metadata standard for their data). As of 2021, the requirement to provide simulation model metadata has been added as part of CCMC model onboarding process, and there are more than 98 models and model versions currently stored in the CCMC Metadata Registry. Such information is being used as the backend of the CCMC model catalog, https://ccmc.gsfc.nasa.gov/models/.

In addition, the process of collecting simulation run metadata, as in Figs. 6 and 7, for all ROR models has been defined. As of 2023, CCMC has collected metadata from 4000 + ROR runs produced by 45 ROR models and model versions. The collected ROR runs metadata will be an important backend as the CCMC works on improving the ROR service for the community, including adding an option for users to browse the data (e.g. the developing 'Browse output data' option on the simulation run webpages, see Fig. 11). Externally, the CCMC is working with the SMWT to add the models and ROR runs metadata into the official SPASE registry. Being an early adopter of the SPASE standard for simulations, the CCMC is working with the SPASE groups to provide feedback as needed on the SPASE standard itself. Therefore, the SPASE standard will continue to improve and evolve to support the needs of the community, including the need

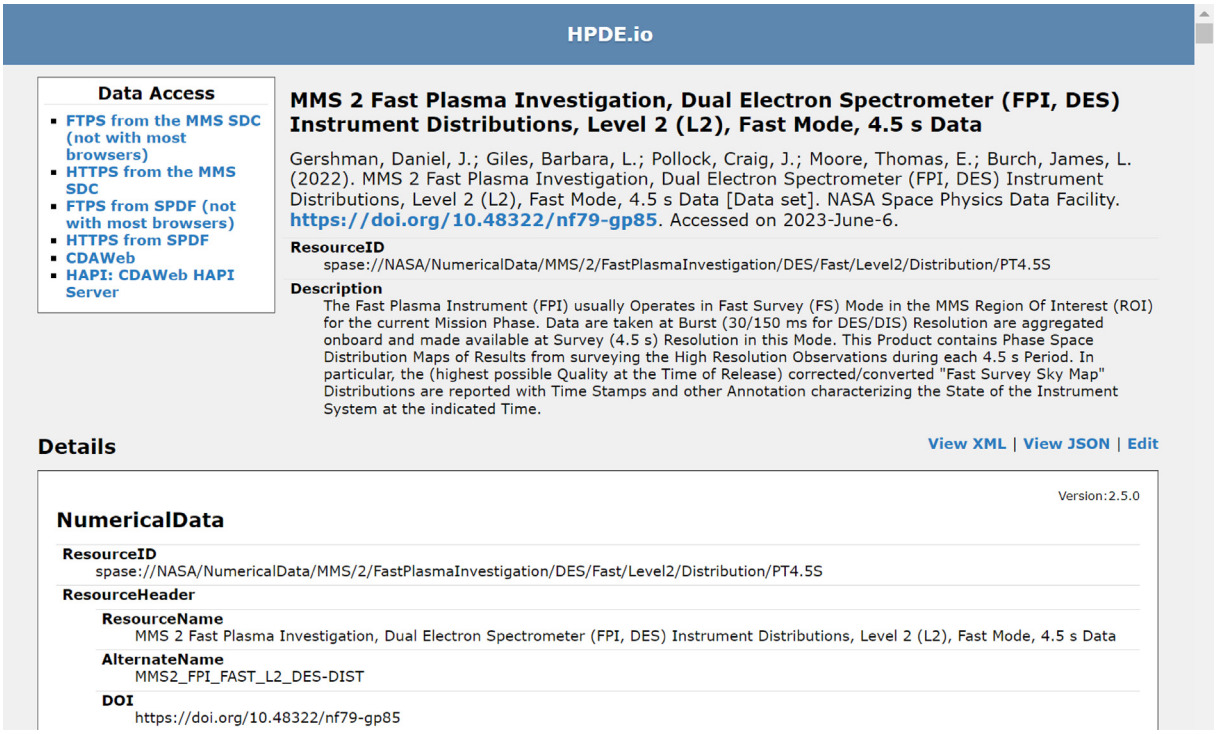

Fig. 10. A screenshot displaying the SPASE landing page for an example observational dataset. The SPASE metadata model supplies all of the information seen here (including the full bibliographical reference with DOI URL just beneath the resource title) and much more than can be seen further down on the page. The URL for this landing page is https://hpde.io/NASA/NumericalData/MMS/2/FastPlasmaInvestigation/DES/Fast/Level2/Distribution/PT4.5S.html (See section 4.1.1 for more details).

to uniquely identify, retrieve and understand modeled data as depicted in Figs. 1 and 9.

4.1.3. Higher-level processed datasets

Sections 4.1.1 and 4.1.2 focused on the digital resources obtained directly from instrument measurements and simulation models, respectively. Higher-level data products can also be derived from instrument data when researchers apply an additional layer of processing or when data from different instruments or platforms are combined to support a specific study or type of analysis. These datasets are also important in advancing heliophysics and thus require the same capabilities as observational and modeled data, namely for users to uniquely identify, retrieve and understand the dataset. We discuss two examples in the subsections below.

4.1.3.1. Composite datasets and SPASE description strategies. Composite datasets, which may contain data from different sources and instrument types, are becoming more common. Cutting edge Heliophysics research studies often require data assimilation, for instance, combining observations and modeled data. Solar wind data such as those available from OMNIWeb (King and Papitashvili, 2005; also see https://omniweb.gsfc.nasa.gov/) represent a classic example of a composite dataset. OMNIWeb is available from 1963 onward and combines spacecraft ephemeris data, solar wind plasma data, and interplanetary magnetic field data from the ACE, Geotail, IMP 8, and Wind spacecraft along with geomagnetic indices and energetic flux measurements from geostationary satellites. Also, almost all space physics journals now require authors to submit the data analyzed in support of published research. It is therefore important to have a way to describe these composite datasets to ensure that they are also FAIR-compliant.

SPASE descriptions of both simple and composite datasets are best exemplified by the *NumericalData* resource (see section 3.3), as shown in Figs. 3 and 8. That said, the same rationale outlined below also applies to the SPASE *Catalog* and *DisplayData* resource types (see left column in Fig. 3).

SPASE 2.5.0, the current version of the SPASE schema (https://spase-group.org/data/model/index.html), includes a new resource type (destined to the data domain in Fig. 2) called *Collection*, which has been added to accommodate the need to describe datasets containing data from multiple, disparate data sources. While we refer the readers to the SPASE model dictionary in general, the SPASE dictionary defines a *Collection* resource in part as:

> ``An aggregation of resources, which may encompass collections of one resourceType as well as those of mixed types. A collection is described as a group; its parts may also be separately described. . .All the resources that are part of the research effort can be described as a Collection.''

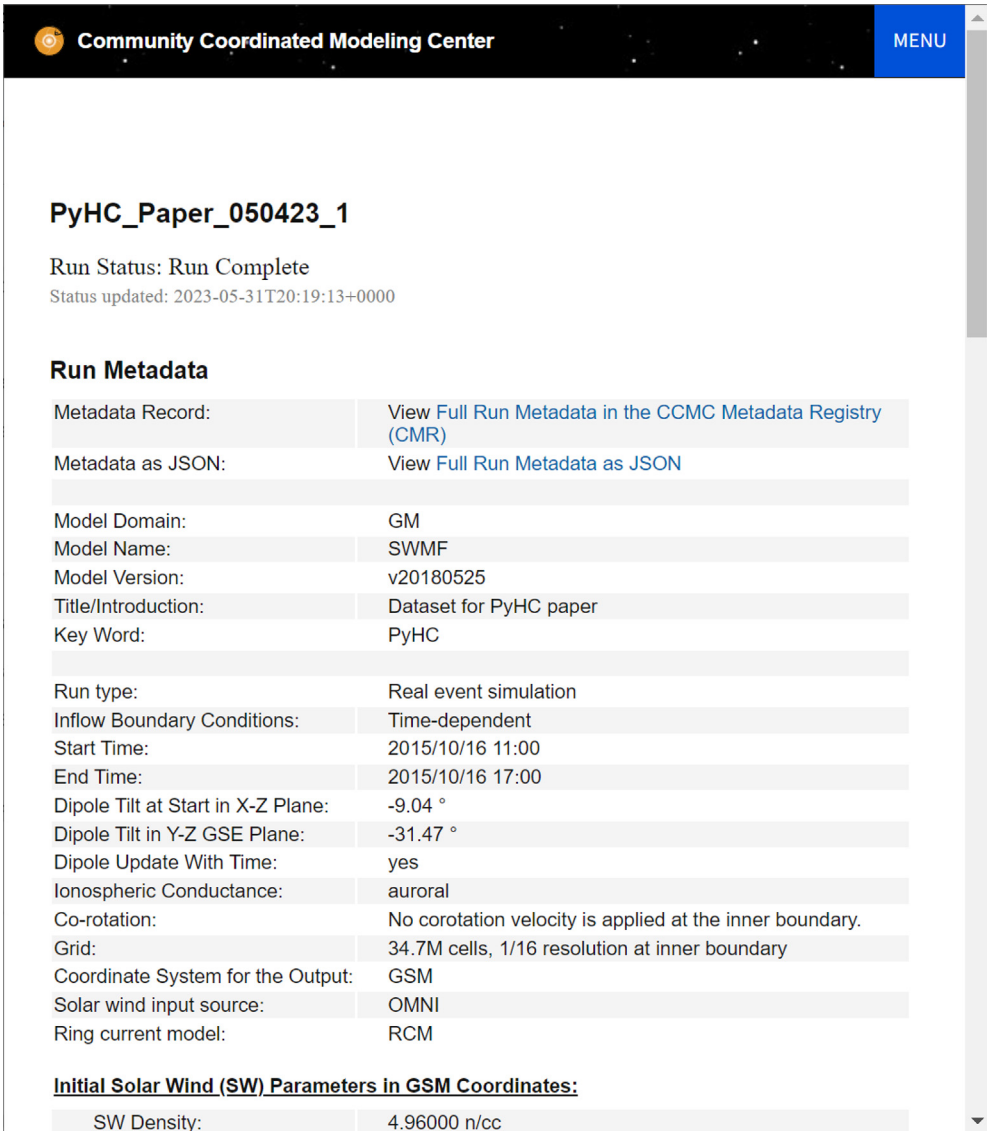

Fig. 11. A screenshot of a landing page for an example simulation output showing a variety of metadata recorded for the output. At the bottom of the page are two links labeled "Request output data as a single archive file", which takes the user to an online form to complete the request, and "Browse output data", which will allow the user to browse the files in the simulation output once the feature is fully developed (See section 4.2.4 for more details).

The *Collection* metadata contains a *ResourceHeader* and an unlimited number of Member containers. Each Member container includes text fields for *ResourceName*, *Description*, and *MemberID* that specify details of the individual components of a given collection. The *MemberID* text fields are to be populated by the *ResourceID* of the SPASE descriptions of the *Catalog*, *DisplayData*, *NumericalData*, etc., comprising the Collection. The Collection metadata resource type directly targets the need to precisely describe compilations of data related to research publications, machine learning, and other composite data sets. To date, only a few SPASE Collection descriptions have been generated to describe a few aggregations of observations, see https://hpde.io/NASA/Collection/index.html. We expect Collection resource descriptions to grow as more and more higher-level products are generated from research projects.

4.1.3.2. Datasets for data science or machine-learning analysis. Artificial Intelligence (AI) and Machine Learning (ML) have become powerful tools in the space physicist's analysis toolbox to assist in scientific discovery, pattern recognition, feature importance, and prediction tasks. Search and access are prerequisites to data science analysis. Metadata and information models are the structure to the information in our world. They give us a way to navigate the deluge of 21st century digital life and permit discovery. All scientists, engineers, researchers require an understanding of metadata and information models. However, a bottleneck in AI/ML applications is in the preparation of the data for the algorithms, which requires gathering data from various archives, understanding the quality of each data point, aligning them spatially, temporally, or both, potentially rescaling or normalizing, and

structuring the integrated dataset for ingestion into the chosen algorithm. While there are numerous space physics examples that describe the process (e.g., McGranaghan et al., 2018b, 2021b; Sadykov et al., 2021) and provide guidance on the development of the metadata that would be needed to identify and support the searching for or discovery of these datasets, there is not yet a metadata model that can completely and uniquely describe AI/ML-ready datasets. Considerations are being made by the SPASE Group and the SMWT for future extensions of the SPASE model schema to enable descriptions of the growing number of these datasets. More discussion on challenges related to AI/ML datasets metadata are provided in the future outlook section (5.2).

4.2. Data search and access

Finding and accessing digital resources are two key elements of the FAIR principles. They represent two basic tasks that researchers must do to obtain the data needed for their research. In the traditional data environment in which users search and access data from a centralized data archive, accessible resources are usually limited to what is available from the archive (see section 2.2.2). In a more modern environment, the searching and querying functionality reside outside of the repository to form a middleware, i.e., a software system servicing information flows between the users and the data sources along paths (1), (2) and (3) in Fig. 1, so that uniform search and access capability can be implemented to query resources from multiple, distributed sources (Merka et al., 2008a;b). As noted in section 2.1, for the middleware approach to be effective, digital resources need to be uniformly and adequately described. Although uniform metadata are available from resources provided in standard, self-describing and self-documenting data formats already in use in Heliophysics as discussed in section 2.2.4.1, metadata embedded in data files are accessible only when the data files are opened, making searching and reading the metadata descriptions of the resources inconvenient and cumbersome.

The middleware approach based on the basic SPASE metadata model had been used to develop several heliophysics virtual observatories (VxOs): the virtual heliospheric (VHO) (Szabo et al. 2007), the virtual magnetospheric (VMO) (Merka, 2006; Meerkat, 2006; Walker, 2007), the virtual energetic particle (VEPO) (Cooper et al., 2007), and virtual wave (VWO) (Fung, 2008; 2010) observatories. The virtual solar observatory (VSO), however, does not use SPASE (Hill et al., 2009). All the SPASE metadata resources of the early SPASE-based VxOs were hosted in a single SPASE registry residing on Github at https://github.com/hpde/. Lessons learned from the pioneering VxO implementations and the availability of more SPASE descriptions of digital resources based on the more recently released and more matured SPASE model have enabled the heliophysics data portal (HDP) mentioned in section 2.2.2 and helped spur the recent development of the Heliophysics Digital Observatory (HDO; https://msqs.gsfc.nasa.gov/hdo/public/) by consolidating and streamlining the previous VxOs (Fung et al., 2021; 2022).

Fig. 12 shows the general layout of the information architecture, illustrating the relationship between the users, the middleware, and different (distributed) sources of digital resources (e.g., data). The green box is the same as the one in the heliophysics information flow diagram in Fig. 9. The middleware, consisting of a query builder and the search engine, communicates with the users via paths (1) and (2) in Fig. 1 and with the data sources via path (3). Direct communication using web services through applications programmer interfaces (API) would take place via path (3a) in Fig. 1. It is important to recognize here that only transfer of metadata is needed along (1), (2), (3), and (3a), making the tasks performed by a middleware much more manageable and efficient even over the internet, and the technical approach more focused and intuitive. If SPASE descriptions of datasets need updating or changing, such as due to reprocessing or relocation of the data, the same middleware would still operate seamlessly.

We do recognize, however, that many functionalities of the traditional data environment are simply replicated by a middleware- and SPASE-based environment. After all, most, if not all, of the functions require only having the right metadata, which is not required to be in SPASE form. Traditional data services are invariably offered by different data repositories or archives, but searching and accessing different archives would require using different interfaces developed and implemented independently by those archives. This multiplicity of interfaces makes getting and using distributed resources extremely inconvenient and inefficient. The advantages of SPASE then are that (1) it is developed specifically for describing heliophysics and space weather resources, so by design it would have all the right terms and fields needed for describing heliophysics products, (2) it has an extensive and extendable scope of information, (3) its standardized and backward-compatible schema enables uniform resource description that facilitates the use and operations of middleware software systems, and it allows full access to metadata without having to open any data files.

We show in Fig. 13 a screenshot of an example of magnetospheric resource selection panels of the HDO, which is completely driven by registered SPASE metadata (Fung et al., 2021; 2022). The panels result simply by specifying a time interval of interest as the beginning input to the HDO query builder (Figs. 1 and 13). With the time interval specified, the Observatory (left) panel would display all the platforms operating within the specified interval as highlighted in blue. Bolded text indicates platform multiplicity that can be expanded by clicking on the expansion triangle. Other observatories not operational during the specified interval would be grayed out and would not be selectable. Upon selecting the observing platforms of interest, the SPASE metadata for the selected observatories, *IMAGE*

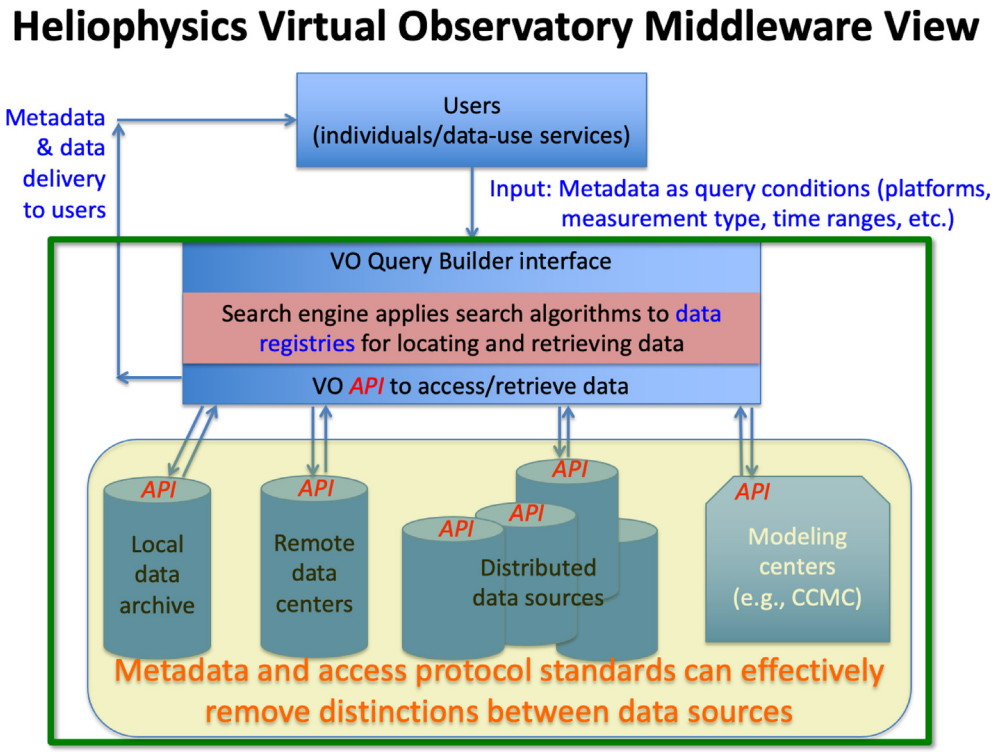

Fig. 12. Information flow in an architecture involving a middleware for handling communication with users for data querying and searching and with distributed data sources, including modeling centers with model execution capabilities, for data accessing and retrieval. The overall architecture with uniform communications supported by a set of metadata standards (yellow enclosure), such as SPASE.

and *Geotail* in this example, would automatically populate the Instrument and Data Product panels with all available resources from the selected observatories. The user can then select the specific data products from the instruments carried by the selected observatories obtained during the specified time interval and obtain the pertinent available data files by clicking the Query button. It is clear from Fig. 13 that the data search and access processes are more streamlined and intuitive than traditional web portal interfaces in which time interval, observatories, instruments, and data products must be selected independently.

The Heliophysics Data Portal (HDP, https://heliophysicsdata.gsfc.nasa.gov) provides a search capability also based on SPASE metadata. It enables searches primarily for data services. It allows searching for availability of data using date ranges, keywords, measured parameters, observatory names and observed regions, and a number of other terms, in any order. Detailed descriptions of data products, which become DOI landing pages as DOIs for the datasets are minted, are always available. Access methods from services for all available data products, including web browser tools, and web service methods such as HAPI, are all directly provided to the extent they are offered by the providers. In many cases, browse product plots can be produced directly from an HDP page.

One important element within a SPASE description is the *AccessURL* (Fig. 3), which lists the access protocols and services available for obtaining the data via path (4) in Fig. 1. This allows SPASE-based search portals such as the one depicted in Fig. 13 to offer immediate connections to the data discovered in a search. For example, if the *AccessURL* points to a human-focused web page, users can visit that page and download data through that page. Most observational datasets are also available through direct HTTP or FTP download (directories of files exposed online). Some *AccessURL* entries point to an API access mechanism, i.e., something intended not for direct clicking but for programmatic access via client software that understands the API (see section 2.2.4.4). The NASA SPDF offers the Coordinated Data Analysis System (CDAS) Web services (see section 2.2.4.4 and Table 2), and they also offer client software that users can run for creating programmatic or automated downloads (Table 2).

Another API-based access mechanism is the Heliophysics Application Programmer's Interface (HAPI) (Weigel, 2021b; HAPI version 3.0.0 Specification, https://doi.org/10.5281/zenodo.4757597), which is also a COSPAR Recommended access protocol standard (see COSPAR, 2021) for serving time series data. A HAPI server has a URL endpoint to describe the datasets it offers (a "catalog" endpoint), and another endpoint to list the parameters available within each dataset (via an "info" endpoint that takes a dataset id). The SPASE *AccessURL* for a dataset accessible via HAPI would list the HAPI server URL and the dataset name, and then any software that knows how to use the HAPI protocol can request and

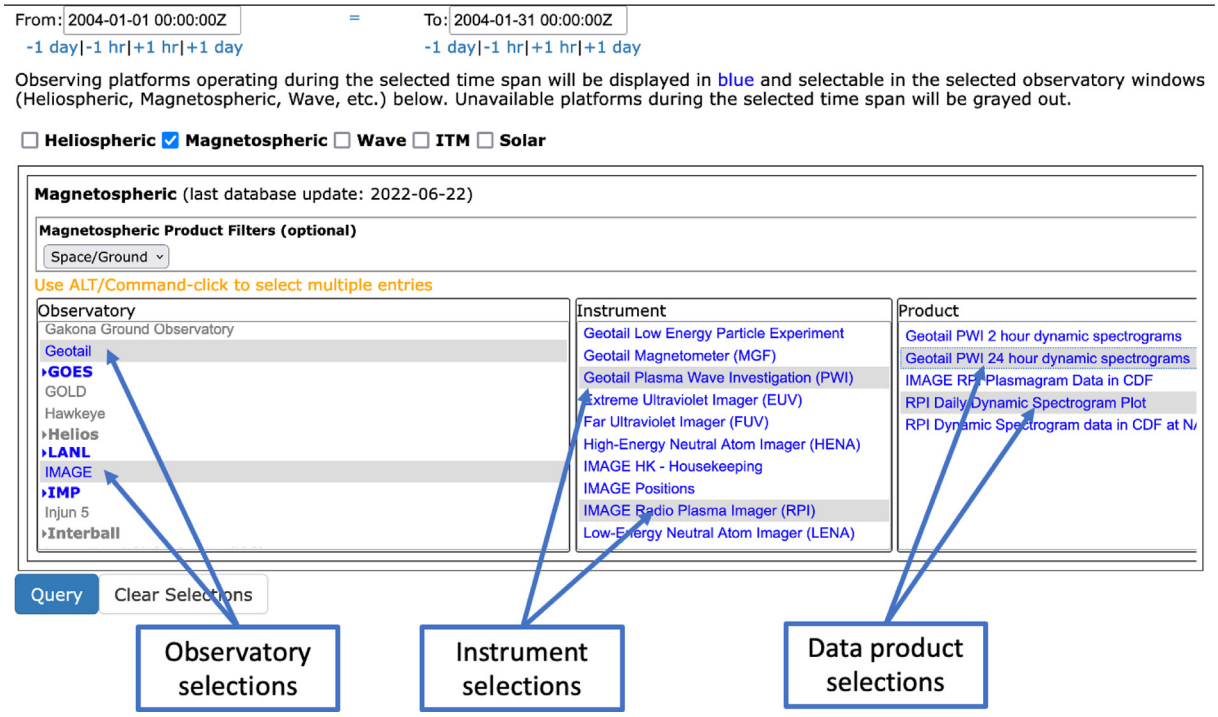

Fig. 13. A screenshot from the Heliophysics Digital Observatory (HDO; https://msqs.gsfc.nasa.gov/hdo/public/), showing the different panels, for example, for magnetospheric resource selections for the specified time interval that are dynamically powered by the underlying SPASE metadata. Grayed-out mission names indicate those missions have no registered resources available for the specified time. Upon clicking on the Query button, a user can retrieve the data files for the selected products available for the specified time interval (at the top of the figure) from the selected instruments and observatories. The HDO would play the role of the Heliophysics Virtual Observatory shown in Fig. 9.

obtain digital data for that dataset. HAPI servers provide data using a very simple streaming format which can be conveyed as simple CSV, or servers can optionally offer JSON or binary versions of the streamed data. A sample HAPI request for data from an example organization has the following standard form:

https://example.org/hapi/data?id = DATASET_ID&start = 2020–01-01 T15:35:00.123Z&stop = 2020–01-05 T00:00:00Z.

The result from a valid URL will be a stream of data covering the entire time range requested. If the time covers data from multiple files, it is the server's job to stitch this data together and serve it as a continuous stream. By hiding these arbitrary file boundaries, the HAPI standard offers a simple, conceptual way to think about time series datasets as collections of parameters. In fact, the goal of HAPI is to capture the lowest common denominator of what is needed to serve time series data, making it easy for data centers to add HAPI alongside existing services. HAPI has been implemented by large and small data providers. A list of active HAPI servers is available at https://hapi-server.org/servers.

HAPI and SPASE are independent but interact in important ways. SPASE focuses on data description and enables discovery, while a HAPI server is focused on access – providing a standard computer-based protocol to get to the numbers. HAPI metadata is minimal and focused on what is needed to scientifically interpret the numeric values. HAPI allows additional metadata (in ways that will not conflict with required HAPI keywords), and it has an optional keyword that can be a link to more rich metadata such as a SPASE document. A search engine for data that uses SPASE lets users find data, and then users can employ other programs in their workflow to obtain data through HAPI (or CDAS, or any other protocol listed in the *AccessURL* element). Many clients and software packages (e.g. in the PyHC) already understand HAPI, and given a top-level HAPI server URL, can let users explore the contents of a HAPI server and load data for user-selected time ranges. Autoplot (Faden, 2010) is one existing HAPI-enabled analysis program outside of the PyHC, and Fig. 14 shows an exploration window where a user can select data to plot. The figure also shows how multiple data sources are easy to combine in one tool since that tool can talk to any HAPI server. There are also IDL, Python, and Matlab codes that can read HAPI data, so users can incorporate data reading into their own workflow.

Das2 is another access API that can be listed in the SPASE *AccessURL* element. It is a second-generation web-based data delivery, visualization, and analysis system from the University of Iowa (Piker et al., 2018). Having various Java libraries and client-side applications that support space science data visualization such as interactive spectrograms, overlays, waveform plots, etc., Das2 has more features than HAPI, making it both more powerful and also somewhat harder to implement (https://das2.

org/). It also permits server-side data resampling on the fly, smoothing the data exploration user experience in tools like Autoplot. Das2 is used for space-based time-series and dynamic spectra, as well as ground based low frequency radio astronomy dynamic spectra.

A solar system data access protocol has been developed following the successful European Union-funded Europlanet research programs, within the VESPA (Virtual European Solar and Planetary Access; see Table 3 in section 2.2.4.4) team. The EPN-TAP (Europlanet Table Access Protocol, see Table 3) interface allows tools and users to discover data products based on science-driven metadata (temporal, spectral and spatial coverage, sampling and resolutions, observed target, instrument, measured parameters, processing level...). The scope of VESPA includes Heliophysics, with several EPN-TAP services providing solar or planetary magnetosphere data collections (for more detail, see https://www.europlanet-vespa.eu/themes2024.shtml).

We discuss in the following subsections a few science task examples to illustrate how SPASE enables applications not initially envisioned, such as supporting event list analysis and data-model comparisons, to further demonstrate the capability of the SPASE metadata model as the digital resource landscape expands.

4.2.1. Event lists

Analysis of a heliophysical event often requires observations from different instruments or even different platforms (ground stations and spacecraft). For example, to study the evolution of a coronal mass ejection (CME), one would need to use remote sensing observations from ground-based and/or space-based solar monitors (magnetogram, ultraviolet observation, white-light coronagraph, etc.) and spacecraft observations in the solar wind (magnetic field, plasma data, radio signal, particle distribution, etc.). SPASE metadata can easily support selection of products of a given MeasurementType produced by instruments of a given InstrumentType across multiple missions or platforms. Furthermore, coordinated modeling is sometimes used along with observational data for in-depth event study. Thus, it is crucial to make the cross-mission and interdisciplinary data easily searchable and accessible, and this can be facilitated by adopting the uniform SPASE metadata standard for both the observational and simulation data (SPASE Group, 2014, 2021).

Based on analyses of similar events, event lists or catalogs can be constructed for studying specific heliophysics phenomena more broadly. As a higher-level science product from missions and/or projects, these event lists or catalogs are important resources of the Heliophysics knowledge base (see Figs. 2 and 3) and should be archived with proper SPASE descriptions and shared and made searchable in a similar manner as typical science data. There are many community-compiled event lists or catalogs which are kept in various locations, online, offline, or in publications. Many mission teams have compiled event lists from mission observations as well. Due to the wide range of sources of event lists, it generally requires an extensive search to find all the event catalogs relevant to a given science phenomenon or topic, even for a subject expert.

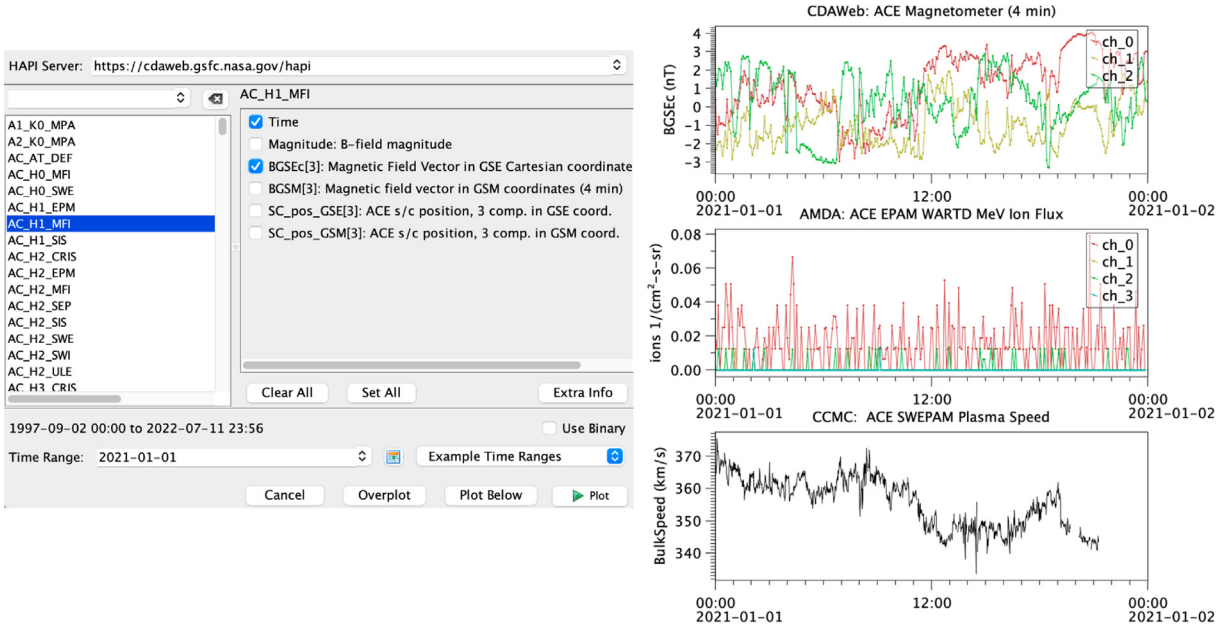

Fig. 14. Shown here on the left panel is the Autoplot data selection window for choosing data from a HAPI server from the NASA CDAWeb service. Each item in the list on the left column is a dataset, while the list on the right shows the parameters from that dataset to be plotted, as shown in the top plot on the right. The right panel are three stacked plots, each coming from a different HAPI data source: CDAWeb, AMDA, and CCMC. HAPI enables clients like Autoplot to display data from many different sources.

According to the NASA Heliophysics data policy (https://science.nasa.gov/science-red/s3fs-public/atoms/files/HPD_Data_Policy_Final_20220209_TAGGED.pdf) and NASA Research Opportunities in Space and Earth Science (ROSES) Data Management Plan (DMP) template of Heliophysics Division, the data products generated by NASA-supported work are required to be described in SPASE, made available to the public, and archived for long-term use. Many publishers of scientific journals also require data and event catalogs supporting a publication to be made available as a condition for publishing, although they do not generally archive the data of large volumes themselves. To ease the searchability and usability of these event lists, the SPASE Group has developed an event list standard format called Heliophysics Event List (HPEvent) (available at https://spase-group.org/docs/conventions/HP-Event-List-Specification.pdf). Some catalog examples following this standard are available at https://hpde.io/NASA/Catalog/index.html and have been given SPASE *ResourceIDs*. For instance, the CDPP tool, AMDA, enables export of event lists in HPEvent format. The catalogs will eventually be minted with DOIs so that they will become citable, enabling usage tracking. This limited collection of catalogs is just a start, and more work is needed to expand the collection by making SPASE descriptions of more event data and catalogs.

There are many Event Lists (a number in the current SPASE registry) that don't follow the HPEvent format. We can make SPASE registrations for the repositories of such event lists, e.g., the Heliophysics Events Knowledgebase (HEK, https://www.lmsal.com/hek/), and bring the list descriptions into the same SPASE database as the event lists that follow the HPEvent format. At the level of simply registering the Event Lists, the current SPASE Catalog Resource is sufficient, with AccessURLs that point to the archived lists of whatever sorts; current lists in HDP are highly varied in this respect. In the longer term, we would like to have tools that operate on Event Lists, aiding in finding overlaps or analyzing characteristics of different classes of events, but to make this possible across Catalogs will require at least a mapping from HPEvent to HEK. The HEK "VOEvent"-based scheme is more flexible than the HPEvent approach which, while very efficient for simple time series, does not have the generality of the VOEvent approach (Seaman et al., 2011). This is a problem beyond the primary use of SPASE for product registration and data search and access.

4.2.2. Gathering events for statistical studies

At present, some data services such as the HDP, the nascent HDO, SPEDAS (Angelopoulos et al., 2019; https://spedas.org/blog/) and the CDPP AMDA offer varied capabilities for searching and retrieving event and event list data. While it is quite feasible to collect the data of a good number of events of a given phenomenon to support statistical studies, it is by no means straightforward to do so routinely. Assembling data to support data science or machine-learning studies is particularly challenging, as we discuss in sections 4.1.3.2 and 5.2. At the CCMC, basic statistical analysis can also be performed with the Comprehensive Assessment of Models and Events based on Library tools (CAMEL, https://webserver1.ccmc.gsfc.nasa.gov/camel/). At COHOWeb (https://omniweb.gsfc.nasa.gov/coho/), one can get scatter plots, linear regression fits, medians, averages, standard deviations, and distribution functions for user-selected solar wind parameters from more than 10 missions covering the heliosphere. Describing these resources with SPASE metadata that includes caveat descriptions (see Table 1 and section 2.2.5) would help identify and acquire the data and determine what additional processing may be needed to prepare and use the data in a statistical analysis. This will also help data services, such as the CCMC and COHOWeb, with further development of analysis tools and user interface to provide more effective support for statistical studies.

Some additional statistical studies are done using event lists. A small number of data sources have been providing event search based on phenomena. For example, users can search active regions, CMEs, filaments, and other solar phenomena for any time interval through the NASA web application Helioviewer.org or the ESA advanced open source software JHelioviewer (https://www.jhelioviewer.org) because it includes the HEK. Some solar and solar wind events are also searchable through the Heliospheric Cataloguing, Analysis and Techniques Service (HELCATS, https://www.helcats-fp7.eu/index.html) and the Coordinated Data Analysis Workshop (CDAW) Data Center (https://cdaw.gsfc.nasa.gov/). The SOLAR VO also allows searching across the HEK and has the capability to search instrument data that overlap with the events from the HEK (https://solarnet.oma.be/web-client/hek_events). Using the SPASE metadata standard, including the HPEvent format, will simplify the aggregation of similar datasets from different data sources and event lists to support statistical studies. Different terminologies of the same phenomenon can sometimes hinder the search and use of datasets including event lists. Therefore, it is important to unify the terminologies and address them in the SPASE model. In addition, as event lists can vary largely based on different selection criteria, it is critical to list the criteria and include the references in the metadata.

4.2.3. Data discovery

As noted in section 4.1.3.2, extensions and rethinking of the portions of SPASE that support the analysis and communication (for instance, visualization) components of the data science taxonomy are a focus of this paper (see sections 4.2.1 and 4.2.2 above). The categories of SPASE allow data to be integrated by keywords and the dereferenceable link (non-ambiguous resource retrieval mechanism used across the internet) to the data for integration. And it also provides a means to reference a resource persistently via DOIs. Common metadata keywords allow disparate sources of data to be found via search and

subsequently used in an analysis together by users. An example of integrating data in this way would be searching for observations of particle precipitation and receiving data from numerous platforms such as the Defense Meteorological Satellite Program (DMSP) and Fast Auroral SnapshoT (FAST) Explorer (McGranaghan et al., 2021b). By searching by the keyword 'particle precipitation', a researcher might discover datasets that might otherwise not have known about, and thus expand the possibilities for scientific advances (e.g., Fig. 15). In this way, the rich and uniform descriptive capability of SPASE could help lessen the complexity of data wrangling and convergence needs for rigorous space physics scientific research across the landscape of heliophysics digital resources, which requires the use of many heterogeneous datasets to piece together a system-level understanding.

4.2.4. Data-model comparison

This section demonstrates the usefulness of metadata for a comprehensive, consistent and reproducible comparison of model solutions with observations in space weather research (Fig. 9). At present, data-model comparison in space weather modeling suffers from three distinct limitations. First, keeping up with the ever-growing number of models and different versions is challenging, including deciding which one should be used for a given scenario. Second, often there is no agreement in the scientific community on forecasting goals and metrics. Third, there is a slow iterative process between model developers and end users, which means that by the time a paper on a new model gets published, the model is most likely already out of date (MacNeice et al., 2018). In what follows, we explain how metadata documentation of models, data and metrics can assist with all three problems in the context of large-scale solar wind modeling.

The Ambient Solar Wind Validation Team embedded in the COSPAR International Space Weather Action Teams (ISWAT; https://www.iswat-cospar.org/) initiative was formed to develop an open platform for the community to more easily compare ambient solar wind models in the community, as conceptually represented by the red oval region in Fig. 9. The platform is actually integrated into the CAMEL tools (https://ccmc.gsfc.nasa.gov/tools/CAMEL/) framework of the CCMC and is publicly available online via NASA's CCMC (Rastätter et al., 2019). A critical backbone of this community effort is the metadata architecture to ensure rigorous documentation of the model settings and reproducibility of the solutions.

The team proposes eight different types of metadata that are necessary to register the models in their database. The metadata components include information on the input observational data, data preprocessing, model description, model setting, model output, model chain, model solution, and comparison with observed data (validation) and metrics. The information for all of these metadata components will be stored in an associated metadata file according to the SPASE metadata model (Fig. 2) and will also be registered in the SPASE metadata registry under the CCMC Naming Authority (see section 3.2).

Fig. 16 shows the key stages involved in an ambient solar wind model validation process and illustrates the information flows throughout the process. Fig. 16 represents an expanded view of Fig. 9 with Validation and Metrics corresponding to the red oval region where model performances are compared against observations. It is clear that the quality and sufficiency of descriptions of the input solar observational data and the models being employed are needed to perform the corresponding preprocessing of the observational data used as input to the model. Each composite model run consists of executing a chain of models to compute the final solution. In this example, the component models in the model chain must be adequately and compatibly described by metadata so that the output from the coronal model can be used as input to the heliospheric model for the combined model to produce the final model solution for comparison with ambient solar wind observations.

As a general guideline, metadata information for this context is defined based on two distinct model domains: the solar corona and the inner heliosphere. The models used in both of these domains need detailed descriptions (Fig. 5), model settings and input (Fig. 6), and model outputs (Fig. 7), as outlined in Fig. 4. It is also possible to register more than one model per domain if required. Running the entire model chain, for example, with a different set of parameters will result in a new model run. The information on how the models are linked to each other will be described in the model chain metadata file. Finally, the model solution metadata will describe the solution of the model chain, which the solution's metadata will reference, and which can then be compared to observed in-situ solar wind data. The validation and metrics metadata define the validation analysis and have been agreed upon by a community effort. All the metadata information of the participating solar wind models will be made available via CCMC's CAMEL web-based system (https://ccmc.gsfc.nasa.gov/tools/CAMEL/), which is currently under development.

We note that this metadata architecture fits many of the semi-empirical and magneto-hydrodynamical models calculating the ambient solar wind, but there are other models which will not fit, e.g., machine-learning models. Those models require their own specific metadata (section 4.1.3.2 and 4.2.3). More details about this validation project can be found in Reiss et al. (2022, this volume). The usage of metadata in the Ambient Solar Wind Validation Team exemplifies how metadata plays an essential supportive role both in this particular model-data comparison and in space weather research in general.

4.3. Data visualization

As a use case example of how SPASE information is used to support data visualization (section 2.2.5), we point

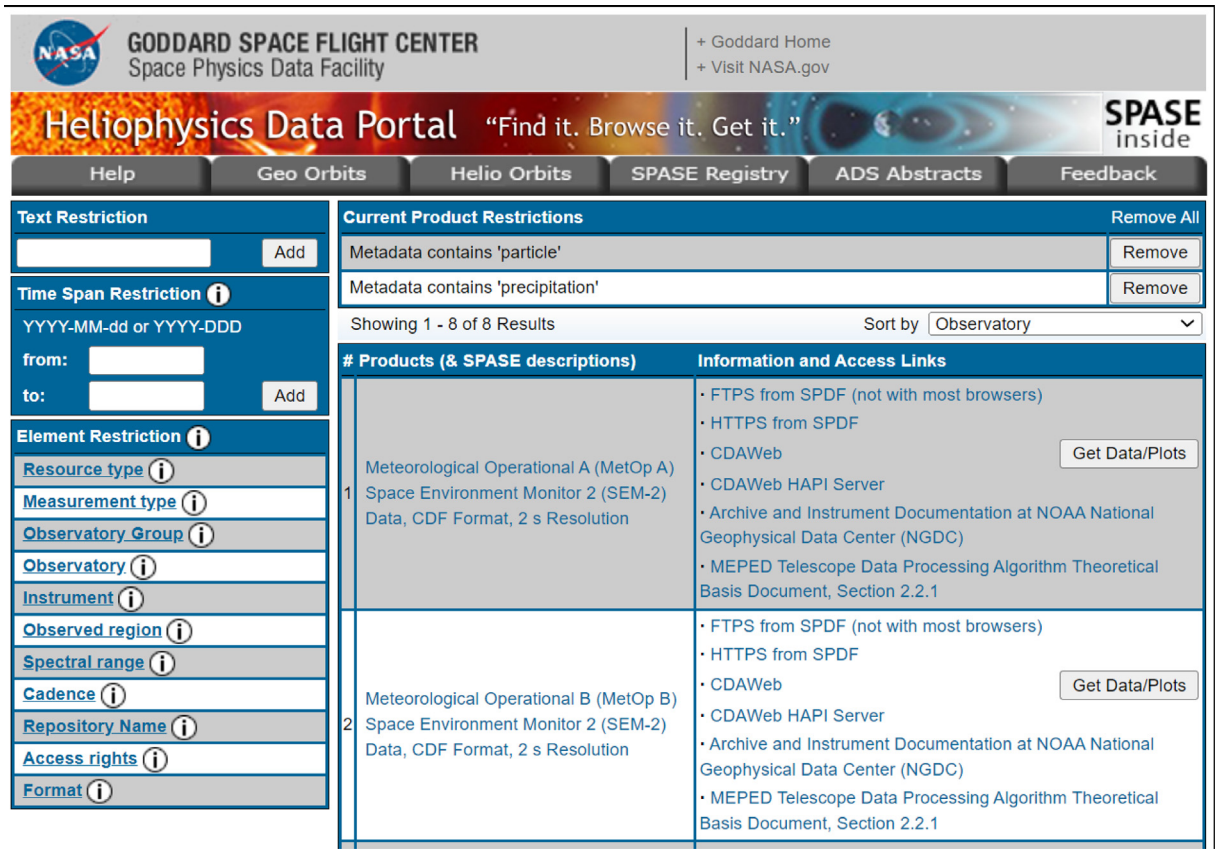

Fig. 15. A screenshot of the Heliophysics Data Portal webpage. By searching by the keywords 'particle' and 'precipitation', a researcher is presented with a list of matching datasets, which might include datasets that they might otherwise not have known about. This SPASE-enabled capability expands the possibilities for scientific advances.

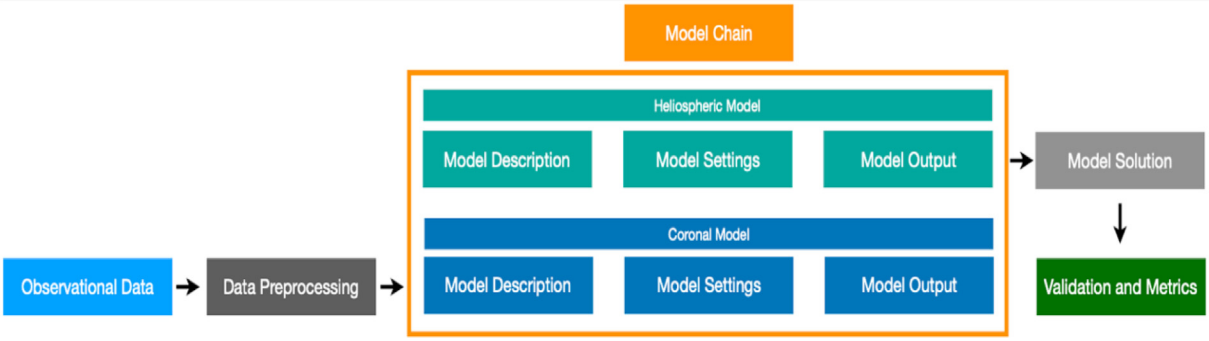

Fig. 16. Graphical representation of the metadata components for the ambient solar wind model validation project. This is an example of a common model framework where metadata for each of the boxes are collected. The orange boundary indicates the model chain which in this case consists of the coronal and the heliospheric domain.

out that SPASE parameter descriptions are used in the Automated Multi-Dataset Analysis tools (AMDA; https://amda.cdpp.eu from the CDPP, see Génot et al., 2021) to access the actual data in its Plot Manager. The SPASE XML files are stored in a registry and used to display information at each level in the Workspace Explorer of the tool (Fig. 17). SPASE metadata are also displayed as titles for axes on the plot (built from Name, Units, Structure, CoordinateSystemName elements of SPASE Parameter container). A user selects a physical quantity or a parameter in the Workspace Explorer that provides SPASE-based information (ResourceName, Description, Acknowledgement, Contact, InformationURL, TemporalDescription, Caveats, ...) on the corresponding mission (SPASE Observatory metadata), instrument (SPASE Instrument metadata), and data set (SPASE NumericalData metadata). The selected parameter is dragged and dropped to the Plot Manager in which it can be plotted over a selected time interval (Fig. 18) (also see discussions and Fig. 3 in Roberts et al. (2018)).

The SPASE Group (https://spase-group.org/connect.html) has also promoted the development of the HAPI with the intention of providing a single route to all heliophysics time series data with easily adopted methods for servers and clients. Very recently this protocol has been integrated in AMDA such that data are now distributed via HAPI thanks to the use of the official node.js (https://nodejs.org/en/knowledge/HTTP/servers/how-to-create-a-HTTP-server/) HAPI server and the implementation of a binding to the AMDA REST web services (https://amda.irap.omp.eu/service/hapi). This will enhance the visibility of AMDA and its datasets via a facilitated access, e.g., plotting AMDA datasets in Autoplot (https://autoplot.org/) is now straightforward (Fig. 19).

In general, software analysis and data visualization tools do not use SPASE metadata with perhaps the exception of *RenderingHints* (under Parameter description; Fig. 8) that describes the attributes to aid in the rendering the display of parameters, e.g. Autoplot in Java and pySPEDAS's PyTplot component in Python (https://pytplot.readthedocs.io/en/latest/index.html). However, the CCMC is planning to include SPASE metadata in Kamodo's plotting and display output in the future (Ringuette et al. 2023).

5. Future outlook

We have so far demonstrated SPASE as a science-enabling tool for several key examples in heliophysics. All of these capabilities are in various stages of development, with some not yet available such as searching for data by heliophysics phenomena, and others already demonstrating exciting applications such as the HDP and the HDO. Here we consider a few rapidly developing areas that may impact the international heliophysics data environment (Fig. 1) and the associated data flow (Fig. 9) in the coming years, and how the SPASE metadata can continue to play an important science-enabling role.

5.1. Open science

There is an increasing call for "open science" within the Heliophysics and most other science communities (Ramachandran et al., 2021; Gentemann et al., 2021). The idea is to make any research result easily accessible and reproducible by other researchers, leading to more collaborative, efficient, and even more rewarding results. This requires that, in addition to the reasoning steps captured in typical journal articles, any data underlying the arguments and any software needed to carry out the analysis must be reported as well. Here, the SPASE metadata and associated DOIs as described in sections 3.1-3.5 will be extremely useful. Specific datasets and registered software can be cited in papers by using the SPASE Software Resource (Fig. 2). DOIs can also be generated for collections of data and software needed for a paper (section 4.1.3.1). Authors will also be associated with the papers, and SPASE descriptions with ORCIDs (https://orcid.org/) will allow uniform referencing services. Uniform access methods, such as HAPI,

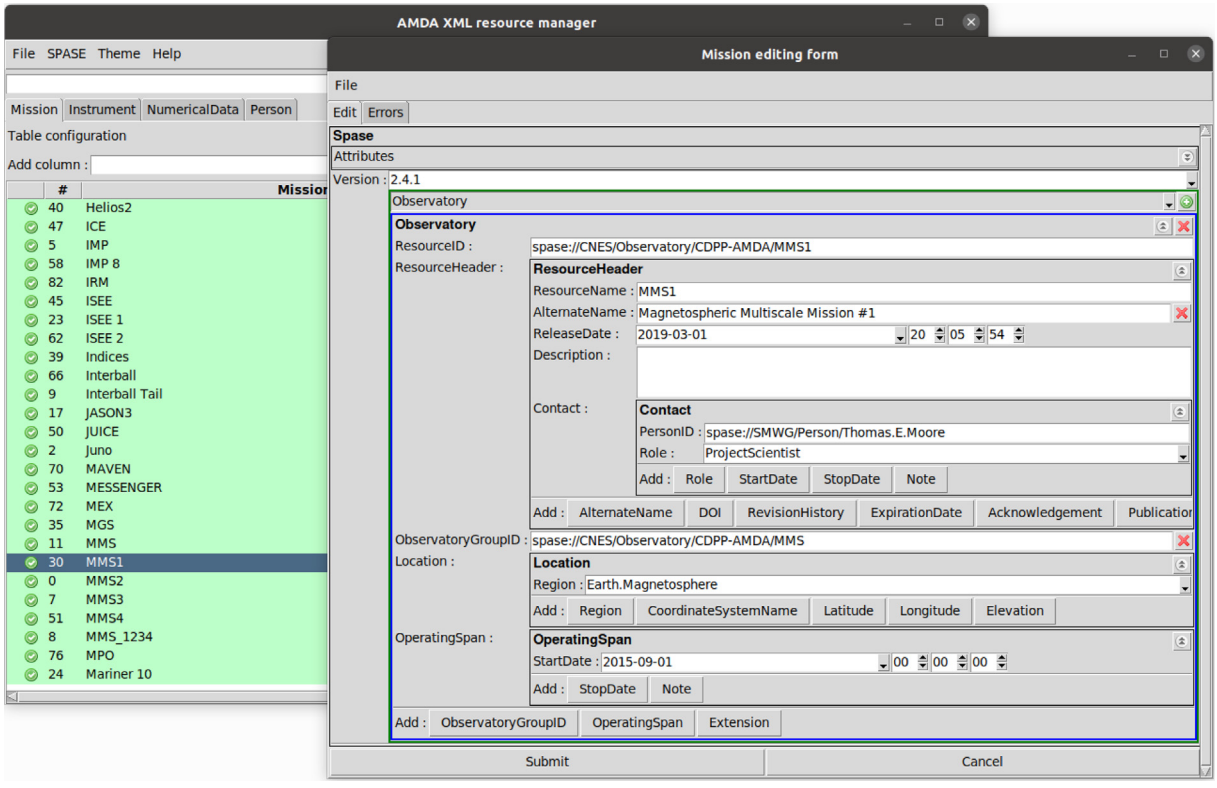

Fig. 17. AMDA XML resource manager and SPASE XML file editor.

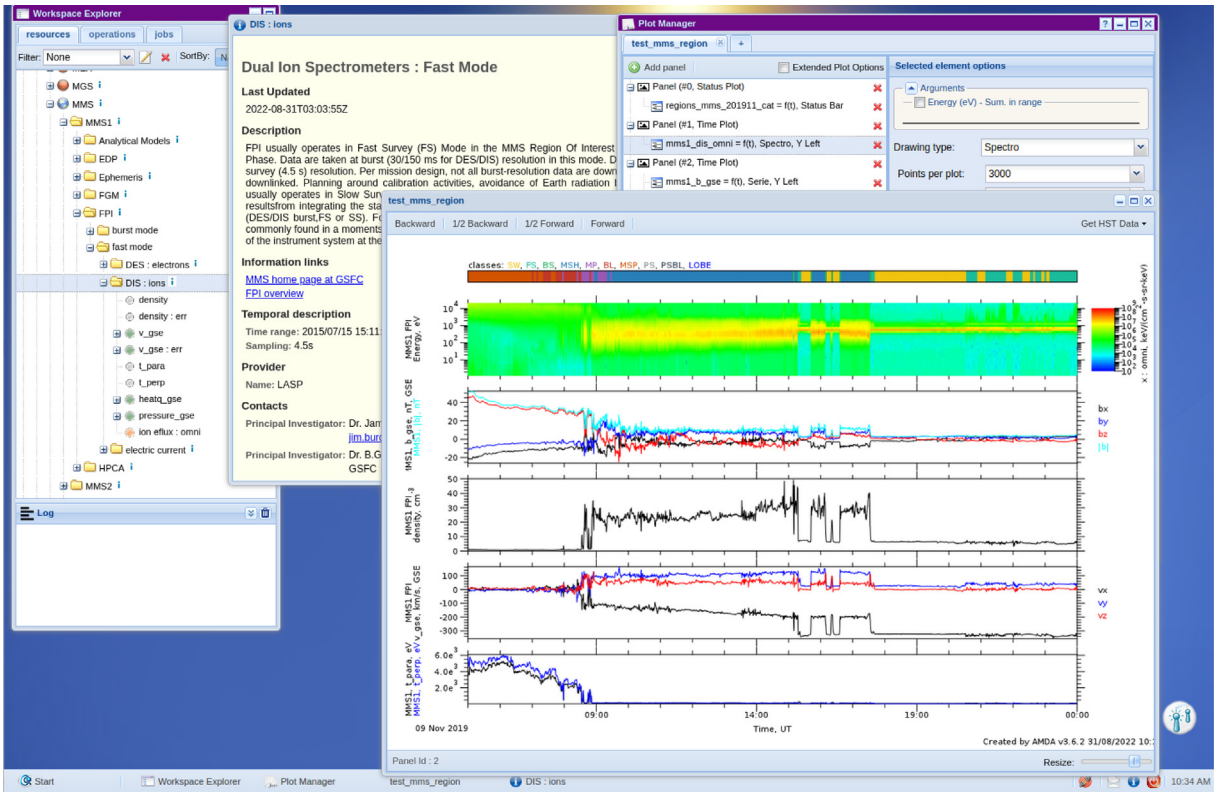

Fig. 18. Overview of AMDA Plot Manager showing SPASE metadata in yellow panellaces.

will make it easier to access the resources cited in papers. There is still a considerable amount to be worked out to make open science a reality that is not overly burdensome to researchers, but the simplification of searches based on uniform metadata will be an important part of the recipe for success.

As mentioned above, reproducibility of previously published research results might be straightforwardly enabled if the resources (data, analysis tools, models, software environment, etc.) used to produce the previous results could be made FAIR and citable, so that the same resources and analyses could simply be reproduced. While it is true that, in general, SPASE (section 3) does not always keep track of the versioning of the files of a given data product (e.g., calibration history) or other digital resource, it does have the optional provision for capturing resource revision history under *ResourceHeader* shown in Figs. 3 and 8 **(see section 3)**.

But resource versioning is a wider problem that cannot be resolved by SPASE alone. Most heliophysics missions and archives tend to serve the best calibrated data available. After reprocessing or updating, older versions of the same product are not always retained and maintained, even though sometimes they had been used in publications during the early phase of a mission. The data retention plan is often linked to the mission data management plan which must follow the funding agency-, or even PI-, dependent data policy. All these factors indeed impact reproducibility. Having recognized the importance of reproducibility, recent missions, such as Solar Orbiter, have started to tackle this issue by storing and providing access to all versions of their data. Version information will be critical for such comparisons, and needs to be kept in the files' metadata to eventually be reflected in SPASE. Enabling reproducibility of a given analysis is a complex issue involving numerous entities and processes beyond those describing or hosting datasets as focused on here, yet we defer full exploration of this issue to the community for a case-by-case discussion, including the process by which science in the Heliophysics community can be performed in the open from the beginning as required by open science. Such discussions are beyond the scope of this work.

5.2. Data science

Data science is becoming ubiquitous in our society (McGranaghan et al., 2017; Ramachandran et al., 2021; Masson et al., 2023). Data science analysis techniques, in particular machine learning, are increasingly being applied to different heliophysics research domains (Galvez et al., 2019; McGranaghan et al., 2018; 2021; Sadykov et al., 2021). As such, the heliophysics community faces both an exciting opportunity and an important imperative to explore a new frontier built at the intersection of traditional approaches and state-of-the-art data-driven sciences and technologies (McGranaghan et al., 2017). In this paper, we take data science to mean scalable architectural approaches, techniques, software, and algorithms that alter

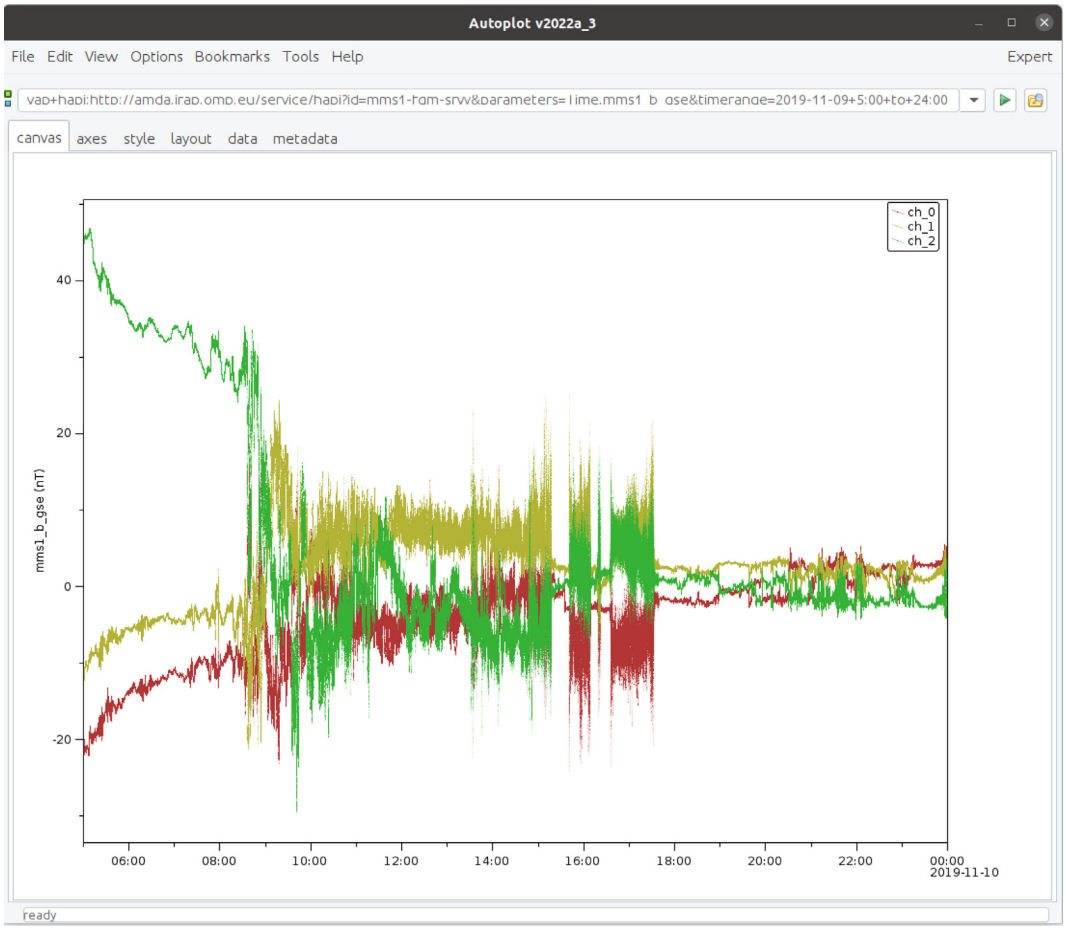

Fig. 19. AMDA datasets plotted in Autoplot through HAPI.

the paradigm by which data is assembled, managed, analyzed, and communicated. It is a combination of statistics, computer science, and domain knowledge.

AI/ML are a subset of data science, falling under the 'data analysis' portion of the full data lifecycle. Nevertheless, they have received widespread attention in Heliophysics, giving rise to workshops (e.g., Camporeale et al. (2018); McGranaghan et al., (2018a)), conferences (https://ml-helio.github.io/), journal special issues (e.g., McGranaghan et al., (2021a)), and an explosion in publications (see Camporeale et al. (2019) for a description of the landscape).

McGranaghan et al., (2021a) provided a taxonomy of data science topics covering the full spectrum of the data lifecycle: collection, management, analysis, and communication. Therefore, improving uniformity in the search and access of digital resources by improving the metadata dictionary and information models (sections 4.1.3.2 and 4.2.3) will make our data more readily consumable by machine-learning algorithms and interoperable across domains, broadening our science and building connections to other domains. As noted in section 4.1.3.2, however, extensions and rethinking of the SPASE schema may still be needed to support unique identification and thus discovery of ML-ready datasets. Engaging the community will be important for extending the SPASE schema and improving its effectiveness in servicing AI/ML data, as discussed in section 4.1.3.2. The categories that are important to add will come into greater focus once the community reaches a minimal agreement on how they shall be defined. Directly and indirectly, therefore, SPASE could help build a broader community by supporting data science methods. This will permit space physics to embrace trends toward open science discussed above.

A challenge is the current lack of consensus about what determines analysis readiness versus AI/ML-readiness. The concept of "analysis ready" data (Ramachandran et al., 2018) is indeed an active area of discussion in the AI/ML community, and has recently been expanded to include conversations about AI-readiness (see discussion across the Earth Sciences https://wiki.esipfed.org/Data_Readiness), analysis ready-cloud optimized (e.g., Abernathey et al., 2021), and numerous other perspectives on the topic. The Heliophysics community has maintained a living discussion of AI-readiness defined in the context of space physics (https://github.com/rmcgranaghan/data_science_tools_and_resources/wiki/Curated-Reference%7CChallenge-Data-Sets) that will be

an important guide for SPASE extensions to include this growing area in our field. While it is beyond the scope of this paper to formally define the SPASE extensions for describing AI/ML datasets, we endeavor to identify several characteristics common to AI/ML-ready data such that the incorporation of their descriptions in SPASE would enable those datasets to be discovered and reused more easily (note that these common characteristics have also emerged from the community discussion in the Github Wiki page linked above):

Data are calibrated, promoted in level, standardized, etc., so that values correspond well to the physical system being studied
Spurious data and non-physical values are either corrected or identified
Data are interpolated, patched, and harmonized, to provide even or consistent sampling and cadence
Data labels or categories (such as features and event indicators), if relevant, are made compatible with the feature set, i.e., labels "Y" made compatible for learning with feature set "X" as an input
Data is then made available in a format that is easy to read into an AI algorithm (e.g., Keras, PyTorch).

The sheer variety of AI/ML techniques requires numerous methods of processing data. Metadata, alone, does not enable AI/ML. However, descriptive data are vital to making an explicit and traceable data transformation pipeline that a data analysis or machine learning process uses and to making the resultant dataset findable. There are several elements of SPASE to make such datasets more findable, including but not limited to:

Resource attributes pertaining to how to access the resource, availability and storage format (such as AccessInformation and AccessURL as shown in Figs. 3 and 8);
Information on the execution platform for a resource, including operating system and necessary hardware (ExecutionEnvironment under Software in Infrastructure domain shown in Fig. 2);
Information about the level of processing at which a given resource is provided (ProcessingLevel as an optional attribute as shown in Fig. 8); and
Information that describes traceability of the data processing pipeline for a dataset (by using the RevisionHistory optional attribute under ResourceHeader as shown in Figs. 3 and 8)

Once a given dataset is discovered, it is expected that more bespoke elements of the data processing are provided by the researcher who created them and can be understood by other researchers. It is indeed important to capture these processing steps, however complex. This is the 'metadata' that a published article's methodology section provides. SPASE as it exists now supports this information, too, through elements such as InformationURL. The role of SPASE in this instance is to provide the metadata categories that help these datasets be found and thus supply a connection to them for the researcher. An example of this process is the 'DMSP Particle Precipitation AI-ready Data' (Galvez et al., 2019; McGranaghan et al., 2020). Enriching these published datasets with SPASE would make the datasets findable separately from the publication, and then the publication itself will provide the full description of the processing and AI/ML analysis.

5.3. High-level data search capability

The VxOs described in section 4.2 and developed during the first decade of 2000 provided valuable lessons for how SPASE metadata could enable users to search remotely and access data stored not only at central archives but also at distributed locations (Merka et al., 2008a, b; Fung, 2008, 2010). These exploratory efforts also provided a proving ground for the middleware architecture (Fig. 12) for virtual observatory operations, representing a paradigm shift from the way data was served primarily from a central archive. Since all the resources registered in a SPASE registry, irrespectively of their storage locations, are visible and accessible to a virtual observatory, there is a higher potential for a given data query to be met and more data to be discovered and returned (section 4.2.3). With the support of SPASE metadata, not only data services can be supported via conventional data queries of time, platform (spacecraft or ground station), and instrument, but science-based queries are also possible. For example, the nascent HDO (Fung et al., 2021) can also provide data searches queried by observing spacecraft location, as in the case of Fig. 13, and by magnetospheric state conditions (Fung et al., 2022).

We have described in sections 4.2.1 and 4.2.2 how metadata data can be searched and used to access resources for heliophysical event and statistical studies. Compiling datasets with multiple events pertaining to a given phenomenon is largely a manual and often laborious process. With community-wide adoption of SPASE metadata standards, particularly with robust descriptions of Annotation resources (Figs. 2 and 3) and applications of machine-learning techniques, it is hoped that querying data by phenomena can be supported in the not-too-distant future. To a lesser extent, however, support for data querying by phenomena and association with publications is an achievable goal in the next few years.

5.4. Software libraries

Publicly available software supporting science is also becoming more and more popular in Heliophysics. While proprietary code-based software like IDL SolarSoftWare (https://www.lmsal.com/solarsoft/; https://www.mssl.ucl.ac.uk/surf/sswdoc/) has been developed since the beginning of the SOHO mission era in the mid-1990 s, a community of developers and scientists supporting python packages for Heliophysics has been growing for more than a decade,

though not fully coordinated at the beginning. The Python in Heliophysics Community (PyHC; https://pyhc.org/) was set up in 2018 to promote and facilitate the use and development of Python in the Heliophysics community. Since then, PyHC has promoted a set of standards of development (https://github.com/heliophysicsPy/standards/blob/main/standards.md) that has been adopted by seven core Python libraries (https://heliopython.org/projects/): SunPy, PlasmaPy, pySPEDAS, HAPI client, Kamodo, SpacePy and pysat. These python packages cover the fields of solar physics (SunPy), solar wind and magnetospheric physics (pySPEDAS), ionosphere, thermosphere and atmospheric physics including ground-based experiments like SuperDARN (pysat), plasma physics (PlasmaPy), space science (SpacePy), the heliophysics API for time series (HAPI) and interpolation through data, especially simulation results (Kamodo), to simplify model-data comparisons.

Let us present here two examples of how these python libraries know where/how to obtain data, and what metadata is extracted from those libraries/routines. For the latter, a submodule of the SpacePy package called pycdf provides a Python interface to the Common Data Format (CDF) library and enables the extraction of all their global attributes and variable attributes (https://spacepy.github.io/pycdf.html), often in the ISTP/SPASE standard. The second example is related to the access of Solar Orbiter data at ESA. A SunPy plugin for accessing data in the Solar Orbiter Archive (SOAR) called sunpy-soar was developed to access the Solar Orbiter Archive (SOAR) at ESA using the unified Finding and Downloading SunPy Fido object (https://github.com/sunpy/sunpy-soar), again based on metadata.

In addition to the more conventional aspects of infrastructure already discussed, heliophysics infrastructure is increasingly dependent on such softwares for data utilization, analysis, and visualization. In some cases, a portion of these capabilities are made possible through an online platform, such as the Virtual Solar Observatory (VSO: https://sdac.virtualsolar.org/cgi/search), while other more flexible options are available as software packages. An increasingly useful collection of software packages in Python is available via the PyHC and on the Heliophysics System Observatory (HSO) Connect website (https://hso-connect.hpde.gsfc.nasa.gov). These and additional options in other languages are typically hosted on GitHub. Efforts are underway to streamline and connect these software resources together as a more interoperable web of tools in the interest of open science.

The PyHC website is now becoming a one-stop shop maintaining an up-to-date list of 60 + known Python packages available in the worldwide community. This website also provides a growing gallery of examples that are particularly useful for newcomers. PyHC has enabled the development of a community of developers working hand-in-hand under the PyHC umbrella. For more information on PyHC, its current status and future outlook, see Barnum et al. (2022) and the PyHC.org website. However, while various space weather services (including simulations) are being provided by various groups and organizations, not all of them are accessible via these libraries. There are also compatibility issues between these libraries where SPASE metadata could play a central role. SPASE is not yet used in any of these python packages but is clearly something on the PyHC to-do list especially as phenomenon keywords become available in SPASE.

Software registration is another area that needs attention. The open science initiative discussed above calls for open data and open software. While open data has become the *de facto* standard practice among the heliophysics research community in recent years, there remain outstanding technical and legal issues surrounding the practicality of open software in terms of licensing, proprietary and intellectual property rights, maintenance, citation, versioning, ethics, security, export control, etc. While the legal issues must be worked out by the source institutions, there is promise for solutions on the technical side through a forming collaboration, e.g., between PyHC and PyOpenSci. Apart from these issues, an open software community must enable users to discover what software is useful and how to access and use a particular piece of software for a particular research project. Software is a SPASE resource type (Fig. 2), so it is possible to develop a software products SPASE registry enabled with proper DOI referencing as discussed in section 3.5, which can then be served by an HDP- or HDO-like middleware (Fig. 12) to provide search and access functionality for software tools.

5.5. International collaborations and coordination

Interest in space weather transcends geopolitical boundaries due to its global impact on both space-based and ground-based infrastructure. Various international scientific organizations have organized programs with emphasis on international collaborations and coordination, such as the COSPAR Panel on Space Weather (https://cosparhq.cnes.fr/scientific-structure/panels/panel-on-space-weather-psw/), the COSPAR ISWAT initiative (https://www.iswat-cospar.org/iswat-cospar), the current SCOSTEP PRESTO Program (https://scostep.org/presto-science-program/), and the United Nations International Space Weather Initiative (ISWI) (https://www.unoosa.org/oosa/en/ourwork/psa/bssi/iswi.html). All of these programs have the goal to share data in order to promote greater utilization and scientific return of the data as well as to encourage collaborative research to benefit the global society. However, up till the late 2010 s, there was no international forum dedicated to the heliophysics information architecture at international level. The International Heliophysics Data Environment Alliance (IHDEA; https://ihdea.net/) was created in 2019 to be an alliance where major data providers and tools developers can discuss, coordinate and promote the use of the various standards and tools contributing to enhancing open science based on FAIR

principles. IHDEA is an open organization, having a yearly plenary session every autumn.

As noted in section 3.2, however, different data services have set up and maintained their own SPASE registries, making it less convenient or efficient for searching and accessing resources that are stored in distributed repositories. Thus, there needs to be an effective coordination of registries for sharing their SPASE metadata, so that they can be used collectively, as implied by Fig. 1. A central SPASE registry, where metadata in different repositories (i.e., under different NamingAuthorities) are shared, like the one maintained by the SPASE Group (https://github.com/hpde/), would make a more efficient environment for data exchange (section 4.2) and data discovery (section 4.2.3). It is hoped that through discussions in international forums such as IHDEA, COSPAR and other national and international organizations, agreements can be reached by the international heliophysics and space weather community to establish a uniform SPASE data dictionary (section 3.1) and a fully open SPASE registry (section 3.2), i.e., a lingua franca for metadata to support sharing of heliophysics and space weather resources across international boundaries.

6. Summary and conclusions

We have examined how metadata is needed to facilitate information flows throughout the heliophysics data environment (Fig. 1) and to support the operations of various information infrastructure architecture (section 2). The SPASE metadata model (section 3) was designed and developed specifically for describing heliophysics and space weather resources (Figs. 2-8), so it is particularly suited for providing uniform descriptions of those resources towards enabling their findability, accessibility, interoperability, and reusability, i.e., to satisfy the FAIR principles. While SPASE is perfectly adequate to support finding data in the traditional fashion, it can be further exploited to develop higher-level data search capabilities and support international collaborations.

Over the past several years many technologies needed for an open data infrastructure have been developed and deployed. SPASE is now an internationally adopted metadata standard (Appendix C) and recommended for adoption by the international space weather community (COSPAR Panel on Space Weather, 2021). Together with DOIs, standard data formats (CDF, FITS, HDF, netCDF), data archives with search portals (SPDF, CCMC, HDP, HDO, etc.), and data visualization and analysis tools (e.g., AMDA, Autoplot, JHelioviewer or PyHC python packages), the capabilities collectively enabled by SPASE can form the basis of an open science infrastructure. As expected for grassroots efforts, the development of each of the data infrastructure technologies occurred asynchronously, each addressing a recognized community need. Appendix C lists the international organizations and data services that have adopted the SPASE metadata for describing their digital resources. As Appendix C shows, however, adoption of the SPASE information model for metadata description remains limited.

With this paper, we have provided the international heliophysics and space weather research community with a utility perspective of SPASE so as to make the information model more understandable and accessible. We certainly hope that the SPASE information model will receive broad adoption by the international community through the promotion and endorsement by organizations like COSPAR Panel on Space Weather (section 5.5), and more digital resources will be described in SPASE, especially as the SPASE information model becomes more mature and stable. Section 3.4 and 4.1.3.1 have shown that the SPASE information model can also be extended in order to accommodate new description requirements. As pointed out in section 4.2.1, the NASA Heliophysics Division has already taken a concrete step in their latest science and data management policy (https://science.nasa.gov/science-red/s3fs-public/atoms/files/HPD_Data_Policy_Final_20220209_TAGGED.pdf) by requiring all NASA-sponsored datasets and data collections to be described by using the SPASE information model. From here, it will be of interest to consider how the SPASE metadata model may work with other metadata systems, such as the Planetary Data System information model (https://pds.nasa.gov/datastandards/documents/im/current/index_1J00.html) to support open science.

The future ahead appears bright. Each of the mentioned technologies complements each other in a growing data environment, especially one based on the FAIR principles. We assert then that a FAIR-compliant information architecture enabled by a standard metadata model, for which SPASE is a strong candidate, would result in a significantly improved and effective Heliophysics data environment. With enhanced end-to-end integration, it will be possible to create a citable data resource by observation or with software, along with a corresponding SPASE description. The SPASE description will be consumed by search portals that provide search engines to locate the desired resources. Visualization and analysis tools will be able to use the portals to locate resources relevant to a research topic, acquire the resources with standard API, perform analysis and display the resources.

Best of all the systems will grow organically through distributed, asynchronous actions. As described in section 3.5, a "genesis" event of the creation of a digital resource with a SPASE description will enable the generation of a landing page that describes the resource. This in turn will allow a DOI to be minted. Data portals and services will add the resource to their search indexes, and the data will be archived. Other services will send out notifications to interested parties. As the resources are used and reused to support various science tasks associated with Fig. 9 as described in this paper, attribution will be given through DOIs and community impact will be assessed. Most importantly, participation and contributions to the heliophysics

data environment will be open to all with a very low threshold. If one has a digital resource to share, one can share it with everyone and receive proper attribution.

Finally, we should note that while this paper is focused on the ongoing development and capabilities of the SPASE information model and its utility within the Heliophysics community, it also provides a means for interconnections with other domains. Focus of those cross-disciplinary collaborations is beyond the scope of the present paper, but they are ongoing within the Earth Science (Buttigieg et al. 2013, 2016; Raskin et al., 2005), geospatial sciences (Janowicz et al., 2022), and other NASA information services (Accomazzi et al., 2014) communities among myriad others. These efforts will help form the basis of discussions on future development and collaborations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

K.M. acknowledges support by the NASA HGI program (80HQTR19T0028), the NASA cooperative agreement NNG11PL10A and 80NSSC21M0180. M.A.R. acknowledges the Austrian Science Fund (FWF): P34437. HDRL including SPDF, SDAC, HDMC, and SPASE groups are supported by NASA. CCMC is a multi-agency partnership of NASA, United States Air Force, National Science Foundation, NOAA, and Office of Naval Research. CDPP Team acknowledges the support of Benjamin Renard, Alexandre Schulz and Myriam Bouchemit for implementing SPASE in AMDA.

The authors would like to express sincere gratitude to the Reviewers and Editor for their constructive comments and considerations during the extended review process that at the end helped make a much-improved paper.

Appendix A. Acronyms and websites

ADAPT, Active Data Archive Product Tracking.

AI, Artificial Intelligence.

AI-readiness discussion across the Earth Sciences (https://wiki.esipfed.org/Data_Readiness).

AI-readiness discussion in the context of space physics (https://github.com/rmcgranaghan/data_science_tools_and_resources/wiki/Curated-Reference%7CChallenge-Data-Sets).

AMDA, Automated Multi-Dataset Analysis (https://amda.cdpp.eu).

AMDA Representational State Transfer (REST) web services http://amda.irap.omp.eu/service/hapi.

Autoplot (http://autoplot.org/).

AUTUMN, Athabasca University THEMIS UCLA Magnetometer Network (https://autumn.athabascau.ca/).

CAMEL, Comprehensive Assessment of Models and Events Using Library Tools (https://ccmc.gsfc.nasa.gov/-camel/).

CCMC, Coordinated Community Modeling Center (https://ccmc.gsfc.nasa.gov/).

CCMC Metadata Registry (CMR) https://kauai.ccmc.gsfc.nasa.gov/CMR/view/metadata.

CDAWeb, Coordinated Data Analysis Web (https://cdaweb.gsfc.nasa.gov/).

CDAS, Coordinated Data Analysis System (https://cdaweb.gsfc.nasa.gov/WebServices/).

CDAS Python library (https://pypi.org/project/cdasws/).

CDF, Common Data Format (https://cdf.gsfc.nasa.-gov/).

CDPP, Centre de Données de la Physique des Plasmas (http://www.cdpp.eu).

CEF, Cluster Exchange Format (https://caa.esac.esa.int/documents/CAA-MDD-0001_v35.pdf).

CHS, The Nagoya University Center for Heliospheric Science (https://chs.isee.nagoya-u.ac.jp/en/about/).

CME, Coronal mass ejection.

CNES, Centre national d'études spatiales (https://cnes.fr/en).

CNSA, Chinese National Space Administration (http://www.cnsa.gov.cn/english/index.html).

Coordinated Data Analysis Workshop (CDAW) Data Center https://cdaw.gsfc.nasa.gov/.

COSPAR International Space Weather Action Teams (ISWAT) https://www.iswat-cospar.org/.

COSPAR Panel on Space Weather (https://iswat-cospar.org/psw; https://cosparhq.cnes.fr/scientific-structure/panels/panel-on-space-weather-psw/).

CrossRef (https://www.crossref.org/).

CSA, ESA Cluster Science Archive (https://csa.esac.esa.int).

DARTS, JAXA Data ARchives and Transmission System (https://darts.isas.jaxa.jp/).

DataCite, https://datacite.org/.

Dataset naming and file naming recommendations (https://spdf.gsfc.nasa.gov/guidelines/filenaming_recommendations.html).

DOI, Digital Object Identifier (https://www.doi.org/).

EPN-TAP, EuroPlaNet-Table Access Protocol (http://www.ivoa.net/documents/epntap).

ESA (European Space Agency) space weather service network (https://swe.ssa.esa.int/current-space-weather).

ESAC, European Space Astronomy Centre (https://www.esa.int/esac).

ESDC, ESAC Science Data Centre (https://www.cosmos.esa.int/esdc).

FITS, Flexible Image Transport System (https://fits.gsfc.nasa.gov/).

FTP, File Transfer Protocol.

Guidelines for Resource ID Formation.

(https://spase-group.org/docs/conventions/Resource-ID-Formation-Rule-v5.pdf).

HAPI, Heliophysics Applications Programming Interface (https://github.com/hapi-server/data-specification; https://Hapi-server.org; https://github.com/hapi-server/).

HAPI servers are available at http://hapi-server.org/servers.

HDMC, Heliophysics Data and Model Consortium (https://hpde.gsfc.nasa.gov/hpde_hdmc_projects.html).

HDP, Heliophysics Data Portal (https://heliophysics-data.gsfc.nasa.gov).

HDO, Heliophysics Digital Observatory (https://msqs.gsfc.nasa.gov/hdo/public).

HDRL, Heliophysics Digital Resources Library (see https://hpde.gsfc.nasa.gov/).

HDF, Hierarchical Data Format (https://www.hdf-group.org/).

HEK, Heliophysics Knowledge Base (https://www.lmsal.com/hek/api.html).

Heliophysics System Observatory (HSO) Connect (https://hsoconnect.hpde.gsfc.nasa.gov).

Heliospheric Cataloguing, Analysis and Techniques Service (HELCATS) https://www.helcats-fp7.eu/index.html.

Heliophysics catalogs https://hpde.io/NASA/Catalog/index.html.

Helioviewer https://Helioviewer.org.

HPEventList specification (https://spase-group.org/docs/conventions/HDMC-Event-List-Specification-v1.0.4.pdf).

IHDEA, International Heliophysics Data Environment Alliance (https://ihdea.net/).

IMPEx, Integrated Medium for Planetary Exploration, a project funded by the European Union under the Seventh Framework Programme (http://impex-fp7.oeaw.ac.at).

Intermagnet, International Real-time Magnetic Observatory Network (https://intermagnet.github.io/).

International DOI Foundation at https://doi.org/.

International Solar-Terrestrial Program (ISTP) metadata guidelines (https://spdf.gsfc.nasa.gov/istp_guide/istp_guide.html, also https://github.com/IHDEAlliance/ISTP_metadata/).

ISS-SOLAR, International Space Station SOLAR package (https://www.esa.int/Science_Exploration/Human_and_Robotic_Exploration/Research/SOLAR_three_years_observing_and_ready_for_solar_maximum).

ISTP Metadata Guidelines, HPDE Data File Internal Metadata Guidelines, https://spdf.gsfc.nasa.gov/sp_use_of_cdf.html.

IVOA, International Virtual Observatory Alliance (https://ivoa.net/).

IUGONET, Inter-university Upper atmosphere Global Observation NETwork (http://www.iugonet.org/index.jsp; http://search.iugonet.org/list.jsp).

JAXA, Japan Aerospace Exploration Agency (https://global.jaxa.jp/).

Magnetospheric Multiscale (MMS) science data center (https://lasp.colorado.edu/mms/sdc/public/search/).

JHelioviewer (https://www.jhelioviewer.org).

NamingAuthority for both space-based and ground-based resources, see https://spase-group.org/services/naming-authority.html.

NASA, National Aeronautics and Space Administration (https://www.nasa.gov/).

NASA Heliophysics Science Data Management Policy.

Version 1.2 (https://hpde.gsfc.nasa.gov/Heliophysics_Data_Policy_v1.2_2016Oct04.html).

Version 2.0 (https://science.nasa.gov/science-red/s3fs-public/atoms/files/HPD_Data_Policy_Final_20220209_TAGGED.pdf).

NASA Satellite Situation Center (SSC) Web services (https://sscweb.gsfc.nasa.gov/WebServices/).

netCDF, Network Common Data Form (https://www.unidata.ucar.edu/software/netcdf/).

NOAA, National Oceanic and Atmospheric Administration (https://www.noaa.gov/).

NSSDC, NASA Space Science Data Coordinated Archive (https://nssdc.gsfc.nasa.gov/).

OMNI Web and Datasets (https://omniweb.gsfc.nasa.gov/).

Open Madrigal Initiative (https://cedar.openmadrigal.org/openmadrigal).

ORCID, Open Researcher and Contributor Identifier (https://orcid.org/).

P2SA, ESA Proba-2 long term archive (http://p2sa.esac.esa.int/p2sa/).

PDS, Planetary Data System (https://pds.nasa.gov/).

Proba-2 science center (https://proba2.sidc.be/).

PyHC, Python in Heliophysics Community (https://heliopython.org/).

PyTplot in Python (https://pytplot.readthedocs.io/en/latest/index.html).

REST, REpresentational State Transfer (https://en.wikipedia.org/wiki/Representational_state_transfer).

ROR, CCMC Run-On-Request service (https://ccmc.gsfc.nasa.gov/requests/requests.php).

SCOSTEP PRESTO Program (https://scostep.org/presto-science-program/).

SDAC, NASA Solar Data Analysis Center (https://umbra.nascom.nasa.gov/).

SOAR, ESA Solar Orbiter ARchive (https://soar.esac.esa.int).

SOHO, Solar & Heliospheric Observatory (https://soho.nascom.nasa.gov/about/about.html).

Solar Orbiter ARchive (https://soar.esac.esa.int).

SolarSoft (https://sohowww.nascom.nasa.gov/solarsoft/).

SPASE dictionary (https://spase-group.org/data/model/search/index.html).

SPASE Group (https://spase-group.org/about.html; https://spase-group.org/connect.html).

SPASE registry on Github https://github.com/hpde/.

SPASE simulation extension data model (https://spase-group.org/data/simulation/spase-sim-1.0.0.pdf).

SPASE, Space Physics Archive Search and Extract (https://spase-group.org/).

SPASE web-based editors (http://xmleditor.spase-group.org/; http://xmleditor.spase-group.org/?simulation=true).

SPDF, NASA Space Physics Data Facility (https://spdf.gsfc.nasa.gov).

SPEDAS, Space Physics Environment Data Analysis System (https://spedas.org/blog/).

Standard filenaming templates, https://github.com/hapi-server/uri-templates/wiki/Specification.

SuperDARN, Super Dual Auroral Radar Network (https://superdarn.jhuapl.edu/).

SuperMAG (https://supermag.jhuapl.edu/).

TAP, IVOA Table Access Protocol (https://www.ivoa.net/documents/TAP/).

TFCat,Time-Frequency Catalog format (https://doi.org/10.25935/6068–8528; https://voparis-tap-maser.obspm.fr/browse/tfcat/q).

THEMIS, Time History of Events and Microscale Interactions During Substorms (http://themis.ssl.berkeley.edu/overview_data.shtml).

UDunits (https://www.unidata.ucar.edu/software/udunits/).

UFA, ESA Ulysses Final Archive, https://ufa.esac.esa.int/ufa/.

United Nations International Space Weather Initiative (ISWI) (https://www.unoosa.org/oosa/en/ourwork/psa/bssi/iswi.html; http://www.iswi-secretariat.org/).

Uniform Resource Identifier (URI) https://en.wikipedia.org/wiki/Uniform_Resource_Identifier.

VO, Virtual Observatory.

VOEvent format (https://www.ivoa.net/documents/VOEvent/).

VOTable format (https://www.ivoa.net/documents/VOTable/).

VSO, Virtual Solar Observatory (https://sdac.virtualsolar.org).

VSWMC, ESA Virtual Space Weather Modelling Centre (https://esa-vswmc.eu/, https://swe.ssa.esa.int/gen_mod).

VxOs, Heliophysics virtual observatories.

WCS, World Coordinate Systems https://fits.gsfc.nasa.gov/fits_wcs.html.

World Wide Web Consortium (W3C) https://www.w3.org/TR/prov-overview/.

XML, eXtensible Markup Language (https://www.w3.org/XML/).

Zenodo (https://zenodo.org).

Appendix B. Comparison between SPASE metadata model and ISTP Guidelines

Table B1 shows how ISTP Guidelines embedded global attribute metadata are used to populate SPASE NumericalData product descriptions. The first column in the table lists the ISTP global attribute/ISTP keyword name while the second column lists where the ISTP metadata maps to within the SPASE XML *NumericalData* schema. The third column denotes whether the ISTP global attribute text string may require any hand editing prior to populating the equivalent SPASE text field. If so, then "Yes" is listed. The rightmost column lists the methods used in order to transform the ISTP metadata prior to mapping into SPASE. The most common form of ISTP metadata transformation is automated via stream editing (e.g., via sed commands in BASH, see section 3.3). Otherwise, hand edits, lookup tables, or IDL programs are used as needed. The acronym SMWG appears five times in the last column. SMWG stands for SPASE Metadata Working Group and the SMWG is the name of the SPASE metadata repository currently used to store SPASE *Document*, *Instrument*, *Observatory*, *Person*, *Repository*, and *Service* resource descriptions. When SMWG appears, it denotes that SPASE resource descriptions stored in the SMWG metadata registry are utilized to populate the new SPASE description. For instance, the ADAPT IDL routines often harvest metadata content from SMWG SPASE *Observatory* or *Instrument* data resource descriptions to assign values to the *NumericalData ResourceID* or *InstrumentID* text fields when generating a new SPASE *NumericalData* resource description.

Table B2 shows how ISTP Guidelines embedded variable attribute metadata are used to populate SPASE *NumericalData Parameter* descriptions. The first column in the table lists the ISTP variable attribute/ISTP keyword name while the second column lists where the ISTP metadata maps to within the SPASE XML *Parameter* schema. The third column denotes whether the ISTP variable attribute text string may require any hand editing prior to populating the equivalent SPASE text field. If so, then "Yes" is listed. The rightmost column lists the methods used in order to transform the ISTP metadata prior to mapping into SPASE. The most common form of ISTP metadata transformation is automated via stream editing (e.g., via sed commands in BASH, see section 3.3). Otherwise, hand edits, lookup tables, or IDL programs are used as needed.

Appendix C. Current list of data systems that utilize SPASE compliant metadata

This appendix lists systems that use SPASE compliant metadata to enable search services, data discovery or SPASE registry services. These services can display a logo declaring "SPASE Inside", which can be found at https://spase-group.org/spase-inside.html. Parties interested in using SPASE or getting involved in the further development of standards and services to enable the open exchange of Heliophysics data can get involved by referring to links found on the "Connect With Us" webpage: https://spase-group.org/connect.html. The following data systems (in

alphabetical order) currently leverage SPASE to provide data to the Heliophysics community.

Australia

ASWS (https://www.sws.bom.gov.au/Geophysical).

Australian Space Weather Service (ASWS) provides information on a range of space weather products and services.

Canada

AUTUMN Virtual Magnetic Observatory (https://autumn.athabascau.ca/).

Provides access to data from the Athabasca University Geophysical Observatory (AUGO).

Canadian Space Science Data Portal (CSSDP) (https://asc-csa.gc.ca/eng/open-data/access-the-data.asp and https://www.spaceweather.gc.ca/data-donnee/sd-en.php).

Enables and simplifies researcher access to space science analytic tools and data.

ESA

ESA's Space Situational Awareness (SSA) Space Weather (SWE) Web Portal provides access to Space Weather (SWE) data using SPASE (https://swe.ssa.esa.int/swe-data-browsing) and programmatically through the HAPI and metadata described using SPASE (Panitzek, K., 2022). Adoption of SPASE by the ESAC Science Data Centre Heliophysics archives, is under consideration.

EU framework 7 projects

Near-Earth space data infrastructure for e-Science (ESPAS) (https://www.espas-fp7.eu/portal/).

An e-Infrastructure to support access to observations, modeling and prediction of the near-Earth space environment extending from the Earth's atmosphere up to the outer radiation belts.

Integrated Medium for Planetary Exploration (IMPEx) (http://impex-fp7.oeaw.ac.at/).

An integrated interactive computational framework where data from planetary missions are interconnected with numerical models.

France

Centre de Données de la Physique des Plasmas (CDPP) (http://www.cdpp.eu/).

Automated Multi-Dataset Analysis (AMDA) (http://amda.cdpp.eu/).

Provide integrated analysis of multi-point and multi-instrument data for case studies and statistical studies of plasmas in space physics.

Germany

German Research Centre for Geosciences (https://www.gfz-potsdam.de/en/home/).

Information System and Data Center for geoscientific data (http://isdc.gfz-potsdam.de/).

An access point for all manner of geoscientific geodata, its corresponding metadata, scientific documentation and software tools. The majority of the data and information, the portal currently offers to the public, are global geomonitoring products such as satellite orbit and Earth gravity field data as well as geomagnetic and atmospheric data for the exploration. Exploring Semantic web ontologies for use with reasoners and semantic searches.

Japan

IUGONET (https://www.iugonet.org/product/).

Provides a unified metadata database and seamless data environment for ground-based observations of the upper atmosphere acquired by a global network of radars, magnetometers, optical sensors, helioscopes, etc., and stored individual databases.

United States: NASA heliophysics Division

HPDE Repository (https://github.com/hpde/).

Provides comprehensive access to all registered metadata in NASA's Heliophysics Division.

HPDE Landing Page (https://hpde.io/).

Formatted information about each registered resource in the HPDE repositories.

Heliophysics Data Portal (http://heliophysicsdata.gsfc.nasa.gov/).

Provides a quick and easy way to find and access a comprehensive set of NASA and other datasets, images, movies, and associated services.

Heliophysics Digital Observatory (HDO) (https://msqs.gsfc.nasa.gov/hdo/public).

The goal of the NASA HDO is to provide a convenient portal for searching and accessing digital resources to support heliophysics and space weather research. To that end, the HDO focuses on providing higher-level search capabilities for meeting digital resource requirements stemming from heliophysics cross-disciplinary, system-science research tasks.

NSSDC SPASE Query (https://nssdc.gsfc.nasa.gov/spase/).

Search for select person, observatory and instrument descriptions.

References

Abernathey, R. P., T. Augspurger, A. Banihirwe, C. C. Blackmon-Luca, T. J. Crone, C. L. Gentemann, J. J. Hamman, N. Henderson, C. Lepore, T. A. McCaie, N. H. Robinson, and R. P. Signell (2021), "Cloud-Native Repositories for Big Scientific Data," in Computing in

Science & Engineering, vol. 23, no. 2, pp. 26–35, 1 March-April 2021. https://doi.org/10.1109/MCSE.2021.3059437.

Accomazzi, A., Gray, N., Erdmann, C., Biemesderfer, C., Frey, K., Soles, J. The Unified Astronomy Thesaurus. Astronomical Data Analysis Software and Systems XXIII. Proceedings of a meeting held 29 September – 3 October 2013 at Waikoloa Beach Marriott, Hawaii, USA. ASP Conference Series, vol. 485, 2014, p.461. 2014ASPC..485..461A.

Angelopoulos, V., Cruce, P., Drozdov, A., et al., 2019. The Space Physics Environment Data Analysis System (SPEDAS). Space Sci. Rev. 215, 9. https://doi.org/10.1007/s11214-018-0576-4.

Bargatze, L. F., Enhanced Interoperability Through Automated Data Archive Product Tracking, ADAPT, OF Space Weather Data Products, COSPAR 2018, Pasadena, CA, July 14-22, 2018.

Bargtze, L. F., R. M. Candey, and S. F. Fung (2022), Enhanced ADAPT support of the heliophysics data environment, presented at the 44th COSPAR Scientific Assembly, Athens, Greece, July 17-23.

Barnum, J. A. Masson, R.J.W. Friedel, A. Roberts, B.A. Thomas, (2022) Python in Heliophysics Community (PyHC): current status and future outlook, Adv. Space Res., in press, 2022. https://doi.org/10.1016/j.asr.2022.10.006.

Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C.J., Lewis, S.E., 2013. The environment ontology: contextualising biological and biomedical entities. J. Biomed. Semant. 4 (1), 43. https://doi.org/10.1186/2041-1480-4-43.

Buttigieg, P.L., Pafilis, E., Lewis, S.E., Schildhauer, M.P., Walls, R.L., Mungall, C.J., 2016. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. J Biomed. Semant. 7 (1), 57.

Camporeale, E., 2019. The challenge of machine learning in Space Weather: Nowcasting and forecasting. Space Weather 17, 1166–1207. https://doi.org/10.1029/2018SW002061.

Camporeale, E., Wing, S., Johnson, J., Jackman, C.M., McGranaghan, R., 2018. Space weather in the machine learning era: A multidisciplinary approach. Space Weather 16, 2–4. https://doi.org/10.1002/2017SW001775.

Cecconi, Baptiste, Erard, Stéphane, André, Nicolas, Jacquey, Christian, Génot, Vincent, Henry, Florence, Bonnin, Xavier, Le Sidaner, Pierre, Chauvin, Cyril, Fuller, Nicolas, Braga, V F, Aboudarham, Jean, Louys, Mireille, Derrière, Sébastien, & Preite-Martinez, Andrea. (2014). Solar System UCDs: Assessment Study of Unified Content Descriptors (UCDs) for the Solar System Resources (Planetary sciences and Heliophysics) (0.6). Zenodo, doi: 10.5281/zenodo.3479165.

Cooper, J. F., T. P. Armstrong, M. E. Hill, N. Lal, R. E. McGuire, R. B. McKibben, T. W. Narock, A. Szabo, and C. Tranquille. 2007. "Virtual Energetic Particle Observatory for the Heliospheric Data Environment." *AGU Fall Meeting Abstracts*, A251-:

COSPAR Panel on Space Weather, Resolutions on Metadata Standards and Data Access, *Space Research Today*, Volume 212, December 2021, Page 19. https://doi.org/10.1016/j.srt.2021.11.014.

Erard , S., Cecconi , B., Le Sidaner , P., Berthier , Henry , J., F., Molinaro , M., Giardino , M., Bourrel , N., André , N., Gangloff , M., Jacquey , and C., Topf , F. (2014). The EPN-TAP protocol for the Planetary Science Virtual Observatory, Astronomy and Computing, Volumes 7–8, Pages 52-61. doi: 10.1016/j.ascom.2014.07.008.

Faden, J.B., Weigel, R.S., Merka, J., et al., 2010. Autoplot: a browser for scientific data on the web. Earth Sci. Inform. 3, 41–49. https://doi.org/10.1007/s12145-010-0049-0.

Fung, S.F., March 2010. The Virtual Wave Observatory (VWO): A portal to heliophysics wave data. Radio Sci. Bull. 332, 89–102.

Fung, Shing F., C. F. Dolan, and L. N. Garcia (2021), The Heliophysics Digital Observatory (HDO): Providing Enhanced High-Level Data Search Capabilities, presented at the fall AGU meeting, December 13-17. (https://agu.confex.com/agu/fm21/meetingapp.cgi/Paper/810730).

Fung, Shing F., C. F. Dolan, and L. N. Garcia (2022), The Nascent Heliophysics Digital Observatory (HDO): a Follow-on to the Heliophysics VxOs, presented at the 44th COSPAR Scientific Assembly, PSW.4-0003-22, July 16-24, 2022. (https://app.cospar-assembly.org/2022/browser/presentation/29100).

Fung, S. F., The Virtual Wave Observatory (VWO), presented at the Fall AGU meeting, San Francisco, CA, December 15-19, 2008.

Galvez, R., D. F. Fouhey, M. Jin, et al. (2019), A Machine-learning Data Set Prepared from the NASA Solar Dynamics Observatory Mission, Astrophys. J. Supplement Series, 242:7 (11pp), https://doi.org/10.3847/1538-4365/ab1005.

Génot, V., M. L. Khodachenko, E. J. Kallio, et al. (2012) Capabilities of the analysis tools of the IMPEx infrastructure, presented at the European Planetary Science Congress, EPSC Abstracts Vol. 7 EPSC2012-152.

Génot, V., E. Budnik, C. Jacquey, et al., (2021) Automated Multi-Dataset Analysis (AMDA): An on-line database and analysis tool for heliospheric and planetary plasma data, Planetary and Space Science, Volume 201, article id. 105214, doi: 10.1016/j.pss.2021.105214.

Gentemann, C. L., Holdgraf, C., Abernathey, R., Crichton, D., Colliander, J., Kearns, E. J., et al. (2021). Science storms the cloud. *AGU Advances*, 2, e2020AV000354, doi: 10.1029/2020AV000354.

Hess, S. L. G., M. L. Khodachenko, E. J. Kallio, et al. (2012a). IMPEx Data Model: a simulation extension to the Spase data model, presented at the European Planetary Science Congress, EPSC Abstracts, Vol. 7 EPSC2012-360. r.

Hess, S. L. G., Khodachenko, M., Kallio, E. J., Genot, V. N., Gangloff, M., Jarvinen, R., Hakkinen, L. V., Topf, F., Al-ubaidi, T., Schmidt, W., Modolo, R., Alexeev, I. (2012b) "IMPEx Simulation Data Model: an extension to SPASE for the description of simulation runs," Poster at the fall AGU meeting, San Francisco, USA, December 2012.

Hill, F., Martens, P., Yoshimura, K., et al., 2009. The virtual solar observatory—A resource for International Heliophysics Research. Earth Moon Planet. 104, 315–330. https://doi.org/10.1007/s11038-008-9274-7.

Janowicz, K., Hitzler, P., Li, W., Rehberger, D., Schildhauer, M., Zhu, R., Shimizu, C., Fisher, C.K., Cai, L., Mai, G., Zalewski, J., Zhou, L., Stephen, S., Gonzalez, S., Mecum, B., Lopez-Carr, A., Schroeder, A., Smith, D., Wright, D., Wang, S., Tian, Y., Liu, Z., Shi, M., D'Onofrio, A., Gu, Z., Currier, K., 2022. Know, KnowWhere, KnowWhere-Graph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence. ". AI Mag. 43, 30–39. https://doi.org/10.1002/aaai.12043.

Khodachenko, M. L., Génot, V., Kallio, E., Alexeev, I., Modolo, R., Al-Ubaidi, T., André, N., Gangloff, M., Schmidt, W., Belenkaya, E., Topf, F., and R. Stoeckler (2011), Integrated Medium for Planetary Exploration (IMPEx): a new EU FP7-SPACE project, EPSC Abstracts, Vol. 6, EPSC-DPS Joint Meeting.

King, J.H., Papitashvili, N.E., 2005. Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data. J. Geophys. Res. 110, A02104. https://doi.org/10.1029/2004JA010649.

MacNeice, P., Jian, L.K., Antiochos, S.K., Arge, C.N., Bussy-Virat, C.D., DeRosa, M.L., et al., 2018. Assessing the quality of models of the ambient solarwind. Space Weather 16, 1644–1667. https://doi.org/10.1029/2018SW002040.

Masson, A., G. De Marchi, B. Merin, M.H. Sarmiento, D.L. Wenzel, B. Martinez, (2021), Google dataset search and DOI for data in the ESA space science archives, Advances in Space Research, 67, 8, 2504-2516. https://doi.org/10.1016/j.asr.2021.01.035.

Masson, A., Fung, S.F., Camporeale, E., et al., 2023. Heliophysics and space weather information architecture and innovative solutions: ways forward. Adv. Space Res.

McGranaghan, R.M., Bhatt, A., Matsuo, T., Mannucci, A.J., Semeter, J. L., Datta-Barua, S., 2017. Ushering in a new frontier in geospace through data science. J. Geophys. Res. Space Phys. 122, 12586–12590. https://doi.org/10.1002/2017JA024835.

McGranaghan, R., Borovsky, J. E., and Denton, M. (2018a), How do we accomplish system science in space?, Eos, 99, October 15, 2018. https://doi.org/10.1029/2018EO107411.

McGranaghan, R. M., Bloch, T., Ziegler, J., Hatch, S., Camporeale, E., Owens, M., Lynch, K., Gjerloev, J., Zhang, B., and S. Skone. (2020).

DMSP Particle Precipitation AI-ready Data (1.0.0-alpha) . Zenodo. https://doi.org/10.5281/zenodo.4281122.

McGranaghan, R. M., Ziegler, J., Bloch, T., Hatch, S., Camporeale, E., Lynch, K., et al. (2021b). Toward a next generation particle precipitation model: Mesoscale prediction through machine learning (a case study and framework for progress). Space Weather, 19, e2020SW002684. https://doi.org/10.1029/2020SW0.

McGranaghan, R.M., Mannucci, A.J., Wilson, B.D., Mattmann, C.A., Chadwick, R., 2018b. New capabilities for prediction of high-latitude ionospheric scintillation: A novel approach with machine learning. Space Weather 16, 1817–1846. https://doi.org/10.1029/2018SW002018.

McGranaghan, R.M., Camporeale, E., Georgoulis, M., Anastasiadis, A., 2021a. Space Weather research in the Digital Age and across the full data lifecycle: Introduction to the Topical Issue. J. Space Weather Space Clim. 11. https://doi.org/10.1051/swsc/2021037.

Meerkat, J. (2006), The Virtual Magnetospheric Observatory VMO, *The Evolving Heliophysics Data Environment: VxO Kickoff Meeting*, Baltimore, Maryland, May 22, 2006. (https://hpde.gsfc.nasa.gov/VMO_UMBC.pdf).

Merka, J., A. Szabo, R. Walker, T. Narock, and T. King (2008a) Uniform data discovery and access with the Virtual Heliospheric and Magnetospheric Observatories, presented at the EGU General Assembly, Geophysical Research Abstracts, Vol. 10, EGU2008-A-10989, 2008. (https://meetings.copernicus.org/www.cosis.net/abstracts/EGU2008/10989/EGU2008-A-10989.pdf).

Merka, J., Narock, T.W., Szabo, A., 2008b. Navigating through SPASE to heliospheric and magnetospheric data. Earth Sci. Inform. 1, 35–42. https://doi.org/10.1007/s12145-008-0004-5.

Merka, J. (2006) The Virtual Magnetospheric Observatory (VMO), presented at the VxO Kickoff Meeting on "The Evolving Heliophysics Data Environment", May 22, 2006. (https://hpde.gsfc.nasa.gov/VMO_UMBC.pdf).

Modolo, R., Hess, S., Genot, V., Leclercq, L., Leblanc, F., Chaufray, J.-Y., Weill, P., Gangloff, M., Fedorov, A., Budnik, E., Bouchemit, M., Steckiewicz, M., André, N., Beigbeder, L., Popescu, D., Toniutti, J.-P., Al-Ubaidi, T., Khodachenko, M., Brain, D., Curry, S., Jakosky, B., Holmström, M., 2018. The LatHyS database for planetary plasma environment investigations: Overview and a case study of data/model comparisons. Planet. Space Sci. 150, 13–21. https://doi.org/10.1016/j.pss.2017.02.015.

Mumford, S. J., Freij, N., Christe, S., et al., 2021. SunPy (v3.0.3). Zenodo. https://doi.org/10.5281/zenodo.5751998.

Panitzek, K., Programmatic access to SWE data within the SSA SWE network using HAPI, ESA technical note, SSA-SWE-HAPI-TN-0001, 2022; https://swe.ssa.esa.int/documents/20182/25484/SSA-SWE-HAPI-TN-0001.pdf/55860c01-a728-4510-9741-c8ccf57ac78f.

Paschmann, G., Melzner, F., Frenzel, R., et al., 1997. The electron drift instrument for Cluster. Space Sci. Rev. 79, 233–269. https://doi.org/10.1023/A:1004917512774.

Piker, C., Granroth, L., Mukherjee, J., Pisa, D., Cecconi, B., Kopf, A. and Faden, J. (2018). Lightweight Federated Data Networks with Das2 Tools. AGU Fall Meeting 2018 posters, Washington DC, USA. https://doi.org/10.1002/essoar.10500359.1Ramachandran, R., Bugbee, K., & Murphy, K. (2021). From open data to open science. Earth and Space Science, 8, e2020EA001562. https://doi.org/10.1029/2020EA001562.

Poedts, S., Lani, A., Scolini, C., et al., 2020. EUropean Heliospheric FORecasting Information Asset 2.0. J. Space Weather Space Clim. 10 57. https://doi.org/10.1051/swsc/2020055.

Preite Martinez, A., M. Louys, B. Cecconi, S. Derriere, F. Ochsenbein, & IVOA Semantic Working Group (2018). The UCD1+ controlled vocabulary Version 1.3 Version 1.3. ivoa.spec, 527. https://doi.org/10.5479/ADS/bib/2018ivoa.spec.0527M.

Ramachandran, R., Lynnes, C., Bingham, A. W., and Quam, B. M. (2018). *Enabling Analytics in the Cloud for Earth Science Data*. In Proceedings of "Workshop on Enabling Analytics in the Cloud for Earth Science Data." February 2018. NTRS - NASA Technical Reports Server Document ID: 20180002954 and Report number: MSFC-E-DAA-TN55638.

Ramachandran, R., Bugbee, K., and Murphy, K. (2021). From open data to open science. *Earth and Space Science*, 8, e2020EA001562. doi: 10.1029/2020EA001562.

Raskin, R.G., Pan, M.J., 2005. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). Comput. Geosci. 31 (9), 1119–1125. https://doi.org/10.1016/j.cageo.2004.12.004.

Rastätter, L., Wiegand, C., Mullinix, R.E., MacNeice, P.J., 2019. Comprehensive Assessment of Models and Events Using LibraryTools (CAMEL) framework: Time series comparisons. Space Weather 17, 845–860. https://doi.org/10.1029/2018SW002043.

Reiss, M.A., Muglach, K., Mullinix, R., et al., 2022. Unifying the validation of ambient solar wind models. Adv. Space Res. https://doi.org/10.1016/j.asr.2022.05.026.

Ringuette, R., Rastaetter, L., De Zeeuw, D., Pembroke, A., 2023. Kamodo: Simplifying model data access and utilization. Adv. Space Res. https://doi.org/10.1016/j.asr.2023.03.033.

Roberts, D.A., Thieman, J., Génot, V., King, T., Gangloff, M., Perry, C., Wiegand, C., De Zeeuw, D., Fung, S.F., Cecconi, B., Hess, S., 2018. The SPASE data model: A metadata standard for registering, finding, accessing, and using Heliophysics data obtained from observations and modeling. Space Weather 16. https://doi.org/10.1029/2018SW002038.

Sadykov, V., A. Kosovichev, I. Kitiashvili, et al (2021), "Prediction of Solar Proton Events with Machine Learning: Comparison with Operational Forecasts and "All-Clear" Perspectives", https://arxiv.org/abs/2107.03911, 2021.

Seaman, R., Williams, R., Allan, A., Barthelmy, S., Bloom, J. S., Brewer, J. M., Denny, R. B., Fitzpatrick, M., Graham, M., Gray, N., Hessman, F., Marka, S., Rots, A., Vestrand, T. and Wozniak, P. (2011). Sky Event Reporting Metadata (VOEvent) Version 2.0. IVOA Specification. https://www.ivoa.net/documents/VOEvent/.

SPASE Group (2014). SPASE Simulation Extensions for the Space Physics Archive Search and Extract (SPASE) Data Model SPASE Group. Version 1.0.0. https://doi.org/10.48322/TXCA-X050. Accessed on 2022-January-26.

SPASE Group (2021). Space Physics Archive Search and Extract (SPASE) Base Information Model. SPASE Group. Version 2.4.0. https://doi.org/10.48322/E72C-5Y75. Accessed on 2022-January-27.

Szabo, A., T. Narock, J. Merka, A. Roberts, J. Vandegriff, G. Ho, J. Raines, P. Schroeder, A. Davis, and J. Kasper. 2007. "The Virtual Heliospheric Observatory (VHO)." *AGU Fall Meeting Abstract*, SH51A-0250.

Torbert, R.B., Vaith, H., Granoff, M., et al., 2016. The electron drift instrument for MMS. Space Sci. Rev. 199, 283–305. https://doi.org/10.1007/s11214-015-0182-7.

Walker, R. J., T. A. King, S. P. Joy, L. F. Bargatze, P. Chi, J. Weygand. 2007. "The Virtual Magnetospheric Observatory at UCLA". *AGU Fall Meeting Abstract id.SH51A-0246.*

Weigel, R. S., Vandegriff, J., Faden, J., King, T., Roberts, D. A., Harris, B., et al. (2021a). HAPI: An API standard for accessing Heliophysics time series data. Journal of Geophysical Research: Space Physics, 126, e2021JA029534. https://doi.org/10.1029/2021JA029534.

Weigel, R. S., Vandegriff, J., Faden, J., Roberts, D. A., King, T., Candey, R., and Harris, B. (2021b). The Heliophysics Application Programmer's Interface Specification 3.0.0. Zenodo. https://doi.org/10.5281/zenodo.4757597.

Wilkinson, M., Dumontier, M., Aalbersberg, I., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3. https://doi.org/10.1038/sdata.2016.18 160018.