HORIZON Research and Innovation Action
HORIZON-CL4-2022-SPACE-01-82
Strategic Autonomy in Developing, Deploying and Using Global Space-based Infrastructures, Services, Applications and Data 2022

101082633

D 1.2

Data Management Plan (DMP)

WP1: Management

Date of preparation (latest version): Jan 30, 2024

DOCUMENT INFORMATION

Deliverable Number	D1.2
Deliverable Name	Data Management Plan (DMP)
Due Date	31/03/2024
Deliverable Lead	KULeuven
Authors	Giovanni Lapenta
Responsible Author	Giovanni Lapenta, Giovanni.lapenta@kuleuven.be
Keywords	DMP
WP/Task	WP 1 /Task 1
Nature	[R]
Dissemination Level	PU
Final Version Date	DAY/MONTH/YEAR
Reviewed by	[NAME, AFFILIATION]

DOCUMENT HISTORY

Partner	Date	Comment	Version
KULeuven	30/01/2024	First draft	0.1

Executive Summary

The policies planned for the management of data in ASAP are outlined following the guidelines of Horizon Europe.

TABLE OF CONTENTS

1. Data summary

ASAP will generate tools to process existing publicly accessible data and publications available to the heliospheric scientific research community. ASAP exploits Artificial Intelligence algorithms (Machine Learning, ML, and Artificial Intelligence, AI), Statistical Analysis, Information theory and focuses on porting them to processors that can function in space. The idea is to develop a new approach that if we succeed can then later be deployed in situ tools of AI and ML in robotic and automatic space missions. The long-term goal is to open a new path in the scientific community to increase the efficiency and capacity of large-scale data analysis.
ASAP will be available open source and will include:

- advanced techniques in Artificial Intelligence (AI) and Machine Learning (ML)
- advanced statistical tools for heliospheric data analysis
- the ability to use past mission data and synthetic data from simulation, using only data that is publicly available

One of the most challenging and time-consuming efforts in AI is the "training" of models. Many of the AI methods developed in ASAP will themselves represent higher-level data products. For instance, trained neural networks will be stored and reused as a database of coefficients. Additionally, the ASAP network will apply all methods described above to a number of high profiles heliospheric problems ranging from the process of plasma turbulence and magnetic reconnection, the solar wind and its coupling with the Earth outer and inner space environments. In table 1, we show the data available on ASAP.

Table 2 shows a preliminary list of data sources used in the project.

Table 1: ASAP data.

Data produced by Simulations of space physics	***Additional enhanced data from space missions***	***AI models and configurations data***
1D, 2D, 3D arrays of binary data. Examples include direct binary format, HDF5 and CDF. Starting from M12 the project will produce continuously additional data. Several additional simulations will be added per year.	1D, 2D, 3D arrays of binary data. Examples include CDF. Starting from M18 with a continuous addition of data. High level data from European space missions by M24. This would include lists of event of scientific interest (e.g., CMEs, magnetopause crossings)	1D, 2D, 3D arrays of binary data. Examples include HDF5 and CDF. Starting from M18 with a continuous addition of data. A few additional AI models, fully trained, per year. AI models (Machine and Deep learning) hyper parameters can be stored in PMML (Predictive Model Markup Language) standard and in Python dictionaries. Deep Learning models can be also saved in the ONNX format that allows to switch models between DL frameworks

Table 2: ASAP data sources.

Mission	***Description***
MMS	Multi-spacecraft studies of the microphysics of MR, Turb, KHI; wave-particle interactions, particle heating and acceleration
THEMIS	Multi-spacecraft studies of magnetic sub-storms, particle acceleration, MR, Turb
CLUSTER	Multi-spacecraft studies of MS regions (e.g., polar cusp, MP, MT), bow shock, Turb, MR, KHI, etc.
SOHO	Solar imaging, CME/flares
SDO	Solar imaging, CME/flares
ACE	SW monitoring at 1AU; energetic particles from the SW and the interplanetary medium; acceleration processes; SW composition
DSCOVR	SW conditions at 1AU
WIND	SW Turb; coherent structures and intermittent events; particle heating and acceleration
ULYSSES	Properties of the Sun at all latitudes; polar wind Turb
VAN ALLEN PROBES (RBSP)	Radiation belt flux enhancement
PROBA-2	Solar images (flares, CMEs) and plasma properties
DOUBLE STAR	Same as CLUSTER
Parker Solar Probe	SW physics between 9-35 Rs (solar radius)
OMNIweb	SW properties at the nose of the bow shock and geomagnetic indices
Digital Ionogram Data Base (DIDBase)	Large database of ionospheric measurements using ionograms at different locations around the Earth

The spacecraft data that will be used will come from open-access databases that are under responsibility of ESA or NASA. Their access will therefore follow the standard ESA and NASA protocol.

2. Fair Data

2.1 Making data findable, including provisions for metadata.

Discoverability: Metadata Provision

As described in paragraph 1 ASAP will produce and reuse a variety of data types (see table 3).

Table 3: ASAP data types.

Simulation results	***Trained machine learning models***	***Enhanced data by applying statistical tools and AI to satellite observations and simulations***
1. 2D, 3D fields (not too large) with typical grids of the order of 1024^2 and of 256^3. Many arrays such as the electromagnetic fields, density, velocities, etc. In total, per simulation, in the range of 100Mb to 5 Gb assuming about 20 fields for a given simulation. 2. Distribution function data (very huge): selected slices of 6D (3D-3V) arrays, very large from 500 Gb up to 10TB for a given simulation. 3. Direct time series virtual satellite measures (small size), up to 1 Gb for a given simulation.	1. Model hyperparameters (python dictionaries). 2. Trained models, mostly a collection of multiple 2D matrices per model.	1. Time series statistics values. 2. Time series extraction from simulations (virtual satellites) 3. Time series enhancement: multiple 1D arrays at a similar cadence as the original satellite time series. 4. Simulation processing with Representer Analysis techniques. 5. Velocity distribution function processing: enhancement of 2D/3D distributions from satellite observations. Multiple distributions over time. 6. Automatic identification and classification of regions of interest in time series, 2D, 3D data and velocity distributions. The output will be of the same size of the input but with additional identification/classification information.

Metadata will be produced for all data, using the schema in table 4.

Table 4: ASAP metadata.

Metadata	***Level***
Name	Entire Dataset
Data or link to the data (URL/DOI)	Entire Dataset
Title of the simulation or analysis containing main objective (e.g. Turb, KHI, Clustering, Classification, …)	Entire Dataset
Date of creation	Entire Dataset
Date of inclusion in the database	Entire Dataset
Software used	Entire Dataset
Version of the software (down to the commit number)	Entire Dataset
Input files or input conditions	Entire Dataset
Boundary and initial conditions	Entire Dataset
Job info: memory size, number of processors, total effective CPU 's and machine where the simulation has run	Entire Dataset
Other data required to duplicate the results	Entire Dataset
Dimensions	Dataset Field
Units	Dataset Field

Type (string, double, integers)	Dataset Field
Format (binary, HDF5, CDF, ASCII, PMML)	Dataset Field

Simulations specific metadata can be added, for example, the title, physical regime, grid size (Table 5).

Table 5: Simulations additional metadata example.

Title	***Regime***	***Grid space***	***Grid velocity***	***N.procs, Tot.mem***	***Machine***
Turb	Kinetic ions	256^3	51^3	32768, 45T	Marconi (Cineca)
KHI	Kinetic ions	256x256x128	51^3	8192, 22T	Marconi (Cineca)
Plume simulation	Fully kinetic (ions and electrons)	25600*10240		16384, 350 T	Curie (TGCC)

Naming convention

Using this information, it is possible to create a naming convention, for example following some of the points in the metadata: SOFTWARE_GITTAG_TYPE_CASENAME. This can change in the future. Versioning is due to the use of git commit numbers for the software and date of inclusion in the database.

Identifiability of data

ASAP is not actually using DOI for produced data. However, the introduction of persistent and unique identifiers using service like DataCite (https://datacite.org/dois.html) will be evaluated during the project for simulations and enhanced data.

2.2 Making data openly accessible

We will rely on publicly accessible repositories, particularly Zenodo and GitLab. Results of the numerical simulations will be made available on disk or tape storage. In particular a description of the numerical code, initial conditions, boundary conditions as well as the electromagnetic fields data obtained directly from the selected simulations will be stored. These data will be immediately accessible to all

components of the Consortium and will be made accessible to people external to the project by the end of the project at maximum, if possible, even before the end. Spacecraft data summarized in the ICD are public and come from ESA or NASA DB and are accessible to any user following the ESA or NASA procedures. ASAP will furthermore produce new high-level data products from the used spacecraft data that will include catalogues of features and events detected by ML and AI algorithms. These data sets will also be stored on Zenodo, and GitLab and will be made available to the public.

For the storage of software product ASAP is using GitLab (https://gitlab.com/). GitLab is a web-based hosting service for version control, mostly used for computer code. It offers the functionality of distributed version control and source code management, providing several collaboration features like task management, feature requests and bug tracking for every project. Projects on GitLab can be accessed using the standard Git command-line and a web interface. GitLab also allows registered and unregistered users to browse public repositories.

2.3 Making data interoperable.

To facilitate interoperability we will enumerate the numerical simulations using different capital letters referring to the physical environment as e.g., solar wind, magnetosphere, magnetotail and so on followed by the main physical process of interest as , e.g. magnetic reconnection, Kelvin-Helmholtz instability and so on. We will also use the L or E letter to distinguish the numerical approach, Lagrangian or Eulerian, respectively. The same terminology will be used for space data. A scientific standard vocabulary will be used for all data types in the data set.

Data quality assurance and quality control will be maintained throughout the project, utilizing strategies for:

- Preventing errors from entering a dataset. Typical types of errors include for example (a) incorrect or inaccurate data entered in the repository, or (b) data or metadata not recorded due to human error or anomalies in the field.
- Ensuring quality of data for entered data in order to identify potential data contamination. Typical examples include (a) checking for null values in the data (like NaN etc.) so as to avoid any missing, impossible or anomalous values within the data and (b) making sure that data line up in proper columns.
- Monitoring and maintaining the quality of data during the project in order to prevent data contamination.

ASAP will study the usability and inclusion of some visual and statistical analysis tools for performing QA on data entered in the repository.

2.4 Increase data re-use

During the project ASAP consortium will choose an OpenSource license for data access through the ASAP software.

Data generated by ASAP will be also reused in other scientific national or international projects or existing databases. An example might be ingesting into ESA science archives several ASAP high-level products obtained from both ESA data and numerical simulations. The focus would be on ESA Heliophysics Science Archives of completed or long-duration missions such as the Ulysses Final Archive, the Cluster and Double Star Science Archive and the SOHO Science Archive. During the ASAP project, discussions will be carried out with ESA and an agreement will be pursued between the project's PI and ESA to manage the interfacing between AIDA's and ESA's databases. This will happen through a relevant ICD (Interface Control Documents) consisting of the rules for pushing-pulling data, together with lists of metadata to be ingested in ESA databases.

3. Other research outputs

More than data, the main outcome of ASAP will be software for the dedicated hardware used in space missions. Our software will be open source and made available on GitLab.

4. Allocation of resources

There are no immediate costs anticipated to make the data produced FAIR. Data will be processed using the tools developed in the ASAP framework and stored in freely available systems. The databases created by ASAP will be public and interfaced with community-based databases such as the databases maintained by ESA.

5. Data security

For the duration of the project, data will be protected against unauthorized access by means of standard authentication. Appropriate access levels will be granted by definition of roles in the project.

6. Ethics

There are no ethical issues in the generation and analysis of simulations and solar data. There are no human subjects or samples involved. Personal data of project participants will be treated following GDPR.

7. Other issues

No use is planned of other national/funder/sectorial/departmental procedures for data management.

Appendix 1: list of acronyms

List of acronyms	
AI	Artificial Intelligence
ASCII	American Standard Code for Information Interchange
AU	Astronomical Unit
CDF	Common Data Format
CME	Coronal Mass Ejection
DL	Deep Learning
DMP	Data Management Plan
DOI	Digital Object Identifier
ESA	European Space Agency
FAIR	Findable Accessible Interoperable Reusable
GPFS	General Parallel File System
GSS	GPFS Storage Server
HDF5	Hierarchical Data Format
HPC	High Performance Computing
ICD	Interface Control Documents
IDL	Interactive Data Language
iRODS	Integrated Rule-Oriented Data System
KH	Kelvin–Helmholtz Instability
LTFS	Linear Tape File System
ML	Machine Learning
MR	Magnetic Reconnection
MP	Magnetopause
MS	MagnetoSphere
MT	MagnetoTail
NaN	Not a Number
NASA	National Aeronautics and Space Administration
ONNX	Open Neural Network Exchange
PI	Principal Investigator
PMML	Predictive Model markup Language

QA	Quality Assessment
SSH	Secure SHell
SW	Solar Wind
Turb	Turbulence
URL	Uniform Resource Locator