Funded by the Horizon 2020 Framework
Programme of the European Union

HEMERA

N° HEMERA/2019/09/RP/001 Issue 1/revision 1

Date: 01/15/2019

HEMERA-2020
Data Management Plan

Draft version	x
Final approved version	

Work package applicability:

WP 1	WP 2	WP 3	WP 4	WP 5	WP 6	WP 7	WP 8	WP 9	WP 10	WP 11	WP12
	x										

DOCUMENT CREATION:

Type of document (Report RP, Note NT, Data DT, Minutes MN, List LT, Plan PL)	RP
Category (PU, CO)	PU
Subject	Data Management Plan
Author(s)	Cathy Boonne, Sébastien Payan
Company	CNRS
Key words	Balloon, Data, Virtual Access

MODIFICATIONS:

Version	Date	Modifications	Observations
0.1	2018/06/20		First Draft
0.2	2018/07/11	Revised document based on version 0.1	add chapter
0.3	2018/07/19	Revised document based on version 0.2	
0.4	2018/09/25	Revised document based on version 0.3	
0.5	2018/09/25	Revised document based on version 0.4	
0.6	2018/10/18	Revised document based on version 0.5	
0.7	2018/10/26	Revised document based on version 0.6	
0.8	2018/11/20	Revised document based on version 0.7	
0.9	2018/12/14	Revised document based on version 0.8	
1.0	2019/01/15	Revised document based on version 0.9	Final document

DISTRIBUTION LIST:

Steering Committee	Y
All partners	Y
Advisory Committee	Y
Open access	Y

HEMERA WP members: Public document

Summary

1. Introduction

The overall HEMERA-2020 project [A1] aims to provide the best balloon measurements. This requires a highly integrated data and information management system. The project is composed by a coordinated set of networking activities, which delivers improved balloon data across the infrastructure, as well as standard protocols for data generation and analysis.

The main objective is to make all the scientific and technological data collected during the flights accessible to the whole European scientific community, upon request to the Data Centre (DC). The data centre will provide free access and services for data archiving including higher level data products, links to large databases of past and ongoing scientific balloon data projects, complemented with access to new data products, together with tools for quality assurance (QA), data analysis and research.

The architecture of the DC will be described in a next deliverable (D2.3).

Currently, the balloon-borne Data Centres are founded on two topical databases:

- Atmospheric balloon-borne database (https://cds-espri.ipsl.upmc.fr/BALLOON),
- Astrophysical balloon-borne database (https://www.asi.it/eng/agency/bases/data-center).

1.1. Purpose of the document

The Data Management Plan (DMP) considers the data management life cycle for the data sets to be collected and processed by HEMERA-2020 project. The DMP outlines the handling of research data during the project, and how and what parts of the data sets will be made available after the project has been completed. This includes an assessment of when and how data can be shared. The DMP describes also the choices that will be made for the metadata standards to be used, database repository, data access policy and data access methods, long term archival and the costs associated to data management.

With regard to access to research data, HEMERA-2020 will make the data and metadata available on the new website DC. From this website, project members and external users will have access to both data and metadata. Research data is originally planned to be archived at IPSL/CNRS.

1.2. Intended readership

This deliverable is intended for use internally in the project and provides guidance on data management to the project partners responsible for data collection. At the current stage of the project, this DMP is just the initial DMP and will evolve throughout the project as new research data sets will be added or modified.

1.3. Document outline

The document consists of the following sections:

- **Section 2** describes the guiding principles for the data management of the overall HEMERA-2020 data sets.
- **Section 3** lists the data sets provided by HEMERA-2020 DC and will provide :
  - the data sets description,
  - the standards and metadata related to,
  - the sharing of the data sets,
  - the procedures for archiving and long-term preservation of the data.
- **Section 4** presents the FAIR data.

1.4. Application area

The prime focus of this document will be on **HEMERA-2020 Virtual Access (WP2)**, as specified in the HEMERA-2020 project document. [A1].

1.5. Applicable documents and reference documents

Applicable documents

[A1] HEMERA-2020 project document

1.6. Abbreviations

ABBREVIATIONS	SIGNIFICATION
ASI	Agenzia Spaziale Italiana
CNRS	Centre National de la Recherche Scientifique
DC	Data Centre
DMP	Data Management Plan
DOI	Digital Object identifier
FAIR	Findable, Accessible, Interoperable and Re-usable
IPSL	Institut Pierre Simon Laplace
QA	Quality assurance
WP	Workpackage

2. Functional guidance principles

The general approach to data management support for HEMERA-2020 project is summarized in a data flow diagram (see Fig.1 below). It is important that the HEMERA-2020 data management strategy be responsive to the needs of the investigators, ensuring that data are accurate and disseminated in a timely fashion. It is also important that the investigators know what is expected of them in this process.

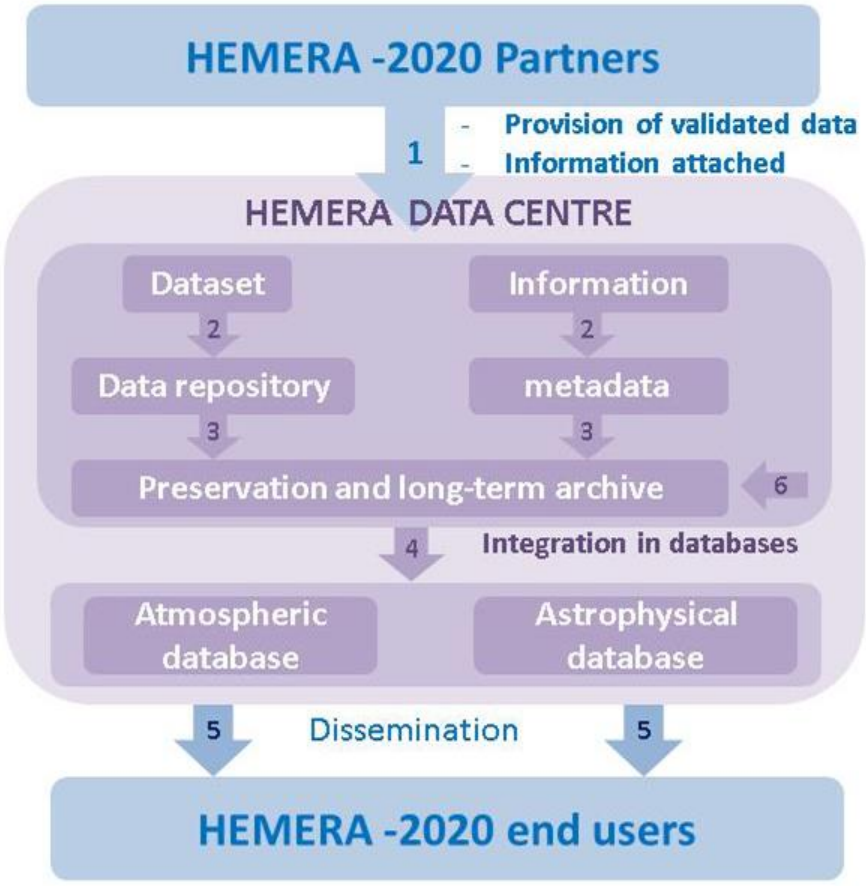

Figure 1: HEMERA-2020 data flow

Step 1: Products provided by HEMERA-2020 institutions are validated and qualified before provisioning AERIS infrastructure. Format of products is going to be described further in the document.

Step 2: Data and information transfer to AERIS infrastructure. Procedures assuming the data transfer will be relied on the specification given by the relevant WP.

Step 3: Data and information management: data populate the data repository and information is treated to create metadata files to populate the HEMERA-2020 catalogue.

Step 4: Data and metadata integration in long-term sustainable databases.

Step 5: Dissemination to end-users. The HEMERA DC infrastructure ensures open access to all data.

Step 6: Preservation and backup of data, information, databases and web site.

The archive, web interface and supporting software will continue to be maintained and updated to ingest new data, and to accommodate changes in the data streams. The archive catalogue record will be maintained to enable dataset-level. These processes will continue until the end of the project and an infrastructure to ensure the long-term availability of the data to the broader community will be set in place. After the end of the project, the data will remain available on a best effort basis.

3. HEMERA-2020 data sets description

In this chapter we describe the different data sets that have been provided by HEMERA partners.

The HEMERA-2020 Data Centre is organized in 2 different databases:

- The Atmospheric database: This database provides a compilation of experimental data obtained from balloon experiments supplied by partners of the consortium.
- The Astrophysical database: This database provides TBC

These databases already exist and were developed within AERIS and ASI data centre respectively.

The table 1 gives an overview of the existing data sets that have been already collected by the new DC within HEMERA-2020. The description of these data sets is given in the following sections.

Table 1: Overview of the data sets collected

Experimental data sets	Brief description
Atmospheric observations: from balloon-borne experiments	These data are time series of chemical or physical variables measured during balloon-borne experiments. These data are in NASA-AMESformat
Astrophysical observations: from balloon-borne experiments	These data are in fits format

All these data sets have been provided by these European institutions:

- CNRS-LPC2E, France
- CNRS-LATMOS, France
- CNRS-LMD, France
- CNRS-LOA, France
- GSMA, Reims University, France
- IAUG, Frankfurt, Germany
- KIT, Karlsruhe, Germany
- INAF, Italy
- …

3.1. Products description

3.1.1 Atmospheric balloon-borne experiments data sets

These data sets are stratospheric chemical variables and related measured during balloon-borne experiments which types are the following:

- Ozone
- NOy chemical family
- Cly chemical family
- Bry chemical family

- HOx chemical family
- Dynamic tracers
- Aerosols

For each set of data, several files have to be provided by partners:

- A pdf file that describes the experiment and provides information on the experimental conditions.
- One or several files containing data in a unique format, called NASA-Ames format. This format is an ASCII based format and is described in Appendix 1. In the case several NASA-Ames files are provided, they can be gathered in a zip file.
- A metadata file indexing the datasets.

These data sets have been collected and managed by AERIS.

Nature and scale of data: vertical profiles of the key stratospheric species (concentrations of gaseous species) that control the mid-latitude ozone budget. The total data volume of this database is currently less than 200 GB and more than 1000 files.

To whom the data set could be useful: These data are of high interest for a large community of users in atmospheric sciences, as well as the private sector. These observations are very useful:

- To compare measurements of the same species recorded by different instruments on the balloon,
- To use them for validation of satellite measurements.

Existence of similar data sets?

Atmospheric balloon-borne data sets can be found elsewhere but this database focuses on molecules of interest for atmospheric sciences.

3.1.2 Astrophysical balloon-borne experiments data sets

All these data are qualified as HEMERA-2020 data only if

- The measurement data files are submitting to the HEMERA data centre in “FITS” format as described in Appendix 2,
- A metadata file indexing the datasets.

Nature and scale of data: TBC

To whom the data set could be useful: TBC

Existence of similar data sets? TBC

3.2 Standards and metadata

The supply of detailed metadata is mandatory for datasets - new or existing - to be referenced by the virtual access in the HEMERA-2020 metadata catalogue. Automatic validation processes ensure the quality and the completeness of the provided information. Each metadata record is associated with a unique universal identifier.

The specification of the metadata profile per datasets is the following:

- resource title
- resource abstract
- id
- temporal extents
- publications
- links
  - type
  - url
  - name
  - description
- contacts
  - name
  - email
  - organization
  - comment
  - address
  - roles
- formats
- data level
- platforms
- parameters
- instrument
- resolution
- type

3.3 Data sharing

Access procedures: Through the web interface, the HEMERA data centre will provide a user-friendly, multi-criteria, research mechanism to discover and preview the datasets of the catalogue.

Interaction with other catalogues is another important point to make our data findable. To achieve this we will use standards for structuring information (e.g. ISO19115), defining vocabularies and for querying our catalogue.

The access procedure will be achieved with a shopping-cart mechanism to select datasets found in our catalogue. In addition to the possibility of a direct download, the data centre will propose scripts to execute the download programmatically. Each downloaded file will be an archive containing additional files recalling metadata, licenses and how to quote and acknowledge data.

Open source tools to manipulate and plot data and corresponding documentation will be included in a dedicated page on the data centre.

To simplify data retrieval for the users, HEMERA data centre will use widely used authentication schemes such as OrcId which is already used in other European research infrastructures.

Document format and availability: The data sets are available in their native format through the HEMERA data centre. From there the fully data are accessible to internal and external users (in and out of the project), free of charge.

3.4 Archiving and preservation (including storage and backup)

Archiving of the data sets by AERIS guarantees a long-term and secure preservation of the data without any additional cost for the project. This access will be freely available all along the years. Free and open access means unrestricted access at no cost for all interested individuals, whether they are with in or outside of the project, but an acceptance of the HEMERA data policy will be required.

Access to all data products and tools will be recorded through web-based user statistics for all virtual access activities.

3.5 Data volume

The new Data Centre infrastructure is correctly sized to be able to accommodate the new datasets, including their different versions and safety copy. The datasets volume represents less than 200 GB and currently such datasets volume is very easy to store and archive in different location.

3.6 Data repository description

The directory structure of the data repository is the following:

---------|/data (root of the hierarchical tree)
----------|/DATB (Database of Atmospheric balloon-borne experiments)
----------|/ID (ID of the metadata describing the dataset)
----------| xxxx.pdf (information file)
----------| xxxx.ames (datafile)
----------|/ID ….
| …
| …
……..
……..
----------|/DASB (Database of Astrophysical balloon-borne experiments)
----------|/ID ((ID of the metadata describing the dataset)
----------| xxxx.pdf (information file)
----------| xxx.fits (datafile)
----------|/ID …
----------| …

3.7 Preliminary data policy

Data are available all along the project and there is no embargo period; as soon as they are on the website, they can be used by internal and external users. The full data policy will be described later in a new revision of this document but the main elements of this policy will comprise:

- Data ownership,
- Data curation,
- Data archiving,
- Open access to data.

4. FAIR DATA

4.1. Findable data

Each metadata record is associated with a unique universal identifier. This will allow the establishment of an automatic link with “Datacite”. Hence, every dataset will be quotable through DOI. We will use the concept of fragment to precisely quote the different versions of a dataset.

Through its web interface, the data centre provides a user-friendly, multi-criteria, research mechanism to discover and preview the datasets of our catalogue. Interaction with other catalogues is another important point to make our data findable. To achieve this we use standards for structuring information (e.g. ISO19115), defining vocabularies (e.g. CF - Climate and Forecast- conventions) and for querying our catalogue (e.g. CSW).

4.1.1 Atmospheric balloon-borne experiments data sets: criteria list

The data research mechanism is based on this multi-criteria list:

- Parameters
- Platforms
- Instruments

4.1.2 Astrophysical balloon-borne experiments data sets: criteria list

The data research mechanism is based on this multi-criteria list: TBC

4.2. Openly accessible data

The web interface of HEMERA DC provides access to all data resulting from the activities of the new infrastructure. This is achieved with a shopping cart mechanism to select datasets found in our catalogue. In addition to the possibility of a direct download, the data centre proposes scripts to execute the download programmatically. Each downloaded file is an archive containing additional files recalling metadata, licenses and how to quote and acknowledge data.

Open source tools to manipulate data and corresponding documentation are included in a dedicated page on the data centre.

To simplify data retrieval for the users, HEMERA DC uses widely used authentication schemes such as ORCID which is already used in other European research infrastructures.

4.3. Interoperable data

Just like the metadata, data are interoperable by adhering to identified open standards and shared vocabularies in the research community.

4.4. Reusable data

The HEMERA data Policy is implemented by the data centre. Its goal is to regulate the sharing of HEMERA data and includes information on dissemination, sharing and potential access restriction. The data policy creation is on going and will also be publically available on the HEMERA DC.

4.5. Data security and long term conservation

The integrity and the security of the collected data are done by mechanisms which are either already existing or currently being developed in the scope of this project. These mechanisms imply multi-site archiving, regular checksum of data files, and automatic data format update if necessary.

4.6. Organization and human resources

As mentioned in the proposal, the AERIS/ESPRI data centre and the HEMERA-2020 data centre involve staff with complementary knowledge and competences. Precise points of contact for both topical and technical questions are indicated in a dedicated web page. They are accessible either via emails or online forms (helpdesk).

APPENDIX

APPENDIX 1. Description of the NASA-Ames format

The NASA Ames format is a text-based, self-describing, portable format. File contents are limited to the printable ASCII character set (ASCII codes 32 to 126). Each NASA Ames file is made up of a file header section and a data section. The file header contains the information needed to make the file self describing, as well as giving information such as the origin of the data. Once the form of a file for a particular instrument has been decided, the file header for that instrument changes little from file to file. The data section lists the data, in a column-oriented format.

For more details see: **http://artefacts1.ceda.ac.uk/formats/NASA-Ames/na-brief-guide.html**

APPENDIX 2 Description of the FITS format

Flexible Image Transport System (**FITS**) is an open standard[3] defining a digital file format useful for storage, transmission and processing of data: formatted as N-dimensional arrays (for example a 2D image), or tables. FITS is the most commonly used digital file format in astronomy. The FITS standard has special (optional) features for scientific data, for example it includes many provisions for describing photometric and spatial calibration information, together with image origin metadata.

For more details see: https://fits.gsfc.nasa.gov/fits_standard.html