Funded by the Horizon 2020 Framework Programme of the European Union

HEMERA

N° HEMERA/2019/9/NT/029 Issue 1/revision 1

Date: 03/12/2019

HEMERA-2020

Definition of HEMERA Data Centre

Draft version	
Final approved version	x

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	Definition of HEMERA Data Centre
Author(s)	Cathy Boonne, Sébastien Payan
Company	CNRS
Key words	Balloon, Data, Virtual Access

MODIFICATIONS:

Version	Date	Modifications	Observations
0.1	2018/11/15		First Draft
0.2	2018/12/20	Revised document based on version 0.1	add chapter
0.3	2019/02/25	Revised document based on version 0.2	Final draft
0.4	2019/03/07	Revised document based on version 0.3	Final document

DISTRIBUTION LIST:

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

HEMERA WP members: Public document

SOMMAIRE

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. Furthermore, a link to the up-to-date catalog of known high-energy astronomical sources as well as the link to the alerts of newly discovered sources will be released.

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

This document aims at describing the general architecture of the HEMERA data centre. The architecture is split in different components that will be described in this document.

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.

1.3. Document outline

The document consists of the following 11 sections:

- Components
- Global considerations
- Data storage
- Data preservation
- Metadata catalogue
- Content management
- Authentication & Authorization
- Administration
- Service layer
- Graphical user interface
- Interoperability

that define the architecture of the HEMERA data centre.

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 contract

1.6. Abbreviations

ABBREVIATIONS	SIGNIFICATION
ASI-INAF	Agenzia Spaziale Italiana – Instituto Nazionale di Astrofisica
CNRS	Centre National de la Recherche Scientifique
DC	Data Centre
DOI	Digital Object identifier
IPSL	Institut Pierre Simon Laplace
P.I.	Principal Investigator
QA	Quality Assurance
WP	Work Package

2. Components

The HEMERA data centre is divided in different components. They can be either functional or transverse components:

- **Functional components**: components which ensure an isolated functionality,
- **Transverse components**: components which either aggregate services coming from functional components or provide services to functional components.

This list below indicates the components of the HEMERA data centre:

Component	Role
Functional components	
Data storage	Stores every produced dataset
Data preservation	Provides redundant storage to ensure long term preservation
Metadata catalogue	Stores metadata records and offers discovery services for them
Content management	Provides features devoted to the management of the content of the web site (screens, news, events...)
Administration	Provides operating features such as user management, parameter management or data reporting
Transversal components	
Authorization and Authentication	Authenticates users and grants roles
Service Layer	Connects the user interface and the functional components to provide added value services
Graphical User Interface	Displays information to the final user

To illustrate this architecture, a general schema is given is the figure 1.

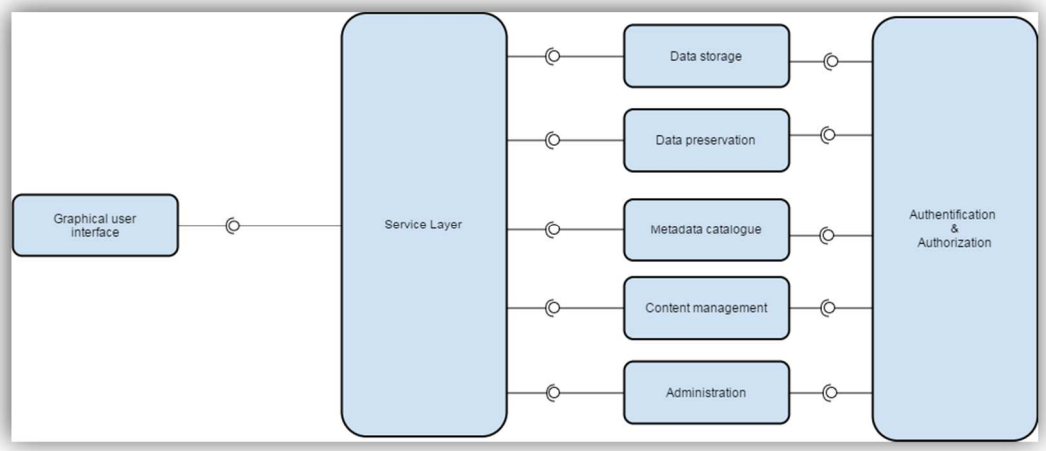


Figure 1: General scheme of the HEMERA data centre architecture

3. Global considerations

3.1. HEMERA data properties

Compared to other Earth observation data, HEMERA data have got the following specificities:

- Fixed sensors,
- Short temporal extensions: HEMERA datasets aren't long time ongoing series,
- Small volumes,
- Isolated datasets: Datasets aren't linked with other datasets.

3.2. HEMERA identification

Due to what has been mentioned above, each produced dataset can easily be isolated and then identified. Meaningful identifiers are generally a bad idea. For instance, such identifiers would certainly be an obstacle to setup interoperability with other partners and data centres.

Hence, the data centre will use **UUIDs** (Universally Unique IDentifiers) to identify each dataset. This UUID will be used in other aspects of the life cycle of the dataset. For instance, it will be used to identify the metadata record associated to the dataset.

Technically, we will use the RFC 4122 standard to generate UUIDs. Hence, they will be composed of 34 alphanumeric characters and four hyphens (e.g.: 123e4567-e89b-12d3-a456-426655440000). More information on UUIDs can be found here:

https://en.wikipedia.org/wiki/Universally_unique_identifier

3.2.1. New datasets

For new datasets, the UUID will be generated when the dataset form is first submitted by the P.I.

3.2.2. Existing datasets

For existing atmospheric and astronomical datasets, an identifier will be allocated with an injective function during the recovery phase.

4. Data storage

4.1.Organisation

Because of the isolated nature of the produced datasets, the data centre will retain the produced files as the corner stone for storage. Hence, data storage will be file-based and won't imply any relational database (such as Oracle, MySQL ...). Thus, in this document, the term *database* will refer to the file tree described in this section.

4.1.1.File server

The data centre will use a server provider by AERIS-ESPRI.

The file server will communicate with other components of the data centre via FTP protocol with a dedicated user authorized to access to the root of the database (figure 2).

It's important to note that the P.I. won't have a direct access to the data storage. All access will pass through the service layer. This is necessary to ensure the integrity of the data centre and to enable richer services, such as file validation or download measurement.

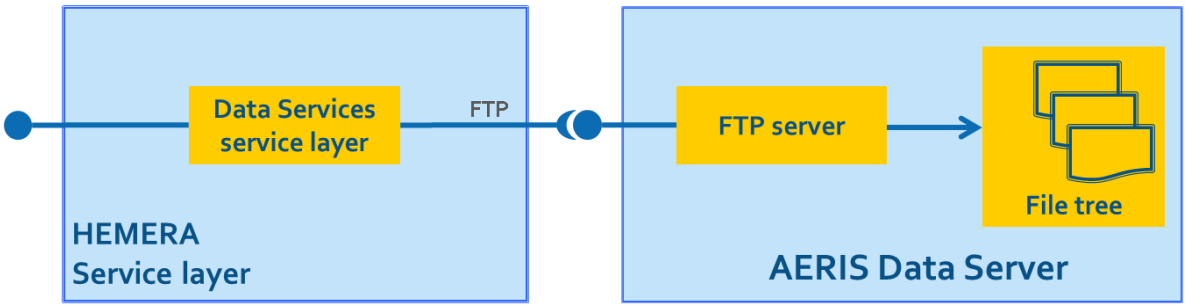


Figure 2: File server general scheme

4.1.2.File tree

The root folder of the database will contain two main subfolders (figure 3):

- DATB (Database of Atmospheric balloon-borne experiments)
- DASB (Database of Astrophysical balloon-borne experiments)

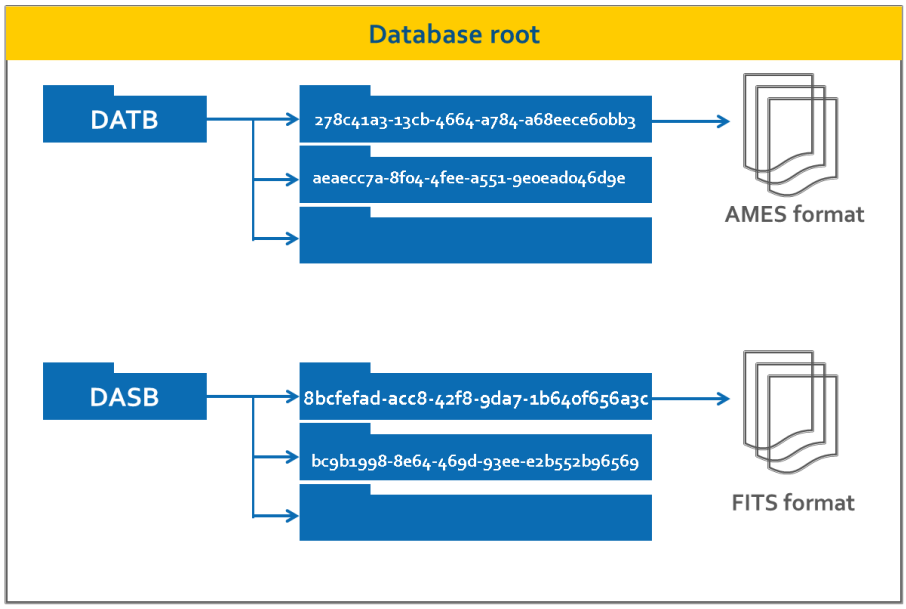


Figure 3: File tree scheme

The leaf folders will contain their respective datasets. These datasets will consist in one or several files which will be stored in a folder named by their UUID.

4.1.3.File format

- *DATB (Database of Atmospheric balloon-borne experiments)*

Datasets will consist in:

- One or several NASA Ames files,
- Optional auxiliary data files (GPS data file)

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**

- *DASB (Database of Atmospheric balloon-borne experiments)*

Datasets will consist in:

- One or several FITS files.

The Flexible Image Transport System (FITS) is an open standard 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**

4.1.4.File names

The data providers are free to choose the name for the file they provide as they don't contain any of the following special character:

- space, tabulation, slash, accented letter

Each file name will be lower cased on the file server.

4.2. Data versioning

The HEMERA data centre can handle dataset versioning.

5. Data preservation

5.1. Data replication

It's important to have several copies of datasets in order to prevent data loss in case of a major incident on the file server. Two replication mechanisms will be organized: a local replication and distant replication (figure 4).

5.1.1. Local replication

At the hardware level, the file server will be back-up every day. The underlying mechanism generates daily, weekly and monthly incremental archives.

5.1.2.Distant replication

Each file addition/modification will trigger of copy on the AERIS-SEDOO long term archiving infrastructure. This infrastructure is located in Toulouse and Tarbes. Thus, AERIS-ESPRI and AERIS-SEDOO are 600 km apart which prevent from data loss in case of major incident on the whole AERIS-ESPRI centre.

5.2. Integrity check

The distant replication implies the computation of an integrity checksum to ensure that the files haven't been degraded during the copy. This checksum will be stored in the AERIS-SEDOO archiving infrastructure. Regularly, the checksum will be re-computed in AERIS-ESPRI and AERIS-SEDOO to confirm the files still match. Otherwise, an alert is sent in order to trigger manual intervention.
More information on checksums can be found here: https://en.wikipedia.org/wiki/Checksum

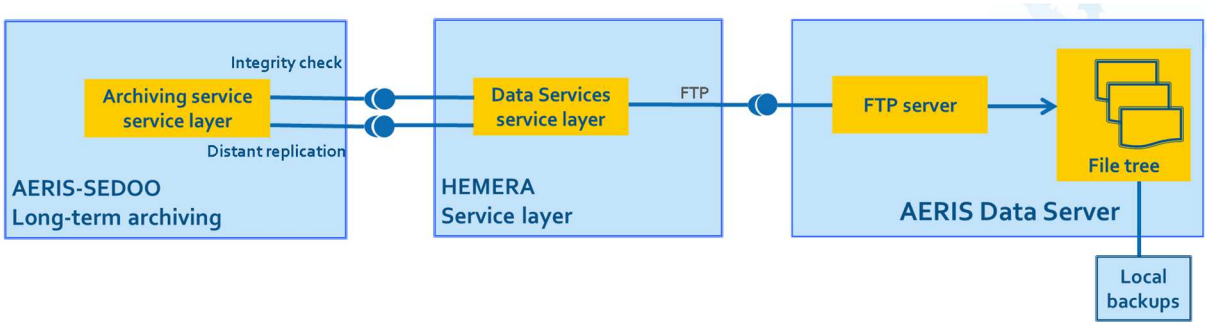


Figure 4: Data preservation scheme.

5.3. Note on other data

As mentioned below, the application will manipulate auxiliary data: screen contents, user lists, application parameters... These data will only be back-up by the hardware level mechanism.

6. Metadata catalogue

Metadata are really critical for the data centre to discover, understand or cite datasets.

6.1. Metadata record

For each produced dataset, a metadata record will be created. This record will have the same UUID than the dataset. In consequence, each metadata record will be associated to a specific URL (http://catalogaddress/*uuid*). This URL will be used as a landing page for DOI (cf. below). The metadata records are specified in Annex 1.

6.2. Catalogue

The data centre will use the AERIS Catalogue to store the metadata records. Though relying on proprietary format and implementation, this catalogue offers classical services such as multi-criteria querying and displaying.

This catalogue is used to store metadata records coming from other atmospheric projects. However, the catalogue supports *project filtering* and will only display information related to HEMERA. Moreover, it can be easily customized to adapt specific needs (search criteria, metadata profile).

Just as data files, metadata records will also be replicated (local and distant replication).

7. Content management

Data portals generally involve two kinds of features:

- Specific domain-related features,
- CMS (Content Management System) features.

Because of their intrinsic differences, specific applications and CMS are usually difficult to reconcile and architectural solutions are rarely satisfying. For instance, developing CMS features in the specific application ends to be costly and hard to maintain. On the other hand, a CMS-based strategy can be also expensive if a migration to another CMS is needed (figure 5).

HEMERA data centre will rely on a hybrid solution which is described in the *Graphical User Interface* section. This solution uses a CMS - WordPress - to provide classical content management features but minimizes its weight in the global system.

However, WordPress will keep a separate place in the data centre architecture with a dedicated server and database and a direct link with the user interface.

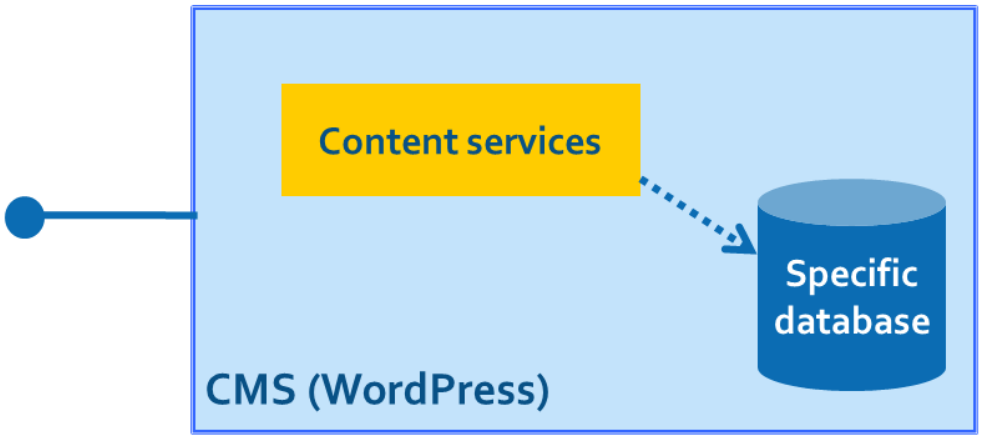
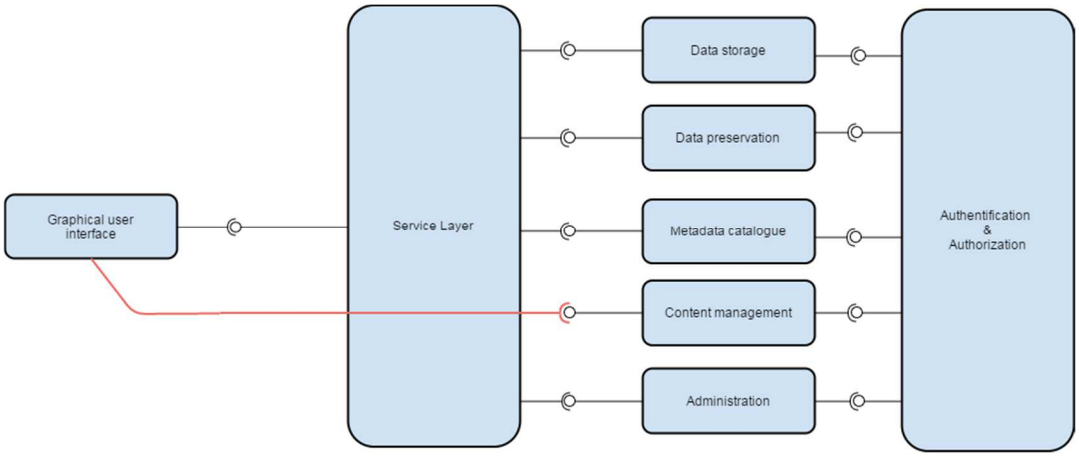


Figure 5: CMS general scheme

In consequence the general component scheme must be corrected like that:



8. Authentication & Authorization

All data centre features won't be available for every user. Indeed, several roles will be put in place:

- Data downloader,
- Data provider,
- Web site editor,
- Administrator,
- ...

Thus, an authentication system - in charge of proving the identity of users - and an authorization system - in charge of granting roles - are mandatory.

8.1. Authentication

The HEMERA data centre will delegate authentication to the ORCID OAuth2 system (cf. Interoperability). This choice is motivated by the fact that this mechanism is also being adopted by several research infrastructures.

It's important to notice that only the ORCID will be stored in the data centre. Principal's information such as name of email address will be asked to ORCID when needed.

8.2. Authorization

Authorizations will be managed internally. For each user, identified by his ORCID, a list of roles will be stored locally and managed via the user interface (figure 6).

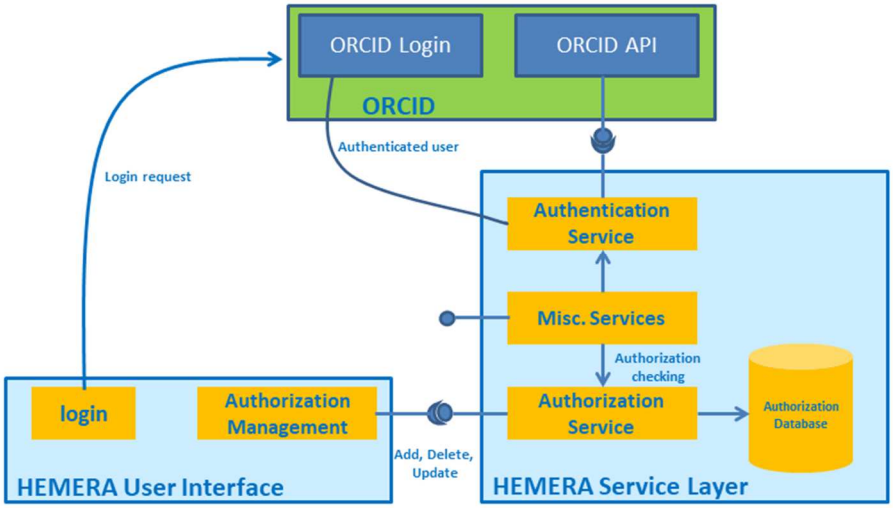


Figure 6: Authorisation scheme

9. Administration

The data portal requires auxiliary administration features such as user management and reporting. These features will be implemented in the service layer and will rely upon a local database (figure 7).

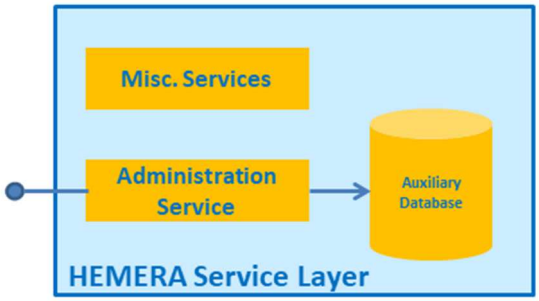


Figure 7: Data centre administration scheme

10. Service layer

The service layer aims at providing modularity and clear role separation which are important to ensure maintainability and evolution of the data centre architecture. Indeed, the service layer is the hub that links the different blocks. Hence, each block can evolve separately without any effect on the others.

The service layer will consist on REST services (https://fr.wikipedia.org/wiki/Representational_state_transfer) hosted on AERIS-ESPRI server.

These services will use JSON standard (https://fr.wikipedia.org/wiki/JavaScript_Object_Notation) to format information.

11. Graphical user interface

The graphical user interface is responsible for collecting information from different services and displaying them to the user's browser.

The skeleton of the web site (static pages, navigation, ...) will be provided by the CMS.

Graphical rendering of domain-specific features won't rely on the CMS framework. They will be developed with a neutral client-side technology:

- web components (https://fr.wikipedia.org/wiki/Composants_web),

which allows to develop custom rich HTML tags. These components will be included in static pages served by the CMS. They will interact directly with the service layer via REST requests.

Consequences of this strategy are:

- Reuse of AERIS existing web components (catalogue),
- Decrease of the link with a specific CMS,
- Improve of the user experience,
- Decrease of the cost of maintenance of the web site.

12. Interoperability

Interoperability is the fact that different systems can automatically exchange information. In geosciences, interoperability is generally associated to the popular OGC protocols (https://en.wikipedia.org/wiki/Open_Geospatial_Consortium).

These protocols enable:

- metadata interoperability: catalogue querying, metadata record harvesting (CSW protocol)
- georeferenced data interoperability: data visualization on map (WMS/WFS protocols)

But interoperability is wider than that and it exists a lot of practical applications of interoperability such as social login (https://en.wikipedia.org/wiki/Social_login).

This section indicates the interoperability mechanism that will be offered or support by the data centre.

12.1. ORCID

ORCID (Open Researcher and Contributor ID) is an alphanumeric code to uniquely identify scientific and other academic authors and contributors (e.g. 0000-0002-4510-0385). The ORCID organization

offers an open and independent registry intended to be the de facto standard for contributor identification in research and academic publishing. ORCID is getting progressively adopted in the European Research Infrastructures via the EnvriPlus project.

The HEMERA data centre will use ORCID in two different ways:

- **In metadata records**: ORCID can be indicated in contact descriptions. In the case of a P.I., the ORCID will be included in the DOI.
- **For authentication**: The data centre won't provide its own authentication mechanism. It will rely on the OAuth2 service provided by ORCID (cf. above).

12.2. Metadata Interoperability

Many standards exist around metadata interoperability:

- ISO standards for metadata content (ISO19115, ISO19115-2, ISO19139...)
- OGC protocol for catalogue querying and harvesting.
- Because of its wide adoption, Geonetwork, can also be considered to be a standard tool in this domain.

Each metadata provider has got his own way to use them: metadata content can vary from a provider to another one with different granularities, vocabularies, languages or structures. Hence, it's difficult to pretend to be globally metadata interoperable.

As mentioned, internally, HEMERA metadata records will be stored in the AERIS catalogue which relies on a proprietary format which might appear to be contradictory with interoperability.

In fact, the proprietary format is needed to enable rich and efficient features in the catalogue and in the user interface, the data centre will target specific metadata interoperability with putting in place the following points:

- Metadata records will be Inspire compliant. This will guarantee that they contain the minimal core of information which is generally required for basic interoperability (title, abstract, contacts, links ...)
- The data centre will offer a default converter to translate AERIS internal metadata format to ISO19139 for any metadata in the INSPIRE core
- The data centre will offer CSW containers dedicated to each partner who want to harvest HEMERA metadata.

Thus, if a partner asks for harvesting, the main necessary action is to adapt the default converter to the specific needs of the partner.

12.3. DOI

DOI are important elements to correctly cite a dataset. With the user interface, P.I. will be able to ask for DOI for their datasets as of all the minimal metadata are provided (figure 8).

The properties of this DOI will be:

Prefix	AERIS prefix
Suffix	UUID of the dataset
Landing page	URL of the metadata record (cf. above)
Metadata	Subset of metadata record. If ORCID is indicated, it will be passed to Datacite.

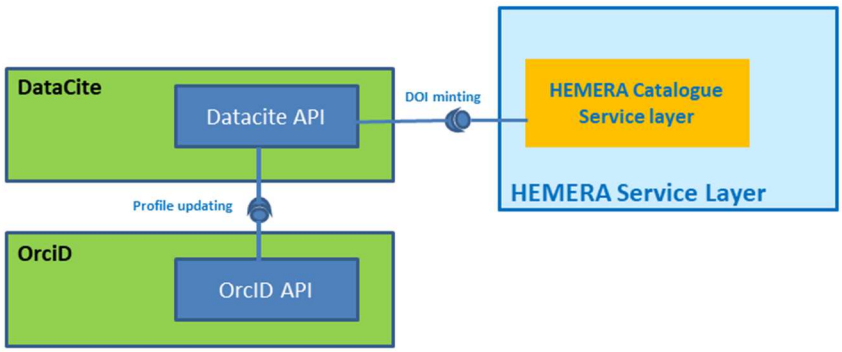


Figure 8: DOI scheme

12.4. Data interoperability

Data interoperability will consist in conversion features. Hence, partners will easily integrate HEMERA data in their own information systems.

ANNEX 1: Metadata records (in bold) specification in JSON format.

```
{
 "resourceTitle" : {
  "en" : " ",
  "fr" : " "
 },
 "resourceAbstract" : {
  "en" : " ",
  "fr" : " "
 },
 "id" : " ",
 "lastModification" : {
  "value" : " "
 },
 "aerisDataCenter" : "ESPRI",
 "temporalExtents" : [ {
  "beginDate" :  “ ",
  "endDate" : " ",
  "comment" : {
   "DEFAULT_VALUE_KEY" : ""
  }
 } ],
 "spatialExtents" : [ {
  "area" : {
   "type" : " ",
   "latitude" : ,
   "longitude" :  ,
   "altitude" :
  },
  "name" : " ",
  "description" : "",
  "additionalData" :  ,
  "comment" : " ",
  "projection" : " ",
  "spatialRepresentation" : ,
  "orbit" :
 } ],
 "publications" : [ {
  "title" : " ",
  "description" : " ",
  "authors" : " ",
  "journalName" : " ",
  "journalSection" : " ",
  "publicationYear" :  ,
  "doi" : ""
 } ],
 "links" : [ {
  "type" : "FTP_DOWNLOAD_LINK",
  "url" : " ",
  "name" : "",
  "description" : {
```

```
    "en" : " ",
    "fr" : " "
   }
  }, {
   "type" : "INFORMATION_LINK",
   "url" : "",
   "name" : "",
   "description" : {
    "en" : "",
    "fr" : ""
   }
  } ],
  "identifiers" : [ {
   "code" : "",
   "codeSpace" : "http://dx.doi.org/"
  } ],
  "status" : "Public",
  "distributionInformation" : {
   "dataPolicyName" : " ",
   "dataPolicyVersion" : " ",
   "dataPolicyUrl" : " ",
   "embargoDuration" :  ,
   "registrationNeeded" : true,
   "description" : {
    "DEFAULT_VALUE_KEY" : " "
   },
   "licenceName" : "",
   "licenceVersion" : "",
   "licencePolicyUrl" : "",
   "accessConstraints" : null,
   "useConstraints" : null
  },
  "language" : "en",
  "contacts" : [ {
   "name" : " ",
   "email" : " ",
   "organisation" : " ",
   "comment" : null,
   "address" : null,
   "roles" : " ",
   "orcId" : " "
  }, {
   "name" : "HEMERA contact",
   "email" : " ",
   "organisation" : "AERIS",
   "comment" : {
    "DEFAULT_VALUE_KEY" : ""
   },
   "address" : null,
   "roles" : [ "pointofcontact" ],
   "orcId" : " "
  } ],
```

```
"quicklooks" : [ {
  "url" : " ",
  "description" : null
} ],
"keywords" : [ {
  "concept" : "HEMERA"
} ],
"modifications" : null,
"genealogy" : null,
"formats" : [ {
  "name" : " ",
  "version" : " ",
  "description" : {
    "DEFAULT_VALUE_KEY" : " "
  },
  "readingInformation" : null,
  "temporalInterval" : null
} ],
"dataLevel" : "L2",
"programName" : "HEMERA",
"collectionName" : "HEMERA",
"clientTemplateName" : "",
"plateforms" :  ,
"parameters" : [ {
  "shortName" : " " ,
  "longName" : " ",
  "uom" : " ",
  "comment" : {
    "en" : " ",
    "fr" : " "
  },
  "type" : " ",
  "cfStandardName" : "",
  "thesaurusConcat" : " ",
  "thesaurusVariable" : {
    "code" : " ",
    "name" : {
      "en" : " ",
      "fr" : " "
    },
    "thesaurusVariable" : {
      "code" : " ",
      "name" : {
        "en" : " ",
        "fr" : " "
      },
      "thesaurusVariable" : {
        "code" : " ",
        "name" : {
          "en" : " ",
          "fr" : ""
        },
```

```
        "thesaurusVariable" : {
         "code" : "NULL",
         "name" : {
          "en" : "",
          "fr" : ""
         },
         "thesaurusVariable" : null
        }
       }
      }
     }
    },
    } ],
    "instruments" : [ {
     "thesaurusConcat" : " ",
     "thesaurusClass" : {
      "code" : " ",
      "name" : {
       "en" : " ",
       "fr" : "  "
      },
      "thesaurusCode" : {
       "code" : "NULL",
       "name" : {
        "en" : "",
        "fr" : ""
       },
       "thesaurusName" : {
        "code" : "NULL",
        "name" : {
         "en" : "",
         "fr" : ""
        },
        "longName" : {
         "en" : "",
         "fr" : ""
        }
       }
      }
     }],
     "manufacturer" : " ",
     "model" : " ",
     "serialNumber" : " ",
     "calibration" : " ",
     "resolution" : null,
     "displayName" : " ",
     "description" : null
    } ],
    "platforms" : [ {
     "thesaurusConcat" : "BALLOONS_ROCKETS.NULL.BALLOONS",
     "thesaurusClass" : {
      "code" : "BALLOONS_ROCKETS",
```

```
    "name" : {
     "en" : "Balloons/Rockets",
     "fr" : "Ballons/Fusées"
    },
    "thesaurusCode" : {
     "code" : "NULL",
     "name" : {
      "en" : "",
      "fr" : ""
     },
     "thesaurusName" : {
      "code" : "BALLOONS",
      "name" : {
       "en" : "Balloons",
       "fr" : "Ballons"
      },
      "longName" : {
       "en" : "",
       "fr" : ""
      }
     }
    }
   },
   "name" : "",
   "description" : {
    "DEFAULT_VALUE_KEY" : ""
   }
  } ],
  "projects" : [ {
   "thesaurusConcat" : "INTERNATIONAL_PROGRAMS.HEMERA",
   "thesaurusCategory" : {
    "code" : "INTERNATIONAL_PROGRAMS",
    "name" : {
     "en" : "International programs",
     "fr" : "Programmes internationaux"
    },
    "thesaurusName" : {
     "code" : "HEMERA",
     "name" : {
      "en" : "Hemera h2020",
      "fr" : "Hemera h2020"
     },
     "longName" : null
    }
   },
   "aerisProjectUuid" : " "
  } ],
  "datasetSpecification" : null,
  "documentRating" : 0,
  "note" : 10,
  "metadataLevel" : "Collection",
  "identifier" : " ",
```

"type" : "COLLECTION"
}