# Milestone M9.1: *Definition of new EUROCHAMP DC Architecture*

## Author(s): ANDRE, François

| | |
|---|---|
| **Work package no** | **WP9** |
| **Milestone no.** | **M9.1** |
| **Lead beneficiary** | |
| **Deliverable type** | ☑ R (document, report)<br>☐ DEC (websites, patent fillings, videos, etc.)<br>☐ OTHER: please specify |
| **Dissemination level** | ☑ PU (public)<br>☐ CO (confidential, only for members of the Consortium, including the Commission) |
| **Estimated delivery date** | **M4** |
| **Actual delivery date** | **31/03/2017** |
| **Version** | **V0.1** |
| **Comments** | |

# TABLE OF CONTENT

## DOCUMENT REVISIONS

| Date | Modifications | Author |
|---|---|---|
| 22/03/2017 | Document creation | FA |
| 29/03/2017 | Minor modifications | BPV |
| 30/03/2017 | V1.0 Final preparation | FA |

# 1. OBJECTIVES

This document aims at describing the general architecture of the Eurochamp data centre. Currently, this document only covers the two databases existing in Eurochamp 2:
The Database of Atmospheric Simulation Chamber Studies (DASCS): This database provides a compilation of experimental and modelled data obtained from experiments in simulation chambers supplied by all partners of the consortium.The Library of Analytical Resources (LAR):It includes infrared and mass spectra.During the project, a new database will be implemented in the Data Center, a Library of Advanced Data Products (LADP). This database provides different types of mature and high level products of chamber experiments.

# 2. COMPONENTS

The 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 aggregates services coming from functional components or provides services to functional components.

This list indicates the components of the data center:

| 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 Authentification | 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 |

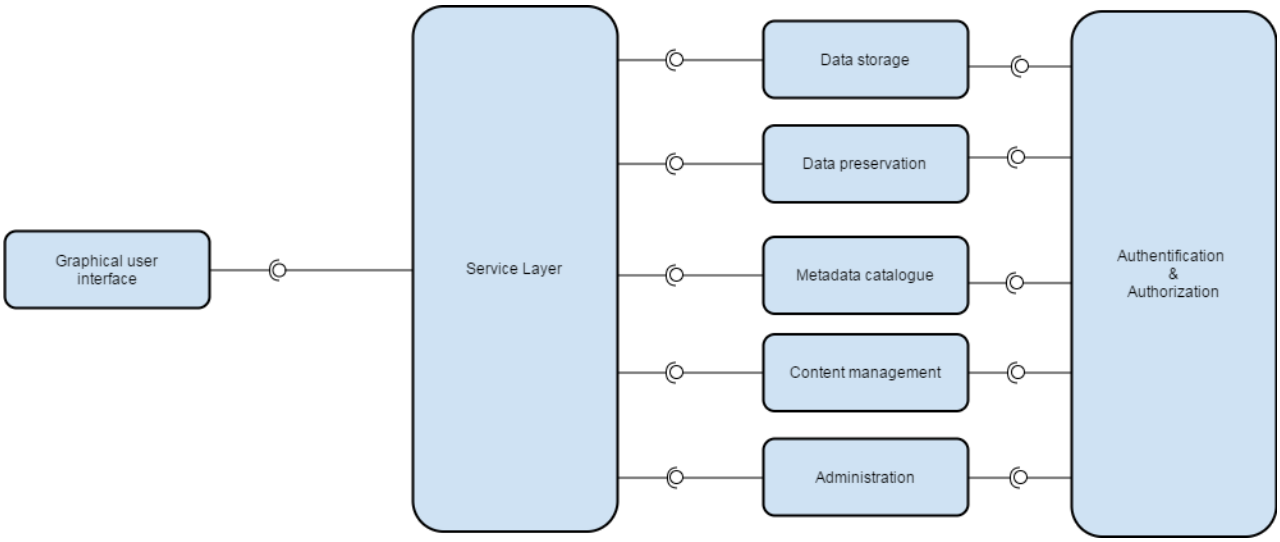



Figure 1: general scheme

# 3. GLOBAL CONSIDERATIONS

## 3.1. EUROCHAMP DATA PROPERTIES

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

- Fixed sensors,
- Short temporal extensions: Eurochamp datasets aren't long time ongoing series,
- Small volumes,
- Not georeferenced,
- Isolated datasets: Datasets aren't linked with other datasets (except for data in LADP which may be linked to data in DASCS).

## 3.2. DATASET 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 centers.

Hence, the data center 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

Within Eurochamp 2 project, datasets were identified by two elements: the name of the database and an incremental number (1,2,...). These identifiers will be converted into UUIDs 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 center 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 center will use a server provider by AERIS-ESPRI.

The file server will communicate with other components of the data center via FTP protocol with a dedicated user authorized to access to the root of the database.

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 center and to enable richer services, such as file validation or download measurement.

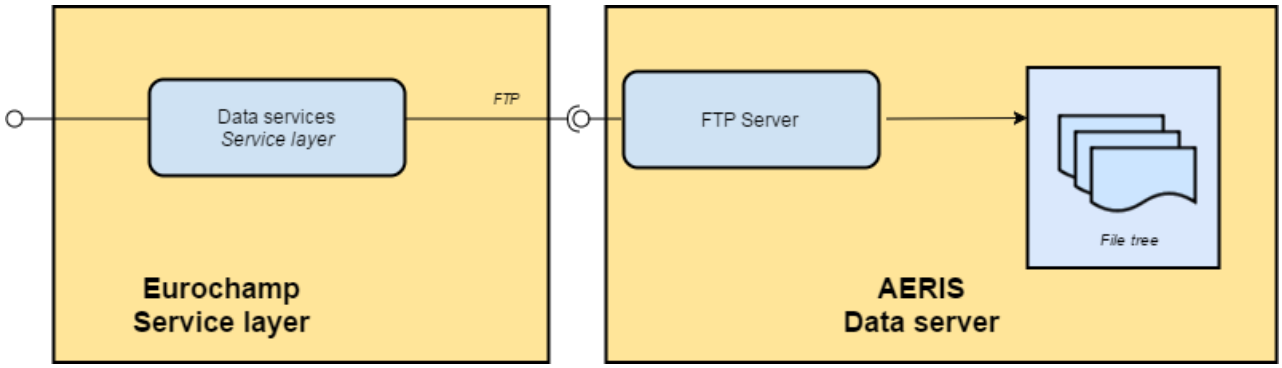



Figure 2: general scheme

### 4.1.2. File tree

In his first version, the root folder of the database will contain two main subfolders, DASCSC (i.e. *chamber experiments)* and LAR (i.e. *analytical library)*, which 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.

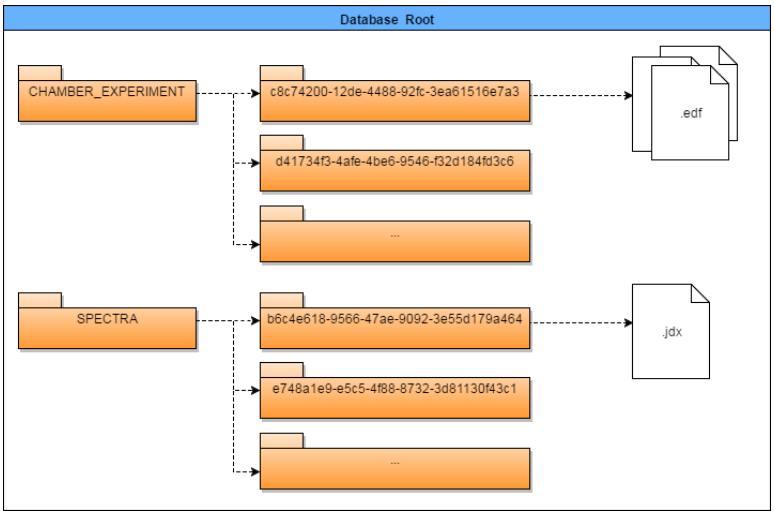



Figure 3: File tree

#### 4.1.3. File format

##### a) Chamber experiments

Datasets will consist in,

- One or several EDF file,
- One description.pdf file.
- Optional auxiliary files

The EDF format is described here: ___To be completed with a link to a page on the new data center___

##### b) IR and mass spectra

Datasets will consist in

- One JCAMP file

The JCAMP format is described here : ___To be completed with a link to a page on the new data center___

#### 4.1.4. File names

P.I. are free to choose the name for the file they provides 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 data center won't handle dataset versionning. Only the last version of each file will be stored.

# 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.

### 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.

### 5.1.1. 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.

Obviously, each deletion will be also impacted on the AERIS-SEDOO infrastructure.

## 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

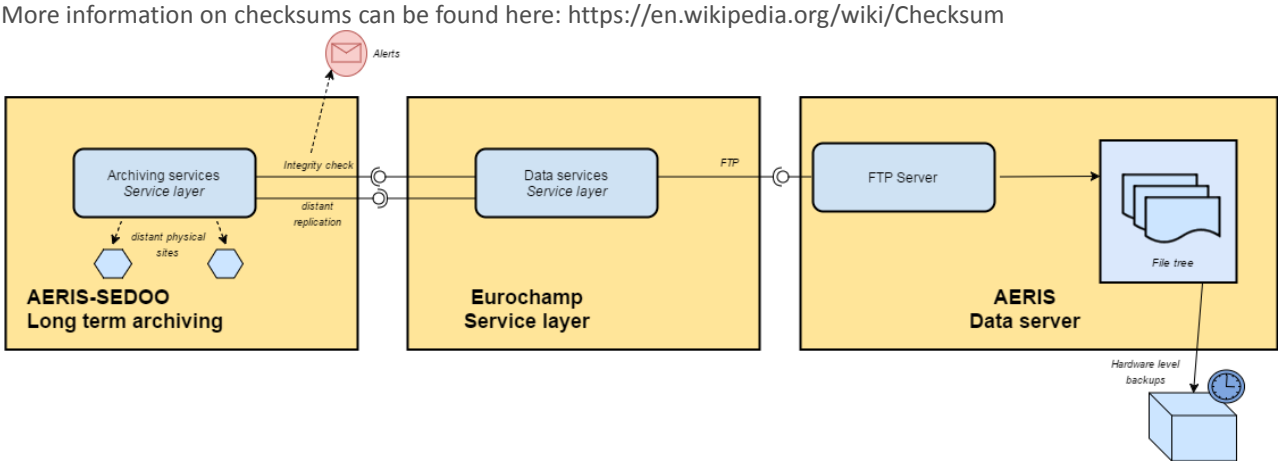



Figure 4: general scheme

## 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 center 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).

Eurochamp metadata profile will extend Inspire profile to include specific information (chamber types, institute...). The precise profile will be defined during the first months of the project.

### 6.2. CATALOGUE

The data center 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 Eurochamp. 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).

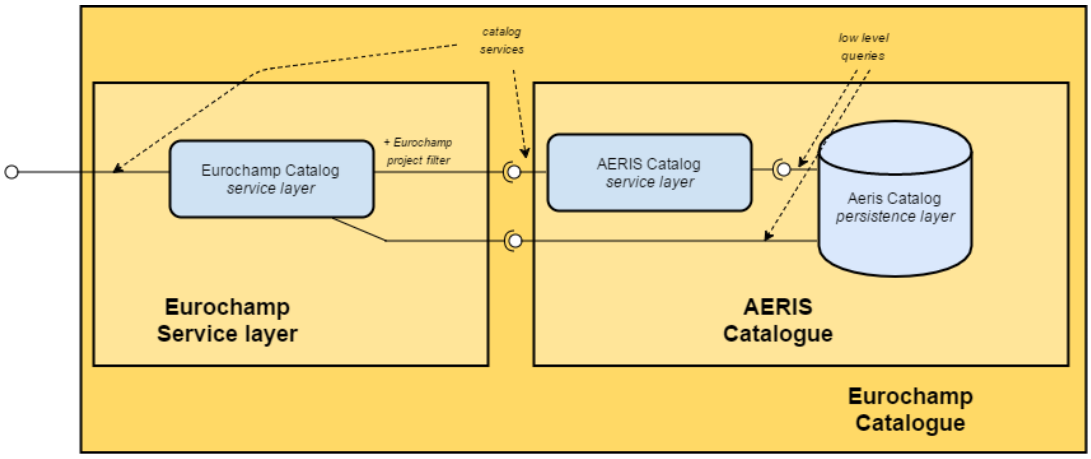



Figure 5: general scheme

## 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.

Eurochamp data center will rely on an 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 center architecture with a dedicated server and database and a direct link with the user interface.

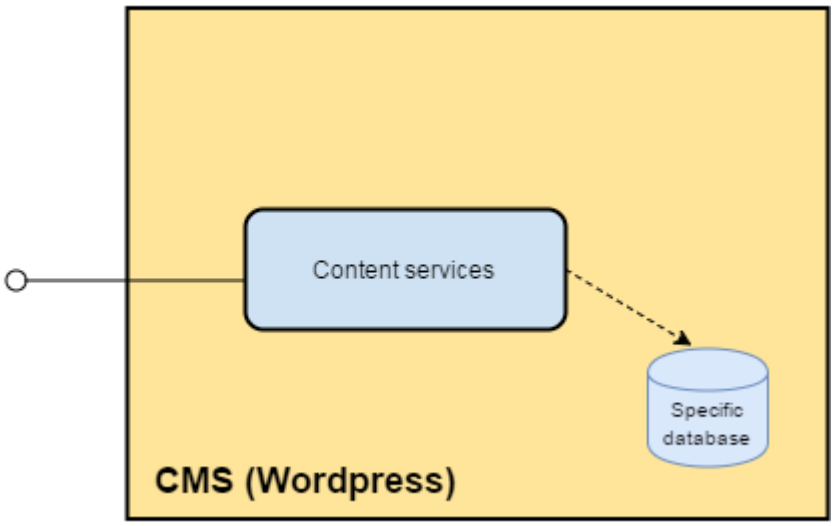
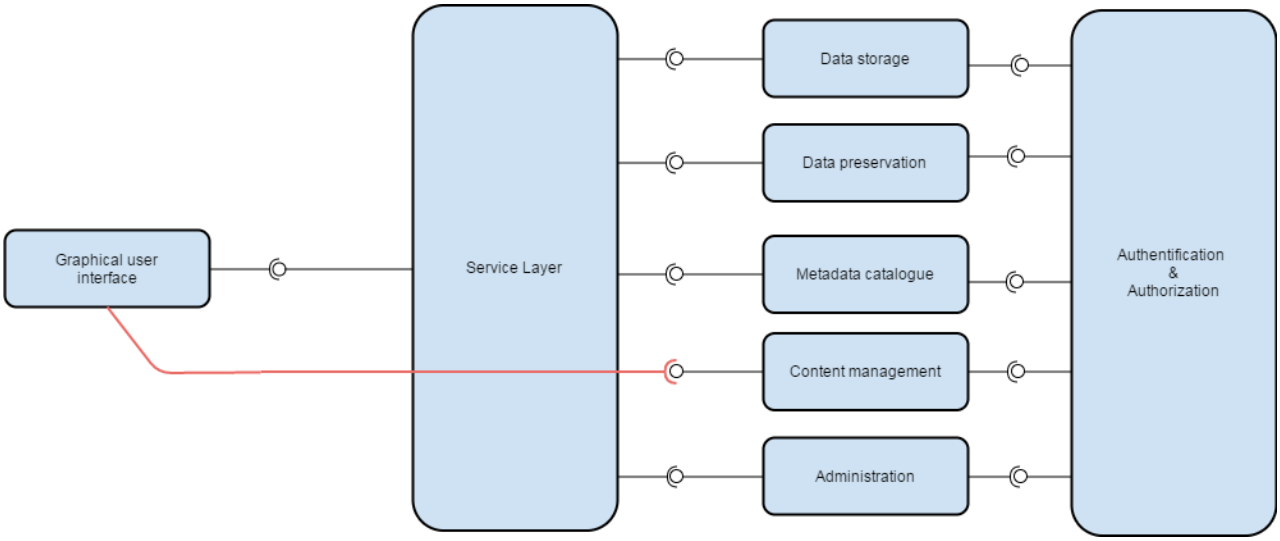



Figure 6: general scheme

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




## 8. 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.

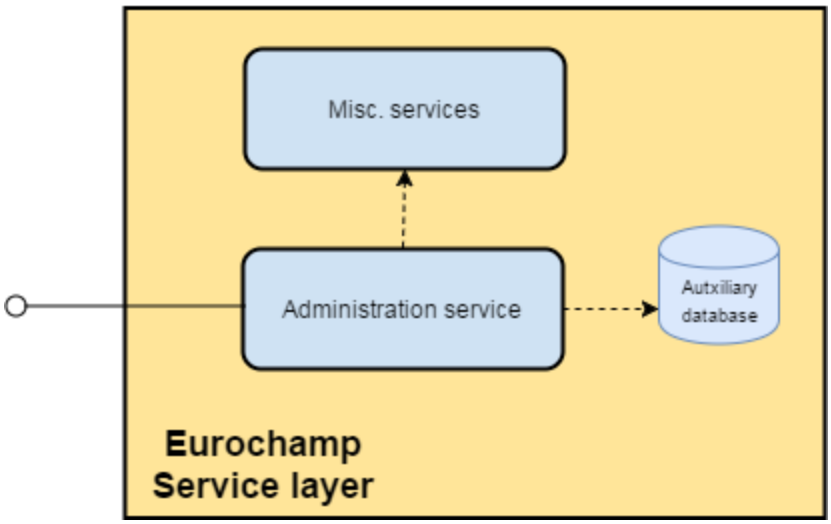



Figure 7: general scheme

# 9. AUTHENTICATION & AUTHORIZATION

All data center 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.

## 9.1. AUTHENTICATION

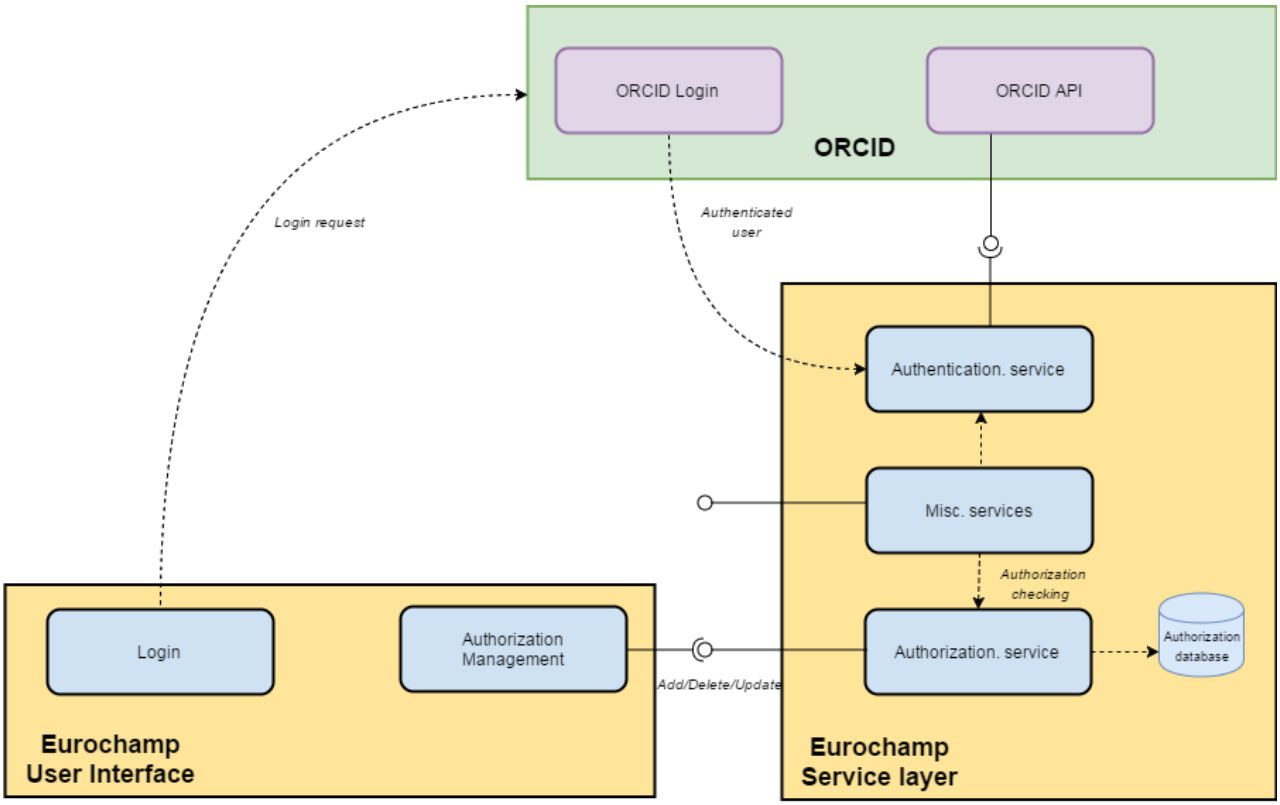
For a long time, the first reflex of software developers has been to implement authentication internal, with a local data system. This choice has the advantage of seeming more secured but has the drawback of implying a complex account management. With the appearing of wide spread social applications (Facebook, Google, ...), the possibility to log on an application with this accounts has been appreciated by both users and administrators.

In consequence, the Eurochamp 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.

## 9.1. 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.




## 10. SERVICE LAYER

The service layer aims at providing modularity and clear role separation which are important to ensure maintainability and evolutivity of the data center 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.

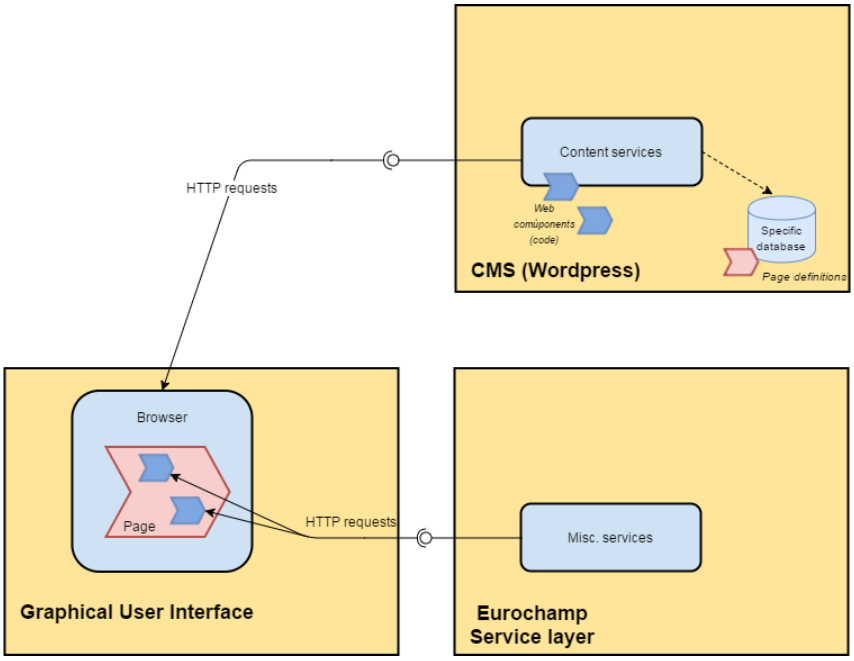



Figure 8: general scheme

# 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).

Indeed, for instance, 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 center.

## 12.1. ORCID

ORCID (Open Researcher and Contributor ID) is a nonproprietary 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 Eurochamp 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.

However, in spite of all these standards, metadata interoperability isn't that obvious. Indeed 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.* It is more realistic to allow *metadata being interoperable with specific partners*.

As mentioned, internally, Eurochamp 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 center 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 center will offer a default converter to translate AERIS internal metadata format to ISO19139 for any metadata in the INSPIRE core.
  - The data center will offer CSW containers dedicated to each partner who want to harvest Eurochamp 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.

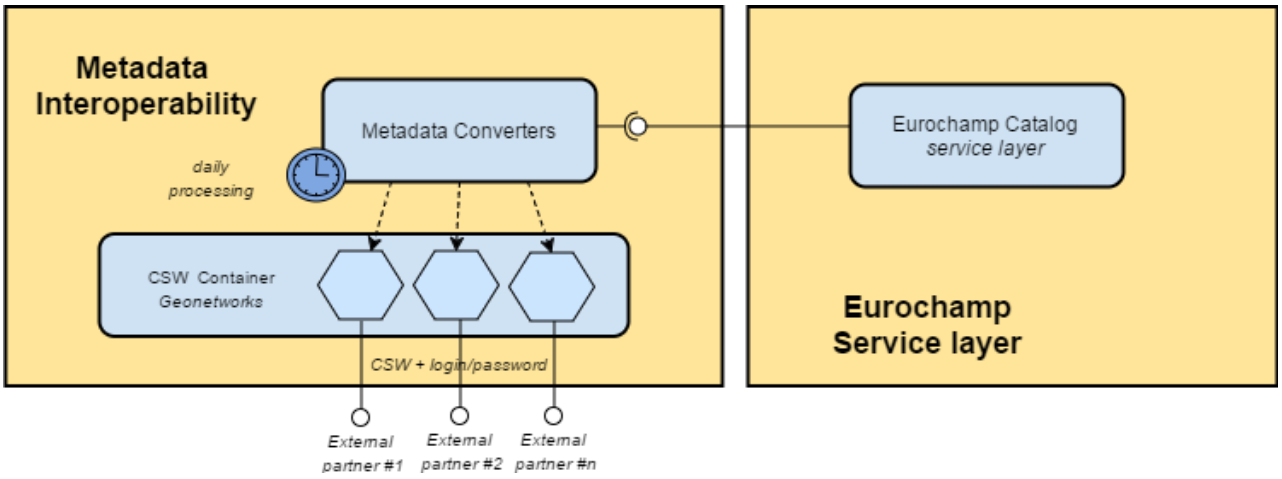



Figure 9: general scheme

## 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.

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. |

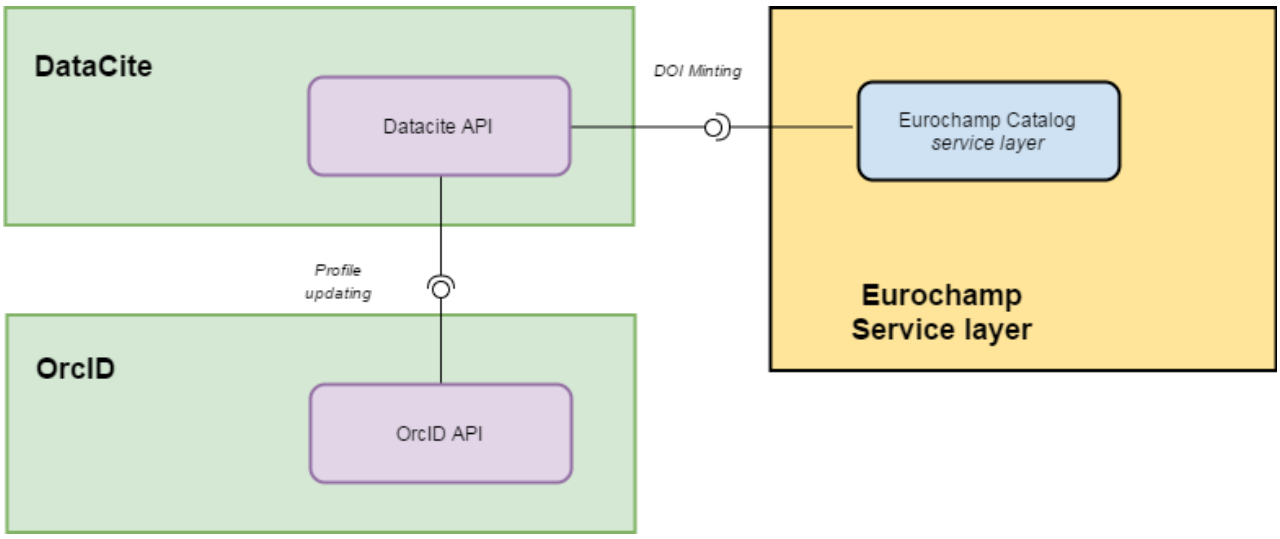



Figure 10: general scheme

## 12.4. DATA INTEROPERABILITY

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