# CNES clearinghouse (SERAD project)

**Danièle Boucon(1), Richard Moreno(1), Dominique Heulet(1), Martine Larroque(1)**

*(1) CNES*

*18 av E. Belin, 31401 Toulouse Cedex 9, France*

*EMail: Daniele.Boucon@cnes.fr*

## ABSTRACT

SERAD (Service for Data Referencing and Archiving) is a project that will allow CNES to dispose of a centralized repository structure that collects, stores and disseminates information and metadata on Data that are under CNES responsibility. The objective for CNES is to better handle and to improve the access to this data patrimony. The objective of this paper is to present in detail the referencing part of the project; in other words, how to create and maintain a centralized and standardized metadata repository that catalogues data within the CNES context and addresses different scientific user communities. In particular the following points will be addressed: the use of ISO19115 standard to describe a large variety of space scientific domains (Earth Observation as well as Universe Sciences), the use of a "off the shelf" product as a clearinghouse, the communication with other clearinghouses and interoperability, the semantic search criteria based on thesaurus, the preliminary feedback.

Keywords: metadata, scientific data, referencing, ISO 19115, clearinghouse, access

## INTRODUCTION

SERAD (Service for Data Referencing and Archiving) is a project that will allow CNES to dispose of a centralized repository structure that collects, stores and disseminates information and metadata on Data that are under CNES responsibility. The objective for CNES is to better handle and to improve the access to this data patrimony.

In order to achieve this goal, it is mandatory to identify all data which are relevant and to verify whether these data are properly archived ; if not, then, to proceed to the archiving of these data.

The SERAD mission is then the following (see figure 1):

- to constitute and maintain an open and centralized metadata repository of all data under CNES responsibility, the SERAD clearinghouse
- to do, when necessary, the archiving of data
- to survey the data production centers in order to guaranty the long term preservation of these data even if (critical case) one of these centers has to be closed

This system will be built upon existing tools that will be customized:

- MDweb, which is an open source tool for cataloguing data, not limited to geographic information, but able to handle data from any other scientific domain
- SIPAD-NG which is a generic tool allowing to give a full access (with search criteria) to any kind of archived data (except files in database)

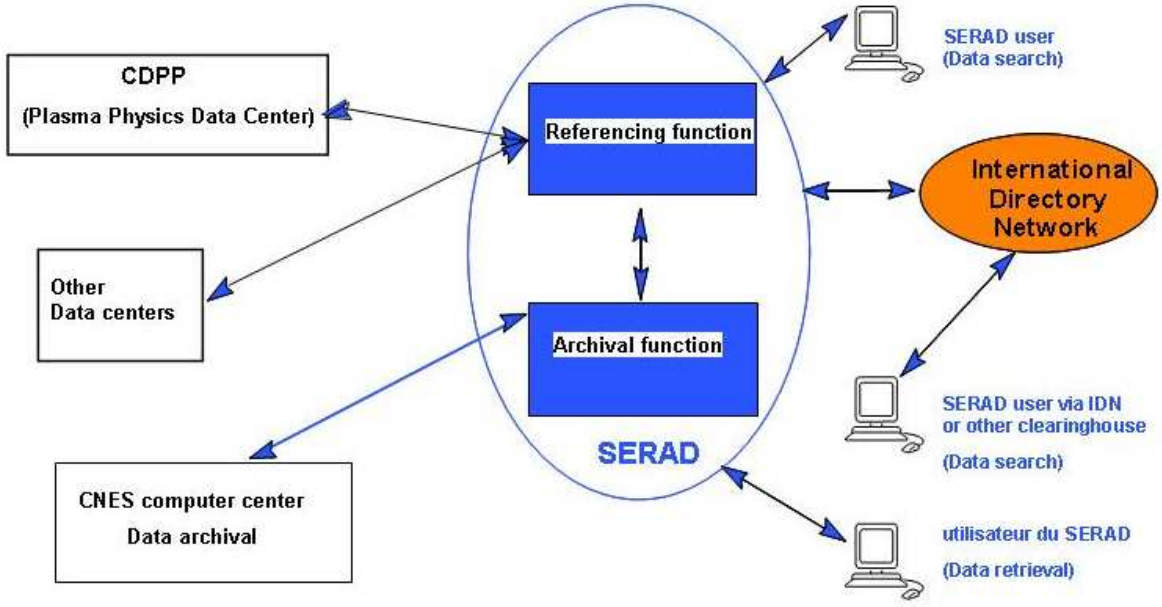


Figure 1: general SERAD overview

This paper presents in detail the SERAD clearinghouse.

## INTRODUCTION TO THE SERAD CLEARINGHOUSE

The SERAD clearinghouse is a repository structure that stores and disseminates metadata on space scientific data. Metadata –fully described later- is used here in the sense of descriptive information given on the referenced data, in order to discover or search sets of data of interest.

The user community includes researchers, scientists, CNES engineers as well as general public.

**The originality and the challenge of the SERAD system is to cover all space domains without restrictions: Earth Observation, Universe Sciences, and Microgravity (Life Sciences and Material Sciences).**

The heritage mission on the long term and the ambition is also to cover **all missions without time limitation**. That means that the SERAD clearinghouse deals with recent as well as very old missions (**the extent today is more than 30 years**).

The objective is to assist users to efficiently locate information on available data on a specific scientific domain, or through different domains. **It will also provide a centralized and overall view on the CNES data patrimony on the long term.**

The clearinghouse provides widespread access to the metadata without access rights. It also gives the link to the server of the data addressed in the metadata. At the end, it is up to the user to fetch data of interest on this pointed data server. The SERAD clearinghouse does not intent to give direct access to the data.

An implicit objective is to show the existence of the data in order that they may be used on the long term. This is the best way to “keep them alive” and ensure their preservation.

## WHAT KIND OF REFERENCED DATA?

As previously said, the data addressed by the clearinghouse metadata are scientific data, not only in Earth Observation, but also in Universe sciences and Microgravity domains. That’s why a classification has been created to address more precisely each space mission.

A list of missions candidate to referencing has been established and is regularly maintained with new missions. There is a large diversity of concerned missions and experiments, depending on time period,

CNES involvement, and the known information on archived data. The data involved here may or may not be archived at CNES.

Because it will need time to take into account all of them in the clearinghouse, priorities should be done among them.

**One of the challenges today is to build a pertinent and sufficient metadata database, available for the opening of the system, and of interest for the user community.**

## Data over time

For 40 years, in CNES, a big number of space missions have been producing a huge amount of data (hundreds of To). These data constitute a valuable heritage that must be preserved because many of them are:

- unique when related to an event that will never happen again or in a very long time (e.g. Halley comet period is 76 years !)
- integrated in long cycles of observations, including cycles for climate change observation
- mandatory to prepare future missions (e.g. GAIA takes benefits for HIPPARCOS experience)

With the arrival of new missions, this amount of data will further increase in volume and complexity

This is mainly due to:

- geographical dispersion of data: the missions are often conducted with partners and the data produced are not necessarily stored at CNES; when they are, they are not always in the same place
- variety of scientific domains (universe sciences, earth sciences, life sciences, microgravity, etc.)
- variety, at a given time, of data production and of data preservation techniques

As a consequence, the level of information on experiments and datasets may vary considerably from one mission to another. This could be a difficulty for the required quality level of the metadata associated to these data.

Some data are directly usable, others require more investigation and/or treatments (that's one of the other SERAD missions, not treated here).

**The number of concerned missions is around 100** (and this number will regularly grow).

All are candidates for referencing, but with an appropriate priority level defined in a plan. This level mainly depends on the scientific interest and the age of the mission (recent missions will be preferred).

That's why we distinguished 3 missions levels: "completed" –no more treatment are made on the data, "ongoing" –data continue to arrive from the mission-, "future"-no data yet available-.

## Data classification

**The system leans on the following classification** to assign the missions, to organize each experiment or dataset., and to define an appropriate set of related keywords. These domains and keywords are used to qualify the data.

These main domains have been divided into two other hierarchical levels, also called "Topic Categories". The following figure gives the example of "Earth Observation" classification tree. The missions are finally assigned to the appropriate tree leaves. For example ENVISAT, JASON1 to 3, PLEIADES, SPOT1 to SPOT5 are the missions related to the sub domain "Glaciology" (see figure 2). The colors correspond to "completed" –blue, "ongoing" –green, "future" –red. Of course, the same mission could be assigned to different tree leaves, depending on the imbedded experiment. The next step will then be to identify the experiments and datasets under consideration. This is then the work of the data inventory.

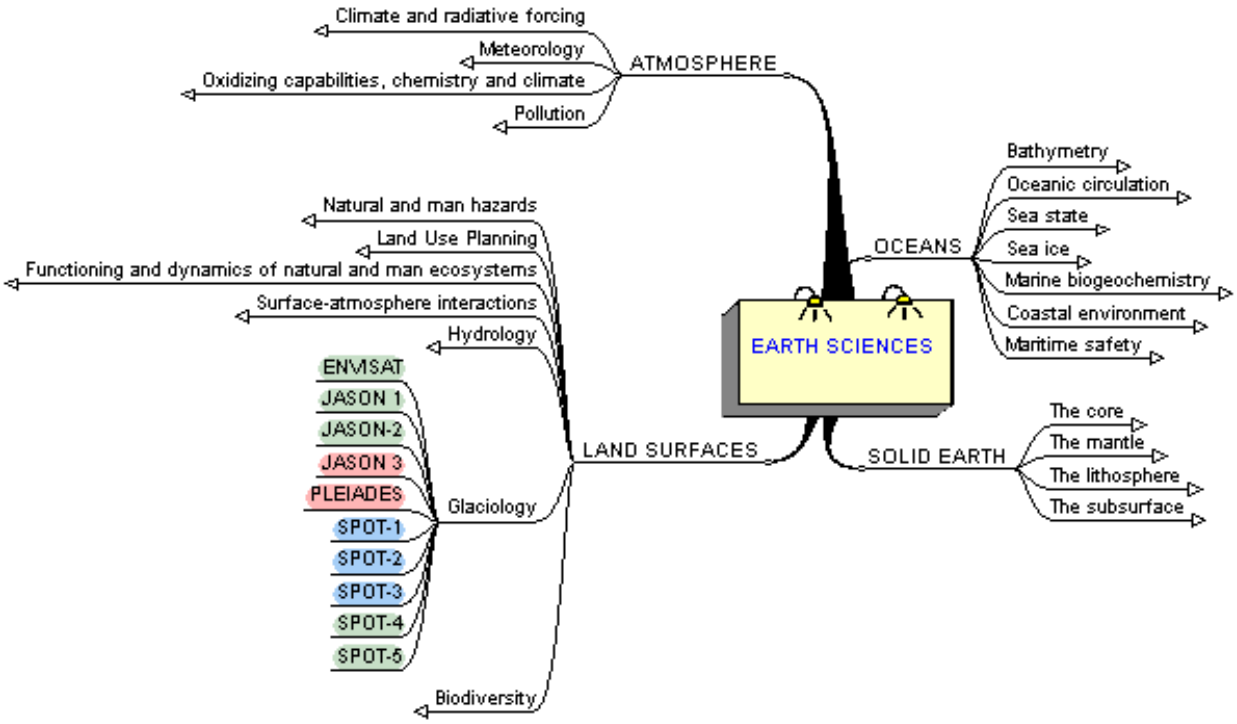


Figure 2: Earth Observation classification

The classification is linked to CNES space domains of interests, but links have been established with other classifications, such as the IDN[1] one. For example, the SERAD "Glaciology" sub domain corresponds to the IDN "Frozen Ground" sub domain.

This terminological consistency is important on the interoperability point of view, to be able to exchange metadata with other systems/organizations in the future.

## Data inventory

**The SERAD data inventory is the centralized repository of all known information on missions in the scope of the SERAD project.**

The content of this inventory is under construction and perpetually evolves (new missions, and updates of old ones). Today it includes more than 50 missions (of the 100 planned), 80 % of which are completed, 20 % are ongoing. The oldest mission dates from 1978 (GEOS mission).

Considering the "data over time" paragraph, 3 inventory levels are distinguished in the inventory, directly related to the level of information required to instantiate it, thus minimizing the maintenance effort: "complete" inventory (for completed missions and data archived at CNES), "minimal" inventory (for ongoing missions, and data not archived at CNES), "reference only" inventory (missions for which only referencing is required). Of course this evolves regularly, for example one "reference only" mission today may be turned to "complete" later with the knowledge of all necessary related information.

**The inventory is based on an organizational model.** It is divided into different parts, and **includes information required to create the clearinghouse metadata,** at each place where referencing is required (mission, experiment, set of experiments or datasets, datasets).

One of the advantages of this inventory is its total independence of systems or specific software. It is made up of XML files, humanly readable, easy to maintain, and **usable over the long term**.

[1] International Directory Network, http://idn.ceos.org/

The data inventory regularly evolves, and is maintained at least once a year.

## SERAD METADATA

Metadata are pieces of information describing resources. Resources are usually data sets but may also be aggregations of data sets, software files, etc.

Initially, in order to create a database that covers a large set of missions for the clearinghouse opening, the SERAD decided that the resources would be **space missions, and data experiments or collections (consistent set of data sets)**, those described in the Data inventory. The space mission metadata level has been created to provide information on the production context of the data. In the clearinghouse, metadata on experiments or collections are linked to space missions metadata by a "parent identifier" link. It should be noted that each experiment or dataset of a mission will not necessarily be referenced, for different reasons (if the experiment or dataset is deemed unusable or irrelevant, or if the referencing information is not yet available).

In a second phase, the data set level may be considered.

Among many metadata standards (Dublin Core, DIF, FGDC, ISO 19115, …), the standard chosen for this clearinghouse was **ISO 19115** [2], because it is to date the most complete, and is a superset of the DIF and Dublin Core.

ISO 19115 is dedicated to the description of geographic data, but has been slightly expanded for the domains of Universe Sciences (for example with an astronomical coverage instead of the spatial coverage) and Microgravity. The ISO19115 is implemented by an XML schema based on the ISO 19139 Technical Specification.

Metadata provide all the necessary information for resource discovery, resource locating, resource usage, resource assessment, resource retrieval. Metadata are no more than text files. Metadata content obviously depends on the resource the metadata refer to. However, given some type of resources and some level of description details, there is usually only one single way to describe a resource because all these resources are designed to meet the needs of a designated community.

The challenge is then to populate this content with information at the right level of complexity and with the correct terminology, in order that software can **find the correct metadata easily and effectively**, and in order that these metadata be **understandable** by the future user.

The SERAD metadata model is included in the data inventory model. The metadata model is quite rich, although only a few information elements are mandatory, including those mandatory for ISO 19115: the title, the abstract, the resource language, the metadata responsible, the metadata publication date, the resource creation date.
Metadata are structured as follows (only the two first categories will always be found, the others will depend on the characteristics of the resource):

- Metadata Information: management of metadata (e.g. File Identifier, Metadata Language, Metadata Standard Name, Reference System name)
- Identification Information: identification and characterization of the resource (e.g. Title, Alternate Title, Abstract, Purpose, Keywords, Topic Categories, Temporal Coverage, Geographical Coverage, Spatial Resolution, Access Constraints, Content Type, Points of contact)
- Data Quality Information: quality and quality assessment of the resource. Descriptions of various quality assessment measures and quality results can be found here, as well as the description of the process that has led to the resource (e.g. resource lineage)
- Content Information (e.g. Information on Wavelength)
- Distribution Information: how to access the data (link towards the data server)
- Resource Acquisition: process by which the resource is acquired (platform and instrument names)

Furthermore, the SERAD metadata are **INSPIRE** compliant: the metadata required by this directive are included in the model. In the case of Earth Observation, the SERAD will also take into account the SBA – Social Benefit Area- and Community of practices (http://www.earthobservations.org/geoss.shtml).

A thesaurus has been created for the SERAD (see also [3], "Searching Data"). It contains the metadata keywords, organized in domains and sub-domains as defined in the Data classification. This thesaurus is compliant with the ISO 2788 standard [1] and takes into account the relationships "Broader term, Narrower term, Related term"[2]. Many experts were involved in this work, to insure its consistency with the terminology of user communities. This thesaurus will continue to evolve, depending on resources description. The thesaurus is in RDF/SKOS format.

As previously written, **all descriptive information is located in the inventory**. Thus, the metadata are directly extracted by script from the data inventory in an XML form, at the level they have been defined. The metadata base thus inherits the properties of the inventory (system independent, humanly readable), and will take benefit of a long term usage. One mission inventory generates at least 2 metadata (the mission plus at least one experiment), and may generate as many metadata as required by the SERAD.

Metadata are created upon request and are stored in a repository, before ingestion in the referencing tool.

## INTEROPERABILITY

The SERAD clearinghouse is able to exchange metadata with other clearinghouses, thanks to the use of standards:

- ISO 19115 metadata standard. As it is a superset of the DIF and DUBLIN CORE metadata standard, it is then easy to transform the SERAD metadata to one of them by an XSLT transformation (for example the IDN clearinghouse uses the DIF form).
- OGC CSW –Catalogue Services for the Web (implemented in the referencing tool). It gives the opportunity to retrieve metadata to compatible systems, or to let other organisations harvest the CNES ones. To promote data knowledge between organisations or even within CNES, this service may be used in the future.

In the case of resources concerned by the directive, the metadata are also INSPIRE compliant.

## THE PROCESS

The general process, from the metadata creation to the final access to data of interest, including interfaces with external systems, is described figure 3. This process is based on independence between the different functional entities, the use of a generic tool and centralized services, and a knowledge base designed for the long term. These principles will facilitate maintenance and impact of the change of a function on other functions. The SERAD functional entities are described as follows:

- "Metadata creation and knowledge base": the classification, the inventory, the thesaurus and the models are managed by the SERAD apart. They are enriched according to the resources needs.
- "Search Metadata and get results": this is represented by the referencing tool MDweb. It receives Metadata from the previous functional entity to update the catalogue. It provides the user search and gives access to metadata describing the data of interest.

[2] Broader term: the term is the name of the broadest class to which the specific concept belongs.

Narrower term: the term refers to the concept with a more specific meaning.

Related term: the term is associated, but is not a synonym, a quasi-synonym, a broader term or a narrower term.

- "Data Access": the interface between MDweb and the data server is made by the data server URL link in the metadata. SERAD will have its own data server, which is the SIPAD-NG[3] generic tool. Other data may be archived in different places around the world, and managed by other organizations.
- "Data Archival" (cited here but not part of the clearinghouse): at CNES, data are archived in the centralized STAF[4] system, directly accessed by SIPAD-NG. Data may also be archived in other systems.

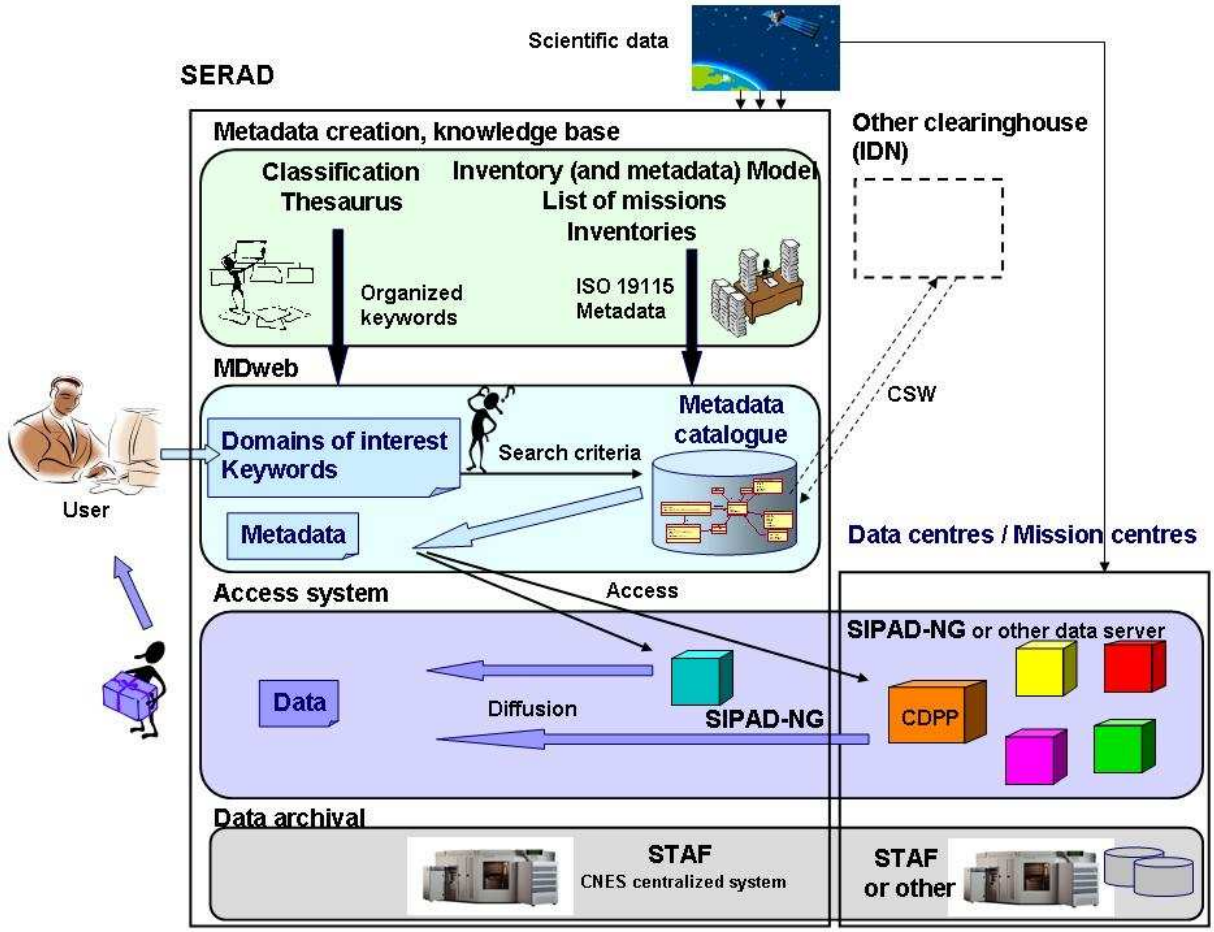


Figure 3: the referencing system process with external interfaces

## THE REFERENCING TOOL

For many years, CNES made its own experience by creating a complete and specific clearinghouse prototype, the BDMS ("Bureau des Métadonnées et des Services"). Starting from this, CNES recently chose to use MDweb ("MetaData web"), a COTS –Commercial Off-the-Shelf-, in order to reduce development and maintenance cost and to use up to date technologies. MDweb is an open source tool[5] for cataloguing and locating resources (data, documents and services) that are accessible over the web. It is

[3] Système d'Information, de Préservation et d'Accès aux Données – Nouvelle Génération (Information System for Data Preservation and Access – New Generation).

[4] Système de Transfert et d'Archivage de Fichiers (Long term data archiving system)

[5] Open source licence LGPL V3

based on current geographic-information metadata (ISO 19115, 19119, 19139) and communication (OGC's CSW-2) standards and conforms to the rules for implementing metadata and the associated discovery services of the INSPIRE directives.

This tool requires a set of configuration and minimal specific developments to adapt to CNES needs:

- five specific profiles: one profile per domain (not only geographic information) - Earth Observation, Universe Sciences, Life Sciences and Material Sciences (Microgravity). For Earth Observation domain, another profile exists to conform to INPIRE directives
- a user interface taking into account the different profiles, and enabling an easy metadata search by domain, or a transversal search among the domains
- a thesaurus editing function, according to ISO 2788 standard. This function will enable a SKOS format export, as well as a PDF format for document management
- in addition to classical search functions, a powerful semantic search module by keywords and relationships between keywords, based on the SERAD thesaurus, will be implemented

A large effort is made to provide the future **user with an efficient search**:

The simple metadata search is a **semantic approach**, based on indexed words extracted from metadata elements such as title, subtitle, abstract, purpose, keywords, and relationships between metadata keywords belonging to the thesaurus.

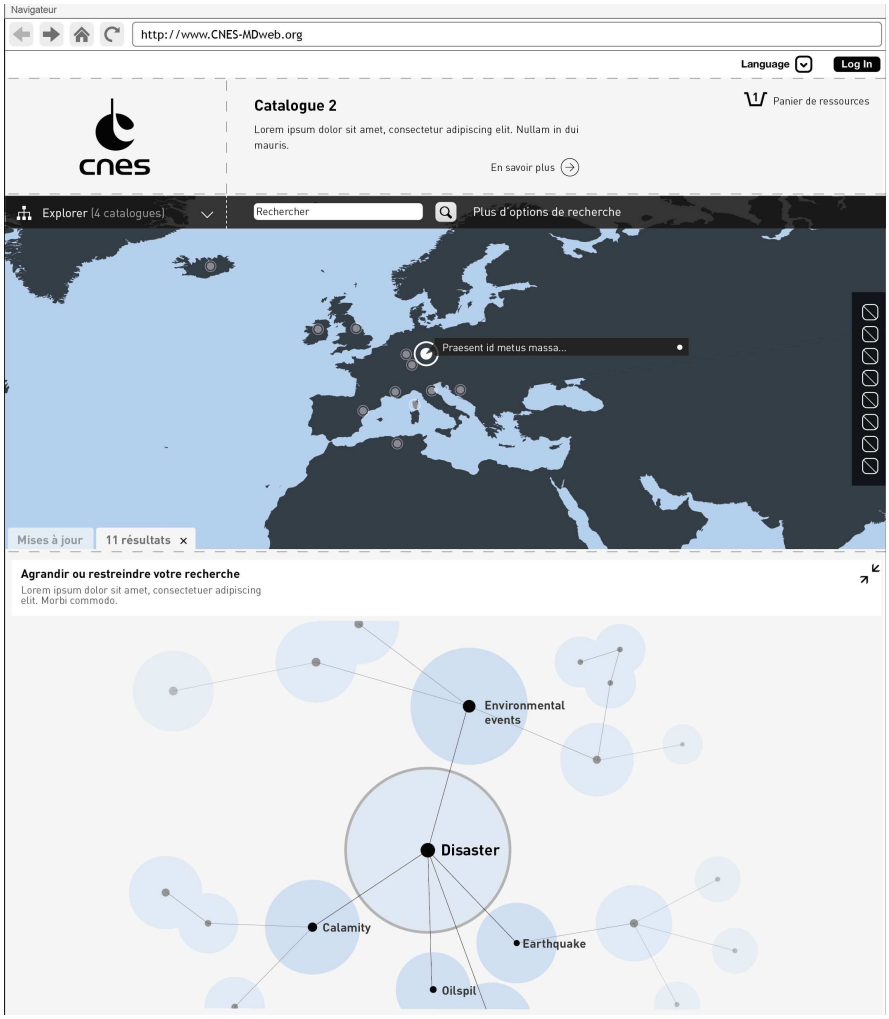


Figure 4: simple search with refinement based on keywords

The advanced metadata search is based on classical criteria such as temporal extent or geographical coverage, and also criteria such as instruments or platforms names.

Domain specificities such as astronomical coverage, stars map (Universe Sciences) or experiment name (Microgravity) are also covered.

The metadata are then available in an XML ISO 19139 or PDF document.

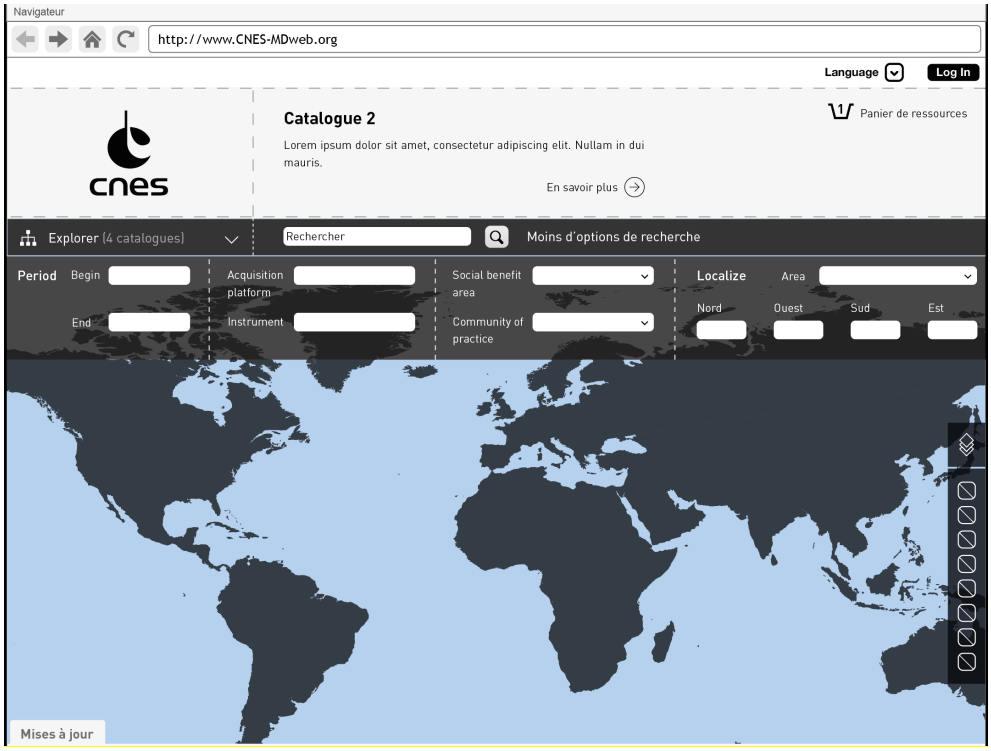


Figure 5: advanced search

MDweb includes as well a service for metadata harvesting, based on OGC CSW-2 specification.

Figure 6 shows MDweb architecture. The “Edition” module is used for metadata ingestion, and not for metadata creation (metadata are created outside the system). MDweb will be installed at CNES on a dedicated machine.

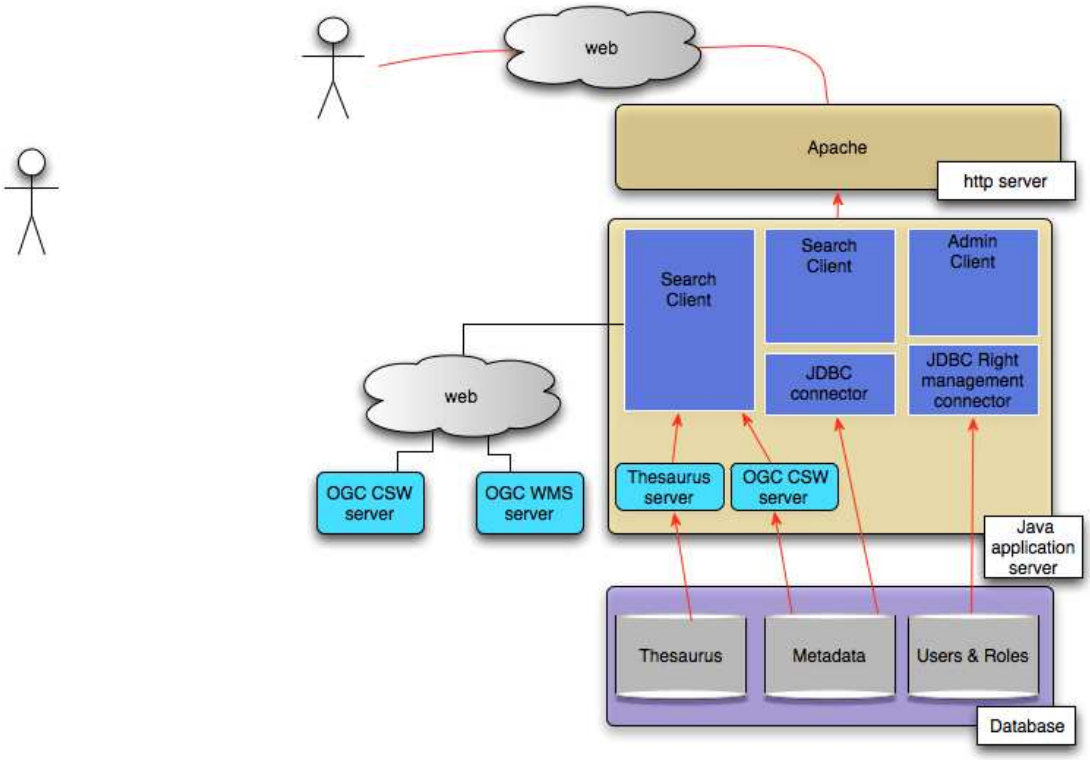


Figure 6: MDweb architecture

## PRELIMINARY FEEDBACK

- It needs a lot of time to build “good” metadata. The participation (and availability …) of experts is always required, in order to complete, validate and insure metadata quality and consistence.
- The right level of information is not always easy to estimate, in order to have general information, but yet precise enough. It is also important to limit potential updates (and minimize metadata repository maintenance) for the future. This means anticipating the user needs and behavior. This is made more difficult as the final users may be a scientist, an engineer, or any curious person.
- It is difficult and costs money to gather information on very old missions, when the associated points of contact are no more available. It should be kept in mind that this should be done during the life of the mission.
- The SERAD has to deal with space missions of different ages and of different scientific domain, which is a challenge.
- The terminology is a crucial point in such a system, specifically it covers different scientific domains and thus potentially different user communities. It took a long time to build a first version of the SERAD thesaurus. For the moment it contains around 1000 words, and the knowledge will continue to evolve regularly. This terminology must be coherent inside the same user community, but sometimes the same word has not exactly the same meaning through different domains!
- It is also important to limit metadata duplication. Currently different kinds of metadata are often created for different needs (data server access, data management, data exploitation …). Pieces of information are dispatched. To avoid recreating them, one solution is to fetch and automatically extract such information by scripts or filter.

## CONCLUSION

It is important to keep in mind that the SERAD is designed for the long term.

The MDweb specific functionalities are under development, and should be operational by end of 2011.

One of the challenges is to build a multi domains and consistent metadata repository by then, validated by experts of the domains. 50 missions have already been inventoried.

The referencing will concern experiments or collections in a near future, and potentially datasets in a longer term.

The resources addressed by the metadata are available using the links in the distribution part of the metadata. In general it is the URL link of the data server.

The SERAD clearinghouse opening is planned for end of 2012.

## REFERENCES

[1] – Documentation – Guidelines for the establishment and development of monolingual thesauri. International standard ISO 2788:1986.

[2] – Geographic information – Metadata. International standard ISO 19115:2003 and ISO 19115-2:2009.

[3] – CEOS WGISS Interoperability Handbook, issue 1.1, February 2008, www.ceos.org.