IDOC General guidelines for dataset management

IDOC-OD-009

Préparation

	Nom et Fonction	Date
Rédacteurs	Gilles Poulleau	Mai 2020
Vérificateur		
Approbateur		

Liste de diffusion

Nom	Fonction	Société

Evolutions

Edition	Date	Modifications
0.0	27/12/2017	1st draft
1.0	30/01/2019	1st released version
1.1	20/05/2021	Updates, still to be better reviewed

TABLE DES MATIERES

1   SCOPE OF THE DOCUMENT

This document is related to the « IDOC-OD-008 IDOC general guide for dataset management » which needs to be handled first.

This present document is applicable for dataset dissemination ie the service in which an application is built to allow users to access a dataset hosted at IDOC. This application can be a simple interface to

Applicable Documents

	Référence	**Titre**
AD1	IDOC-LI-000	IDOC item list

2   REFERENCE DOCUMENTS

	Référence	Titre
RD1	IDOC-EX-001	IDOC description executive summary
RD2	IDOC-OD-002	IDOC Risk analysis and management
RD3	IDOC-OD-003	IDOC General principles applicable to project design
RD4	IDOC-OD-004	IDOC Guidelines for new Data integration
RD5	IDOC-OD-005	IDOC Guidelines for new Pipeline Data Production
RD6	IDOC-OD-006	IDOC Guidelines for new archive long term preservation
RD7	IDOC-OD-007	IDOC Guidelines for new instrument operations
RD8	IDOC-OD-008	IDOC Guidelines for new services
RD9	IDOC-INF-009	IDOC General guidelines for dataset dissemination
RD10	IDOC-INF-010	IDOC Organigramme
RD11		REGARDS – A generic CATALOG ACCESS SYSTEM AND data VALORIZATION tool

3   GENERAL PRINCIPLES APPLICABLE TO DATASET DISSEMINATION

The dataset dissemination is related to the "Access" element of the OAIS model.
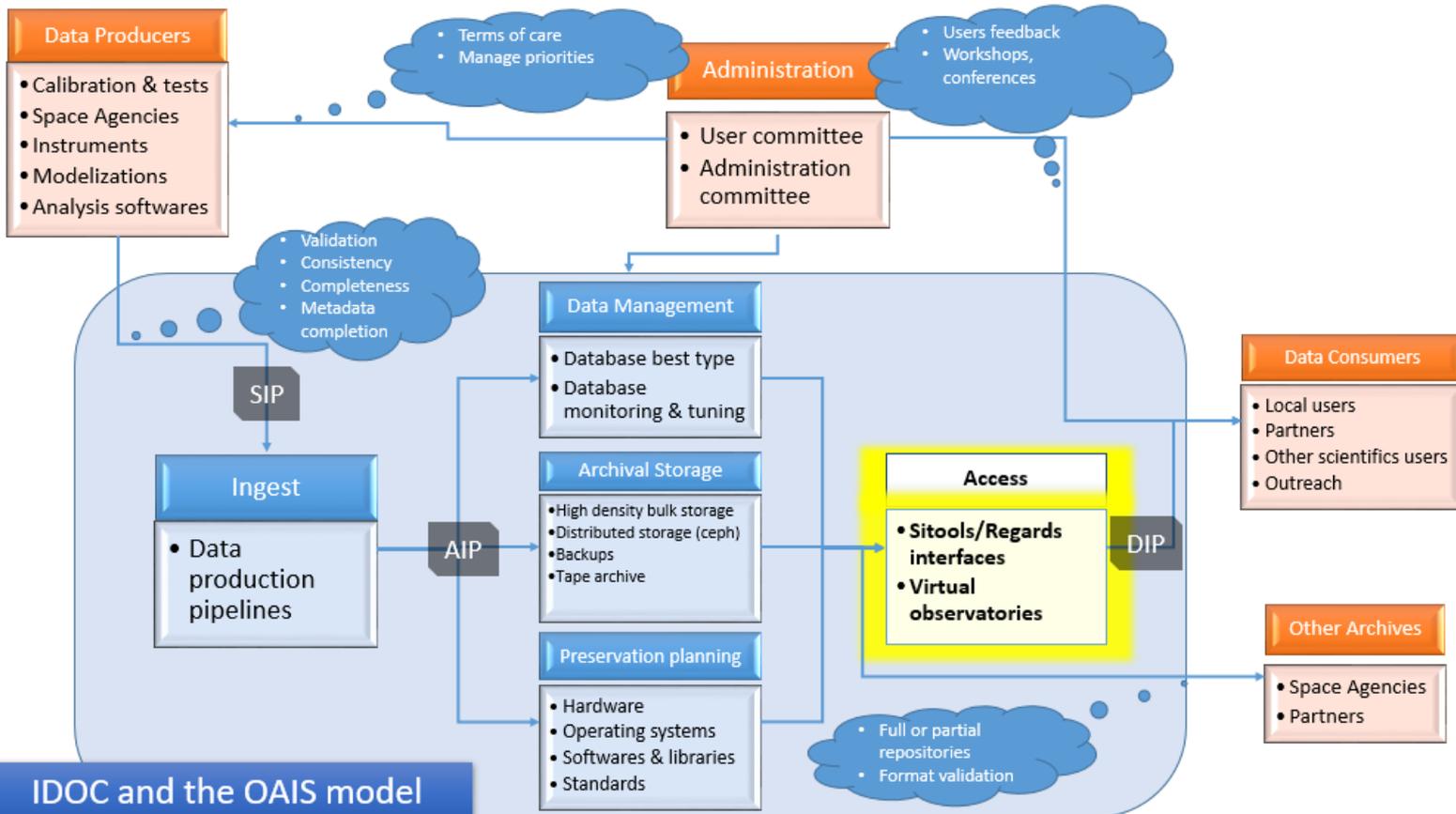
3.1   RESPECT OF DMP

Whatever the service IDOC implements, it has to be done in accordance with the terms agreed with the producers of the data.

Moreover, IDOC's dataset management ensures integrity and authenticity during the processes of ingest, archival storage, and data access: changes to data and metadata are documented and the relationship of the dataset with the original data is maintained.

Data management OAIS description: https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file

The Data Management function maintains databases of descriptive metadata identifying and describing the archived information in support of the OAIS's finding aids; it also manages the administrative data supporting the OAIS's internal system operations, such as system performance data or access statistics. The primary functions of Data Management include maintaining the databases for which it is responsible; performing queries on these databases and generating reports in response to requests from other functional entities within the OAIS; and conducting updates to the databases as new information arrives, or existing information is modified or deleted. In managing these databases,

the Data Management function supports search and retrieval of the OAIS's archived content, and administration of the OAIS's internal operations.

3.2 SECURITY AND ACCESS CONTROL MECHANISM

Data management ? (Roles, rights….)

Continuity of access : (see document IDOC-OD-002)
Check availability of servers (application servers, database servers, archive servers) :
Check request working on a dataset : selenium ?

3.3 CITATIONS

The interfaces developed must explicitly advise IDOC users to include in their articles and publications an explicit reference to the IDOC site that has enabled them to progress in their work.

4   HOW TO ACCESS TO IDOC DATA

ACCES is the primary mechanism by which the OAIS meets its responsibility to make its archived information available to the user community.

IDOC data can be accessed via different tools and services.

4.1   ACCESS THROUGH WEB INTERFACES : SITOOLS / REGARDS OSS / MIZAR / MAGYC

List of current IDOC interfaces :
https://idoc.osups.universite-paris-saclay.fr/RESOURCES/Online%20Catalogs

4.1.1   SITOOLS 2

SITools2 is a generic tool performed by a joint effort between CNES and space research laboratories. The aim of SITools2 is to provide a self-manageable data access layer deployed on already existing scientific laboratory databases. It is a secure Client/Sever application allowing User and Data sources management and access through a WEB 2.0 interface.

CNES and IDOC at IAS have been collaborating in defining and implementing the needed layers and plugins for astronomy, such as web services for the Virtual Observatory, interoperability, advanced research forms with resolver names and map extractors.
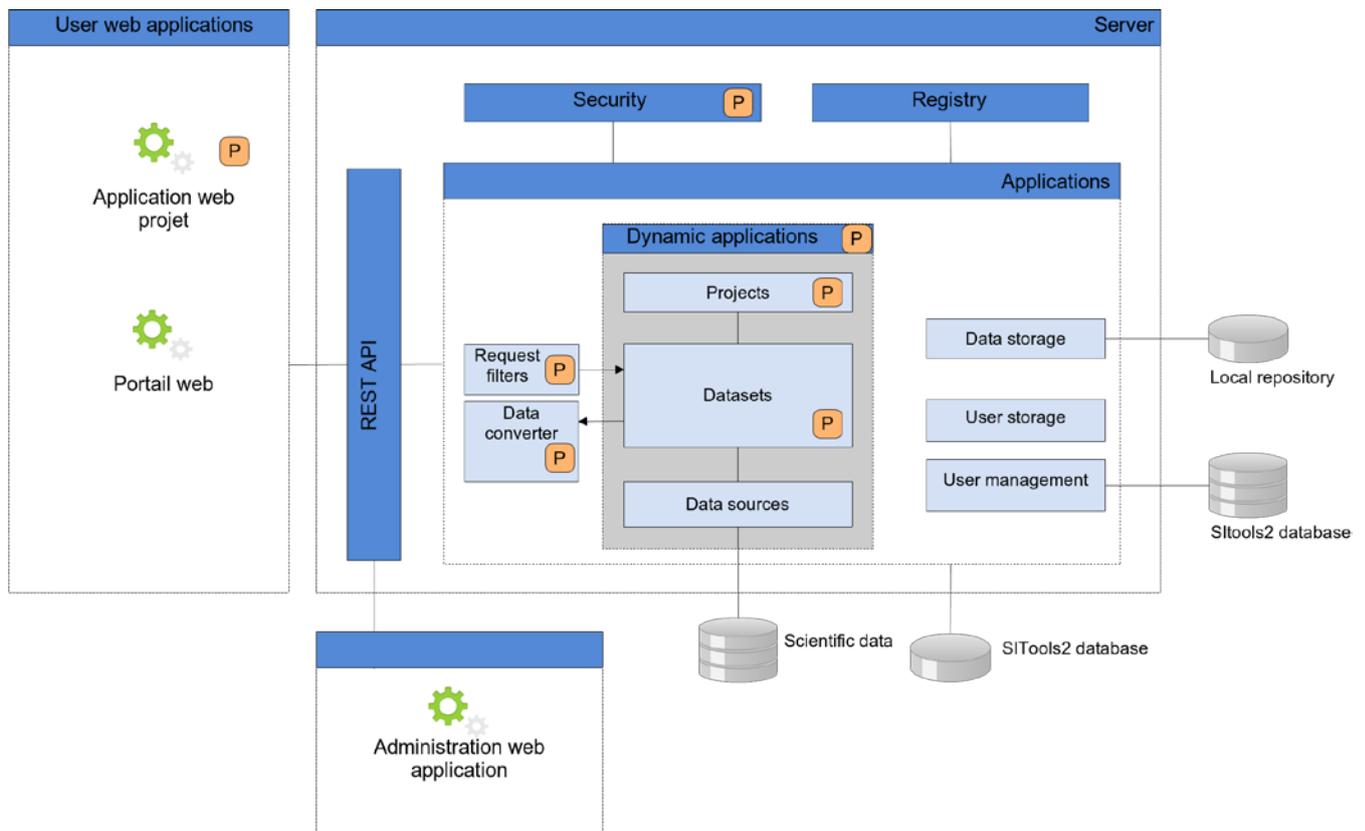

Figure 1 SiTools2 architecture

4.1.2   REGARDS

IDOC is now in the process of testing the REGARDS framework as new framework for dissemination. It is not only dedicated to dissemination, but a focus is done one the access part in this section.

REGARDS architecture is composed of five main components corresponding more or less to the functions of the OAIS model. Those components deals with « Ingestion », « Storage », « Data management », « Administration ». and « Access ».
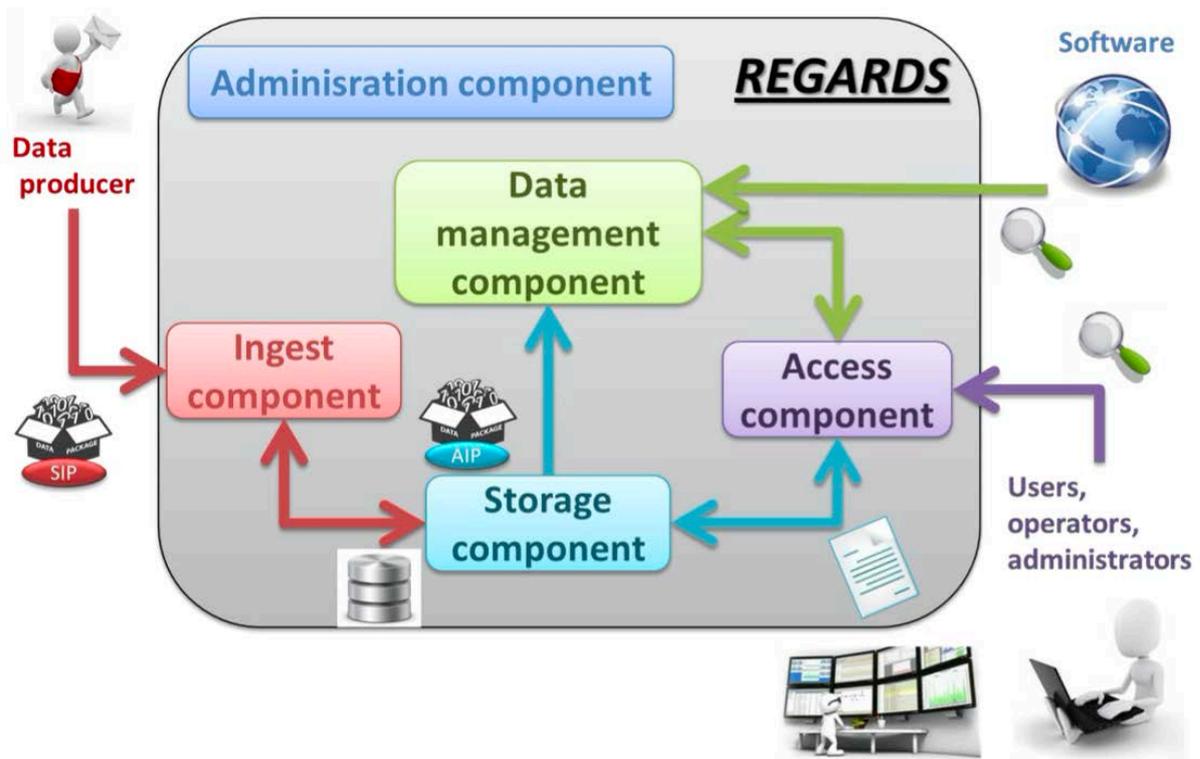

Figure 2 REGARDS architecture

The access function includes three parts: user interface configuration, interoperability access and data ordering.

1. User Interface configuration:
For each mission to be archived with REGARDS, the data (products) to be archived and distributed, criteria to search available data for users and web GUI look and feel need to be defined. This process is iterative with the project team and the REGARDS team specialized in long term archiving.
This process is facilitated by the structuration in modules. Each module is a plugin: the construction of the user interface is done without any development, directly by assembling modules.

2. Data ordering:
The ordering (order micro-service) provides access to the data either through the local archive or through the external "data sources" (URL or web services processing). The data download can be synchronous or asynchronous (ie. with the possibility to process the data (processing micro-service) or in case of high volume to download it as a background task).
Dedicated plugins can be developed depending on the exposed services of the "data source".

3. Interoperability access (example of MIZAR cartographic component):
REGARDS cartographic component for visualization is the OpenSource product MIZAR (https://github.com/MizarWeb).
MIZAR can be embedded in the REGARDS GUI and be connected to REGARDS catalog using the exposed web services (Opensearch web services). Doing so, the users can research data from REGARDS using MIZAR GUI and its cartographic search capabilities.
Selected products can then be ordered (to be downloaded) by users through MIZAR, using the order capabilities of REGARDS.
MIZAR is able to interface several catalogs and could be used to provide search capabilities into various catalogs including different REGARDS projects or external catalogs.

REGARDS also contribute to better implement the FAIR principles for IDOC Archives:
- Findable: better referencing of CNES archives
- Accessible and Interoperable: pertinence of search and selection tools, compatibility with standard protocols of interoperability
- Reusable: standard metadata descriptions, documentation, services (visualization, web processing services)

Please refer to [RD11] for more details on REGARDS

4.1.3   Data Visualization

4.1.3.1   MIZAR

Mizar is a Javascript library developed by CNES. This library is mainly based on GlobWeb library developed by Telespazio France using the WebGL standard. WebGL allows embedding 3D visualisation in a browser without any plugin and it is supported on recent versions of Mozilla Firefox and Google Chrome.

This library is designed for developers who wants to integrate the core api of Mizar in their web page.

This core API provides functionnalities needed to :

- display GIS and astronomy data
- navigate through them
- connect to main interoperability OGC and IVOA protocols

The core API is a Javascript API.

This library is used by several IDOC instances (planetary data, Herschel data, SZ Clsuter…)

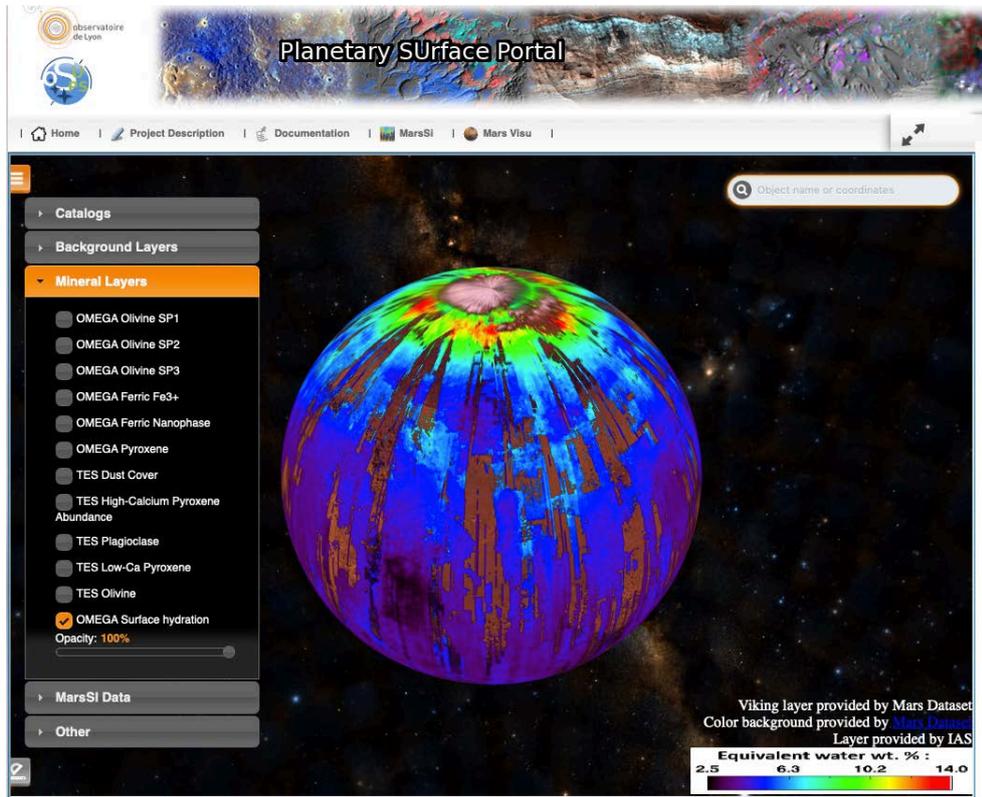Example using MIZAR inside the PSUP (Planetary Surface Portal) SITools2 instance :

Figure 3 Mizar with PSUP

4.1.3.2 Datacube Explorer

CNES / IDOC DataCube is a javascript application to explore data cubes with two positional dimensions and one spectral dimension. Its main functionalities are :

- a Slider to view each spectral picture of the cube with different color gradients
- a Spectrum graph to visualize the value of selected pixel though each slide
- an Histogram to select the pixel values to focus on and adjust the color gradients accordingly
- a Metadata bloc to display the file informations
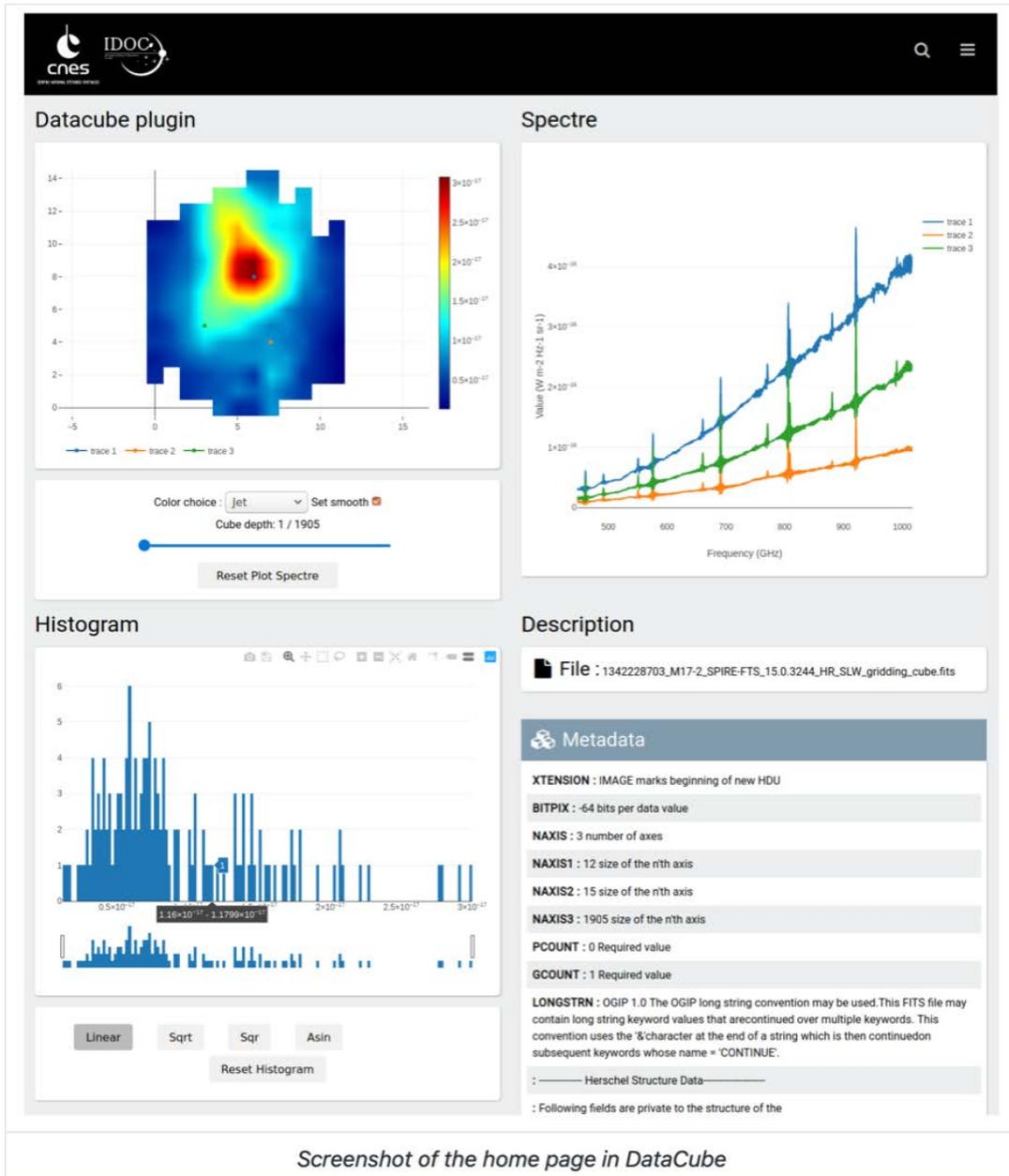
Screenshot of the home page in DataCube

Figure 4 Datacube Explorer

DataCube is a javascript application that can be used as a standalone app (see image herein before) or as a plugin in MizarWidget.

Screenshot MizarWidget with DataCube installed as a plugin
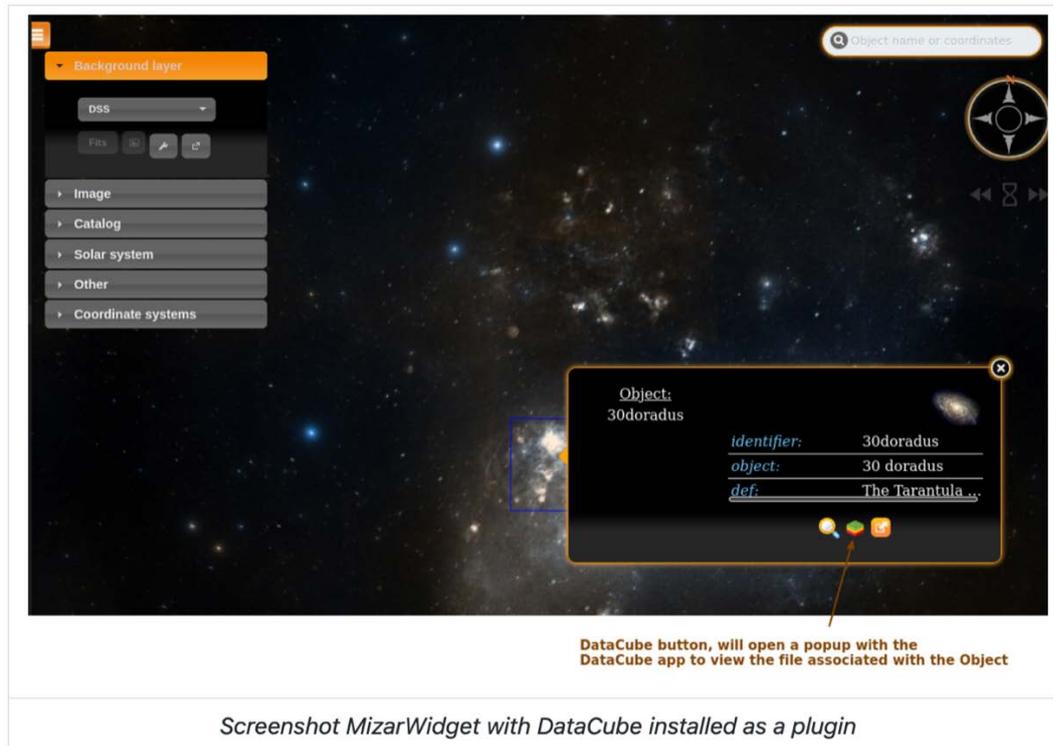
More information on this tool can be found on the dedicated github wiki page: https://github.com/MizarWeb/DataCube/wiki

This tool will be integrated on all IDOC instances with 3D data (FITS or NETCDF files)
There is already a cube explorer inside the HESIOD instance.

4.1.4 MAGYC

MAGYC is a web service portal dedicated to MultiwAvelenght GalaxY Clusters.
http://idoc-magyc.ias.u-psud.fr:443/magyc/

It is based on the multiwavelenght expertise of the IAS and OCA teams and uses VO standards and tools such as UWS, TAP, Vizier and Sitools2.
The data are processed on demand using the UWS protocol.

UWS (Universal Worker Service) is an IVOA standard for asynchronous web services with job management. That means users do not get their results immediately after a request, but they send jobs to the web service. The service puts these jobs into a queue and they are processed consecutively. The job results are then stored on the server and can be retrieved by the user at a later time.

Webservices include search for optical and near infrared potential clusters in survey data with redshift estimate, cutouts of SZ and Xray maps, conesearch in Galaxy Cluster catalogs and Observatory logs. Functionalities are evolving and will be updated with incoming datasets. Necessary inputs are positions and search radius.

The portal is currently based on SDSS DR12 data, Planck 2015 ymap, Rosat all sky survey, SZDB, MCXC, Wen12 cluster catalogs and XMM, Chandra and HST logs (see full references in the Credits tab) .

1. VO

At IDOC, all datasets have to be remotely available through webservices and be compliant with IVOA recommendations.
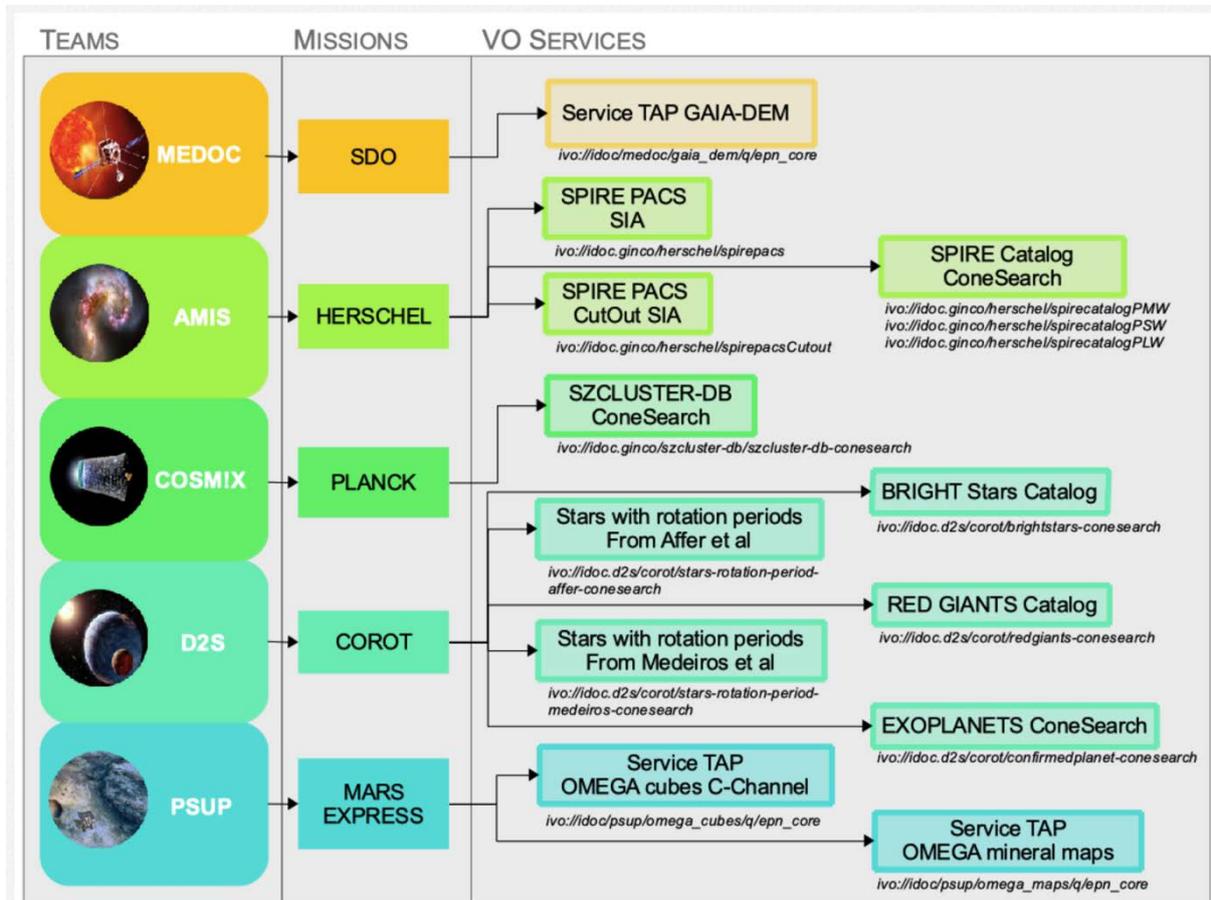

Figure 5 IDOC VO

TAP services are implemented using DACHS.
SIAP and Conesearch are implemented using SITools plugin.
All services have been validated through the eurovo validator, except for TAPs services.

The Dissemination Information Package, or DIP, is the version of the information package delivered to the Consumer in response to an access request. In order to be compliant to the VO standards, original IDOC metadata are modified to match the mandatory metadata, especially for the EPN TAP protocol.

Example with GAIA-DEM: information from the IDOC database is extracted to create the mandatory Granule UID, obs_id and granule_gid mandatory EPN tap parameters
This allows the user to discover and identify IDOC data through generic and recognized VO tools as TOPCAT, Aladin or VESPA for Solar and planetary data.

IDOC has also implemented the IVOA UWS (Universal Worker Service) as described in 4.1.4

4.2 DISSEMINATION THROUGH THE PARIS SACLAY CENTER FOR DATA SCIENCE

https://io.datascience-paris-saclay.fr/map.php
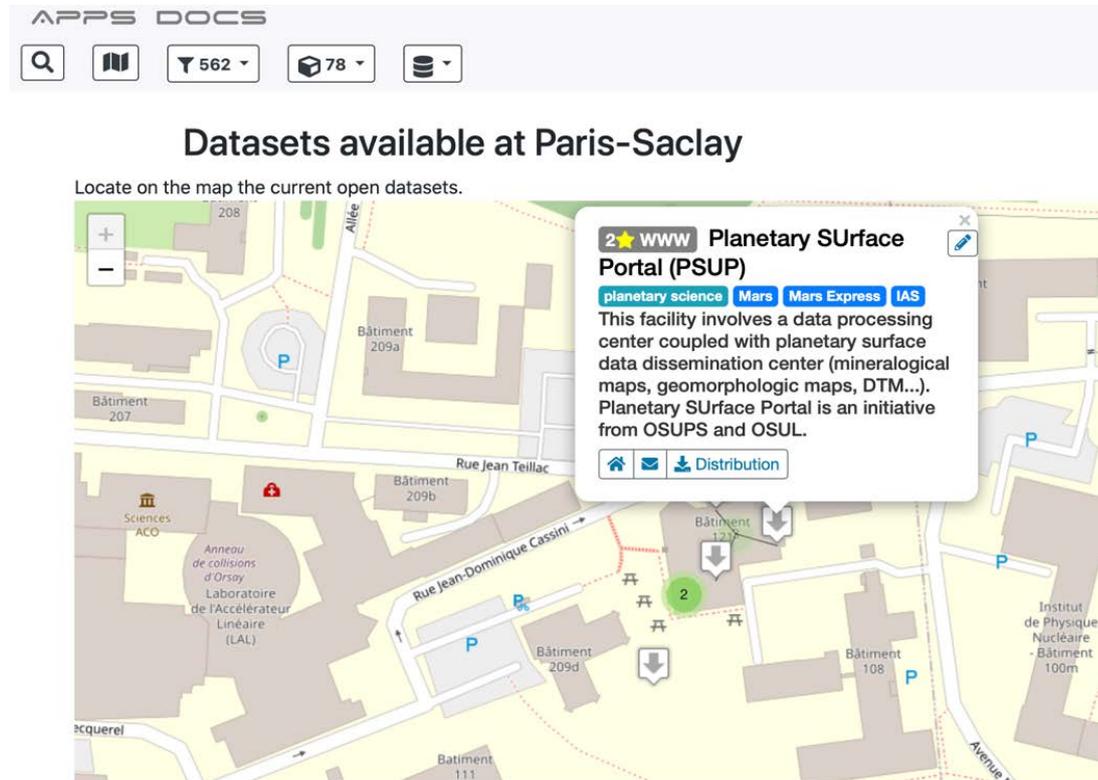
4.3 DISSEMINATION THROUGH EUROPEAN ARCHIVES

The European Space Astronomy Centre, near Madrid, Spain, hosts the science operations centres and scientific archives for ESA's astronomy and Solar System missions.

Final product are sent to European archival center, which also have a data access infrastructure. This is then an indirect access to IDOC data.

Planeto

HSA (Herschel Archive)

SDO ?

5   PROCEDURE TO PREPARE DATASET DISSEMINATION

1.1   ACCESS TO THE DATASET

a)   Data immediately available (production and storage « online » or not « nearline »)

b)   What are the desired technical processes for making the dataset available to the communities concerned (man-machine interfaces, level of ergonomics, machine-machine...)?

c)   The type of access

- The access model: human-machine (manual or semi-automatic search) or machine-machine (fully automatable independently of the specific organization of the dataset - web-service type)

- Describe any possible « Use Case » of desired uses

- What software tools allow scientific access (once the data is selected using the interface described in 3.2.4), understanding and using the data? Specify whether some of these tools involve commercial commitments.

d)   What are the documentation associated with the dataset for its understanding and use?

e)   In which catalogs, directories ... are these data stored   ?