IDOC-DATA Instructions for Data Provision

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
2.0	03/08/2022	IDOC-DATA

TABLE DES MATIERES

1 SCOPE OF THE DOCUMENT

This document is related to the « IDOC-OD-008 Instructions for new services » 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

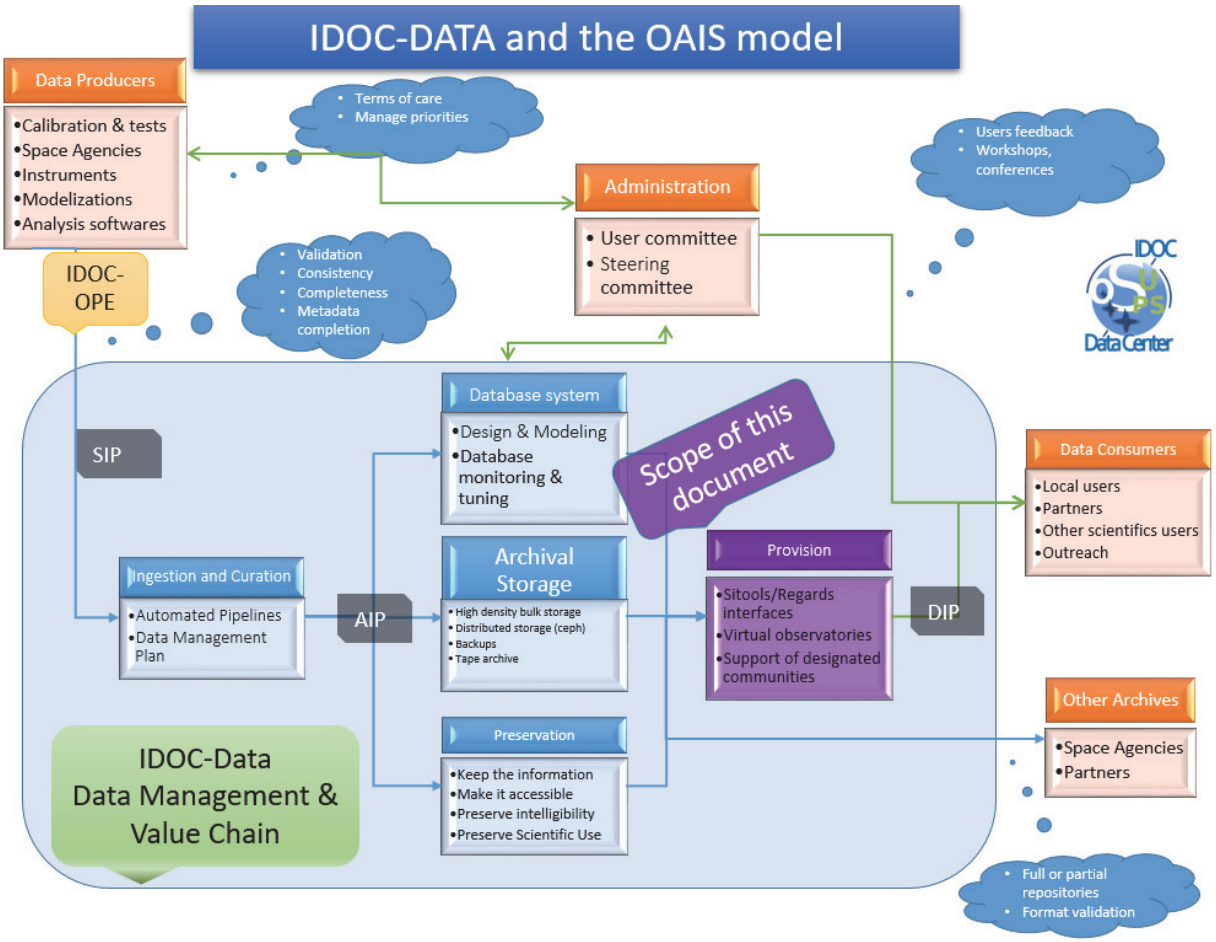
2 REFERENCE DOCUMENTS

Acronym	Reference of the document	Document full name
RD1	IDOC-EX-001	IDOC-EX-001 IDOC executive summary
RD2	IDOC-OD-002	IDOC-OD-002 IDOC Risk analysis and management
RD3	IDOC-INS-003	IDOC-INS-003 IDOC Instructions applicable to project design
RD4	IDOC-INS-004	IDOC-INS-004 IDOC-DATA Instructions for Data Ingestion and Curation
RD5	IDOC-INS-005	IDOC-INS-005 IDOC-OPE Instructions for Ground Segments
RD6	IDOC-INS-006	IDOC-INS-006 IDOC-DATA Instructions for Data Preservation
RD7	IDOC-INS-007	IDOC-INS-007 IDOC-OPE Instructions for Instrument Operations
RD8	IDOC-INS-008	IDOC-INS-008 IDOC Instructions for Services
RD9	IDOC-INS-009	IDOC-INS-009 IDOC-DATA Instructions for Data Provision
RD10	IDOC-INF-010	IDOC-INF-010 IDOC Organigrammes
RD11	IDOC-DW-011	IDOC-DW-011 Diverses schemas for documentation
RD12	IDOC-INS-012	IDOC-INS-012 IDOC instructions for architecture and coding practices
RD16	IDOC-EX-016	IDOC-EX-016 OSUPS Schéma Stratégique Numérique
RD17	IDOC-OD-017	IDOC-OD-017 Services offerts par IDOC
RD30	IDOC-HO-030	IDOC-HO-030 Presentation IDOC-public-english

RD31	**IDOC-HO-031**	**IDOC-HO-031 Presentation IDOC Français**

3 GENERAL PRINCIPLES APPLICABLE TO DATA PROVISION

The dataset provision is related to the "Access" element of the OAIS model.



3.1 IDENTIFYING THE EXPECTATIONS OF DESIGNATED COMMUNITIES

IDOC-DATA staff is composed of engineers deeply participating in the research of the scientific teams building the datasets. They thus have the necessary knowledge to accommodate the data evolutions. Most of the times, they even are the dynamic of these evolutions. Moreover, the IDOC-DATA organization largely promotes the technological monitoring to explore, in advance, future technical developments that seems promising for its present and future requirements. This allows to be prepared to integrate the right tools at the right time, if and when their benefits are validated. The participation of IDOC-DATA members in regional and national networks allows reinforcing the effectiveness of this technology monitoring.

For each of the five scientific themes it hosts (solar physics, interstellar medium physics, cosmology, stellar physics, planetary surfaces, and other thematic of GEOPS and AIM), IDOC-DATA includes a scientific leader. Each of those leaders is a recognized senior scientist who acts as an adviser irrespective of each specific delimited project requirements.

As already stated in R5 a group of independent experts validates the technical and scientific orientations and the quality of the actions undertaken.

During the definition phase of each project, the leader is usually building a team of (national or international) experts in the given theme who will help in defining the requirements and act as beta

testers. After the service in online, the leader presents the datasets or tools in conferences and workshops and collects feedbacks.
These feedbacks lead to changes in the interface or integrates its FAQs. A "contacts and credits" page is always available, which also allows contributors to be thanked.
The experts team also act as reviewers periodically to ensure that the tools and datasets evolves in phase with the community needs. Further communication is done through interface sites, exchange forums or the implementation of collaborative tools.

In practice, each of the interfaces giving access to the 63 datasets distributed by IDOC-DATA allows:
- to find the first level of help for the use of these data
- to contact the experts of this interface
- to find the DOI of the dataset and the contact information of the creator(s) of this dataset
- to participate in a possible collaborative exchange place.

IDOC-DATA also organizes workshops to help you get to grips with the most complex aspects or new categories of data.

At last, IDOC-DATA governance is assured by the IDOC-DATA steering committee. It is designated by the OSUPS Governing Board which gives its recommendations. The IDOC-DATA steering committee nominates both the IDOC-DATA technical and scientific leaders.

3.2 Compliance with specific contracts or Data Management Plans

Whatever the service IDOC implements, it has to be done in accordance with the terms agreed with the producers of the data or project funders (e.g. space agencies).

Moreover, IDOC's dataset management ensures integrity and authenticity during the processes of ingestion, storage, data access and preservation: 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 Provision 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.

3.3 Virtual Observatories

The international "Virtual Observatory" initiative was born out of the need to exploit the enormous amount of data produced by ground and space-based observation programmes, the volume of which has increased enormously due to the implementation of new generations of high density sensors.
It was also noted that the data from one instrument were sometimes not sufficient to fully describe an observed object or to analyse a phenomenon, but that, thanks to the multiplication of available instruments, the combined use of all the data relating to this object or phenomenon made it possible to resolve many questions.

Hence the founding idea of the Virtual Observatory (VO): to be able to cross-reference information regardless of the original source of production.

The VO has thus become an essential framework for the organisation of databases and services in astrophysics, through the standardisation of data structures and access protocols; it constitutes a response to the challenge posed by the management of large quantities of data, their analysis and distribution.

This integration into a standardised framework allows different systems to communicate, i.e. to have system interoperability. This allows the user to seamlessly correlate data from different sources and use generic tools. These possibilities have opened up new horizons in astronomical research.

The evolution of VO, at the international level, on which the developments carried out at IDOC-DATA are based, is done within the framework of the IVOA (International Virtual Observatory Alliance), within which expertise is shared, and VO standards are discussed and formalised.

It is therefore essential to ensure that the tools for making data available within IDOC-DATA are fully compliant and operational with the existing VOs.

Any new VO service has to b
https://idoc.osups.u-psud.fr/Resources/VO.

3.4 SECURITY AND ACCESS CONTROL MECHANISM

A lot of information about how IDOC-DATA deals with this matter can be found in the document IDOC-OD-002.

3.5 DATA IDENTIFICATION

Each object (datasets, software, DMPs,...) available through IDOC-DATA must be identified through a DOI.
DOIs in IDOC-DATA are managed by the DATACITE platform under the cover of the CNRS representative to this platform (this representative is currently INIST)
IDOC-DATA has decided on a nomenclature for the management of the DOIs assigned under its aegis:
The common prefix is 10.48326
10.48326/IDOC has been defined for IDOC as a platform (thus to identify for example the IDOC website)

3.6 CITATIONS

The interfaces developed 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.

As the debate is not easily settled between user identification through the creation of accounts giving access to data and free access, the exact monitoring of the exact uses of the data is not uniform.
Indeed, a registration request may put off some visitors and discourage them from using the data presented. On the contrary, logging in with a (free) account seems to be more efficient for IDOC-DATA to be quoted in articles or other types of publications resulting from the use of the data or tools made available.

Consequently, the choice implemented is the result of a discussion between the actors of the dataset: providers, relevant user committee...

3.7 LICENSES

Data access is mainly open and free of charge for IDOC-DATA datasets, services and codes.
The general license for the data, unless otherwise specified in the dataset-specific copyright (according to french government recommendations https://www.data.gouv.fr/en/licences) is the following:
https://www.etalab.gouv.fr/wp-content/uploads/2018/11/open-licence.pdf
The general license for available software codes, unless otherwise specified in the specific download page is the following:
https://spdx.org/licenses/GPL-3.0-or-later.html#licenseText
Note: A small part of the hosted data may be subject to a temporary embargo or moratorium = delay before open access to provide time for publication to scientist
In the French law about “open science”, this delay is under a maximum of 6 month for science, technical and medical (STM) disciplines, after scientific qualification of data. (Article 30 - THE FRENCH LAW FOR A DIGITAL REPUBLIC - OCTOBRE 2016)
This moratorium depends mainly on the contracts proposed by the space agencies concerning the data from the missions they have financed.
- As far as NASA is concerned, the data are immediately open.
- ESA usually proposes a period of 6 months. The data life cycle of the space missions to which IDOC-DATA contributes is as follows: As soon as a slice of 6 months of data is available, it is the subject of a data package in the OAIS definition which is delivered in parallel to ESA and made available on the IDOC-DATA interfaces. Therefore, a part - which can be large – of the dataset is available making almost 99% of the IDOC-DATA available data publicly accessible.
IDOC-DATA puts in place the elements to guarantee the moratorium period, and then makes this information accessible to all.
During the period of ownership, either the means of access given are concealed from the visitor without the necessary authorizations, or explicit signs describe the restricted nature of such access.
example of an interface request authentication:

4 MEANS AVAILABLE FOR ACCESSING OR MANIPULATING DATA AT IDOC

Access 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 had 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.

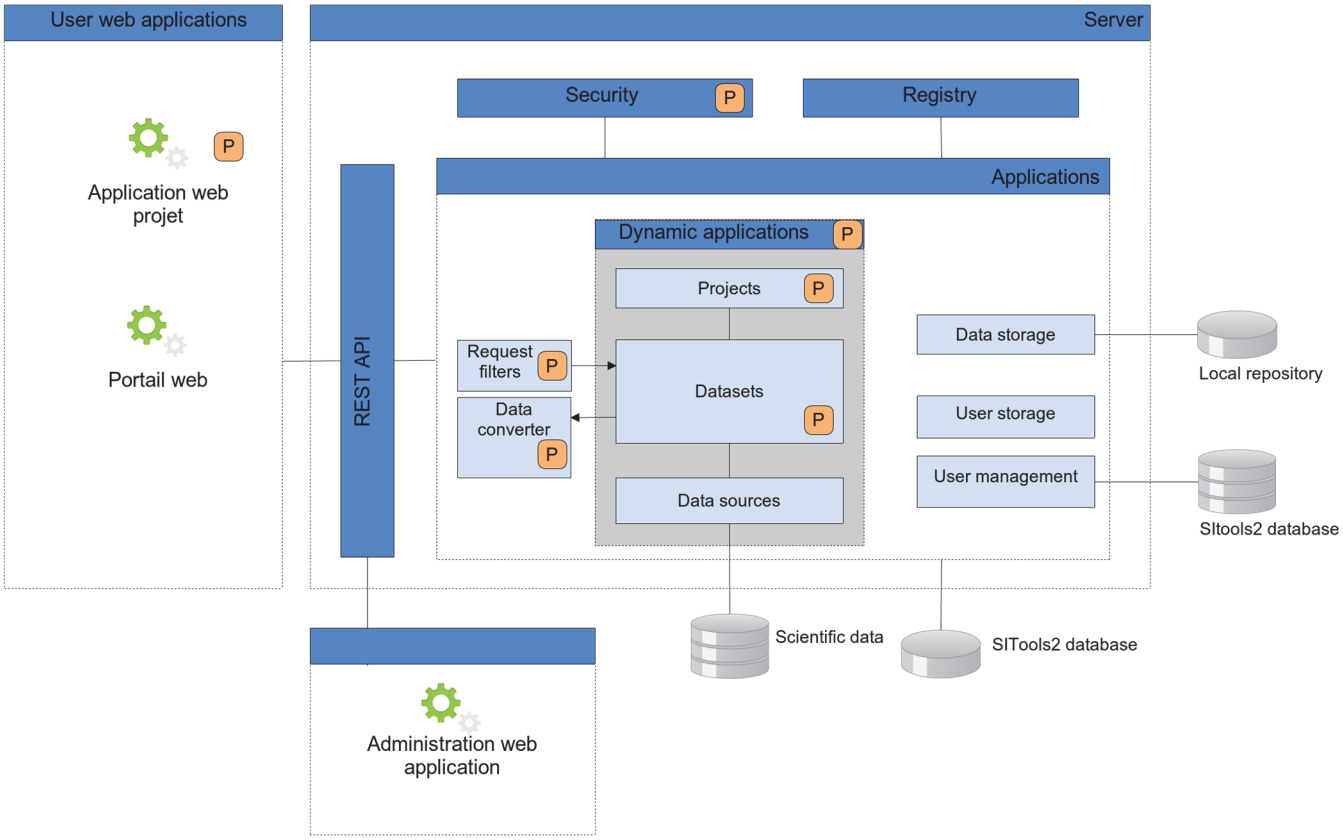


Figure 1 SiTools2 architecture

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 ».

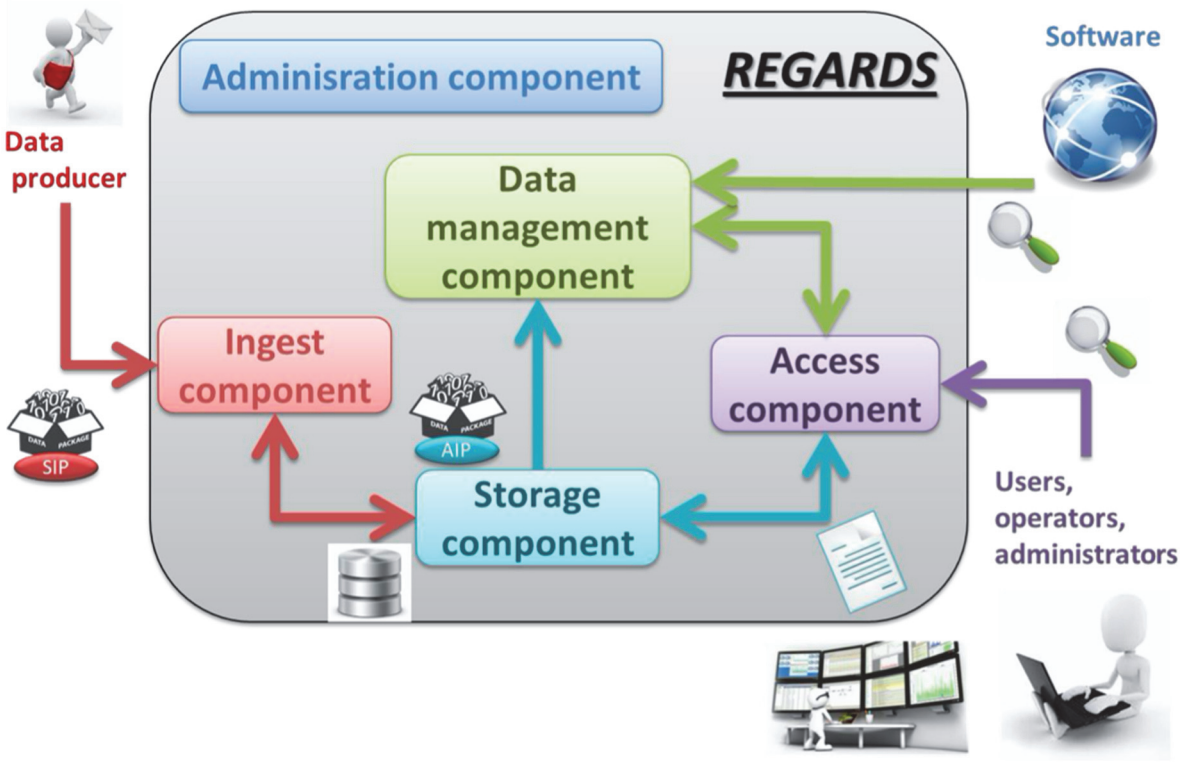


Figure 2 REGARDS architecture

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:

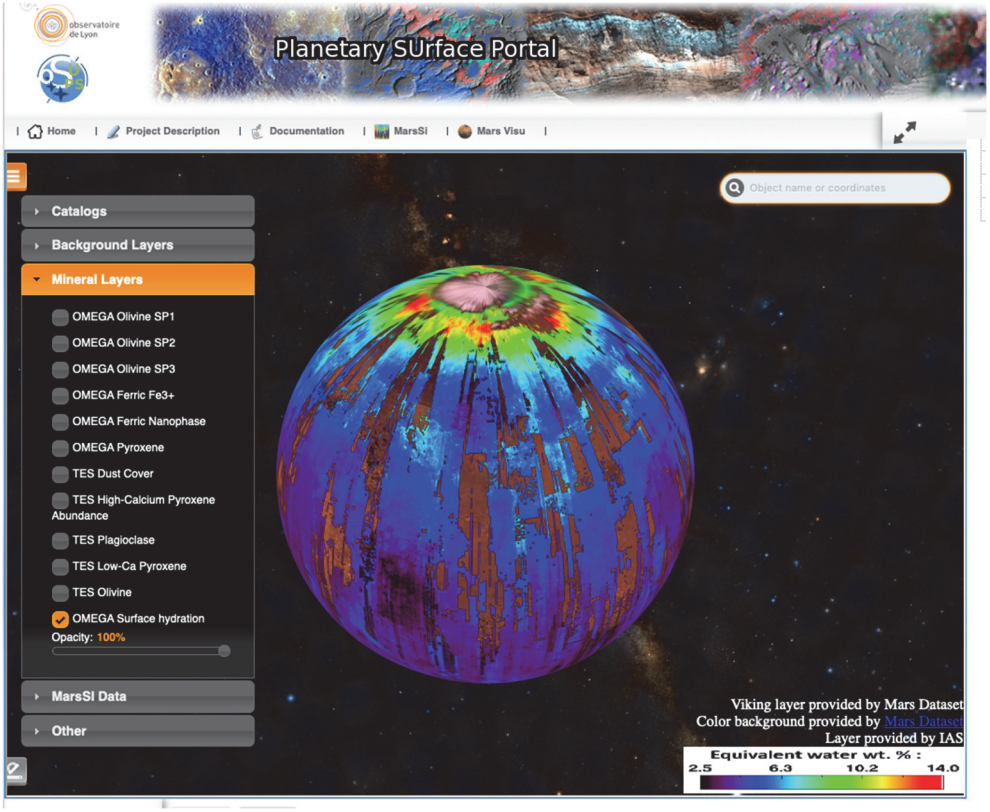


Figure 3 Mizar with PSUP

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

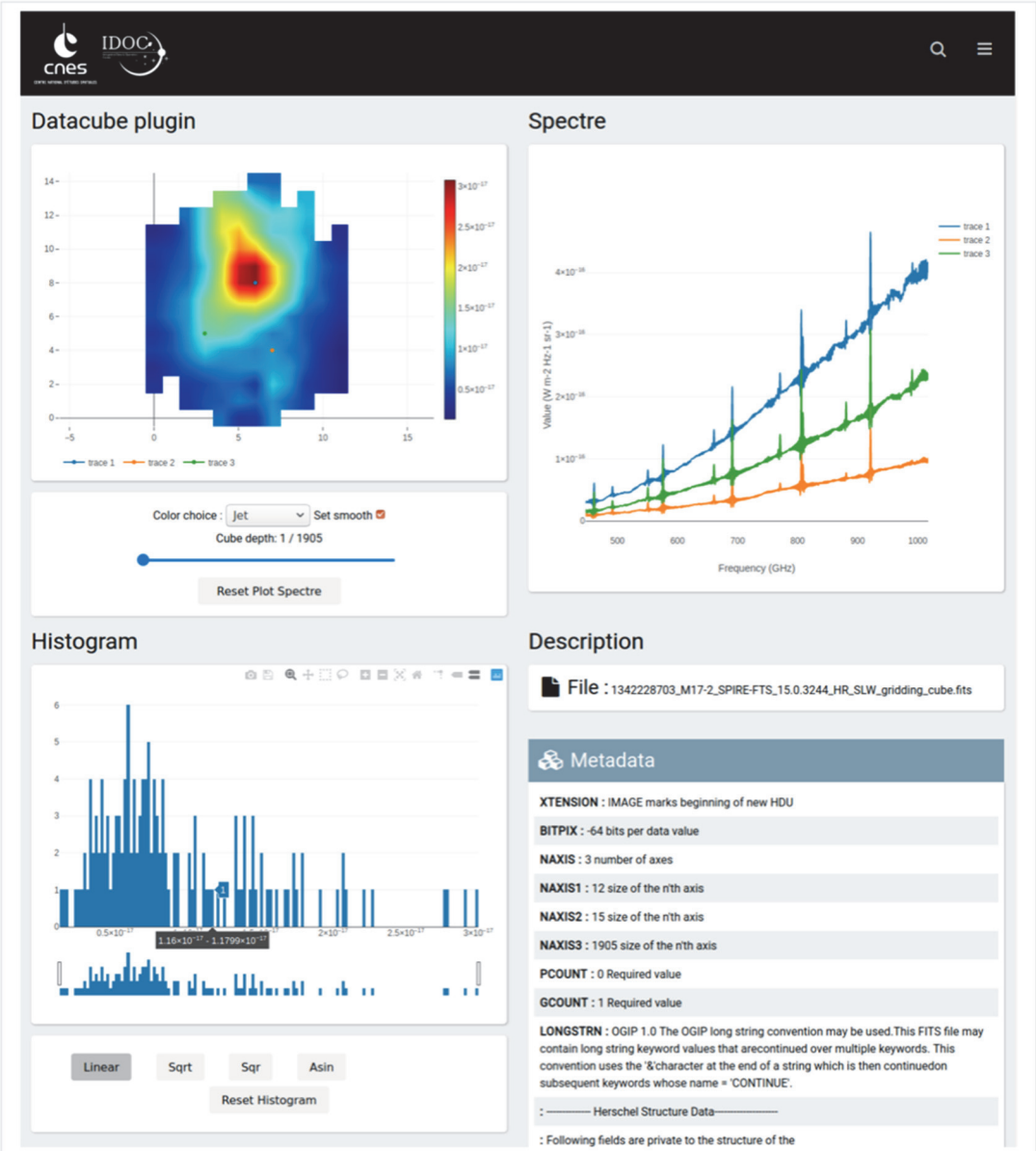


Figure 4 Datacube Explorer

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.

4.2 VIRTUAL OBSERAVATORIES

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

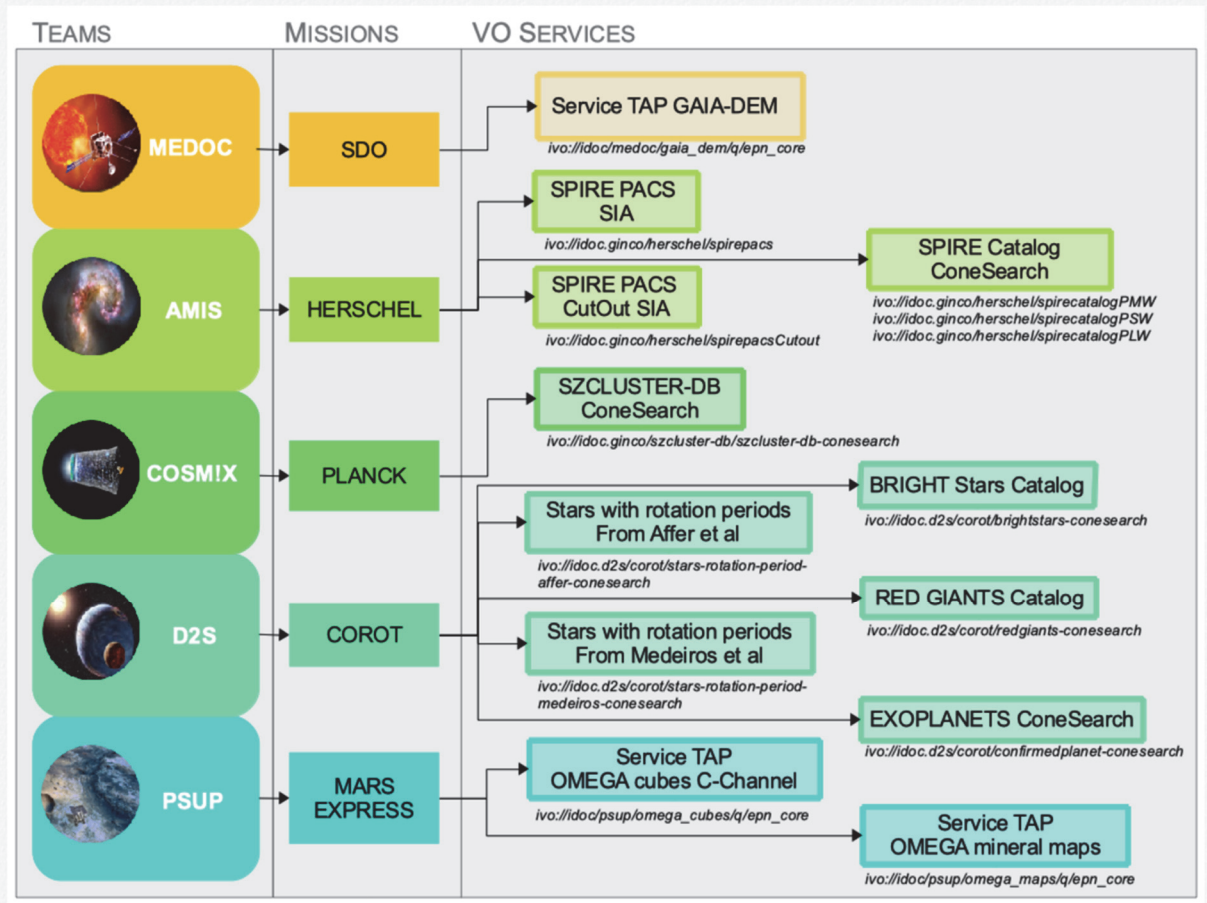


Figure 5 IDOC VO

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 **Erreur ! Source du renvoi introuvable.**

4.3 DISSEMINATION THROUGH THE PARIS SACLAY CENTER FOR DATA SCIENCE

https://io.datascience-paris-saclay.fr/map.php

4.4 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.

4.5 COMBINED ACCESS TO DATA AND PROCESSING

4.5.1 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) .

5 ANNEX 1: PROCEDURE TO PREPARE DATASET PROVISION

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 and by which means are they available ?