Regards - SWH Catalog & Datalake API

Petiton Julien, Benoît Chausserie Lapree and Dominique Heulet

March 24, 2021

REGARDS - SWH CATALOG & DATALAKE API

Julien Petiton, Benoît Chausserie-Lapree, Dominique Heulet

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

ABSTRACT

CNES developed a Framework called REGARDS (REnewal of Generic tools to Access and aRchive Space Data) to catalog and archive space mission’s data, addressing various topics from Earth Observation to Astronomy, as well as physical sciences and technology. REGARDS is highly configurable and adaptable for Mission Centers (Life of Mission Archive) and for the thematic Data Centers (Long Term Archives).

This paper presents the main characteristics of REGARDS (functions, architecture, configuration and adaptation to meet the requirements of a project) and two use cases for the SWH (SPOT World Heritage) archive and for the renewal of the CNES long term storage facility (STAF).

For SWH, the paper presents the steps of the deployment of a REGARDS catalog containing about 20 million of products (30 years of data from SPOT 1 to SPOT 5 missions): prototyping, configuration and plugin development.

The CNES long term storage facility (STAF) relies on hardware (tape libraries) and a dedicated API, that will both be obsolete in a few years. One of the aims of the CNES DATALAKE project is to move archived data to a new infrastructure and to replace the STAF API by REGARDS. The paper shortly describes the evolution of the STAF service (from 1995) and the aims of the DATALAKE project. The paper focuses on the main steps of migration to a new API based on REGARDS.

Index Terms— *data, catalog, archives, storage, datalake, STAF, SWH, SPOT, Long Term Archives.*

1. CONTEXT

REGARDS development started in 2015. Current release (V1.4) is being integrated by two Mission Centers being developed: SWOT (altimetry and hydrology) and MICROCARB (quantify CO2 in the atmosphere). This release is also in operation or pre-operation for CDPP (French Data Center for space Plasma Physics), AVISO (Data Center for altimetry missions), SERAD (Data Archive of several space missions).

Due to its high capabilities of configuration and adaptability, the use of REGARDS framework optimizes development and maintenance costs. REGARDS is able to cope with huge data volumes expected from space missions in the 2020 and beyond. REGARDS addresses the interoperability needs and meets the need to bring the processing as close as possible to the data.

REGARDS is an open source software (under GPLv3 license): https://github.com/RegardsOss

2. REGARDS ARCHITECTURE

REGARDS [1], [2] application architecture relies on micro services. These micro services (developed in JAVA) are highly cohesive and loosely coupled web services, interacting inside the cluster using synchronous (REST through HTTP protocol) and asynchronous communications. A web user interface (front end) provides user and administration functions and interacts with the micro services (back end).

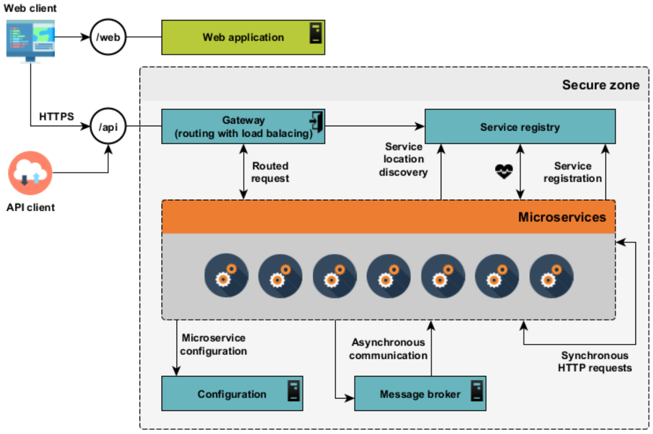


Fig. 1. REGARDS Architecture

The administration interface provides classical administration functions (for example, users and access rights management), but also a large number of configuration functions. Plug in mechanism allows to extend functions of micro services and web interface.

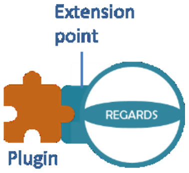


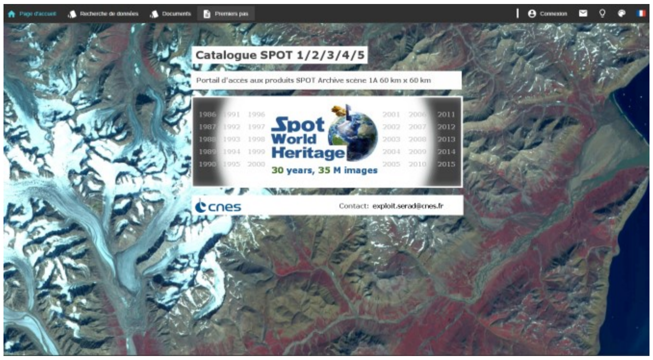
Fig. 2. REGARDS Plugin

The next section presents an example of application of these possibilities to build a catalog access system.

3. REGARDS SWH

SWH [3] is an initiative from CNES to preserve and publish the SPOT products. SPOT 1 to 5 satellites have collected more than 20 million of images all over the world during the last 30 years (from 1986 to 2015). The current L0 archived products represent a total size of about 700 Tbytes.

In 2019, all L0 data have been reprocessed to produce L1A data (despatialized images in the GEOTIFF 'standard image format' associated with an XML descriptive file in the 'DIMAP format'). These L1A products represent 1 Pbyte online data. In 2020, REGARDS was used to build an access catalog to these data.



Catalogue SPOT 1/2/3/4/5

Portail d'accès aux produits SPOT Archive scène 1A 60 km x 60 km

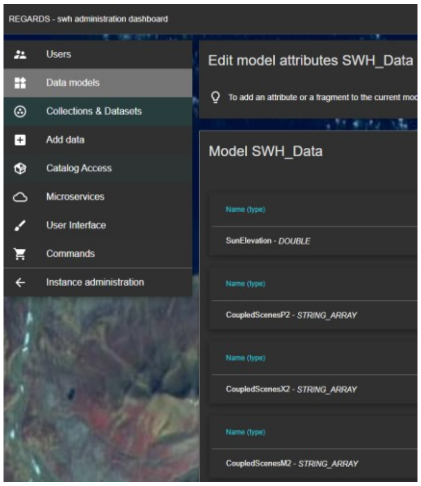
Contact: exploit.serad@cnes.fr

Fig. 3. REGARDS SWH home page

3.1.1. STEP 1: data model configuration

An SWH product is described by about 40 attributes. Some of them are standard attributes (*DataFileSize* for instance), but most of them are specific to earth observation thematic (*CloudCover* for instance).

The web administration interface allows to describe this model and to define the layout in the web user interface, as shown in fig. 4 and 5.


REGARDS - swh administration dashboard

Edit model attributes SWH_Data

To add an attribute or a fragment to the current mod

Model SWH_Data

Name (type)

SunElevation - *DOUBLE*

Name (type)

CoupledScenesP2 - *STRING_ARRAY*

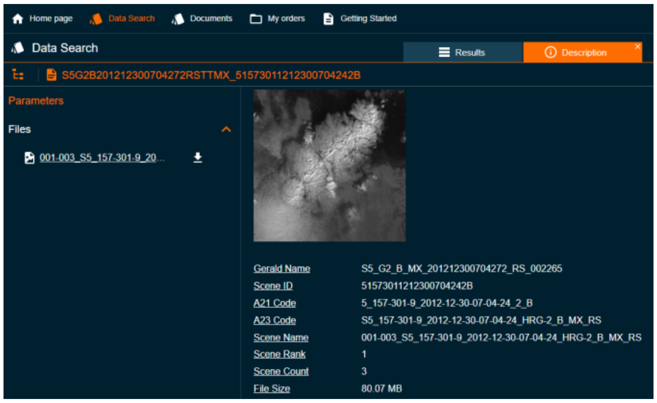
Name (type)

CoupledScenesX2 - *STRING_ARRAY*

Name (type)

CoupledScenesM2 - *STRING_ARRAY*

Fig. 4. admin interface: data model configuration



S5G2B201212300704272RSTTMX_51573011212300704242B

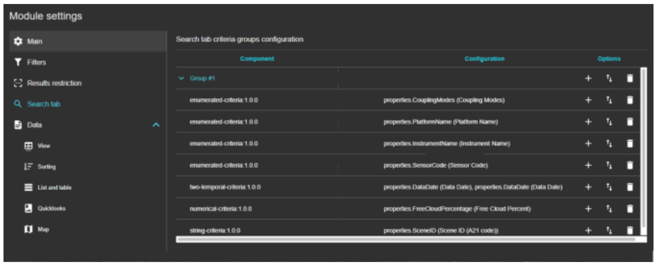
001-003_S5_157-301-9_20...

Gerald Name	S5_G2_B_MX_201212300704272_RS_002265
Scene ID	51573011212300704242B
A21 Code	5_157-301-9_2012-12-30-07-04-24_2_B
A23 Code	S5_157-301-9_2012-12-30-07-04-24_HRG-2_B_MX_RS
Scene Name	001-003_S5_157-301-9_2012-12-30-07-04-24_HRG-2_B_MX_RS
Scene Rank	1
Scene Count	3
File Size	80.07 MB

Fig. 5. user interface: attributes of SWH product

3.1.2. STEP 2: search criteria

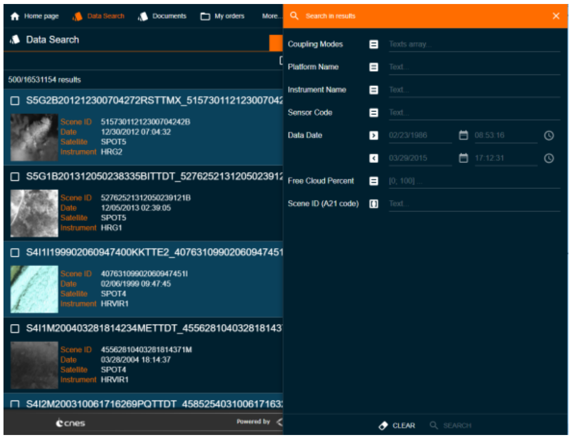
An important step is to configure the search forms. REGARDS web administration interface allows to choose attributes that will be used in search criteria. Search forms are dynamically built in the user interface as shown in fig. 6 and 7.


Module settings

Search tab criteria groups configuration

Component	Configuration	Options
Group #1		
enumerated-criteria:1.0.0	properties.CouplingModes (Coupling Modes)	
enumerated-criteria:1.0.0	properties.PlatformName (Platform Name)	
enumerated-criteria:1.0.0	properties.InstrumentName (Instrument Name)	
enumerated-criteria:1.0.0	properties.SensorCode (Sensor Code)	
two-temporal-criteria:1.0.0	properties.DataDate (Data Date), properties.DataDate (Data Date)	
numerical-criteria:1.0.0	properties.FreeCloudPercentage (Free Cloud Percent)	
string-criteria:1.0.0	properties.SceneID (Scene ID (A21 code))	

Fig. 6. admin interface: search form configuration



500/16531154 results

☐ S5G2B201212300704272RSTTMX_515730112123007042

Scene ID 51573011212300704242B
Date 12/30/2012 07:04:32
Satellite SPOT5
Instrument HRG2

☐ S5G1B201312050238335BITTDT_527625213120502391

Scene ID 52762521312050239121B
Date 12/05/2013 02:39:05
Satellite SPOT5
Instrument HRG1

☐ S4I1I199902060947400KKTTE2_407631099020609474 51

Scene ID 40763109902060947451I
Date 02/06/1999 09:47:45
Satellite SPOT4
Instrument HRVIR1

☐ S4I1M200403281814234METTDT_455628104032818143

Scene ID 45562810403281814371M
Date 03/28/2004 18:14:37
Satellite SPOT4
Instrument HRVIR1

☐ S4I2M200310061716269PQTTDT_458525403100617163

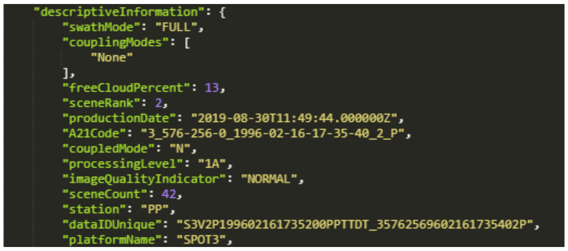
Coupling Modes	Texts array...
Platform Name	Text...
Instrument Name	Text...
Sensor Code	Text...
Data Date	02/23/1986 08:53:16
	03/29/2015 17:12:31
Free Cloud Percent	[0; 100] ...
Scene ID (A21 code)	Text...

Fig. 7. user interface: search form

3.1.3. Step 3: metadata ingest

Step 1 and 2 illustrate the REGARDS parametrization capabilities. Step 3 illustrates REGARDS adaptation capabilities through plug in development.

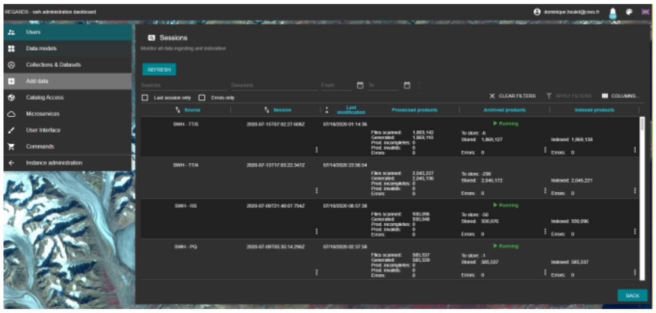
To produce and ingest metadata, the *dataprovider* micro service browses the files arborescence and, for each DIMAP file, applies a plug in that will produce a SIP (Submission Information Package). A SIP is a GeoJson file containing the values of the attributes (read in the DIMAP file) corresponding to data model as defined in step 1.



```
"descriptiveInformation": {
    "swathMode": "FULL",
    "couplingModes": [
        "None"
    ],
    "freeCloudPercent": 13,
    "sceneRank": 2,
    "productionDate": "2019-08-30T11:49:44.000000Z",
    "A21Code": "3_576-256-0_1996-02-16-17-35-40_2_P",
    "coupledMode": "N",
    "processingLevel": "1A",
    "imageQualityIndicator": "NORMAL",
    "sceneCount": 42,
    "station": "PP",
    "dataIDUnique": "S3V2P199602161735200PPTTDT_35762569602161735402P",
    "platformName": "SPOT3",
```

Fig. 8. Detail of a SIP

This plug in is the only piece of code that has been developed for building SWH catalog. The web administration interface allows to define acquisition processing chains that apply the plug in on the 20 million SWH products.



Source	Session	Last modification	Processed products	Archived products	Indexed products
SWH - TT6	2020-07-15T07:02:27.608Z	07/18/2020 01:14:36	Files scanned: 1,803,142 Generated: 1,869,110 Prod. incompletes: 0 Prod. invalids: 0 Errors: 0	▶ Running To store: -6 Stored: 1,869,127 Errors: 0	Indexed: 1,869,138 Errors: 0
SWH - TT4	2020-07-13T17:03:22.347Z	07/14/2020 23:56:54	Files scanned: 2,045,227 Generated: 2,045,136 Prod. incompletes: 0 Prod. invalids: 0 Errors: 0	To store: -250 Stored: 2,045,172 Errors: 0	Indexed: 2,045,221 Errors: 0
SWH - RS	2020-07-09T21:40:07.794Z	07/10/2020 08:57:58	Files scanned: 930,096 Generated: 930,048 Prod. incompletes: 0 Prod. invalids: 0 Errors: 0	▶ Running To store: -50 Stored: 930,075 Errors: 0	Indexed: 930,096 Errors: 0
SWH - PQ	2020-07-09T05:35:14.290Z	07/10/2020 02:37:58	Files scanned: 585,537 Generated: 585,539 Prod. incompletes: 0 Prod. invalids: 0 Errors: 0	▶ Running To store: -1 Stored: 585,537 Errors: 0	Indexed: 585,537 Errors: 0

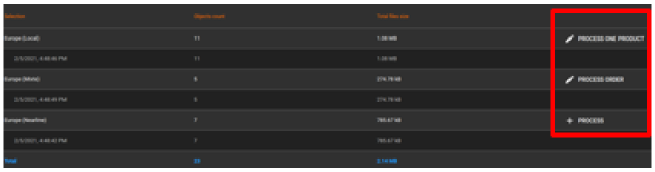
Fig. 9. Admin interface : acquisition chains management

3.1.4. Following steps

The previous paragraphs describe the main steps to create a REGARDS catalog for a project. The web administration interface provides other functions, such as defining the layout of the user interface, configuring user's groups and data access rights, adding web pages and documents, etc.

3.1.5. On demand processing

As explained in the introduction, only the L1A products are available. A plug-in is currently being developed to allow on demand processing to generate L1B and L1C products. This plug in calls the MUSCATE [4] processing chain deployed on the CNES HPC (High Performance Computing Center). The objective of this development is to allow the launching of an external processing chain when downloading products. The fig. 10 displays the mockup of this function.



Selection	Objects count	Total files size	
Europe (Local)	11	1.08 MB	PROCESS ONE PRODUCT
2/5/2021, 4:48:46 PM	11	1.08 MB	
Europe (Mixte)	5	274.78 kB	PROCESS ORDER
2/5/2021, 4:48:49 PM	5	274.78 kB	
Europe (Nearline)	7	785.47 kB	+ PROCESS
2/5/2021, 4:48:42 PM	7	785.47 kB	
Total	23	2.14 MB	

Fig. 10. User interface: cart and processing selection

4. REGARDS AS DATALAKE API

The STAF (*Service Transfert et Archive Fichier*) is the CNES Long Term Storage facility. The STAF is in operation since 1995, and evolved to cope with hardware and software obsolescence. A main step of STAF evolution is to be conducted through the DATALAKE project.

The objectives of the DATALAKE project are to replace the storage infrastructure (tape libraries and disk) by an object storage facility. The DATALAKE project offers several storage classes as shown in fig. 11.

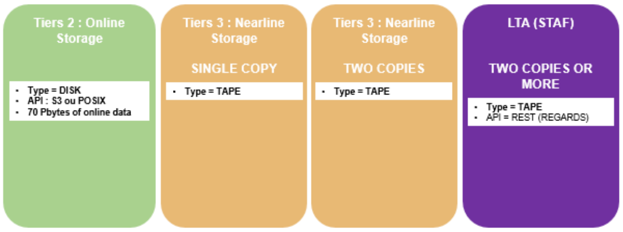


Fig. 11. DATALAKE Storage classes

The STAF provides client server API that will be soon obsolete. Another objective of the DATALAKE project is to replace this API by REGARDS in order to provide a REST API for accessing STAF, to limit the impacts of the evolution of infrastructure and to preserve business metadata.

Currently the REGARDS *storage* micro service uses a plug in to store data into STAF or to restore data from STAF as shown in fig. 12.

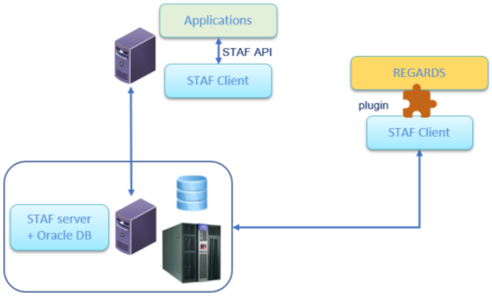


Fig. 12. STAF interface

In 2021, a new plug in will be developed to interface REGARDS and DATALAKE. So REGARDS will be able to interface both STAF and DATALAKE as shown in fig. 13.

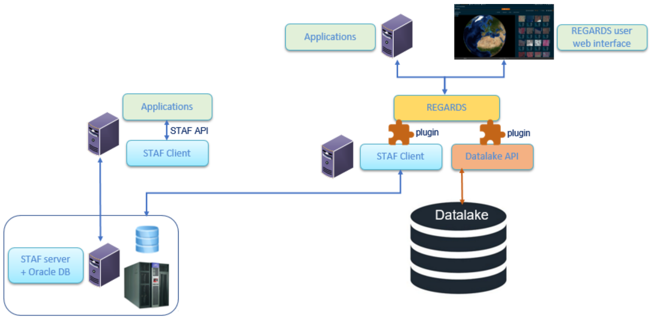


Fig. 13 REGARDS interface with STAF and DATALAKE

Migration from STAF to DATALAKE is foreseen to be completed in 2023. For long term archiving, REGARDS REST API will be the mandatory interface to access DATALAKE LTA storage class to store and retrieve data as shown in fig. 14.

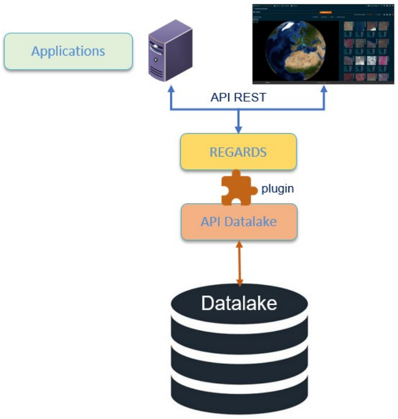


Fig. 14 STAF interface in 2023

Migration between STAF and DATALAKE is foreseen to begin mid-2022. Data files will be moved from the STAF tape library to DATALAKE. No migration is needed for REGARDS catalogs (data access is managed by the REGARDS storage micro-service and uses STAF or DATALAKE plug in). Migration to REGARDS will be necessary for other catalogs as shown in fig. 15.

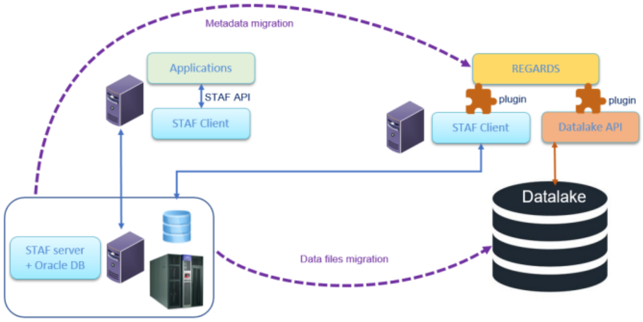


Fig. 15 Migration from STAF to DATALAKE

5. REFERENCES

[1] REGARDS: the new CNES generic system to access and archive space data, BIDS 2019 Conference

[2] REGARDS documentation, https://regardsoss.github.io/

[3] SPOT WORLD HERITAGE – SPOT 1-5 DATA CURATION AND VALORIZATION WITH NEW ENHANCED SWH PRODUCTS, BIDS 2017 Conference

[4] MUSCATE: A VERSATILE DATA AND SERVICES INFRASTRUCTURE COMPATIBLE WITH PUBLIC CLOUD COMPUTING, BIDS 2017 Conference