SPOT WORLD HERITAGE: SPOT 1-5 DATA CURATION AND VALORIZATION WITH NEW ENHANCED SWH PRODUCTS

Julien Nosavan, Agathe Moreau, Antoine Masse, Benoît Chausserie-Laprée, Claire Caillet

CNES

ABSTRACT

SPOT 1-5 satellites have collected more than 25 million images all over the word during the last 30 years from 1986 to 2015 which represents a unique historical dataset.
Spot World Heritage (SWH) is the CNES initiative to preserve and promote this SPOT archive by providing new enhanced products to users.
A first step has begun in 2015 with the start of the repatriation of the SPOT data hosted in the Direct Receiving Stations spread across the world. From 2017, the SWH initiative is moving into a new operational phase with the launch of the official CNES SWH project and the development of first activities.
SWH processing will take place on CNES High Performance Computing Centre to take advantage of SPOT archive proximity and will use Big Data technologies to manage this volume of data, such as Docker for deployment and Elastic Stack for cataloguing and supervision.
First SWH products are expected to be distributed on CNES Web platforms in 2018 while the whole archive is expected to be processed within 2 years until 2020. Access will be free and controlled as defined in SWH licence agreement.

Index Terms— preservation, SPOT, SWH, long term archive, GeoTIFF, GERALD, reprocessing, curation, processing, Elasticsearch, Docker, Web platform

1. INTRODUCTION

SWH has been announced by France at the GEO Plenary, in January 2014 at Geneva. The announcement included a commitment to make available to the public the archive of the SPOT 1 to 5 satellites images.
Since then, CNES launched a first experiment in partnership with ADS to make SPOT L1C orthorectified products, issued from SPOT 1 to 5 satellite images available freely to the public for non-commercial use: following this work, about 100.000 orthorectified SWH images have been produced at the L1C level by CNES and made available on the French Land products data Centre THEIA (https://theia.cnes.fr).
In 2016, SWH has been refocused as a dedicated project and 2 main objectives have been set up: the first objective is to build a new Long Term Archive of SPOT data at level L1A more accessible to users. Indeed, current data are proprietary raw data only accessible through private and commercial ADS platform and L1A data are GeoTIFF images with radiometric processing applied and geometry of the product kept unchanged.
The second objective of SWH is to provide and diffuse enhanced SWH products in line with ESA Sentinel-2 "standards", meaning L1B (product with additional corrections and geometric model refined) and L1C (orthorectified product in Top Of Atmosphere reflectance) in order to extend temporal and comparative analysis.
First SWH activities have then started such as the repatriation of remote SPOT data from Direct Receiving Stations to CNES archive system, the extraction of these data on a shared workspace and the integration of the first processing chains on CNES infrastructure.
In 2018, the development of the SWH-PRODCENTER dedicated to SWH activities will follow up and will gather all SWH activities, parallelizing extraction, processing chains and product diffusion to users. First SWH products are expected to be distributed in the end of 2018.

2. SPOT SATELLITES

SPOT 1-5 is a CNES programme, the last satellite of which has now ended its commercial exploitation.
The SPOT satellites main characteristics are the following: each SPOT satellite is composed of two imager instruments (four for SPOT 5) producing squared "scene" images (60 x 60 km), with two modes of acquisition:

- Panchromatic, black and white with a 10m spatial resolution for SPOT 1-4, or 5m/2,5m spatial resolution for SPOT 5 (2,5m with specific THR high resolution mode using 2 images PAN at 5m resolution)
- Multispectral, colour, with a 20m spatial resolution (for SPOT 1-4), or 10m spatial resolution (for SPOT 5) in three bands (green, red and near infrared (SPOT 1-3) or mean infrared (SPOT 4 - 5)).

Table 1 : SPOT main characteristics

Satellite	SPOT 1-3		SPOT 4		SPOT 5				
Instrument	HRV		HRVIR		HRG				HRS
Spectral Band	PAN	G/R/NIR	PAN	G/R/NIR/MIR	THR	PAN	G/R/NIR	MIR	PAN
Resolution	10m	20m	10m	20 m	2,5m	5m	10m	20m	10m

About 25 million SPOT data have been acquired in 30 years, between 1986 and 2015.

3. SPOT CURRENT ARCHIVE

The SPOT satellite images archive is stored in CNES Long Term Archive system STAF (Système d'Archivage et de Transfert de Fichiers) in CNES Toulouse. The STAF is a high capacity archive system based on robotics manipulating magnetic tapes with redundancy and specific operations to ensure the long term archiving.

Despite the 25 million SPOT products referenced, only 7 million are physically stored in STAF because archiving function in SPOT Ground Segment deployed in the Direct Receiving Stations was not automatic. Repatriation is thus actually performed by ADS and CNES to update the STAF archive with about 5 million images still stored in these distant stations. In the end, it is expected to store 12 million images on STAF at the end of the DRS transfer phase, foreseen for the end of 2018. This means that unfortunately, about 13 million data are still stored in Distant Receiving Stations and are not believed – at the moment – to be easily transferable to the STAF (bad storage conditions, corrupted tapes, format compatibility, data losses …).

The SPOT images are stored as level 0 products under the dedicated GERALD (Generic Exchange for raw Archive Level Data) format, which is constituted as follows:

- One descriptive file (.desc) which describes the acquired segment (or data strip) and its associated scene framing along the Spot Reference Grid (GRS): the .desc file is an ASCII file format
- One or several image files (.ima) which represent the acquired segment: the .ima files are binary files in a proprietary format; they correspond to the first level of processing, completely reversible (auxiliary and ancillary data extractions, quality data analysis and on-board compression for SPOT 5 only).

The characteristics of the GERALD archive format are the following:

- It is the exchange format between the DRS, the former ADS production Centre and the CNES STAF archive,
- It is a long-term format, totally independent from exchange media, self-contained (it contains all the auxiliary and ancillary data to elaborate final products towards users) and without loss on information (it preserves the content of the data as acquired by the satellite and it contains quality indicators to define the original quality of the data).

At last, it shall be noted that the GERALD descriptive file can be formally described using the CNES XIF format thanks to the BEST framework (BEST is a software framework dedicated to data modelling and simulation).

The main drawback of the current GERALD archive format is that it is not directly useable: it is a proprietary format with the image data stored in a raw format, and with on-board characteristics. The level 0 data is not "despatialized".

4. SWH PRODUCTS

In order to facilitate the use of SPOT products, 3 new SWH products have been defined: L1A, L1B and L1C.

L1A product will be the first image product (GeoTIFF) including basic radiometric corrections and preliminary cloud cover estimation; this product will be principally based on the current SPOT N1A scene format which is the reference for years. This L1A level will replace the current SPOT raw GERALD archive level and will form the new official SPOT archive available on CNES open Web platform with criteria extracted from metadata file (DIMAP).

L1B product will be in segment format and will provide geometric corrections in line with Sentinel-2 L1B product. First of all, inter-bands registration will be reprocessed with optimized L1B ground parameters and geometric model will be refined with Sentinel-2 Global Reference Images and Digital Elevation Model (Planet Observer). On radiometric side, L1B will include new corrections based on first THEIA experiment providing for instance technical masks (water, cloud …). L1B will also include new algorithms optimized and ready to deal with huge volume of data; in particular, a new denoising algorithm based on Non Local Bayes technique [1] has been developed and optimized for SPOT 5 THR mode production with reduction of computational complexity and a tile-ready processing. Example of this SPOT THR product is shown in Fig.2:

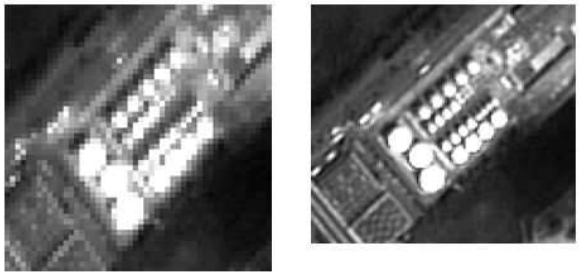

Figure 2: Example of SPOT THR mode result: (left) SPOT 5 HMA image (5m) and (right) SPOT THR product (2.5m)

L1C product will be the orthorectified product in Top Of Atmosphere reflectance, still in line with Sentinel-2 L1C product with a specific split in 20km x 20km tiles (sub-tiling of 100 km x 100 km S2 L1C tiles).

All these products will be based on the current SPOT format (including DIMAP metadata file) and will also include evolutions to take into account Sentinel-2 specificities.

5. SWH SOFTWARE ARCHITECTURE

SWH software architecture relies on a strong reuse of existing tools to minimize development costs and secure validation phases regarding to the volume of data to process.

SWH development is organized around a main Centre dedicated to SWH activities: the SWH-PRODCENTER.

This Centre gather all SWH activities delegated to SWH components managed with the help of a centralized database SWH-DBREF based on Elasticsearch to deal with the volume of the data to index. Supervision will be ensured by

Kibana software that natively allows the setting of customizable “views” and “dashboards” for operators. It is described in Fig.3:

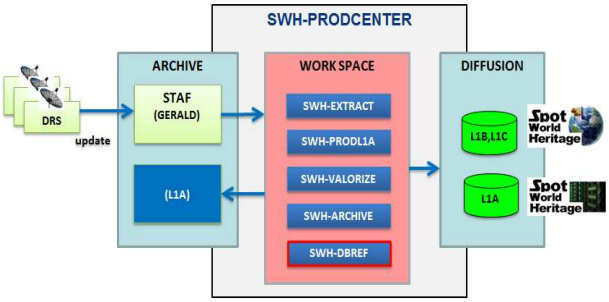

Figure 3 : SWH-PRODCENTER

SWH-EXTRACT is the data extraction chain based on an internal CNES development using itself CNES internal tools (VDLIB, STAF client). It is described in Fig.4:

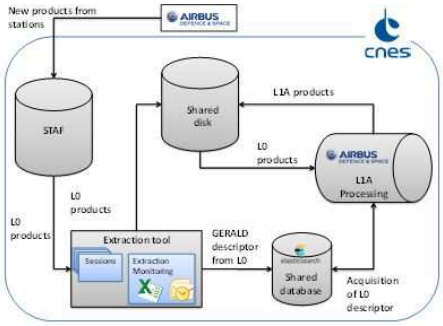

Figure 4 : Extraction and processing flow

The objective of the tool is to extract GERALD data from the STAF to a shared SWH workspace. The tool notifies the L1A processing chain of the availability of the GERALD by adding indexes in SWH-DBREF. The tool can be configured to extract all the SPOT data archive and can be easily monitored through the extractions reports and e-mails.SWH-PRODL1A is the L1A processing chain based on the integration and the automation of the operational ADS SPOT N1A processing chain on CNES infrastructure to take advantage of SPOT archive proximity and CNES High Processing Centre. It is described in Fig.5:

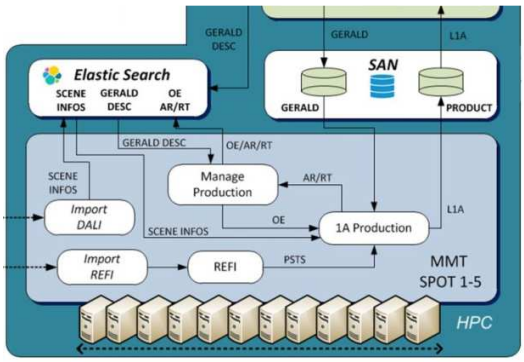

Figure 5: SWH-PRODL1A processing chain

The objective of the tool is to generate L1A products from GERALD products using the shared SWH workspace. The tool notifies the availability of the product by adding indexes in SWH-DBREF. L1A processing chain is scalable and will use Docker software to deploy Docker images on CNES HPC servers to be – when necessary - independent of the system environment.

SWH-VALORIZE is the L1B/L1C processing chain based on the reuse of CNES MUSCATE software already used in THEIA experiment. It is described in Fig.6:

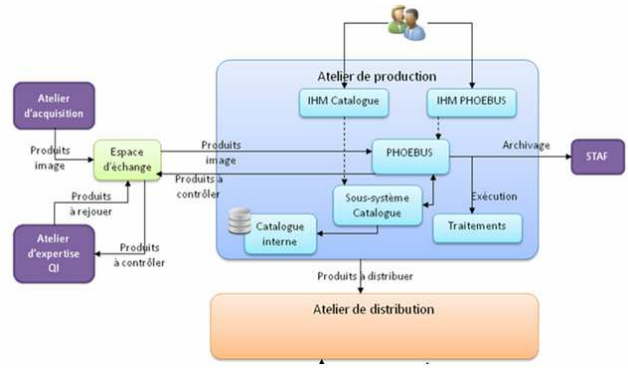

Figure 6: MUSCATE functionalities

The objective of the tool is to generate and diffuse L1B and L1C products from L1A products using the shared SWH workspace. The tool notifies the availability of the products by adding indexes in SWH-DBREF. First THEIA SWH processing chains based on CNES SIGMA software for geometric refining and orthorectification will also be reused and adapted to SWH/Sentinel-2 specificities. Finally, L1B and L1C products will be diffused on a dedicated archive catalogue. SWH-VALORIZE will extract some of these metadata to index the L1B and L1C products and provide search criteria to the users. A cartographic search will also be provided like THEIA platform.

SWH-ARCHIVE is the archiving processing chain based on the reuse of CNES REGARDS software. It is described in Fig.7:

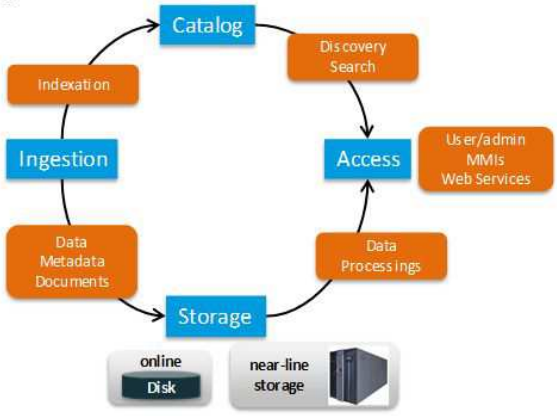

Figure 7: REGARDS functionalities

REGARDS is a new CNES product currently in development. It is based on micro-services architecture with external interfaces using HTTP Restful services. REGARDS is a generic product which means it will be used by several missions in several data centres (earth observation, astronomy, space sciences, etc.). For each project a new implementation will be deployed allowing project specificities. For SWH, REGARDS will be configured and a specific plugin will be developed to interface SWH-DBREF to identify which L1A products are to archive. The L1A product will be a zip containing the GeoTIFF images as well as a DIMAP descriptive file and additional information (cloud coverage, traceability …). REGARDS will extract some of these metadata to index the L1A product and provide search criteria to the users. A cartographic search will also be provided by REGARDS. Once the L1A product is archived, the associated GERALD will be deleted from the online storage.

The objective of SWH-ARCHIVE is to archive back L1A products in replacement of GERALD archive from the shared SWH workspace to STAF. The tool notifies the archiving status by adding indexes in SWH-DBREF.

6. SWH CHALLENGES AND IT ARCHITECTURE

The current estimation of 12 million images SPOT represents a volume of ~700 TBytes for GERALD data, 1 PByte for L1A products, 1 PByte for both L1B and L1C products (using compression).

GERALD extraction is expected to be performed during one and a half years depending on the timing of SPOT data retrieval from the Direct Reception Stations.

L1A processing capacity has been estimated with benchmarks on ADS SPOT N1A processing chain and a cluster of 17 nodes (24 core, 128 Go RAM) has been identified to generate the whole L1A archive in one year meaning a daily production of 90.000 scenes.

L1B/L1C processing capacity has been estimated with current performances observed on THEIA prototype and ongoing optimizations on CNES MUSCATE software. 60% of L1A expected to be processed in L1B and L1C level, depending of the Cloud coverage and the Quality image configuration put in place in THEIA processing. The whole L1B and L1C products are then expected to be processed in less than 2 years.

L1A archiving is expected to be performed during one and a half years but

All these components will be launched in background with a weekly configuration (TBC) and a daily supervision. However, priority can then be taken into account if a specific area is asked for, depending on the partnership put in place.

SWH IT architecture is based on the use of ~20 nodes (24-core processors) and optimized GFPS shared workspace for computation. A final diffusion workspace will host data for the SWH Web platform.

Here is displayed in Fig. 8 the global IT architecture of SWH-PRODCENTER built on CNES facilities:

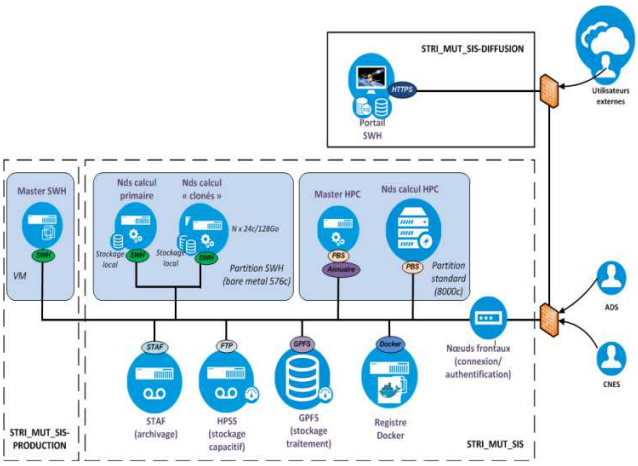

Figure 8: SWH IT architecture

7. SWH SCHEDULE AND PRODUCT DIFFUSION

SWH-EXTRACT software is validated and operations have started in August 2017: GERALD data are currently being extracted.

L1A processing chain development has started in July 2017 and the operational version is expected in summer 2018.

All remaining activities (SWH-VALORIZE, SWH-ARCHIVE and SWH-DBREF) are gathered in the SWH-PRODCENTER development that will start in the beginning of 2018.

Production is expected to start in the end of 2018 and first products are then also expected to be distributed in the end of 2018 while the whole archive is expected to be processed within 2 years until 2020.

SWH products will be accessible through two dedicated Web platforms: one dedicated to L1A archive products, one dedicated to L1B and L1C products.

SWH licensing is being defined but the main goal is to provide a free and controlled access to all SPOT data. The new generated SWH products will then be accessible free of charge to registered users.

8. REFERENCES

[1] A. Masse, S. Lefèvre, R. Binet, S. Artigues, P. Lassalle, G. Blanchet, and S. Baillarin, "Fast and accurate denoising method applied to very high resolution optical remote sensing images", Proc. SPIE 10427, Image and Signal Processing for Remote Sensing XXIII, 1042703 (4 October 2017), doi:10.1117/12.2277705.