The VITO Earth Observation Long Term Data Archiving Facility

Paepen Martine [(1)], Goor Erwin [(1)] , Deronde Bart [(1)], Ooms Bart [(1)]

[(1)] VITO, Flemish Institute for Technological Research
Boeretang 200, 2400 Mol, Belgium
EMail: martine.paepen@vito.be

ABSTRACT

The image processing and archiving centre (CVB), hosted at VITO, processes all data received from the VEGETATION instrument on board of the SPOT4 and SPOT5 satellites, archives the processed data, compiles the image catalogue, and forwards finished products to the users. In addition to the products from the SPOT-VEGETATION sensor, VITO also archives higher level products produced by various processing chains derived from SPOT-VEGETATION and complementary satellite instruments like AVHRR, MERIS, MODIS and AATSR. These products are mainly used for environmental monitoring on a global scale. The archiving facility will also serve other missions in the future such as PROBA-V, a new Belgian satellite which will be launched in April 2012.

This paper briefly describes the architectural concepts of the archiving facility and the specific implementation at VITO for long term data preservation of heterogenous earth observation data.

Keywords: CVB, CVB Archiving Facility, SPOT-VEGETATION, PROBA-V, VITO

INTRODUCTION

Since 1998 VITO is hosting the image processing, archiving and dissemination centre for the SPOT-VEGETATION data (CTIV) [1]. The SPOT-VEGETATION instruments acquire a nearly daily coverage of the earth's land surfaces at a spatial resolution of 1 square kilometer in four spectral bands. The CTIV archive contains raw images, VGT-P (P= Physical) products for scientific applications requiring highly accurate physical measurements, VGT-S (S= Synthesis) products ensuring coverage of all landmasses worldwide with a minimum effect of cloud cover for 1 day (VGT-S1) and 10-day (VGT-S10) synthesis products. The role of the archive is to preserve the data and to provide data on request, more specifically near real time provision or requests for long time series of a specific ROI (Region Of Interest).The operational archive of CTIV was designed as a dedicated archive for the specific data received from the SPOT-VEGETATION sensors.

Because there was a need for a more flexible archiving facility in a more multi mission context to be able to archive also higher level products from SPOT-VEGETATION and complementary satellite instruments, the CVB Archiving Facility (CAF) was initially developed as a sub-system in the Flexsys project within an ESA contract by VITO, Trasys Space and Spacebel. Later, the CVB Archiving Facility evolved into a fully operational infrastructure at VITO.

The architectural concepts of the CVB Archiving Facility are based on the principles of harmonisation, interaction with other systems (in the Service Oriented Architectural way), standardisation and modularity. It is built as a modular, generic and portable Java application extendible with a variety of dynamically pluggable modules and is independent from the selected database and storage facilities.

The CVB Archiving Facility is today a fully operational system serving products based on input data from SPOT-VEGETATION, MERIS, AVHRR, AATSR and MODIS sensors. It is continuously evolving in to an up-to-date Archiving Facility for long term data preservation. In the future the facility

will also be used as the Long Term Data Archive for the products of the PROBA-V mission and also for high resolution data from airborne sensors.

CVB ARCHIVING FACILITY (CAF)

The CVB Archiving Facility is developed as one of the sub-systems of the Flexsys project. The goal of the Flexsys project was to define and build a new infrastructure for the development and deployment of new earth observation services at VITO in a cost-effective way. We designed the infrastructure of the Flexsys project with four main concepts in mind: multi mission, harmonization, interoperability and modularity at all levels. The system integrates independent sub-systems, also called services, like the Processing Facility, Product Distribution Facility , Archiving Facility, etc. which are loosely coupled. The Archiving Facility has later evolved to the CAF which is a fully operational and independent Archiving Facility.

Architectural Concepts of the CAF

- **Archival Information Packages**: we developed the CAF according to the OAIS principles. The OAIS is the reference model for an 'Open Archival Information System' [2]. The CAF archives Archival Information Packages (AIP).
- **Heterogeneous data and metadata** : we designed the CAF in a manner that it can ingest any kind of data and metadata without any configuration, providing the type of each metadata can be represented by an integer, a real, a string, a date or a node structure that is made up of several metadata. We also provide a structure for more complex types such as the bounding box and segment types, which offer specific spatial operators.
- **Complete archive functionality:** the CAF provides application services to permit the ingestion, storage, retrieval and management of the data and metadata.
- **Two layers:** we designed the CAF with two layers: the abstraction layer and the physical storage layer. The abstraction layer is built as a modular, generic and portable JAVA application. The abstraction layer makes the CAF independent from the selected database and storage facilities which are called the physical storage layer.
- **Distributed Architecture:** instances of the application can run on different computers to increase parallelism and performance. Each application instance can be configured according to:
    - The storage resource that can be accessed from the computer.
    - The ingestion and storage plug-ins installed on the computer.
    - The functionality that the instance provides (ingestion, storage, interface computing).
- **Service Oriented Architecture:** the collaboration of the application services (ingestion, storage, retrieval and management) is achieved with the help of the services concept. Each instance can execute the services it is configured for. The first service request comes from the user (another system like a processing facility or a human operator) and triggers the execution of a workflow made up of the execution of other services.
- **Central Relational Database:** the core of the archiving facility is a relational database which is used for both the generic catalogue, and the persistent layer of the data management of the archiving facility. As such it will contain information such as:
    - Configuration data for the different instances.
    - The service request repository. An application instance which is configured as the workflow manager puts service requests in the database and any available application instance can request a pending service from the database to execute the service.
    - The metadata describing the object which is ingested together with the data as an xml file. This metadata can be used for the generic catalogue.

- The metadata produced by the application itself like the physical storage location , the lifecycle state, …. of the different objects.

There is no dependency on the type of database. At VITO we currently use Oracle as database but the design of the archiving facility allows to change the database without major modifications of the archival sub-system.

- **Highly configurable lifecycle management:** an XML configuration file describes projects and lifecycles to define the amount of copies, the rewriting time, etc. of an object. This is a very flexible way to insert data for new projects in an independent way. Life cycles indicate which resources are used for short or long term preservation and how long an object is stored on a specific storage device before automatically being moved to another medium or deleted.
- **Uniform Storage Operations:** the CAF provides uniform storage operations like copy, delete, check, …. These operations are independent of the underlying infrastructure. Plug-ins (scripts) are implemented to make it possible to add or change the physical storage layer. With the CAF we are able to store data on disk, on tape or any other storage medium. Even for one object it is possible to store one or more copies on disk for short term or middle long term data preservation and one or more copies on tape for long term data preservation.
- **Extensible Query Language:** the CAF uses the query language to query the database. We defined it as a set of operators that can be used on metadata names or on constants. The CAF uses rules to dynamically translate the original user request into the instructions directly usable in SQL requests. These rules are stored in the central database to make it possible to change and add new operations without changing the code.
- **External Interfaces:** the CAF provides two asynchronous uniform interfaces: a Command Line Interface and a Web Service Interface. Both interfaces provide the same services with the same parameters. In addition, the Web Service Interface notifies the requester when the execution is completed. As the service structure is normalized, new interfaces can easily be added, new developments then only deal with the interface protocol itself.
- **File System Access:** the objects stored in the CAF can be made accessible through a virtual read-only file system and classified as directory according to their metadata. This directory structure used to present the objects is customizable and can be changed / adapted if needed. This file system can be mounted in the usual way for a UNIX/Linux system and can be exported as NFS, Windows CIFS or FTP. Several structures can be defined at the same time so that the user can browse the archive and access objects using a file explorer, a FTP client, as with any other file system.

Figure 1 shows the high level architecture of the Archiving Facility. The central database hosts the catalogue and the service management repository. The physical storage services are not directly accessible via the clients but can interact with network accessible resources, remote devices and local devices to store the data. The other application instances which are responsible for the ingestion, retrieval and management are accessible through the clients and provide the Command Line Interface and the Web Services Interface.

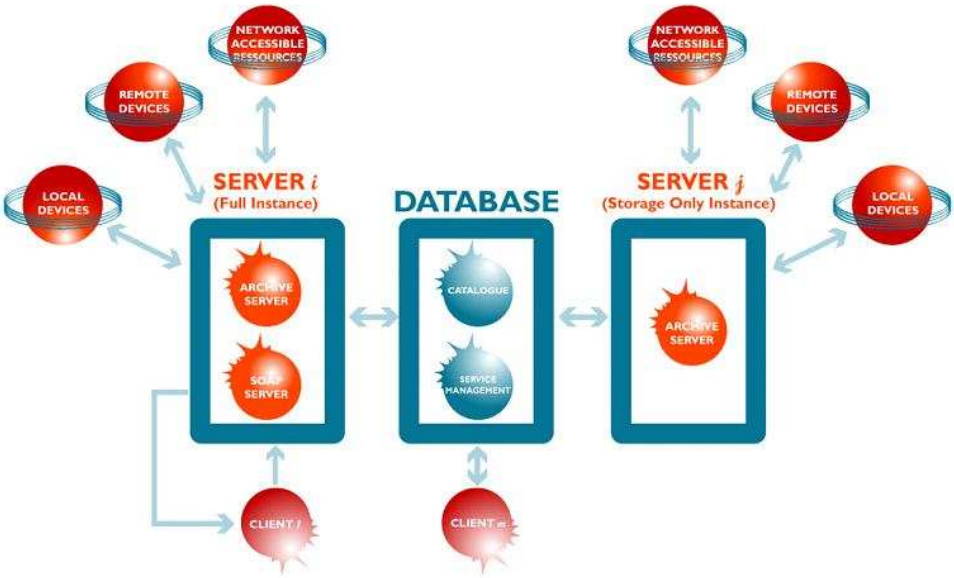

Figure 1: Architecture of the Archiving Facility

Fully Operational System

Today VITO uses the CAF to archive valuable derived earth observation products in addition to the products from the SPOT-VEGETATION sensors. At this moment the CAF mirrors the rolling archive of AATSR and MERIS, contains NOAA AVHRR-products (RD, S1, S10) and higher level products from the MODIS sensor (L1 and L2). We use the data of these sensors in multi mission earth observation applications at VITO.

The CAF is deployed as an independent application on a Linux Server, with acceptance environments both on a Linux Server and on a Windows Server. We integrated the CAF with the CVB Processing Facility. The CVB Processing Facility contains archival and retrieval processes integrated in multi mission processing chains. These archival and retrieval processes interact with the CAF through the Command Line Interface or the Web Service Interface. The end user can also send requests via the Command Line Interface (ssh client) or he can use a graphical user interface (GUI) which requests the services from the CAF via the Web Services Interface.

At any time the operators at VITO are able to monitor the Archiving Facility with the console of the CAF. They can query the current state of service executions and of the service execution stacks. The console also offers a synthetic view of the status and configuration of the running archival instances.

The EMC Networker tool manages the storage on tape for long term data preservation. Generally we make two copies of the products on LTO-tapes for long term data preservation (one archive tape and one clone tape which is stored on a remote location) and one copy on disk for short term data preservation. The lifecycle specifications to define the storage resource (tape or disk) and the time before being automatically moved to an other medium or deleted, can be configured by project via the configuration file.

The CAF file system makes all the objects stored in the archiving database accessible in a virtual read-only mode. The directory structure used to present the objects is customizable and can be adapted whenever needed.

Figure 2 shows a diagram with the operational deployment of the CAF at VITO.

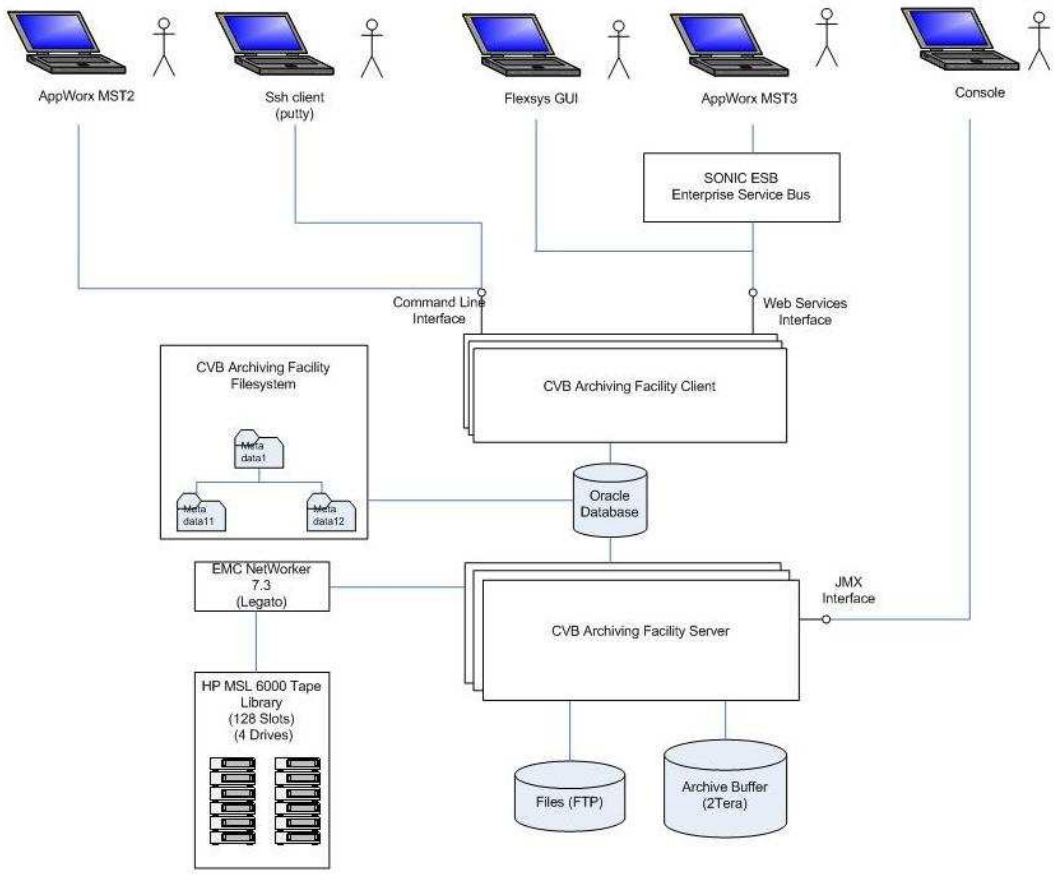

Figure 2: CVB Archiving Facility in operational use at VITO

The CAF is continuously evolving into an up-to-date Archiving Facility used for the ingestion, storage, retrieval and management of heterogeneous data and metadata to ensure the long term data preservation and the accessibility and usability of the preserved earth observation data.

Today we are implementing an activity reporting tool to produce reports on the CAF. For example: monthly reports on the number of ingested objects (globally, per day, per projects, …), the number of objects retrieved per request, the mean time of retrieval, the number of tape failures.

We are investigating the difference in performance and spatial/locator features between the most common databases like Oracle, MySQL and PostgreSQL. And we are also studying new hardware possibilities for the long term archive (tape or disk) and for the middle long term archive where there is a need for faster (online) access.

LONG TERM DATA ARCHIVE FOR PROBA-V PRODUCTS

The lifetime of the SPOT-VEGETATION sensors is not ensured after 2012. To ensure the continuation of low resolution (1km) earth observation products the PROBA-V mission is started. PROBA-V is a new Belgian satellite, built under the authority of ESA which will be launched in April 2012 [3].

VITO is responsible for the development of the PROBA-V User Segment. The main objectives of the PROBA-V User Segment are :

- processing the PROBA-V Level 0 and ancillary data as received from the Secondary Ground Station up to level 3 with a spatial resolution of both 300m and 1 km (for continuity with the VEGETATION data).

- archiving the received and processed data.
- distributing the various products (both at 300 and at 1 km spatial resolution) to the user community.
- assuring the image quality and high level programming of the VEGETATION instrument operations.

Figure 3 shows a high level diagram with the different blocks and interfaces of the PROBA-V User Segment. For the Long Term Data Archive VITO will reuse the Software Application of the CVB Archiving Facility.

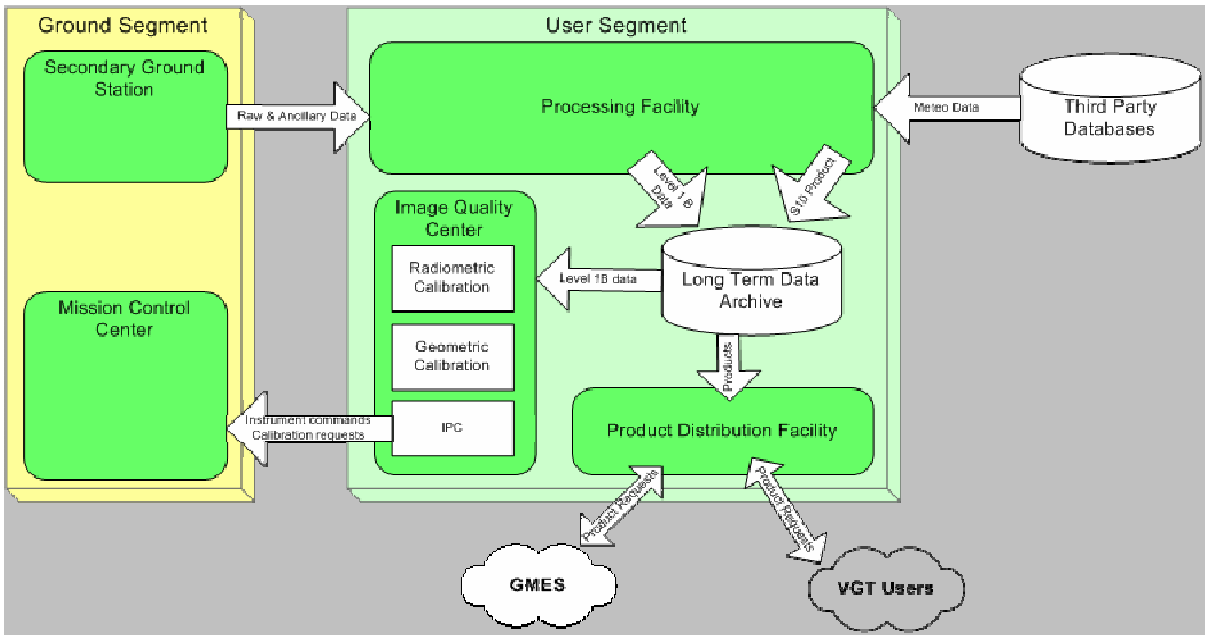

Figure 3: PROBA-V User Segment Blocks and Interfaces

We will re-use the Archiving Facility and integrate it in the Product Distribution Facility to make the archived data discoverable and accessible by Web Services according to standards as proposed by INSPIRE and/or GMES.

CONCLUSION

The CVB Archiving Facility is a flexible and independent facility that covers all main archiving functionality i.e. ingestion, storage, retrieval and management of heterogeneous data and metadata from multi mission earth observation sensors. The two-layer service oriented architecture makes the application independent from the selected database and storage facilities. The external interfaces and especially the Web Service Interface make the Archiving Facility accessible from 'anywhere'.

Currently, we integrated the CAF with the CVB Processing Facility at VITO to archive products from various processing chains for higher level products derived from heterogeneous satellite instruments like SPOT-VEGETION, AVHRR, MERIS, MODIS and AATSR.

In the future we will use the CAF to implement the Long Term Data Archive for the earth observation products from the PROBA-V mission.

We will integrate the Archiving Facility into our Product Distribution Facility to make the archived data discoverable and accessible by Web Services according to standards proposed by INSPIRE and/or GMES.

VITO can also integrate the historical archive of CTIV in the CVB Archiving Facility to create a more standardised and up-to-date way to access and distribute the VEGETATION products for future environmental research.

REFERENCES

[1] – SPOT-VEGETATION: http://www.spot-vegetation.com/

[2] – OAIS: Open Archival Information System, International Organization For Standardization, ISO 14721:2003, http://public.ccsds.org/publications/archive/650x0b1.pdf (version of 11 June 2007)

[3] – PROBA-V: http://www.esa.int/SPECIALS/Proba/SEMD16ZVNUF_0.html