Providing Access to Terabytes of Earth Observation Data in an International Organization - Infrastructure and Services

Paul Hasenohr[1], Armin Burger[2]

[1]Agriculture Unit, Institute for the Protection and Security of the Citizen, Joint Research Centre, European Commission, TP 266, Via E. Fermi 2749, 21027 Ispra (VA), Italy, paul.hasenohr@jrc.it
[2]Agriculture Unit, IPSC, JRC, European Commission, armin.burger@jrc.it

Abstract

The Joint Research Centre (JRC) of the European Commission stores over 60 TB of low, medium, high and very high resolution satellite imagery in an heterogeneous manner. The scientific units within the JRC are the users of these data. Often they are also managing them either by choice or by lack of an alternative solution. An internal project called Community Image Data portal (CID) has been set up to rationalize the situation. One of its core activities was to create a central repository with catalogue, processing and dissemination facilities favouring the use of open source technologies and is now to keep it running and expand it further.

The user requirements have been collected by means of a survey last year and have been combined with the requirements from the IT department, from the management and finally from the CID team.

Earth Observation data users mainly require to have a central catalogue referencing all datasets available at the JRC, as well as a central archive which should have a back-up facility, provide fast file based access to data from both Windows and Linux and also be reliable. Furthermore, data in this central archive should be available via geographic web services and web mapping while using flexible authentication and authorization schemes.

The paper will describe the user requirements, the overall architecture and its technical implementation with focus on the FOSS aspects.

1 Introduction

A large amount of satellite remote sensing data (over 60 TB) is stored at the Joint Research Centre of the European Commission by numerous scientific units in an heterogeneous manner. An internal project called Community Image Data portal (CID) has been set up to rationalize the situation. We will present in this paper in a first part the requirements collected from the data users and the constraints associated with them. Then, in a second part, we will give an overview of the system and services architectures before going into the details of their technical implementation in the last part.

2 Needs and Constraints

2.1 Earth Observation Data at the JRC in 2007

In order to get an overview about the usage of satellite remote sensing (SRS) data at the JRC, a survey (Åstrand et al, 2007) took place in December 2006 – January 2007. The objectives were to make an inventory of existing satellite data and future requirements; to obtain an overview of how data is acquired, used and stored; to quantify human and financial resources engaged in this process; to quantify storage needs and to query the staff involved in image acquisition and management on their needs and ideas for improvements.

From this survey, it appeared that in 2006 an annual 15 person-years spread over around 20 projects are placed on image acquisition and data management for a global expenditure on SRS data of 7.2 M€. At that time, the total amount of image data stored was 55 TB with an expected increase of 80% in the following two years.

With respect to the data itself, a wide variety of platform/sensors of all spatial resolutions are available (mostly MERIS, MSG, AVHRR, Seawifs, Spot Vegetation, Modis, Aster, IRS, Landsat, Spot 1/2/4/5, Eros, Formosat, Ikonos, Quickbird, ERS, Radarsat) in many different file formats, projections and processing levels. Furthermore, several types of licences govern data usage.

2.2 Requirements from our Earth Observation Data Users

The scientific officers interviewed during the survey expressed a strong interest in the set up of a service at JRC level operated by the CID project which would offer a Long Term Archive made of a central catalogue, common storage repository with backup facility and access to data via WMS, WCS, FTP and file-based protocols (NFS for Unix/Linux and CIFS for Windows).

In addition to these general requirements, emphasis was placed on one hand on the possibility for the users to add custom metadata about their data and, on the other hand on the reliability of the service as some users are running operational services such as e.g. crop monitoring for yield estimates over Europe or forest fire monitoring.

As all Earth observation data users at JRC have already systems set up to manage their data, they are interested in a common service only if it is easy to use and its performances are at least equivalent to what they already have in place.

2.3 Constraints implied by the user requirements

Data access from Unix/Linux and Windows systems via native protocols implies that the storage must directly support NFS and CIFS. In addition, the path to a specific dataset shall be the same for both protocols (with the exception of the protocol specific part).

Considering that CID will store tens of terabytes of data, staff must be able to access easily their own data in a flexible way. This is true for both OGC and file-based protocols.

2.4 Constraints related to the data itself

As described previously, several Intellectual Property Rights (IPR's) apply to data available at JRC and they must be enforced accordingly via authentication and authorization mechanisms.

In case of a time series (e.g. MODIS or MERIS time series), each dataset part of the time series must be identified individually but the time series as a whole must also be registered in the catalogue and some of its characteristics (i.e. starting and ending data) shall be updated as required.

Constraints related to the management of mosaics such as Image 2006 are similar to those for time series.

A challenge has been to provide the users with a system able to accommodate most kinds of data available at the JRC (single datasets, time series, mosaics, derived products).

2.5 Constraints related to the environment

From the start-up, the project had to tackle limitations in staff and budget allocation while at the same time it had to go operational quickly.

Having data accessible at high speed from both the Internet (via web services or file download) and the internal network implied to comply with some very strict security measures established by the network security team.

3 Architecture

3.1 Main guidelines

We decided to use Open Source solutions over proprietary ones every time it was possible for several reasons:

- Constraints related to the environment as explained previously
- Experience of the staff with Open Source products
- Successful use of Open Source in previous projects
- Bad experiences in the past with proprietary software (technical support useless or not timely available)
- Possibility to easily test several products in order to choose the most suitable
- Open Source software did not require long administrative procedures for procurement of every single piece of software
- Possibility to combine several specialized pieces of software together, instead of purchasing one single product and to adapt our needs to the possibilities offered by this proprietary software.

Furthermore, CID intends to set up a Long Term Archive. Therefore it is convenient not to be bound to commercial products which might be discontinued at any point in time.

3.2 System architecture

In order to address user requirements about reliability with minimal downtime and to be in line with the concept of a Long Term Archive, all systems have been set up in High Availability (HA). All services are implemented as master/slave or in load balancing. The underlying hardware is located in a data centre with secured power supply, redundant network connections and internet access.

One of the requirements of the network security team has been to place all CID systems in a dedicated network container (half class C subnet) isolated from other JRC networks and accessible only through a reverse proxy supporting HTTP/HTTPS/FTP.

In order to reduce the number of interfaces towards the external IT environment we set up redundant DNS, NTP, email servers to be used by all machines within the CID subnet. These servers are then communicating with servers outside the CID subnet.

Accessing data via CIFS and NFS with identical permissions for both protocols requires the use of NFSv4. Therefore providing file based access to the data requires setting up a Windows Domain Controller for CIFS, Kerberos for NFSv4 and an LDAP user directory for both. The file servers are provided by Network Attached Storage from NetApp.

To keep the running costs low and to facilitate commissioning of new servers as needed for end-user services (web or ftp servers) we decided to use virtual servers. All pre-deployment servers are virtual. Core base services such as DNS, NTP, Windows Primary Domain Controller, Kerberos Admin and one Kerberos Key Distribution Center rely on physical machines as they prove to be more reliable than virtual ones.

3.3 Services Architecture

3.3.1 Authentication and Authorization Service

The authentication and authorization services ensure that all access to data in the Image Portal is taking into account existing IPR's. This requires a fine-grained definition of user credentials and permissions, depending on data properties (like platform type, geo-location, date, etc.) and access protocol. This service is tightly integrated with the portal catalogue database (see section below).

3.3.2 Portal Catalogue and Image Search

All images that shall be made available via the Image Portal need to be imported into the system via a data loading framework. It allows image upload and registration both via a desktop interface and via Web services in batch mode. During data loading all possible metadata are extracted from the images. Together with additional data descriptions specified by the user these metadata are stored in a database that forms the core of the portal catalogue.

Various services are set up to allow scientific users to search for images in the catalogue and to get access to the image data. The main entry point to the services is performed via a single user interface. The catalogue search had to be suitable for both experienced researchers in remote sensing and standard users. This is achieved via different search modes. The expert mode allows virtually any characteristics of an image to be specified in the search.

An additional possibility to query the metadata catalogue is available via an OGC Catalogue Service for the Web (CSW). This enables the metadata harvesting by external services based on the CSW protocol.

3.3.3 Concept of Portfolios

The implementation of data access services needed to take into account constraints related to existing data IPR's as explained in section 2.4. For access via file system protocols NFS and CIFS this is realized through the generic functionality of access control lists (ACL) of the storage solution.

To achieve this also for OGC protocols (WMS & WCS) and FTP the concept of *portfolios* has been implemented. Portfolios contain references to the data in the catalogue. They are defined and managed in the main Image Portal interface. Portfolios for WMS and WCS tackle some problematic issues with regard to the nature of image data: every image in the catalogue represents a layer in WMS/WCS notation. With ten thousands of images available, a default 'GetCapabilities' request would typically exceed the parsing capabilities of most OGC clients, and an orientation in the returned layer list would be nearly impossible. Here the concept of portfolios comes into play: portfolios let the users define the dataset layers they would like to have available via OGC services. In addition, the portfolios allow proper authorized access to restricted data. Only data a user is allowed to access via WMS/WCS can be added to a portfolio. This circumvents missing authentication features of WMS/WCS protocols. The access to data via FTP is managed in a similar way via FTP portfolios.

3.3.4 Virtual File Structure

The data in the Image Portal is stored in a standardized physical file structure according to image metadata (platform/sensor, date), regardless of data ownership. While this facilitates the navigation through the file structure for every user, it causes inconveniences for users who need to have access to data purchased for their specific project using their custom data structure. A possibility to define a custom *virtual file structure* has been set up in order to cope with this issue.

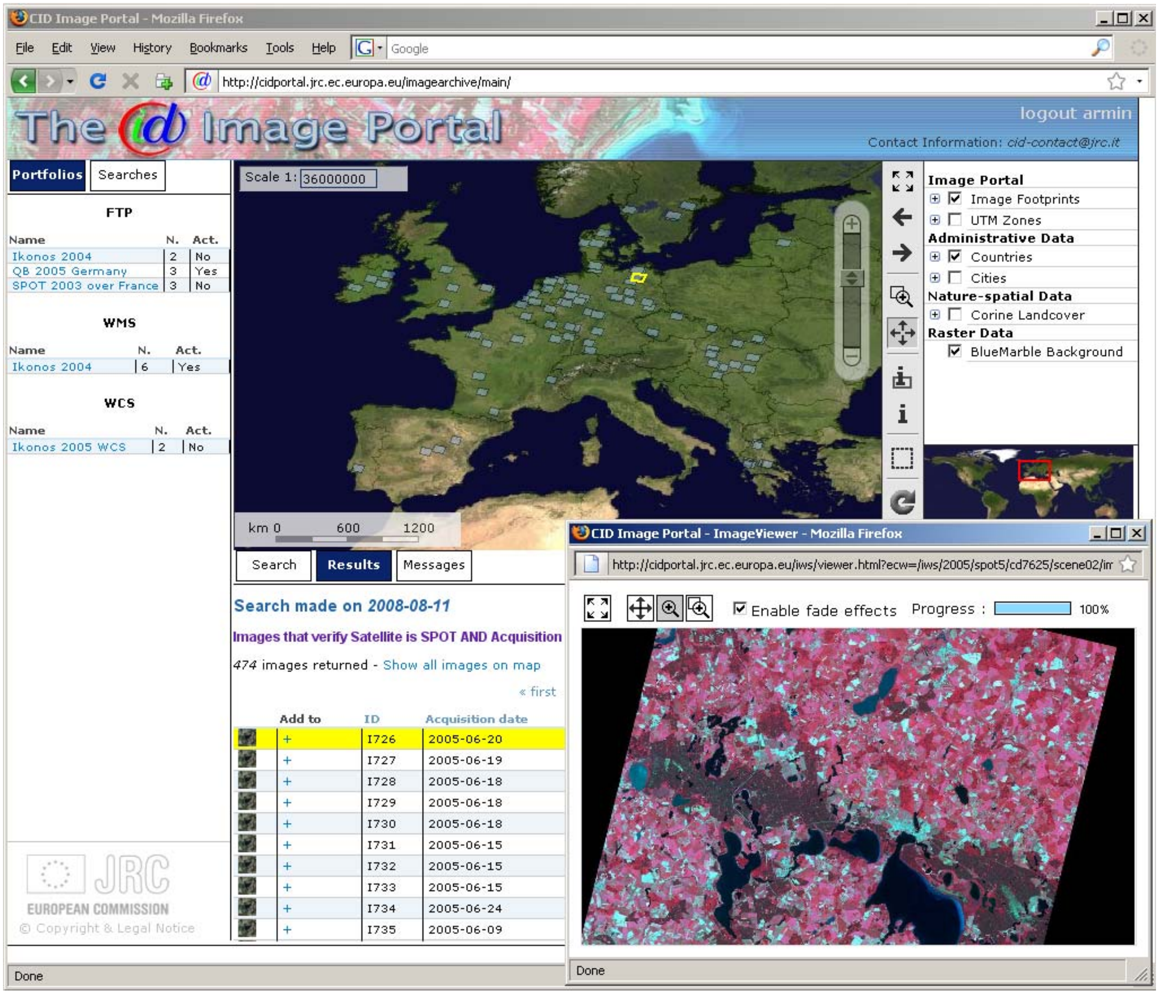
4 Technical implementation

4.1 Overview of end user services

4.1.1 Image Portal Interface

The Image Portal Web interface is implemented on top of the Drupal PHP framework (Drupal 2008) with mapping functionality based on UMN MapServer (MapServer 2008) and p.mapper (p.mapper 2008). The portal interfaces with the metadata database running on PostgreSQL (PostgreSQL 2008) with PostGIS (Refractions Research 2008) as spatial extension. Users can search for imagery in the catalogue based on various data properties, like platform, image resolution, acquisition date, data description and geo-location (see figure 2). An expert search allows a flexible definition of search parameters using nearly all available metadata fields from the database as search filters.

Search results are automatically saved for a limited time and can be modified later on and permanently stored in the user's profile. Search and thumbnail preview will be available to the public. More advanced access methods like full resolution viewing, download or access via OGC services will be available according to data IPR's and user credentials defined in the authorization application.


CID Image Portal - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Google
http://cidportal.jrc.ec.europa.eu/imagearchive/main/
The cid Image Portal
logout armin
Contact Information: cid-contact@jrc.it
Portfolios
Searches
Scale 1: 36000000
km 0 600 1200
Image Portal
☑ Image Footprints
□ UTM Zones
Administrative Data
☑ Countries
□ Cities
Nature-spatial Data
□ Corine Landcover
Raster Data
☑ BlueMarble Background
Search
Results
Messages
EUROPEAN COMMISSION
JRC
© Copyright & Legal Notice
Done
CID Image Portal - ImageViewer - Mozilla Firefox
http://cidportal.jrc.ec.europa.eu/iws/viewer.html?ecw=/iws/2005/spot5/cd7625/scene02/im
☑ Enable fade effects
Progress : 100%
Done

FTP

Name	N.	Act.
Ikonos 2004	2	No
QB 2005 Germany	3	Yes
SPOT 2003 over France	3	No

WMS

Name	N.	Act.
Ikonos 2004	6	Yes

WCS

Name	N.	Act.
Ikonos 2005 WCS	2	No

Search made on *2008-08-11*

Images that verify Satellite is SPOT AND Acquisition

474 images returned - Show all images on map

« first

	Add to	ID	Acquisition date
	+	I726	2005-06-20
	+	I727	2005-06-19
	+	I728	2005-06-18
	+	I729	2005-06-18
	+	I730	2005-06-18
	+	I731	2005-06-15
	+	I732	2005-06-15
	+	I733	2005-06-15
	+	I734	2005-06-24
	+	I735	2005-06-09

Figure 2. Image Portal Web interface

4.1.2 Data Access via OGC Services

Access to repository data via OGC protocols WMS and WCS is managed via portfolios. After an image search the user can add images from the result list to his portfolios. The user can only add data to the different portfolio types that he is allowed to access via the respective protocol. Every image added will be available as a separate layer in the OGC service. On activation of the portfolio a temporary configuration file (i.e. a map file for UMN MapServer) is created. The URL to the service identifying the corresponding map file is displayed and can be inserted in a WMS/WCS client. This URL for a WMS/WCS portfolio remains constant, even after adding or removing images. The activation of the OGC services requires a login via the Image Portal interface. The services remain active until a predefined idle time-out is reached and the map files are deleted.

The WMS/WCS services themselves are implemented in Python MapScript (MapScript 2008) based on the OWSRequest class, using as configuration input the map files dynamically created for the portfolios. This set up allows a very flexible management of OGC services and handling of the WMS/WCS requests before sending the server response back to the client.

Furthermore, the usage of the MapServer framework with GDAL (GDAL 2008) as library for reading raster data provides a probably unbeaten solution with regard to the number of supported raster formats. This avoids the necessity to convert and duplicate data in order to visualize or serve them to the clients. The section 'MARS Stat ImageServer' in Genovese et al, 2007, provides an example for the usage of this feature.

4.1.3 Data Access via FTP

The main file access method to the original (especially raw and unprocessed) data files for external users is via FTP download. In a similar way as for OGC services the user creates FTP portfolios and adds images to them from the result list of a search. The portfolio activation launches the creation of symbolic links to the image directories below the user's FTP home directory. The URL to the FTP site is displayed and can be copy-pasted in an FTP client. The FTP links are deleted after a predefined time out.

4.1.4 Catalogue Service for the Web (CSW)

The CSW solution used in the Image Portal is based on GeoNetwork (GeoNetwork 2008). Metadata for all images referenced in the portal database are automatically exported in ISO 19115/19139 XML format to a WebDAV directory that is regularly inspected by GeoNetwork. Newly added datasets are identified by GeoNetwork and propagated via its CSW protocol.

4.2 User requirements addressed in details

4.2.1 Custom File Structure

As mentioned in section 3.3.4 users required the possibility to access images in the repository following a custom file structure that is defined for their particular purposes and project needs. Since the physical file structure cannot be changed and data shall not be duplicated, a virtual file structure approach needed to be implemented, providing the same structure to both Windows and Unix clients. None of the analysed existing virtual file system implementations were regarded as suitable for this task. The solution chosen was the definition of the virtual file structure via Unix-style symbolic links created and managed with generic PHP functionality inside the Drupal framework.

This virtual file structure is mainly rule-based and relies on data search definitions. Its definition is integrated into the Image Portal Web interface and works in a similar way as the definition of the portfolios. Users can specify that the resulting images of a search are linked below a user-defined virtual directory via symbolic links to the physical files. Any image that will be added later to the portal and that falls under the search criteria defining the virtual directory will be automatically linked via a Web service called by the data loading system. In addition, the user interface allows to add single images from the result list to virtual directories.

5 Conclusion

The chosen Open Source software pieces proved to be good choices for the required tasks and under the existing environment. The Open Source systems were at least equal or superior to proprietary software, providing stable, flexible and powerful solutions. Support by the user community was in most cases available quickly and valuable. A big advantage to proprietary software is the direct support by the main developers of the Open Source software which leads to better and more precise identifications of problems together with possible solutions.

The modular approach using a combination of smaller components to perform a bigger task leads to better manageable, maintainable and more sustainable systems. One drawback of this approach, however, is the more complicated selection of the single components and their integration. This offers a wide field for specialized commercial support for Open Source software.

References

Åstrand, P, Burger, A, Hasenohr, P, Loekkemyhr, P 2007, CID Survey Support, ISSN 1018-5593, European Commission Joint Research Centre, Ispra/Italy

Drupal 2008, A Web development framework and CMS, viewed 14 August 2008, <http://www.drupal.org>

GDAL 2008, Geospatial Data Abstraction Library, viewed 14 August 2008, <http://www.gdal.org/>

Genovese, G, Baruth, B, Royer, A, Burger, A 2007, MARS Stat Action of the European Commission, Crop and Yield Monitoring Activities, *GEO Informatics*, vol. 10, no. 4, pp. 20-22.

GeoNetwork 2008, A standards based, Free and Open Source catalog application, viewed 14 August 2008 <http://geonetwork-opensource.org/>

MapScript 2008, MapScript API Reference, viewed 14 August 2008, <http://mapserver.gis.umn.edu/docs/reference/mapscript/>

MapServer 2008, An Open Source development environment for building spatially-enabled internet applications, viewed 14 August 2008, <http://mapserver.gis.umn.edu/>

p.mapper 2008, A PHP MapScript mapping framework, viewed 14 August 2008, <http://www.pmapper.net/>

PostgreSQL 2008, The worlds most advanced open source database, viewed 14 August 2008, <http://www.postgresql.org/>

Refractions Research 2008, PostGIS - a spatial extension to PostgreSQL, Victoria, viewed 14 August 2008, <http://postgis.refractions.net/>