Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

GEOSUD SDI: accessing Earth Observation data collections with semantic-based services

Mathieu Kazmierski
UMR ESPACE-DEV (IRD)
mathieu.kazmierski@ird.fr

Jean-Christophe Desconnets
UMR ESPACE-DEV (IRD)
jean-christophe.desconnets@ird.fr

Bertrand Guerrero
UMR ESPACE-DEV (IRD)
bertrand.guerrero@ird.fr

Dominique Briand
UMR ESPACE-DEV (IRD)
dominique.briand@ird.fr

Abstract

Ecosystems and territories are complex systems involving multidisciplinary approaches on different time scales and different locations. Yet, mastering the spatial information on these systems is critical to lead a relevant environmental research, as well as addressing efficient public policies. Although the volume of Earth Observation (EO) data generated these last years greatly increased, their usages are still too limited when it comes to environmental issues. So as to remedy to this underutilisation, the GEOSUD (GEOinformation for Sustainable Development) project had undertaken to deploy a national Spatial Data Infrastructure (SDI) to ease access to high resolution and very high resolution satellite images for public stakeholders and scientists. This paper gives an overview of this infrastructure and places emphasis on the innovative components that fit the specific needs in terms of images discovery of the GEOSUD community. These components relies on domain controlled vocabularies, which can assist both experts and non-experts in the field of remote sensing in their search of the appropriate material to suit their needs.

Keywords: satellite images, discovery services, metadata, faceted search, geoprocessing

1 Introduction and background

Ecosystems and territories are complex systems requiring multidisciplinary approaches on different time scales and different areas. Yet, mastering the spatial information on these systems is critical to lead a relevant environmental research, as well as addressing efficient public policies. Considering this statement, European initiatives such as the INSPIRE directive or the even broader Open Data initiative[1] have been set up in the past few years and are now well established. Thus, spatial data usage of vector data from large and well-known repositories has considerably developed within the targeted community of users in the past few years.

Earth Observation (EO) data strengthen these repositories and are offering observations data at relevant spatial and spectral resolutions with high acquisition frequency that eventually allow to carry specific studies on dynamics of territories. Although the volume of EO data generated these last years greatly increased [1], their usages by public stakeholders and scientists are still limited when it comes to environmental issues.

The core problems are well identified. The first concerns the high costs for user licence of satellite images. Indeed, many offers for high resolution or very high resolution images require to pay for using images with a quite restrictive and expensive licence that eventually brings quite substantial financial costs.

Besides, the lack of awareness of what is on offer, regarding to the amount of satellite images, and the varying degrees of capacity and knowledge skills in the field of remote sensing of end-users make their choice of a sensor and associated product challenging. In fact, there are numerous dedicated applications for discovery and access to satellite images although they are provided with hardly comprehensible user interfaces for non-expert audiences. Finally, the multiplicity and the lack of standardisation in nomenclatures (e.g. the multiple names of a processing level depending of the image provider) as well as in image descriptions are major obstacles for users to easily access to a clear view of the wide range of imagery products available on a given territory and to quickly evaluate if it fits their needs.

Issues related to the access of distributed and heterogeneous data are quite common. Usually, it is solved by the deployment of a Spatial Data Infrastructure [2,3]. The COPERNICUS[2] initiative on a European level and the GEOSS [4] on a global level had implemented this principles and give now access to products on a regional, continental or global scale.

In France, the GEOSUD project started in the finding that public stakeholders working in the field of environmental management and public policies underuse satellite images. It had undertaken to deploy equivalent measures as the ones stated above to ease the access to high resolution and very high resolution EO data for public stakeholders and scientists. This project began in 2011 and is led by a consortium of 12 organisations among which public structures, universities, research institutes, companies and spatial data end-user communities. In addition to the acquisition of national annual high resolution coverages the first five years, the satellite images offer will be broadened by a receiving antenna GEOSUD that will allow to program and acquire images from different types of high resolution or very high resolution sensors.

The main goals of the GEOSUD project are to guarantee and ease the access to satellite images, by simplifying their use licences, the discovery and the download of its resources, and in a second phase, by guaranteeing access to on-line geoprocessing that serves the working domains of GEOSUD end-users through an image analysis application. So as to fit the needs of a heterogeneous users community, regarding

[1] European Open Data: https://open-data.europa.eu/en/data/

[2] COPERNICUS : http://www.copernicus.eu/

their comprehension level of remote sensing concepts, this project has to ensure that the different user interfaces and services are adapted to the various degree of expertise.

Moreover, the GEOSUD SDI will be one part of the institutional sector dedicated to satellite image access (called the "Pôle THEIA"). The latter aims at providing a wide range of satellite data on continental surfaces [5]. These data are produced thanks to different projects funded within the French scientific community, among which the main ones are GEOSUD, Postel, Kalideos, Hydroweb, Take Five, Spirit,…

The THEIA infrastructure will be built as a federation of data and services centres for satellite images, in respect of each access conditions and for a broader targeted audience than GEOSUD. Common services are in the heart of this federation, including in particular an image discovery application. It aims to give a unique access point to all available data in the federation, with increased transparency. An identification and authentication common mechanism is considered. The GEOSUD user database would have to be interoperable with the latter.

This paper presents the GEOSUD SDI. In particular, we focus on the innovative components that meet the specific needs in terms of data access and data discovery for non-expert end-users. These semantic components rely on a set of controlled vocabularies.

The paper is organised as follows. After a brief presentation of the GEOSUD context and remind the fundamental principles underlying this SDI in the Section 2, the Section 3 details the two main innovative components of the SDI: the data standardisation and semantic annotation service, and the data discovery application, which make use of annotations to facilitate both the discovery process and the image selection process for end-users. The Section 4 gives an overview of the technical choices that will be implemented in the GEOSUD infrastructure this year. The section 5 concludes this paper by reminding all the expected benefits from this infrastructure for GEOSUD end-user community, the contribution of this SDI in the national infrastructure THEIA as well as of the expected use of high performance computing for large-scale geoprocessing that will be handled as the next step toward innovative services for public policies.

2 Interoperability of access services

The principles that led to the design of the GEOSUD infrastructure is based on the definition of a spatial data infrastructure as proposed by the INSPIRE directive: « the metadata, spatial data sets and spatial data services; network services and technologies; agreements on sharing, access and use ...operated or made available in an interoperable manner »[6]. Thus, a SDI that follows these rules must give access to data through discovery, visualisation and download interoperable services.

The adoption of international standards, that both ensure data harmonisation and access services standardisation, allows on one hand the aggregation of heterogeneous data sources from multiple satellite images providers and, on the other hand, the unification of their description so as to offer a broad and homogeneous vision of available data.

The ISO Technical Comity TC/211 and the Open Geospatial Consortium (OGC) are the main designers of the standards in the field of spatial data and services. In the particular context of Earth Observation data, spatial agencies such as ESA (European Spatial Agency) have highly contributed to define these specifications (e.g. Heterogeneous Mission Accessibility specifications).

To provide images both to the European environmental management community and to the EO community, we have committed ourselves to take into account the recommendations of the INSPIRE directive as well as the specifications emitted by the EO community when we designed the access components and the underlying metadata models. As for the visualisation and download services, they were designed according the OGC standards: WMS (Web Map Service), WMTS (Web Map Tile Service) for the first one; WCS (Web Coverage Service) for the second one.

3 Enhancing images discovery by enriching metadata from heterogeneous sources

Images discovery web-services make use of the information contained in the metadata. Most of the existing web-services for images discovery, since they are based on standards, whether from OGC, as the Catalog Service for the Web (CSW) standard [7], or as OpenSearch with its EO extension (EO OpenSearch) [8], use a reduced set of metadata. On the one hand, this reduced set does not reflect the richness offered by metadata of image providers. On the other hand, it often offers unsatisfactory expressiveness to build requests on a specific characteristic of an image e.g. its spatial resolution or spectral bands. It also limits the results filtering and ranking, which are critical factors when the web-service gives access to a large number of images. Moreover, it offers little if any metadata on the image semantic, which may be of key importance for the selection process depending on its intended purpose [9].

3.1 Abstract metadata model for EO metadata insertion

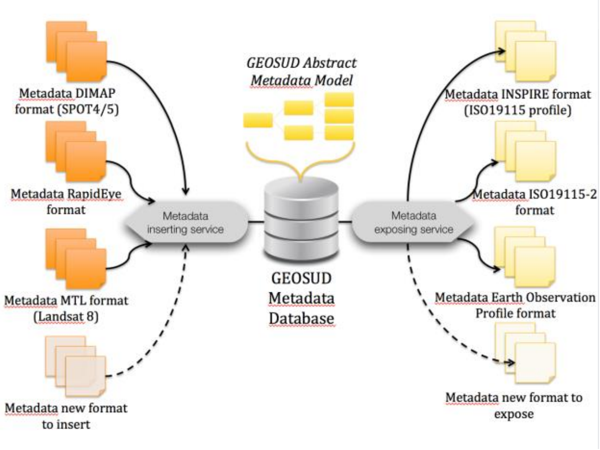
Image providers are given metadata in non-generic models: DIMAP for SPOT and Pleiades images, MTL for Landsat. Other producers adopted metadata description standards such as ISO 191115-2 [10] or the OGC Earth Observation Profile [11].

Moreover, the GEOSUD SDI addresses as well to the EO users community than to other thematic communities. Then it must provide metadata and interoperable discovery services for these communities, by assuring access to metadata compliant with both INSPIRE and OpenSearchGeo Spatial and Temporal Extensions specifications.

In this context, the multiplicity of input and output formats requires many transformations. Many methods are offered to deal with heterogeneous metadata. This reference [12] describes some of them to assure metadata interoperability at the schema level. The switching-across method appears to be the most efficient in our case. It has been developed from the crosswalking method, which consists in the mapping of syntactic and semantic elements from one model to another.

The latter works well when the number of metadata models is relatively low, what is not the case in our context. So as to make the crosswalking more efficient when input and output models are numerous, the switching-across method consists in channelling transformations through a switching schema from the input models to the output models. By doing so, it limits the amount of transformations by avoiding model-to-model mapping. Thus, so as to minimize costs and effort of transforming metadata, the adopted method is to base the transformations on an abstract model, which is given a "switching-across" role (see Figure 1). In later stage of the project, the insertion of images from new sensors will be possible and will be eased by this approach.

Figure 1: Abstract metadata model GEOSUD for the insertion and export of metadata in various standardised or not models


Metadata DIMAP format (SPOT4/5)
Metadata RapidEye format
Metadata MTL format (Landsat 8)
Metadata new format to insert
GEOSUD Abstract Metadata Model
Metadata inserting service
GEOSUD Metadata Database
Metadata exposing service
Metadata INSPIRE format (ISO19115 profile)
Metadata ISO19115-2 format
Metadata Earth Observation Profile format
Metadata new format to expose

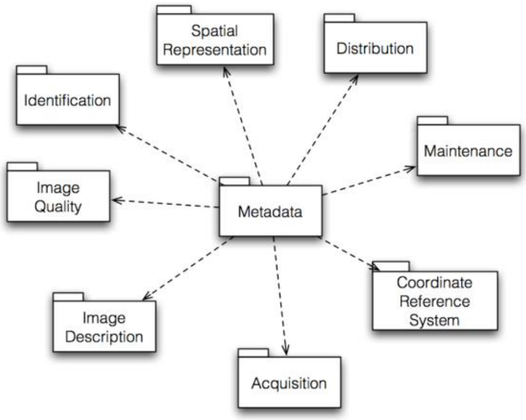
Source: UMR ESPACE-DEV, IRD

In the Figure 2, we present a general view of the GEOSUD abstract model. So as to cover a broad range of functionalities (discovery, visualisation, processing, perennial archiving), it gives a great deal of information. It covers information on image identification such as its *geographicalExtent*, its description (*imagingCondition*, *processingLevelCode*, *pixelResolution*), its acquisition condition (*GSD_instrument*, *GSD_SpectralBand* classes). It also provides content on the evaluation of the image quality (*GSD_QuantitativeResults*, *GSD_QualityReportDocument* classes) and their lineage (*GSD_Lineage* class).

It allows designing lasting components to insert metadata in the SDI and will ease the metadata insertion from new sensors.

The abstract metadata model is deployed through the metadata insertion service. Mapping schemes between source models (DIMAP, MTL) to the abstract model are also provided. The insertion service read the producers metadata and execute the mapping for each metadata according to its native model. The resulting metadata is stored in a database and exposed on demand in an interoperable format (ISO19115 INSPIRE, EOP) through standardised web-services that fits the user needs.

Figure 2: GEOSUD abstract model (packages view)



Source: UMR ESPACE-DEV, IRD

So as to illustrate the use of the abstract model in its switching schema role, we give an extract of a mapping between models, implemented in the insertion service of the GEOSUD SDI (see Table 1). These mappings assure the transformation of elements based on the DIMAP model used to describe SPOT or PLEIADES images toward the ISO 19915 INSPIRE model, so to as to expose metadata through a CSW 2.0.2 AP ISO discovery service.

3.2 Automatic annotation of images metadata

The image metadata provided by data producers deal essentially with their intrinsic features such as their footprint, represented as a polygon or a bounding box, their acquisition date or the spectral bands that compose image. They also deal

Table 1: Extract of crosswalks between DIMAP model (SPOT, PLEIADES products), GEOSUD abstract Model and ISO 19115 model

DIMAP elements	Geosud Abstract Model elements	ISO 19115 elements

Min(Dataset_Frame/Vertex/FRAME_LON)
Min(Dataset_Frame/Vertex/FRAME_LAT)
Max(Dataset_Frame/Vertex/FRAME_LON)
Max(Dataset_Frame/Vertex/FRAME_LAT)
	GSD_Identification.geographicalExtent	MD_DataIdentification.extent.geographicElement

Dataset_Sources/Scene_Source/INSTRUMENT Dataset_Sources/Scene_Source/INSTRUMENT_INDEX	GSD_Intrument.instrumentShortName	N/A
Production/PRODUCT_INFO	GSD_ImageDescription.processingLevelCode	MD_ImageDescription.processingLevelCode
Dataset_Sources/Scene_Source/MISSION_INDEX Dataset_Sources/Scene_Source /SENSOR_CODE*	GSD_GridSpatialRepresentation.pixelResolution	MD_Resolution.distance

Source: UMR ESPACE-DEV, IRD

with the image acquisition and image production conditions, such as the processing level (e.g. 1A, 2A, 2B...).

Consequently, if an expert in the field of remote sensing can achieve to search efficiently these images by relying on its knowledge, most end-users will not, since their knowledge of concepts or specific vocabulary would be insufficient.

Adapting a “provider” vocabulary to a “consumer” vocabulary may be necessary to enhance users search experience.

It is indeed simpler to make a request based on a toponym like “I am looking for all the images that cover the city of Toulouse” than to draw a bounding box using its coordinates. In the same way it is often more relevant to give the possibility to “look for all the images containing urban area” when the search purpose is to look for urban dynamics assessments.

To ensure the vocabulary adaptation and the enrichment of metadata, we rely on internal and external controlled vocabulary. To adapt footprints, we rely on the GEONAMES ontology [13]. It gives access through a REST service to all toponyms across the French territory. For example, when inserting a SPOT5 image with the bounding box {north: 44.49321, south: 43.81890, east:-0.26412, west: -1.21798} into the SDI, the insertion service execute the following HTTP request to the GEONAMES API :
http://api.geonames.org/search?north=44.49321&south=43.81890&east=-0.26412&west=-1.21798&username=geosud

It returns all the toponyms and extra-information on each one of them within the specified bounding box (see Table 2).

Table 2. Example of a Geonames API response in XML format to an HTTP search request

Line number	XML response extract
1	<**geonames** style="MEDIUM">
2	...
3	**<geoname>**
4	<**toponymName**>Mont-de-Marsan</**toponymName**>
5	<**name**>Mont-de-Marsan</**name**>
6	<**lat**>43.89028</**lat**>
7	<**lng**>-0.50056</**lng**>
8	<**geonameId**>6433897</**geonameId**>
9	<**countryCode**>FR</**countryCode**>
10	<**countryName**>France</**countryName**>
11	<**fcl**>A</**fcl**>
12	<**fcode**>ADM4</**fcode**>
13	</**geoname**>
14	...
15	**</geonames>**

Source : Geonames

The latter response is then consumed by the insertion service. It extracts the content from the <toponymName> tag and populates fields of the image metadata GEOSUD database. In this case, the toponym name “*Mont-de-Marsan”* is inserted into *GSD_Identification.geographicIdentifier* field.

Based on the same principle, the enrichment of metadata with land cover information relies on the Corine Land Cover 2006 classification [14]. Adapting vocabulary on spatial resolution or on processing level will be taken in charge by another component of the SDI, which will execute these operations simultaneously to the metadata indexing.

3.3 User faceted search application

Discovery applications on which is based the image search are usually complex for they often offer an expert approach to the search process, where a large number of search criteria are offered through non-intuitive interfaces. Moreover, the semantic of the criteria is not always readily understandable for end-users. To overcome these limitations, we choose to base the search process on an interactive filtering mechanism to retrieve information, which is widely used on the Internet: the faceted search [15,16].

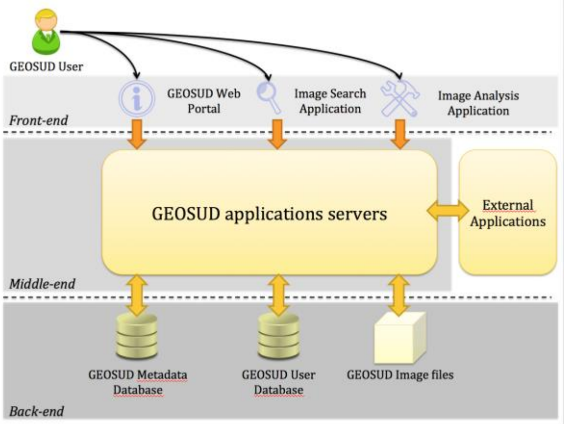
Also referred as faceted navigation or faceted classification, the faceted search is defined as a method to access a data collection by allowing user to explore the latter through selected filters. It is based on a classification system in where multiple categories can be assigned to the same data and where the filtering can be enabled in different ways [17]. In our example, a filter could be an image property : acquisition date, spatial resolution, location,…

The enrichment of metadata and the adaptation of vocabularies during the metadata insertion phase discussed above provide categories for faceted search with a less expert-oriented semantic, close to the various audiences of GEOSUD. For example, we provide a hierarchical facet “spatial location” which rely on administrative toponyms such as region or city names, from which have been extracted the metadata value *geographicIdentifier*.

4 Implementation

So as to deploy these services under the conditions emitted above and to allow external applications to access in a controlled way to GEOSUD data, the logical architecture of GEOSUD SDI will used a widely accepted 3-tier architecture principles (see Figure 3): front-end user applications, middle-end services that gives access to data and a back-end composed of databases and the image files.

Figure 3: Simplified view of the GEOSUD 3-tier architecture



Source: UMR ESPACE-DEV, IRD

5 Conclusion

High resolution and very high resolution EO data have become essential to undertake environmental research and address efficient public policies. The GEOSUD SDI brings an original and comprehensive solution for public stakeholders and scientists by allowing them to access in a standardised way to satellite images from a wide range of sensors.

One of the original aspects of this infrastructure is that it focuses and adapts to the various degree of expertise of its end-users, which is also a major constraint to search efficiently images in their everyday work. The adaptation and enrichment of metadata from image providers by the use of controlled vocabularies (GEONAMES, Corine Land Cover) allow the search process to share a semantic that is close to a non-expert user. This also helps to build a discovery application that is based on these vocabularies. The choice of a faceted-centred discovery mechanism will enhance the user experience and increase the relevance of returned results.

Today, the user needs consist in the exploitation of images. Thus, their analysis is used to build complex environmental indicators that require specific tools (ENVI, eCognition) as well as large computational and data resources. Yet, these tools are still out of reach of a large part of public stakeholders or scientists. Consequently, the next challenge for GEOSUD SDI is to give access to a satellite image-processing platform that fits the latter audience needs with specific processing chain (e.g. detection of nitrate-fixing intermediate crops). For this purpose, an online computational platform combined with high performance computing environment is considered. In this context, we will attach importance to tackle the barriers created by the various level of expertise of GEOSUD end-users, either in the geo-processes discovery or their configuration and execution.

References

[1] A.J. Tatem, S.J. Goetz; S.I. Hay. Fifty years of earth observation satellites. American Scientist, 96(5):390-398, 2008/9.

[2] Global Spatial Data Infrastructure : Developing Spatial Data Infrastructures: The SDI Cookbook. Version 2.0. Consulted the 25 January 2004 at : http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf

[3] Yang C., Raskin R., Goodchild M., Gahegan M. : Geospatial Cyberinfrastructure : Past , present and future. *Computers, Environment and Urban Systems* 34(2010). 264-277.

[4] Christian E.J. : GEOSS Architecture Principles and the GEOSS ClearingHouse. *IEEE systems journal*, 2(3). September 2008.

[5] Theia Land data center: M.Leroy, P.Kosuth, O.Hagolle, S.Cherchali, P.Maurel, J.Desconnets. ESA Living Planet Symposium. Edimburgh, UK. 9-13 September 2013.

[6] European Parliament, 2007. DIRECTIVE 2007/2/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Available at: http://eurlex.europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML.

[7] OGC : OpenGIS Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile, Rapport, Open Geospatial Consortium, 2007.

[8] OGC 10-032 : OpenSearch Geo Spatial and Temporal Extensions. 2014

[9] Boisson P., Clerc S., Desconnets JC, Libourel T. : Using a semantic approach for a Cataloguing Service. OTM workshops (2) 2006: 1712-1722. LNCS Springer Heildelberg.

[10] ISO TC/211 : ISO 19115-2 - Geographic information — Metadata —Part 2, Extensions for imagery and gridded data. First edition. 2009

[11] OGC OGC10-157r3 - Earth Observation Metadata profile of Observations & Measurements Standard, 2012.

[12] Chan L.M., Zeng M.L. : Metadata interoperability and standardisation – a study of methodology Part 1. Achieving interoperability at the schema level. *D-Lib magazine* 12(6) ISSN 1082-9873. June 2006

[13] Geonames API web service : http://www.geonames.org

[14] Corine Land Cover WFS web service : http://sd1878-2.sivit.org/geoserver/wfs?

[15] Uddin, M.N., Janecek, P.: Faceted classification in web information architecture: A framework for using semantic web tools. *Electronic Library*, 25(2), 2007

[16] Denton, W.: How to make a faceted classification and put it on the web. See http://www.miskatonic.org/library/facet-web-howto. html (2011).

[17] Laporte, M.-A., Mougenot, I., & Garnier, E. (2013) A faceted search system for facilitating discovery-driven scientific activities: a use case from functional ecology. Workshop S4BioDiv, ESWC 2013, CEUR-WS.org.