Towards a Spatial Data Infrastructure for Big Spatiotemporal Data Sets

CONFERENCE PAPER · APRIL 2015

READS
62

8 AUTHORS, INCLUDING:

Lubia Vinhas
National Institute for Space Research, Brazil
36 PUBLICATIONS **95** CITATIONS
SEE PROFILE

Gilberto Câmara
National Institute for Space Research, Brazil
326 PUBLICATIONS **2,811** CITATIONS
SEE PROFILE

Ricardo Cartaxo Modesto de Souza
National Institute for Space Research, Brazil
65 PUBLICATIONS **807** CITATIONS
SEE PROFILE

Alber Sanchez
National Institute for Space Research, Brazil
12 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Towards a Spatial Data Infrastructure for Big Spatiotemporal Data Sets

Karine Reis Ferreira[1]
Gilberto Ribeiro de Queiroz[1]
Lúbia Vinhas[1]
Gilberto Câmara[1]
Luis Eduardo Maurano[1]
Ricardo Cartaxo Modesto Souza[1]
Alber Sanchez [2]

[1]Instituto Nacional de Pesquisas Espaciais - INPE
São José dos Campos - SP, Brasil
{karine, gribeiro, lubia, gilberto, maurano, cartaxo}@dpi.inpe.br

[2]Institute for Geoinformatics (ifgi)
University of Münster, Heisenbergstraße 2, 48149 Münster, Germany
{alber-sanchez}@uni-munster.edu.de

Abstract. The recent technological advances in geospatial data collection, such as mobile phones, Earth observation and GPS (Global Positioning System) satellites, have created bigger spatiotemporal data sets than ever. This novel scenario has motivated new infrastructures and technologies to efficiently store, process, analyze and disseminate large data sets. In this paper, we propose a Spatial Data Infrastructure (SDI) architecture for big spatiotemporal data sets and describe a pilot implementation of it. SDI is a sharing platform that facilitates the access and integration of multi-source spatial data in a holistic framework with a number of technological components including policies and standards.

Key words: spatial data infrastructure, big data, spatiotemporal data, database systems.

1. Introduction

The age of big geospatial data has come. The recent technological advances in geospatial data collection, such as mobile phones, Earth observation and GPS (Global Positioning System) satellites, have created massive data sets with better spatial and temporal resolution than ever. Space agencies worldwide plan to launch around 260 Earth observation satellites over the next 15 years. This novel scenario has motivated new infrastructures and technologies to efficiently store, process, analyze and disseminate large spatiotemporal data sets.

A Spatial Data Infrastructure (SDI) is a platform that facilitates the interaction between people and data by providing required technologies, policies and standards (Rajabifard et al., 2002). SDI as a sharing platform aims to facilitate the access and integration of multi-source spatial data within a holistic framework with a number of technological components including policies, standards, access and the interaction between spatial data stakeholders and spatial data (Mohammadi, 2008).

In this paper, we propose a SDI architecture for big spatiotemporal data sets and describe a pilot implementation of it. The SDI is divided in four components: Databases, Web services, Users and Geographical Information System (GIS) tools. We use the term "GIS tools" to refer to different types of software systems that handle geospatial information, such as spatial statistical packages, GIS software libraries, Desktop GIS and web geospatial portals.

2. Spatial Data Interoperability and Standards

Since the beginning of the 2000s, the GIS community has made a serious effort towards spatial data interoperability. The International Organization for Standardization (ISO) and the Open Geospatial Consortium (OGC) have proposed standards to represent and store spatial information in data files and database systems as well as to serve spatial data, metadata and processes via web services.

Geography Markup Language (GML) (OGC, 2007a) and Keyhole Markup Language (KML) (OGC, 2008a) are examples of data formats proposed by OGC for spatial data interchange. Spatial extensions of traditional object-relational Database Management Systems (Spatial DBMS), such as PostGIS and Oracle Spatial, deal with spatial information, vector and raster, in compliance with the OGC Simple Feature Access (SFA) specification (OGC, 2006a; OGC, 2006b).

Regarding web services, we can group the standards in three groups: data, metadata and processing. Web service specifications for spatial data include Web Map Server (WMS) (OGC, 2006c), Web Feature Service (WFS) (OGC, 2010) and Web Coverage Service (WCS) (OGC, 2012). Catalogue Service Web (CSW) is a standard to publish and search collections of metadata for spatial data, services and related objects (OGC, 2007b). Specifications for geospatial processing services include Web Processing Service (WPS) (OGC, 2007c) and Web Coverage Processing Service (WCPS) (OGC, 2008b).

Standards play a key role in SDIs. They assure spatial data interoperability among distinct SDIs as well as between SDIs and GIS tools. Nowadays, many data providers throughout the world have created their own SDIs, organizing and disseminating their geospatial data sets and metadata on the Internet using well-known OGC standards. At the same time, many GIS tools, such as spatial statistical packages, desktop GIS and web portals, have became compliant with OGC standards. They are able to access and combine spatial data sets from different data sources via OGC web services.

3. The SDI Architecture and its Pilot Implementation

This section presents the proposed SDI architecture, shown in Figure 1, and describes its pilot implementation. The SDI has four components: Databases, Web services, Users and GIS tools.

3.1. Databases

The big spatiotemporal data sets, vector and raster, as well as their metadata are stored in multiple server-side databases. Vector data is stored in spatial extensions of object-relational DBMSs (Spatial DBMS) compliant with the OGC SFA specification. In the pilot implementation, we are using the well-known spatial DBMS PostGIS[1]. It is an open source spatial extension for PostgreSQL object-relational database system that can store, index, query, transform and process big amount of vector spatial data sets efficiently.

For raster data, such as remote sensing images, we propose the use of array databases. Array databases organize data as a collection of arrays, instead of tables used in object-relational DBMSs. Arrays are multidimensional and uniform, as each array cell holds the same user-defined number of attributes. Attributes can be of any primitive data type such as integers, floats, strings or date and time types. To achieve scalability, array databases strive

[1] Available at: http://www.postgis.net/

for efficiency of data retrieval of individual cells. Examples of array databases include RasDaMan (Baumann et al, 1999) and SciDB (Stonebraker et al, 2013).

In the pilot implementation, we are using the array database SciDB. SciDB splits big arrays into chunks that are distributed among different servers; each server controls a local data storage. It provides an efficient storage mechanism based on chunks and vertical partitions. Array databases have no semantics, making no distinction between spatial and temporal indexes. Thus, we have designed a collection of relational tables to store all the metadata needed by spatial applications built on topo of SciDB.

We prepared, organized and stored a set of spatiotemporal information in the pilot SDI. In SciDB, we stored MODIS images. We have selected 31 grid tiles covering the South America for three MODIS products: MOD09Q1, MOD13Q1 and MCD43A4 (Rudorff et al., 2007). An array holding more than 21,300 images, from 02/18/2000 to 08/13/2014, has been created for MOD09Q1. Based on this same design we are also building arrays for product MOD13Q1 and product MCD43A4.

In PostgreSQL/PostGIS databases, we stored spatiotemporal data sets produced by INPE, such as deforested regions detected by PRODES (Monitoring of Brazilian Amazon Rainforest) and DETER (Real Time Deforestation Detection System) projects, the land use of the deforested areas identified by the TerraClass project and fire spots detected from satellite images. PRODES has been monitoring deforestation for the whole Brazilian Amazon yearly since 1988, whereas DETER has been producing near real-time deforestation and forest degradation alerts for more than 5 million Km^2 in the Brazilian Legal Amazon. The TerraClass project classifies the deforested areas following the thematic categories: agriculture, clean pasture, woody pasture, pasture with exposed soil, regeneration with pasture, second-growth forest, occupations mosaic, mining and urban area.

3.2. Web services

To disseminate the databases on the Internet, we propose the use of web services for raster and vector data sets as well as for their metadata. In most cases, the well-established OGC standards for web services, such as WMS, WFS, WCS and CSW, are suitable for making the data sets and their metadata available on the Web. In some particular cases, we need to define and create web services dedicated for specific tasks.

In the pilot implementation, we are using the GeoServer software[2] to create WMS and WFS web services from a PostgreSQL database with the PostGIS extension. GeoServer is an open source software server that allows users to create web services from spatial DBMS databases easily. To create, edit and serve metadata via CSW service, we are using the open source tool GeoNetwork[3] and the metadata structure recommended by the Brazilian Spatial Data Infrastructure (INDE – *Infraestrutura Nacional de Dados Espaciais*) initiative (CONCAR, 2009). INDE established in November 2009 a profile of the ISO standard 19115:2003 Geographic Metadata. GeoNetwork also works in an integrated way with PostGIS databases.

For raster data, we are augmenting the support of web services through the development of OGC compliant web services on top of SciDB. For instance, a dynamic visualization service, based on WMS Application Profile for EO Products, is under development for visualizing remote sensing data as dynamic maps from 3D arrays (space + time) stored in SciDB. An interface based on Web Coverage Service 2.0.1 is another ongoing effort to bridge

[2] Available at: http://www.geoserver.org/
[3] Available at: http://www.geonetwork-opensource.org/

the gap between an array database such as SciDB and GIS tools. Other development includes in-house web services designed for handling time series of remote sensing imagery.

Besides the services described above, we propose the development of a web services responsible for discovering data needed by the user. This service must access and combine metadata about all data sets stored in the SDI and help users to find what they require.

3.3. Users

Users can access the spatiotemporal data sets through GIS tools or web geospatial portals. In this SDI, we define two types of users, *internal* and *external*. Internal user refers to people that work on building the databases. They are responsible for organizing, validating and putting spatiotemporal data sets into the SDI databases. Besides that, they are in charge of informing all metadata about these data sets. Internal users have direct access to the SDI databases and their metadata. External users are people that search and access the SDI databases through web services. They do not have direct access to the databases; they only obtain data and metadata subsets that we make available on the Internet via web services.

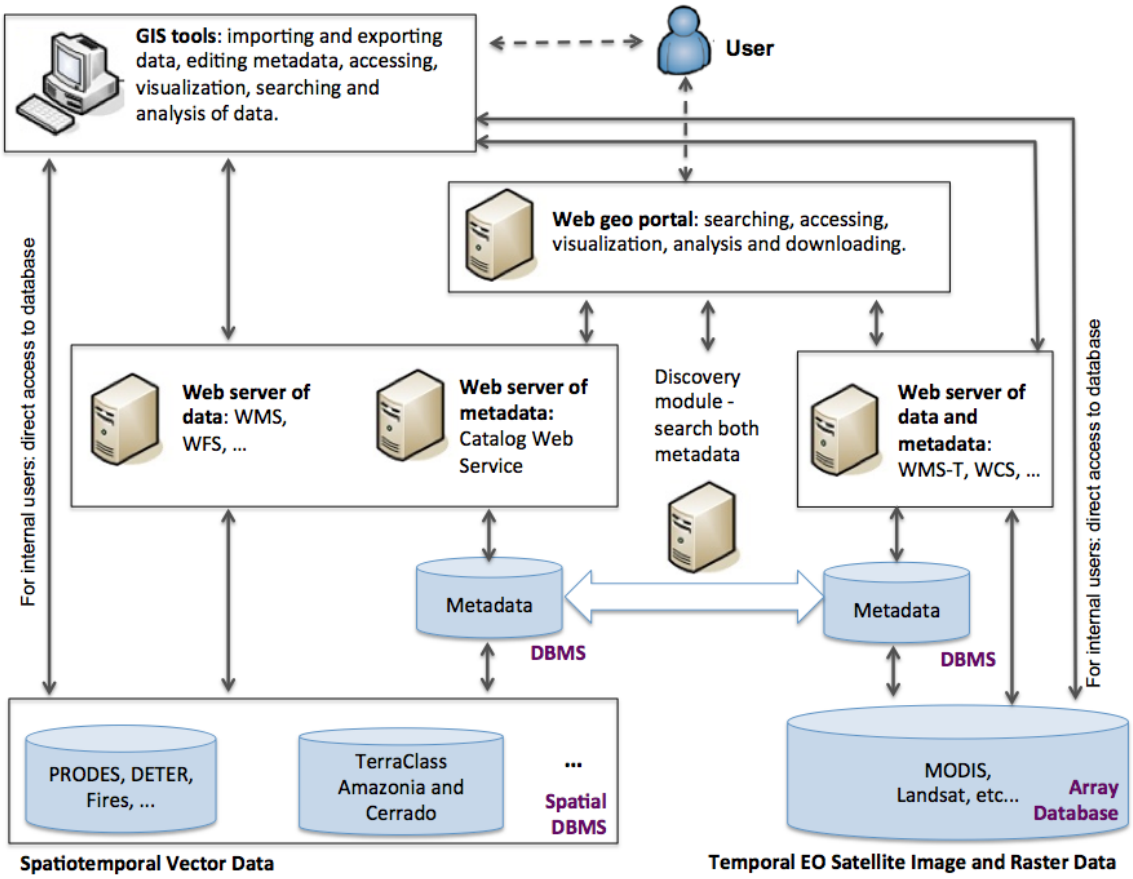

Figure 1. The Spatial Data Infrastructure (SDI) Architecture

3.4. GIS Tools

To efficiently deal with big spatiotemporal data sets, GIS tools must provide the following features:

1. **Mechanisms to access and combine data sets from different types of data sources**. These tools must be able to access data sets from database systems, data files as well as web services and to process them in an integrated way. In these cases, they must

provide mechanisms that split the processing into small parts, taking into account the resources available in each data source.

2. **Take advantage of all database systems and web services resources**. These tools must use as much as possible all resources provided by each type of server-side database systems and web services to deal with spatiotemporal data efficiently. For example, it is important to take advantage of all spatial DBMS functionalities for spatial vector data treating as well as of array database systems resources to handle multidimensional raster data.
3. **Server-side processing mechanisms**. GIS tools must allow users to visualize, process and analyze data in the server-side, without needing to download big amount of data from servers to user local machines.
4. **Script language to express complex processing**. These tools must allow users to express complex processing through script languages, such as Python[4] and LUA[5]. Graphical User Interface (GUI) is not sufficient for users that want to execute complex processing and analysis.
5. **Spatiotemporal data handling**. GIS tools must represent and treat the temporal component of the geospatial data sets. They must provide dynamic visualization using new techniques to highlight the temporal dimension of these data sets. Besides that, these tools must have new algorithms and functions to process and analyze spatiotemporal data, such as classifiers for temporal images.

In the pilot implementation, we are improving the TerraLib GIS library and the TerraView system to provide the five features described above and, therefore, to deal efficiently with big spatiotemporal data sets. TerraLib is a C++ software library base to develop geographical applications and TerraView is a general-purpose GIS built using TerraLib (Camara et al. 2008). Both are developed by INPE and are available as free and open source systems.

TerraLib and TerraView are now able to access and combine data from different kinds of data sources, such as web services, database systems and files. Besides that, they are able to represent and handle spatiotemporal information based on the data model proposed by Ferreira et al (2014). This data model defines three spatiotemporal data types, *time series*, *trajectory* and *coverage*, grounded on *observations*. It also defines the data type *event* that can be derived from these three types.

We are also developing a web geospatial portal called TerraBrasilis, presented in Figure 2. This portal provides a web interface to access, query, visualize, process and download the spatiotemporal data sets stored in our pilot SDI. It is being developing using TerraLib library and following the five features described above.

Figure 2 shows TerraBrasilis accessing and presenting the deforested regions detected by PRODES project from 2011 to 2012. In this case, the portal is using the OGC WMS service that was built on top of the PostGIS database that contains this data set. This portal provides a slider (Figure 2 on the top and right side) that allows users to define a temporal filter, visualize the data sets dynamically and download them.

4. Final Remarks

In this paper, we propose a SDI for big spatiotemporal data sets and describe a pilot implementation of it as a proof of concept. The pilot SDI is being developed using only free

[4] Available at: https://www.python.org/
[5] Available at: http://www.lua.org/

and open source systems, such as PostGIS/PostgreSQL, SciDB, GeoServer, GeoNetwork and TerraLib. We can download them without costs and customize them according to our needs.

In the SDI, we propose the use of spatial DBMS to store vector spatial data and array databases to raster data. Spatial DBMSs are able to store, index, query and retrieve vector data efficiently, whereas array databases are efficient to deal with multidimensional arrays. As future work, we intend to exploit the SciDB array database for vector data too and evaluate its performance for that.

On top of the databases, we propose the use of web services to disseminate them on the Internet. In most case, we use well-established OGC standards for web services, such as WMS, WFS and WCS. This compliance assures spatial data interoperability among distinct SDIs as well as between SDIs and GIS tools. In some particular cases, we need to define and create web services dedicated for specific tasks. An example is an in-house web service responsible for extracting time series from remote sensing images that was built on top of the SciDB database system.

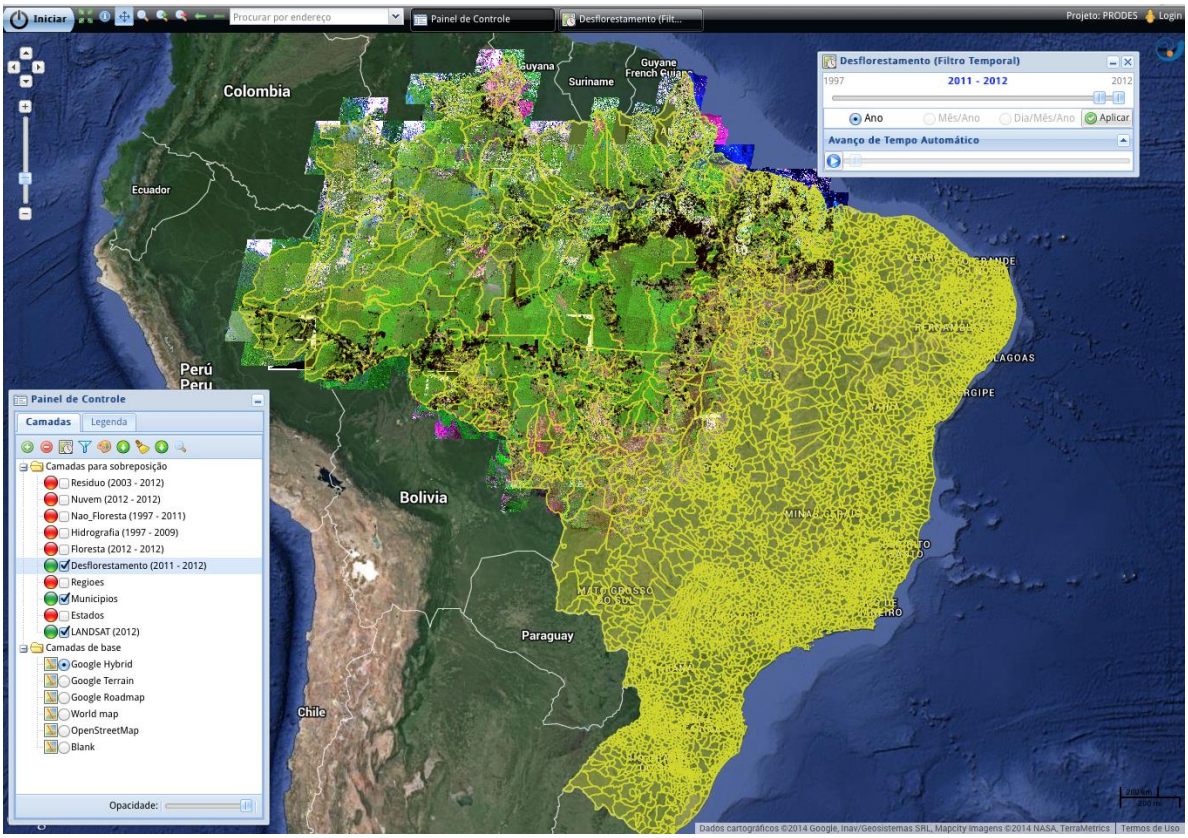

Figure 2 – The web geospatial portal TerraBrasilis

References

Baumann, P., Dehmel, A., Furtado, P., Ritsch, R., Widmann, N.: **Spatio-temporal retrieval with RasDaMan**. In: Proceedings of the 25th International Conference on Very Large Data Bases. pp. 746–749, 1999.

Camara, G.; Vinhas, L.; Queiroz, G. R.; Ferreira, K. R.; Monteiro, A. M. V.; Carvalho, M. T. M.; Casanova, M. A. TerraLib: An open-source GIS library for large-scale environmental and sócio-economic applications. **Open Source Approaches to Spatial Data Handling**. Berlin: Springer-Verlag, 2008.

CONCAR. **Perfil de Metadados Geoespaciais do Brasil (MGB)**, Comissão Nacional de Cartografia (CONCAR): 194p, 2009.

Ferreira, K. R.; Camara, G. and Monteiro, A. M. V. An algebra for spatiotemporal data: From observations to events. **Transactions in GIS**, v. 18(2), p. 253-269, 2014.

Mohammadi, H. **The Integration of Multi-source Spatial Datasets in the Context of SDI Initiatives**. PhD thesis, University of Melbourne, 2008. Available at: http://www.csdila.unimelb.edu.au/publication/theses/hossein-PhD.pdf (accessed in July 2014)

Open Geospatial Consortium – OGC. **OGC WCS 2.0 Interface Standard- Core: Corrigendum**, 2012. Available at: http://www.opengeospatial.org/ (accessed in November 2014)

Open Geospatial Consortium – OGC. **OpenGIS Web Feature Service 2.0 Interface Standard,** 2010. Available at: http://www.opengeospatial.org/ (accessed in November 2014)

Open Geospatial Consortium – OGC. **OGC KML**, 2008 (a). Available at: http://www.opengeospatial.org/ (accessed in November 2014)

Open Geospatial Consortium – OGC. **Web Coverage Processing Service (WCPS) Language Interface Standard**, 2008 (b). Available at: http://www.opengeospatial.org/ (accessed in November 2014)

Open Geospatial Consortium – OGC. **OpenGIS Geography Markup Language (GML) Encoding Standard**, 2007 (a). Available at: http://www.opengeospatial.org/ (accessed in November 2014)
Open Geospatial Consortium – OGC. **OpenGIS Catalogue Services Specification**, 2007 (b). Available at: http://www.opengeospatial.org/ (accessed in November 2014)

Open Geospatial Consortium – OGC. **OpenGIS Web Processing Service**, 2007 (c) Available at: http://www.opengeospatial.org/ (accessed in November 2014)

Open Geospatial Consortium – OGC. **OpenGIS Implementation Specification for Geographic information - Simple feature access - Part 1: Common architecture**, 2006 (a) Available at: http://www.opengeospatial.org/ (accessed in November 2014)

Open Geospatial Consortium – OGC. **OpenGIS Implementation Specification for Geographic information - Simple feature access - Part 2: SQL option**, 2006 (b) Available at: http://www.opengeospatial.org/ (accessed in November 2014)

Open Geospatial Consortium – OGC. **OpenGIS Web Map Server Implementation Specification**, 2006 (c). Available at: http://www.opengeospatial.org/ (accessed in November 2014)

Rajabifard, A., Feeny, M. E., Williamson, I. "Future Directions for SDI Development". **International Journal of Applied Earth Observation and Geoinformation**, v. 4 (1), p. 11-22, 2002.

Rudorff, B. F. T.; Shimabukuro, Y. E. and Ceballos, C. J. **O sensor MODIS e suas aplicações ambientais no Brasil**. Editora Parêntese, 2007.

Stonebraker, M., Brown, P., Zhang, D., Becla, J.: SciDB: A database management system for applications with complex analytics. **Computing in Science & Engineering** v. 15(3), p. 54–62, 2013.