e-Sensing:
Big Earth observation data analytics for land use and land cover change information

Thematic Project (48 months)
November 2014 – October 2018

Project leader:

Prof. Dr. Gilberto Câmara (INPE – Instituto Nacional de Pesquisas Espaciais)

Principal investigators:

Prof. Dr. Leila Fonseca (INPE)

Prof. Dr. Lúbia Vinhas (INPE)

Prof. Dr. Maria Isabel Sobral Escada (INPE)

Prof. Dr. João Vianei Soares (INPE)

Other researchers involved:

Prof. Dr. Karine Reis Ferreira (INPE)

Dr. Gilberto Ribeiro de Queiroz (INPE)

Prof. Dr. Pedro Andrade (INPE)

Eng. Ricardo Cartaxo Modesto de Souza (INPE)

Eng. Luiz Eduardo Maurano (INPE)

MsC. Emiliano Ferreira Castejon (INPE)

Dr. Julio Cesar de Lima d'Alge (INPE)

Dr. Thales Seth Körting (INPE)

Dra. Ieda Sanches (INPE)

Table of Contents

1 Abstract

Earth Observation satellites are the only source that provides a continuous and consistent set of information about the Earth's land and oceans. However, current scientific methods for extracting information for Earth observation data lag far behind our capacity to build sophisticated satellites. These satellites produce massive amounts of data, but only a fraction of that data is effectively used for scientific research and operational applications. Most published scientific results are based on experiments made in small data sets and have not been properly tested and validated. Thus, the project addresses a key scientific problem: *How can we use e-science methods and techniques to substantially improve the extraction of land use and land cover change information from big Earth Observation data sets in an open and reproducible way?* In response to this challenge, our project will conceive, build and deploy a completely new type of **knowledge platform for organization, access, processing and analysis of big Earth Observation data**. We will show that this knowledge platform allows scientists to produce information in a completely new way. Since our platform is fully based on open source software, we will also show that it promotes data sharing and reproducibility of results.

Os satélites de observação da Terra são a única fonte de dados que fornece um conjunto contínuo e consistente de informações sobre nosso planeta Terra. Contudo, os atuais métodos científicos para extração de informações desses dados estão muito aquém da nossa capacidade de construir satélites sofisticados. Embora esses satélites produzam grandes quantidades de dados, apenas uma pequena parte dele é efetivamente usada para a pesquisa científica e aplicações operacionais. A maior parte dos resultados científicos publicados na literatura são baseados em experiências feitas em pequenos conjuntos de dados e assim não foram devidamente testados e validados. Temos então um desafio científico importante: *Como podemos usar métodos de e-science para melhorar substancialmente a extração de informações sobre a mudança de uso e cobertura do solo a partir de grandes conjuntos de dados de observação da Terra em uma forma aberta e reprodutível?* Em resposta a este desafio, **nosso projeto vai conceber, construir e implantar um tipo completamente novo de plataforma de conhecimento para a organização, acesso, processamento e análise de grandes dados de observação da Terra**. Vamos mostrar que esta plataforma de conhecimento permite aos cientistas para produzir informação de forma inovadora. Como a nossa plataforma é totalmente baseada em software livre, vamos também mostrar que promove o compartilhamento de dados e reprodutibilidade dos resultados.

2

3 Statement of the Scientific Problem

3.1 Description of the scientific challenge

Motivation

Humanity is changing rural and urban landscapes at an unprecedented pace. Humans control directly or indirectly more than 50% of the Earth's terrestrial ecosystems (Vitousek et al., 1997). Global population will increase to around 8.5 billion by mid-century. Crop and livestock demand and production will rise by around 40% between 2008 and 2030. Growing pressures on food, water, and energy threaten the planet, at the same time we need to mitigate and adapt to climate change (Beddington, 1999).

Given the size of the global challenges, citizens and politicians are pressing scientists to provide qualified information that would allow wise decisions about the future of our planet. FAPESP is a member of the International Group of Funding Agencies for Global Change Research, known as the "Belmont Forum", and that launched recently "The Belmont Challenge". This document sets outs the priorities for future research on global change (IFGA, 2011). The Belmont Forum considers that we need:

> *"Enhanced environmental information service provision to users through knowledge platforms: Delivering applied knowledge to support innovative adaptation and mitigation solutions, based on the observations and predictions."*

One of the most immediate consequences of humanity's transformation the Earth's ecosystems and landscapes is *land use change*. Thus, one of the responses to the "Belmont Challenge" is to develop a *knowledge platform for land use change*. To build this knowledge platform, we need data from Earth Observation satellites, the only source that provides a continuous and consistent set of information about the Earth's land and oceans. These satellites produce vast amounts of data. The Landsat archive alone holds more than five million images of the Earth's land surface, corresponding to about 1 petabyte of data (Wulder et al., 2012). From 2014 onwards, new satellites from Europe, USA, China, Brazil, and India will each produce in a year as much data as one Landsat satellite in ten years. Most of this data will be freely available, since Brazil, USA and the European Commission have all set up open access data policies for their Earth Observation satellites.

Scientific challenge

Currently, Earth observation data analysis methods lag far behind our capacity to build sophisticated satellites. Currently, most scientific data analysis methods for Earth

observation data are file-based. After data is collected by a satellite, it is downlinked to ground stations of data providers such as INPE and NASA. These data providers offer data to their users as individual files. Scientific and application users download scenes one by one. For large-scale analysis, users need to obtain hundreds or even thousands of files.

For example, the MODIS vegetation index archive for Brazil from 2002 to 2014 has 12.000 independent files. To analyse such large data set, a program has to open each file, extract the relevant data and then move to the next file. The program can only begin its analysis when all the relevant data has been gathered in memory or in intermediate files. Data analysis on large datasets organized as individual files will run slower and slower as data volumes increase. This practice has put severe limits on the scientific uses of Earth Observation data.

The file-based research performed by most Earth observation scientists has led to a lack of large-scale validation and reproducibility. Since researchers cannot assess large EO data sets, they test their algorithms in a limited number of data sets (usually one). Thus, most new methods published in leading journals of the field have never been validated in large-scale data sets. This makes for a huge gap between the scientific results and the operational use of Earth Observation data.

Additionally, despite the inherent nature of Earth observations as systematic data collection, scientists do not organize remote sensing images as space-time arrays due to the lack to computational support. Mostly, they produce land cover maps taking either a single or at most two time references. Scientists thus ignore the time reference inherent to Earth observation data.

As a result of ignoring the temporal nature of Earth Observation data and doing research with limited data sets, most methods developed by scientists of the field do not match the requirements of operational applications. For example, the National Land Cover Database of the United States has a reported accuracy level of 78% (Wickham et al., 2010). An assessment of the *GlobCover* global land cover map produced by the European Space Agency and the EC Joint Research Centre states "*the GlobCover 2009 land cover map cannot be used for any change detection application*" (Bontemps et al., 2011). Global land cover/use data sets such as MODIS, GLC2000 and GLOBCOVER have many mismatches on the spatial distribution of their land classes (McCallum et al., 2006). In Africa, the agreement between these land cover products is only around 60% (Kaptué-Tchuenté et al., 2011).

By contrast, consider the case of PRODES, the Amazon monitoring systems developed by INPE. PRODES uses medium spatial resolution satellites (30 meter

resolution) to produce yearly maps of complete forest cover removal and is recognized as a benchmark application for deforestation measurement (Hansen et al., 2008). PRODES is based on visual interpretation, requiring an estimated effort of 60 person-months per year. PRODES needs to achieve a 95% level of accuracy, because it is the official Brazilian government estimate and is the basis for law enforcement and public policies. So, to help improve PRODES and similar large-scale and high-impact uses of Earth Observation data, there is a need for much better and more reliable algorithms for extracting information from remote sensing images.

The scientific challenge in Earth Observation is significant. *How can scientists trust or validate methods that have been tested on a single data set? How can they assess if a result that was valid for one geographical area and one time instance is useful and valid for other areas and for different times? How can we develop and test information extraction methods in a statistically significant number of data sets?* We need to remove these limitations to achieve radical progress in Earth observation-related research.

Thus we have a key scientific challenge: *How can we use e-science methods and techniques to extract land change information from big Earth Observation data sets in an open and reproducible way?*

Our proposal

Our solution to solve the state-of-the-art limitations is to put together a highly innovative set of technologies and methods. We believe the time has come to set up a network of high-performance centres for global Earth observation data processing and analysis. These centres will open EO data to science domains that need this data mostly, including Agriculture, Hydrology, Ecology, and Biodiversity.

We propose an innovative *knowledge platform* totally based on open source software. Users will no longer need to download hundreds or thousands of images to do their analysis. Our knowledge platform will allow scientists to perform data analysis directly on big data servers. These servers will use the innovative technology of array databases. These databases manage multidimensional arrays, each holding many images joined in space and time. Researchers will have an unprecedented open access data set to validate their methods. Scientists will be then able to develop completely new algorithms that can seamlessly span partitions in space, time, and spectral dimensions.

We share the vision for big scientific data computing expressed by the late database researcher Jim Gray: "*Petascale data sets require a new work style. Today the typical*

scientist copies files to a local server and operates on the data sets using his own resources. Increasingly, the data sets are so large, and the application programs are so complex, that it is much more economical to move the end-user's programs to the data and only communicate questions and answers rather than moving the source data and its applications to the user's local system" (Gray et al., 2005).

Project objective

Our project will conceive, build and deploy a completely new type of ***knowledge platform for organization, access, processing and analysis of big Earth Observation data****. We will show that this knowledge platform allows scientists to produce information on land use and land cover change in a completely innovative way. Since our platform is fully based on open source software, we will also show that it promotes data sharing and reproducibility of results.*

We will make two important contributions to Computer Science and to Earth Observation related sciences, as further described in the proposal:

1. New spatial databases methods and techniques that use array databases to build a geographical information system that handles big spatial data.
2. New data analysis, data mining, and image processing methods to extract land change information from large Earth observation data sets.

3.2 Contribution to Computer Science: Big Data in Geoinformatics

The focus of this project is in using *Big Data in Geoinformatics*. We take Geoinformatics to be the subfield of Computer Science that develops and implements methods for storing, retrieving, organising and analysing spatiotemporal and geographic data, and that studies how resulting insights are used and understood. Geoinformatics draws heavily on core Computer Science areas such as Databases, Computational Geometry, Computer Graphics, Image Processing and Software Engineering, and also has increasing ties to emerging areas such as Data Mining, Ontologies, and Knowledge Engineering. The discipline also interacts strongly with Temporal and Spatial Statistics and Cognitive Science.

Recent advances in technologies for data collection have changed the scope of Geoinformatics. More and more, Geoinformatics research and applications has to deal with *big spatial data*. Current and future technologies of Earth Observation satellites, mobile phones, geosensors and GPS devices produce an unprecedented amount of data with space and time references. Big spatial data allows researchers to ask new scientific

questions about our cities and our environment (Lynch, 2008). Governments can benefit from big spatial data by increasing their power to take good decisions to improve social, economic, educational, and health condition of their citizens (Boyd and Crawford, 2012).

Requirements for EO data

Dealing with Earth Observation data requires methods capable of handling terabytes of data simultaneously. Geoinformatics researchers and practitioners need to develop methods to handle and analyse big data, as well as understanding the societal impact of big data collection. However, existing approaches for EO data analysis and data management are not suitable for big EO data sets. In our view, there is a need for a new type of knowledge platform that:

1. Manages large EO data sets in an efficient way and allows remote access for data analysis and exploration.
2. Allows existing spatial (image processing) and temporal (time series analysis) methods to be applied to large data sets.
3. Enables development and testing of new methods for space-time analyses of big EO data.
4. Organizes databases and analysis methods to enable reproducibility of analysis procedures, and hence can be easily shared, published and replicated.

Array databases

To manage large scientific data sets, leading database researchers put forward a set of requirements for scientific data management systems (Gray et al., 2005; Stonebraker at al., 2013):

1. A data model based on multidimensional arrays, and not on sets of tuples.
2. A storage model based on versions.
3. Scalability to 100s of petabytes and 1,000s of nodes with high degrees of tolerance to failures.
4. Open source to insure that data is never "locked up".

With this motivation, an MIT team led by Michael Stonebraker (the designer of INGRES and POSTGRES) developed SciDB (Stonebraker et al., 2013), an open source array database optimized for management of big data and for complex analytics. It splits large volumes of data in distributed servers in a "shared nothing" way. A big array is broken into "chunks" that are distributed among different servers; each server controls its local data storage. Arrays are multidimensional and uniform, as each array cell holds the same user-defined number of attributes. Since arrays are a natural data structure to

store Earth Observation images, using SciDB researchers and institutions can break the "image-as-a-snapshot" paradigm. Entire collections of image data will be archived as single spatiotemporal arrays.

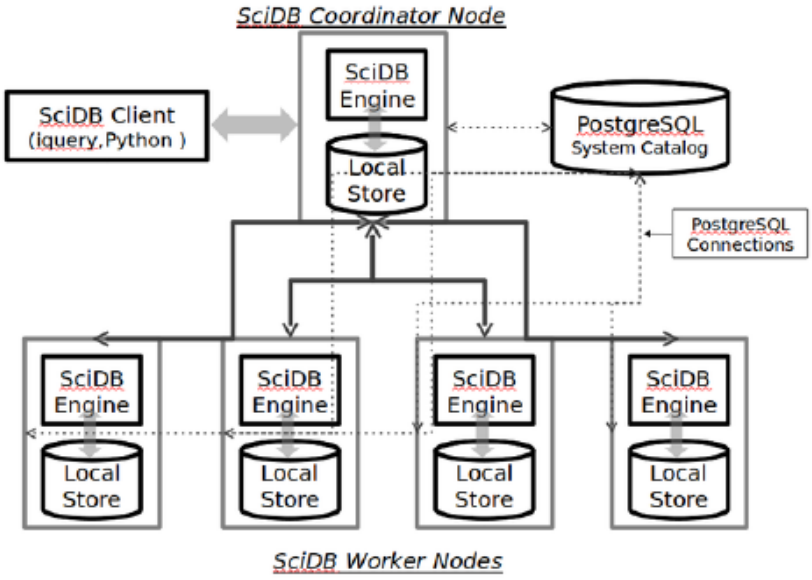

Figure 1 – Architecture of SciDB: large data sets are broken into chunks which are sent to distributed servers for efficient parallel processing of data analytics. Source: Stonebraker et al., 2013.

Extending SciDB for GIS

Array databases have no semantics. 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. Currently, SciDB has only functions for generic array processing, and has no specific functions for spatial and spatiotemporal processing. It does not distinguish spatial and temporal dimension, has no support for cartographical projections, and does not support neighbourhood operations. So we will need to extend SciDB with a core set of functions that are common to most applications that deal with Earth observation data. This core set of functions will be developed to work efficiently on large distributed data sets on the server side.

To transform the SciDB array database manager into a spatial database manager, we will do the following actions:

1. Conceive and implement the *TerraScript language,* which has a set of concepts that are fit for spatiotemporal Earth Observation applications and thus are

easily understood by domain scientists. This scripting language will be based in a formal algebra proposed by Ferreira et al. (2014) and will be developed as an extension of Lua (Ierusalimschy, 1996).

2. Implement an adapter data type that provides the interface between the algebra of Ferreira et al. (2014) and the multidimensional arrays used by SciDB. This interface is proposed in Câmara et al. (2014).
3. Convert the 100+ image processing algorithms available in the TerraLib GIS library (Câmara et al., 2008) to work on SciDB for server-side data processing.
4. Provide an interface to the R data analysis language (Ihaka and Gentleman, 1996) for those scientists that are familiar with the R environment.
5. Develop new methods for time series analysis and spatiotemporal analysis of Earth Observation data, described in more detail in the next section.

In Ferreira et al. (2014), we have developed a formal algebra for spatiotemporal data types, based on three basic data types:

1. *Time Series*: Given a sensor in a fixed location, we measure values of a property at specified times. Examples are meteorological stations and hydrological sensors.
2. *Trajectory*: Given a moving object, we measure its location and specified times. Examples are cars and migratory animals.
3. *Coverage*: Given a predefined area (spatial extent), we fix a time for data collection, and given a spatial resolution, measure a value. Examples are remote sensing images and digital terrain models.

The model is set forth as an algebraic specification, describing data types and operations in a language-independent and formal way. The presented algebra is extensible, specifying data types as building blocks for other types. Using these operations, Ferreira et al. (2014) argue that this algebra is able to describe different kinds of spatiotemporal data and show how to define *events* such as deforestation and floods.

To map the spatiotemporal algebra of Ferreira et al. (2014) into SciDB, we will use an interface that acts as an adapter between spatiotemporal types such as *Coverages* and *TimeSeries* and SciDB multidimensional arrays. In Câmara et al. (2014), we introduce a generic *field data type* that can represent different types of spatiotemporal data such as trajectories, time series, remote sensing and climate data. Our generic field allows different semantics of multidimensional arrays. A time series of rainfall is mapped to a 1D array, whose indexes are time instants, and values are the precipitation counts. A set of remote sensing image is implemented as a 3D array, where one of the dimensions represents time and the other two the spatial extent. Logistic and trajectory models

record moving objects by taking positions as time instances; its values are the object's locations in space.

To complete the groundwork necessary for transforming an array database such as SciDB into a geographical information system (GIS), we need to design a new type of GIS. The current generation of GIS consists of a front-end client with a visualisation and query interface, and a back-end database. Knowledge platforms for Earth Observation will be composed of distributed databases. This needs rethinking GIS architecture. Users will no longer be responsible for database creation and maintenance.

We propose a system with four modules (see Figure 2). The *User Interface* module converts data into an informative graphics, text and images and allows the user to send commands for remote processing. The *Data Discovery* module finds the remote servers that contain the data needed by the user. The *Data Processing* part retrieves data from the remote servers or sends software to be executed remotely. The *Remote analysis* module executes the server-side processing.

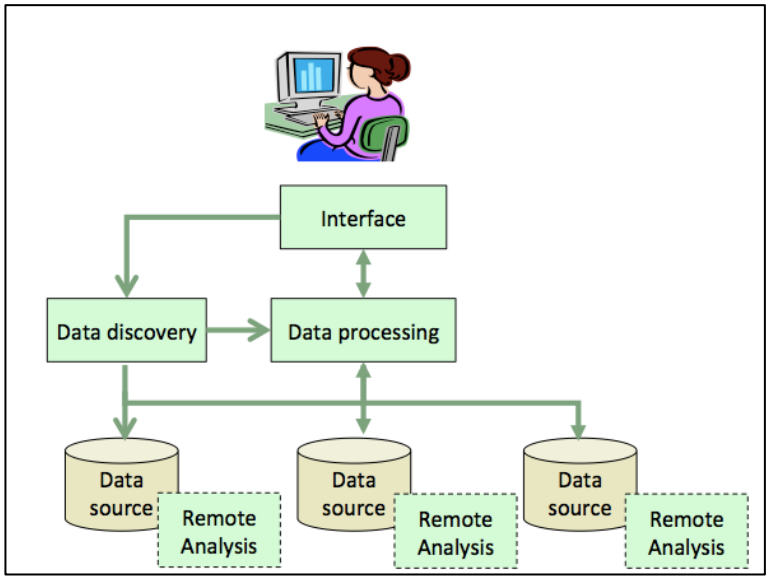

Figure 2 – Proposed architecture for big Earth Observation data processing and analysis.

To find out what data sources exists, the *Data Discovery* module uses a broker that searches for geospatial metadata information and finds out information about available data services. Development of data discovery tools draws on recent relevant results on geospatial semantics, especially using Linked Open Data (Bartle and Kolas, 2012; Koubarakis et al., 2012). The *Data Processing* module will offer the scientist a choice of using the TerraScript language, based on the algebra of Ferreira et al. (2014) or to use directly the R data analysis language. In both cases, the commands will be executed in the remote server.

Our choice of R is motivated by the fact that SciDB provides a native interface to it. R is the *lingua franca* of data analytics, providing a wide variety of statistical and graphical tools, including spatial analysis, time-series analysis, classification, clustering, and data mining. R is easily extensible through functions and extensions, and the R community is noted for its active communication through high quality extension packages (Pebesma et al., 2012).

To build the big Earth Observation data processing and analysis, we will develop updated versions of the TerraLib GIS library. TerraLib is a library that contains hundreds of functions, including algorithms for image processing, vector geometries, spatial database queries, cartographic projection and geographical metadata (Câmara et al., 2007). We will develop a version of TerraLib that interfaces with SciDB, manages distributed data sources and executes its algorithms on the server side. The new version of TerraLib will allow the user to write scripts in TerraScipt, derived from the formal algebra of Ferreira et al. (2014). We will also extend the existing interface between R and TerraLib (Andrade et al., 2005) to allow easy interface between the algorithms available in both sides.

Our proposed approach to transform an array database such as SciDB into a full geographical information system for large data sets will require a substantial amount of work, a fact that is reflected on the project's duration and team composition.

Alternative approaches

We considered three alternatives to the use of SciDB for large Earth Observation data handling: object-relational databases, the *MapReduce* programming paradigm, and the *Google Earth Engine*.

As an alternative for file-based image analysis, some spatial databases use object-relational database managers, adopting a mixed model where the array data is stored as binary objects inside relational tables. Examples of these solutions are the current version of the TerraLib GIS developed by INPE (Vinhas et al., 2003) and the Rasdaman array manager (Baumann et al., 1998). In this approach, large arrays are broken into chunks; each chunk is stored as a binary object in a line of a relational table. Each chunk is retrieved based on a query that fetches the content of one line of a relational table. Each query is first passed on to the database query analyser and then executed. Thus, to get the data, there is a performance penalty caused by the query processor of the database manager. When performing analyses on a large data set, this cost can be significant.

Compared to object-relational databases, the SciDB solution provides significant performance gains. Benchmarks comparing object-relational databases and array databases for big scientific data have shown gains in performance of up to three orders of magnitude in favour of the latter (Cudre-Maroux et al., 2010; Planthaber, 2012). Our experiments (reported below) provide additional evidence that object-relational DBMS cannot solve the challenge of handling large scientific arrays.

Google and Facebook use the *MapReduce* model to process social network data or text documents (Dean and Ghemawat, 2010). Social network data comes in real time and fast loading and processing of documents is critical. Typical tasks in Google require finding occurrences of a search term in billions of pages stored in thousands of servers. Scientific data processing is a different problem. Scientific databases are "read-mostly", with algorithms that are much more sophisticated than social network data processing. Scientific data processing is more localized, with tasks that typically involve applying the same function to all elements of a subset of an array. Thus, we consider that MapReduce model does not support our needs.

The other alternative we considered was to take dedicated services of the *Google Earth Engine*. This engine provides a programming interface that allows access to Earth observation data stored by Google. From a technical viewpoint, Google's solution is similar to ours. From an ethical and practical perspective, it has strong limitations. Google controls the data and the access, and the whole technology is proprietary. Users cannot organize their own datasets and would have to negotiate legal and commercial conditions with Google for the company to store and process their data. The end user license agreements do not meet even minimal standards for scientific collaboration. Furthermore, this service is provided on an as-is basis. Google does not make any long-term commitments regarding service maintenance.

3.3 Contribution to the target domain: Big Earth Observation Data Analysis

We take a space-time perspective of Earth Observation data, considering that each sensor revisits the same place at regular intervals. Sensor data can, in principle, be calibrated so that observations of the same place in different times are comparable. These observation can be organized in regular time intervals, so that each measure from sensor is mapped into a three dimensional array in space-time (cf. Figure 3). From a data analysis perspective, observations of the sensor are mapped into points in space-time. Researchers can then design algorithms that can address any location in the space-time box and perform operations in arbitrary space-time partitions.

For illustration, let's consider the MODIS sensor that covers the Earth daily at 250-meter resolution. One of the land products generated by NASA is the MOD09Q1, which contains the values of two spectral bands at 8-day resolution. Data is available as individual files from the NASA archive and can be organized as a single SciDB array: each pixel is identified by its spatial position in a 250m grid and by its time position in an 8-day grid (Figure 3). Other data sets such as SENTINEL, MERIS, PROBA-V, VEGETATION, LANDSAT can all be organized in a similar arrangement.

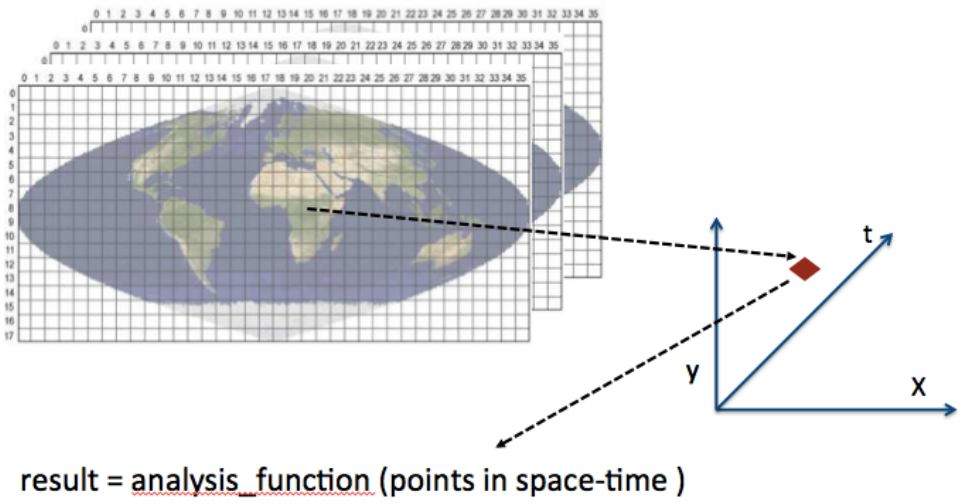

Figure 3 – Mapping data from a sensor into a position in space-time. Any remote sensing satellite can be converted into a tri-dimensional array, thus allowing for data analytics to be designed in a generic and sensor independent way.

Organizing sensor data as space-time array, we can develop better analysis methods. We need to capture subtle changes associated with forest degradation and temporary or mixed agricultural regimes (Broich et al., 2011). Since land is becoming scarce, intensification and extensification will dominate future land change. Gradual change will prevail, and will have a large impact on climate, ecosystems, and society. As deforestation in Brazil no longer occurs by fast clear-cuts, forest transition areas have become more complex to describe and measure (Perz 2007). Recent research shows that much of the recent increase of agricultural productivity in Brazil is due to double cropping-practices (Arvor et al. 2012). These results motivate us to explore high temporal resolution and multiple date remote sensing data to improve land use and land cover classification in Brazil.

We will investigate two complementary research approaches in the project. The first approach is to use a time series of vegetation indices derived from MODIS images

(Galford et al., 2008; Verbesselt et al., 2010). The other line of research works with long-term combinations of LANDSAT images (Griffiths et al., 2013).

The vegetation indexes derived from MODIS support tracking of land change trends in tropical forests (Anderson et al., 2005) and mapping crop frequency changes in agriculture (Epiphanio et al., 2010). In this project, we want to use these indexes to support the real-time deforestation alert system DETER and to improve information about agriculture crops in Brazil.

A second line of research works with multi-temporal combinations of LANDSAT data. Long time series of multi-sensor remote sensing data have a huge potential to perform coherent land change analyses across large areas. Analyses based on multi-temporal LANDSAT data can describe subtle land change; they are especially useful for detecting degradation and succession in sparsely vegetated ecosystems. Recent research shows how to derive pixel-based image composites that allow wall-to-wall monitoring of large regions at high spatial resolution (Griffiths et al., 2013). Using image composition, we get multi-temporal training samples whose statistical properties improve information extraction from LANDSAT time series data (Griffiths et al., 2012; Zhu et al., 2012). We will use the LANDSAT time series data to detect forest degradation in Amazonia and to complement MODIS-based information on agriculture.

As shown in Figure 3 above, the use of array databases is particularly well suited for time series analysis of remote sensing imagery. By organizing successive images as 3D arrays, users will be able to develop algorithms that can seamlessly span partitions in space, time, and spectral dimensions, and arbitrary combinations of those. These algorithms will provide new insights into changes in the landscape, including local interactions such as crop rotations, or shifts of a land use type in a particular direction, change in structure, or shrinking and expanding land uses.

In resume, our project will develop algorithms that use the spatiotemporal nature of Earth Observation data in a comprehensive way. Although the subject of time series analysis of remote sensing images is being studied in the literature with promising results, it will be the first time (to our knowledge) that scientists will have a platform that will allow them to do large-scale analysis of time series of remote sensing data. Thus, we are requesting an appropriate amount of resources and dedicating a lot of time to achieve the expected results.

3.4 Summary of preliminary results

On this session, we will describe preliminary results specifically related to the project proposal.

Array databases

Our preliminary work with the array database SciDB shows that it is efficient for dealing with EO data. As test data we used the MODIS09 land product with three spectral bands (visible, near infrared, and quality). Each MODIS09 tile covers 4800 x 4800 pixels in the Earth's surface at 250 meters ground resolution. We took three time steps of 22 MODIS images and merged the 66 images into an array of 1,520,640,000 cells. Each cell contains three values, one for each band.

On this organised data we tested different relevant SciDB functions. The *subarray* selects subsets of the large arrays. The *apply* function allows the application of a function to all elements of an array. The *filter* operation selects from an array those cells that match a predicate. The *aggregate* function calculates a combined value (e.g., the average) for all elements of an array. Fig. 4 shows timings of these operations as a function of array size. The results show a linear behaviour of the SciDB algorithms, which is very encouraging.

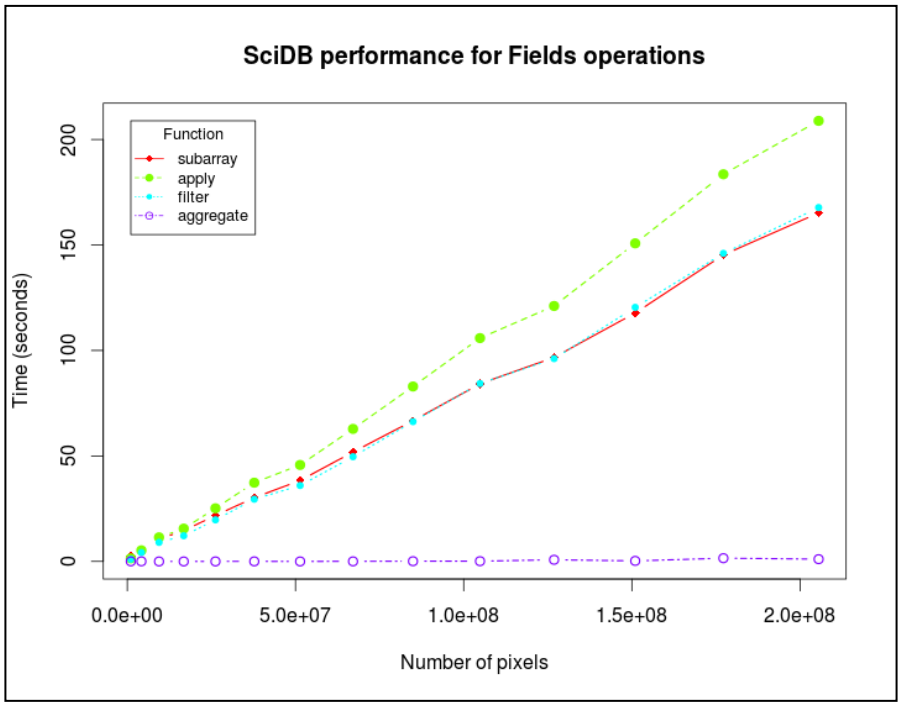

Figure 4 – Performance measures for image operations in SciDB (source: Câmara et al, 2014). The results were obtained in an Intel Xeon CPU@ 2. 00GHz, with 8 cores and 32 GB memory. The performance results are satisfactory because the processing time grew linearly with array size. With a bigger server configuration, we expect better results.

Data analytics

As part of our aims to develop time-series analysis methods for Earth Observation data, we have already implemented the Dynamic Time Warping (DTW) algorithm for land change monitoring and classification (Maus and Câmara, 2013). DTW is an algorithm measuring similarity between two temporal sequences (Keogh and Ratanamahatana, 2005). It works by comparing a temporal signature of a known event (such as a person's speech) to an unknown time series (such as a speech record of unknown origin). DTW provides a robust distance measure for comparing time series, allowing similar shapes to match even if they are out of phase in the time axis. DTW is a good method for similarity search in big time series data (Rakthanmanon et al, 2013).

DTW is well suited for classification of land cover and land changes in Earth observation data. Figure 5 shows early results of our on-going work. Using DTW, it was possible to correctly distinguish between forest, pasture and agriculture in the Brazilian Amazonia. We show the results for three areas, each associated to a time series derived from MODIS from 2000 to 2012. Area 1 was covered by tropical forest until 2004, when it was cut for the first time. Following a second cut in 2006, the area was used as pasture until 2011, when pasture was replaced by agriculture. Area 2 is a place where the forest has been preserved. In area 3, the forest was cut in 2003, and after a period of further cleaning, this land is being used for soybean agriculture since 2005. There has been one crop per year from 2005 until 2012.

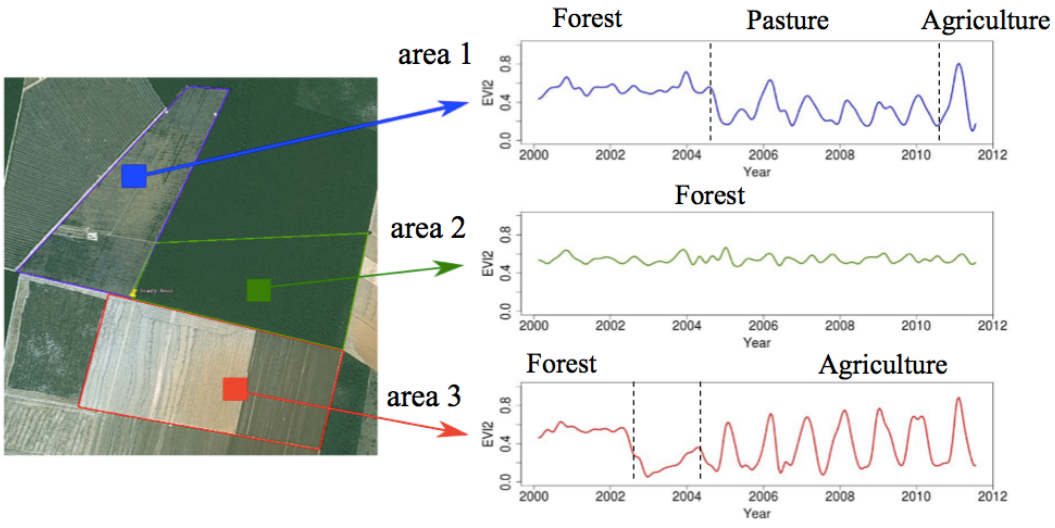

Figure 5 – Detection of land cover changes using time series of vegetation indexes in Mato Grosso, Brazil. Left: Google Earth image. Right: EVI2 8-day time series derived from MODIS images for three areas for the period from 2000 to 2012 (source: Maus and Câmara, 2013).

3.5 Significance and Relevance to the FAPESP e-science program

The aim of the FAPESP e-science call is to "*identify, select and broaden world class, basic and applied research. The intention is to explore and create new knowledge and technology. Bold, novel, and unconventional approaches to the core science and technology challenges in the research areas approached are encouraged".*

This will be the first time the technology of array databases like SciDB and of time-series and space-time series analysis methods will be applied to large-scale Earth Observation data. The project will also be the first to conceive and build a geographical information system that works with array databases.

The FAPESP call also requires "*Engagement of both computer scientists and scientists from the target domain*" and "*Evidence of benefits to research in the target domain"*

The project will address use cases of high importance for Brazil. We will develop novel ways to extract information from Earth observation data, and test these against relevant existing solutions. We have set up an experienced interdisciplinary project team, combining experts in Geoinformatics with researchers in Agriculture and Forestry. Our team will not only develop new algorithms and methods; it will also validate them by comparing the results with the operational products of INPE's Amazon monitoring systems.

FAPESP also asks for "*evidence of offering training in eScience practices*". Our team has been engaged in interdisciplinary research for more than 20 years. We offer courses and advise thesis and dissertations in INPE's graduate programs in Computer Science, Remote Sensing and Earth System Science. Our curricula provides substantial evidence that we offer Computer Science students an environment where they can fully understand the needs and practices of interdisciplinary research.

Furthermore, FAPESP requests emphasis on "*dissemination of results*" and for an open "*data management policy*". Our team at INPE has been very active in promoting open access to scientific data and developing open source software. INPE was the first institution to fully open its LANDSAT data archive in 2005, preceding the USA who only did so in 2008. All our Earth observation data is fully available on the internet. We also have a 20-year history of developing open source software for geographical applications. When selecting the new technologies for this project, we chose to work only with open source software, to ensure effective result dissemination.

4 Specific Aims and Expected Results

4.1 Expected Results

Our project has two expected results:

(1) Conceive, build and deploy a completely new type of **knowledge platform** for organization, access, processing and analysis of big Earth Observation data.

(2) Show that this **knowledge platform** allows scientists to produce information on land use and land cover change in a completely innovative way. Since our platform is fully based on open source software, we will also show that it promotes data sharing and reproducibility of results.

Expected Result 1

Our knowledge platform will provide scientists with a completely new way for easy access, processing and analysis of big Earth Observation satellite image. We will use the SciDB array database manager for storing and processing large sets of remote sensing images. We will extend SciDB from a pure array database manager to become a spatiotemporal database manager. This platform will provide new space-time data analysis, data mining, and image processing methods. These algorithms will extract land change information from large Earth observation data sets. Scientists will be able to develop algorithms in high-level programming languages and process them on the server side.

Expected Result 2

We will demonstrate that the proposed knowledge platform can be used for improve information extraction from Earth observation data. To do this, we will develop use cases related of land use and land cover change in Brazil. The use cases will be carried on using the data and the analysis tools available in the platform so they can be reproduced. They will be related to deforestation monitoring, early detection of forest degradation, production of land cover and land use maps. They will be designed with a validation possibility in mind.

Forest use case

The first use case will focus on Brazil to benefit from INPE's expertise in large-scale tropical forest monitoring. INPE has developed applications, for monitoring in the Amazon using LANDSAT and MODIS images, which are taken as standard references on Forestry monitoring by the scientific community (Hansen et al., 2008; Espindola et al., 2012). INPE's operational applications include: (a) the PRODES yearly maps of forest

removal, using LANDSAT-class data; (b) the DEGRAD system for monitoring forest degradation using LANDSAT-class data; (c) the DETER daily alerts of forest change areas, using MODIS-class data. Given the expertise of INPE, we expect the Brazilian use case to provide us a very solid reference to validate our novel techniques and to formulate best practice.

Our use cases address two important problems: (1) how to use MODIS time series data to try to replicate and improve the DETER system for daily alerts of forest change; (2) how to use LANDASAT time series and spatiotemporal analysis to support the DEGRAD system of monitoring degradation. We will take INPE's operational products as our references for validation. INPE's data is openly available online, thus allowing assessment of our methods compared to INPE's results that rely on visual interpretation. Our aim is not to replace INPE's existing systems, but to explore how automated methods can complement and enhance them.

Agriculture use case

The second use case is to use MODIS and LANDSAT time series to improve the mapping and monitoring of Brazilian major agricultural commodities: soy, maize, sugarcane, rice and wheat. We will pay special attention to the Amazonia and Cerrado regions, building on experiences that provided a huge amount of land use classes verified in the field to serve for validation of the new algorithms and the promise of big-EO-data. We will work to produce a comprehensive land cover map showing where the crops are.

This task, particularly in big territories and spread crop production areas like Brazil, can benefit greatly from the big EO data development, as it needs to be updated constantly with crop calendars changing driven by climatic variations. The monitoring of growing conditions during the cycle is also a need for yield forecasting and hence the framework of the big EO will allow quick response credibility to operational use. The project team will benefit from running technical cooperation agreements established between INPE and CONAB (Companhia Nacional do Abastecimento) in the context of Brazil´s participation in GEO Global Agricultural Monitoring initiative, GEOGLAM. CONAB is responsible for Brazil's national agricultural information system and is incorporating Earth Observation methods into its work routine for providing crop assessment on a monthly basis. We will compare and validate our results with those of CONAB.

4.2 Knowledge creation and dissemination

The project will create a completely new knowledge platform for Earth Observation data, based on an array database manager and new algorithms and methods for spatiotemporal data analysis. To ensure reproducibility and sharing of the results, we will put together all of the software, methods and lessons learned in a shareable open source environment that will include:

(a) the databases developed by the project and the executable procedures to build these from scenes;

(b) the GIS toolkit for big Earth observation data;

(c) the space-time package for forestry alert and agriculture mapping using big Earth Observation data;

(d) documents containing results of the validations of the use cases done by the project;

(e) documents with the lessons learned with the deployment of the proposed IT infrastructure and best practice recommendations.

In the first two years of the project, we expect to have consolidated a first version of our knowledge platform. We will then hold an international workshop at INPE where we will invite leading experts on big data and Earth observation. We will exchange experiences and open the platform to the scientific community. We also will hold a second workshop close to the project's end, to present to the community the main results, and to further exchange experiences.

We also plan to hold workshops and presentations at major national and international conferences on Remote Sensing, including Brazilian Symposium on Remote Sensing, the meetings of the American Geophysical Union and of the European Geophysical Union, the GIScience and ACM GIS conferences.

Multidimensional arrays are the most common data structure in science, used in areas as diverse as Particle Physics, Astronomy, Medicine, Genomics and Meteorology. We believe that array databases are applicable to those and many other scientific fields. Thus, we will take special care to interface with other communities, directly or through the seminars that are planned as part of FAPESP's e-science program. We believe our experience with the SciDB array manager and with spacetime data analysis will be relevant to many Brazilian scientists outside the Earth observation and Geoinformatics domains.

4.3 Expected impacts

The proposal will provide a way to support analysis and processing of large Earth observation data sets for non-specialists. Currently, the vast majority of scientific data analysis methods do not undergo sufficient testing and validation. Scientists develop a new method and then test it in a limited number of cases. Prospective users are skeptical to adopt such unproven methods in their application. The result is a deadlock. After decades of Earth observation satellites, there are few large-scale operational applications. The project's proposed IT infrastructure provides a way out of this deadlock. Innovative proposals will be validated in large and significant use cases thus increasing user's confidence in the methods that will be proven valid. As a result, it will become easier for users to innovate and prove methods in operational application.

National impacts

1. *Monitoring changes in Amazon rain forest and the TerraAmazon software*

INPE leads a successful operational application of Earth observation data to monitor the Brazilian Amazon rain forest. INPE has developed TerraAmazon (Ribeiro et al., 2007), an open source forest monitoring, reporting and verification (MRV) system that is now part of the UN-REDD program. TerraAmazon supports three complementary remote sensing applications (PRODES, DEGRAD and DETER) to monitor deforestation and forest degradation in the Brazilian Amazon. The results of the project will be taken into account by INPE when designing the future versions of TerraAmazon software, thus providing a strong link between the project's results and the services to be provided by FAO to its member states through the UN-REDD programme (see below).

2. *Transforming INPE's Remote Sensing Data Centre*

INPE has a remote sensing data centre that is responsible for reception, processing, archival and free distribution of many Earth observation satellites, including LANDSAT and MODIS, and the Chinese-Brazilian CBERS. INPE has hundreds of terabytes of data, and its archive dates back to 1973. Since 2005, INPE has distributed more than 2 million images. In the next years, INPE will further increase its data centre by receiving data from national and international satellites such as CBERS-4, LANDSAT-8, and SENTINEL-1/2/3. INPE is currently planning a major upgrade of its data centre, so it becomes not only a data distribution centre, but also a major data processing hub, where scientists and users could run application programs that will explore INPE's vast data archive. The INPE data centre team will follow this project closely, since they will draw on our experience to design a new high-performance centre for Earth observation data processing and analysis.

3. *Transforming Remote Sensing analysis in major Brazilian institutions*

Many Brazilian institutions hold significant Earth Observation data archives, such as EMBRAPA, ANA (National Water Agency), INMET (National Institute for Meteorology) and IBAMA (Environmental Agency). These archives are also organized as collections of files. We way pay special attention to ensure they are aware of project results, so they can consider setting up their own data processing centres.

4. *Transforming big data processing in other scientific areas*

As we argued in section 3.2, we will make specific efforts to make our experience with SciDB array manager well-known in the broad scientific community, especially in the state of São Paulo. We have much hope that many other research teams can benefit from our results.

International impacts

1. *UN-REDD Programme*

The United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD Programme) assists developing countries to build capacity to reduce emissions and to participate in a future REDD+ mechanism. The Programme supports national REDD+ readiness efforts in 50 partner countries, spanning Africa, Asia-Pacific and Latin America. One of the main work areas of the programme is the establishment of MRV (monitoring, reporting and verification) systems. For this, the leading agency is FAO, the United Nations Food and Agriculture Organization. In 2009, INPE and FAO signed a memorandum of understanding to make the Brazilian TerraAmazon software available to other developing tropical nations. The future development of the TerraAmazon software provides a strong link between the project's results and the services to be provided by FAO to its member states. The project will thus leverage an existing partnership and will provide support for big EO data sets to be used for global applications.

2. *GEOSS (Global Earth Observation System of Systems)*

One important area where we expect our project to have a large impact is global initiatives for sharing Earth observation data, such as the Group on Earth Observations, GEO (an intergovernmental organization with 90 member countries and 67 regional organizations including UN agencies). GEO's mandate is "*to achieve comprehensive, coordinated and sustained observations of the Earth*" and to "*to improve monitoring of the Earth, increase understanding of Earth processes and enhance prediction of the*

behaviour of the Earth system"[1]. GEO has launched relevant initiatives in the areas of Forest (GFOI - Global Forest Observations Initiative) and GEOGLAM (GEO Global Agriculture Monitoring initiative). Brazil is an active participant in GEO, having held the GEO Plenary Session in Foz do Iguaçu in 2012. The results of the project have direct impact on GEOSS. Consider a situation where major data providers, such as ESA, USGS, NOAA, NASA, and INPE, and major international Earth observation institutions, such as FAO and UNEP, will have organized their Earth observation data as open access archives, with analysis and processing services as we propose. Researchers and users of Earth observation data would be able to produce information to support global applications such as GFOI and GEOGLAM. Furthermore, our proposed solution does not require the development of a unique, centralized service. There would be a network of data processing centres, all of them offering compatible services.

5 Means and methods

This section describes the tasks to be carried out by the project team to achieve the expected results. We have organized the project in three work packages (WP):

- WP 1 – Databases: research and development associated with using array databases to store large EO data sets and developing workflows and methods for efficient storage, access and processing of large data, reproducibly.
- WP 2 – Data analysis: R&D on spatiotemporal techniques for extracting change information on large Earth observation data sets, relevant for forestry applications; include novel time series applications for remote sensing data, and combined time series and multi-temporal image processing.
- WP 3 – Use case development: case studies of forestry and agriculture applications that use large Earth observation data sets. These use cases will validate the methods and data developed by the other work packages.

5.1 Work Package 1 - Big Earth Observation Databases

Task 1.1 - Building and deployment of Big Earth Observation databases to support data analysis and use cases

Team: Lúbia Vinhas (lead), Gilberto Queiroz, Ricardo Cartaxo, 1 TT-4A scholarship.

This task is concerned with building databases required for the data analysis (WP 2) and use cases (WP 3) packages. INPE holds one of largest continuous LANDSAT

[1] GEO Geneva Ministerial Summit Declaration, 2014 (source: GEO website).

archives outside the US, with data that covers most of South America since 1973 (LANDSAT-1). INPE was the pioneer in open access to medium-resolution imagery, having made its data archive openly available since 2006. In this task, INPE will deploy the project's IT infrastructure and will build an open access database with SciDB that will hold MODIS, LANDSAT, SENTINEL, and CBERS images for South America. INPE will use the locally deployed IT infrastructure to provide support for validation of the methods and techniques developed in the project. This will be the first time that this data is available in an array database, ready for joint analyses. These databases will include: (a) All MODIS MOD09Q1 images at 250 meter resolution from 2000 to 2014; (b) The EVI2 vegetation index at 250 meter resolution for the same period, with raw and smoothed data.; (c) The TRMM precipitation data for the same period; (d) The forest fires data produced by INPE and NASA; (e) selections of INPE LANDSAT data from 1973 to 2014; (f) selected SENTINEL-2 data; (g) data from INPE's DETER, PRODES and TerraClass systems.

Task 1.2 – Extend SciDB for geographical data handling

Team: Gilberto Câmara (lead), Lúbia Vinhas, Karine Ferreira, Julio D'Alge, Emiliano Castejón, Gilberto Queiroz, Ricardo Cartaxo, 1 TT-4A scholarship, 1 DR scholarship.

Task 1.2 is research-oriented, since its expected result is a producing a new type of data manager for big spatial data. This task will develop software that extends the SciDB array database to build a spatial database manager that will provide the capabilities needed to access, analyse and visualise big Earth observation data sets. Extending SciDB to become a spatial data manager will require substantial new research in Geoinformatics.

Our first activity will be to develop a link between TerraLib and SciDB. Since SciDB is a "pure" array database, it has no specific information about geographical data. Linking TerraLib with SciDB will allow arrays in SciDB to have additional information about satellite image metadata, cartographical projections, and temporal information. TerraLib will store this additional information; applications developed using TerraLib will use its facilities to handle the metadata associated to SciDB arrays.

Then, we will extend SciDB for server-side processing of spatial and spatio-temporal data. The core processing part of array-based TerraLib algorithms will moved to the server to operate in a distributed and efficient way. We will implement the generic fields data type proposed by Câmara et al. (2014) that provides the interface between the algebra of Ferreira et al. (2014) and the multidimensional arrays used by SciDB. Using this adapter class, we will convert the 100+ image processing algorithms available in the

TerraLib GIS library (Câmara et al., 2008) to work on SciDB for server-side data processing.

The third step is the conception and development of *TerraScript*. The *TerraScript language* will have a set of concepts that are fit for spatiotemporal Earth Observation applications and thus are easily understood by domain scientists. This scripting language will be based in a formal algebra proposed by Ferreira et al. (2014) and will be developed as an extension of Lua (Ierusalimschy, 1996). We will also implement *TerraScript* to work with SciDB for server-side processing.

5.2 Work Package 2 - Data analysis for big Earth observation data

This work package will develop new methods for space-time change analysis of big Earth observation data. These methods aim to break the current paradigm of file-based Earth observation data analysis, which have severely limited the quality of current methods for information extraction. WP2 has two tasks. The first task ("*Integration between SciDB, TerraLib and R*") is more technological, and is a required step for the second task ("*Space-time analysis of big EO data for land change monitoring*") that is research-oriented.

Task 2.1 - Integration between SciDB, TerraLib and R

Team: Pedro Andrade (lead), Karine Ferreira, Gilberto Queiroz, Ricardo Cartaxo, 1 TT-4A scholarship.

This task will develop the integration between SciDB, TerraLib and the R software. R is an open-source platform and language for statistics and graphics. Many researchers in statistics around the world implement their methodologies in R, making them freely available in the internet as packages. More than 5,000 R packages are available, covering a wide range of modern statistics. Among those packages, it is important to cite the spacetime package (Pebesma et al., 2012) that provides data structures for spatio-temporal objects. Our research team has already developed aRT, a package for integrating spatial databases managed by TerraLib with R functions (Andrade and Ribeiro, 2005). In this task, we will extend aRT to be able to use the link between TerraLib and SciDB (see Task T1.2). The integration between SciDB, TerraLib and R will be available as an R package to create, read, write, and query big geospatial databases without needing to load all data at once into R.

Task 2.2 - Space-time analysis of big EO data for land change monitoring

Team: Leila Fonseca (lead), Gilberto Câmara, Pedro Andrade, Emiliano Castejon, 2 DR scholarships (36 months), 1 PD scholarship (40 months).

This task will develop new methods for space-time analysis of big Earth observation data. It is expected to produce new research results, since it will be the first time that EO scientists have full access to large data sets to validate their data analysis methods.

We will start by using existing methods of remote sensing time series data analysis, such as the BFAST algorithm by Verbesselt et al (2012), the DTW (Dynamic Time Warping) algorithm used by Maus and Câmara (2013), and the set of features based on polar coordinates by Körting et al (2013).

Subsequent versions of the package will include new methods for space-time data analysis that combine spatial and temporal properties of remote sensing images to detect and understand change (e.g. change types like deforestation versus degradation). These methods may include segmentation algorithms to partition the time series into homogeneous regions with a similarity in time and space. It will also be able to analyse multiple remotely sensed data sets (MODIS, Landsat, SENTINEL-2, Proba-V, etc.) and other time series data (climate, fire activity) at once to improve change detection capacity and enable differentiation of different change types and drivers (e.g. human-induces deforestation versus drought stress and anomalies). All the work on this task will be released as open source software, therefore we expect that this software will be maintained and improved by the community.

We hope the methods will provide significant scientific innovations for:

1. Analysis of big EO data sets.
2. Change detection for high spatial resolution satellite data while dealing with high spatial detail and inherent noise (shadow, soil effects).
3. Differentiation between more local e.g. human induced deforestation events and region changes e.g. climate induces events like drought stress.
4. Characterization of land-cover land-use before and after the change finding out what the change driver is (e.g. climate, human, etc.).
5. Analysis of multiple EO data sets e.g. high spatial resolution (e.g. Landsat, and Sentinel-2) and high temporal resolution (Proba-V, MODIS, MERIS).

5.3 Work Package 3 - Use case development

The WP will define the requirements and perform validation for the use cases to be developed by the project team. The team will select a number of representative case studies for use in monitoring forest and agriculture areas in Brazil.

Task 3.1 - Specification and Validation of Tropical Forest Change Alert Methods and Data

Team: Isabel Escada (lead), Luis Maurano, Julio D'Alge, Silvana Amaral, 1 PD scholarship (36 months).

This task will specify and validate methods and databases for rapid detection of change in tropical forests. It will take as its reference the Brazilian PRODES and DEGRAD systems for clear-cut and forest degradation, respectively. These systems are based on visual interpretation of Landsat images with a finer spatial resolution (30 m) than Modis (250 m), providing a ground truth database. We will compare the DETER, a system used for real-time alerts of new deforestation and forest degradation based on visual interpretation and the data set produced using the new methods developed in the project with PRODES and DEGRAD data, comparing their performance.

On year 1, we will select large regions in the Brazilian Amazon rain forest where DETER is being used. We will compare data and methods provided by WP 1 and 2 with the results of manual interpretation done by DETER for selected areas during the years 2005-2013, using PRODES and DEGRAD data as references. The result will be a set of recommendations to other teams for methods and data improvements.

On years 2 and 3, the team will apply the improved version of the methods again for Brazil. We will compare data and methods provided by WP 1 and 2 with the results of DETER, PRODES and for selected areas. The result will be a set of recommendations to project teams for methods and data improvements.

On years 3 and 4, the team will validate the improved versions of the forest change alert system again in Brazil, and will assess the usefulness and potential for these methods to be used to support and aid INPE's operational monitoring systems. Field works will be carried out along the four years to solve doubts in specific regions of Brazilian Amazon where confusion in the classification results will be eventually detected. By the end of year 4, the project will have a detailed assessment of the usefulness and validity of the forest change alert methods developed by the project. The team will then make recommendations on the usefulness and applicability of such methods for global tropical forest monitoring. Our team will work closely with the expert team of GFOI (Global Forest Observation Initiative) that has been set up by GEO.

Task 3.2 - Specification and Validation of Tropical Agriculture Monitoring Methods and Data

Team: João Vianei Soares (lead), Ieda Sanches, 1 PD scholarship (36 months).

This task will specify and validate methods for monitoring agricultural production in tropical areas. It will take as its reference the Brazilian Agricultural Monitoring System, set up by Brazil's Ministry of Agriculture CONAB with the help of INPE, that maps grain crops and biofuel production using a mix of remote sensing and field work. This work is part of Brazil's contribution to the GEOGLAM (Global Agricultural Monitoring Initiative), launched by the Ministries of Agriculture and endorsed by the head of states of the G20 members

In year 1, we will develop requirements for Land Use mapping based on the needs of Brazil´s participation in GEOGLAM and start gaining experience on how to implement BIG EO routines, as they develop, into mapping land use for agricultural areas.

In year 2, we will use methods to detect the planted area of soybeans, maize and sugarcane crops in selected states of Brazil. We will compare and validate our results with the CONAB´s up-to-date crop masks. The result will be a set of recommendations to the teams in charge of WP1 and 2 for methods and data improvements.

In the third year, the team will apply the new version of the methods for planted area detection again in Brazil (including rice and wheat). During the year, the team will perform further evaluations, which will then be feedback to the project team, for improvement on data analysis and database production by the WP 1 and WP 2 team.

In the fourth year, the team will test and validate the improved versions for agricultural monitoring in Brazil to include other tropical areas that are part of GEOGLAM. By the end of year 4, the project will have a detailed assessment of the usefulness and validity of the methods developed by the project in a global basis. The team will then make recommendations on the usefulness and applicability of such methods for global tropical agriculture monitoring.

6 Timetable

MILESTONES

TASK	Month 12	Month 24	Month 36	Month 48
T1.1 Big EO databases	Version 1 of the database for use cases in Brazil	Version 2 of database for use cases in Brazil		
T1.2 Extend SciDB	Integration of TerraLib and SciDB	TerraLib algorithms for SciDB server-side processing	*TerraScript* available for SciDB server-side processing	Completed extension of SciDB as a spatial data manager
T2.1 Integrate SciDB, TerraLib and R		Version 1 of aRT-SciDB package	Version 2 of aRT-SciDB package	
T2.2 Space-time analysis of big EO data	Big-EO time series R package (V1)	Big-EO time series R package (V2)	Big-EO space-time R package (version 1)	Big-EO space-time multi-sensor R package (version 2)
T3.1 Forestry use case	Identification and selection of areas	Preliminary detection of clear cut and degradation	Detection of clear cut and degradation: final adjustment	Assessment of the forest change alert methods
T3.2 Agriculture use case	Identification and selection of areas	Detection of planted area of soybeans, maize and sugarcane	Detection of planted area of soybeans, maize, sugarcane, rice and wheat	Assessment of the agricultural mapping methods

6.1 Milestones for Work Package 1

6.1.1 Task 1.1 – Big Earth Observation Databases

Milestone M1.1.1 - Version 1 of the database for use cases in Brazil (month 12)

This database will be built using the SciDB array manager, containing the data needed for the use cases in Brazil in years 1 and 2.

Milestone M1.1.2 - Version 2 of database for use cases in Brazil – year 2 (month 24)

This database will contain the data for the use cases in Brazil in years 3 and 4.

6.1.2 Task 1.2 – Extend SciDB for geographical data handling

Milestone M1.2.1 - Integration of TerraLib and SciDB (month 12)

SciDB to be accessible as a data source from TerraLib, thus allowing SciDB to be extended with metadata information about its multidimensional arrays.

Milestone M1.2.2 TerraLib algorithms for SciDB server-side processing (month 24)

We will implement the generic fields data type proposed by Câmara et al. (2014) that between the algebra of Ferreira et al. (2014) and the multidimensional arrays used by SciDB. Using this adapter class, we will convert the 100+ image processing algorithms available in the TerraLib GIS library (Câmara et al., 2008) to work on SciDB for server-side data processing.

Milestone M1.2.3 - TerraScript available for SciDB server-side processing (month 36)

The *TerraScript language* will have a set of concepts that are fit for spatiotemporal Earth Observation applications. It will be developed as an extension of Lua. We will make a binding between Lua and SciDB, to allow *TerraScript language* programs to be executed directly in SciDB for server-side processing.

Milestone M1.2.4 - Completed SciDB as a spatial data manager (month 48)

The final result of the project will be a consolidated extension of SciDB to transform an array database into a geographical data manager.

6.2 Milestones for Work Package 2

6.2.1 Task 2.1 – Integration between SciDB, TerraLib and R

Milestone M2.1.1 - Version 1 of new aRT-SciDB package (month 12)

First version of the R package to access SciDB, combined with GIS databases, with scripts to access the database developed in Task 1.1.

Milestone M2.1.2 - Version 2 of new aRT-SciDB package (month 24)

Second version of the R package to access SciDB, combined with GIS databases, with scripts to access the database developed in Task 1.1.

6.2.2 Task 2.1 – Integration between SciDB, TerraLib and R

Milestone M2.2.1 - Version 1 of Big-EO space-time R package for land change monitoring and its application – time series models (month 12)

This will be the version 1 of the Big-EO space-time R package for land change monitoring. It implements BFAST and DTW models for time series analysis, allowing them to be applied to use cases in Brazil in year 1.

Milestone M2.2.2 - Version 2 of Big-EO space-time R package for land change monitoring and its application – time series models (month 24)

Version 2 of the R package to access SciDB databases combined with TerraLib, with improved version of the time series models.

Milestone M2.2.3 - Big-EO space-time R package for land change monitoring and its application – spatiotemporal models (month 36)

First version of the Big-EO space-time R package for land change monitoring, implementing spatiotemporal models.

Milestone M2.2.4 - Big-EO space-time R package for land change monitoring and its application – multi-temporal and multi-sensor models (month 48)

Second version of the Big-EO space-time R package for land change monitoring, implementing spatiotemporal models that deal with multiple sensors.

6.3 Milestones for Work Package 3

6.3.1 Forestry use case

Milestone M3.1.1 - Identification and selection of areas with different patterns of deforestation in Brazilian Amazon (month 12)

Study areas in Amazonia will be selected based on the analysis of spatial and temporal deforestation patterns based on existing deforestation data from PRODES, DETER and DEGRAD, for the period of 2005-2013. With this analysis we aim to point out areas which present different spatial and temporal pattern of forest conversion to better test the performance of new methods and data generated in WP1 and 2.

Milestone M3.1.2 - Preliminary detection of clear cut and forest degradation: comparison with PRODES, DETER and DEGRAD deforestation data (month 24).

Using statistics methods we will compare preliminary data and methods provided by WP 1 and 2 with the results of DETER for selected area, using PRODES and DEGRAD data as reference. The result will be a set of recommendations to the teams for methods and data improvements.

Milestone M3.1.3 Detection of clear cut and forest degradation: final adjustment recommendations (month 36).

This deliverable is the same as the previous one using new methods an data provided by WP1 and 2, but incorporating the improvements recommended in

M.3.1.2. The result will be a new set of recommendations to the teams of WP1 and 2 considering final adjustments in the methods developed and data obtained.

M3.1.4 Assessment of the usefulness and validity of the forest change alert methods developed by the project. (month 48).

The improved versions of the forest change alert system will be validated, and will assess the usefulness and potential for these methods to be used to support and aid INPE's operational forest monitoring systems.

6.3.2 Agriculture use case

Milestone M3.2.1 - Identification and selection of areas and requirements for organizing the EO data structure (month 12)

This deliverable will produce requirements to maximize the usability of the BIG EO data approach to detect agricultural fields in the most efficient manner. It will be based on the identification of suitable Agricultural land cover available from Terraclass (http://www.inpe.br/cra/ingles/project_research/terraclass.php) and from CONAB, for the commodities of interest.

Milestone M3.2.2 - Detection of planted area of soybeans, maize and sugarcane crops (month 24)

This deliverable will produce agricultural maps for soya, maize and sugarcane for selected areas of Brazil and compare the results with TerraClass and CONAB. It will produce of a set of recommendations to the teams of WP 1 and 2 for improvement on data analysis and database production.

Milestone M3.2.3 - Detection of planted area of soybeans, maize and sugarcane, rice and wheat (month 36)

This deliverable will extend the previous results by including rice and wheat crops, and using the new algorithms produced by WP 1 and WP2 as a response to the evaluation done in month 24.

Milestone M3.2.4 - Assessment of the usefulness and validity of the forest change alert methods developed by the project. (month 48).

This deliverable with produce reports and papers with validation of the improved versions for agricultural monitoring in Brazil. It will also include other areas included in GEOGLAM.

7 Data Management Policy

Our policy will be to deal with the databases and software created by this project as a resource to be shared with the Brazilian Earth Observation community. Thus, we will open the database after month 24 of the project to the community. We will encourage scientists to develop new data analysis methods and to use the methods and algorithms we will build to develop new applications. We will maintain the database accessible and updated for long-term use by the scientific community.

INPE is well known in Brazil and abroad as a leader in open access to scientific data. All remote sensing images from INPE's archive are available on-line without restriction, as well as all data from the PRODES, DETER, DEGRAD and TerraCLASS systems. INPE has permanent staff members dedicated to the maintenance and upgrade of its databases.

The team will develop a collaborative open source environment to share the project's results without constraints. The team will organize a web-based platform for disseminating all project materials as well as for supporting communication and collaboration with and within the stakeholder community.

The big EO databases and all tools will be published as open source software. As for the documents and software the most appropriate OSI certified OS license will be used (http://opensource.org/). The type of license will take all existing licensing dependencies between software components into account. This will include:

1. The databases developed by the project (Task 1.1)
2. The SciDB-R-EO GIS toolkit (Task 1.2)
3. The Big-EO space-time R package (Task 2.1)
4. Documents such as interactive demonstrations and instructional videos on the concepts and technologies.
5. Documents describing the lessons learned with the deployment of the proposed IT solution and best practice recommendations.

Furthermore, we will base many of our analysis methods on the R statistical language. R packages and scripts are particularly suited for sharing methods and workflows between scientists and practitioners. R allows for combining code and reporting text seamlessly. These tools will be used to generate documentation showing the validity of the reproducibility approach and making both online and document-based records transparent and open. We will also take considerable care to ensure that our experience in using SciDB is well documented and easily reproducible.

8 Dissemination and Evaluation

The project's dissemination and evaluation plan will be done by publishing papers in conferences and journals, and by actively interacting with the community. Some of our planned activities are described below.

8.1 Scientific papers

The project is requesting 3 PhD scholarships and 3 post-doctorate scholarships, as well as 3 TT scholarships. Thus, there will be a lot of young researchers associated with the project. We expect that each PhD and each post-doc researcher will be associated to at least one paper in an indexed journal per year. Thus, we expect to produce about 20 papers in journals, and similar numbers for refereed scientific conferences. Considering the potential impact of our work, we will target some of the papers to journals with high visibility, such as *Nature*, *Science* and *PNAS*.

8.2 Expert workshops

We plan to hold two workshops on months 18 and 42, and will gather the project members, and a selection of invited experts from the fields of EO data analysis, array databases, and spatiotemporal analysis with R. These workshops will allow the project team to interact with Computer Science and Earth Observation scientists with similar interests.

8.3 Interacting with the Earth observation community

The project is addressing core concerns of the Earth observation community. This relates to new and improved EO information products as well as to new and improved IT technologies and standards. Hence we will assure an intense interaction with Earth Observation community to discuss the approach, requirements, as well as findings and our recommendations. This will be done by actively interacting with international initiatives such as UN-REDD (led by FAO), the Global Forest Observation Initiative (GFOI) and CEOS (Committee on Earth Observation Satellites). We will also organize dedicated sessions on large interdisciplinary geoscience conferences such as Brazilian Symposium on Remote Sensing.

8.4 General outreach

The aim of this action is to make the project results reach the widest possible community. We recognise that there is a substantial number of beneficiaries of the results derived from Earth observation data. These include:

a) National decision-makers that rely on information about land change, food production, carbon emissions, energy production, water quality and biodiversity conservation.

b) Organised societal groups that represent interest groups that are concerned with both local and global issues. A particular important subgroup is environmental NGOs.

c) The general media outlets committed to provide good information for the society at large.

d) Citizens whose livelihoods and interests are affected by energy, food, and environmental matters.

Reaching out to these diverse groups requires a completely different strategy than that of traditional scientific venues. The experience of INPE will be extremely valuable for this task. INPE's Amazon information systems have been featured in media outlets such as The New York Times, The Economist, BBC, The Guardian, Le Monde, Nature, and specialized sites like Mongabay. We will organize specific outreach activities targeted to the media and to environmental NGOs, where the project results will be presented to show how to improve EO information production. These events will include the production of media kits and press releases that will show why our project is relevant to the society at large.

9 Additional Funds and Resources

This project is being submitted to FAPESP as a counterpart and complement to a project submitted to the European Comission's Horizon 2020 (H2020) research program. Both projects have been designed in joint consultation beteween Brazilian and European teams, due to the fact that Prof.Dr. Gilberto Câmara (INPE) is currently the Brazil Chair at the Institute for Geoinfomatics (IFGI) at the University of Münster (Germany) for the period from June 2013 to May 2015.

The H2020 submission is called "Big-EO-Data: Big Earth observation data analytics for land change monitoring" and is lead by Prof. Dr. Edzer Pebesma from IFGI (University of Münster), with Prof. Dr. Jan Verbesselt from University of Wageningen (Netherlands), and expert teams from VITO (Belgium), UN-FAO (Italy), the 52 North Initiative (Germany), and the TerraDue company (Italy). The project has requested EU 2,9 million for a three-year research. INPE will not receive funds directly from the Horizon 2020 program, since the European Commission rules do not allow paying for research in Brazil. There is no direct duplication of efforts, since support for INPE's work has been only requested to FAPESP.

INPE experts will participate in the project meetings and will cooperate closely with the European research teams. In particular, Prof. Dr. Gilberto Câmara has been appointed as Deputy Coordinator of the H2020. Gilberto will return to Brazil in June 2015, but will continue to work in close cooperation with the European partners.

The H2020 will complement the proposal sent to FAPESP. The two projects share similar goals: advancing Earth Observation science by developing a new e-science knowledge platform based on array databases (SciDB) and time series analysis. They will also emphasize the development of new data analysis methods based on R.

The main differences between the H2020 proposal and the FAPESP one are our stronger emphasis on the development of an integrated geographical information system (GIS) that combines the capabilities of TerraLib, R, and SciDB (Task 1.2). As INPE is the main developer of TerraLib, this task was not emphasized in the H2020 proposal. The second difference is the use of INPE's expertise on image processing to develop space-time segmentation methods (Task 2.2). The third difference is the emphasis on the Brazilian use cases in the FAPESP proposal (Work Package 3). A final difference is INPE's commitment to FAPESP to ensure long-term data preservation and allow open access to the Brazilian research community to the databases created by the project.

10 References

Adami, Marcos, Bernardo Friedrich Theodor Rudorff, Ramon Morais Freitas, Daniel Alves Aguiar, Luciana Miura Sugawara, and Marcio Pupin Mello. "Remote sensing time series to evaluate direct land use change of recent expanded sugarcane crop in Brazil." *Sustainability* 4(4): 574-585, 2012.

Anderson, Liana, Yosio Shimabukuro, Ruth Defries, and Dave Morton, "Assessment of deforestation in near real time over the Brazilian Amazon using multitemporal fraction images derived from MODIS." *IEEE Geoscience and Remote Sensing Letters*, 2 (3):315-318, 2005.

Andrade, Pedro, and Paulo Justiniano Ribeiro Jr. "A Process and Environment for Embedding the R Software into TerraLib." In *Brazilian Symposium of Geoinformatics (GeoInfo)*, 2005.

Arvor, Damien, Margareth Meirelles, Vincent Dubreil, Agnès Bégué, and Yosio Shimabukuro, "Analyzing the agricultural transition in Mato Grosso, Brazil, using satellite-derived indices". Applied Geography, 32: 702-713, 2012.

Battle, Robert, and Dave Kolas. "Enabling the geospatial semantic web with Parliament and GeoSPARQL." *Semantic Web* 3(4): 355-370, 2012.

Baumann, Peter, Andreas Dehmel, Paula Furtado, Roland Ritsch, and Norbert Widmann. "The multidimensional database system RasDaMan." In *ACM SIGMOD Record*, 27(2): 575-577, 1998.

Beddington, John: Food, energy, water and the climate: A perfect storm of global events. *Sustainable development* UK 9, 2009.

Bontemps, Sophie, Pierre Defourny, Eric V. Bogaert, Olivier Arino, Vasileios Kalogirou, and Jose R. Perez. "GLOBCOVER 2009-Products Description and Validation Report." ESA (European Space Agency) Report, Frascati, 2011.

Boyd, Danah and Kate Crawford. "Critical Questions for Big Data." *Information, Communication & Society* 15(5): 662-679, 2012.

Broich, Mark, Matthew Hansen, Peter Potapov, Bernard Adusei, Erik Lindquist, and Stephen Stehman. "Time-series analysis of multi-resolution optical imagery for quantifying forest cover loss in Sumatra and Kalimantan, Indonesia." *International Journal of Applied Earth Observation and Geoinformation* no. 13 (2):277-291, 2011.

Câmara, Gilberto, Lúbia Vinhas, Karine Ferreira, Gilberto Queiroz, Ricardo Cartaxo, Miguel Monteiro, Marcelo Carvalho, Marco Casanova, and Ubirajara Freitas. "TerraLib: An open source GIS library for large-scale environmental and socio-economic applications." In Brent Hall (ed.), *Open Source Approaches in Spatial Data Handling*, pp. 247-270. Springer Berlin Heidelberg, 2008.

Câmara, Gilberto, Max Egenhofer, Karine Ferreira, Pedro Andrade, Gilberto Queiroz, Alber Sanchez, Jim Jones, Lúbia Vinhas, "Fields as a Generic Data Type for Big Spatial Data". In: Edzer Pebesma, Matt Duckham (eds), Proceedings *GIScience 2014 Conference*, Vienna, 2014.

Cudre-Maroux, Philippe, Hideaki Kimura, Kian-Tat Lim, Jennie Rogers, Samuel Madden, Mike Stonebraker, Stan Zodnik, Paul Brown, "SS-DB: A Standard Science DBMS Benchmark". In: XLDB 2010 (Extremely Large Databases Conference), Stanford, CA, USA, 2010. (available on-line: http://www-conf.slac.stanford.edu/xldb10/docs/ssdb_benchmark.pdf).

Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: a flexible data processing tool." *Communications of the ACM* 53(1): 72-77, 2010.

Epiphanio, Rui, Antonio Formaggio, Bernardo Rudorff, Eduardo Maeda, and Alfredo Luiz. "Estimating soybean crop areas using spectral-temporal surfaces derived from MODIS images in Mato Grosso, Brazil." *Pesquisa Agropecuária Brasileira*, 45 (1):72-80, 2010.

Espindola Giovana, Ana Aguiar, Edzer Pebesma, Gilberto Camara, and Leila Fonseca, "Agricultural land use dynamics in the Brazilian Amazon based on remote sensing and census data". *Applied Geography* 32(2):240-252, 2012.

Ferreira, Karine, Gilberto Câmara, and Miguel Monteiro, "An algebra for spatiotemporal data: from observations to events". *Transactions in GIS*, 18(2): 253–269, 2014.

Galford, Gillian L., John F. Mustard, Jerry Melillo, Aline Gendrin, Carlos C. Cerri, and Carlos EP Cerri. "Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil." *Remote Sensing of Environment,* 112(2): 576-587, 2008.

Gray, Jim, David T. Liu, Maria Nieto-Santisteban, Alex Szalay, David J. DeWitt, and Gerd Heber. "Scientific data management in the coming decade." *ACM SIGMOD Record* 34(4): 34-41, 2005.

Griffiths, Patrick, Tobias Kuemmerle, Robert E. Kennedy, Ioan V. Abrudan, Jan Knorn, and Patrick Hostert, "Using annual time-series of Landsat images to assess the effects of forest restitution in post-socialist Romania." *Remote Sensing of Environment* 118:199-214, 2012.

Griffiths, Patrick, Sebastian van der Linden, Tobias Kuemmerle, and Patrick Hostert. "A Pixel-Based Landsat Compositing Algorithm for Large Area Land Cover Mapping," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5): 2088-2101, 2013.

Hansen, Matt, Yosio Shimabukuro, Peter Potapov, and Kyle Pittman, "Comparing annual MODIS and PRODES forest cover change data for advancing monitoring of Brazilian forest cover". *Remote Sensing of Environment* 112(10):3784-3793, 2008.

Ierusalimschy, Roberto, Luiz Henrique Figueiredo, and Waldemar Celes Filho. "Lua-an extensible extension language." *Software, Practice and Experience* 26(6): 635-652, 1996.

Ihaka, Ross, and Robert Gentleman. "R: a language for data analysis and graphics." *Journal of computational and graphical statistics* 5, no. 3 (1996): 299-314.

IGFA: Belmont Challenge White Paper. *International Group of Funding Agencies for Global Change Research* (2011).

Keogh, Eamonn, and Chotirat Ann Ratanamahatana. "Exact indexing of dynamic time warping." *Knowledge and information systems* 7(3): 358-386, 2005.

Koubarakis, Manolis, Manos Karpathiotakis, Kostis Kyzirakos, Charalampos Nikolaou, and Michael Sioutis. "Data Models and Query Languages for Linked Geospatial Data." *In Reasoning Web. Semantic Technologies for Advanced Query Answering,* pp. 290-328. Springer Berlin Heidelberg, 2012.

Kaptué-Tchuenté, A.T., J.L. Roujean, and S.M. De Jong. "Comparison and relative quality assessment of the GLC2000, GLOBCOVER, MODIS and ECOCLIMAP land cover data sets at the African continental scale." *International Journal of Applied Earth Observation and Geoinformation* no. 13 (2):207-219, 2011.

Körting, T.S., Fonseca, L.M.G., Câmara, G. "GeoDMA – Geographic Data Mining Analyst". *Computers & Geosciences* no. 57: 133-145, 2013.

Lynch, Clifford. "Big Data: How Do Your Data Grow?" *Nature* 455, 7209: 28-29, 2008.

Maus, Victor, and Gilberto Câmara, "Satellite time series analysis for land use/cover change detection". INPE Technical Report, São José dos Campos, INPE, 2013.

McCallum, Ian, Michael Obersteiner, Sven Nilsson, and Anatoly Shvidenko, "A spatial comparison of four satellite derived 1km global land cover datasets." *International Journal of Applied Earth Observation and Geoinformation* no. 8 (4):246-255, 2006.

Pebesma, Edzer. "spacetime: Spatio-temporal data in R." *Journal of Statistical Software* 51(7): 1-30, 2012.

Pebesma, Edzer, Daniel Nüst and Roger Bivand. "The R Software Environment in Reproducible Geoscientific Research." *EOS, Transactions American Geophysical Union* 93(16): 163–163, 2012.

Perz, Steve, "Grand theory and context-specificity in the study of forest dynamics: forest transition theory and other directions." *The Professional Geographer* no. 59 (1):105-114, 2007.

Planthaber, Gary, Michael Stonebraker, and James Frew. "EarthDB: scalable analysis of MODIS data using SciDB." In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, pp. 11-19. ACM, 2012.

Rakthanmanon, Thanawin, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. "Addressing big data time series: mining trillions of time series subsequences under dynamic time warping." *ACM Transactions on Knowledge Discovery from Data*, 7 (3): 2013.

Ribeiro, Vanildes, Ubirajara Freitas, Gilberto Queiroz, Mário Petinatti, and Eric Abreu. "The Amazon Deforestation Monitoring System: a large environmental database developed on TerraLib and PostgreSQL." *OSGEO Journal* 3:70-75, 2007.

Sinton, David, "The Inherent structure of information as a constraint to analysis: Mapped thematic data as a case study". In Geoffrey Dutton (ed) *Harvard Papers on Geographic Information Systems*. Reading, MA, Addison-Wesley, vol.7: 1-17, 1978.

Stonebraker, Michael, Paul Brown, Donghui Zhang, and Jacek Becla. "SciDB: A Database Management System for Applications with Complex Analytics." *Computing in Science & Engineering* 15(3): 54-62, 2013.

Wickham, James, S. V. Stehman, J. A. Fry, J. H. Smith, and C. G. Homer. "Thematic accuracy of the NLCD 2001 land cover for the conterminous United States." Remote Sensing of Environment 114(6): 1286-1296, 2010.

Wulder, Michael, Jeffrey Masek, Warren Cohen, Thomas Loveland, and Curtis Woodcock. "Opening the archive: How free data has enabled the science and monitoring promise of Landsat." *Remote Sensing of Environment* 122 (2): 2-10, 2012.

Verbesselt, Jan, Rob Hyndman, Glenn Newnham, and Darius Culvenor. "Detecting trend and seasonal changes in satellite image time series." *Remote Sensing of Environment* 114(1): 106-115, 2010.

Verbesselt, Jan, Achim Zeileis, and Martin Herold, "Near real-time disturbance detection using satellite image time series". *Remote Sensing Of Environment* 123(3):98-108, 2012.

Vinhas, Lubia, Ricardo Cartaxo Modesto de Souza, and Gilberto Câmara. "Image Data Handling in Spatial Databases." In *Brazilian Symposium on Geoinformatics*, *GeoInfo* 2003.

Vitousek, Peter, Harold Mooney, Jane Lubchenco, and Jerry Melillo. "Human domination of Earth's ecosystems." *Science* 277(5325): 494-499, 1997.

Zhu, Z., C.E. Woodcock, and P. Olofsson, "Continuous monitoring of forest disturbance using all available Landsat imagery." *Remote Sensing of Environment* 122:75-91, 2012.

ANNEX I - WORK PLAN FOR REQUESTED FELLOWSHIPS

TECHICAL TRAINING FELOWSHIPS

1. **Building and deployment of big Earth observation databases to support data analysis research and use case development**

   Associated Task: Task 1.1
   Scholarship: TT-4A Scholarship, 24 months
   Starting date: month 1
   Duration: 24 months
   Supervisors: Gilberto Ribeiro, Ricardo Cartaxo and Luis Maurano

   This scholarship is concerned with building databases based on SciDB, which will be shared with the project partners for use case development and validation and for assessment of the proposed IT infrastructure. To accomplish the task it will be necessary to design and develop data processing and ingestion tools for SciDB, considering particularities of each selected data set in the scope of the project.

   These databases will include:

   (a) All MODIS MOD09Q1 images at 250 meter resolution from 2000 to 2014;

   (b) The EVI2 vegetation index at 250 meter resolution for the same period, with raw and smoothed data.;

   (c) The TRMM precipitation data for the same period;

   (d) The forest fires data produced by INPE and NASA;

   (e) selections of INPE LANDSAT data from 1973 to 2014;

   (f) selected SENTINEL-2 data;

   (g) data from INPE's DETER, PRODES and TerraClass systems.

2. Interface between TerraLib and SciDB

Associated Task: Task 1.2

Scholarship: TT-4A Scholarship, 24 months

Starting date: month 1

Duration: 24 months

Supervisors: *Karine Reis Ferreira, Gilberto Queiroz and Lubia Vinhas*

TerraLib is a spatial library developed by INPE with a large set of algorithms and methods for spatial querying, image processing, spatial analysis, access to spatial databases, digital terrain models and network analysis, as well as support for cartographical projects and geospatial metadata. SciDB is an array database that is able to store large multidimensional arrays in a distributed way and to perform server-side processing of arrays. In this task an interface between TerraLib and SciDB will designed and implemented. The core concept is TerraLib's "abstract data source" a data source contains geographical data, which is described using a generic method such as XML or Linked Open Data.

The task will build a Linked Open Data description of a SciDB multidimensional array and develop methods in TerraLib so that this description is understood as a TerraLib data source. In this way, multidimensional arrays in SciDB will be associated to geographical metadata information stored by TerraLib.

3. Integration between SciDB/TerraLib and R

Associated Task: Task 2.1

Scholarship: TT-4A Scholarship, 24 months

Starting date: month 13

Duration: 24 months

Supervisors: Gilberto Câmara and Pedro Ribeiro de Andrade

This scholarship will be used to develop the integration between the spatial extension of SciDB (developed by the "Interface between TerraLib and SciDB" scholarship) and the R software. R is an open-source platform and language for statistics and graphics. Many researchers in statistics around the world implement their methodologies in R, making them freely available in the Internet as packages.

The integration between the spatial extension of SciDB and R will be available as an extension of aRT package to create, read, write, and query big geospatial databases without needing to load all data at once into R. Concepts from the SciDB GIS extension provided by TerraLib will be mapped into R classes so that R users that already know TerraLib will have no difficulty to use the package. The current aRT version uses the data structures of package sp, which does not have spatial-temporal representations. This scholarship will adapt aRT package to support the spatial-temporal data structures developed by spacetime package. It will investigate the best map between the data structures for spatial-temporal representation in spacetime and the ones available in TerraLib. The implementation will use the methodologies to integrate R and C++ developed by the packages Rcpp and RInside.

This package will be useful even for pure R users as it overcomes R memory problems to deal with big data. The package will be implemented following R requirements for documentation (manual and vignettes), examples, and automatic tests.

DOCTORATE SCHOLARSHIPS

1. Conception and development of TerraScript

Associated Task: Task 1.2

Scholarship: Doctorate scholarship

Starting date: month 1

Duration: 36 months

Supervisors: Lúbia Vinhas, Karine Reis Ferreira and Julio D'Alge

This scholarship will support the research to embed LUA language into SciDB, creating an extension named SciDB-LUA. This extension will turn SciDB in a high customizable system. This environment will be the basis for running the TerraScript in the server side and also in the client side of applications. TerraScript will have a set of concepts that are fit for spatiotemporal Earth Observation applications and thus are easily understood by domain scientists. This scripting language will be based in a formal algebra proposed by Ferreira et al. (2014) and will be developed as an extension of Lua. The spatiotemporal data types proposed by Ferreira et al. (2014), such as Coverages and TimeSeries, will be stored in SciDB multidimensional arrays. The interface between the spatiotemporal data types of Ferreira et al. (2014) and the SciDB multidimensional arrays will be based on the generic field data type specified in Camara et al. (2014) that can represent different types of spatiotemporal data such as trajectories, time series, remote sensing and climate data. We expect that this PhD thesis will also give rise to at least three scientific papers in high-quality venues.

2. Big Earth observation data space-time R package for deforestation and degradation

Associated Task: Task 2.2

Scholarship: Doctorate scholarship

Starting date: month 1

Duration: 36 months

Supervisors: Gilberto Camara, Pedro Ribeiro de Andrade and Silvana Amaral

This task will focus on the development of space-time data analysis methods for monitoring deforestation and degradation in large data sets of Earth observation images. The methods will be validated for the forest change monitoring as specified in the use cases of Amazonia, but it will be developed as a generic multi-purpose open-source toolbox for change monitoring of any kind of EO data (e.g. drought detection, agriculture yield anomaly monitoring, etc.).

The novel R package builds upon experience of already developed and widely used toolkits such those proposed in Verbesselt el al. (2012) and Griffiths et al. (2012). The new Big EO space-time R package will further improve existing algorithms, implement within SciDB environment and optimize algorithms for large satellite data sets.

The development of the R package and cutting-edge change-monitoring approaches will enable analysis of large archives of high spatial and temporal resolution satellite data (Sentinel sensor and the Landsat Archive) and provide a proof-of-concept for highly demanded capacity to analysis large EO big data sets for wide variety of applications (deforestation monitoring, drought detection, fire risk analysis).

3. Segmentation of Earth observation image time-series

Associated Task: Task 2.2

Scholarship: Doctorate scholarship

Starting date: month 1

Duration: 36 months

Supervisors: Leila Fonseca, Thales Korting and Emiliano Castejon

Image segmentation is a way to separate the image into simple regions with homogeneous behavior. In a single image, methods generally observe the similarity between neighboring pixels. When we introduce the time component, by observing image time-series, this neighborhood must be extended to include the similarity between pixels along time. Methods that consider spectral features and the time dimension define a temporal segmentation.

The problem of segmenting multitemporal is the research focus of this working plan. The PhD student will investigate and propose new methods for detecting homogeneous regions in image time-series. By introducing the time dimension, algorithms must deal with a different neighborhood, and also different similarity measures and threshold intervals.

Most current time-series analysis methods use the single pixel along time to evaluate its behavior in a classification process. Therefore, like in single (static) image analysis, a well-known problem is the common “salt and pepper” effect in the classification maps. With a preliminary analysis made by segmentation, it would be possible to delineate homogeneous regions in time.

Expected results include a comparative study of state-of-art segmentation methods, a temporal segmentation algorithm and scientific papers showing the method and experiments using real data.

POS-DOCTORATE SCHOLARSHIPS

1. Spatial extensions to Digital Time Warping for land use mapping

Associated Task: Task 2.2

Scholarship: Post-doctorate scholarship

Starting date: month 6

Duration: 36 months

Supervisors: Gilberto Camara, João Vianei Soares and Pedro Ribeiro de Andrade

This tasks is concerned with developments of improved versions of the Dynamic Time Warping (DTW) algorithm for land change monitoring. DTW is an algorithm measuring similarity between two temporal sequences (Keogh and Ratanamahatana, 2005). It works by comparing a temporal signature of a known event (such as a person's speech) to an unknown time series (such as a speech record of unknown origin). DTW provides a robust distance measure for comparing time series, allowing similar shapes to match even if they are out of phase in the time axis. DTW is a good method for similarity search in big time series data (Rakthanmanon et al, 2013).

DTW is well suited for classification of land cover and land changes in Earth observation data. We have already obtained good results by applying DTW to time series of vegetation indexes to distinguish between forest, pasture and agriculture in the Brazilian Amazonia. However, currently DTW is a pure time series analysis methods. In this post-doctorate work, we will investigate methods to improve the performance of DTW methods by including neighbourhood information in the DTW analysis. We consider that combining the temporal clustering techniques of DTW with contextual information provided by the regions around each pixel being analysed will result in a better performance of land use mapping applications.

2. **Use of Time Series and Space-time data to monitor forest change in Amazonia**

Associated Task: Task 3.1

Scholarship: Post-doctorate scholarship

Starting date: month 1

Duration: 36 months

Supervisors: Isabel Escada, Karine Reis Ferreira and Silvana Amaral

This Post-Doctoral researcher will analyse deforestation trajectories across the case studies. The focus will be in the analysis of spatial and temporal patterns of the deforestation, including clear cut and forest degradation, in the use cases. This analysis will help defining the Amazon use cases and the sampling design to the validation step. In the first year, an analysis of deforestation data from DETER, PRODES and DEGRAD will be performed for the period of 2005 to 2013 to help defining the Amazon use cases. Contributions in the development of validation methods and analysis should be given, including planning and participation of fieldwork campaigns. In the Year 2, the focus will be in the comparative analysis of the results of DETER, PRODES and DEGRAD aiming to recommend actions to improve the new methods and data. Activities for final validation should be carried out to support the assessment of the potential use of the new methods in INPE's operational monitoring systems. Interactions with researchers of WP1 and WP2, which will develop and produce new methods and data, are expected. This position will also be expected to contribute and lead research articles associated with the project.

3. Specification and Validation of Tropical Agriculture Monitoring Methods and Data

Associated Task: Task3.2

Scholarship: Post-doctorate scholarship

Starting date: month 6

Duration: 36 months

Supervisors: João Soares, Ieda Sanches and Luis Maurano

This Post Doctoral appointment will focus in the specification and validation activities of BIG EO data new approach and methods for Agricultural monitoring. The PD fellow will work under the supervision of Dr. Joao Vianei Soares and Dr. Ieda Sanches over the entire period of the project to work on the four main deliverables intended for the Agricultural sub task:

1) Identification of requirements for organizing the EO data structure;

2) Detection of planted area of soybeans, maize and sugarcane crops in selected areas and compare with classes produced by TerraClass and CONAB;

3) Include rice and wheat and,

4) Detailed assessment of the usefulness of BIG EO data approach for national and global scales.

Deliverable 1 aims at producing requirements to maximize the usability of the BIG EO data approach to detect agricultural fields in the most efficient manner. It will be based on the identification of suitable Agricultural land cover available from Terraclass (http://www.inpe.br/cra/ingles/project_research/terraclass.php) and from CONAB, for the commodities of interest (Soybean, Corn and Sugar cane).

Deliverable 2 will do formal comparisons between agricultural land class identifications using the BOIG EO data algorithms developed in WP 1 and 2 and will feed back with a set of recommendations to the teams of WP 1 and 2 for improvement on data analysis and database production.

Deliverable 3 will include rice and wheat as targets for mapping as well to comply with the strategic national engagement with GEOGLAM under the GEO and G20 agendas (produce information for the 4 main agricultural commodities); keep developing new sets of recommendations to the teams of WP1 and 2.

In deliverable 4 at the end of the project lifetime we should have improved versions for agricultural monitoring in Brazil including other areas in GEOGLAM. The goal is a detailed assessment of the usefulness and validity of the methods developed by the project at national and global basis.