Fiscalização a serviço da sociedade

REVISTA do TCU

Federal Court of Accounts Journal • Brazil • year 48 • Issue nº 137 • September/December 2016

Evolution of Control in the Digital Age

Fiscalização a serviço da sociedade

REVISTA do TCU

Federal Court of Accounts Journal • Brazil • year 48 • Issue nº 137 • September/December 2016

The concepts and opinions expressed in signed doctrinal works are the sole responsibility of the authors.

www.tcu.gov.br

Mission Statement
To improve public administration for the benefit of society through government audit

Vision Statement
To be a reference in promoting a more efficient, ethical, responsive and responsible public administration

Federal Court of Accounts – Brazil Journal, v.1, n.1 (1970) - . – Brasília : TCU, 1970- .
v.
From 1970 to 1972, annual; from 1973 to 1975, triannual; from 1976 to 1988, biannual; from 1990 to 2005, quarterly; 2006, annual; as of 2007, triannual

ISSN 0103-1090

1. Oversight of public expenditure- Brazil Journal, v.1, n.1 (1970) - . – Brasília : TCU, 1970-

Catalogued by Biblioteca Ministro Ruben Rosa

TRIBUNAL DE CONTAS DA UNIÃO

FOUNDER
Minister Iberê Gilson

EDITORIAL COUNCIL SUPERVISOR
Minister Aroldo Cedraz de Oliveira

EDITORIAL COUNCIL
Substitute-Minister Augusto Sherman Cavalcanti
Prosecutor General Paulo Soares Bugarin
Eduardo Monteiro de Rezende
Rainério Rodrigues Leite
Flávia Lacerda Franco Melo Oliveira

EDITORIAL RESPONSIBILITY
Serzedello Corrêa Institute
Documentation Center

CONTRIBUTORS
Biblioteca Ministro Ruben Rosa

TRANSLATION
Department of International Relations

Images
iStock

DESIGN
Pablo Frioli

Editorial Design and photomontage
Vanessa Vieira - ISC/Seducont

Documentation Center
SAFS Quadra 4 Lote 1
Edifício Anexo III - Sala 21
Brasília-DF
70.042-900
revista@tcu.gov.br

Printed by Sesap/Segedam

Letter to the Reader

Dear reader,

Public Administration is going through a period marked by important challenges brought about by technological evolution of the Digital Era and by the growing involvement of society in public policy formulation and in discussions on the efficacy and transparency in the use of public resources. We are witnessing a transformation in the relationships between government and citizens. External control cannot be indifferent to this.

Those of you who follow the *TCU Journal* know that the topic is not unprecedented in this publication. Even so, we are far from exhausting the subject in view of the numerous possibilities and perspectives that this new era brings with it. This issue, in particular, is devoted to a look at the near future – at the advances that may be conquered based on the most modern data mining techniques, semantic analysis of texts, geoprocessing and virtual reality.

Investments in science and technology are essential to sustainable development and to building productive economies, with societies that are fairer and more inclusive. According to the Organization for Economic Cooperation and Development (OECD), "one of the most important lessons of the past two decades has been the pivotal role of innovation in economic development". Thus, by intensifying the use of Information Technology in external control and share their experiences, the Brazilian Federal Court of Audit (*Tribunal de Contas da União* - TCU) shows paths the State should follow to become more efficient, transparent and effective in catering to the needs of the population.

There is no doubt that auditing, in the Digital Era, has to consider intensive use of the so-called Big Data Analytics. Processing colossal amounts of data to extract from them knowledge that is useful to public management and its oversight, is more and more present thanks to the evolution of machine learning and artificial intelligence. As Cezar Taurion, the interviewee of this edition, reminds us, advances in this area are so quick that it is difficult for us to even imagine the results that we can achieve in the next few years.

By taking an avant-garde position in using these instruments, TCU has already reaped the benefits of predictive models and intelligent algorithms produced by our own technical teams. In the "Opinion" column, Wesley Vaz, Head of the Department of Information Management for the External Control, in charge of these initiatives, it is a question of incorporating new tools to the work routine of the auditors. In this manner, they will choose better the control activities to be performed by the Court and immediately execute them in an efficient, timely and precise way, thus making the Court's performance more effective.

Similarly, new work methods are being developed based on capturing, treating and analyzing images – either through traditional geoprocessing techniques or virtual reality innovations – that promise to revolutionize in a short time the performance of environmental and public works auditing as well as audits of other ventures involving large territorial extensions.

Finally, as we can see in the several articles published in this edition, it is important to highlight that digital transformation of the control activity is not a subject of interest exclusively for Courts of Accounts and similar agencies. Brazilian universities, such as the *Universidade de Brasília* (UnB) and the *Universidade Federal de Minas Gerais* (UFMG), as well as international centers of excellence in accounting and audit - such as Rutgers University, in the United States - have research teams devoted to developing solutions that potentialize the use of technology and large masses of data as tools to combat misuse and frauds involving public resources with more efficacy.

Therefore, we face the challenge of continuously innovating, focused on results oriented by the strategy of the Court, viewing at consolidating itself as a protagonist in the national and international control system. Moreover, it is imperative for us to be able to carry out the mission given to us.

Enjoy reading this issue!

Bruno Spada

Aroldo Cedraz de Oliveira
President of the Federal Court of Accounts – Brazil and Supervisor of the Editorial Council of the TCU Journal

Interview 07

Opinion 13

Highlights 17

Interview

Opinion

Highlights

Articles 17

Articles

Technological innovations in government auditing

Vinicius Magalhães

Cezar Taurion

Partner and Head of Digital Transformation at Kick Ventures and VP of Innovation at the Smart City Business Institute

In addition to being an Information Technology professional, Cezar Taurion is an avid scholar in this area since the late 1970s, having produced nine books on the most varied branches of IT, such as open-source, grid computing, embedded software, innovation, cloud computing, big data and digital processing. His academic training includes knowledge in Economy, Computer Science and the Marketing of Services. He was a professor of the MBA in Strategic IT Management at FGV/ RJ and of Internet Entrepreneurship at NCE/ UFRJ. In recent years, he has actively participated in the development and application of new technologies, both in Brazil and abroad, which allows him to follow real cases with the most diverse characteristics and complexities.

While performing these activities, Taurion seeks to understand and evaluate the impacts of technological innovations on organizations and their business processes. In his opinion, thinking outside the box is insufficient because "the box still restrains us to a source of reference, limiting innovative thinking." In this interview with the TCU Journal, the IT expert tackles topics such as digital transformation, exponential technologies, Artificial Intelligence (AI) and the impact of all this on society, business and on the activities of the Federal Court of Accounts of Brazil.

It is the Federal Court of Accounts´ responsibility to verify if government institutions are using public resources effectively and efficiently. How can technological innovations help TCU fulfil this mission and "do more with less"?

Technological evolution advances at an exponential pace, surprising those who are most unsuspecting. The computing capacity of a $10 million supercomputer in 1985 is now embedded in a smartphone. The Internet is becoming more and more popular. A quick look at the recent past shows how impressive the pace of change has been. In 1995, a mere 21 years ago, there were 35 million Internet surfers worldwide. Today we amount to almost three billion, about 40% of the world's population. 80 million used cell phones - today three out of four people in the world have one, or more than 5.2 billion people. We should not think of the digital revolution as a simple automation of current processes. We are in a whole new game. Work, business models, organizational structures, products and services will be transformed effectively through digital technologies. Today, technologies make it much easier and cheaper than 10 years ago to cross-reference

data from a variety of sources, ranging from conventional databases to social networks, where we all leave our fingerprints. Citizens are more empowered and, with their smartphones, are connected to thousands of others through social networks, sharing their viewpoints, opinions and comments. Thus, in addition to technologies cheaper and more available technologies, oversight activities can rely on the collaboration of society itself.

Which technologies would you rate as most important for External Control? Why?

Some technologies like the Internet and smartphones allow for quick and easy access at any time. All governing bodies can and should be connected, continuously sending information. Why wait weeks and months to analyze data if such data can be analyzed continuously? The actions will have a much more immediate effect. Analyses can be much more effective if we use algorithms and artificial intelligence. Algorithms are already everywhere, from recommending books and movies to fraud detection by credit card operators. We are immersed in an ocean of data and little use is made of them. Estimates indicate that in 2020 we will be generating 73.5 zettabytes of data throughout the globe or 73 followed by 21 zeros! Results from algorithm applications show that good results can be achieved in almost any situation. For example, an American Psychological Association study analyzing 17 case studies on hiring practices by large firms showed that algorithms beat best practices (usually based on intuition) by 25% when considering the success of the hiring; in other words, hiring a new employee at the company. With the use of predictive algorithms, we can not only content ourselves with describing and analyzing past facts, but also making predictions for the behavior of a public agency based on the patterns observed so far.

"All governing bodies can and should be connected, continuously sending information. Why wait weeks and months to analyze data if such data can be analyzed continuously?"

In your opinion, what is the biggest obstacle when introducing technological innovations at the workplace? Cultural barriers or financial constraints? Why?

Of course, there are financial constraints. However, the cost of technology has been dropping and will continue to fall dramatically. Nowadays, you can buy for about a hundred reais one-terabyte (one trillion bytes) storage devices that cost hundreds of thousands of dollars a little over ten years can now be bought for about a hundred reais. The great barrier is the cultural one. Thinking in a linear way, when evolution is exponential, leads us to the terrible trap of underestimating the impact of transformations. An example of how exponential change is underestimated was the Human Genome Project. It was launched in 1990 and estimated to be completed in 15 years at the cost of $6 billion. By 1997, or half the timeframe, only 1% of the human genome had been sequenced. Through the linear planning that we adopted, assuming 1% in 7 years, it would take us 700 years to complete the sequencing. Sounds logical, right? The pressure to end the project was immense, but when the futurist Ray Kurzweil was asked, he said "1% means halfway. Go ahead!" He thought exponentially. One percent doubling every year means reaching 100% in 7 years. The project was completed in 2001, four years ahead of schedule and costing much less than what was estimated. Linear, traditional thinking has missed the mark for 696 years! Often thinking outside the box is insufficient because the box still attaches us to a reference source, limiting innovative thinking.

The involvement of a citizen in the analyses and decisions taken by the government seems to be an approach that has come to stay. Nowadays, this participation is considered an essential element in a democracy. What technological tools do you think are most effective in promoting a dialogue between public institutions and society?

Social networks are already a part of everyday life for much of society, and this number is trending upwards. If we take a closer look at this phenomenon, we find that the Internet and social networks can be seen as decentralising cognitive technologies through which more and more people can express their ideas and opinions very quickly away from the control of the centra-

lising mechanisms. Let us compare an event broadcast live on a unidirectional TV with a social network. On TV, we are obligated to see the images and listen to the voice that the broadcaster emits, while on the networks, everyone who is present at the event films from a certain angle and emits their own perceptions of the fact, sharing with thousands of others in real time. It creates a less vertical and much more open power topology. Centralized control of the distribution channels of ideas no longer exists, and horizontal exchanges have replaced the vertical and centralized model.

In recent years, oversight activities have migrated from a *lack of information to an abundance of data*. What is the most viable way for an institution - whose main activity depends on the analysis of information - to identify the best data and decide how to use and/or interpret it?

The immense availability of data causes changes in the way we think about data. When we move from thinking based on scarcity to an abundance of data, we must think differently. Because of the difficulty and technological limitation that existed, we ended up building a mental model of data scarcity. Thus, we refine a series of practices, such as statistical analysis by sampling. From a small sample of data, we extrapolate to a broader scenario. With time, we refined the models and they are pretty reliable today. There are, however, some gaps. Accuracy depends greatly on sampling. For example, an opinion poll based on a random sample of landline users has a bias: if the collection was done during work hours, whoever is going to answer the phone does not necessarily represent the opinion of the people who work outside their home. They may have a very different point of view from those that can answer their home telephone during the day. In addition, if we want a more detailed research, a small sample universe, as we do today, will not be statistically representative.

"With the use of predictive algorithms, we can not only content ourselves with describing and analyzing past facts, but also making predictions for the behavior of a public agency based on the patterns observed so far."

What would be a good example of this?

Surveys on voting intentions. They usually take about two thousand people and a general picture is presented. However, if we want to detail to the point where we want to know the intention of young people between the ages of 18 and 25 from the state of Paraíba, the sampling will be insufficient. We are stuck with the initial questions and we cannot get out of them. This thought is different with large volumes. When the variable becomes "N = all" we can make unimaginable granulations on the scarcity model. We can identify trends and detect correlations never thought of before. We can ask new questions and go down to new levels of segmentation. We go to a more opportune mindset; that is, we take advantage of opportunities to ask questions not thought of before analyzing the data. Another interesting feature that affects our way of looking at data is that large volumes do not demand extra precision from each piece of data. In fact, we already do that today. Think of a large number, such as the GDP of a country. We do not detail the cents; we restrict ourselves to the large numbers and the trends they point out.

What is needed in order to get concrete data when faced with an abundance of information?

We have to follow a few steps. First, having a high-level sponsor in the organization is essential. It cannot be an action limited to a certain area or restricted to the IT sector. Afterwards, to define clearly the business objectives and which problems the initiatives for massive data analysis will solve. One key factor: the team. It will be tough for us to get people who meet the hackers + deep statistical and mathematical knowledge + good business knowledge equation. One suggestion is to set up a multidisciplinary team and operationalize the processes involving data analysis projects. Of course the team must have a good manager, one who can understand the different languages spoken by professionals that are quite different from each other and who is passionate about the concept of analyzing data. A bureaucrat manager will not be able to unravel the inevitable problems of inter and intra-team communication. Moreover, setting up a team with only hackers, for example, could generate a sensational predictive algorithm, but of little value to the organization. After all, the goal is not to generate

fantastic analytical models, but to solve business problems like oversight effectiveness and efficiency. Finally, there is governance. Due to the characteristics of dealing with very large and varied volumes with unstructured data (antithesis of the structured and relational model we are accustomed to), the tendency is to not document or create governance processes. Thereby, there is the risk of constantly reinventing the wheel.

IT tools are very efficient in collecting and sorting data, but do you believe they can be trusted when dealing with interpretation?

It really depends on the efficiency of the algorithm. However, the evolution is very rapid. In 2009, when Google was talking about self-driving cars, it was a futuristic curiosity. Today it is almost a reality. These days, Facebook recognition and imaging systems are more efficient than a human being at recognising faces in photographs is. Therefore, we have to think exponentially. That means that if a digital technology is not as effective today, maybe it will be ten or twenty times better in two to three years.

Artificial Intelligence (AI) and Advanced Machine Learning are among the so-called Strategic Technology Trends, as they allow the creation of systems capable of understanding, learning and predicting events that can improve decision making. In your opinion, what are the challenges and benefits of using these technologies for a Court of Accounts?

The revolution led by AI is so fast that we have trouble figuring out how it will turn out. Science fiction imaginary still prevails. I remember one phrase from one of the

> *"Often thinking outside the box is insufficient because the box still attaches us to a reference source, limiting innovative thinking."*

movies that I had at home, "2001: A Space Odyssey," where the computer intruded in the life of the astronauts aboard the vessel: "Just what do you think you're doing, Dave?" Since then, we have seen robot movies like "Terminator", I, Robot, Her and Jarvis, Tony Spark's personal assistant in Iron Man. By the way, Jarvis means 'Just Another Rather Very Intelligent System'. Science fiction might very well be defined as cinematic anticipation. Much of what we see in Jarvis is already in our daily lives in some way.

Jeff Hawkins, Founder of Numenta (and the inventor of the Palm Pilot), says that AI is at a point today similar to where computing was in the early 1950s, when pioneers established the basic ideas of computers. Less than 20 years later, computers made airline reservation systems and bank ATMs possible and helped NASA put a man on the moon, results that no one could have predicted in the 1950s. Guessing the impact of AI and robots in a decade or two is becoming even more difficult." In 20 years, this technology will be one of the main drivers of innovation and technology, if not the primary one," Hawkins says. One cannot be deny that AI will affect society and employment, as we know it. At its inception, automation affected production lines in factories. Now, the risk of unemployment affects functions that were previously reserved for humans. For example, a truck driver. It is one of the most common jobs in the world. There are 3.5 million of them in the United States, and we have more than one million registered to transport cargo here in Brazil. The Dutch government has already conducted a successful test of unmanned trucks crossing Europe. Uber recently paid US$ 680 million to buy Otto, a start-up that develops technology for self-driving trucks founded by Google's AI specialists. The consultancy McKinsey predicted that, within eight years, a third of all trucks on the road will be self-driven, running without drivers. In perhaps 15 years, the truck driver, like the elevator operator, will be an anachronism. Uber invested in Otto not only to operate trucks, but because it wants to operate self-driving car fleets. In September, it began tests on this fleet in Pittsburgh. Canada's postal service wants to send drones, instead of vans, to deliver rural mail.

At the TCU, it seems clear to me that the role of AI will be that of an extremely effective aid in account analysis, analyzing broadly and quickly millions of documents, making comparisons and identifying correlations that we cannot always do as humans. We do not have the ability to handle large volumes of data very quickly. Thus, our decisions are often influenced by personal experiences.

Government institutions are increasingly dependent on systems, technology, and data. With this in mind, the security of information and cyber security play an extreme-

ly important role. In your opinion, what are the main challenges and barriers to making systems and information more secure?

This is one of the biggest challenges. We have two points to validate. One is the guarantee of data security. Another is ensuring privacy. This is a subject that is quite fluid. We leave a huge digital footprint in our day-to-day life. Social networks like Twitter and Facebook are gigantic repositories of opinions and comments. For example, there will be more words written on Twitter in the next two years than those contained in all the books that have already been printed. Facebook, with its more than 1.4 billion users (936 million log on to it every day), is today the largest network in the history of humankind. Looking at this huge amount of users and considering that approximately a quarter of the world is less than 14 years old, this means that about 25% of adults on the planet have a Facebook account. Another inexhaustible source of information about who we are is the famous Google homepage, containing just a single field for data entry. It is a repository for the collective id of humanity. It listens to our confessions, concerns and secrets. We type in what we want into that rectangle, without censorship. Since they are searches that people can do without censorship, they express feelings of hatred and prejudice in them, which are usually camouflaged by social behavior in public. This is the social scientist's problem: what they most want to know is precisely what their objects of study try most to hide. The simple act of asking something uncomfortable creates self-censorship. In the searches, you get information that

"It seems clear to me that the role of AI will be that of an extremely effective aid in account analysis, analyzing broadly and quickly millions of documents, making comparisons and identifying correlations that we cannot always do as humans."

is practically impossible to obtain through traditional searches. Google Trends can generate excellent samples of the private mind, what people really want to know and do not always share with others. They search alone. Only Google knows ... Several projects already show how much can be known with Google Trends, such as stock market forecasting, what drives economic productivity, influenza and dengue epidemics, incidence of racism and entrenched prejudices. As we are discussing something very recent, there is still a long road of learning ahead. The Internet is very young as the predominant human registry and Facebook itself has only gained this share over the past six years. Information on human behavior is still being built, and maybe ten or twenty years from now we might be able to respond more precisely to questions such as how we relate, how new ideas infiltrate and spread through society, how Facebook's timeline will express a person's life (today it receives an average of 0.6 Mb of new data per user per day), how Tweets will show society's reaction to certain events and how Google's rectangle will express the corners of our minds.

Nowadays, innovation is a critical success factor for organisations, including the government. The adoption of new methodologies and technologies is often an essential requirement for innovation. In this view, what methods and techniques, in their assessment, would have the potential to promote innovative control practices?

The impact of digital transformation should be much greater than the concept of e-commerce some 15 years ago. Most of the transactions of whole sectors of industries with their customers today are already done via the Internet, such as banks, commerce and airlines. Buying online is routine. New business models emerge, like AirBnB, and put consolidated sectors in check such as the hotel industry. In many countries, new lending models like those provided by LendingTree (United States) and Kiva, are changing the relationship between society and traditional lending providers: financial institutions. In Germany, Friendsurance is a rupture in the traditional model of the insurance industry.

It is clear that society is already used to using the Web and smartphone apps for their day-to-day activities, whether to take a taxi, buy a product, check-in a flight or transfer money between checking accounts. Cosmetics is an interesting example. Five years ago, Brazilians did not purchase using

the Web. In 2014, they spent 1.3 billion reais. This is still only 1.5% of sales in this market, compared to 6% in the US. Imagine the potential for growth. Again, the one who started it was a business outside the traditional companies from the industry, BelezanaWeb. Using apps like WhatsApp has changed the way people interact. Families exchange more messages between them than speak by phone. Practically all of us share, even compulsively, our ideas, opinions and just about everything we do on social media platforms like Facebook and Instagram. Society is becoming more technological and the consumerization process of IT is a movement that puts pressure on companies to provide the same technologies that their customers already use. They adopt them first!

Organizations have to adjust to the speed of digital transformation. Transforming an organization depends on a change in the mindset of upper management. Managers need to understand the urgency for change in order to provoke it. With this support, a new mentality that encourages change leads to the hiring of talents that do not exist today. Digital transformation requires an organization to move out of its comfort zone. The speed of change no longer allows for long-term planning based only on incremental developments, such as expanding a market or launching similar new products. The unexpected appears at every turn. The business scenario becomes increasingly volatile, ambiguous, uncertain and complex. Big banks only moved quickly towards the digital bank because they were provoked by the Fintechs. The speed of change was not their choice, but

“Transforming an organisation depends on a change in the mindset of upper management.”

a necessary reaction. To visualise and build the future, we have to be optimistic and realistic at the same time. It is not easy, but we have to build it so we will not be run over by it.

Knowledge Management (KM) is still a major challenge for organisations like the TCU, whose main raw material to carry out their activities is information and knowledge. Which approach do you believe will have the greatest impact on KM strategies in the coming years and why?

Knowledge management has a lot to do with the use of algorithms and AI. For comparison sake, let us take a traditional profession, law. One provocation we can make is: “will there still be lawyers in the future?” The rate of success in futuristic predictions is the same as for chimpanzees playing darts and hitting the target, but we can debate some ideas and draw conclusions for ourselves. That is, as long as we do not cling to beliefs and paradigms that limit our critical eye. Let us look at the context. The work practices of lawyers have not changed much in the last few decades. In general, lawyers offer high-cost, personalized advice, and partners in prestigious firms preside over pyramid-shaped organisations receiving high commissions, while battalions of incoming lawyers do the hard work of searching precedents and drawing up contracts. The high costs of these firms and their fees provide a scenario that is open to disruption. There are already some very interesting initiatives that, although still disdained by traditional lawyers, can provoke an Uber effect in the next few years.

Some U.S. law firms already use AI as a “digital associate”, delegating predictive algorithms to the task of performing intelligent searches for documents, opinions, and case law related to the cases under review. Interestingly, an analysis carried out in Europe and the U.S. of the use of AI in law shows that, with rare exceptions, it is not traditional law firms who invest in the concept but new entrants. We can see that the Uber, Airbnb, Skype and WhatsApp moment repeats itself. Established businesses tend to be conservative and strive to preserve their business model. We are finally facing significant changes in society, and virtually no function or business sector will be safe from the transformations. The turbulent scenario, like the context we see today involving Uber and taxi drivers, is bound to happen when law firms feel real threats to their current model. However, some will understand that the process is irreversible and the winners will be the ones who can make the right mix between lawyers and technology. We can say the same about TCU. Aggregate knowledge becomes quite useful when it is handled with agility by AI engines, supporting the work of the ministers.

Opinion

The pillars of the data analysis strategy and consumption of information at the TCU

Efficiency and effectiveness: words that summarize the fundamental needs imposed on all organisations, public or private - that is, to deliver relevant results with the use of the least amount of resources. Doing in the best way possible what should best be done makes efficiency and effectiveness concepts strongly connected. Misalignment between the two can represent excellence to the contrary: to be doing very well something that should not be done. The TCU, with its mission of improving public administration for the benefit of society through external control, must also seek to perform its mission more and better.

Specifically regarding the TCU, as is an auxiliary body of the power to oversee the legality and legitimacy of spending, as well as the proper operation of public policies, we can synthesise efficiency and effectiveness as the need to "do better and choose better what to do." Given the limitations of resources, it is becoming increasingly important to be accurate in the planning phase, designing control activities based on the risk, materiality and relevance identified in each of the objects. "Doing better" represents the proper functioning of the work processes related to external control activities, which demand the full functioning of the methods and availability of tools and trained professionals.

In order to plan and execute work in the most desired way, one of the essential inputs is information. It is the correct and timely use of information that allows organizations to make more appropriate decisions about what to do and how.

Despite the challenge of efficiency and effectiveness being perennial in the corporate world, the scenario today requires us to think about new alternatives. Considering that the amount of information systems has increased to the point where the main administrative acts and facts (contracting, bidding, payments, benefit granting, etc.) can be represented digitally, information on the state's functioning and external control interests are increasingly available - ready to be analyzed and consumed.

The trend towards the massive use of public data is in place and is similar to the current reality of private organizations. Information is and will be increasingly more abundant, distributed and fluid, adaptable to the interpretation of several players. It will thus be increasingly difficult (if not impossible) to retain information or prevent it from being combined to generate other information. The *abundance of information, coupled with the endless ways of combining it, imposes another challenge on us:* to consume it. Having the fastest route to get from one point to another on torrential rainy days is encouraging, even though you know that it only occurs because there is a community of users who contribute by producing and consuming information simply and cheaply. Effective and efficient, is it not?

Whereas the correct use of abundant information on the functioning of the State is a fundamental condition for the proper management of public resources, how to enable public agencies, including control agencies, to make full use of this input in order to increase their efficiency and effectiveness at possible costs? In other words, what needs to be done, in a pragmatic way, for public institutions to adapt to the current reality of demanding better products and services and work better, from the point of

Wesley Vaz Silva
Head of the Department of Information Management for the External Control (SGI)

view of using information as an essential resource? How can we deal with already recognised difficulties and make a move towards instrumentalising work processes with information that is necessary and sufficient to fulfil its purpose?

The following are three pillars inspired by the functioning of public and private national and international institutions that have served as the basis for TCU's strategy, directed towards the full use of information for the sake of control. The execution of coordinated actions has allowed the TCU to advance in the use of data analysis techniques to produce and consume useful information for planning and executing control actions. The belief is that the foundations presented here can be considered by other control bodies, as well as by any Brazilian public institution that wishes to rethink its operation based on the intensive consumption of information.

Two premises were essential to establish the pillars. Both directly confront some myths concerning Public Administration and, for this reason, can be counterintuitive at a first glance. The myths:

1. **Stability**. It must be questioned whether, as was (or still is) widely understood, public institutions have institutional stability and sufficient resources to implement public policies under their responsibility. The global public and private corporate environment is one of uncertainty, subject to constant and increasingly drastic changes.

2. **Infallibility**. We need to recognise that complex solutions are naturally subject to glitches and failures. In certain environments, particularly for public institutions operating in a highly regulated environment, failures can generate discomfort and, ultimately, accountability. The reaction to this is imposed as an inhibition to the emergence of new initiatives for fear of failure. In this context, a new relationship with the risks must be established without, obviously, challenging the principle of legality to which all public initiatives are bound.

Therefore, it is necessary to find solutions that provide new products and services for the proper functioning of institutions, characterised by *innovation in an environment of uncertainty*. Presented by Eric Ries in his book The Lean Startup, this is precisely the concept of start-up, upon which the pillars of promotion of the full use of information by and for control are being sedimented in the TCU's strategy. Let us look at the pillars:

I. Governance

Support from top management and leadership. These are often the most important enablers of corporate governance because they represent the prerequisites for the superior guidelines that support the mission. Without the clear support of top management, transformative initiatives are no more than good ideas that are about to be overcome by fatigue.

However, guidance alone is not enough. The guidelines need to be turned into strategies. In the TCU's case, which was directed towards promoting the use of information for control activity, the choice was to bolster the decentralisation of data analysis activities to the technical units. The departments of external control began to train their auditors to begin the data analysis work, as well as to seek

access to information that would be useful to their work, to use technological support tools and to form an interest and research group on the subject.

The sum of support from top management with the formation of teams of auditors has generated positive results and reinforced promotion of the construction of a new culture, where decentralisation and coordination among the roles encourage the consumption of information for external control activities. The success of this strategy depends, among other factors, on the clarity of the roles and responsibilities of each stakeholder, on the trust among stakeholders, on the experimentation and agile testing of data procedures and on the rapid assessment of results, whether they are positive or negative.

In two years, many TCU technical units led audit work using data analysis techniques. Without claiming to be exhaustive, they are:

- Ruling 539/2015-Plenary, derived from the Centralization-Oriented Audit led by the Department of External Control of Mato Grosso (Secex-MT) whose rapporteur was Minister Benjamin Zymler. The objective of the audit was to validate the selection of control objects based on a predictive data analysis of voluntary transfers.

- Ruling 718/2016-Plenary, carried out by the Department of External Control - Social Security, Labour and Welfare (SecexPrevi) having Minister Vital do Rêgo as rapporteur, which aimed at structuring a continuous oversight of social security benefits regarding their concession, maintenance and payment in the scope of the National Institute of Social Security.

In 2016, some TCU technical units (like Secex-MT, SecexPrevi, Secex-CE and Secex-Education) formalised the existence of their own data analysis centre, made up of auditors capable of using data analysis techniques in their tasks. The work of these groups is based on collaboration and integration with the other TCU support units, seeking to insert new methods and tools into the daily work processes with a light and results-oriented structure.

New institutional groups have emerged at the external level to enhance collaboration and joint work related to data analysis. We highlight the national strategic information network for external control (*Rede Infocontas*), created in 2013 and which recently bolstered the creation of an information analysis unit in each of the state courts of accounts in Brazil. At the international level, a working group on Big Data and Analytics, made up of several countries including Brazil, was created at the last International Congress of Supreme Audit Institutions, held in December of 2016. The objective is to promote collaboration and the exchange of good practices, as well as to discuss and encourage the design of new strategies and the use of new methods and tools based on information to improve the effectiveness and efficiency of the Supreme Audit Institutions. In addition, like what already occurs in TCU with the role assigned to the Information Management Secretariat for External Control (Seginf), Supreme Audit Institutions of the U.S., India and China have created organiztional structures responsible for establishing and executing the corporate strategy for the use of information and data analysis techniques in their operations.

Once the guidelines have been established, the strategy defined, the professionals trained and the objectives defined, it is necessary to get the tools and inputs.

II. Plataform

In the words of researcher Doug Laney, "information is only useful when it can be fully used." If information is the input, the raw material, its full use does not only depend on its existence or availability. The task of making information consumption truly simple involves developing new methods, tools and professional skills, but first information must be obtained and made fully available for use.

This is the essence of Labcontas, a virtual environment created and managed by the TCU, which allows auditors to access internalised information from dozens of databases resulting from cooperation agreements between Federal

Public Administration institutions and the TCU itself.

In order to consolidate itself as a de facto collaborative platform, Labcontas is also used by external partners; mainly state courts of accounts (TCEs). There are currently 25 of these partners, of which 20 are TCEs, which use the software and the public information of the platform and, in return, contribute public information about the functioning of institutions from their jurisdiction and of mutual interest to control partners (contracting a determined state from the federation, for example).

Having a central point that gathers information and software for all its collaborating professionals is essential for data analysis and the consumption of information to be more efficient. The recent stimulus to open the data in public government, embodied with the publication of regulations that support the exchange of information between the public agencies (see Decree 8.789/2016) added to the increase in the quantity of information available (not always quality), suggests the emergence of structuring initiatives geared towards *the construction of an open public data platform* of the State, focused on the advanced analysis of information by public organizations, which would serve as a point of contact for initiatives both in the area of control actions and in the intention to become an effective instrument for management.

With the environment ready, information available and the structure assembled, it time for products.

III. Information-based solutions

The global organisations that have consolidated their operations based on the production and consumption of information have a true neurosis in attempting to keep their information ready for use in the simplest possible way and at a marginal cost, close to zero. In other words, the main goal is to make information very easy to use and inexpensive.

The point is that, though we have qualified teams and data at our disposal, if the intent is to foster the internal and external information-consuming ecosystem, solutions that solve the problem need to be built (weather it is to plan an audit or have inputs to hold a manager accountable or not) in a very simple way, in the vision of the one who will use this input to make decisions. Even if solutions use complex algorithms based on machine learning and cognitive processing or more recent indexing and content relevance calculation techniques, what should matter to the auditor are the results that are obtained from those products and how reliable they are for each purpose.

Concerning the TCU, it is worth mentioning the DGI Consultations, a solution that provides a search interface for information produced and under the stewardship of the TCU in a simple and direct way and that exemplifies how auditors for the sake of their daily work can access a great volume of information in a very simple way. Inspired by this example and made possible through cooperation with the Brazilian Ministry of Transparency, Oversight and the Office of the Comptroller General, ALICE (Analyzer of Bids, Contracts and Public Calls) produces and sends automatic electronic messages containing a risk warning on bids published on the previous day, considering aspects such as the amounts involved and indications of irregularities obtained directly from the texts of the public calls for bid in comparison with TCU precedents.

These are two examples of information-based solutions that serve as a further tool to support risk-based planning processes, as well as providing relevant inputs during the execution of controls and process instructions through the comprehensive use of internal and external information available at TCU. The time is not far when even automatic preliminary analysis of the evidence can - why not? - be delivered to auditors, to support them in their evaluation work.

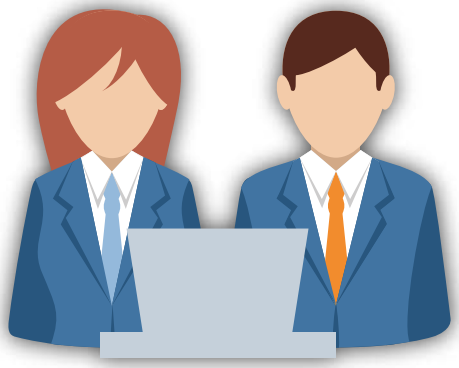
Recognising the importance of the full use of information in public control institutions is a fact, but carries with it many questions. The heterogeneous maturity of organiztions and the passive problems that affect their full functioning make any change of strategy in the way work is done very challenging. However, today's demands do not seem to provide us with alternatives other than to enter the reality of the fourth industrial revolution, of digital thinking, where the consumption of goods and services is characterizd by the requirement for low cost and quality. Thse societal desires emerge and are increasingly directed towards the functioning of the State.

In addition to a critical analysis on the common sense of the functioning of public organizations, the need to shape up to this new reality imposes on public professionals an honest reflection on their role as part of an organization that will never be ready enough, but that needs to deliver more and operate better.

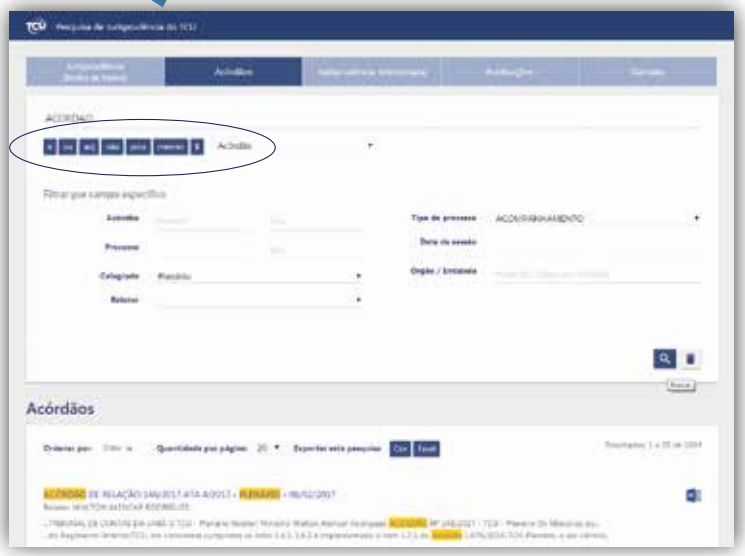
In a pragmatic way, it is already possible to see good results from the execution of the strategy of using information and data analysis in TCU's work processes. Apart from the occasional scaremongering, but keeping alive the ability to be inspired and compare oneself to a global and local reality, there is no alternative but the constant rethinking of the way the mission of public institutions must be fully fulfilled. The desired increase in efficiency and effectiveness certainly involves the full use of information, which demands involvement and increase of the interaction between professionals with complementary profiles, the use of new technologies, the management of information and, above all, a change in the way of planning and executing the work processes.

New precedents search

Same standard adopted by the Supreme Courts



Simpler, faster and more precise search



Among the main changes we highlight:

- Simultaneous search in all of the precedents databases;
- Logic operators used by the supreme courts
("and", "or", "adj", "no", "*prox*" (next), "*mesmo*" (same) and "$";
- New resources: search history, index and browsing by result;
- In Selected Precedent: search by legal reference, browsing by the classification tree, and use of the External Control Vocabulary (VCE) to allow searching by synonyms;
- Cleaning the content of bases.

Fundamentals of auditing in mixed reality

Aroldo Cedraz de Oliveira
is president of the Federal Court of Accounts-Brazil. He has a PhD in Veterinary Medicine from the University of Hannover, Germany, and an M.A. in Veterinary Medicine from the Federal University of Santa Maria/RS. He is also a Honorary Professor (h.c.) of Nanjing Audit University (NAU) and was Congressman during four terms. He is editorial coordinator of the publication *"Public administration control in the digital era"*.

Francisco Osório Ramos
is an employee of the Federal Court of Accounts – Brazil. He has a B.A. in Electric Engineering from the University of Brasilia (UnB) and an M.A. in Physics from the same institution. He is a Physics Professor at UnB, an Electric Engineering professor at Paulista University (UNIP) and at the Brasilia Higher Education Institute (IESB), both located in Brasília. He is also a professor of Computer Science at the Brasilia University Center (UniCEUB), where he teaches Artificial Intelligence, Computer Graphics and Computer Architecture.

SUMMARY

Supported by information collected from the website of the National Aeronautics and Space Administration–NASA and the Jet Propulsion Laboratory–JPL, we analyze the computer graphics techniques employed in the project OnSight–NASA and how they can be applied to the audit of public works. Based on this theoretical reconstruction, the foundations of a new audit modality called MRA – Mixed Reality Auditing are established. This consists of a remote audit procedure in real time that uses images from digital cameras at the place of the audit, transmitted to Head-mounted Displays-HMDs, or special mixed reality glasses, used by an observer who is at a separate location.

Keywords: Computer graphics. Mixed reality. Rendering. Frustum. Quadtree. Game engine. Audit of public works.

1. INTRODUCTION

The Jet Propulsion Laboratory-JPL, a technological center for research and development in partnership with NASA, has been offering to the public a guided tour of an area on planet Mars since March, 2016 (JPL, 2016). This is the mission Destination: Mars, an interactive exhibit using mixed reality glasses, one of the recent advances in virtual reality. Visitors can explore

various locations on Mars, reconstructed from real images sent to Earth by Curiosity, a motorized robot jeep that explores the surface of the red planet since August 2012 (Ibid., 2016).

Destination: Mars is an adaptation of the project OnSight, a software tool for missions on Mars. The data and images transmitted daily by the vehicle Curiosity are picked up on Earth by three giant 80-meter antennas. Then, using a mixed-reality device, this information is transmitted from the surface of Mars to a laboratory on Earth. Scientists are enveloped by the images captured in the vicinity of the vehicle on Mars and may wander around rocky surfaces or crouch down to examine geological formations at different angles. OnSight provides scientists with a way to plan and, together with the vehicle Curiosity, conduct operations on Mars, exploring the planet directly from their offices in Pasadena, California (Ibid., 2016).

2. HMDS AND DIGITAL REALITY

HMD or Head-mounted display is a video device used over the head like a helmet, containing a wide display, headphones and an interface through which the user can experience a virtual reality environment (MELZER; MOFFITT, 2011). It consists of a transparent display mounted on the head of the observer, where the information is presented without requiring the user to look outside of their normal field of vision, as shown in Figure 1 (CANON, 2016). Since its pioneer studies, researchers in the field of virtual reality have used HMDs as essential devices for visual display.

Virtual reality consists of an immersive multimedia or computer-generated reality that replicates an environment, simulating physical presence in real-world or imaginary-world locations, allowing the user to interact within that world. In other words, it is a generic

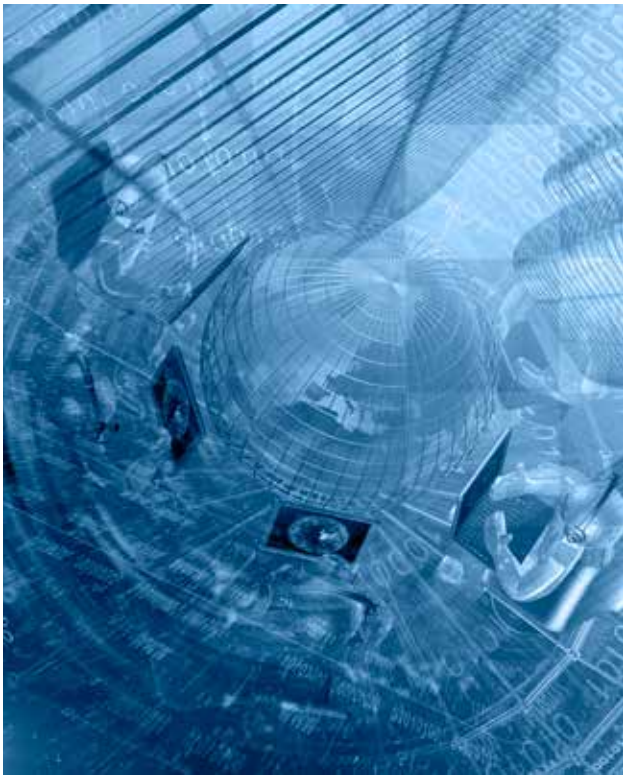
Figure 1:
HMD device

Source: CANON (2016)

term, applicable to all kinds of immersive experiences, created using only real or synthetic world content, or a hybrid of both (LACKEY; SHUMAKER, 2016). Examples of virtual reality goggles include Oculus Rift, by Facebook; Gear VR by Samsung; and Project Morpheus by Sony.

According to Lackey and Shumaker (2016), augmented reality is a direct or indirect view of a physical real-world environment whose elements are augmented or supplemented with entries generated by sensors, such as sound, video, graphics or GPS data. It is synthetic content superimposed on the real world, where that content is neither attached to nor part of it. The content of the real world and the computer graphics are not able to interact with each other. Examples of augmented reality glasses are Google Glass, the Daqri Smart Helmet and the Moverio by Epson.

Mixed reality, also referred to as hybrid reality, consists of a fusion between real and virtual worlds to produce new environments and visualizations where physical and digital objects co-exist and interact in real time. It is a synthetic content superimposition on the real world of the user, the world connected to him or with which he interacts (Ibid., 2016). Among the mixed reality glasses are the HoloLens, the Canon MREAL Display MD-10 and the Magic Leap.



Modernly, it is admitted that virtual, augmented and mixed realities can be gathered under a more generic name: digital reality.

3. LEVEL OF DETAIL AND RENDERING

Computerized stereoscopic vision (stereovision) consists of the extraction of 3D information from images such as those taken by a digital camera. Comparing the scene from two selected points, 3D information can be extracted by examining the relative positions of objects in both scenes. The information collected from these stereoscopic images is projected onto a free surface, created for this purpose (AKENINE-MÖLLER; HAINES; HOFFMAN, 2008).

Two cameras, spaced horizontally from each other, are used to obtain two different views of a scene, in a way similar to human binocular vision. Comparing these two images, relative depth information can be obtained in the form of a disparity map, which encodes the difference between the horizontal coordinates of the corresponding points in the images. The values in this disparity map are inversely proportional to the depth at the corresponding location of the pixels (Ibid., 2008).

With the use of computerized stereoscopic vision, for each image that reaches the observer's HMD a completely new 3D model is generated through automated algorithms that construct a mathematical and statistically accurate model of the analyzed surface, referred to as stereoscopic correlation (COLANER, 2016). Observing two images and finding the differences between them with the aid of a disparity map and the use of a particular model of camera, the distance between each point is ascertained and these images' range maps are generated.

In order to speed up image processing, all objects that are outside the observer's field of vision are removed in the rendering process (the process of generating an image from a 2D or 3D model through digital processing, consisting of three basic steps: determining the virtual camera, visible surface and light sources). Only the pixels that are visible in the *frustum* need to be rendered. All objects that are visible from a particular point are precomputed, whereupon all non-visible objects are immediately removed, thus reducing the number of objects that intersect the *frustum* (Ibid., 2016).

The *frustum* is the field of vision from the eye of the beholder and can be represented by an imaginary spatial volume containing all that is visible in a three-dimensional scene. It can be represented by a truncated

pyramid, consisting of six planes. Four of these planes correspond to the sides of the screen, called the right, left, top and base *frustum*. The two remaining planes are called the near and far planes of the *frustum*, and define the minimum and maximum distances that the objects of a scene are visible to the observer (DUNN; PARBERRY, 2002). Figure 2 illustrates the geometry of the *frustum*.

To manage the level of detail, the surface of the scene is divided into tiles of different sizes, each with a certain level of detail, following a partition in the surface (2D space) known as the quadtree (COLANER, 2016).

The quadtree consists of a tree data structure in which each internal node has exactly four children, used to partition a two-dimensional (2D) space; for example, a surface with recursive subdivision into four quadrants, regions or adaptable cells. Each cell or reservoir has a maximum capacity. When this capacity is reached, the reservoir is divided. The following tree structure follows the spatial decomposition of the quadtree, i.e. each node undergoes a division into four sub-nodes (LENGYEL, 2004).

Figure 3 shows a quadtree built for an area containing a single object. The illustration on the right shows how the corresponding data structure is organized. Each node has four sub-nodes. If no geometry of world intercepts a quadrant, then this quadrant is not subdivided. Any quadrant that does not contain objects is removed from the tree. It is also assumed that any missing quadrants are empty.

According to Lengyel (2004), organizing the geometry in a quadtree brings the benefit that whenever a node of the tree can be determined as not visible, it becomes immediately known that all its sub-nodes are also not visible and can be eliminated, therefore improving management of the level of detail.

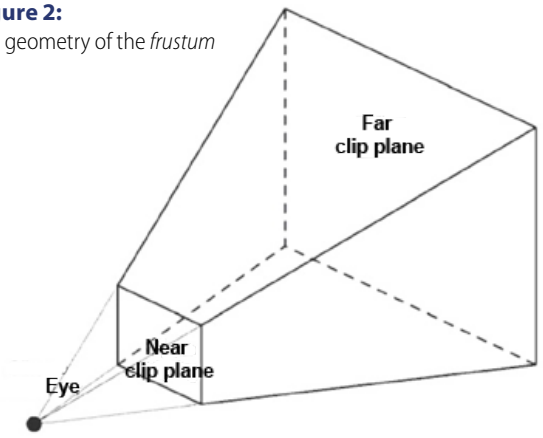
4. GAME ENGINE

An engine for creating video games such as Unity[1] creates software for integration with the hardware involved, proving to be appropriate for user interaction with mixed reality, as the available glasses have an interface with the aforementioned engine. A frame rate of 60 fps (frames per second) is used. Frame rate is the rate at which an imaging-processing device consecutively displays images called frames.

The Unity game engine employs the technique of mipmaps, which consist of sequences of 3D world textures, optimized and pre-calculated, each of them a representation with progressively lower resolution of the same texture, which increases rendering speed. High resolution mipmap textures are used for high-density samples, such as objects close to the camera. Low-resolution textures are used when the object appears farther away. The mipmap technique can improve rendering performance by up to 33% in Unity.

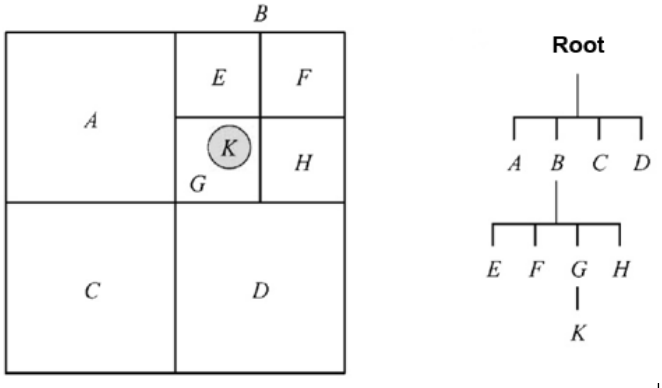
Thus, for the tiles closest to the observer, resolution is maximized. For tiles farther away from one's field of vision, the resolution decreases exponentially. When the observer turns his gaze further ahead, lower resolution textures are superimposed over higher resolution textures. With this technique, images captured on location by cameras on the vehicle Curiosity can be combined with images from satellites orbiting Mars over the place of observation, considering that their resolution is constant in the region (COLANER, 2016).

Figure 2:
The geometry of the *frustum*



Source: Dunn e Parberry (2002)

Figure 3:
Partitioning a 2D space with quadtree



Source: Lengyel (2004)

Additionally, if the viewer clicks on a detail in the image, other zoomed-in images can be displayed, taken from a different angle or even from a satellite in orbit. For their part, mixed reality glasses (HMDs) support the viewer's gaze, gestures and voice commands.

5. AUDITING IN MIXED REALITY

Auditing in mixed reality consists of performing an audit in which the auditor is not physically present at the construction site, but at another distant location. Electromagnetic signals from the surface images captured by digital cameras on site are transmitted to the HMD of the auditor who receives those images and can browse through the site virtually and in real time, although at a distance; for example, in a room at the headquarters of TCU (Federal Court of Accounts of Brazil). In this way, TCU Auditors can perform audits of public works at locations far away from Brasilia without ever having to leave the headquarters building in the federal capital.

The auditor is enclosed by multiple images taken in the vicinity of the work site, all with high levels of detail. He is also able to take various vantage positions and look around at different angles as he browses through a given surface area, even crouching down to examine details from different perspectives and gather audit-related findings.

Audits of road works the following flaws on highways can be observed remotely, using Mixed Reality Auditing –MRA: pothole formations, dips, embankment slippage, worn-down and cracked coating, ruptured drainage components, exudation on curved stretches, patches, degradation, ruptured and eroded road edges, etc.

In regard to audits of railway works, visual observation in MRA can detect the following flaws both in the construction of earthworks, bridges and flyovers, train station platforms as in the railway itself: rot, fractures or excessive holes in wooden railway ties; burns, corrosion or buckling of steel rail tracks. Longitudinal displacement (or drag) of railway tracks, unevenness in the joints, disappearance of clearance between them, contraction, as well as track misalignment can also occur.

As for building audits using, irregularities can be found in the faulty positioning of the construction; in the earthworks (excavation, transportation and landfills); in the foundation (cracks in pipes or in support columns); in reinforced or prestressed concrete structures (inaccuracies of bobs in external corners, columns and elevator shafts); in masonry (cracks or fissures, the use of damaged bricks, inaccurate wall and frames positioning); on the roofing (leaks, broken tiles, roof slope and gutters), as well as in the coverings (i.e. loose boards).

All information relating to basic and executive projects can be inserted into mixed reality software, such as amount of materials, quality and prices per unit to be conferred with the visual findings of the works being audited, simply by staring at a target object and using gesture-based controls to select the commands on the menu bar. Similarly, a virtual measuring ruler

can be inserted to measure the distance between two points.

Internationally, the audit of aid resources in the event of major natural disasters is a recurring theme in the *International Journal of Government Auditing* of INTOSAI. With audit techniques in mixed reality, experts from all over the world, among them SAI auditors, geologists, seismologists, physicists and environmentalists, can monitor occurrences *in loco*, without leaving their countries, simultaneously and in real time.

6. CONCLUSION

This study aimed to analyze computer graphics techniques employed in the project OnSight– NASA, applying them to the audit of public works. Based on this theoretical reconstruction, the foundations of a new audit modality is established called MRA – Mixed Reality Auditing.

The creation of this new auditing concept by the Federal Court of Accounts of Brazil is a major breakthrough in the field of public works and environmental auditing. It will have an immense impact on how this work has been being carried out so far because it eliminates the need for specialists to travel long distances to construction sites or the location of natural disasters, with obvious economy of time and resources for traveling and other related expenses.

The groundwork launched in this study would be a guideline for the TCU to establish a partnership with NASA, initiating a joint project in the area of interest. TCU could also possibly initiate its own project, including the development of the necessary software as well as the selection of the most suitable HMD and the consequent integration with the hardware to be used.

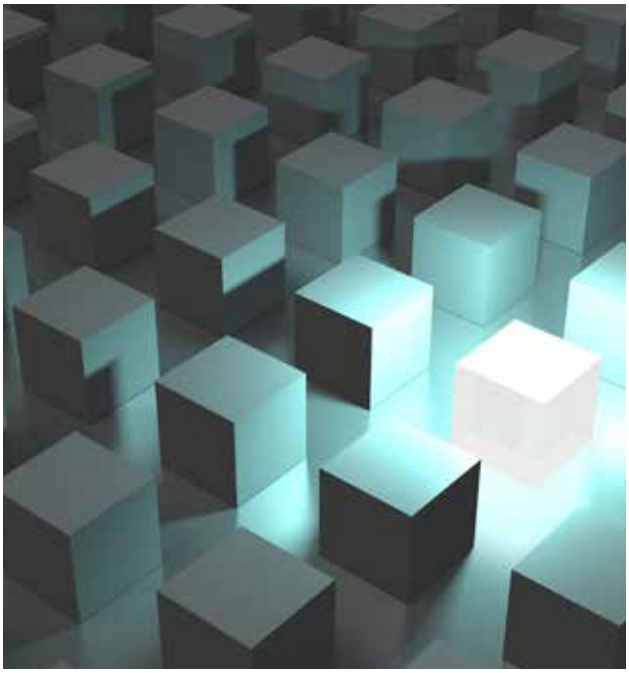
NOTES

1 Available at: <http://unity3d.com>..

BIBLIOGRAPHY

AKENINE-MÖLLER, T.; HAINES, E.; HOFFMAN, N. Real-Time Rendering. 3rd ed. Massachusetts: A K Peters; Florida: CRC Press, 2008.

CANON. MREAL Display MD-10, 2016. Available at: <https://goo.gl/wBeCih>. Accessed on Sept. 14, 2016.



COLANER, S. VR and AR go to Mars: Interview with NASA Scientists Jeff Norris and Alex Menzies. Tom's Hardware, New York, April 7, 2016. Available at: <https://goo.gl/gvLN7H>. Accessed on Sept. 14, 2016.

DUNN, F.; PARBERRY, I. 3D Math Primer for Graphics and Game Development. Texas: Wordware Publishing, Inc., 2002.

JPL. 'Mixed Reality' Technology Brings Mars to Earth. Jet Propulsion Laboratory–NASA, California, March 30, 2016. Available at: <https://goo.gl/RYrCrZ>. Accessed on Sept. 14, 2016.

LACKEY, S.; SHUMAKER, R. Virtual, Augmented and Mixed Reality: 8th International Conference on Human-Computer Interaction. 1st ed. New York: Springer, 2016.

LENGYEL, E. Mathematics for 3D Game Programming & Computer Graphics. 2nd ed. Massachusetts: Charles River Media Group, 2004.

MELZER, J. E.; MOFFITT, K. Head-mounted Displays: Designing for the User. South Carolina: Create Space Independent Publishing, 2011.

TCU open data services platform: crowdsourcing, civic cloud and civic applications

Monique Louise de Barros Monteiro
is an employee of the Federal Court of Accounts – Brazil. She has an MSc. and a BSc in Computer Science from the Federal University of the State of Pernambuco (UFPE). She holds 12 professional certifications: Big Data Certified Science Professional, Big Data Certified Professional, SOA Certified Governance Specialist, SOA Certified Architect, SOACP, PSM, SAFe Agilist, Oracle Master Java EE 5 Enterprise Architect, MTS, MSP, SCJP e IBM OOAD.

Remis Balaniuk
is an employee of the Federal Court of Accounts – Brazil. He has a postdoc in Virtual Reality from Stanford University, U.S.A, and a PhD in Computer Science from the Grenoble Institute of Technology (*Institut National Polytechnique de Grenoble*), France, and an M.A. in Computer Science from the Federal University of the State of Rio Grande do Sul (UFRGS). He has a technology degree in Data Processing from the University of Brasilia (UnB). Currently, he is a professor and researcher at the Catholic University of Brasília.

Marcelo Pacote
is an employee of the Federal Court of Accounts – Brazil. He has an MSc and a BSc in Computer Science from the Federal University of Brasilia (UnB). He has a specialization degree in Development of Mobile Applications, from the University of Araraquara (Uniara), and 17 professional certifications: PMP, CBDP, SOACP, SOACC, CSM, CSD, SCJD, SCJP, SCJA, SCEA (I), SCBCD, SCWCD, RUPF, IRIP, CTFL, ITILF and Oracle SQL Expert. He is the author of a book in the field of Information Technology.

SUMMARY

This article describes the development, availability and sustainability of a platform for open data exposure and information gathering through crowdsourcing, based on its use by civic applications. Among the results expected from the implementation of this platform, we can mention the opportunities of 1) centralization and availability of open data and 2) obtainment of contributions and perceptions of citizens. Such information may be very useful both for the definition of indicators for public services and policies and also to increase knowledge about citizens' wishes in different localities and population spheres, including the society's perception in relation to the services provided to them..

Keywords: civic applications, web services, crowdsourcing, civic cloud, open data.

1. INTRODUCTION

A recurring theme in Public Administration in Brazil and in other nations is the sharing of open data for greater transparency of government actions.

For example, we currently have in Brazil the **Portal Brasileiro de Dados Abertos**[1] (Brazilian Open Data Portal), a tool made available by the government so that everyone can find and use data and public information. The portal includes data on budget execution, cartographic information and service units, as well as different indicators and statistics.

Among the factors responsible for catalyzing initiatives such as the *Portal Brasileiro de Dados Abertos* (Brazilian Open Data Portal) is Law 12.527 - *Lei de Acesso à Informação Pública* (Law on Access to Public Information), sanctioned on November 18, 2011. In accordance with the guidelines established by this law, the general rule is to classify information of collective interest produced or guarded by the State as public, in order to guarantee the fundamental right of access to information. Thus, information should be classified as confidential in exceptional cases, when it is essential to the security of society and the State.

In its planning, the *Tribunal de Contas da União* – TCU (Federal Court of Accounts - TCU) includes actions to promote the disclosure of open data through its information vehicles. In this context, since the beginning of 2015 the TCU has internalized databases from various agencies. This includes data from the *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira* – INEP (Anísio Teixeira National Institute of Educational Studies and Research– INEP), the *Cadastro Nacional de Estabelecimentos de Saúde – CNES*[2] (National Registry of Health Establishments – CNES), Ministry of Social and Agrarian Development (social assistance area), of the *Sistema Nacional de Emprego* – SINE (National Employment System - SINE), *Portal de Compras do Governo Federal* - SIASG/Comprasnet[3] (Federal Government Pro-

curement Portal - SIASG/Comprasnet), of the *Sistema Integrado de Administração Financeira do Governo Federal* - SIAFI[4] (Integrated System of Financial Administration of the Federal Government – SIAFI), of the *Cadastro Nacional de Empresas* - CNE[5] (National Business Register – CNE), among others.

Once the bases are internalized, the data received is made available to society, through consultation by end users or by software systems. The methodology adopted for the provision of the so-called open data is detailed below.

2. METHODOLOGY

In spite of their usefulness, tools such as the Brazilian Open Data Portal traditionally have as a main characteristic the availability of information in a static and / or crude form (e.g. spreadsheets, files in XML, CSV or HTML format), with a large number of records (e.g. spreadsheets with tens of thousands of lines or files with hundreds of megabytes). However, raw government electronic data are often difficult for ordinary citizens to understand. Therefore, the intervention of people, groups and / or companies with the capacity and availability to develop applications and other technological elements that can translate open governmental data into products of interest and public or private utility is necessary.

Appropriation of much of the data set available by developers of civic technologies is difficult, as these databases, typically offered in raw format, are often of low information value and are available in the form of static files. This feature requires the developer to periodically transfer, understand, process, adapt and make them available in applications. Another common requirement is the need for remote and continuously available IT infrastructure, which implies hosting costs (e.g. contracting cloud services) that developers have to pay for. Such difficulties limit the potential for open data usage by civic applications.

Aware of this scenario, the TCU adopted a model of performance in the open data ecosystem based on the need to provide civic technology developers with a service platform where their applications can remotely access processed and updated data in a more adequate and optimized way to be processed by client software systems. Another premise is that the platform offered allows applications to store the data they generate.

First, the data received from the databases cited in the introduction are persisted in a corporate relational database.

Once the data persistence step in the TCU storage infrastructure is complete, they are ready to be accessed by both end users and applications through web services. According to Erl et al. (2012), "from a general perspective, a service is a software program that makes its functionality available via a published technical interface, called a service contract". Considering that such software services may be implemented by different means, we will use in this article the name web service to refer to services available on the World Wide Web.

Each web service is responsible for grouping and exposing a set of related operations. Such operations can be accessed by externally developed applications to obtain data through standardized protocols according to World Wide Web standards. Such protocols include, for example, Hypertext Transfer Protocol (HTTP) (FIELDING et al., 1999).

Figure 1 shows the concept of web service and its operations at a conceptual level. We have two web services represented and each of them deals with a certain "theme", grouping query operations that will be executed on the persisted database. Each operation, in turn, supports the receipt of parameters that allow you

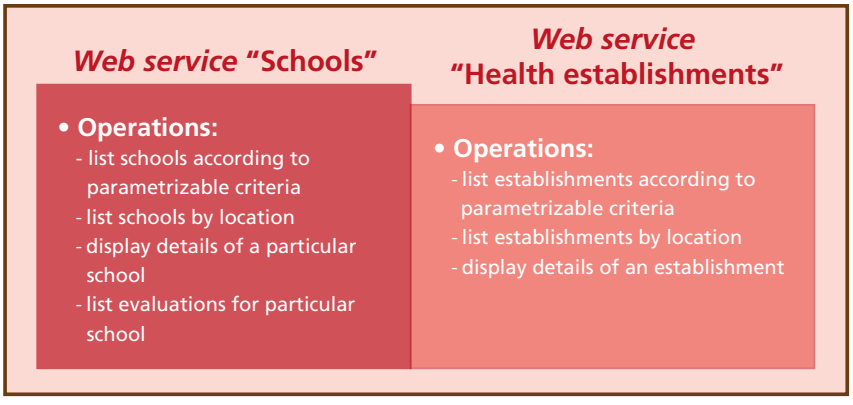


Figure 1:
High-level representation of two web services. One groups query operations on schools, and another, query operations on health establishments

Source: Own elaboration

to filter dynamically the search criteria as well as the attributes and amount of data returned.

3. DEVELOPMENT

The web services were initially designed as a support tool for the development of civic mobile applications, in the context of the *Projeto BEPiD - Programa Educacional Brasileiro de Desenvolvimento para IOS* (BEPiD Project - Brazilian Educational Development Program for IOS), through a partnership between TCU and *Universidade Católica de Brasília* - UCB[6] (Catholic University of Brasilia – UCB). This partnership involved the development of civic applications by students of that university, who used the information channels provided by the web services developed by TCU.

Web services packages were specified and developed for the following groups of information:

- **Health**: information on medical specialties, health establishments, professionals, medicines and specialized services, all within the scope of public services.

- **Social assistance**: information from referral centers and specialist reference centers for social assistance (CRAS and CREAS, respectively).

- **Jobs**: information regarding SINE posts.

- **Education**: education information, including basically registry data of educational establishments and their resources.

We will name the set of web services mentioned above "civic cloud". From here on, to refer specifically to the set of operations defined in the contract or interface provided by web services, we will use the acronym API, acronym for Application Programming Interface.

The API initially proposed and made available followed the requirements expected by the "*Mapa da Saúde*" (Health Map) and "*Nossa Escola*" (Our School) applications developed by UCB students.

Due mainly to the need of collecting feedback from citizens - a practice known worldwide as crowdsourcing - the applications were developed following a social network model, thus requiring a registry of users and of public service evaluations provided by them. The **metamodel for civic applications** emerged from this. It is an additional set of web services, which allows consultation, addition, change, and deletion of information related to users, applications, user groups, posts, hashtags, notifications, and so on.

In addition, in order for the same model to be adopted both for public health and for education, as well as for other areas of expertise of the State, it is possible

for users to submit assessments, in the form of **postings**, of generic entities denominated here as **objects**. Thus, a user can evaluate an object, which could be, for example, a school or a hospital, depending on the usage context determined by the application in use.

The masses of data generated by civic crowdsourcing will be invaluable sources of information, in particular for the design of indicators that will allow the discovery of knowledge about the functioning of the public administrative machine, the citizen's needs and perception. Crowdsourcing, when used as a practice of transparency, allows large amounts of data to make sense, and can generate new ideas for the development of projects for society.

This knowledge will serve in the future as input top TCU when planning control actions. Finally, the collection and storage of data in a format defined by the TCU itself facilitates future initiatives of statistical analysis and data mining, both predictive and prescriptive, ranging from simpler analyses to more complex applications based on machine learning.

As an example of a civic application, Figures 2 to 3 show screen shots of the Health Map application. The app has search features for health establishments within a radius of location, display of detailed data for each establishment (including, but not shown in captures, listing of medical specialties and quantitative of professionals by function) and also a functionality for the user to evaluate the quality of care provided.

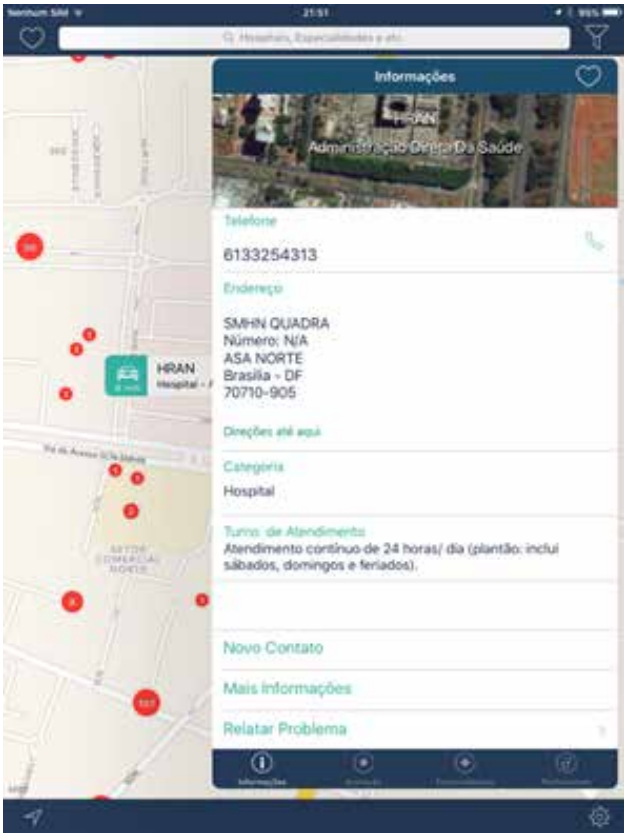
4. TECHNOLOGICAL ASPECTS

The civic cloud was developed according to the REST style - Representational State Transfer (FIELDING, 2000). REST encompasses an architectural style for the construction of applications and APIs, and roughly, provides a mechanism in which, in the context of

Figure 2:
Health Map - search for health establishment in radius, displayed on a map (satellite view)

Source: Own elaboration

Figure 3:
Health Map - detail of an establishment



Source: Own elaboration

web services referenced here, each operation of a web service can be represented as follows:

Method HTTP **(GET | POST | PUT | DELETE) + URI** (*Uniform Resource Identifier*)

Each of the above **methods** represents one of the possible methods or operations supported by the HTTP protocol, to be executed on a **resource** identified by a **URI**. Briefly, the methods have the following semantics:

- GET: Used to retrieve information (ex.: query operations);
- POST: Used to create new features (ex.: register new user);
- PUT: Used to update features (ex.: update an existing user's registration data);
- DELETE: Used to exclude features (ex.: delete a user).

Examples of calls to API operations that can be executed by applications:

- GET <http://contas.tcu.gov.br/nossaEscolaRS/escolas/123>: retrieves information from the school externally identified by the URL <http://contas.tcu.gov.br/nossaEscolaRS/escolas/123>. It is considered in this example that "123" is the internal school identifier to the application (ex .: key in a database);
- DELETE <http://contas.tcu.gov.br/appCivicoRS/pessoas/456>: delete the user externally identified by the URI <http://contas.tcu.gov.br/appCivicoRS/pessoas/456>. It is considered in this example that "456" is the internal school identifier to the application (ex: key in a database);
- POST <http://contas.tcu.gov.br/appCivicoRS/pessoas> {"nome":"José da Silva", "email": "josedasilva@tcu.gov.br", "dataNascimento":"12/12/1950"}: register a new user, sending their information in the body of the message;
- PUT <http://contas.tcu.gov.br/appCivicoRS/pessoas/789> {"nome":"José da Silva", "email": "novoemail@tcu.gov.br", "dataNascimento":"12/12/1950"}:

The API is documented in the Swagger format. This format allows not only the visualization of the contract of the civic cloud (operations supported, formats of the URIs, format of the data trafficked, parameters for each operation) as well as the test of the functionalities. In addition, the documentation is generated dynamically, being updated with each change in the API contract.

Figure 4:
Example of web services listing (metamodel for civic applications)

AppCivicoRS
Web Services referring to the common template for civic applications.

Contact the developer TCU			
applications: Applications	Show/Hide	List Operations	Expand Operations
groups: Groups	Show/Hide	List Operations	Expand Operations
hashtags : Hashtags	Show/Hide	List Operations	Expand Operations
installations : Installations	Show/Hide	List Operations	Expand Operations
notifications: Notifications	Show/Hide	List Operations	Expand Operations
people: People	Show/Hide	List Operations	Expand Operations
posts: Posts	Show/Hide	List Operations	Expand Operations
object types: Object Types	Show/Hide	List Operations	Expand Operations
post types: Post Types	Show/Hide	List Operations	Expand Operations

Source: Own elaboration

Figures 4 to 6 illustrate the contracts of the metamodel API for civic applications and their visualization through the Swagger interface. The other APIs - health, education, social work, and employment - have similar documentation interfaces.

5. CONCLUSION AND FUTURE PERSPECTIVES

The Health Map, Our School, My Medication, Mami and Vacin App applications are real examples of applications developed based on the API available. They are available in the App Store and Google Play stores.

In 2016, the TCU provided another opportunity to validate the model for civic applications described here: the Civic Applications Challenge[7]. It is a hackathon for external developers to register applications that make use of the web services available.

At the time this article was written, the contest was in progress with about ninety applications registered. Throughout the Challenge, we have maintained a direct open channel with the developers, through an institutional mailbox and a website in GitHub. Through this channel, we have received more than forty suggestions for adjustments and extensions to the operations provided by the API. Most of the suggestions were implemented and some are in the list of improvements to be implemented in the next version of the API.

Although TCU proposed and implemented the web services platform, it is open to contributions and partnerships. Its success will depend on the convergence of efforts and ideas in the construction of innovative and useful solutions to the citizen. If successful, it will bring results that can contribute to the improvement of public services and policies, providing new services and resources for society as a whole.

NOTES

1 Available on: <http://dados.gov.br>. Access on: 10 out. 2016.

2 Available on: <http://dados.gov.br/dataset/cnes>. Access on: 10 out. 2016.

3 Available on: <http://www.comprasgovernamentais.gov.br/acesso-aos-sistemas/comprasnet-siasg>. Access on: 10 out. 2016.

4 Available on: <http://www.tesouro.fazenda.gov.br/siafi>. Access on: 10 out. 2016.

5 Available on: <http://cne.smpe.gov.br/>. Access on: 10 out. 2016.

6 Available on: <http://www.bepiducb.com.br/index.html>. Access on: 10 out. 2016.

Figure 5:
Example of a web service operations listing

applications: Applications		Show/Hide	List Operations	Expand Operations
GET	/rest/aplicativos	Returns the set of registered applications		
POST	/rest/aplicativos	Registers a new application		
GET	/rest/aplicativos/pessoa/{codPessoa}	Returns the application set of an accountable agent		
GET	/rest/aplicativos/{codAplicativo}	Retrieves the data of a given application by code		
PUT	/rest/aplicativos/{codAplicativo}	Updates the data of an application already registered		
GET	/rest/aplicativos/{codAplicativo}/hashtags	Retrieves the registered hashtags of a given application		
GET	/rest/aplicativos/{codAplicativo}/tipos-perfil	Retrieves profile types created for a given application		
POST	/rest/aplicativos/{codAplicativo}/tipos-perfil	Registers a new profile type for a given application		
GET	/rest/aplicativos/{codAplicativo}/tipos-perfil/{codTipoPerfil}	Retrieves the data of a certain profile		
PUT	/rest/aplicativos/{codAplicativo}/tipos-perfil/{codTipoPerfil}	Updates the data of a profile type already registered		

Source: Own elaboration

Figure 6:
Example of detailing of the contract of an operation

applications: Applications	Show/Hide	List Operations	Expand Operations
GET /rest/aplicativos	Returns the set of registered applications		

Response Class (Status 200)
Model Model Schema

Application {
cod (long, optional),
descricao (string, optional),
links (Array[Link], optional),
nome (string, optional),
}
Link {
href (string, optional),
rel (string, optional),
templated (boolean, optional),
}

Response Content Type application/json
Registers a new application

Source: Own elaboration

7 Available at: <http://portal.tcu.gov.br/desafio-aplicativos-civicos/>. Access on: 10 out. 2016..

6. REFERENCES

BOOTH, D. et al. (Eds.). W3C. Web Services Architecture. February 2004. Section 1.4. Available at: <https://www.w3.org/TR/ws-arch/#whatis>. Access on: 10 Oct 2016.

BRASIL. Law No. 12,527, of November 18, 2011. Provides for procedures to be observed by the Federal Government, States, Federal District and Municipalities, in order to guarantee access to information, and provides other measures. Diário Oficial da União (Federal Official Bulletin), Brasília, DF, 18 Nov 2011. Extra edition.

ECMA. The JSON Data Interchange Format. Geneva, Oct 2013. Available at: <http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf>. Access in: 10 Oct 2016.

ERL, T. et al. SOA with REST: Principles, Patterns & Constraints for Building Enterprise Solutions with REST. New Jersey: Prentice Hall, 2012. p. 24.

FIELDING, R. T. Architectural styles and the design of network-based software architectures. 2000. Dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Information and Computer Science. University of California, Irvine, 2000.

FIELDING, R. et al. Hypertext Transfer Protocol – HTTP/1.1. June 1999. Available at: <https://www.w3.org/Protocols/rfc2616/rfc2616.txt>. Access in: 10 Oct 2016.

The semantic retrieval of information in the context of external control

Márcia Martins de Araújo Altounian
is an employee of the Federal Court of Accounts – Brazil. She has a B.A. in Library Science and Documentation from the University of Brasília (UnB) and a specialization degree in Knowledge Management from the Blaise Pascal Institute and in Information Architecture and Organization from UFMG.

Beatriz Pinheiro de Melo Gomes
is an employee of the Federal Court of Accounts – Brazil. She has B.A. in in Library Science and Documentation from the University of Brasília (UnB) and a specialization degree in Strategic Knowledge Management and Business Intelligence from the Catholic University in the State of Paraná (PUC).

SUMMARY

In the information retrieval systems (IRS), syntactic base techniques are being supplanted by the increasing exploitation of semantic retrieval techniques, which enable the understanding of the concepts in their context and purpose. Some technologies have contributed to this reality, such as the data semantic markup, used in the semantic web, natural language processing and neural networks. The thesaurus is also a semantic component that affects IRS performance. Thesauruses are artificial language tools in a specific domain, formed by a system of related concepts. This article presents the application of the TCU thesaurus, called External Control Vocabulary, in some corporate information systems.

Keywords: Semantics. Information retrieval. Thesaurus.

1. INTRODUCTION

Information retrieval (IR) is a field of common interest in computer science and information science that is concerned with developing and studying aspects of efficiency and effectiveness of searches in an information system, so that results are coherent with its search expression and, above all, relevant to the user of the system.

Searching in large repositories of unstructured and non-standardized information becomes an arduous task, of-

ten leading to ambiguities, out-of-context content, and even failure to retrieve desired information. The search tools, when using the natural language, need knowledge about the meaning of the expressions used and the relations between them, which allows contextualization and treatment of linguistic phenomena that affect the quality of the retrieval.

For the process of classification and retrieval of information, several automatic methods have been developed in order to obtain relevant answers for any research carried out. To date, most of them are based on syntactic and statistical aspects, taking into account frequency and distribution of words present in documents. Based on the syntactic aspects of the information, the classification performed by these methods is still very distant, in terms of quality, from the classification based on indexation of the subjects of the documents carried out by specialists.

Recently, new methods of automatic document classification have been explored based on semantic aspects of language that refer to the meaning of the terms in their context and purpose. One of these initiatives consists in the use of instruments of representation of semantic and conceptual relationships such as thesauruses, which are characterized by the domain of knowledge and aim to solve problems of ambiguity inherent in everyday words.

The External Control Vocabulary, TCU thesaurus, was developed with the objective of standardizing the treatment of specialized contents and contributing to semantic recovery in the context of external control, in addition to making searches in the Court's information systems more agile and precise.

2. SEMANTICS

The term "semantics", traditionally studied in linguistics and philosophy, has been widely used in a wide range of fields, especially in information technology. The so-called semantic web (SW), which originates in the expansion of the web and in the limitations of the search tools based on syntax, is certainly one of the reasons for this association.

But what is semantics? The word originates in the Greek term semantiké and many meanings can be found according to the perspective used, since there are, among others, textual, cognitive, lexical, formal and argumentative semantics. In general, all converge on the same point: they study meaning, or significance. The wide range of possibilities testifies that the study of meaning can be done from various angles.

Three basic properties stand out in semantics. They are synonymy, antonyms, and polysemy. Synonymy is the division of semantics that studies the relation between linguistic expressions that have the same meaning. For example, "girl" and "lassie" are nouns that have the same meaning, referring to the figure of a young woman; so it is with verbs like "renounce," "refuse," and "reject," which convey the idea of revulsion. Except for the rare occurrence of perfect synonymy, one can be affirm that synonymy is the relation between words and expressions that have a common meaning and significance.

If on the one hand synonymy studies the words with similar meanings in the language, antonyms deal with the study of words that indicate opposite meanings. In the same line as the previous example, we can cite the nouns "girl" and "lady", as well as the verbs "renounce" and "accept" - all are terms with opposite meanings.

It is also possible for the same word to assume different meanings - in which case, the context in which it is inserted will dictate its meaning. A good example of daily life is the word "orange", from which we can quickly think of two very different senses: a citrus fruit and a reddish-yellow color. This is a common example of polysemy, a term which, formed by the prefix "poly" ("many") and by the suffix "semy" ("meanings"), is the part of semantics that studies the significations that a word assumes in a particular linguistic context.

There are still two other semantic properties, in the same line of study of signification, that deserve to be mentioned: homonymy and paronymy. The homonymy studies the relation of two or more words that, although they have different meanings, have the same form and the same sound - it is the case of terms like ad/add and to/too (preposition/adverb), among many others. Paronymy studies the particularities of words that, on the contrary, have similar spelling and pronunciation, but different meanings - are examples: eminent/imminent and elicit/illicit.

Finally, semantics also studies the properties of denotation and connotation of words. Denotation is the property that a word has to be limited to its own concept - for example, the term "stars" in "the stars of the sky". Connotation is the property that a word has to be expanded in its semantic field, within a context, generating several interpretative possibilities - for example, the same term "stars" in "movie stars".

When we transpose the concepts explored here to what we now know as SW, it is easy to see that we are talking about promoting improvements in the processes of representation and retrieval of information on the web. Since 1990, the web is characterized by the use of markup languages that aim at the presentation and reading by people and search engines based on algorithms with syntax orientation. The use of semantics can increase the possibility of document associations to their meanings by means of descriptive metadata. Therefore, the question of meaning in the semantics is fundamental to SW.

Of the several lines of study related to semantics, the one of formal semantics seems to be the most relevant to information technology. Three main aspects permeate the studies on formal semantics:

1. The principle of compositionality, which states that the meaning of sentences depends on the meaning of the words that compose them - that is, the meaning of the whole is a function of the meaning of the parts and the syntactic combination between them. To know the meaning of a sentence, it is necessary to know the meaning of its parts, as well as the rules that define its combination.

2. The condition of truth, which determines the conditions under which such a sentence is true. In this

context, knowing the meaning of a sentence is equivalent to knowing its truth conditions, which is not the same as knowing its truth value, that is, whether the fact is true or false.

3. The models in semantics, in which simple systems are constructed, in relation to the complex systems that are to be studied. A logical theory is constructed for the model and, if the results are reasonably positive with respect to the complex system, it is said that the simple system is a good model; otherwise it is abandoned.

Formal semantics also considers the fact that natural languages are used to talk about objects, individuals, facts, events and properties, described as external to the language itself - thus, referentiality is one of its fundamental aspects. For this reason, in formal semantics meaning is understood, on the one hand, as a relation to language, and, on the other, to what language speaks about.

Formal semantics seeks to answer the following questions: what do linguistic expressions represent or denote? how do we calculate the meaning of complex expressions from the meanings of their parts?

3. SEMANTIC RECOVERY OF INFORMATION

The Information Retrieval Systems (IRS) seek to represent the content of documents and present them to the user to meet their information needs quickly.

To do so, IR researches techniques for treatment, organization and content search based on the use of standards. IR tools generally work with indexing techniques that can quickly indicate and access documents from a textual database.

There are three main types of indexing:

- traditional indexing, in which the descriptive or characterizing terms of the documents are determined;
- full-text indexing (or indexing of the whole text), in which all the terms that make up the document are part of the index;
- indexing by tags (by parts of the text), in which only some parts of the text are chosen to generate the entries in the index (only those considered more important or more characterizing).

Searches are usually carried out using terms provided by the user or chosen by him/her among those presented.

These terms can mean the subject or class to which the desired documents belong (in traditional indexing) or the terms that must be present in the desired documents (in full-text indexes and by tags indexes).

In conventional search systems, the techniques used are syntactically based. However, when the user search involves information whose relevance cannot be given by keywords, this model is not efficient. Therefore, we note the increasing exploitation of semantic information, which enables the understanding of the concepts in their context and purpose.

Semantic marking of data at source is an example of the new technologies used for IR. SW has used this strategy to implement metadata standards that add meaningful information to the data about their contexts, marking them semantically.

The exploration of the intrinsic semantics of the data seeks the foundations of linguistics and information science to expand the universe of information retrieved and the measurement of contexts, through the use of natural language structures, such as verbal and nominal phrases, and tools for representing semantic and conceptual relationships.

There is increasing demand for Natural Language Processing (NLP) techniques that allow the construction of algorithms to search for relevant information in a large number of documents. The foundation for efficient NLP methods should be basic knowledge about language properties and especially about the semantics of concepts. This context arises from the idea of semantic memory, a theme that has been the object of psycholinguistic theories and is a rich source for the development of computational

models that intend to approach the mental processes used by the human brain to understand language. In addition to algorithms, computational models require data representing knowledge about language and common sense associations between lexical concepts and their properties. Currently, this is very difficult because this task can only be produced and verified by humans. Semantic memory works with a mental lexicon - that is, concepts and units of knowledge - and contains information about the relations between concepts, forming a conceptual network of elements connected to each other by different types of associations.

Neural networks, in turn, are a representation that have many characteristics common to human memory: they can handle incomplete or distorted information, allow automatic generalizations and display content based on context. These functions allow several applications in the IR stored in human memory, essential in precise contexts. These applications are we intend to reproduce in the computational environment.

A semantic component that influences the performance of an IRS is the thesaurus, an artificial language tool of a known domain, built by experts to represent the informational content through concepts, specifying the relations between them. It is a system of concepts that relate to each other and are represented by terms.

Each term must be linked to another, and it is this link that forms the structure of the thesaurus. Terms are used by indexers at the time of indexing and should be made available to the user at the time of IR.

Silveira and Ribeiro-Neto (2004) study the automatic use of thesaurus to improve web search results, through a ranking based on concepts that have been studied for IR in specific domains. In an experiment conducted by the authors, the search terms used in an IRS on the web were compared to the concepts of a thesaurus, used to find related concepts. Each related concept was interpreted independently and processed separately, and then combined into a Bayesian network[1], to allow a final, concept-based ranking. The aim was to see if increased concept information would increase the average accuracy of the search results.

Six sources of information were used in the study: key word, concept, specific term, general term, related term and synonym. The authors verified, among other things, that the use of a specific thesaurus for a specific domain is fundamental to improving search performance.

The experiment demonstrated a 30% increase in the average accuracy of the search results. Thus, it proposes the view that thesauruses improve both recall and accuracy in an IRS.

4. THESAURUS

According to Moreira, Alvarenga and Oliveira (2004), the term "thesaurus" originates from the Greek *thesaurós* and means treasure or repository. This term came about with the publication of the analogical dictionary of Peter Mark Roget, London, in 1852, Thesaurus of English words and phrases. The term also designates vocabulary, dictionary or lexicon, but Roget's dictionary

differs from others because it is a vocabulary organized according to its meaning, not in alphabetical order. The work had the merit of establishing the denomination for vocabularies that relate their terms through some kind of relation of meaning.

In his introduction, Roget defines his dictionary as a "classification of ideas" and explains that, unlike the others, his allows one to arrive at the most adequate word or the one that best fits the needs of the writer without him/her knowing, at first, what it is (GOMES, 1996).

In the 1960s, information scientist Brian Campbell Vickery presented four meanings for the term "thesaurus" in the literature of his area, the most common meaning being that of an alphabetical list of words in which each word is followed by others related to it (VICKERY, 1980 apud MOREIRA; ALVARENGA; OLIVEIRA, 2004).

Currás (1995, p.88) defines thesaurus as "a specialized, normalized, post-coordinated language, used for documentary purposes, where the linguistic elements that compose it - simple or compound terms - are syntactically and semantically related to each other. "

Tristan (2004, p. 167) defines it as "a vocabulary of terms, which is nothing more than a selection of terms, based on concept analysis, in which is defined the general greater scope term and its relation to more specific terms, which represent the minor concepts". The National Organization of Information Standards specifies:

> A thesaurus is a controlled vocabulary organized in a pre-established order and structured so that the relationships of equivalence, homography, hierarchy, and association between terms are clearly indicated and identified by standardized relationship indicators used reciprocally. The primary purposes of a thesaurus are (a) to facilitate retrieval of documents and (b) to achieve consistency in indexing of written or otherwise recorded documents and other types, particularly post-coordinated information storage and retrieval systems. (ANSI / NISO Z39.19, 2003 apud SALES; CAFÉ, 2008).

According to Campos and Gomes (2006), the evolution in the construction of thesauruses is based on two perspectives, the American and the European. The thesaurus elaborated in the United States from the 1950s onwards was the result of the development that took place from the heading of subjects to the uniterm, moving from a pre-coordinated system to post-coordinated systems.

Silva (2008) states that at the same time, in England, the Classification Research Group (CRG) – based on the Faceted Classification Theory developed by the Indian mathematician and librarian Shiyali Ramamrita Ranganathan - expanded the categories of personality, matter, energy, space and time (PMEST) and developed several classification tables. This gave rise to a technique called Thesaurofacet, which allowed better posi-

tioning of the concept in the system of concepts in a given subject area, through the use of its categories. In addition, the terminological thesauruses are also based on Faceted Classification Theory, Concept Theory and some terminological principles. These instruments find in the characteristics of the concept an essential element to show the relations between the concepts and their positioning in the system, besides defining it.

The aforementioned Concept Theory, focused on the referent and originally called Analytical Theory of Concept, was launched in the late 1970s by information scientist Ingetraut Dahlberg, adding terminological principles related to the conceptual content and its definition. For Campos and Gomes (2006), this is a consolidated theory to determine what would be understood as the smaller unit in a thesaurus: the concept represented by a term. In addition, Moreira (2003) points out as innovation the use of definitions to position the concept in the system.

Bräscher (2010) points out as a function of the thesaurus the translation of the language of documents, indexers and researchers into a controlled language, used in indexing and in IR in information systems. According to Sales and Café (2008), the ANSI / NISO Z39.19, 2003, emphasizes that thesauruses are not only used by information specialists at the time of indexing, but also by information users when searching for documents.

According to Café and Bräscher (2011), in thesauruses the "semantic relations are established through the analysis of the characteristics or properties of the concepts, which allow identification of differences and similarities that evidence certain types of relationships". A term present in a thesaurus can be characterized in different ways depending on the subject in question and also the type of system one wishes to construct. The thesaurus structure comprises three main types of semantic relations to relate the terms: hierarchy, equivalence, and association.

In hierarchical relations the terms are organized into genus/species. The equivalence relations are of synonymy, that is, there are synonymous terms in the thesaurus and it should be indicated which one is the appropriate term to represent a certain concept. The association relations, in turn, present associations between terms, without specifying what kind of relation actually exists - they are just terms that relate in some way.

Thesauruses also deal with cases of ambiguity (the possibility that a linguistic communication is open to more than one interpretation) and polysemy (the possibility that a word admits more than one meaning).

5. EXTERNAL CONTROL VOCABULARY

The TCU thesaurus, called *Vocabulário de Controle Externo* - VCE (External Control Vocabulary - VCE), was launched in 2015 and aims to be a terminological control instrument that allows standardization of technical treatment and greater accuracy in the retrieval of the contents in the TCU information systems.

The interrelationship of the concepts in the VCE was expressed through relations of three types: equivalence, hierarchical and associative. The relations aim to present the descriptors in their semantic context.

- **Equivalence relation:** if it is considered synonymous or almost synonymous, , represent

the same concept for vocabulary purposes, one of them is chosen as descriptor and the others are prohibited.

- **Hierarchical relation**: relationship that expresses degrees or levels of superordination and subordination of terms; the superordinate term represents the genus of which the subordinate term is type or species.

- **Associative relation**: a gathering of related concepts that deserve to be related, but which are not bound by equivalence or hierarchical relationships.

In the VCE, each term corresponds to a concept and all terms have relationships. The relationship is determined by the meaning of the term. The relations between terms help to understand the specific concepts of external control and related areas that make up the thesaurus.

Beyond a hierarchical list of words, the VCE is composed of three distinct but interrelated parts. The first one is formed by key words, related to the fields of performance of the Court. They are accompanied by definitions and synonyms. The second part corresponds to the TCU clientele and the Supreme Audit Institutions (SAIs) associated with the International Organization of Supreme Audit Institutions (Intosai), and includes information such as history, alternative names, CNPJ and similar institutions. The third part considers the national toponymy formed by the regions, mesoregions, federative units and Brazilian municipalities.

6. VCE APPLICATIONS

6.1 E-JURIS

The E-Juris is a corporate tool that is part of the e-TCU, which aggregates all the Court's systems related to issuing opinions and controlling court cases, containing the same logic, structure and presentation of the other court cases systems. In addition, it is integrated with the TCU portal, the corporate search system and other TCU systems such as Sagas and VCE. The premises adopted for the new system were selectivity, quality, relevance, timeliness and simplicity.

The main purpose of the e-Juris is to disseminate the relevant theses, from the point of view of precedents, which based TCU sentences. This is done through periodical publications (bulletins and newsletters) and creation of a database containing the Court precedents and making it available for research and consultations. This way, the new system unifies work processes that were previously done separately and without integration.

The relevant precedents are represented by statements, in the form of a summary. The statements repre-

sent jurisprudential precedents, not the "understanding" or prevailing Court precedent on a given issue.

The adoption of indexing by e-Juris allows greater precision in the retrieval of statements. In addition to predicting the search for synonyms, the VCE adds to the system the suggestion functionality of correlated subjects to be searched, since the tool is structured in a system of interrelated concepts by hierarchy, equivalence and association.

6.2 TCU PORTAL DIGITAL LIBRARY

Corporate repositories of knowledge are used to disseminate the information produced internally, to allow access to the organizational culture and to subsidize the information transformed into knowledge. Since they are more than a depository of documents, repositories can act decisively, supporting the development of new products and services and capacity building and training of the organization's workforce. In addition, they usually serve as an official source of information to partners and collaborators, support daily activities and help the decision-making process.

At TCU, the digital library is one of easiest to use and most accessible corporate repositories of knowledge. Developed to organize, handle and disseminate information that can generate new knowledge, the digital library allows the insertion of material with several documentary typologies. Thus, in the same environment it is possible to find books, academic papers, presentations, booklets, periodicals, agreements, contracts, official correspondence and norms, as well as various images and other types of resources. The tool allows two levels of access to the deposited contents to be established: it is possible to allow free access to documents of a public nature, in the same way that it is possible to limit access to documents of an internal nature or that have some type of restriction.

The insertion of documents into the environment is done in a decentralized way and there are content managers responsible for approving what will be available in the repository. The digital library also has a data entry form that can be used throughout the Court. This form is composed of controlled metadata and allows the description of the content with elements such as title, authorship and date. Moreover, it requires that the documents be classified in a thematic tree and that the subjects treated be translated by keywords derived from a controlled vocabulary.

In other words, TCU digital library is a corporate repository of knowledge that requires information to be classified and indexed. In order to index content, the environment is integrated with the VCE. In addition, since the library is also integrated with the textual search tool of the corporate portal, it is possible to directly search the portal and retrieve the content deposited in the library environment.

6.3 GUIDANCE SYSTEM

The Guidance System was designed as a tool to guide, manage and disseminate knowledge about external

control. It allows any Court employee to forward questions about pre-selected topics of external control, such as auditing, planning, annual accounts, special accounts, representations, complaints, requests, quality evaluation, executive debt collection, external control standards and other process procedures, as well to ask questions about the Fiscalis system

After selecting the theme, the system directs the question to the unit responsible for the area. Based on the questions and answers collected, each respondent unit automatically creates a database of frequently asked questions (FAQ), which is stored in the system and available for queries and searches by all servers.

Several TCU units have already been registered as respondent units and the system is integrated with the VCE. Adoption of the controlled vocabulary allows greater accuracy in the retrieval of the desired information, since index terms that represent the subjects treated are assigned to both questions and answers.

6.4 EXTERNAL CONTROL WIKI

Among the various possibilities of storing the knowledge of an institution, one in particular deserves to be highlighted in an era in which we produce knowledge through various collaborators and partners: wikis - a technology tool known as "social software" and designed with a set of characteristics which allow the creation and organization of knowledge in the collaborative world. The use of wikis has proved to be a low-cost solution with a high degree of efficiency, to foster the cooperative creation of knowledge within organizations.

The wiki environment is the evolution of the concept of computer supported cooperative work (CSCW), which arose from the need for organizations to have people working in different physical locations and, at the same time, needing to achieve quick results together; that is; it has emerged to facilitate the communication and productivity of remote groups.

Since 2009, the TCU uses the free software Mediawiki to manage an external control wiki and restricted access to its employees. It is an important collaborative space of knowledge construction that brings together informal tutorials and specialized entries coming from the VCE.

The Wiki can be accessed and edited by all employees; they aggregate information to entries and tutorials on topics covering norm, law, and doctrine. The Wiki is an important knowledge management tool insofar as it provides information and documents useful to the daily work of the auditors, according to their area of activity. It is configured as a genuine collective intelligence environment of the organization.

6.5 FUTURE APPLICATIONS

The VCE also has potential applications in various information systems of the Court. The integration of terminology into the TCU portal search tool, through the adoption of controlled vocabulary for both treatment and search, is a key factor for increasing precision and speed in IR.

Another possibility of using VCE is as a dictionary in software for automatic indexing of large volumes of documents, serving as a terminological parameter in the area of external control. This functionality has already been tested during the process of automatic indexing of statements, which was part of a conceptual test to acquire data mining and semantic data analysis software.

In an ideal scenario, we envisage that the Court will adopt standards such as subject metadata and the terminological control tool to ensure the improvement of its of IRS performance.

NOTES

1 Bayesian networks constitute a graphical model that simply represents the causal relations of the variables of a system. In summary, Bayesian networks, also known as networks of opinion, causal networks and probabilistic dependency graphs, are graphical models for reasoning (conclusions) based on uncertainty, in which the nodes represent the variables (discrete or continuous), and the arcs represent the direct connection between them (SILVEIRA; RIBEIRO NETO, 2004).

REFERENCES

ALMEIDA, M. B.; SOUZA, R. R. Avaliação do espectro semântico de instrumentos para organização da informação (Evaluation of the semantic spectrum of instruments for the organization of information). Encontros Bibli – Revista Eletrônica de Biblioteconomia e Ciência da Informação (Meetings Bibli - Electronic Journal of Library Science and Information Science), Florianópolis, v. 16, n. 31, p. 25 50, 2011. Available in: <http://mba.eci.ufmg.br/downloads/11963-60907-1-PB.pdf>. Access in: Nov 23 2016.

ALTOUNIAN, M.; ZAULI, A. A semântica na recuperação da informação na web: novas tendências (Semantics in the retrieval of information on the web: new trends). 2013. Work submitted as a partial requirement for approval in the discipline Recuperação da Informação (Information Retrieval), Escola de Ciência da Informação (School of Information Science), Universidade Federal de Minas Gerais (Federal University of Minas Gerais), Belo Horizonte, 2013.

BRÄSCHER, M. Elaboração de tesauros (Elaboration of thesaurus). Brasília, 2010.

CAFÉ, L.; BRÄSCHER, M. Organização do conhecimento: teorias semânticas como base para estudo e representação de conceitos (Organization of knowledge: semantic theories as a basis for study and representation of concepts). Informação & Informação (Information & Information), Londrina, v. 16, n. 3, p. 25 51, Jan./jun. 2011.

CAMPOS, M. L. A.; GOMES, H. E. Metodologia de elaboração de tesauro conceitual: a categorização como princípio norteador (Conceptual thesaurus elaboration methodology: categorization as guiding principle). Perspectivas em Ciência da Informação (Perspectives in Information Science), Belo Horizonte, v. 11, n. 3, p. 348 359, Sept./dec. 2006.

CAPRI, D.; GARRIDO, I.; DUARTE, R. Recuperação semântica da informação (Semantic information Retrieval). 2009. Work submitted as a partial requirement for approval in the discipline Recuperação da Informação (Information Retrieval), Centro de Ciências da Educação (Center of Educational Sciences), Universidade Federal de Santa Catarina (Federal University of Santa Catarina), Florianópolis, 2009. Available in: <http://pt.slideshare.net/doritchka/angel-recuperao-semntica-da-informao>. Access in: 23 Nov. 2016.

CURRÁS, E. Tesauros: linguagens terminológicas (Thesaurus: terminological languages). Brasília, DF: CNPq; Ibict, 1995.

GOMES, H.E. Classificação, tesauro e terminologia; fundamentos comuns (Classification, thesaurus and terminology; common fundamentals). Rio de Janeiro: UNIRIO, 1996.

MOREIRA, A. Tesauros e ontologias: estudo de definições presentes na literatura das áreas das ciências da computação e da informação, utilizando-se o método analítico-sintético (Thesaurus and ontology: a study of the definitions found in the Computer and Information Science Literature, by means of an analytical-synthetic method). Perspectivas em Ciência da Informação (Perspectives in Information Science), Belo Horizonte, v. 8, n. 2, p. 216-226, Jul / dec.2003.

_____; ALVARENGA, L.; OLIVEIRA, A. P. Thesaurus and ontology: a study of the definitions found in the Computer and Information Science Literature, by means of an analytical-synthetic method. Knowledge Organization, v. 31, n. 4, p. 231-244, 2004.

SALES, R., CAFÉ, Lígia. Semelhanças e Diferenças entre Tesauros e Ontologias (Similarities and Differences between Thesaurus and Ontologies). DataGramaZero - Revista de Ciência da Informação (DataGramaZero - Revista de Ciência da Informação), v.9 n.4 Aug 2008.

SILVA, A. Análise das relações semânticas em tesauros jurídicos brasileiros: orientações das normas e aplicação prática (Analysis of semantic relations in Brazilian legal thesaurus: guidelines of the norms and practical application). 2013. Completion of course work Bacharel em Biblioteconomia (Bachelor of Librarianship) – Centro de Ciências da Educação (Center for Educational Sciences), Universidade Federal de Santa Catarina (Federal University of Santa Catarina), Florianópolis, 2013. Available in: <https://repositorio.ufsc.br/bitstream/handle/123456789/103801/TCC_Aline_da_Silva_20131PDFA.pdf?sequence=1>. Access in: 23 Nov. 2016.

SILVA, D., SOUZA, R., ALMEIDA, M.B. Ontologias e vocabulários controlados: comparação de metodologias para construção (Ontologies and controlled vocabularies: comparison of methodologies for construction). Ciência da Informação (Ciência da Informação), v. 37, n. 3, p. 60-75, Sept / dec. 2008.

SILVEIRA, M. L.; RIBEIRO-NETO, B. Concept-based ranking: a case study in the juridical domain. Information Processing & Management, Doha, v. 4, n. 5, p. 791 805, Sept. 2004.

SOUZA, A. et al. Recuperação semântica de objetos de aprendizagem: uma abordagem baseada em tesauros de propósito genérico (Semantic retrieval of learning objects: an approach based on thesaurus of general purpose). In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (BRAZILIAN SYMPOSIUM ON COMPUTER EDUCATION), 19., 2008, Uberlândia. Anais… Uberlândia: SBIE, 2008. p. 603 612.

SOUZA, A.; ROCHA, R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências (Information retrieval systems and web search engines: current outlook and trends). Perspectivas em Ciência da Informação (Perspectives in Information Science), Belo Horizonte, v. 11, n. 2, p. 161 173, May / Aug. 2006. Available in: <http://www.scielo.br/pdf/pci/v11n2/v11n2a02.pdf>. Access in: 23 Nov. 2016.

SZYMANSKI, J.; DUCH, W. Information retrieval with semantic memory model. Cognitive Systems Research, Kalamazoo, v. 14, n. 1, p. 84 100, Apr. 2012. Available in: <http://ethologie.unige.ch/etho5.10/themes/semantic.memory/szymanski.duch.2011.information.retrieval.in.sematic.memory.neural.network.models.pdf>. Access in: 23 Nov. 2016.

Geotechnologies and monitoring of Sustainable Development Goals by Supreme Audit Institutions

Rherman Radicchi Teixeira Vieira
is an employee of the Federal Court of Accounts –Brazil. He works in the Department of External Control – Ports, Water and Railway Infrastructure. He has a B.A. in Agronomy from the University of Brasília (UnB).

Leonardo Pereira Garcia
is an employee of the Federal Court of Accounts – Brazil. He works in the Research and Innovation Center/ISC. He has a B.A. in Administration from the University of Brasília and a specialization degree in Corporate Education.

André Delgado de Souza
is an employee of the Federal Court of Accounts – Brazil. He works in the Department of External Control in the State of Paraíba. He has a B.A., M.A. and PhD in Civil Engineering from the Federal University of the State of Pernambuco and a sandwich PhD from the Berlin Technical University.

Erick Fonseca dos Santos
is an employee of the Federal Court of Accounts – TCU/Brazil. He is working towards an M.A. in applied computing focused on computational vision and recognition of objects in satellite images, associated to the TCU GeoControl project aimed at monitoring infrastructure works through remote sensing. He has a B.A. in Computing.

ABSTRACT

One of the initiatives proposed by the United Nations (UN) so that Sustainable Development Goals (SDGs) can overcome challenges faced by Millennium Development Goals (MDGs) is the use of geospatial data and investments in training to use new technologies. In addition, International Standards of Supreme Audit Institutions (ISSAIs) acknowledge the importance of training its employees to meet new requirements. The proposal of concrete actions in this work aims at supporting the strategy of investing in geotechnologies so that Supreme Audit Institutions (SAIs) can monitor SDGs. This activity was accomplished by reviewing literature on United Nations and Intosai technical references that might contribute to the use of geotechnologies for SDG monitoring by SAIs.

Keywords: Sustainable Development Goals. Supreme Audit Institutions. Geotechnologies. Remote Sensing. Geographic Information System (GIS). *Global Positioning System* (GPS). Intosai. ISSAI. Audit International Standards. Technical Skills. Qualification. Diagnosis.

1. INTRODUCTION

The United Nations Organization (UN) 2030 Agenda for Sustainable Development poses a challenge on activities carried out by Supreme Audit Institutions (SAIs). Recent resolutions of the UN General Assembly emphasize the key role of SAIs and of the International Organization of Supreme Audit Institutions (INTOSAI) in meeting Sustainable Development Goals (SDGs) (UNITED NATIONS, 2016).

UN Resolution A66/209, dated 2011 (Id., 2011), points out that SAIs perform an important role in promoting the efficiency, accountability, effectiveness and transparency of public administration, fostering national development as to SDGs. In addition, Resolution A69/228, dated 2014 (Id., 2014a), strengthens explicitly the key role of SAIs in the 2030 Agenda.

The 23rd United Nations/Intosai Symposium report (Id., 2015d) on the role of SAIs regarding the 2030 Agenda emphasizes the importance of the intensive use of data analytics[1]. One of the initiatives proposed by the UN so that SDGs can overcome challenges faced by Millennium Development Goals (MDGs) is the use of geospatial data (Id., 2015c). In addition, the International Standards of Supreme Audit Institutions (ISSAIs) highlight the possible application of geotechnologies to several audit phases (INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2013b).

However, in bibliographical information sources researched, there are no sources joining in a single system UN technical references on the use of geotechnologies for monitoring SDGs and Intosai technical references on the topic of geotechnologies.

For those reasons and in order to improve the use of new technologies applied to control, this paper aims at reviewing literature on United Nations and Intosai technical references that may contribute to the use of geotechnologies for SAI monitoring of SDGs.

2. GEOTECHNOLOGIES AS SUPPORT TOOLS FOR MONITORING SDGs

Geotechnologies optimize a universal and standardized approach for monitoring other relevant information, including economic, educational, environmental and health indicators. Geographic information allows data modelling and analysis, map creation and the detection and monitoring of its modifications along the time on a consistent and standardized basis (UNITED NATIONS, 2015a).

Souza (2016, p. 40) describes geotechnologies as "a set of technologies for collecting, processing, analyzing and making available georeferenced information". Some of them are the Global Navigation Satellite System (GNSS), Remote Sensing and Geographic Information Systems (GISs). GNSS enabled people to know accurately their location in Earth, and Global Positioning System (GPS) distinguishes from other systems of this type (SOUZA, 2016).

Souza (2016, p. 41) defines remote sensing "as the use of sensors to capture and record the distance, without direct contact, of power reflected or absorbed by the target surface". By using computer software, data generated by such technology can be stored, handled and analyzed. Unmanned aerial vehicles (UAV), small unmanned aircrafts, may monitor areas through air photographs and recordings (SOUZA, 2016).

Thus, GISs are "systems that computer process geographic information" (SOUZA, 2016, p. 41). Câmara (2015, p. 2) explains that "the main difference between a GIS and a traditional information system is its ability to store both descriptive attributes and geometries of different types of geographic information". GISs enable integration of data from various sources, making possible deeper analyses (SOUZA, 2016). It can be entered data from different sources in only one database, and that data can be combined by using analysis and manipulation algorithms (CÂMARA, 2005).

The document "The millennium development goals: report 2015" (UNITED NATIONS, 2015e) summarizes the progresses, challenges and lessons learned along fifteen years monitoring MDGs. When covering monitoring challenges, it can be said that large data gaps, poor data, outdated and non-disaggregated data are some of the main challenges. One of the initiatives proposed to overcome those challenges is the **use of geospatial data** (UNITED NATIONS, 2015e).

In report "A world that counts: mobilizing the data revolution for sustainable development" (INDEPENDENT EXPERT ADVISORY GROUP, 2015), UN calls signatory Member States to 2030 Agenda, Corporations and civil society to coordinate efforts to improve information availability, quality, update and disaggregation, in order to support implementation in all levels.

According to document "Assessing gaps in indicator availability and coverage" (CASSIDY, 2014), one major challenge to be faced is the low general coverage of indicators proposed. Almost one third of indicators lacks data in more than half of the countries (SHUANG et al., 2013) and, on average, only 46% of data were collected, posing an international challenge as for statistical data production (UNITED NATIONS, 2014e).

The report "Indicators and a monitoring framework for the sustainable development goals" (UNITED NATIONS, 2015c) highlights opportunities created by data revolution, via big data, geophysical and social data and new forms of data sharing. In addition, that report presents several examples of geotechnologies supporting SDGs, as shown in Exhibit 1 at the end of this paper.

Authors such as Jeffrey Sachs (2012) point out numerous differences between MDGs and SDGs, as detailed in Table 1.

In document "Data for development: an action plan to finance the data revolution for sustainable development" (OPEN DATA WATCH, 2015, p. 8, translation added), prepared by UN and by Open Data Watch, is it declared that

> The SDGs will depend on more geospatial and earth observations data than the MDGs. Satellite imagery is increasingly available for free at a moderate resolution, and at a cost for high-resolution sources. Satellite products have the potential to be utilized in monitoring more than 23 potential SDG indicators, ranging from measuring global air quality to crop and forest cover, to disaster impacts, and water resources. New satellite imagery is one example of emerging technology that offers significant opportunities for a global water monitoring platform.

There is no question that data collection poses a major challenge. As a consequence, it should be used new technologies and methods available, including those provided by big data and geographic information technologies (UNITED NATIONS, 2015a). Since there is more and more information available, the study of methods using databases has distinguished itself (FREITAS; DACORSO, 2014).

In view of comprehensiveness and complexity of 2030 Agenda, it is required several types of data with different cover levels (SUSTAINABLE DEVELOPMENT SOLUTIONS NETWORK, 2015). Thus, each type of data upholds and supports the other types.

For the purpose of monitoring SDGs, it is required data ecosystem fostering and composition. For that reason, UN specified the data typology in report "Data for development: A Needs Assessment for SDG Monitoring and Statistical Capacity Development" (SDSN, 2015), making clear what are primary data sources and setting key principles for selecting robust monitoring indicators.

Table 1:
Differences between Millennium Development Goals (MDGs) and Sustainable Development Goals (SDGs)

	MDG	SDG
Period	2000-2015	2016-2030
Intermediate control points (SACHS, 2012)	Absent	Present
Monitoring Depth (INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2015)	8 goals, 20 targets	17 goals, 196 targets
Geographic scope (SACHS, 2012)	Developing countries	All signatory Member States
Monitoring Forms	Not defined previously	Agreed by governments; monitoring system development
Technology role (SACHS, 2012)	Secondary	Primary
Focus (UNITED NATIONS, 2014e)	Poverty Alleviation	Sustainable Development in a broad sense

Source: Authors' Making

Data typology refers to census data, vital records and vital statistics for births, household surveys, agricultural surveys, administrative data, economic statistics and environmental data, including geospatial data. The latter are critical to determine SDG environmental indicators, and to conduct a disaggregated analysis of their social and economic indicators (SDSN, 2015).

Out of primary sources of indicators proposed by such report (SDSN, 2015), 11% are related to environmental data, including geospatial data. Table 2 shows primary data sources for those indicators.

Another data ecology component refers to key principles for selecting global monitoring indicators, as proposed in report "Follow-up and Review of the SDGs: fulfilling our commitments" (UNITED NATIONS, 2015b). According to those principles, indicators need to be: limited in number and globally harmonized (for global monitoring indicators), simple, i.e. single-variable indicators with straightforward policy implications, high-frequency, allowing regular monitoring, preferably on an annual basis, consensus-based, in line with international standards and system-based information, constructed from well-established data sources, disaggregated to the greatest extent possible, universal, mainly outcome-focused, science-based and forward-looking, a good proxy for broader issues or conditions.

Tools and methodologies for data collection, analysis and communication are an integral part of data ecology. Those tools and methodologies are inherently connected with geospatial fields, including photogrammetry, cartography, geographic information systems (GISs) and geospatial analysis, remote sensing and geospatial intelligence.

Table 2:
Primary data sources for SDG indicators

Primary data sources	
Administrative data	33%
Household surveys	26%
International information	13%
Environmental data (agricultural surveys or geospatial data)	**11%**
Vital Records and vital statistics for births	8%
TBD	6%
Workforce surveys	2%
Other economic data	2%
Census	Cross-cutting
TOTAL	100%

Source: Sustainable Development Solutions Network (2015, translation and emphasis added)

In those fields, the following are particularly relevant to SDG monitoring by SAIs: cartography[2]; geographic information systems[3] and remote sensing[4]. Those fields are referred to hereunder as geotechnologies, i.e.: "the set of technologies to collecting, processing, analyzing and making available georeferenced information" (SOUZA, 2016, p. 40).

At event "Unleashing the power of 'Where' to make the world a better place: How geographic information contributes to achieving the SDGs" (UNITED NATIONS, 2015f), Lawrence Friedl, representing National Aeronautics and Space Administration (NASA), said that SDGs arrive at a prime convergence moment for seizing the power of spatial data. Geographic information is a key element in the complex context of implementation and monitoring of 2030 Agenda (Id., 2015a), being mentioned in target number 18 of SDG number 17, that seeks the strengthening of means of implementation and the revitalization of the global partnership for sustainable development as one of relevant data sources. The "target 17.18" seeks to enhance capacity building in developing countries by 2020, to increase significantly the availability of high-quality, timely and reliable data disaggregated, for instance, by geographic location (Id., 2015c).

By reviewing national indicators proposed by United Nations System in Brazil (UNITED NATIONS DEVELOPMENT PROGRAMME, 2015), it can be noticed that a significant amount of those indicators required geographic disaggregation. Most indicators proposed use some form of geographic disaggregation and, pursuant to report "Follow-up of 2030 Agenda for Sustainable Development", the data disaggregation present in the new agenda poses a challenge to be overcome.

3. INTOSAI TECHNICAL REFERENCES ON THE TOPIC OF GEOTECHNOLOGIES

Document "ISSAI 5130 – Sustainable development: the role of Supreme Audit Institutions" (INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2004), covering new assignments arisen from SDGs, acknowledges that institutions need to question whether existing methodologies for conducting audits are appropriate for such context. It also affirms that, depending on particularities of each SAI,

experts may be included in the process, whether as hired employees or advisors.

"Auditing forests: guidance for Supreme Audit Institutions" ((INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2010), prepared by Intosai Working Group on Environmental Audit (WGEA), describes how SAIs may use the technology of Geographic Information System (GIS). Covering methodological aspects, the document said that "Computer-based technologies can be exceptionally useful in audits. Two examples of those technologies include GPS and GIS" ((INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2010, p. 9). Geographic information may be used for many purposes and in different audit phases. GIS may be used in planning phase and GPS may be used in execution phase as a support tool. Geographic information may be used for many purposes, including different planning phases, street network based applications, natural resource based applications, viewshed analysis, facilities management, and others (INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2010).

According to document "Environmental data: resources and options for Supreme Audit Institutions" (INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2013a), both audit governmental managers and SAIs will benefit from GIS use. Said document describes geospatial data on an independent basis, in a special section, because, in its author's view, data provide SAIs with unique considerations. Due to governmental budget restrictions, governmental managers feel pressured to show results, leading them to use environmental data on a larger scale to demonstrate that their programs met goals set. That type of change may influence how performance is measured and how governmental managers and SAIs evaluate those programs (INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2013a).

Said document declares that spatial data sources may be exceptionally useful to SAIs when those are checking environmental issues with clear geographic aspects, such as environmental protection areas or polluted area location. Geospatial data is also useful to select samples from different sites, finding high risk areas and standards in data, what would not be possible without its spatial component. SAIs may also use spatial data to present its results, making them more tangible.

At last, in "Environmental Data: Resources and Options for Supreme Audit Institutions" (INTOSAI, 2013a), it is said that the use of data tied to geographic locations makes databases much more complex because of the need to record both what is happening and where, using geographic coordinates. The result is a greater demand on quality control. The challenge for SAIs of assessing the quality of the

database is also greater. Finally, SAIs that are considering using spatial data sources need access to the tools and also have to invest in the development of technical skills required.

In report "The 7th survey on environmental auditing" (Id., 2012), WGEA presents the result of survey taken by over 112 SAIs. The survey asked SAIs to describe innovating work methodologies that they have been applying to environmental audits. Geospatial technology was the most voted item[5].

ISSAI 5540 (INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2013b), that covers issues on the use of geospatial data in environmental disaster management audits, says that GISs and remote sensing may provide added value to all audit stages, as shown in Table 3[6]. According to the document, GISs may help in extensive and complex data analytics in many ways: present data spatially, filter data by spatial location, analyze spatial location, store and view data in tiers etc. GISs enable users to produce high quality maps in any scale, store large amounts of geographic information, view complex data and produce new data based on existing data.

ISSAI 5540 proposes a checklist for the use of geospatial data in audit, fully shown in Table 4.

Said ISSAI also says that the use of geospatial information and GIS in public sector has increased for many reasons. One of the main ones is the extent and complexity of information to be considered and assessed when making decisions. Numerous decisions require geospatial information and GISs support the assessment of that type of information. In the context of public policies, geospatial information has many

Table 3:
Geotechnology application to several audit phases pursuant to *ISSAI 5540*

AUDIT STAGE	GEOTECHNOLOGY APPLICATION
Evaluating relevant risks	GISs enable the analysis of comprehensive data or data tiers in a geographic context. Remote sensing may be used to check information in databases using field information
Planning audits	GISs and remote sensing may help to decide the audit focus at this point
Conducting audits	The team may use GPS devices and satellite-based maps to connect field audit data to geographic data. This field data can be assessed right after entering and combining data with maps
Assessing audit results	GISs enable the assessment of different geographic information tiers, allowing performance to be measured. In addition, the viewing of results through GISs allows to find geographic differences in governmental organizations performance
Disclosing results	Using GISs and remote sensing, audit results may be mapped and presented to support the main audit conclusions and recommendations and to make easier the disclosure of results. The viewing of audit results in maps makes the message clearer and stronger than a message only in writing

Source: Adapted from International Organisation of Supreme Audit Institutions (2013b, translation added)

Table 4:
Checklist for the use of geospatial data in audit

***Checklist*: use of geospatial data in audit**
What geospatial data is needed to answer the audit questions?
What accuracy is required of the geospatial data?
What is the required timeframe of the geospatial data?
What geospatial data is available?
From which sources can the required geospatial data been derived from and how reliable are they?
What is the quality of the available geospatial data?
What are the costs of the available geospatial data?
If the required geospatial data are not available, could they be gathered as part of the audit process and budge?
Do the auditors involved have the required knowledge to gather and analyse the required geospatial data or should external expertise be insourced?

Source: International Organisation of Supreme Audit Institutions (2013b, translation added)

purposes, such as goal setting, establishment of measures, monitoring and evaluation. In addition, geospatial information can be used in many public policy areas. Some examples are management of natural resources, environmental protection, economy, education, safety and health (INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2013b).

4. FINAL CONSIDERATIONS

One of the initiatives proposed by UN so that SDGs can overcome challenges faced by Millennium Development Goals is the use of geospatial data (UNITED NATIONS, 2015e). In addition, ISSAI (INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS, 2004) International Standards acknowledge the importance of its employees' qualification to meet new requirements.

This paper aimed at reviewing literature on United Nations and Intosai technical references that may contribute to the use of geotechnologies for monitoring SDGs by Supreme Audit Institutions. We hope this systematization assists further studies to be developed in the scope of SAIs on this subject-matter.

NOTES

1 For instance, satellite imaging application or the use of remote sensing for control actions.

2 Science of designing, producing, disclosing and studying maps as tangible and digital objects.

3 Any system that captures, stores, manages and views location data.

4 Science of obtaining measurement information about an object or phenomenon through a sensor without making physical contact with the studied object/ phenomenon.

5 The use of external expertise had the same vote counting, sharing the first place as an innovating methodology, according to that survey.

6 It should be noticed similarities among audit steps and some functions described in DACUM Research Chart for Geospatial Analyst (*capture data, manage data, analyse data, produce deliverables*). Available at: <http://bit.ly/2gqLhvT>. Web: Jul 12. 2016.

REFERENCES

CÂMARA, G. Representação computacional de dados geográficos. In: CASANOVA, M. et al. Bancos de Dados geográficos. Curitiba: MundoGEO, 2005. p. 1-44.

CASSIDY, M. Assessing gaps in indicator availability and coverage. 2014. Available at: <http://bit.ly/2glULel>. Web: Jul 8, 2016.

Exhibit 1:
Examples of geotechnology support to Sustainable Development Goals (United Nations, 2015)

SDG		INDICATOR		INDICATOR DESCRIPTION	DISAGGREGATION	SOURCE[1]
1	Eradicate poverty in all its forms everywhere	6	Losses from natural disasters by climate and non-climate-related events (in US$ and lives lost)	Measures human and economic losses in rural and urban areas due to natural disasters, disaggregated by climate and non-climate-related events	This indicator may be disaggregated spatially (including urban and rural segregation)	Non-geospatial sources
2	End hunger, achieve food security and improve nutrition, and promote sustainable agriculture	13	Gap in income of cultures	Tracks development gaps in the main cultures, i.e., current income against expected income in optimum conditions	Appropriate to spatial disaggregation, global to local scales	Geospatial data, including remote sensing and satellite
9	Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation	58	Access to clear roads throughout the year (% access to roads within [x] km distance to road)	Access to clear roads throughout the year is critical to rural development processes, including access to inputs, markets, education and health services. This indicator tracks the share of population living within [x] km distance to clear roads throughout the year	This indicator may be disaggregated spatially	Geospatial data, including remote sensing and satellite
11	Make cities and human settlements inclusive, safe, resilient and sustainable	69	Mean urban air pollution in particulate matter (MP10 and MP2.5)	Follows up mean urban air pollution	Per city and state	Geospatial data, including remote sensing and satellite
12	Ensure sustainable consumption and production patterns	75	Aerosol Optical Depth	Measures total aerosols (e.g.: sea salt, dust and smoke particles) distributed within a column of air from Earth surface to the top of the atmosphere	This indicator may be reported with high level spatial disaggregation (including cities and neighborhoods)	Geospatial data, including remote sensing and satellite
15	Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss	83	Annual change in forest area and land under cultivation (modified MDG Indicator)	Tracks the net change of forest area and the expansion of agriculture into natural ecosystems, as well as the loss of productive agricultural land to the growth of urban areas, industry, roads, and other uses	This indicator can be disaggregated spatially	Geospatial data, including remote sensing and satellite
		85	Annual change in desertification in land degradation (% or ha)	Components of land degradation include salinization, erosion, loss of soil nutrients, and sand dune encroachment	Geographic disaggregation by sub-region	Geospatial data, including remote sensing and satellite

Source: Adapted from United Nations (2015c, translation added)

FREITAS, R. K. V. de; DACORSO, A. L. R. Inovação aberta na gestão pública: análise do plano de ação brasileiro para a Open Government Partnership. Revista de Administração Pública, Rio de Janeiro, v. 48, n. 4, p. 869-888, 2014.

INDEPENDENT EXPERT ADVISORY GROUP. A world that counts: mobilising the data revolution for sustainable development UN report. 2015. 30 p. Available at: <http://bit.ly/1DcbDol>. Web: Jun 8, 2016.

INTERNATIONAL ORGANISATION OF SUPREME AUDIT INSTITUTIONS. Auditing forests: guidance for supreme audit institutions. 2010. 71 p. Available at: <http://bit.ly/2fJOrK5>. Web: Jul 12, 2016.

_____. Environmental data: resources and options for supreme audit institutions. 2013a. 64 p. Available at: <http://bit.ly/2gEOStn>. Web: Jul 12, 2016.

_____. ISSAI 5130. Exposure draft. Sustainable on development: the role of supreme audit institutions. 2004. 53 p. Available at: <http://bit.ly/2fcy2Br>. Web: Jul 12, 2016.

_____. ISSAI 5540. Use of geospatial information in auditing disaster management and disaster-related aid. 2013b. 36 p. Available at: <http://bit.ly/2glYicn>. Web: Jul 14, 2016.

_____. The 7th survey on environmental auditing. 2012. 77 p. Available at: <http://bit.ly/2fvUVOT>. Web: Jul 12, 2016.

_____. The role of sais and means of implementation for sustainable development (lessons learned from MDGs). 2015. Available at: <http://bit.ly/2gkLcc0>. Web: Jul 9, 2016.

UNITED NATIONS. Ambassador statement: unleash ing the power of "Where" to make the world a better place: How geographic information contributes to achieving the SDGs. 2015a. 3 p. Available at: <http://bit.ly/2fJOZ2z>. Web: Jul 5, 2016.

_____. Assessing gaps in indicator availability and coverage. 2014. 9 p. Available at: <http://bit.ly/2glULel> Web: Jun 5, 2016.

_____. Follow-up and Review of the SDGs: fulfilling our commitments. 2015b. 39 p. Available at: <http://bit.ly/2fvQkMQ>. Web: Jul 9, 2016.

_____. Indicators and a monitoring framework for the sustainable development goals: launching a data revolution for the SDGs. 2015c. 225 p. Available at: <http://bit.ly/1DMsAfp>. Web: Jul 9, 2016.

_____. Resolution A/66/209. Promoting the efficiency, accountability, effectiveness and transparency of public administration by strengthening supreme audit institutions. 2011. Available at: <http://bit.ly/2fctLy9> Web: Jul 5, 2016.

_____. Resolution A/69/228. Promoting and fostering the efficiency, accountability, effectiveness and transparency of public administration by strengthening supreme audit institutions, 2014a. 3 p. Available at: <http://bit.ly/2gETHmm>. Web: Jun 5, 2016.

_____. 23e Symposium ONU/INTOSAI. 2015d. 8 p. Available at: <http://bit.ly/2fJRstI>. Web: Jun 5, 2016.

_____. The millennium development goals report 2015. 2015e. 73 p. Available at: <http://bit.ly/1Rg5uUm>. Web: Jul 5, 2016.

_____. Unleashing the power of "Where": how geographic information contributes to achieving the SDGs. 2015f. 24 p. Available at: <http://bit.ly/2fcsLtH>. Web: Jul 9, 2016.

_____. NoTheme I: Sustainable Development Goals. How INTOSAI can contribute to the UN post 2015 agenda including good governance in order to strengthen the fight against corruption? Abu Dhabi: Incosai, 2016.

OPEN DATA WATCH. Data for development: an action plan to finance the data revolution for sustainable development. 2015. 81 p. Available at: <http://bit.ly/2fvQaVD>. Web: Jul 9, 2016.

PROGRAMA DAS NAÇÕES UNIDAS PARA O DESENVOLVIMENTO. Acompanhando a agenda 2030 para o desenvolvimento sustentável [UNITED NATIONS DEVELOPMENT PROGRAMME. Follow-up of 2030 Agenda for Sustainable Development]: subsídios iniciais do Sistema das Nações Unidas no Brasil sobre a identificação de indicadores nacionais referentes aos Objetivos de Desenvolvimento Sustentável. Brasília: PNUD, 2015. 250 p

SACHS, J. D. From millennium development goals to sustainable development goals. The Lancet, London, v. 379, p. 2206-2211, 2012.

SUSTAINABLE DEVELOPMENT SOLUTIONS NETWORK. Data for development: a needs assessment for SDG monitoring and statistical capacity development. 2015. 81 p. Available at: <http://bit.ly/1zeUVcf>. Web: Jul 9 2016.

SHUANG, C. et al. Towards a Post-2015 Framework that Counts: Development National Statistical Capacity. 2013. Available at: <http://bit.ly/2g8V35c>. Web: Jul 9, 2016.

SOUZA, A. D. Modelo de controle para obras de esgotamento sanitário utilizando sistema de information geográficas. 2016. Tese (Doutorado em Engenharia Civil) [Doctoral Thesis in Civil Engineering] – Universidade Federal de Pernambuco, Recife, 2016.

InfoSAS: a data mining system for production control of SUS [Brazilian public healthcare system]

Osvaldo Carvalho
has a PhD in Computer Science from the Pierre et Marie Curie University, France, and is an associate professor at the Federal University of the State of Minas Gerais (UFMG). He worked with distributed algorithms and currently teaches Computer Programing and develops information systems.

Marcos Prates
has a B.A. in Computational Mathematics from the Federal University of the State of Minas Gerais (UFMG), an M.A. in Statistics from the same institution and a PhD in Statistics from the University of Connecticut. He develops statistic methods and algorithms for the analysis of spatial statistics and machine learning.

Raquel Minardi
has a B.A. in Computer Science and a PhD from the Federal University of the State of Minas Gerais (UFMG). She also has a postdoc from the French Alternative Energies and Atomic Energy Commission (CEA). She develops models and algorithms in data visualization and in computational biology. She has published over 20 articles in international journals and in national and international conferences.

Wagner Meira Jr.
has a PhD from the University of Rochester and is a Computer Science professor at the Federal University of the State of Minas Gerais (UFMG). His areas of interest are data mining, parallel and distributed systems and their applications.

Renato M. Assunção
has a PhD in Statistics from the University of Washington and is a professor in the Computer Science Departament of UFMG. He is also a specialist in algorithms development and methods for the analysis of statistic data, especially those with georeferencing.

José Nagib Cotrim Árabe
has a PhD in Computer Science from the University of California, Los Angeles (UCLA), U.S.A., and is a professor at the Federal University of the State of Minas Gerais (UFMG). He is head of the Computer Science Department.

ABSTRACT

This paper introduces InfoSAS, a system for detection of statistical anomalies in SUS production records. InfoSAS finds *per capita* inhabitant service rates that are much higher than the Brazilian average or hospitalization prices way over those charged by most institutions for the same procedure. Results show that hundreds of millions of Brazilian Reais spent by SUS are destined for treatments considered anomalous based on conservative criteria. Statistical anomalies may result from fraud, but also from healthcare intensive programs, epidemics, or poor distribution of healthcare services. In any case, serious anomalies should be investigated or explained.

Keywords: SUS, Brazilian public healthcare system, Anomaly detection, Data mining, Service rate per inhabitant, Hospitalization average price, Log-normal distribution, Audit prioritization.

1. INTRODUCTION

The large size of the healthcare sector and its huge amounts of funds turn healthcare into an attractive target for fraud in the entire world. In the United States, for instance, over 270 billion dollars are lost annually due to fraud (THE ECONOMIST, 2014). There is no reason to believe that the scenario in Brazil is different. In this paper, we introduce InfoSAS, a system for mining and viewing SUS information that, like many other systems (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), can find varied data models and standards and could be very useful for SUS management. In particular, for identifying statistical anomalies that might indicate deviations. This paper focuses on its use for audit purposes.

InfoSAS can find many facts requiring managers' attention, including service rates per inhabitant much higher than Brazilian average levels, as shown in Figure 2; or hospitalization prices way over those charged by most institutions, as shown in Figure 4. As seen in Section 5, we have reason to believe that using such system in audit prioritization may produce reimbursements to SUS of BRL400 million per year. It is no exaggeration to estimate that the amount recovered could reach billions of Brazilian Reals as a result of future audits of production data for a period of 8 years (from 2009 through 2016), as well as from reducing anomalous behaviors in future production.

2. STATISTICAL DISCREPANCIES

Services delivered by SUS (jointly called SUS production) are recorded by means of several instruments and those records are entered into the SIA[1] and SIH[2] databases. SUS also keeps the CNES[3] database, a record of healthcare institutions. InfoSAS searches those SUS databases as well as population information provided by IBGE [Brazilian Institute of Geography and Statistics], and produces what we call factsheets.

These databases are massive and those factsheets provide focus when examining them. They match a mining target, an examination period and an institution, showing charts and tables found when examining those databases. A mining target is a SUS production subset, defined by one or more procedures listed in the SUS Table[4] and, sometimes, by patient age group as well. It is worth noting that facts from those sheets can also be verified on an independent basis by checking other sources of information provided by SUS, such as TabWin[5].

Figure 1 shows an example of a factsheet chart. In the chart, we can see that, for the target "Treatment of Visual Apparatus Diseases", all services in Curitiba from September 2013 through August 2014 met expectations based on the Brazilian average, that the numbers of those services are higher than the average of the State of Paraná, and that their provider contributed to less than half of those services. This information is consistent with a regular scenario.

However, that is not always the case. Some facts call the attention of audit professionals. In Figure 2, some cities have service rates per inhabitant much higher than Brazilian and State rates. The institution is practically the only one servicing those cities, a clear

Figure 1:
Examination of the participation of an institution in services in Curitiba for the target "Treatment of Visual Apparatus Diseases"

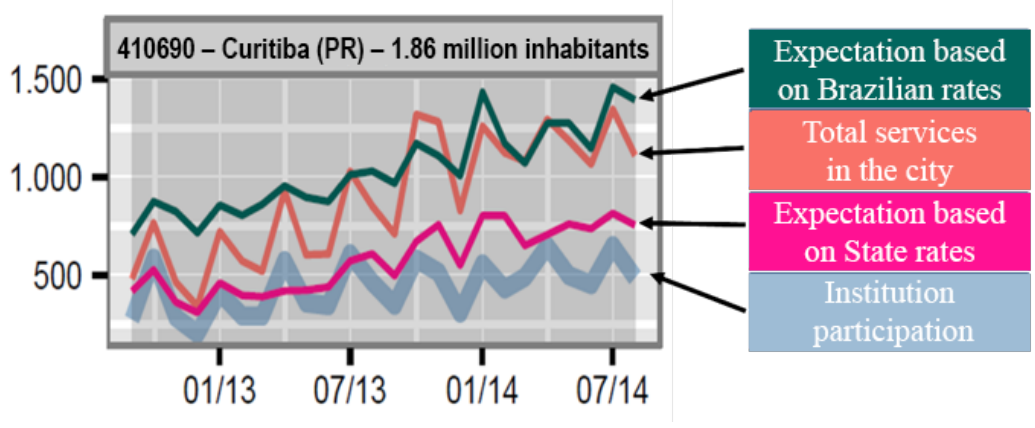


Figure 2:
Analysis of service rates per inhabitant for six cities serviced by an institution for the target "Treatment of Visual Apparatus Diseases"

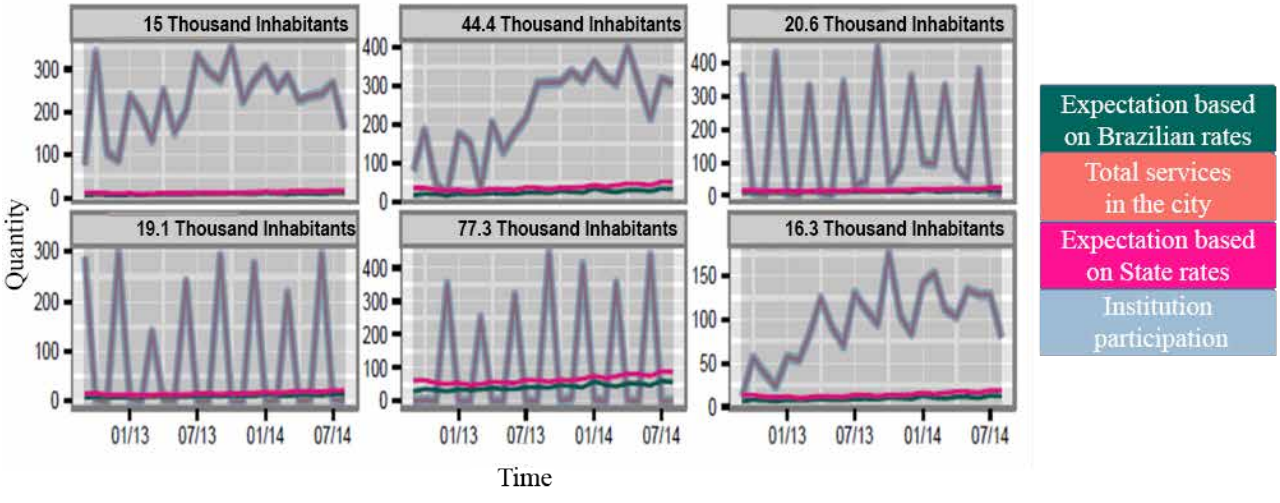
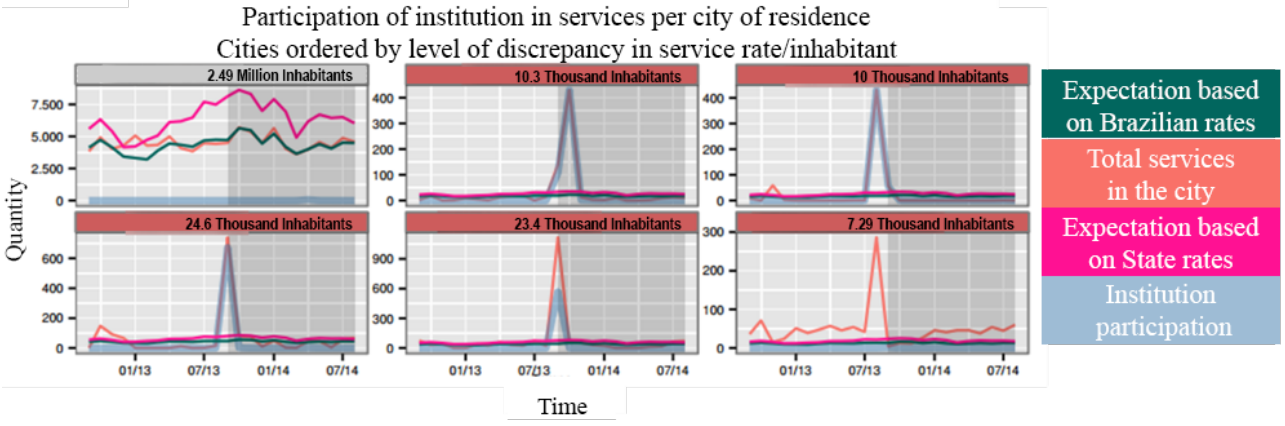


Figure 3:
Statistical discrepancies explained by target characteristics, "Mammogram bilateral screening", and by provider, a Women's Mobile Medical Unit

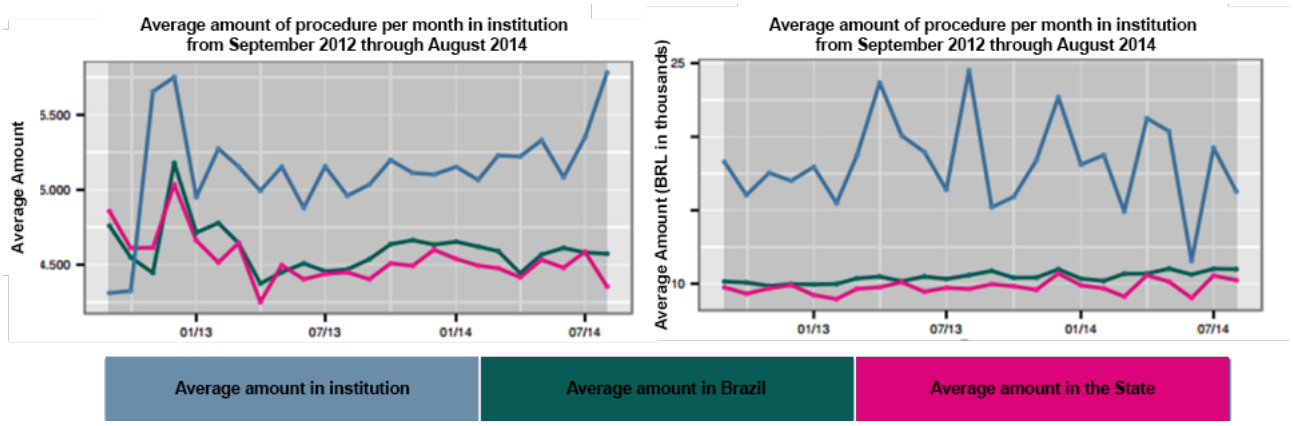


example of what we consider a statistical anomaly or discrepancy.

Statistical discrepancies should be considered cautiously, as shown in the example in Figure 3. Each of the six charts refers to one city and all of them refer to only one provider. We note that only the top left chart (which refers to the city where the relevant institution is headquartered) does not have unusual peaks. It has amounts much higher than expected for Brazilian rates and for the State where those cities are located. Is that fraud evidence? No. Since the target is "Mammogram" and the provider is a mobile medical unit, the statistical anomaly is explained as follows: the provider is a bus equipped with a mammography machine that performs intensive healthcare programs in small towns, leading to service peaks when visiting those towns.

It is worth pointing out that, in general, a statistical anomaly can actually be fraud evidence, but it may also result from correct actions, as shown above, or from incorrect information, or epidemics. It is also possible that cities with high service rates are actually those few ones that do provide good services in the country. Thus, good judgment should be exercised when analyzing a factsheet, which should be done by a public healthcare professional aware of the local context.

Figure 4:
Analysis of amount charged by one institution for hospitalizations for targets "Cementless/hybrid primary total hip arthroplasty", on the left, and by another institution for "Cardiac surgery – Pacemaker cardiac surgery", on the right



In Figure 4, there are other examples of statistical discrepancies, wherein institutions charge average hospitalization prices that are much higher than prices charged in the remainder of Brazil for two mining targets. Once again, the manager shall exercise caution in each case of statistical discrepancy, since the institution could have a profile of patients with above average complications, or follow practices that can increase hospitalization cost but also reduce the mortality rate.

3. AUTOMATIC DETECTION OF STATISTICAL DISCREPANCIES

Finding factsheets that are of interest to oversight in a Healthcare Secretariat is a convincing factor for the decision to investigate. On the other hand, searching relevant discrepancies randomly is like finding the needle in the haystack. We have over 5000 targets and approximately 6000 providers. Considering only 12-month periods, for 3 years of production, there would be 36 possible time windows. A simple calculation shows that, literally, billions of factsheets may be extracted from the databases examined.

That is why InfoSAS is so invaluable. It uses various algorithms to search discrepancies, producing scores that can sort and prioritize factsheets. InfoSAS also allows its user to focus in areas defined by geographic filters and by exam period and medical target, since an audit and evaluation professional many times directs his/her attention to particular healthcare fields, such as cardiology or orthopedics and, of course, tends to have more interest in the region where they work.

InfoSAS analyses the time series of average monthly value per procedure and of monthly production for each target desired. These time series are calculated per institution and patients' city of residence. Various anomaly detection algorithms are used and each of them calculates a score. Those scores are combined subsequently. Since there is not enough space here to detail all scores, we will define only two of them below. A more detailed description of another algorithm can be found in (CARVALHO et al., 2015).

The purpose of an algorithm is to detect sudden variations in a provider's production. Therefore, the institution production is compared to its own historic production series (or average cost). In the event of any abrupt deviation from its historic series, the algorithm assigns a discrepancy score to the institution. More specifically, in for each month i and for each institution l, we use the formula $score_{li}=(t_{li}+1)/(mean_{li}+1)$, where t_{li} is the interest level (that may be the cost or number of services) for institution l in month *i*, $mean_{li}$ is the mean of the last m months for the interest level of institution *l* in month *i*. Typically, we adopt m=12 or m=6 months.

Another algorithm seeks to find discrepancies in service rates per inhabitant. Based on Brazilian production per 100 thousand inhabitants, the production time series carried out by all institutions for residents within one city are examined. In the event this city has a production that is higher than threshold, the algorithm assigns a discrepancy score to it. Afterwards, the score is assigned to institutions proportionally to the participation of each one in delivering services to that city.

An institution score is the result of the addition of all scores assigned to it in each city serviced. More formally, a city production score is computed as the last 12 months cumulative sum, $score_{li}=\sum_{j=i-11}^{i} dif_{lj}$, where $dif_{li}=max\{0,tBayes_{li}-threshold\}$ and $threshold=k*Brazilianrate_i$. The value of $Brazilianrate_i$ is Brazilian monthly production rate per 100 thousand inhabitants in month *i*, $tBayes_{li}$ is the empirical Bayesian rate per 100 thousand inhabitants in month *i* for the city *l*, and *k* is a constant previously defined. In our studies, we adopt k=3. The empirical Bayesian rate (ASSUNÇÃO et al., 1998; MARSHALL, 1991) is a statistical

technique to calculate rates and ratios not affected by statistical fluctuations due to small populations.

4. INFOSAS OPERATION

Figure 5 shows InfoSAS dataflow. Every month, SIA, SIH and CNES databases, and IBGE population information, are entered into the data mining server, which is guided by a table of mining targets and runs algorithms to find statistical discrepancies. The output of this mining stage is referred to as the mined fact cube. This cube is explored by InfoSAS users through a BI tool. The user selects a report model and specifies parameters to filter target sets, analyses periods and geographical cuts. A report is issued using the selected model and filters, and it bears discrepancy scores calculated by mining algorithms. The user extracts from the report those factsheets drawing his/her attention to carry out more in-depth analyses.

5. CONCLUSION AND FUTURE WORKS

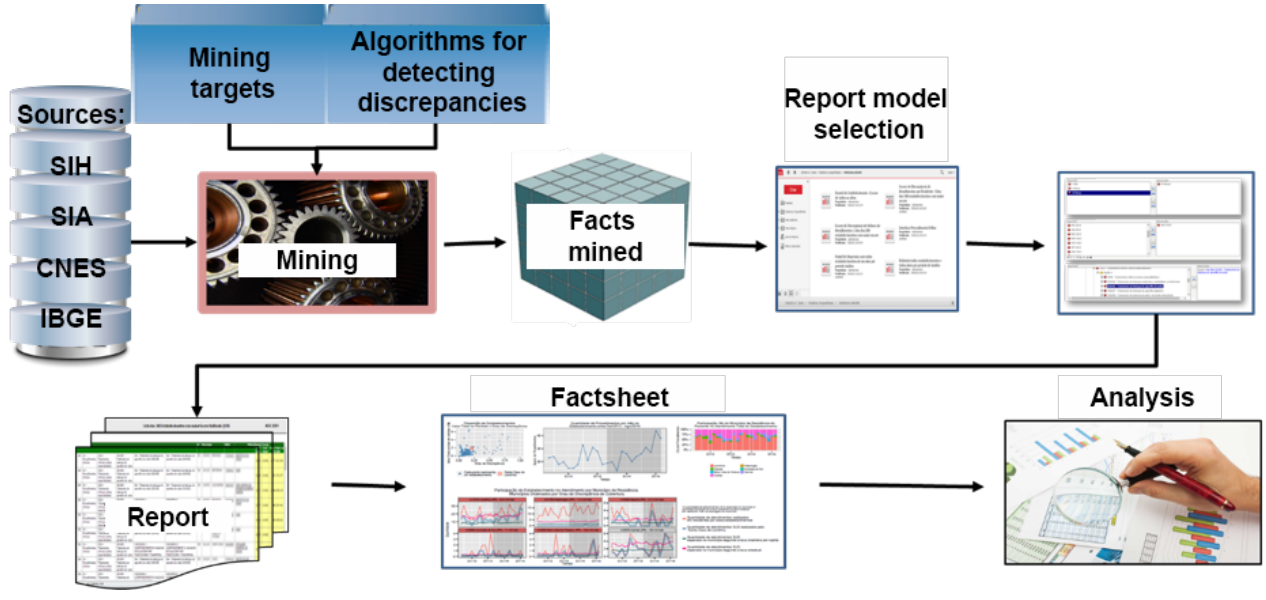
Currently (October 2016), InfoSAS is installed and running on DATASUS. A distance-learning course to train managers to use InfoSAS is being prepared and expected to be ready by November 2016 and offered by the end of this year. InfoSAS system has already been presented at several Brazilian events (DRAC SAS, 2015). In all of them, public health specialists found the results very interesting and useful.

Results obtained by project InfoSAS so far are important and are a step forward in modernizing selection processes for audit and control items. In addition to the need for system maintenance, continuous updates, fixes and fine calibration of algorithms already used, there are many courses for future developments. We envision a new interface with a more flexible and interactive preview of results; a statistical characterization of targets identifying healthcare gaps; the use of SUS population in statistical computation; the calculation of upper and lower limits in order to match a city anomalous production to its providers; the adapted selection of geographic cuts with proper sizing of the frequency of occurrence of each target; and the use of parallelism in preview and mining processing.

In addition, we point out two developments that, in our opinion, are the most important ones. The first one would be feasible by conducting audits, perhaps oriented by InfoSAS alerts, classifying each service as regular or fraudulent. Such classification would make possible the use of supervised learning algorithms, enabling alerts to be issued on a more accurate and complete basis.

The second development consists of a consolidated analysis of anomalies in service rates per inhabitant. The current version of InfoSAS uses several algorithms that try to capture statistical anomalies. We also have a mining target framework where one target may in-

Figure 5:
Data analysis stages by InfoSAS system



clude other targets. Thus, we obtain many findings but we cannot classify institutions by target set nor obtain global estimates of values for anomalous production.

To produce a new report overcoming these problems, we decided to have as targets only those sets of procedures from the same *organization form* of the SUS Table. This causes the intersection between any two targets to be empty. We have also decided to use only one algorithm with a statistical methodology to estimate excessive service rates per inhabitant. This makes possible to find global estimates (in all targets) for anomalous production for each provider, allowing prioritization of fact finding by SUS managers.

The starting point of the methodology we intend to use and already applied to data for 2013 is the fact that the distribution of service rates per inhabitant observed in cities follows a log-normal distribution, as shown in Figure 6. This makes possible to define, for any target and any city, a cut-off point to distinguish what is normal from what is not in service rates per inhabitant.

We regard service rates per inhabitant as abnormal when located to the right of the point dividing the area under a curve in a "normal" part, with 99% probability; and under a curve in an "abnormal" part, with 1% probability, which we deem as a very prudent criterion. In the example shown in Figure 7, rates above 3.2 services per 1,000 inhabitants/month are considered abnormal. Assuming, for that target, that a city with 100,000 inhabitants received 400 services in a given month. By using the maximum rate of 3.2 services/1,000 inhabitants, the population in that city should have received, at most, 320 services. Therefore, we consider that the city had 80 anomalous services, and this excess amount is allocated among providers in the city, unchanging each provider service shares.

The results of applying this analysis to the production entered in SUS in 2013 are very strong. Out of a total of BRL19,912,491,904.00, services considered abnormal totaled BRL413,920,365.56, corresponding to 2.08% of the total. Excess in targets entered into SIH-SUS was BRL350,354,969.25, and in targets entered into SIA was BRL63,565,396.30.

Another result of this study that may be used to prioritize audits is an analysis of allocations among providers of services regarded as anomalous. Throughout Brazil, only 5 providers account for almost 9% of total anomalies; 100 providers account for over 50% of that total. One single provider received almost 10 million Brazilian Reals in services deemed as anomalous in 2013.

6. ACKNOWLEDGMENTS

First of all, we would like to thank the Department of Regulation, Evaluation and Control of the Office of Healthcare Attention, Ministry of Health, which requested and funded the whole project.

Figure 6:
Density of probability of service rates per inhabitant: theoretical log-normal vs. observed data

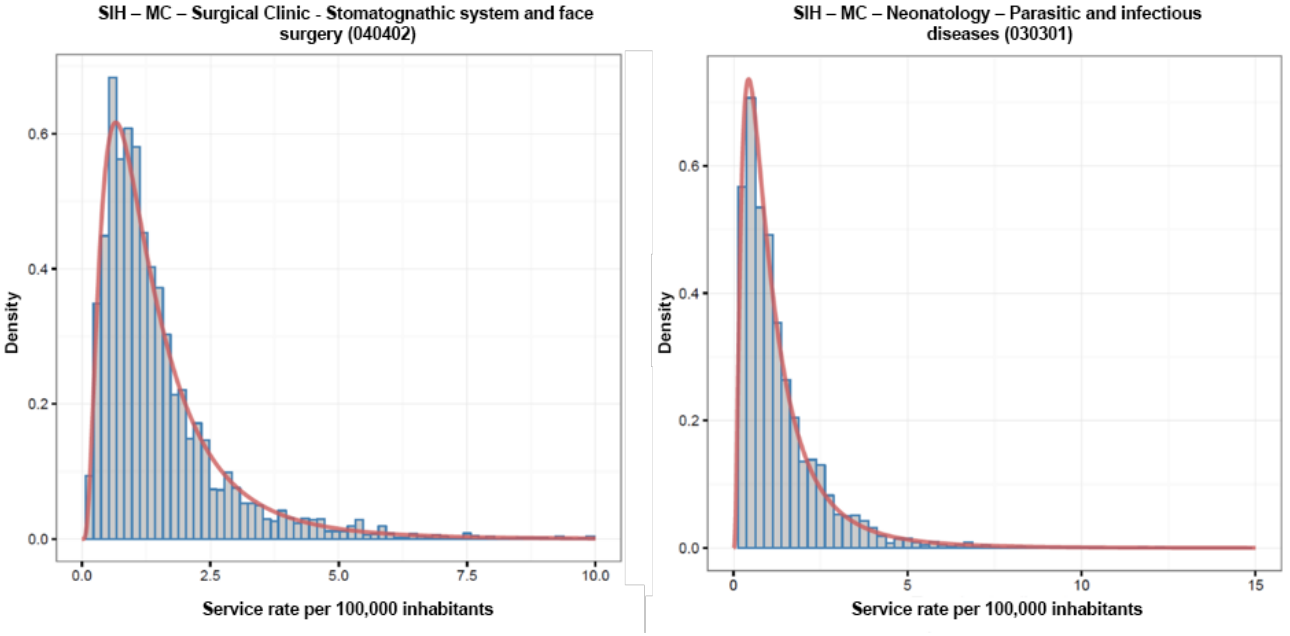
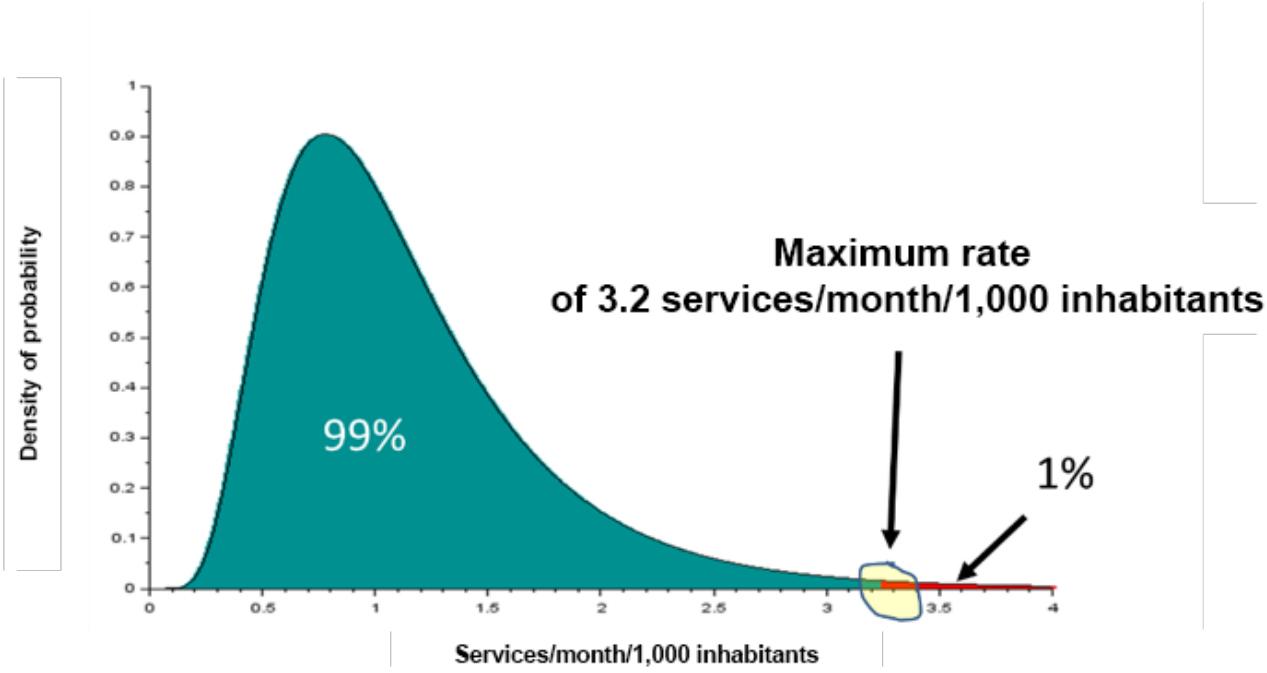


Figure 7:
Definition of maximum service rate per inhabitant at the cut-off point of 1% of log-normal distribution



InfoSAS was built by a large multidisciplinary team. Maria Helena Brandão conducted the project for DRAC. Ester Dias, Marcelo Campos, Sônia Gesteira and Suzana Rattes, from SMSA-BH, Luciana Morais, from Funed, and Mônica Castro, from Unimed-BH, were our health consultants. Fabiana Peixoto and Letícia Neto were the managers, and Tomas Schweizer was the technical leader. Edré Moreira, Carlos Teixeira, Douglas Azevedo, Larissa Santos, Luiz Fernando Carvalho, Luiz Gustavo Silva, Maurício Nascimento Jr., Milton Ferreira, José Carlos Serufo Jr., Pablo Fonseca, Renan Xavier and Raquel Ferreira, graduate scholarship holders participated proactively in the research, building and implementation of various algorithms. Felipe Caetano, Ícaro Braga, Geraldo Franciscani, João Paulo Pesce, João Victor Bárbara, Raphael de Faria and Wicriton Silva were in charge of development, tests, previews and databases. Our sincere thanks go to all the members of the team, the ones actually responsible for building the system.

REFERENCES

ASSUNÇÃO, R. M. et al. Maps of epidemiological rates: a Bayesian approach. Cadernos de Saúde Pública, Rio de Janeiro, v. 14, n. 4, p. 713–723, Oct/Dec 1998.

CARVALHO, L. F. M. et al. A simple and effective method for anomaly detection in healthcare. In: 4TH WORKSHOP ON DATA MINING FOR MEDICINE AND HEALTHCARE, IN CONJUNCTION WITH THE 15TH SIAM INTERNATIONAL CONFERENCE ON DATA MINING, 2015, Vancouver, May 2015. Available at http://homepages.dcc.ufmg.br/~carlos/papers/sdm/dmmh2015.pdf. Web: Nov 29, 2016.

DRAC SAS. Ministério da Saúde. Ciclo de Oficinas do DRAC – Controle de Avaliação. Brasília, DF, Sept 1, 2015. Available at: <https://www.youtube.com/watch?v=vIaR_Q7T-Us>. Web: Jul 4, 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. AI magazine, California, v. 17, n. 3, p. 37-54, 1996.

MARSHALL, R. J. Mapping disease and mortality rates using empirical Bayes estimators. Applied Statistics, Auckland, v. 40, n. 2, p. 283–294, 1991.

THE ECONOMIST. The $272 billion swindle. London, May 31, 2014. Available at http://www.economist.com/news/united-states/21603078-why-thieves-love-americas-health-care-system-272-billion-swindle. Web: Nov 29, 2016.

Audit App: an effective tool for government procurement assurance[1]

Qiao Li
is a PhD candidate at Rutgers University, in New Jersey, U.S.A. The research work developed by Li includes analytical audit, planning and risk management in audits, audit support systems, and machine learning.

Jun Dai
is an assistant professor at the Southwestern University of Finance and Economics, China, and a PhD candidate at Rutgers University, in New Jersey, U.S.A. The research work developed by Dai includes analytics audit, automation in audit, audit applications and blockchain.

ABSTRACT

Recently, governments in many countries have started open data initiatives to make their operations more transparent to citizens. With the open data, anyone who has an interest in monitoring government spending can apply technologies to perform analyses on open data. This study proposes 29 audit apps that could assist various parties in analyzing open government procurement data. These apps could help investigating procurement data from different perspectives such as validating contractor qualification, detecting defective pricing etc. This study uses Brazilian Federal Government procurement contract data to illustrate the functionality of these apps; however, the apps could be applied to open government data in a variety of other nations.

Keywords: Audit Apps; Data Analytics; Government Procurement.

1. INTRODUCTION

According to the World Trade Organization (WTO), government procurement accounts for an average of 10% to 15% of the Gross Domestic Product (GDP) of an economy[2]. Governments are large purchasers of goods and services, which makes them widely abused victims of procurement fraud[3]. Due

to complicated bidding and contracting processes, various contract types, confidentiality of related information and decentralization of data storage, it is very difficult for interested parties to organize, analyze, and monitor the procurement contracts of federal and local governments. In recent years, governments in many countries, such as the United States of America, Canada, and Brazil, have started open data initiatives, aiming at making information about government operations more readily available, useful, and transparent for their citizens[4].

Although open data sources make governments' information available to the public, few studies or methodologies have examined how interested parties can collect and analyze the data. Therefore, this study aims to develop and propose audit apps that could serve as efficient tools for monitoring government expenditures and detecting potential contract anomalies. Audit apps are formalized audit procedures performed through computer scripts, which have seen a recent increase in popularity. This analysis proposes a framework that provides guidance to design effective audit apps that examine government procurement contracts. To illustrate this we also designed 29 audit apps that identify potentially high-risk contracts. Eight of these apps are developed to demonstrate their utility and benefits using the contract data from the Brazilian federal government.

2. BACKGROUND

2.1 OPEN GOVERNMENT DATA AND RELATED ISSUES

In the term open data, "open" indicates that the data are freely available to everyone to use and redistribute (AUER et al., 2007). Such data are gathered and maintained by governments and can be accessed by citizens through online websites. Many countries have built open databases to make data available to the public. For example, Brazil has published a federal procurement information system called "SIASG". Brazil also provides Application Program Interfaces (APIs) for citizens to download data about federal procurement contracts, associated suppliers, goods etc. Although a large variety of government data have been disclosed, this study focuses only on open data that relate to government procurement. Examples of countries that have built websites and databases containing government procurement data include the U.S.A., China, Australia, Canada, Brazil, and the U.K. Open data have had some successes. For example, the British government published government contract data in 2010. Using this, a British official found duplicate purchase records in several government departments that cost over £4 million ($6 million). In another example, officials in San Francisco made transport data public in 2012. They estimate the re-

sulting fall in phone queries has already saved over $1 million (ECONOMIST, 2015). There are, however, some issues with open data that hamper further success. Firstly, data quality is not good enough. Although governments of many countries are required to disclose related data, the level of disclosure varies. Are all the useful data fields given? Are the data details adequate? Is there any missing information? These issues could alter the effectiveness of open data analysis. Secondly, some open data are not prepared in a machine-readable format such as scanned PDF format (which are actually pictures). This makes it harder to collect and analyze such data. Thirdly, the amount of data is huge. This means that searching open data portals and extracting relevant and useful information is often an arduous task. Another issue is that few individuals have the skills to mine data, interpret data, and then put those interpretations or conclusions to good use. Finally, unpublished data could be very valuable, but they are hardly accessible due to data privacy issue (ECONOMIST, 2015).

2.2 INTERESTED USERS OF OPEN DATA

Parties interested in exploring open data, identifying anomalies, discovering irregularities, and detecting frauds include, but are not limited to: citizens, the press, business competitors, and political competitors. There are at least two challenges to address before these groups start working with open data. First, in addition to government open data, they may also need to collect useful data from other sources. Such sources may include social media, the news, government reports, and analysts' reports. It is difficult for interested parties to collect this information and integrate it with government open data. An even bigger challenge exists in analyzing the large amount of data available. Few analytical tools have been specifically designed for the variety of users who want to investigate government data (O'LEARY, 2015). Although there are many general data analytics software programs on the market, the inherent complexity of these analytical tools may impede users from understanding and using the software. This problem is especially pronounced for users with limited auditing and data analytics background. Therefore, developing efficient and effective data analytical tools is a critical issue.

2.3 APPS IN THE AUDITING DOMAIN

Audit apps are formalized analytical routines performed by a computerized tool (DAI; KRAHEL; VASARHELYI, 2014). Each audit app often performs a single analytical audit test, usually requiring few user interactions. Users only need to load data into audit apps to obtain results without many complicated operations. Auditors can even create customized audit apps that accomplish special audit tasks. Audit apps have seen a recent increase in popularity. This is in part due to software developers and audit ser-

vice providers, who have devoted efforts to create audit apps[5]. Few apps on the market, however, are specially developed for citizens wishing to analyze open government data.

Audit apps could be among the favorite tools to analyze open data for several reasons. One reason is that apps in general can be operated easily and with only minimal training. This allows various interested parties from all education backgrounds to analyze data with ease. The low cost of apps is another reason why they are an attractive option. Most users are ordinary citizens or business competitors and not professional auditors. As a result, it is difficult for them to afford expensive professional audit software. Audit apps are cost-effective substitutes that allow users to perform a wide variety of analytics-based audit tests. Customization is another important advantage of apps, because it allows extending apps to fulfill user-specific tasks. Users could create customized apps using professional Software Development Kits.

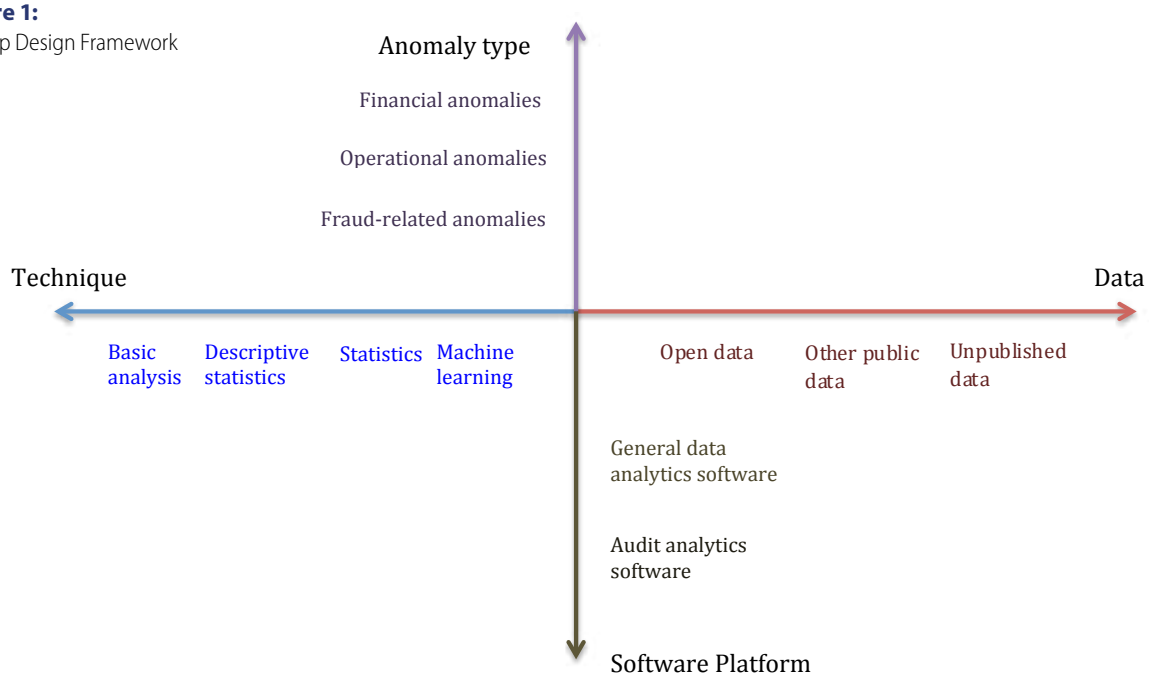
3. AN APP DESIGN FRAMEWORK

Guidance is needed for the design of efficient and effective apps. We propose a framework that provides app design guidance for government procurement audits. The framework (shown in Figure 1) contains four dimensions: anomaly type, data type, software platform, and technique.

The first dimension is anomaly type. Anomalies could be categorized into three types: financial, operational, and fraud-related. Financial anomalies refer to data anomalies that may affect government financial statements or reports. These may include missing data or incorrect/outdated values. Operational anomalies focus on the inefficiency of government operation and activities. These include the purchase of luxury office products (such as an extreme expensive massage chair), obviously unnecessary goods (such as jewelry), or necessary goods priced significantly higher than the market rate. Fraud-related anomalies are unusual data patterns that could result from frauds. For example, if all bidders offer a uniform price and refuse to negotiate during the government bidding process.

The second dimension is the software platform. This will be used to develop and operate audit apps. Generally there are two types of platforms: Audit Data Analytics Software (ADAS)[6] and Generalized Data Analytics Software (GDAS)[7]. ADAS contains pre-programed audit functions making it easy to build apps based on those functions. GDAS usually has the capability of handling high-volume data such as various open data. GDAS also contains a wide range of statistics and machine learning models making it useful

Figure 1:
An App Design Framework



for app development. The drawback of this type of software is that it requires technical background or training of users in order to understand those sophisticated models. Because each software platform has its special characteristics, developers may create several versions of apps that allow users to run them on different software.

The third dimension is data type. Recently, many studies have discussed the use of "big data" for assurance purposes (VASARHELYI; KOGAN; TUTTLE, 2015; CAO; CHYCHYLA; STEWART, 2015; YOON; HOOGDUIN; ZHANG, 2015; ALLES, 2015). Since more countries and organizations start to open their data, such open data will quickly move towards big data (O'LEARY, 2015). In addition, other public data, such as articles of the news, social media, and data collected from machines or various sensors can be collected. These data can all be combined and used to detect anomalies or support investigations (VASARHELYI et al., 2015). Furthermore, certain unpublished information, such as internal operation policies of governmental units, could also provide insight into identifying abnormal government activities. When combined with public data, unpublished information can be used to locate risky government operation processes, and identify potential fraudulent contracts or transactions. Apps need to be developed to facilitate the analyses of a wide variety of data.

The last dimension is technique. Basic analytic techniques could be used to detect unusual patterns. Such techniques include summarization, query, and data matching. Descriptive statistics could provide a wide view of data. Statistics (e.g., regression and time series) and machine learning (e.g., clustering and classification) are advanced techniques that can effectively uncover patterns hidden within complex data. Developers should take advantage of such techniques to create effective and efficient apps.

4. DESIGNING APPS FOR GOVERNMENT EXPENDITURE AUDIT

We proposed 29 apps that could assist with identifying potential anomalies in government expenditure. Such anomalies include suspicious contracts or the use of unqualified suppliers. Table 1 shows the proposed apps.

The first three apps provide a preliminary check on the reliability and completeness of government contract data. They also demonstrate any special data patterns. Abnormal data patterns, such as missing values and abnormal fluctuation, may indicate risks in the contracting process. These apps provide preliminary data-level assurance by identifying contracts with missing or abnormal values. Apps 4 to 10 assist in identifying suspicious suppliers. Since suppliers could be involved in many types of frauds, it is necessary to develop apps to monitor suppliers' behaviors during the bidding and contracting processes. These apps are therefore designed to detect frauds such as bribery,

Table 1:
A list of proposed apps

N°	Purpose of App	Anomaly Indicator	Data	Techniques
Apps for Data Incompleteness/Unreliability Check and Data Patterns Discovery				
1	**Check abnormal contract value**	Unusual number such as 0, 0.01,0.05	Contract data	Query
2	**Check data completeness and integrity**	Missing suppliers/bidding mode/dates/etc.	Contract data	Query
3	**Discover special data patterns (using dashboard)**	Abnormal data fluctuation, extreme large or small numbers etc.	Contract data	Descriptive statistics
Apps for Suspicious Suppliers				
4	**Check supplier information**	Suppliers do not exist in the supplier master file	Contract data, Supplier data	Data matching
5	**Check supplier qualification**	Suppliers, or their parent companies/subsidiaries, are on the list of suspended companies	Contract data, Suspended companies	Data matching
6	**Check relationships**	Family members of suppliers work for the government	Information of suppliers' family members	Query
7	**Check "waived bidding"**	Suppliers have a high proportion of contracts that did not go through normal bidding processes	Contract data	Query
8	**Check abnormal bidding winners**	Some companies always win, or all suppliers win equal number of bids	Contract data, Bidding data	Summarization
9	**Check regional distribution of suppliers**	Suppliers in specific geographical areas win most contracts	Suppliers' geo- information	Visualization
10	**Check abnormal bidder combination**	The same bidders always or never bid with each other	Bidding data	Summarization
Apps for Abnormal Prices and Initial Contract Value				
11	**Check abnormal contract values**	The initial values of contracts do not conform to Benford's Law	Contract values	Benford's Law
12	**Compare contract prices**	Suppliers submit much higher prices in government bids than market price	Price data	Data matching, query
13	**Detect split purchase**	Contracts with same suppliers, dates, and goods/services	Contract data	Duplicate detection
14	**Predict and detect abnormal winning prices**	Winning prices are much higher than the predicted prices	Contract data	Regression
15	**Detect abnormal price gaps**	A large gap between the winner's bid price and others' prices	Pricing information	Query
16	**Check the standard deviation of bidding prices**	All suppliers' prices appear uniform. Suppliers refuse to negotiate the prices	Bidding prices	Standard deviation
17	**Check contract changes**	The initial values of contracts are largely changed	Contract data	Data matching
Apps for Abnormal Bidding Procedure and Mode				
18	**Check valid bidders**	Only a very few valid suppliers	Bidding process	Query
19	**Detect withdrawal of bidders**	Qualified bidders inexplicably withdraw valid bids	Bidding process, supplier data	Query, data matching
20	**Detect duplicate bidding offers**	Companies submit bids with identical individual line items or lump sums	Bidding process	Duplicate detection
21	**Detect duplicate bidders**	Bidders with duplicate addresses, fax or phone numbers, or other demographic data	Demographic information of bidders	Duplicate detection
22	**Check abnormal fluctuation of bid price**	Competitors announce price increases at the same time, at the same amount	Bid prices during the bidding process	Duplicate detection
Abnormal Product/Service Implementation				
23	**Check luxury products**	Government purchase luxury goods	Contract data	Text mining
24	**Check obviously unnecessary items**	Government purchase many unnecessary items such as gift cards	Contract data	Text mining
25	**Check excessive costs**	Costs greatly exceed estimates	Contract data	Query
26	**Check working hours**	Employees are billed for more hours than typically working hours	Invoices	Query
27	**Check duplicate billings**	Duplicate billings for the same products or services	Billings	Duplicate detection
28	**Check abnormal delivery location**	The delivery location is not an office, plant, or job site	Delivery address	Visualization
29	**Check geographic information of invoices**	Employees are billed at multiple distant job sites on same day	Invoices	Visualization

kickback schemes, and bid rigging practices. Apps 11 to 17 are designed to identify unusual price patterns. Since risky contracts are usually associated with abnormal prices, those apps could draw attention to high-risk contracts. Apps 18 to 22 can be used to analyze and monitor the complex bidding procedure. These apps could quickly identify suspicious bidding behaviors. Suspicious bidding behavior may include when only a few valid bidders participate in a bid, or when there are any inexplicable bid withdrawal behaviors. Apps 23 to 29 help to identify purchases of abnormal or unnecessary products, as well as charges of services not rendered. This includes over-spending on office products or misusing public funds.

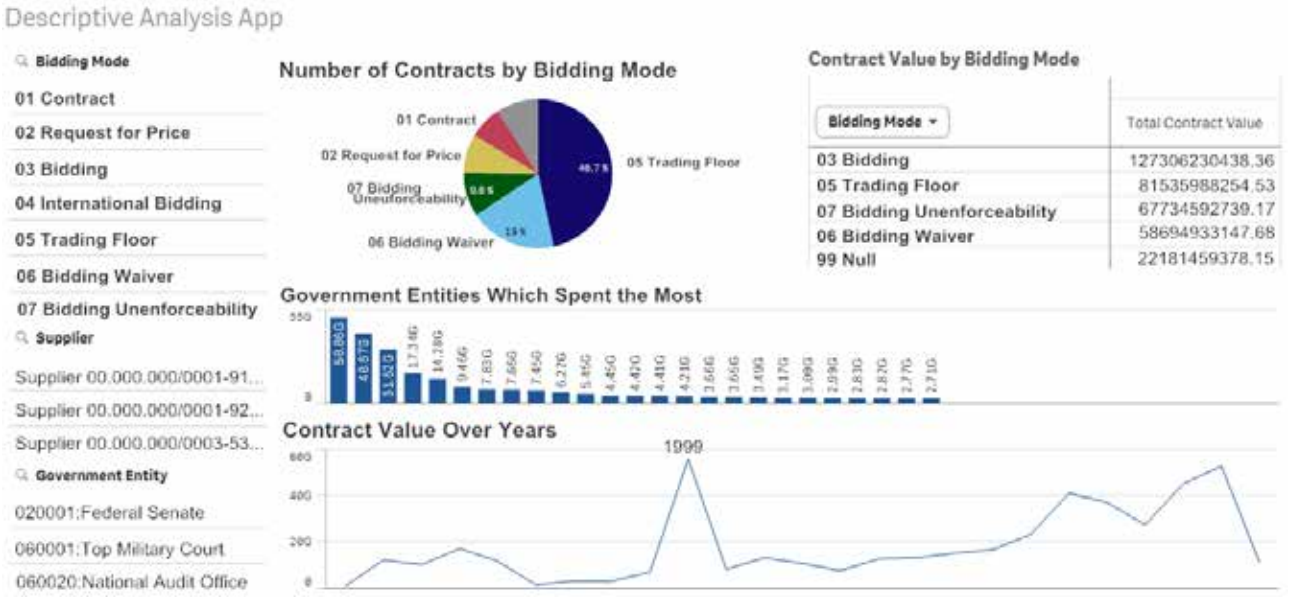
5. ILLUSTRATION OF APPS USING BRAZILIAN FEDERAL PROCUREMENT CONTRACTS

In order to demonstrate the usefulness of audit apps in government contract audit, we collect contract and supplier data from the Integrated System of General Services Administration (SIASG), and the National Registry of Suspended Companies (CEIS), from 1989 to 2014. The contract file mainly contains information about government entities, goods/service suppliers, bidding methods, starting/ending dates, and the initial values of the contracts. The supplier file records companies that preregistered to participate in the bidding processes. Data from this file includes companies' CNPJ and their demographic information. The CEIS system records the CNPJ of companies that are banned from selling products or services to the Brazilian federal government and includes the start and end date of their sanctions. Based on the data, we developed eight apps to perform descriptive analysis, data incompleteness/integrity checks, and anomaly detection.

5.1 DESCRIPTIVE ANALYSIS DASHBOARD

An audit app[8] is developed aiming to perform descriptive analysis on the government procurement contract data. Figure 2 uses a dashboard to show the results. The left panel lists the important fields in the dataset. The right panel uses several charts to show the descriptive analysis of contract values, bidding modes, and government entities. The pie chart shows that 9.6% and 19% of purchases were associated with a bidding mode of 06 (bidding unenforceability) or 07 (bidding waiver), respectively. The table summarizes the initial value of contracts in each bidding mode. According to the table, contracts with bidding mode 07 and 06 rank the third and fourth respectively

Figure 2:
Audit App for Descriptive Analysis (Adapted from Dai; Li, 2016)



in terms of initial contract values. The information from the pie chart and the table indicates that contracts with bidding mode 07 or 06 could be at a high risk of fraud. This is because kickback and bribery schemes are likely to occur when goods or services are purchased without regular bidding processes. The bar chart shows the spending of each government entity between 1989 and 2014. The top three government entities are outliers who spent much more than the rest. The line chart shows the changes in total contract values over time. A peak is found in 1999, the year when the first financial crisis hit Brazil[9]. This indicates a potential risk concern for government expenditures.

5.2 DATA INCOMPLETENESS AND INTEGRITY CHECK

The important fields in contract data are the suppliers, bidding mode, and the start and end dates of the contracts. Therefore, the integrity of those critical data should be checked before performing advanced analytics. Three apps[10] are created for checking the integrity of each of the three fields, and report contracts with missing values. The results show that 35,516 (out of 470,683) contracts do not have supplier information; 16,167 contracts are missing bidding mode, and the start or end dates are not shown in 1,000 contracts. Missing values in the critical fields could result from simple input errors, or fraudulent activities such as bribery or kickback schemes. Further audit tests should be performed to identify and verify the reasons for such missing values.

5.3 ANOMALY DETECTION

High-risk contracts are usually associated with special data patterns that rarely occur in normal situations. For instance, a supplier may win a bit by offering an extremely low price at first, only to raise that price later on. Another example would be a company continuing to sell goods and services, despite being penalized for a breach of obligation. Abnormal patterns in contracts usually indicate fraudulent behavior on the part of suppliers, government entities, or both. Such frauds can cause huge losses to governments. These contracts should be identified and flagged for further investigation. In this study, four apps are developed for detecting contracts having abnormal initial values or suppliers.

An audit app[11] is developed to detect contracts with extremely small initial values (less than 0.1 Brazilian real). The results show that a total of 9,334 contracts have initial values under 0.1 Brazilian real. Among those contracts, 8,678 contracts have 0.00 initial value, 625 contracts have initial values larger than 0.00, but smaller than 0.05, and the initial value of the remaining 31 contracts fall in the 0.05 to 0.10 range. All those contracts should be flagged. Further tests should also be performed on these to examine if the extremely small initial values are reasonable.

Another app[12] is developed to perform Benford's Law analysis on the initial values of the contract. Benford's Law has been widely used in accounting for fraud detection (NIGRINI, 1999; NIGRINI; MILLER, 2009). The frequencies of the first one or two digits of the contracts' initial values should not exceed those suggested by Benford's Law. If this is not the case, it may indicate a potential fraud involving the contracts. The results of the app suggest that there are more contracts that have initial values starting with "60", "79", and "80" than expected. This indicates potential risks in these contracts. The internal policy of the Brazilian government allows simplified bidding procedures if a contract value is lower than 80,000 Brazilian reals. Direct purchases without bidding are allowed if the value is no more than 8,000 Brazilian reals. Results of the Benford's Law analysis indicate that companies and government agencies may have colluded in reducing the initial contract value in order to conform to these limits which simplify purchasing processes.

Examining suppliers is an important aspect of government procurement audit: auditors should pay more attention to purchases made with companies on the CEIS registry. This is because the use of such high-risk suppliers is more likely to result in breach of contract, or the supply of low quality of products or services. Contracts signed with the subsidiaries or parent companies of the suspended companies also need careful examination. These companies may subcontract the work to those related firms under sanction. An audit app[13] is developed to identify the contracts whose suppliers or the subsidiaries/parent of the suppliers are on the CEIS registry. The results show that a total of 25,100 contracts have been signed with companies or their subsidiaries/parent that are, or were, on the CEIS list. For example, the supplier "33.000.118" and their subsidiaries have signed 1,717 contracts with Brazilian government entities from 1989 to 2014. This occurred despite the fact that this supplier was temporarily suspended by a specific government agency from December 15, 2014 to December 14, 2016. Maybe this supplier serves other government agencies; however, the contracts signed by company and its subsidiaries should be carefully examined.

The SIASG system contains an independent supplier file that records preregistered companies participating in the bidding processes. Although registration is not mandated, companies in the supplier file have a lower risk of fraud as compared to non-registered companies. This occurs because registered companies are checked during the centralized registration process. They are therefore guaranteed to meet the legal requirements for participating in government contract bidding. On the contrary, contracts signed with companies that did not register are exposed to a higher risk of fraud. There is still a possibility that the companies will win contracts even if they do not meet all the legal requirements. This may occur when there is collusion between the companies and the government agencies. An app[14] is developed to help detect contracts with suppliers that did not preregister in the system. The results show that a total of 40,942 contracts were signed with non-registered suppliers. Further audit tests need to be performed such as examining whether the companies charge more than the market price, whether there is a personal relationship between companies' management and the government agency etc.

6. CONCLUSION

Since government procurement averages approximately 10%-15% of a country's GDP (OECD, 2015), procurement audits and the detection of potential anomalies or frauds are important issues. This study dis-

cussed the use and benefits of apps for government procurement audits. Twenty-nine specific apps are proposed to facilitate audit government expenditure. Eight apps are developed to demonstrate their usefulness in contract audits. This article could provide insights on how to create effective apps, and how to use apps in government procurement auditing.

NOTES

1 This article is based on the study that is published in Journal of Emerging Technologies in Accounting (DAI; LI, 2016). All the figures and tables in this article came originally from that study and are modified to fit this article by the authors.

2 Available from: <https://www.wto.org/english/tratop_e/gproc_e/gproc_e.htm>. Access on: Nov 25. 2016.

3 Available from: <http://www.whistleblower-lawfirm.com/types-of-fraud-Government-Procurement-Fraud.html>. Access on: Nov. 25, 2016.

4 Available from: <https://www.data.gov/open-gov/>. Access on: Nov. 25, 2016.

5 For example, Caseware has over 50 audit apps on an online market (DAI et al., 2014). Qliksense allows users to develop audit apps by creating customized dashboard. Other companies, such as Forestpin and TeamMate Analytics, also developed some app-like products.

6 Examples of ADAS include ACL and CaseWare IDEA.

7 Examples of GDAS include R, Weka, SPSS, SAS.

8 This app is developed using Qlik Sense.

9 http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/16274/Panel%202%20-%20Gabriel%20Palma%202_0.pdf?sequence=1.

10 These apps are developed using Caseware IDEA.

11 This app is developed using Caseware IDEA.

12 This app is developed using Caseware IDEA.

13 This app is built on the SAS Enterprise.

14 This app is built on the SAS Enterprise.

REFERENCES

ALLES, M. G. Drivers of the Use and Facilitators and Obstacles of the Evolution of Big Data by the Audit Profession. *Accounting Horizons,* v. 29, n. 2, p. 439-449, 2015.

AUER, S.; BIZER, C.; KOBILAROV, G.; LEHMANN, J.; CYGANIAK, R.; IVES, Z. DBpedia: a nucleus for a web of open data. *Lecture Notes in Computer Science*, v. 4825, p. 722-735, 2007.

CAO, M.; CHYCHYLA, R.; STEWART, T. Big Data analytics in financial statement audits. *Accounting Horizons*, v. 29, n. 2, p. 423-429, 2015.

DAI, J.; KRAHEL, J. P.; VASARHELYI, M. A. Which audit app(s) should auditors use? An exploratory study of using recommender systems for audit app selection. Working Paper. *Rutgers, The State University of New Jersey*, USA, 2014.

DAI, J.; LI, Q. Designing audit apps for armchair auditors to analyze government procurement contracts. *Journal of Emerging Technologies in Accounting*. Online early, 2016. Available from: <http://dx.doi.org/10.2308/jeta-51598>. Access on: Nov. 30, 2016

Economist. 2015. Open government data: Out of the box. Available from: http://www.economist.com/node/21678833/print. Access on: Nov. 25 2016.NIGRINI, M. J. I've got your number: how a mathematical phenomenon can help CPAs uncover fraud and other irregularities. *Journal of Accountancy* [Internet], May 1999. Available from: <http://www.journalofaccountancy.com/issues/1999/may/nigrini.html>. Access on: Nov. 25, 2016

NIGRINI, M.; MILLER, S. J. Data diagnostics using second order tests of Benford's Law. *Auditing: A Journal of Practice and Theory*, v. 28, n. 2, 305-324, 2009.

OECD."Size of public procurement"in Government at a Glance. Paris: OECD Publishing, 2015. Available from: <http://dx.doi.org/10.1787/gov_glance-2015-42-en>. Access on: Nov. 25 2016

O'LEARY, D. E. Armchair auditors: crowdsourcing analysis of government expenditures. *Journal of Emerging Technologies in Accounting*, v. 12, n. 1, p. 71-91, 2015.

VASARHELYI, M. A.; KOGAN, A.; TUTTLE, B. M. Big data in accounting: an overview. *Accounting Horizons*, v. 29, n. 2, p. 381-396, 2015.YOON, K.; HOOGDUIN, L.; ZHANG, L. Big Data as complementary audit evidence. Accounting Horizons, v. 29, n. 2, p.431-438, 2015.

Geographic data modeling to define alternative transport corridors to bypass the Metropolitan Region of Belo Horizonte: comparative scenarios

José Irley Ferreira Júnior is a geographer with specialization degrees and an M.A. in the fields of Geosciences and Environment. He is an autonomous consultant in geotechnologies and transportation.

Rodrigo Affonso de Albuquerque Nóbrega is a cartographic engineer and has a PhD in Transportation Engineering. He is also an adjunct professor at the Federal University of the State of Minas Gerais (UFMG).

Leise Kelli de Oliveira is a mathematician and an associate professor at the Federal University of the State of Minas Gerais (UFMG). She is the co-author of the book Logística Urbana: Fundamentos e Aplicações (Urban Logistics: Fundamentals and Applications).

SUMMARY

This paper aims to present the methodology used to compute feasible corridor alternatives to bypass the Metropolitan Region of Belo Horizonte (RMBH). Its computations were based on predictive scenarios developed with the use of geoprocessing and multi-criteria analysis. The study area encompasses extreme concerns of physical, biological, economic, social and logistic natures. In spite of the full cooperation of nature conservation units, designated areas for managed aquifer recharge, and land with very rugged topography, the region has suffered strong anthropogenic pressure, particularly with the fierce growth of the urban stain, the installation of industries and warehouses, and mining activity. From a logistic point of view, the region is strategically important as it connects the highways BR-040 and BR-381, which interconnect from the RMBH to Rio de Janeiro and São Paulo, respectively. Although a public notice and terms of reference were released in 2011, a study of technical, economic and environmental feasibility and the engineering design of the south segment of the ring road have not been concluded nor submitted. In this sense, and in order to promote elements for analysis, control and discussion, the purpose of this paper is to produce material suitable for qualifying and quantifying different transport corridor alternatives for the development of this delineation in transportation infrastructure. The developed model made use of information from the public notice and the terms of reference of the project, and the data used were all official and pertaining to the public domain. Multi-criteria analysis was implemented in a geographic information system setting using the AHP technique in hierarchical levels of decision-making. Four scenarios were produced that reflect distinct and competing interests: biophysical, environmental restrictions, socioeconomic and commercial/logistic. In each scenario, the cost surface and the feasibility corridors were computed. The corridors were compared in relation to their extension, declivity, urban area, conservation areas and vegetated area. The study displayed considerable applicability potential for external control. The results showed that predictive scenarios can be used to promote qualitative and quantitative analyses as to the viability of linear infrastructure projects, even in the phase of publishing public notice.

Keywords: Geographic information system. Multi-criteria analysis. Decision-making rules. Transparency. Evaluation of public works.

1. INTRODUCTION

Despite its long historical evolution, the development of road infrastructure projects in Brazil faces considerable difficulties to this day, principally during the planning and implementation phases. According to Nóbrega (2013), the majority of projects are marked

by inadequacies in their planning phase. In addition, a lack of transparency concerning the data and methods employed in the analyses all cause technical and budgetary problems, which reflect in the rising costs and prolonged periods for completion.

In order to achieve effective results in the activity of transportation infrastructure planning, in a transparent manner in both the public and private sphere, it is essential that the projects be guided by systematic thinking processes. Thus, it is vital that the professionals involved make decisions as a team and consider the interaction of the variables present in all stages of the process. This form of reasoning is based on the principle of multi-criteria analysis, aided by the technique of the Analytic Hierarchy Process (AHP) (Saaty, 1995). This model has been explored with geographic information Systems (GIS) that allow for spatial modeling of variables in the decision-making process.

According to DNIT (BRAZIL, 2012), the proposed construction of the ring road features the project in three segments: south, north and east. Based on the above-mentioned methods of analysis, it is the aim of this paper to investigate the feasibility corridors for the "South Ring", a route whose feasibility study is in development and has yet to present preliminary results. In this context, this article presents the results of geographic data modeling to define alternative transport corridors that represent greater economic, technical and environmental feasibility for the implementation of this transportation infrastructure. In addition to the public notice and the terms of reference (BRAZIL, 2012), we extracted from domain geographic data environmental, logistic, commercial and socio-economic variables to be used in the model.

2. MULTI-CRITERIA ANALYSIS AND GIS MODELING IN TRANSPORTATION

The demand for methodologies to modernize transportation planning is widely known and geoprocessing has been a key factor in integrating, in a coordinated manner, the numerous spatial variables of this process. The inclusion of GIS to assist in the planning of transport corridors requires great deal of data. According to Longley et al. (2013), a GIS is characterized by a set of constructors to represent objects, tools and processes in a computerized environment, usually operating in the form of geographic data models.

One of the models discussed and used together with GIS to assist in the decision-making process is multi-criteria analysis. It provides a structured integration of geographic variables, the opinions of the agents involved, even if different or divergent, and weight of the variables in the decision-making rules, with the goal of reproducing diagnostic and prognostic scenarios. In order to subsidize multi-criteria analysis, methodologies have been developed to optimize the workflow, which are then implemented in GIS modeling. Worth mentioning in this context is the analytic hierarchy process AHP, a technique developed by Saaty (1995) to ensure that subjectivity originating from human decisions is minimized by applying mathematical rules in the process of assigning importance to the given variables. According to Sadasivuni et al. (2009), this technique is applied as a method of variable comparison for multi-criteria analysis and uses mathematical modeling to determine priorities--defined as something important in relation to the organization of disparities in the values, opinions and interests of the agents involved in the planning of transport corridors.

While the AHP method uses paired values as input data, alternatively its output information corresponds to a numerical ranking, which lists, orders and assigns importance to the given preferences. The main role of GIS has been in relation to the composition of values assigned to pixels in digital maps in matrix format (NOBREGA et al., 2009).

Although significant advances in the geographic contextualization of decision-making processes in transportation were achieved in the 1990s and the first decade of the twenty-first century, AHP methodology, coupled with GIS, became an area of interest for practical projects of feasibility corridors only a few years ago. In Brazil, the combined use of GIS and AHP in the planning of transport corridor projects is not exclusive to the academy. Recent studies applied to the planning of railways were developed under federal administration with transport and control managers (BERBERIAN et al., 2015). These initiatives show the interest of transportation managers and technicians in the modernization of the planning process. The results demonstrate the potential of geoprocessing in catalyzing not only a huge range of variables involved in transportation planning, but also in helping to model solutions in the face of the complexity of the public, environmental and transportation policies involved in the process.

3. STUDY AREA

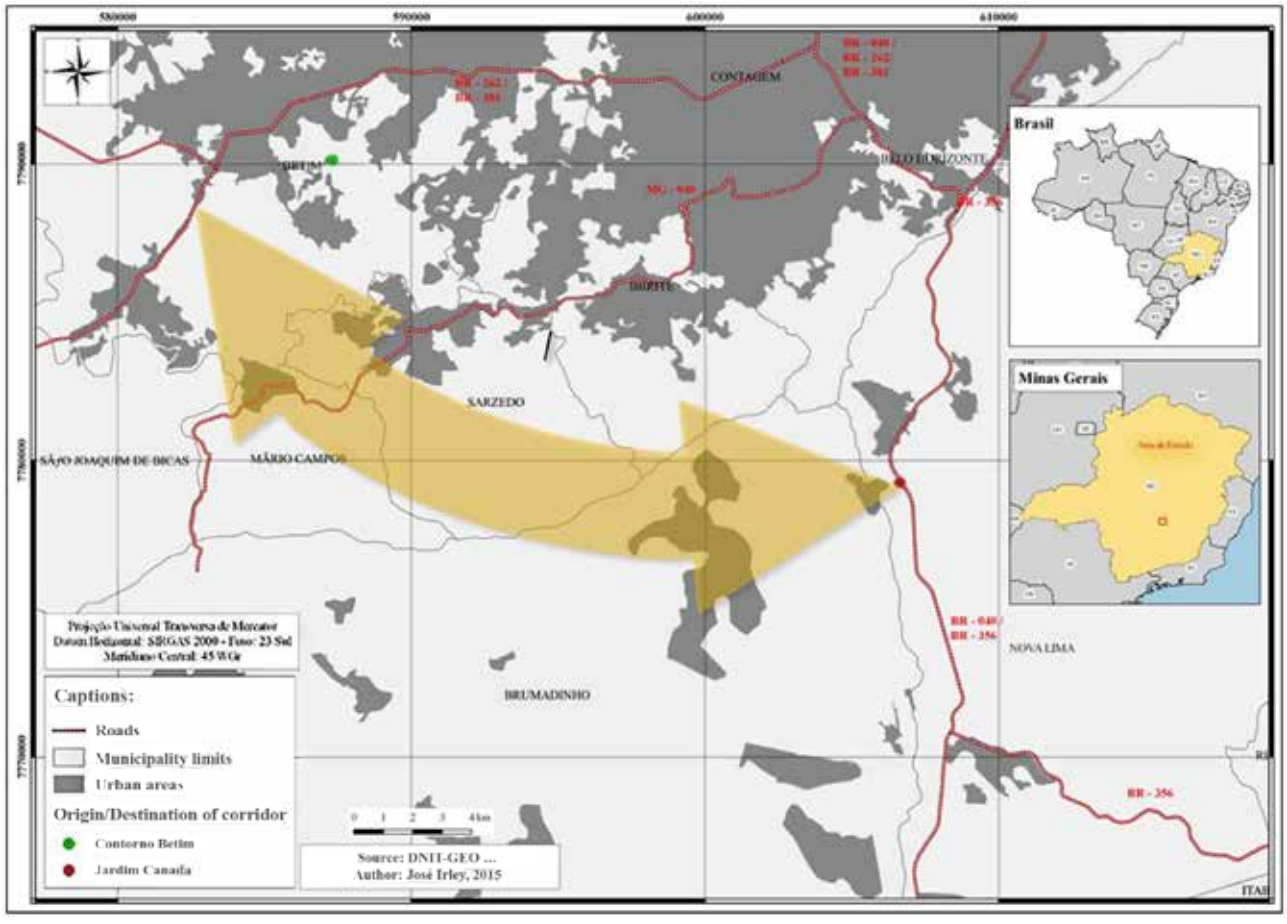
RMBH is composed of thirty-four municipalities. It is an area of high traffic flow density due to its high transportation demands to meet the needs of commerce, industry, mineral prospection activities and services.

Important federal highways such as BR-040, BR-262, BR-381 and BR-356, on which small, medium and large sized vehicles circulate, intercept it. Seeking to mitigate traffic congestion and security problems related to excessive amounts of vehicles which necessarily intersect these urban areas, a ring road was proposed in order to provide an alternative beyond the RMBH. According to DNIT (BRAZIL, 2012), the South Ring Road is part of a highway interchange project including the North Ring Road (the Krupp-Ravenna intersection, 67 km long) and the East Ring Road (Olhos D'agua-Sabará, 22 km long). The region elected for the southern section of the ring road has high population density, industrial activities, vulnerable areas and environmental conservation areas. According to the public notice and terms of reference, DNIT elected the location points in Contorno de Betim and Jardim Canadá in the municipal of Nova Lima, to be the origin and destination of the South Ring Road. Figure 1 illustrates the proposal, linking the traffic flows of BR-040 and BR-381, and their respective exits to Rio de Janeiro and São Paulo.

4. METHODOLOGY

In order to develop this study, data collection and the construction of the database, both tabular and geospatial, were necessary. In regards to technical documentation, the following was used to guide the corridor modeling: "Terms of reference for the study of the delineation and development of the engineering executive project for the southbound highway of the metropolitan region of Belo Horizonte BR-040/MG" (BRAZIL, 2012). This document is identified as Annex I – Basic model for service contracting according to subparagraph (I), paragraph 2, Art. 7, of law No. 8,666 of 21/6/93, under process number 50600.032686/2011-78 (BRAZIL, 2012).

Figure1:
location of the study area



These terms serve to guide the determination of variables found in the text itself, as well as indicate additional manuals and technical documents for consultation, such as the Service Instructions (SI) and Basic Scopes (BS), while also having been used as guidelines for the geographic modeling of this paper.

The geographic bases (Chart 1) of this study were acquired through telephone contacts and personal visits. To facilitate organization, the data were separated into subsets, as described in the terms of reference of the South Ring of RMBH, stating that the delineation should simultaneously consider "the environmental, cultural, social, community, geographical, financial and engineering issues involved in the study of the project" (BRAZIL, 2012). The architecture of the model followed the guidelines of Nóbrega (2014).

4.1 TREATMENT AND PROCESSING OF DATA

One essential procedure that preceded the processing of spatial data was the standardization of the system of coordinates, considering that the simple and correct use of map projection can prevent inconsistencies in the measurements of a linear engineering project. In this study, all the data were redesigned for the Sirgas 2000 Geodetic System with UTM-23S projection.

Subsequently, due to the need for the use of data in matrix format for the multicriteria analysis utilized in the model, the original data in vector and tabular format (Chart 1) were converted to matrix format. The transformation of vector data (discrete) into matrix data (continuous) makes it possible to produce algebra maps. Figure 2 illustrates an example of data transformation from vector to matrix format and their integration with other matrix data for elaborating the accumulated cost surface through algebra maps.

4.2 MULTICRITERIA ANALYSIS, COST SURFACES AND FEASIBILITY CORRIDORS

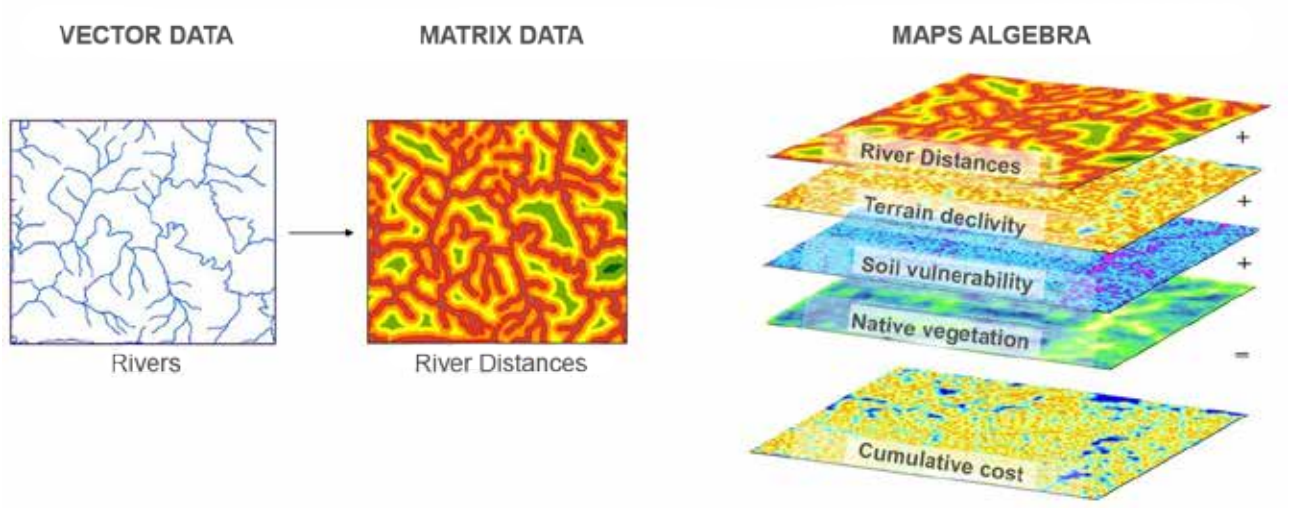
This step consisted of the application of the AHP technique to standardize variables and construction of the accumulated cost surface. This implementation occurs at three levels: intravariables, intervariables and intergroups, the latter being responsible for the integration of the final cost surface. The crux of this process consists in assigning importance ratings to variables of the model that require multidisciplinary knowledge.

At the intravariable level, each incoming data was analyzed before it could be converted into information plans for the model. The technical documentation of the project was consulted to see how each variable

Chart 1:
Database organization

Data	Scale	Source	Date	Type	Subset
Urbanized areas	1:50000	ZEE-MG (2009)	2009	Vector	
Population	1:500000	IBGE (2010)	2010	Chart	
Rural settlements	1:10000	INCRA (2015)	2015	Vector	Socioeconomic
Archaeological heritage – Distance	1:50000	IPHAN (2015)	2015	Chart	
1. Historical, artistic and cultural heritage – Distance	1:50000	Prefectures (2015)	2015	Chart	
Highways – Density	1:10000	DNIT – GEO/Vectorization	2015	Vector	
Urban streets – Density	1:50000	Open Street Map	2015	Vector	
Gas networks – Distance	1:10000	GASMIG	2015	Vector	Marketing and logistic support
Transmission lines – Distance	1:10000	CEMIG	2015	Vector	
Mineral resources	1:1000000	Geodiversity – CPRM	2010	Vector	
Mineral interest	1:1000000	DNPM - SIGMINE	2015	Vector	
VRemaining vegetation	1:150000	Landsat 8 - NDVI	2015	Raster	
Springs - distance from water springs	1:1000000	IGAM	2014	Vector	
Hydrography - density and distance of hydrography	1:1000000	IGAM	2014	Vector	
DTM - Digital Terrain Model (Slope)	1:10000	IGTEC	2009	Raster	Biophysical
Vulnerability to erosion	1:1500000	ZEE-MG	2009	Vector	
Risk of erosion - phyllite, karst and mass movement	1:1000000	Geodiversity - CPRM	2010	Vector	
Reflecting pool	1:1000000	Vectorization	2015	Vector	
Conservation unit – integral protection	1:50000	ZEE-MG	2009	Vector	Environmental restrictions
Conservation unit – Sustainable use	1:50000	ZEE-MG	2009	Vector	
Caves – Distance	1:50000	SECAV	2015	Vector	

Figure 2:
Illustration of the treatment and processing of the data model



could be utilized and to ascertain the level of importance of each class present in the data. For instance, the Euclidean distances of the watercourses in Figure 2 can be seen, while the vector data shows only the presence or absence of a river. The matrix data reveals how far away it is. These distances have been categorized and their importance weighed according to criteria, which indirectly reflect on the possible presence of riparian forest, soft and collapsible soils or in the high costs of crossing. The conversion of vector-matrix data and the use of the AHP technique for each variable were conducted, as shown in Figure 3 (above), thus creating information plans that were integrated by group for the development of the second level of multicriteria analysis, as proposed by Nóbrega (2013).

For the intervariable level, information plans were organized into four groups: environmental restrictions, biophysical, marketing, and logistical and socioeconomic support. Weighting among the different information plans per group resulted from consultations with specialists. Paired comparison analysis was adopted to prevent inconsistencies in the results, as described in Sadasivuni et al. (2009). As a result, cumulative cost surfaces were generated that correspond to the maps in matrix format where each cell is represented by the value calculated from its respective cumulative effort (or implementation cost) of the variables that participated in the composition of each group. Figure 3 (Center) illustrates this process for the variables of a biophysical group. The process was reproduced for the other groups, resulting in four areas of cumulative effort, which served as input for the third level of the AHP process – intergroups.

Figure 3 (below) illustrates the integration of the groups in four different value settings, obtained in turn through distinct perspectives that sought to focus on the feasibility of the corridors in the preservation of self-interests. Each perspective adopted was intentionally focused on defending environmental concerns, reducing costs of engineering, meeting market/logistic demands, or on minimizing negative socioeconomic impacts. As a result, four distinct scenarios were produced, each conservative in their interests, in order for the alternative transport corridors to then be computed.

Once the integrated cost surfaces for each scenario were computed, they were used as the basis for computationally simulating the anticipated effort for connecting the points of origin and destination, the beginning and ending points of the project, located on BR-040 and BR-381, respectively. This calculation is done in two stages, in which the costs of removal of the two extreme points are initially computed (effort vs. distance), to then integrate the two resulting maps on a surface to reveal the corridor of least effort, and consequently of greatest feasibility according to the perspective of the adopted scenario (Figure 4).

5. RESULTS

The methodological development, had the objective of generating transport corridors, the results of

Figure 3:
Intravariable, intervariable and intergroup decision rules used in the model

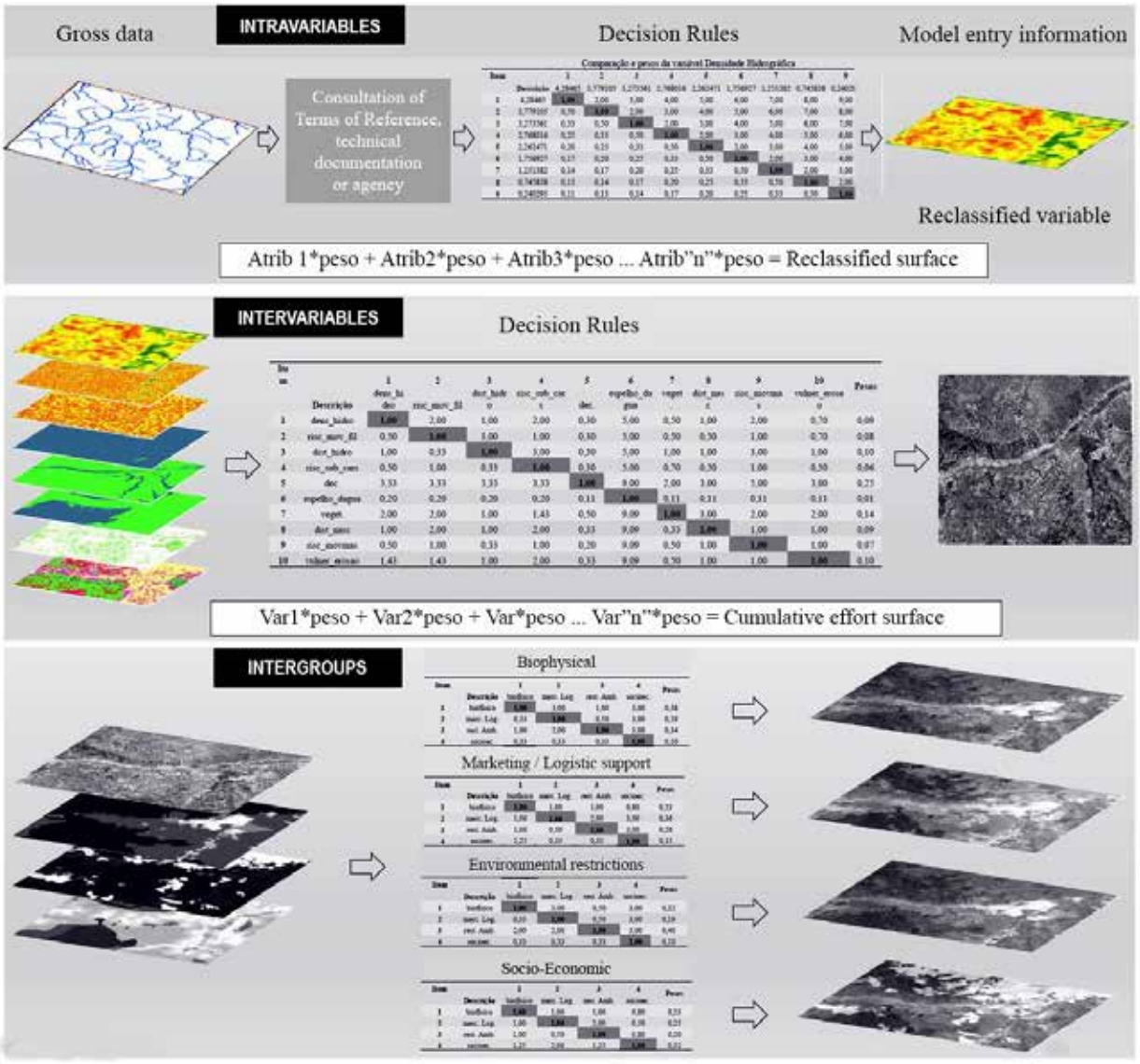
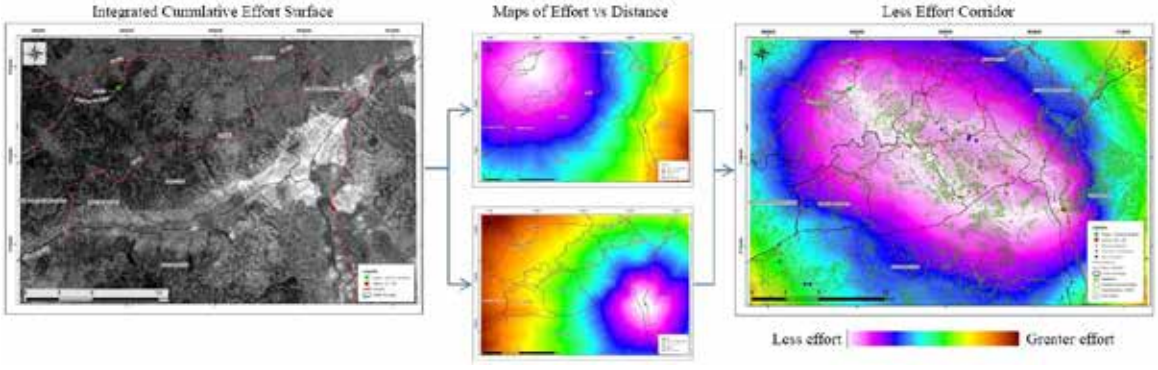


Figure 4:
Illustration of the calculation process of a feasibility corridor



which are presented in Figure 5. As such, four scenarios were produced: biophysical, socioeconomic, environmental restrictions and marketing and logistical support. These scenarios maintained the origin and destination in the terms of reference.

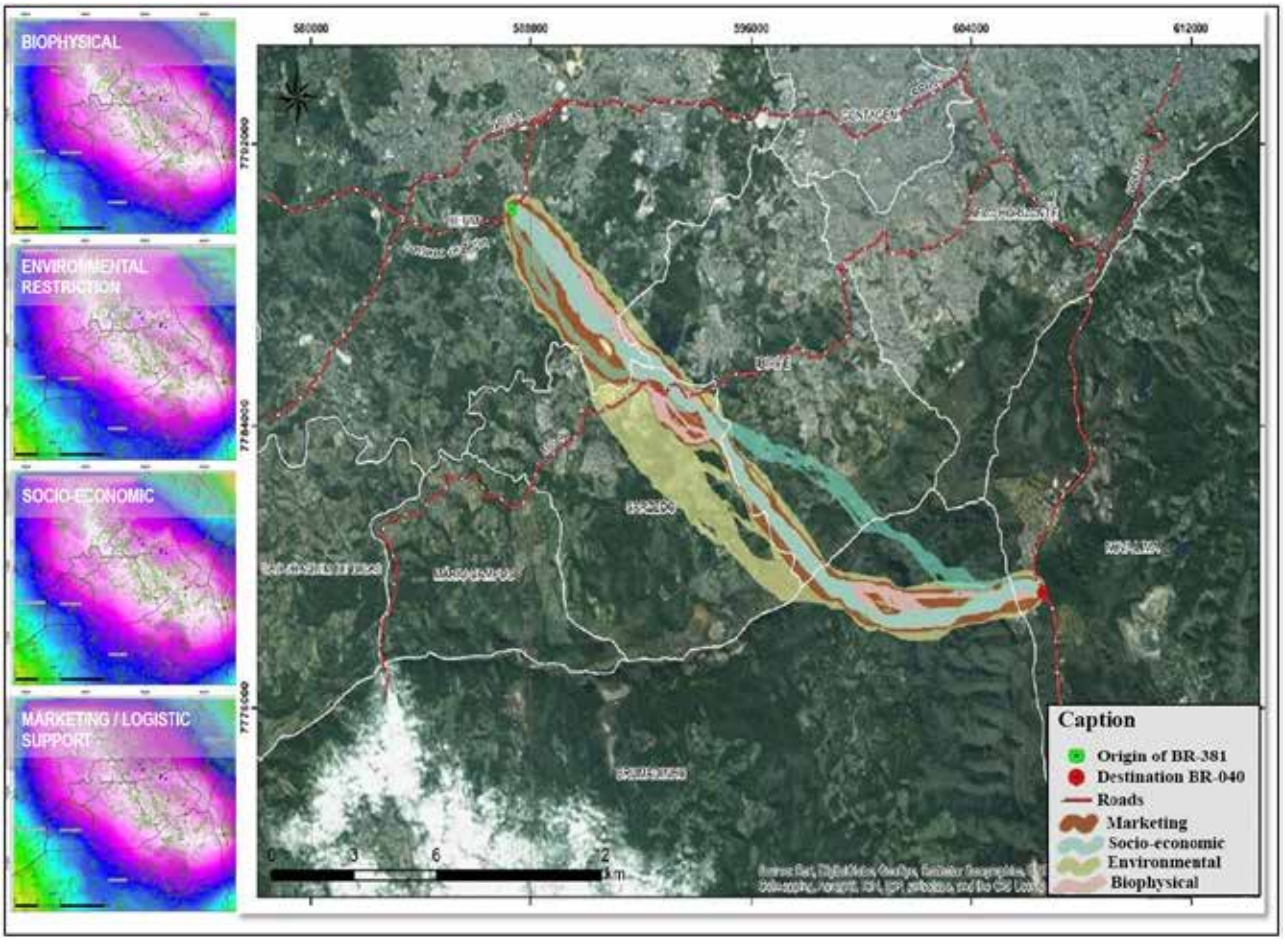
With the design of the transport corridors, it was possible to calculate comparative metrics, such as the full extension of the project and their intersections, with other databases in order to quantify values, assess impacts and compare alternatives. To demonstrate this sensitivity analysis, four variables were chosen which are generally prevalent in transport corridors: slope, urban area, conservation areas and vegetated area. However, the methodology can also be applied to quantify the number of residences to be affected and rivers to be transposed or even to monetize the impact of every alternative corridor, depending on the availability of data present in the study area.

With regard to the extension of the alternatives, the environmental scenario produced a corridor of 24.8 km, the biophysical of 25 km, marketing with 25.2 km and the socioeconomic scenario produced two alternatives of 25.2 km and 23.7 km long, respectively.

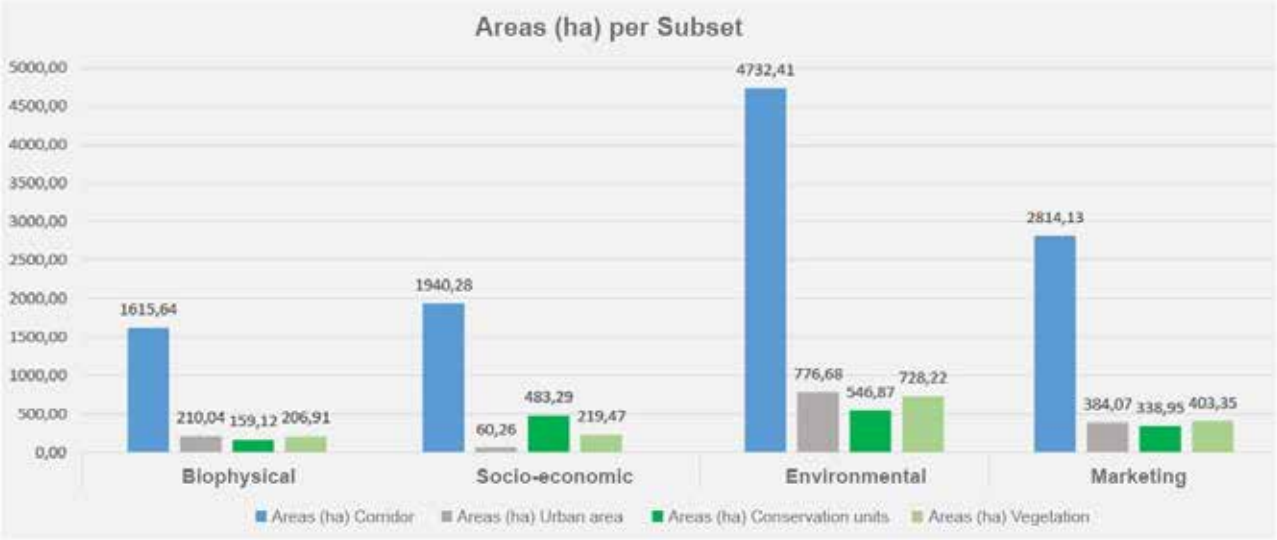
While results from the terrain slope, analysis showed that, inevitably, the delineation of the corridors runs through areas of high declivity, due to the geomorphological nature of the region. However, the model considered the high level of effort to transpose these areas and the construction of the corridors avoided areas with sharp rises in the terrain. A summary of the average slope per corridor of the four evaluated scenarios showed that the biophysical setting scored the lowest average (13.7%), followed by the marketing/logistics scenario (13.9%) and lastly environmental restrictions (14.1%). The socioeconomic scenario presented the highest average (14.6%).

For the variables of urban area, conservation areas and vegetation, the affected areas can be seen in Graph 1. In regards to the biophysical setting, for which weighting of the variables of vegetation and conservation units was

Figure 5: Ttransport corridors computed for each scenario



Graph 1:
Impact metrics for urban areas, vegetated areas and those pertaining to conservation units intercepted by the transport corridors of the analyzed scenarios.



greater, it was observed that when the model was free to construct the corridor, the results showed improvement, corresponding to less impacted areas. In the socioeconomic subset, whose preference for distance from urban areas is evidenced, the results achieved the lowest impact among all the scenarios evaluated.

However, the analyses show the existence of various possibilities for its use once we identify what will undergo intervention and when. The model enables perfected risk analysis and optimized filling out of the environmental impact matrix used to evaluate project alternatives.

Two other proposals were developed taking into account the biophysical and socio-economic scenarios, designating the point of origin as established by DNIT on BR-381 and the destination on BR-040 without defining a specific location (Figure 7). The idea was to verify whether the location on BR-040 published in the public notice corresponds to the location of greatest feasibility according to the criteria adopted in this investigative research. Although not quantified, the preliminary analyses indicate that the point of origin published in the public notice of the project is justified when considered the need for greater distance of the ring road from existing urban areas. Notwithstanding, the model indicated alternative connections between the Ring Road and BR-040 which would cause lesser environmental impacts and possibly lower engineering costs in virtue of running through flatter areas.

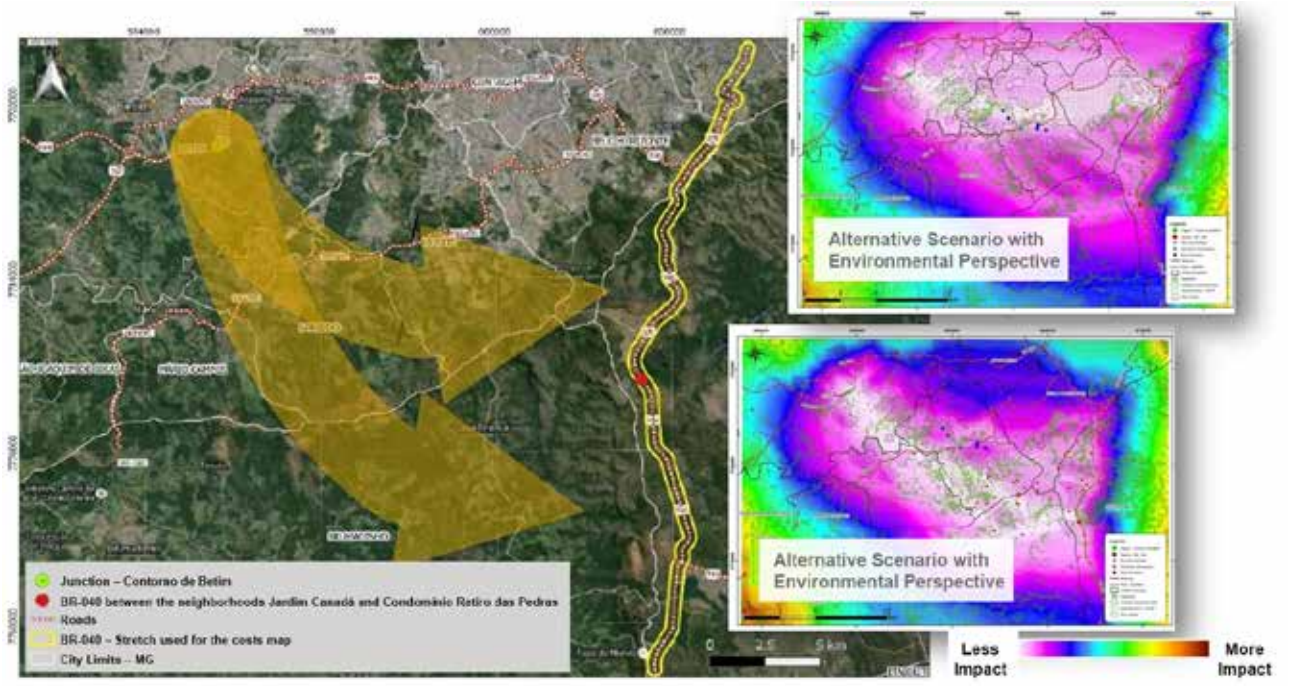
6. CONCLUSION

We faced various challenges throughout this research project regarding interpretation of the terms of reference and the technical documentation as well as data processing. Nevertheless, the central objective of obtaining alternative delineations for the South Ring Road of the metropolitan region of Belo Horizonte by way of distinct, realistic and confrontational perspectives was achieved. Transport corridors were generated considering different variables, supported by the use of geotechnology and multicriteria analysis. The results made possible a harmonic interaction between the application of this technique and the modeling of geographic data. The operational capacity of the model for processing large areas with extensive contextual details of the analyses, as well as time optimization and the analytical capacity of the agents involved in the model, all contribute positively to the modernization of the transportation planning process.

The spatial multi-criteria analysis pilot work showed considerable applicability potential for external control. We believe that, with this model and its operational knowledge, audit activities and external control may be carried out with greater speed and accuracy.

The transport corridors resulting from this study still cannot be compared to the official delineation of the RMBH Ring Road, considering that the latter has not conducted a feasibility study. We expect that the comparative

Figure 7:
Alternative processing to verify adherence of the points of origin and destination as published in the public notice of the project.



metrics can be quantified and analyzed after the project is completed by DNIT or it is presented to the community.

7. ACKNOWLEDGMENTS

The authors thank CNPq for their support in the development of the research.

BIBLIOGRAPHY

BERBERIAN, C. F. Q. et al. O uso de geotecnologias como uma nova ferramenta para o controle externo. Revista do Tribunal de Contas da União, Brasília, DF, n. 133, p. 40-53, 2015.

BRAZIL. DNIT – National Transport Infrastructure Department. Central Coordination Office of Registry and Biddings. Termos de referência para o estudo do traçado e elaboração do projeto executivo de engenharia para o contorno rodoviário sul da região metropolitana de Belo Horizonte BR-040/MG. DNIT, Belo Horizonte, public notice nº 0240/2012-00, processo nº 50600.032686/2011-78, p. 79-157, 2012.

LONGLEY, P. A. et al. Sistemas e ciência da informação geográfica. 3. ed. Porto Alegre: Bookman, 2013.

NOBREGA, R. A. A. Ferrovia Norte-Sul: estudo dos traçados alternativos para escoamento eficiente da produção. In: ENCONTRO REGIONAL DA 3ª CÂMARA DE COORDENAÇÃO E REVISÃO DO MINISTÉRIO PÚBLICO FEDERAL – REGIÕES SUL/SUDESTE, 2., 2013, Florianópolis. Anais… Florianópolis, 17 out. 2013.

_____. Relatório de fechamento de projeto de extensão: ligação Pato Branco/PR – Cruz Alta/RS. Belo Horizonte: UFMG, 2014.

NOBREGA, R. A. A. et al. Bridging decision-making process and environmental needs in corridor planning. Manegement of Environmental Quality: an International Journal, Mississippi, v. 20, n. 6, p. 622-637, 2009.

SAATY, T. L. Transport planning with multiple criteria: the analytic hierarchy process applications and progress review. Journal of Advanced Transportation, New York, v. 29, n. 1, p. 81-126, 1995.

SADASIVUNI, R. et al. A Transportation corridor case study for multi-criteria decision analysis. In: ASPRS ANNUAL CONFERENCE, 2009, Maryland. Anais… Maryland, 9-13 mar. 2009.

The potential of remote sensing data in public works audit

Osmar Abílio de Carvalho Júnior
is a geologist and has a PhD in Geology from the University of Brasília (UnB) where he is a professor.

Renato Fontes Guimarães
is a cartographic engineer with a PhD in Geology. He is also a professor at the University of Brasília (UnB).

Roberto Arnaldo Trancoso Gomes
is a geographer and has a PhD in Geography from the Federal University of the State of Rio de Janeiro (UFRJ). He is also a professor at the University of Brasília (UnB).

SUMMARY

This paper analyzes the potential applications of remote sensing techniques in the audit and monitoring of infrastructure works. Currently, there is a wide availability of remote sensing data from different sensors and platforms, providing a fast and useful source of information to describe the composition of the elements present on the surface and their changes over time. The enhancement of remote sensing with increased spatial, temporal and spectral resolution on different platforms (satellites, aircraft, and unmanned aerial vehicles) has broadened the perspectives of studies and applications of remote sensing data, including the monitoring of public infrastructures in construction or concessions. The extraction of digital elevation models from remote sensors also constitutes an important attribute to describe the features of infrastructure works. The studies most commonly found in the literature are related to urban features and highways. Studies on audit and monitoring of construction works are still little reported, which constitutes a broad field of research and innovation. Several techniques of change detection have been proposed and evaluated for different environments and targets, and in each type of environment and target, they must compare and analyze precision to define the best procedure to be adopted. Specific research for each type of work should be undertaken demonstrating the real potential of remote sensing for oversight in urban or rural environments.

Keywords: Remote Sensing, change detection, digital elevation model, spectral classification.

1. INTRODUCTION

The effective oversight of public works is a key factor to minimize public expenditures. For example, audits carried out by the Fiscobras (the TCU annual plan for public works audit) in Rio Grande do Norte State during the 2011-2012 period provided savings to the public coffers of R$119,529,497.78 (SOUZA; BATISTA, 2013). The Fiscobras in 2015 carried out 97 audits, totaling R$ 31 billion in audited resources, in which 61 works (62.9%) demonstrated indications of serious irregularities (BRASIL, 2015).

However, the continental dimension of the Brazilian territory makes it difficult to conduct traditional audits, which requires constant work of professionals on-site. The growing need for infrastructures aiming at the long-term economic growth of the country combined with the high degree of irregularities makes it imperative to improve technology to inspect continuously public works in progress or under concession throughout the country (MIRANDA; MATOS, 2015; VITAL et al., 2015). During the audit, it is essential to obtain accurate information on the evolution of public works

(construction, renovation, manufacture, restoring or expansion of public property), to detect any inconsistencies or lack of elements in the basic design and technical specifications. In this context, remote sensing can be an important tool, providing periodic monitoring of large areas at low cost. In recent years, extensive research efforts have been made to change detection using remote sensing images in different scenarios: (a) urban (HEGAZY; KALOOP, 2015; SUN et al., 2013); (b) agricultural (MENKE et al., 2009; OLIVEIRA et al., 2014); and (c) natural and environmental preservation areas (COPPIN; BAUER, 1996; YADAV; KAPOOR; SARMA, 2012). Therefore, remote sensing has been widely used to assess the spatial dynamics of the earth's surface and the effectiveness of territorial planning. Specifically, studies on oversight of public works using remote sensing are not frequently reported in scientific journals. The lack of research to apply this technique, demonstrates a large field for scientific innovations, having economic relevance and immediate return to society. The collection of multitemporal images enables providing subsidies for comparing what was planned with what was executed in each work, both in the spatial and temporal dimensions. This approach allows comparing the pre-defined work schedule with the actual work in progress, allowing total or partial reconstruction of the financial and executive background.

However, the use of images to inspect public works has difficulties similar to those described in the mapping of features of urban areas. Some of them are (CAVALLI et al., 2008; WENTZ et al., 2014): (a) the spectra of the pixels of urban environments are constituted by spectral mixtures due to the high heterogeneity of these environments, and may generate confusion among the classes in the classification process (b) the physical structures of the urban classes vary spatially, taking into account different compositions of roofs, pavements, and architectural forms. Therefore, infrastructure works (buildings, bridges, roads, waterways, railways, etc.) are made of different materials (asphalt, paints, concrete, metal, glass, tiles, vegetation and soil); which are combined in various proportions (JENSEN; COWEN, 1999). For example, both the type of materials used and their architectural differences may differentiate two bridges. In addition, another factor of complexity is that during a construction work in progress an intense modification of the elements and patterns takes place. Therefore, the monitoring of works by remote sensing requires obtaining consistent and detailed information, as well as the elaboration of a specific methodology of digital image processing. Different techniques for the detection of urban features consolidated or under construction may be used. Usually, the method of visual interpretation is considered the most accurate, but it is also the most time consuming and expensive. An alternative to visual interpretation is to use supervised classifications, unsupervised classifications, and knowledge-based expert system approaches (JAT; GARG;

KHARE, 2006). In addition, change detection techniques should be adopted considering the different pre-classification or post-classification approaches.

This paper discusses the issues related to the use of remote sensing in the oversight of public works. The definition of the images to be used is evaluated considering the advances made in the improvement of spatial, spectral, temporal attributes and the extraction of altimetric data. The progress obtained with the increase of the different resolutions of the images result in new methods for the treatment and analysis of the data, with implications in the efficiency of the inspection by remote sensing. Besides, this paper reviews the main methods of change detection, which allows monitoring changes during the works and evaluating their adjustments to the initial project.

2. CHARACTERISTICS OF THE TEMPORARY, SPECTRAL AND SPATIAL RESOLUTION OF THE REMOTE SENSORS IN THE OVERSIGHT OF WORKS

2.1 SPATIAL RESOLUTION

The definition of classes related to engineering works by remote sensing and their implications in legal compliance is highly dependent on the spatial (pixel size), spectral (number of spectral bands) and temporal (revisit period of the same place on the earth's surface) resolution of an image. These three factors are important, but the image must contain high spatial resolution so that the objects in a construction work may be individualized. It is useless to have a high spectral resolution if a pixel contains different urban elements and mixed spectral behaviors. Normally, the identification of an urban object in an image must have a minimum representation of four pixels (COWEN et al., 1995; JENSEN; COWEN, 1999). According to Small (2003), the minimum spatial resolution for capturing urban structures is 5 meters, also applicable to engineering works.

High spatial resolution images in urban environments enable the use of the basic elements of interpretation (tone, color, texture, shape, size, orientation, pattern, shadow, location, and location of objects in the urban landscape) to identify and judge their meaning. Currently, different images of high spatial resolution orbital sensors (less than 4 meters) are available in the market. Among them, we highlight: GeoEye-1 (0.46m), WorldView-1 and 2 (0.46 m), WorldView-3 and 4 (0.31 m), Pleiades-1A and 1B (0.5 m), Kompsat-3A (0.55 m) and 3 (0.7 m), QuickBird (0.65), Gaofen-2 (0.8 m), TripleSat (0.8 m), Ikonos

(0.82m), SkySat-1 and 2 (0.9m) and Spot-6 and 7 (1.5 m) (Table 1). The increase in the availability of high-resolution spatial images from commercial satellites has led to the growth of digital image processing techniques for infrastructure studies, road networks, and urban elements.

A significant innovation in the mapping of urban areas using high spatial resolution images is the use of geospatial object-based image analysis (Geobia), which differs from traditional pixel-based methods. In Geobia, the image is segmented into relatively homogeneous regions (image objects) before classification (BLASCHKE, 2010; MYINT et al., 2011). Thus, the classification uses as a basic unit segments and their attributes instead of pixels. The high degree of spectral variability within a class (shadows, solar elevation angle, tree canopy gaps, etc.) may hamper pixel-based classification and favor object-based techniques that are represented by average segment values (YU et al., 2006).

A problem of object-based classification is its dependence on the segmentation stage, which can generate excessive or reduced segments of terrain features (LIU; XIA, 2010). Usually, the lack of segments is considered worse than their excess (KIM; MADDEN; WARNER, 2009). The minimization of this type of error may be obtained through successive segmentation tests prior to classification (TRIAS-SANZ; STAMON; LOUCHET, 2008) and through the analysis of segment accuracy (DORREN; MAIER; SEIJMONSBERGEN, 2003; KIM; MADDEN; XU, 2010).

Table 1:
Description of the main orbital satellites. Images available for free are marked with an asterisk (*)

AVAILABLE SATTELITES	SPATIAL RESOLUTION	TEMPORAL RESOLUTION	SPECTRAL RESOLUTION (Band)
ALOS	2,5 m	Variable	1 panchromatic
	10 m	Variable	3 visible
	10 a 100 m	Variable	1 infrared
CARTOSAT	2,5 m	5 days	1 panchromatic
CBERS*	5 m	Variable	1 panchromatic
	10 m	Variable	2 visible
			1 infrared
	20 m	26 days	3 visible
			1 infrared
	40 m	26 days	1 panchromatic
			2 infrared
	80 m	26 days	1 thermal
	64 m	5 days	3 visible
			1 infrared
EROS	0,7 m	Scheduled	1 panchromatic
FORMOSAT	2 m	Scheduled	1 panchromatic
	8 m	Scheduled	3 visible
			1 infrared
GAOFEN	0,8 m	5 days	1 panchromatic
	3,2 m	5 days	3 visible
			1 infrared
GEOEYE	0,5 m	Scheduled	1 panchromatic
	2 m	Scheduled	3 visible
			1 infrared
IKONOS	1 m	Variable	1 panchromatic
	4 m	Variable	3 visible
			1 infrared
KAZEOSAT-1	1 m	Scheduled	1 panchromatic
	4 m	Scheduled	3 visible
			1 infrared
KOMPSAT 2	1 m	Scheduled	1 panchromatic
	4 m	Scheduled	3 visible
			1 infrared
KOMPSAT 3	0,7 m	Daily (possible)	1 panchromatic
	2,8 m	Daily (possible)	3 visible
			1 infrared
KOMPSAT 3A	0,55 m	Daily (possible)	1 panchromatic
	2,2 m	Daily (possible)	3 visible
	5,5 m	Daily (possible)	2 infrared
LANDSAT 5*	30 m	16 days	3 visible
			3 infrared
	120 m	16 days	1 thermal
LANDSAT 7*	15 m	16 days	1 panchromatic
	30 m	16 days	3 visible
			3 infrared
	60 m	16 days	1 thermal
LANDSAT 8	15 m	16 days	1 panchromatic
	30 m	16 days	4 visible
			3 infrared
			1 aerosol
			1 cirrus
	100 m	16 days	2 thermal

AVAILABLE SATTELITES	SPATIAL RESOLUTION	TEMPORAL RESOLUTION	SPECTRAL RESOLUTION (Band)
PLEIADES	0,5 m	Daily (possible)	1 panchromatic
	2 m	Daily (possible)	3 visible
			1 infrared
RAPIDEYE	5 m	Scheduled	4 visible
			1 infrared
SENTINEL 1*	5 a 20 m	12 days	Radar
SENTINEL 2*	10 m	10 days with possibility of 5 days	3 visible
			1 infrared
	20 m		4 red-edge
			2 infrared
	60 m		1 aerosol
			1 cirrus
			1 water vapor
SPOT	1,5 m	Daily (possible)	1 panchromatic
	6 m	Daily (possible)	3 visible
			1 infrared
SKYSAT	1,1 m	Scheduled	1 panchromatic (possibility of creating a 90-second video)
	2 m	Scheduled	3 visible
			1 infrared
TERRASAR-X	0,25 a 40 m	Scheduled	Radar
TH-01	2 m	Scheduled	1 panchromatic
	10 m	Scheduled	3 visible
			1 infrared
TRIPLESAT	1 m	Scheduled	1 panchromatic
	4 m	Scheduled	3 visible
			1 infrared
WORLD VIEW	0,3 m	Scheduled	1 panchromatic
	1,24 m	Scheduled	1 aerosol
			4 visible
			1 red-edge
			2 infrared
	3,7 m	Scheduled	8 infrared
	30 m	Scheduled	12 Cavis (cloud, aerosol, vapor, ice and snow)
TERRA/AQUA (Sensor-MODIS)*	250 m	1-2 days	1 visible
			1 infrared
	500 m		2 visible
			3 infrared
	1000 m		7 visible
			16 infrared
			6 thermal
TERRA (Sensor ASTER)*	15 m	Variable	2 visible
			1 infrared
	30 m		6 infrared
	90 m		5 thermal

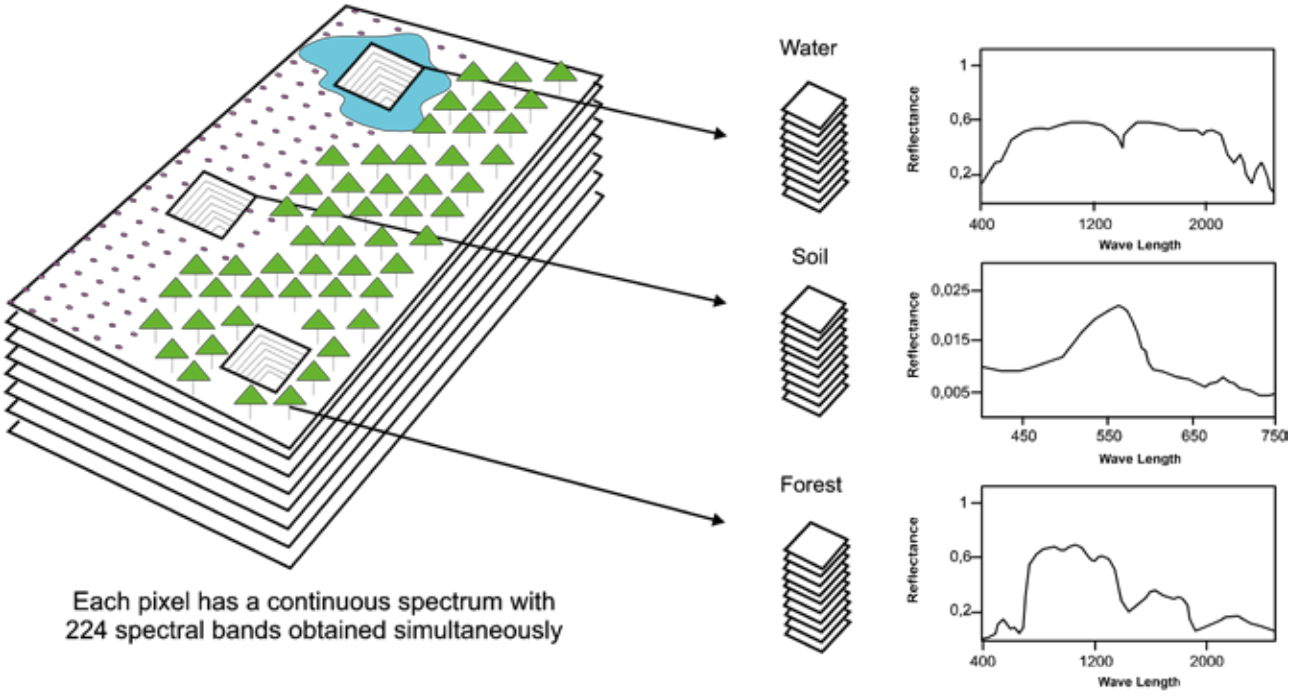
2.2 SPECTRAL RESOLUTION

The detailed knowledge of spectral characteristics enables an accurate identification of surface elements. To this end, several studies have developed specific spectral libraries for different targets from field or laboratory spectrometers to support classification, such as urban elements (BEN-DOR; LEVIN; SAARONI, 2001; HEROLD et al., 2004), mineral elements (CLARK et al., 2007), plantations (RAO; GARG; GHOSH, 2007) and flooded areas(ZOMER; TRABUCCO; USTIN, 2009).

The emergence of hyperspectral images (Figure 1), characterized by hundreds of narrow and continuous spectral bands in the reflected solar spectrum, and enabled the development of new techniques that improved the detection and quantification of materials (CARVALHO JÚNIOR et al., 2003). Spectral classification seeks to convert the spectral signals that reflect the urban coverage into categories that represent the physical nature of the surface. In addition, these images enable a subpixel spectral analysis through the estimate of abundance of the surface materials contained in pixels from the linear mixing analysis (PHINN et al., 2002; WU; MURRAY, 2003) and its Multiple Endmember Spectral Mixture Models (MESMA) variation (DEMARCHI et al., 2012; FRANKE et al., 2009). This approach makes it possible to address the problem of spectral heterogeneity within the same class, which is one of the main difficulties in urban areas (HEIDEN; SEGL; KAUFMANN, 2007). A same type of class may be characterized by several spectrally distinct materials or have differentiated compositions due to deterioration over time (DEMARCHI et al., 2012). For example, roofs may consist of different materials and, in addition, modify their spectral behavior due to the accumulation of fungi and dirt.

Normally, the most used hyperspectral images in urban studies are the Hyperion sensor onboard the EO-1 satellite and those from aircraft-embedded sensors such as Airborne Visible Infrared Imaging Spectrometer (AVIRIS), Compact Airborne Spectrographic Imager (CASI), and Hyperspectral Mapper (HyMap).

Figure 1:
Design of AVIRIS hyperspectral sensor, which high-spectral resolution makes the information of a given pixel similar to that obtained by means of laboratory and / or field measurements. (Modified from GREEN et al., 1998).



2.3 TEMPORAL RESOLUTION

The advent of orbital sensors with high temporal resolution has promoted a new approach to the classification of superficial targets, which consider aspects of cyclical changes or with known trajectories of an event or an object that may be described and identified by a temporal signature. The spectral similarity between the different types of vegetation, which are constituted by the same absorption features, may be difficult to distinguish from a single image in time. However, time series of remote sensing enable the establishment of a typical phenological signature, identifying different types of natural ecosystems (ABADE et al., 2015), phytophysiognomies (CARVALHO JÚNIOR; HERMUCHE; GUIMARÃES, 2006; CARVALHO JÚNIOR et al., 2008) and plantations (COUTO JÚNIOR; CARVALHO JÚNIOR; MARTINS, 2012; SAKAMOTO et al., 2005). This approach enables the detection of individual natural events, such as fire (CARVALHO JÚNIOR et al. 2015) and floods (AIRES et al., 2014).

In the oversight of engineering works, it is fundamental to evaluate the temporal resolution of remote sensors. Construction works evolve through a programmed schedule that enables the definition of which sensors have adequate temporal resolutions for their audit. Weather conditions (atmospheric interference and cloud coverage) also affect the acquisition of images and may be essential to optical sensors in some locations, such as in the Amazon region. In this case, high frequency revisit satellites should be sought for the selection of images or radar image sensors should be used.

High temporal resolution sensors usually have low spatial resolution, such as the MODIS sensor and the Advanced Very High Resolution Radiometer (AVHRR). However, many missions use a constellation of identical satellites that orbit the Earth in a synchronized manner, enabling a succession of high-resolution images with high spatial resolution, such as the following orbital programs: Rapideye (5 satellites); Triplesat (3 satellites); Pleiades (2 satellites) and Spot 6-7 (2 satellites). Although time series constituted with images with the same specifications facilitate the development of automated methods, studies that reconcile images of different sensors have become a research challenge.

2.4 DIGITAL ELEVATION MODELS

Digital elevation models (DEM) are 3D representations of the terrestrial surface that reproduce the natural and anthropic features of landscapes. The construction of these models starts from analog aerial photographs and is then developed by other technologies, such as: digital aerial photography, high-resolution orbital optical sensor, interferometric radar and airborne laser radar. Currently, unmanned aerial vehicles (UAVs) carry on board the different sensors that make DEMs. Recently, mobile laser radar sensors may be mounted to any platform (a boat, a car, etc.).

Optical sensors on board satellites map the Earth's surface from two different points of view, ena-

bling the extraction of DEMs. Among them are satellites with different spatial resolutions, developed by space agencies such as the Spot series, IRS, Cartosat-1, Alos-Prism, WorldView-2, QuickBird-2, Ikonos-2, Aster, among many others. Data may be acquired across a single line of scanning (Spot) or acquired in parallel with a superposition area (QuickBird) (POLI; TOUTIN, 2012). Orbital images have filled the gap in the production of accurate DEMs, but next-generation airborne cameras such as the ADS 80 now produce precision compatible with high-resolution remote sensor images such as Worldview (HOBI; GINZLER, 2012).

DEMs made by Interferometric Synthetic Aperture Radar (InSAR) are obtained from the return of the phase differences of the waves to the satellite. Among the DEMs from this method, we highlight Radarsat 1 and 2, Sentinel 1 and SRTM, and the latter is the most used worldwide (JARIHANI et al., 2015; MUSA; POPESCU; MYNETT, 2015). These products have the advantage of a continuous mapping, even with cloud coverage. **Figure 2** shows an example of an urban area DEM (A) and an aerial photograph overlapping this DEM (B).

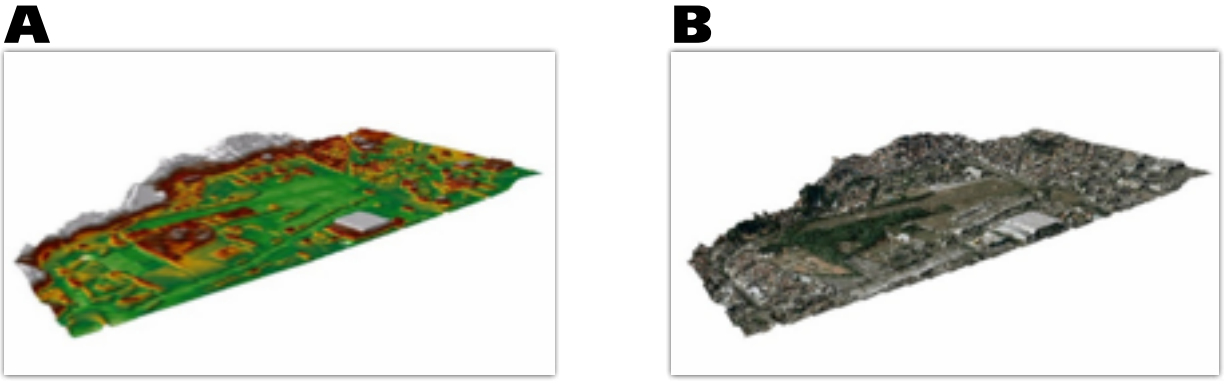
The production of DEMs from laser radar, which measures the distance between a sensor and a target based on half the time spent between pulse emission and detection (BALTSAVIAS, 1999), has already existed for a few decades. However, great technological advances in recent years have made it possible to identify subtle elevation changes from a thick cloud of points, making it possible to map and distinguish objects with small texture variations (MENG; CURRIT; ZHAO, 2010). Laser sensors may be mounted to aircraft, satellites and unmanned aerial vehicles. Recently, a mobile laser scanning system that acquires data in 2D and 3D was developed. Such data is inserted in any land or sea mobile platform (PUENTE et al., 2013). This technology has major advantages, such as: (a) cost reduction due to high-speed data capture; (b) high density of points, ensuring a complete planialtimetric survey that reduces the number of questionable data; and (c) 3D visualization that enables to verify if the mapped objects correspond to the conditions of the real world.

The DEMs are widely used to detect and describe urban features (KIM; NEVATIA, 2004). However, obstructions in dense urban environments are still a major obstacle to mapping, and methodologies are essential to cover the lack of information and insufficient texture to identify features (DURUPT; TAILLANDIER, 2006). New technologies that use laser radar, such as Light Detection and Ranging (Lidar), are being developed to reproduce in 3D high-resolution anthropic features, but their coverage is limited, and data acquisition and processing demand a high cost (BAUGH et al., 2013).

3. DIGITAL IMAGE PROCESSING AND CHANGE

Remote sensing change detection techniques may provide important information for the monitoring of engineering and infrastructure works. Some examples are (a) the area and rate at which buildings are developed; (b) the distribution and spatial relationship of types of changes by evaluating the relative performance of civil construction activities with the planned schedule; (c) the definition of the trajectory of change establishing a sequence of events and possible stoppages; and (d) the preparation of a cartographic representation that helps and shows existing spatial problems, favoring oversight

Figure 2:
urban area DEM (A) and aerial photograph overlapping this DEM (B).



actions. For this purpose, a detailed analysis should be performed concerning the existing methods, performing tests and combinations of procedures to obtain the best treatment of the data from an accuracy analysis.

3.1 PREPROCESSING

Usually, two preprocessing steps are described in change detection: (a) geometric correction; and (b) radiometric calibration (ALMUTAIRI; WARNER, 2010). The geometric correction transform the image coordinate system in a spatial coordinate system. This process enables to compare pixels values over time, eliminating the existence of some systematic distortion. Incorrect geometric registration impairs the accuracy of change detection, generating false artificial artifacts that do not match surface features. Thus, the images must have a perfect overlap, with a root mean square error of at most 0.2 pixels to reach an error of only 10% (TOWNSHEND et al., 1992; DAI; KHORRAM, 1999).

Radiometric calibration removes changes caused by external factors, such as: changes in sensor calibration over time, sun elevation angle, variability in Earth-Sun distance, and atmospheric interference. These methods are subdivided into two types: absolute and relative.

Absolute radiometric correction uses radiative transfer codes to obtain reflectance on the Earth's surface, eliminating atmospheric interference. Key atmospheric interferences derive from two effects: (a) scattering (diffusion or dispersion) that changes the direction of propagation of solar radiation by elastic interaction with particulates, mainly aerosols; and (b) atmospheric absorption, with effective loss of energy at specific wavelengths, mainly from seven gases: water vapor (H2O), carbon dioxide (CO2), ozone (O3), nitric oxide (N2O), carbon monoxide (CO), methane (CH4) and oxygen (O2) (ZULLO JÚNIOR, 1994). However, the correction of the atmospheric effects is more effective using hyperspectral sensors that contain bands exclusive to the different gases and makes a correction for each pixel. In relation to multispectral images, constant values are adopted for the whole image.

Relative corrections normalize digital values to a common scale, in which invariant features between two images are adjusted to a single reference, assuming that these pixels are linearly correlated (DU; TEILLET; CIHLAR, 2002). Therefore, the central question is to obtain the invariant features to conduct the linear regression between temporal images, and they may be determined by visual or computational inspection. Several

automated methods were proposed to find the invariant points and to normalize the images. We highlight the following methods: robust linear regression (HEO; FITZHUGH, 2000) frequency density diagram (SONG et al., 2001; CHEN; VIERLING; DEERING, 2005); distance and similarity spectral measures (CARVALHO JÚNIOR et al., 2013), spectral measurements among canonical components (CANTY; NIELSEN; SCHMIDT, 2004); and non-change zones around the regression line (ELVIDGE et al., 1995) or main components (DU et al., 2002). With the purpose of agglutinating all these techniques, Carvalho Júnior et al. (2013) proposed a sequential method to determine invariant points composed of the following methods: (a) spectral measurements in the original temporal images or in the canonical components; (b) density of the dispersion diagram; and (c) robust regression. The current method is in the Abílio program.

3.2 CHANGE DETECTION

Different methods have been proposed to change detect from remote sensing images, with important bibliographical reviews on the subject (COPPIN et al., 2001; HALL; HAY, 2003; LAMBIN, 1999; LU et al., 2003; SINGH, 1989; TEWKESBURY et al., 2015).

Change detection algorithms are composed of two processing steps: (a) classification; and (b) change detection. A classification of the change detection methods considers whether the change detection step comes before or after the classification step. They are called (YUAN et al., 2005): (a) pre-classification, in whi-

ch a new image highlighting the change features is created and then the classification stage is performed; and (b) post-classification, in which an independent classification is initially performed for each period and then extraction and quantification of changing areas are performed from the cross-tab between temporal images.

Many methods of pre-classification were proposed: (a) the use of algebraic techniques for subtracting and dividing multitemporal images (COPPIN et al., 2001; FRANKLIN et al., 2003; SKAKUN; WULDER; FRANKLIN, 2003); (b) change vector analysis (CARVALHO JÚNIOR et al., 2011; JOHNSON; KASISCHKE, 1998); (c) spectral mixing (ADAMS et al., 1995); and (d) several linear transformations, such as principal components analysis (BYRNE et al., 1980; FUNG e LEDREW, 1987), correspondence analysis (CAKIR et al., 2006), canonical analysis (NIELSEN; CONRADSEN; SIMPSON, 1998) and Tasseled-Cap (HEALEY et al., 2005). Pre-classification methods, although effective in locating changes, are often difficult to identify the nature of the change, thus needing another stage in classification.

Thus, the post-classification method is the most widely used method in change detection studies in urban environments, because it is effective in describing the magnitude, location and nature of the changes that have occurred (HARDIN; JACKSON; OTTERSTROM, 2007). The main advantages of the post-classification method are: (a) an independence in the classification process of temporal images compensates the variations in atmospheric conditions, phenological changes and soil moisture; (b) the process of updating data is simple, benefiting monitoring; (c) it enables comparing sensor data with different types of resolutions; and (d) enables individualizing the different categories of change, not restricted to highlighting the features of changes (COPPIN et al., 2001; MENKE et al., 2009). In contrast, the two main disadvantages of this method are: (a) it usually is not fully automatic, which makes the method time-consuming; and (b) the precision of change detection depends on the accuracy of the classification at each period, and this may facilitate the propagation of errors (YUAN et al., 2005; MENKE et al., 2009).

According to Silva et al. (2012), direct classification of the various spectrum-temporal bands does not fit into the post-classification and pre-classification methods, since the stages of classification and change detection are synchronized. In this type of classification, group analysis (WEISMILLER et al., 1977) and artificial neural networks (DAI; KORRAM, 1999) are normally used.

4. CONCLUSIONS

This paper reviewed the main potentialities and challenges regarding the use of remote sensing in the oversight of works within TCU activities, focusing on the following aspects: (a) quality of image attributes; (b) steps of digital image processing; and (c) possible adjustments and efforts required to quantify and understand the stages of public works. Auditing public works through remote sensing is a complex process, with some degree of interference between classes and a strong component of spatial-temporal changes. Only

a continuous representation in the time of the construction work elements enable precise analysis and oversight. For this purpose, repetitive measurements of the spectral and spatial components of the Earth's surface should preferably be obtained in high resolution. Each attribute provides a specific type of information about the construction work and should be combined for a detailed description of the surface processes. Different models of change detection should be tested, considering the conditions of the surroundings and the environments.

Oversight and monitoring from geoprocessing and remote sensing techniques enable monitoring different works simultaneously, virtually in real time. This new approach requires a set of methodological investigations in order to reinforce the relationship between the original design of the works and the magnitude of change detected in the image. The advance of expert models to detect automated or semi-automated change in public works will enable the establishment of an alert system focusing on field inspection. To this end, computational research efforts should be employed to develop a set of standard detection techniques that will enable improving the management of the construction work phases, reducing uncertainties. This technological arsenal adapted to the different targets will allow the definition of strategies to curb the action of possible fraud or delays in schedule.

The key factor for successful mapping of engineering works is to use high-resolution images (spectral, spatial and temporal) and the DEM integrated with the intended response described by the executive project, containing all factors specific to the activity. Differently, from other studies of remote sensing in urban areas, this case has previously provided the spatial location and the intended changes, enabling a new approach to the development of automated techniques considering a previous model.

REFERENCES

ABADE, N. A. et al. Comparative Analysis of MODIS Time-Series Classification Using Support Vector Machines and Methods Based upon Distance and Similarity Measures in the Brazilian Cerrado-Caatinga Boundary. Remote Sensing, [S.l.], v. 7, n. 9, p. 12160-12191, 2015.

ADAMS, J. B. et al. Classification of Multispectral Images Based on Fractions of Endmembers: Application to Land-Cover Change in the Brazilian Amazon. Remote Sensing of Environment, [S.l.], v. 52, p. 137-154, 1995.

AIRES, F. et al. Characterization and Space-Time Downscaling of the Inundation Extent over the Inner Niger Delta Using GIEMS and MODIS Data. Journal of Hydrometeorology, Washington, DC, v. 15, n. 1, p. 171-192, 2014.

ALMUTAIRI, A.; WARNER, T. A. Change Detection Accuracy and Image Properties: A Study Using Simulated Data. Remote Sensing, [S.l.], v. 2, p. 1508-1529, 2010.

BALTSAVIAS, E. P. Airborne Laser Scanning: Basic Relations and Formulas. ISPRS Journal of Photogrammetry and Remote Sensing, [S.l.], v. 54, p. 199-214, 1999.

BAUGH, C. A. et al. SRTM Vegetation Removal and Hydrodynamic Modeling Accuracy. Water Resource Research, Washington, DC, v. 49, p. 5276-5289, 2013.

BEN-DOR, E.; LEVIN, N.; SAARONI, H. A Spectral Based Recognition of the Urban Environment Using the Visible and Near-Infrared Spectral Region (0.4-1.1 μm): A Case Study Over Tel-Aviv. International Journal of Remote Sensing, Londres, v. 22, n. 11, p. 2193-2218, 2001.

BLASCHKE, T. Object Based Image Analysis for Remote Sensing. ISPRS Journal of Photogrammetry and Remote Sensing, [S.l.], v. 65, n. 1, p. 2-16, 2010.

BRASIL. Tribunal de Contas da União. Fiscobras 2015. 2015. Retrieved from <http://portal.tcu.gov.br/lumis/portal/file/fileDownload.jsp?fileId=8A8182A250C885960150CD7694B146CC&inline=1>. Accessed October 20, 2016.

BYRNE, G. F.; CRAPPER, P. F.; MAYO, K. K. Monitoring Land Cover Change by Principal Component Analysis of Multitemporal Landsat Data. Remote Sensing of Environment, [S.l.], v. 10, p. 175-184, 1980.

CAKIR, H. I.; KHORRAM, S.; NELSON, S. A. C. Correspondence Analysis for Detecting Land Cover Change. Remote Sensing of Environment, [S.l.], v. 102, n. 3-4, p. 306-317, 2006.

CANTY, M. J.; NIELSEN, A. A.; SCHMIDT, M. Automatic Radiometric Normalization of Multitemporal Satellite Imagery. Remote Sensing of Environment, [S.l.], v. 91, p. 441-451, 2004.

CARVALHO JÚNIOR, A. O. et al. A New Approach to Change Vector Analysis Using Distance and Similarity Measures. Remote Sensing, [S.l.], v. 3, p. 2473-2493, 2011.

_____. Análise de imagens hiperespectrais pelo método Multiple Endmember Spectral Mixture Analysis (MESMA) em depósito supergênico de níquel. Brazilian Journal of Geology, São Paulo, v. 33, n. 1, p. 63-74, 2003.

_____. Classificação de padrões de savana usando assinaturas temporais NDVI do sensor MODIS no Parque Nacional Chapada dos Veadeiros. Revista Brasileira de Geofísica, [S.l.], v. 26, n. 4, p. 505-517, 2008.

_____. Standardized Time-Series and Interannual Phenological Deviation: New Techniques for Burned-Area Detection Using Long-Term MODIS-NBR Dataset. Remote Sensing, [S.l.], v. 7, n. 6, p. 6950-6985, 2015.

_____. Radiometric Normalization of Temporal Images Combining Automatic Detection of Pseudo-Invariant Features from the Distance and Similarity Spectral Measures, Density Scatterplot Analysis, and Robust Regression. Remote Sensing, [S.l.], v. 5, n. 6, p. 2763-2794, 2013.

CAVALLI, R. M. et al. Hyperspectral Sensor Data Capability for Retrieving Complex Urban Land Cover in Comparison with Multispectral Data: Venice City Case Study (Italy). Sensors, [S.l.], v. 8, n. 5, p. 3299-3320, 2008.

CHEN, X.; VIERLING, L.; DEERING, D. A Simple and Effective Radiometric Correction Method to Improve Landscape Change Detection across Sensors and across Time. Remote Sensing of Environment, [S.l.], v. 98, p. 63-79, 2005.

CLARK, R. N. et al. USGS Digital Spectral Library Splib06a. US Geological Survey, Digital Data Series, [S.l.], n. 231, 2007.

COPPIN, P. et al. Operational Monitoring of Green Biomass Change for Forest Management. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 67, p. 603-611, 2001.

COPPIN, P. R.; BAUER, M. E. Digital Change Detection in Forest Ecosystems with Remote Sensing Imagery. Remote Sensing Reviews, Londres, v. 13, n. 3-4, p. 207-234, 1996.

COUTO JÚNIOR, A. F.; CARVALHO JÚNIOR, O. A.; MARTINS, E. S. Séries temporais MODIS aplicadas em sucessão de culturas de soja (Glycine max (L.) Merrill) e milho (Zeamays L.) em sistema de plantio direto. Revista Brasileira de Cartografia, Rio de Janeiro, v. 64, n. 3, p. 405-418, 2012.

COWEN, D. J. et al. The Design and Implementation of an Integrated Gis for Environmental Applications. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 61, p. 1393-1404, 1995.

DAI, X. L.; KHORRAM, S. Remotely Sensed Change Detection Based on Artificial Neural Networks. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 65, n. 10, p. 1187-1194, 1999.

_____. The Effects of Image Misregistration on the Accuracy of Remotely Sensed Change Detection. IEEE Transactions on Geoscience and Remote Sensing, [S.l.], v. 36, p. 1566-1577, 1998.

CARVALHO JÚNIOR, O. A., HERMUCHE, P. M., GUIMARÃES, R. F. Identificação regional da floresta estacional decidual na bacia do Rio Paranã a partir da análise multitemporal de imagens MODIS. Revista Brasileira de Geofísica, [S.l.], v. 24, n. 3, p. 319-332, 2006.

DEMARCHI, L. et al. Multiple Endmember Unmixing of CHRIS/Proba Imagery for Mapping impervious Surfaces in Urban and Suburban Environments. IEEE Transactions on Geoscience and Remote Sensing, [S.l.], v. 50, n. 9, p. 3409-3424, 2012.

DORREN, L.; MAIER, B.; SEIJMONSBERGEN, A. Improved Landsat-Based Forest Mapping in Steep Mountainous Terrain Using Object-Based Classification. Forest Ecology and Management, [S.l.], v. 183, n. 1-3, p. 31-46, 2003.

DU, Y.; TEILLET, P. M.; CIHLAR, J. Radiometric Normalization of Multitemporal High-Resolution Satellite Images with Quality Control for Land Cover Change Detection. Remote Sensing of Environment, [S.l.], v. 82, p. 123-134, 2002.

DURUPT, M.; TAILLANDIER, F. Automatic Building Reconstruction from a Digital Elevation Model and Cadastral Data: An Operational Approach. 2006. Retrieved from <http://www.isprs.org/proceedings/XXXVI/part3/singlepapers/O_14.pdf>. Accessed November 25, 2016.

ELVIDGE, C. D. et al. Relative Radiometric Normalization of Landsat Multispectral Scanner (MSS) Data Using an Automatic Scattergram-Controlled Regression. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 61, p. 1255-1260, 1995.

FRANKE, J. et al. Hierarchical Multiple Endmember Spectral Mixture Analysis (MESMA) of Hyperspectral Imagery for Urban Environments. Remote Sensing of Environment, [S.l.], v. 113, n. 8, p. 1712-1723, 2009.

FRANKLIN, S. E. et al. Mountain Pine Beetle Red-Attack Forest Damage Classification Using Stratified Landsat TM Data in British Columbia, Canada. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 69, n. 3 p. 283-288, 2003.

FUNG, T.; LE DREW, E. Application of Principal Components Analysis to Change Detection. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 53, n. 12, p. 1649-1658, 1987.

GREEN, R. O. et al. Imaging Spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). Remote Sensing of Environment, [S.l.], v. 65, p. 227-248, 1998.

HALL, O.; HAY, G. J. A Multiscale Object-Specific Approach to Digital Change Detection. International Journal of Applied Earth Observation and Geoinformation, [S.l.], v. 4, p. 311-327, 2003.

HARDIN, P. J.; JACKSON, M. W.; OTTERSTROM, S. M. Mapping, Measuring, and Modeling Urban Growth. In: JENSEN, R. R.; GATRELL J. D.; MCLEAN D. (Org.). Geo-Spatial Technologies in Urban Environments: Policy, Practice and Pixels. 2. ed. Heidelberg: Springer-Verlag, 2007. p. 141-176.

HEALEY, S. P. et al. Comparison of Tasseled Cap-based Landsat Data Structures for Use in Forest Disturbance Detection. Remote Sensing of Environment, [S.l.], v. 97, n. 3, p. 301-310, 2005.

HEGAZY, I. R.; KALOOP, M. R. Monitoring Urban Growth and Land Use Change Detection with GIS and Remote Sensing Techniques in Daqahlia Governorate Egypt. International Journal of Sustainable Built Environment, [S.l.], v. 4, p. 117-124, 2015.

HEIDEN, U.; SEGL, K.; KAUFMANN, H. Determination of Robust Spectral Features for Identification of Urban Surface Materials in Hyperspectral Remote Sensing Data. Remote Sensing of Environment, [S.l.], v. 111, n. 4, p. 537-552, 2007.

HEO, J.; FITZHUGH, T. W. A Standardized Radiometric Normalization Method for Change Detection Using Remotely Sensed Imagery. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 66, n. 2, p. 173-181, 2000.

HEROLD, M. et al. Spectrometry for Urban Area Remote Sensing – Development and Analysis of a Spectral Library from 350 to 2400 nm. Remote Sensing of Environment, [S.l.], v. 91, n. 3-4, p. 304-319, 2004.

HOBI, M. L.; GINZLER, C. Accuracy Assessment of Digital Surface Models Based on WorldView-2 and ADS80 Stereo Remote Sensing Data. Sensors, [S.l.], v. 12, p. 6347-6368, 2012.

JARIHANI, A. A. et al. Satellite-Derived Digital Elevation Model (DEM) Selection, Preparation and Correction for Hydrodynamic Modelling in Large, Low-Gradient and Data-Sparse Catchments. Journal of Hydrology, [S.l.], v. 524, p. 489-506, 2015.

JAT, M. K.; GARG, P. K.; KHARE, D. Monitoring and Modelling of Urban Sprawl Using Remote Sensing and GIS Techniques. International Journal of Applied Earth Observation and Geoinformation, [S.l.], v. 10, n. 1, p. 26-43, 2008.

JENSEN, J. R. et al. An Evaluation of Coastwatch Change Detection Protocol in South Carolina. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 59, n. 4, p. 519-525, 1993.

JENSEN, J. R.; COWEN, D. C. Remote Sensing of Urban/ Suburban Infrastructure and Socio-Economic Attributes. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 65, p. 611-622, 1999.

JOHNSON, R. D.; KASISCHKE, E. S. Change Vector Analysis: A Technique for the Multispectral Monitoring of Land Cover and Condition. International Journal of Remote Sensing, London, v. 19, p. 3, n. 411-426, 1998.

KIM, M.; MADDEN, M.; XU, B. GEOBIA Vegetation Mapping in Great Smoky Mountains National Park with Spectral and Non-Spectral Ancillary Information. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 76, n. 2, p. 137-149, 2010.

KIM, M.; MADDEN, M.; WARNER, T. Forest Type Mapping Using Object-Specific Texture Measures from Multispectral IKONOS Imagery: Segmentation Quality and Image Classification Issues. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 75, n. 7, p. 819-830, 2009.

KIM, Z. W.; NEVATIA, R. Automatic Description of Complex Buildings from Multiple Images. Computer Vision Image Understanding, v. 96, p. 60-95, 2004.

LAMBIN, E. F. Monitoring Forest Degradation in Tropical Regions by Remote Sensing: Some Methodological Issues. Global Ecology and Biogeography, [S.l.], v. 8, n. 3-4, p. 191-198, 1999.

LIU, D.; XIA, F. Assessing Object-Based Classification: Advantages and Limitations. Remote Sensing Letters, Melbourne, v. 1, n. 4, v. 187-194, 2010.

LU, D. et al. Change Detection Techniques. International Journal of Remote Sensing, Londres, v. 25, n. 12, p. 2365-2407, 2003.

MENG, X.; CURRIT, N.; ZHAO, K. Ground Filtering Algorithms for Airborne LiDAR Data: A Review of Critical Issues. Remote Sensing, [S.l.], v. 2, p. 833-860, 2010.

MENKE, A. B. et al. Análise das mudanças do uso agrícola da terra a partir de dados de sensoriamento remoto multitemporal no município de Luis Eduardo Magalhães (BA-Brasil). Sociedade & Natureza, Uberlândia, v. 21, n. 3, p. 315-326, 2009.

MIRANDA, A. C. O.; MATOS; C. R. Potencial uso do BIM na fiscalização de obras públicas. Revista do Tribunal de Contas da União, Brasília, DF, v. 133, p. 22-31, 2015.

MUSA, Z. N.; POPESCU, I.; MYNETT, A. A Review of Applications of Satellite SAR, Optical, Altimetry and DEM Data for Surface Water Modelling, Mapping and Parameter Estimation. Hydrology and Earth System Sciences, [S.l.], v. 19, p. 3755-3769, 2015.

MYINT, S. et al. Per-Pixel vs. Object-Based Classification of Urban Land Cover Extraction Using High Spatial Resolution Imagery. Remote Sensing of Environment, [S.l.], v. 115, n. 5, p. 1145-1161, 2011.

NIELSEN, A. A.; CONRADSEN, K.; SIMPSON, J. J. Multivariate Alteration Detection (MAD) and MAF Postprocessing in Multispectral, Bitemporal Image Data: New Approaches to Change Detection Studies. Remote Sensing of Environment, [S.l.], v. 64, p. 1-19, 1998.

OLIVEIRA, S. N et al. Detecção de mudança do uso e cobertura da terra usando o método de pós-classificação na fronteira agrícola do oeste da Bahia sobre o Grupo Urucuia durante o período 1988-2011. Revista Brasileira de Cartografia, Rio de Janeiro, v. 66, n. 5, p. 1157-1176, 2014.

PHINN, S. R. et al. Monitoring the Composition of Urban Environments Based on the Vegetation-Impervious Surface-Soil (VIS) Model by Subpixel Analysis Techniques. International Journal of Remote Sensing, London, v. 23, n. 20, p. 4131-4153, 2002.

POLI, D.; TOUTIN, T. Review of Developments in Geometric Modelling for High-Resolution Satellite Pushbroom Sensors. Photogrammetric Record, [S.l.], v. 27, p. 58-73, 2012.

PUENTE, I. et al. Review of Mobile Mapping and Surveying Technologies. Measurement: Journal of the International Measurement Confederation, [S.l.], v. 46, n. 7, p. 2127-2145, 2013.

RAO, N. R.; GARG, P. K.; GHOSH, S. K. Development of an Agricultural Crops Spectral Library and Classification of Crops at Cultivar Level Using Hyperspectral Data. Precision Agriculture, [S.l.], v. 8, n. 4-5, p. 173-185, 2007.

SAKAMOTO, T. et al. A Crop Phenology Detection Method Using Time-Series MODIS Data. Remote sensing of environment, [S.l.], v. 96, n. 3, p. 366-374, 2005.

SINGH, A. Digital Change Detection Techniques Using Remotely-Sensed Data. International Journal of Remote Sensing, London, v. 10, p. 989-1003, 1989.

SILVA, N. C. et al. Change Detection Software Using Self-Organizing Feature Maps. Revista Brasileira de Geofísica, [S.l.], v. 30, n. 4, p. 505-518, 2012.

SKAKUN, R. S.; WULDER, M. A.; FRANKLIN, S. E. Sensitivity of the Thematic Mapper Enhanced Wetness Difference Index to Detect Mountain Pine Beetle Red-Attack Damage. Remote Sensing of Environment, [S.l.], v. 86, p. 433-443, 2003.

SMALL, C. High Resolution Spectral Mixture Analysis of Urban Reflectance. Remote Sensing of Environment, [S.l.], v. 88, p. 170-186, 2003.

SONG, C. et al. Classification and Change Detection Using Landsat TM Data: When and How to Correct Atmospheric Effects? Remote Sensing of Environment, [S.l.], v. 75, p. 230–244, 2001.

SOUZA, I. V. N.; BATISTA, H. M. Estudo dos benefícios econômicos gerados pelas fiscalizações de obras públicas, realizadas pelo Tribunal de Contas da União, no estado do Rio Grande do Norte, no período de 2011 e 2012. 2013. 67 f. (Trabalho de Conclusão de Curso) – Departamento de Ciências Contábeis da Universidade Federal do Rio Grande do Norte, Natal, 2013.

SUN, C. et al. Quantifying Different Types of Urban Growth and the Change Dynamic in Guangzhou Using Multi-Temporal Remote Sensing Data. International Journal of Applied Earth Observation and Geoinformation, [S.l.], v. 21, p. 409-417, 2013.

TEWKESBURY, A. P. et al. A Critical Synthesis of Remotely Sensed Optical Image Change Detection Techniques. Remote Sensing of Environment, [S.l.], v. 160, p. 1-14, 2015.

THOMAS, N.; HENDRIX, C.; CONGALTON, R. G. A Comparison of Urban Mapping Methods Using High-Resolution Digital Imagery. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 69, n. 9, p. 963-972, 2003.

TOWNSHEND, J. R. G. et al. The Impact of Misregistration on Change Detection. IEEE Transactions on Geoscience and Remote Sensing, [S.l.], v. 30, n. 5, p. 1054-1060, 1992.

TRIAS-SANZ, R., STAMON, G., LOUCHET, J. Using Colour, Texture, and Hierarchical Segmentation for High-Resolution Remote Sensing. ISPRS Journal of Photogrammetry and Remote Sensing, [S.l.], v. 63, n. 2, p. 156-168, 2008.

VITAL, A. L. F. et al. Fiscobras: uma obra em construção. Revista do Tribunal de Contas da União, Brasília, DF, v. 133, p. 32-39, 2015.

WEISMILLER, R. A. et al. Change Detection in Coastal Zone Environments. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 43, p. 1533-1539, 1977.

WENTZ, E. A. et al. Supporting Global Environmental Change Research: A Review of Trends and Knowledge Gaps in Urban Remote Sensing. Remote Sensing, [S.l.], v. 6, n. 5, p. 3879-3905, 2014.

WU, C.; MURRAY, T. A. Estimating Impervious Surface Distribution by Spectral Mixture Analysis. Remote Sensing of Environment, [S.l.], v. 84, n. 4, p. 493-505, 2003.

YADAV, P. K.; KAPOOR, M.; SARMA, K. Land Use Land Cover Mapping, Change Detection and Conflict Analysis of Nagzira-Navegaon Corridor, Central India Using Geospatial Technology. International Journal of Remote Sensing and GIS, Deli, v. 1, n. 2, p. 90-98, 2012.

YU, Q. et al. Object-Based Detailed Vegetation Classification with Airborne High Spatial Resolution Remote Sensing Imagery. Photogrammetric Engineering and Remote Sensing, Bethesda, v. 72, n. 7, p. 799-811, 2006.

YUAN, F. et al. Land Cover Classification and Change Analysis of the Twin Cities (Minnesota) Metropolitan Area by Multitemporal Landsat Remote Sensing. Remote Sensing of Environment, [S.l.], v. 98, n. 2, p. 317-328, 2005.

ZOMER, R. J.; TRABUCCO, A.; USTIN, S. L. Building Spectral Libraries for Wetlands Land Cover Classification and Hyperspectral Remote Sensing. Journal of Environmental Management, [S.l.], v. 90, n. 7, p. 2170-2177, 2009.

ZULLO JÚNIOR, J. Correção atmosférica de imagens de satélite e aplicações. 1994. 191 p. Tese (Doutorado em Engenharia Elétrica) – Universidade Estadual de Campinas, Campinas, 1994.

Implementation of a geocatalogue to assist location and recovery of open geographic data

Drausio Gomes dos Santos is an employee of the Federal Court of Accounts – Brazil. He has an M.A. in Applied Computing from the University of Brasília (UnB) and a specialist degree in IT Management from the Federal University of the State of Pernambuco (Cin/UFPE) and a technological degree in Data Processing from the São Paulo Technology College (FATEC/SP) and a B.A. in Geography from the State University of São Paulo (USP).

Alexandre Zaghetto has an M. A. and a PhD in Electric Engineering from the University of Brasília (UnB). He has a B.A. in Electric Engineering, focused on Electronics, from the Federal University of the State of Rio de Janeiro (UFRJ) He is also an adjunct professor of the Computer Science Department of UnB and leader of the Research Group on Biometric Systems

ABSTRACT

This paper approaches the creation of a geocatalogue, which uses semantics and ontologies to assist in the process of discovering geographic resources in services of an open spatial data infrastructure. Its architecture incorporates a non-relational graphs-oriented database and it is subject to implementation in cloud computing. To conceive it, the SERVUS systems development methodology, specialized in architecture oriented to geospatial services, was used.

Keywords: Software Engineering. Geocatalogue. SERVUS. SDI. Semantics. Open data.

1. INTRODUCTION

It has been a challenge to the public administration bodies to obtain a spatial view of the result of governmental actions in a simplified and updated way, as data are decentralized, produced by several sources and in different moments. The Government have taken some actions in order to standardize, optimize resources and integrate data, such as the creation of the National Spatial Data Infrastructure of Brazil (NSDI – BRA).

The NSDI-BRA is an initiative of creating an architecture for dissemination and regulation of use of geogra-

phic data within the public administration. We highlight one of the goals set in its conception, which is to

> "avoid the duplicity of actions and the waste of resources in obtaining aerospatial data by public administration bodies, through the dissemination of metadata related to those data available at the public entities and bodies of the federal, state, district and municipal spheres" (BRASIL, 2008).

Camboim (2013) highlights the governmental initiative of creating the National Open Data Infrastructure (NODI) and the Brazilian Open Data Portal as a strategy to adop the Linked Open Data. This author proposes an architecture capable of making data in the NSDI-BRA available in the NODI in an integrated manner. Such proposal of architecture stands out due to the use of semantic layers and ontologies by joining the semantic web techniques and geospatial data, entering the geosemantics field.

Architectures from the current SDI are, as a rule, service-oriented and they adopt technological patterns established by the Open Geospatial Consortium (OGC). Some of these patterns define the operation of geoservices, which allow the collection of cartographic maps, thematic maps, geographic data and metadata.

One of the essential compounds in an SDI architecture is the services catalogue. The SERVUS software development methodology (USLÄNDER, 2010), specially created to deal with the use and composition of geospatial services, proposes the use of a semantic geocatalogue to assist in locating geographic network connected resources.

To study this aspect, this document is divided in the following way: Section 2 presents some correlated works; Section 3 presents the theoretical foundation and addresses SDI and SERVUS methodology; Section 4 presents the proposed model, details of the geocatalogue and an example of use with methodology SERVUS and finally, Section 5 presents the conclusions.

2. CORRELATED WORKS

Andrade and Baptista (2011) followed a path that included semantic search, listing great amounts of information from several sources, establishment of metrics and generation of a range of resources available in a spatial data infrastructure. The authors assumed that the use of ontologies would be the way to make the searches more precise and to facilitate the automation process. As contributions, we have had the establishment of a combined metric and the creation of a semantic network.

Daltio and Carvalho (2012) proposed the creation of a framework for semantic recovery of spatial data. Such framework was based on a process

of semantic annotation of the geographic resources and use of an ontology management service. In this paper, it is worth noting the importance given to the process of selection of the ontologies.

Gimenez, Tanaka and Baião (2013) present a proposal of semantic integration for the NSDI-BRA by using geo-ontologies of domain incorporated in semantic layers acting on the SDIs.

3. THEORETICAL FOUNDATION

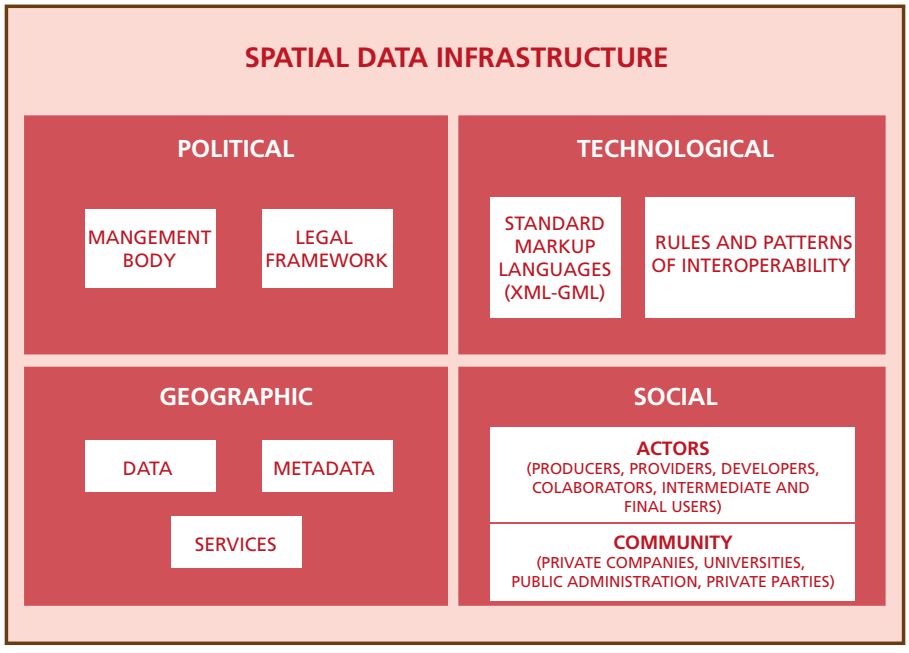
3.1 SPATIAL DATA INFRASTRUCTURE (SDI)

Bernabé-Proveda and López Vázquez (2012, p. 57-59) conceptualize Spatial Data Infrastructure (SDI) as an infrastructure necessary to access, share, exchange, match and analyze geographic data in a standardized and interoperable way. They have also considered the need of those data being available in a network, through a set of systems which use standard protocols and interfaces to promote the creation of applications that can be seen by users as a single system. The SDI can be classified as a structure of technological, geographic, social and political components as illustrated in Figure 1.

The technological component is represented by an architecture based on patterns of interoperability capable of sharing geographic data and information. The markup languages XML and GML have a key role in this component. The social component is represented by a set of actors, amongst them, data producers, service providers, users, software developers and those responsible for patterns and rules, besides a great community composed by private companies, government, universities and society as a whole. The geographic component is represented by data, their metadata and geoservices. The political component is represented by the persons and by the body responsible for establishing the regulation framework and its operating rules.

OGC Web Services Common Standard is a technological standard that proposes a common interface for a set of geoservices: Web Map Service (WMS), Web Feature Service (WFS) and Web Coverage Service (WCS). In this proposal, the intention is to standardize the ways of access, operations, mandatory and optional parameters, besides data structures, aiming at reducing the efforts of interoperability. One of the operations defined is the GetCapabilities, which recovers the metadata re-

Figure 1:
Components of an SDI



lated to the capabilities of the geoservices (GREENWOOD; WHITESIDE, 2010).

WMS consists in a service that produces dynamic maps, in which geographic information are organized in layers. The WFS service deals with geographic entities with discrete or vector data in GML format, representing attributes and geometries. WCS is the service that supports the recovery of spatial data as coverages.

A catalogue of services allows the location of data or geographic services through a range of operations, amongst them, the GetRecords operation, which recovers a set of records of metadata (OGC, 2016).

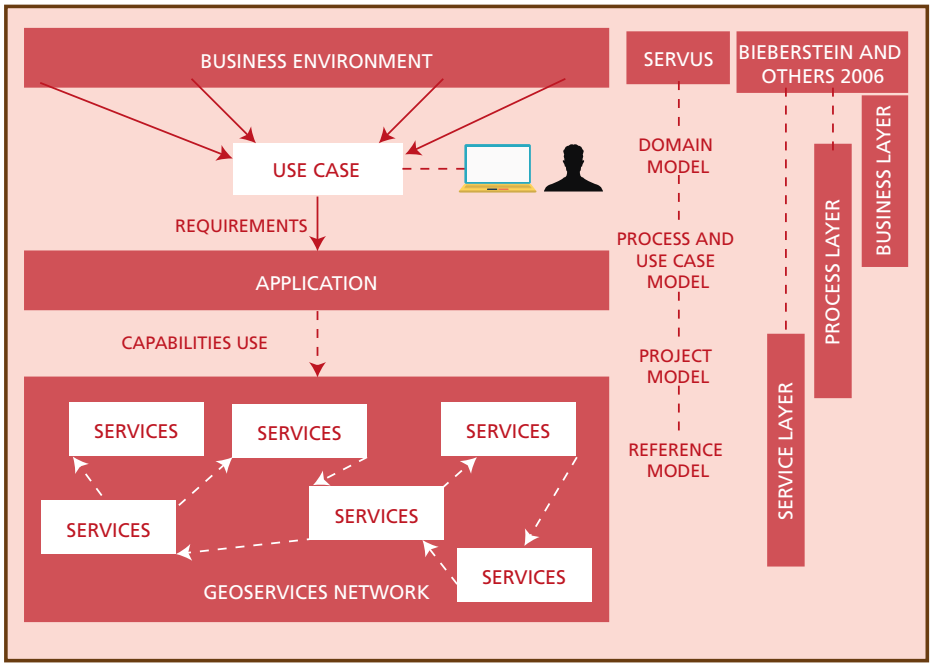
3.2 SERVUS DEVELOPMENT METHODOLOGY

The acronym SERVUS originates from two software engineering terms: "SERvice" and "USe case". The central problem such methodology proposes to solve is – considering a set of use cases which meet the business requirements and demand capacity of resources in a network of geospatial services – how to discover and associate required resources to offered resources, composing an application that supplies quality, functional and informational requirements. Two other elements compose this scenery: semantic descriptions in a network of semantic resources, and side conditions to the development of environmental systems (processes of discovery of resources, the use of OGC patterns, resources matching) (USLANDER, 2010).

According to Uslander (2010, p.90), "SERVUS understands the environmental information systems project as an interactive discovery and matching activity: available capabilities are discovered and associated to the users requirements formulated as use cases".

A modeling language, a development process and a reference architecture compose methodology. During the development process, artifacts are created, such as a domain model, a process/use case model and a project model. The main model is the project model and it expresses the requirements and capabilities in the form of resources. The generated models are based on the abstraction layers proposed by Bieberstein (2006). Figure 2 displays the mapping between requirements and capabilities, the models produced and the relation with the abstraction layers.

Figure 2:
SERVUS model hierarchy



Source: Adapted from Uslander (2010, p.92)

The domain model of the SERVUS methodology represents the thematic domain of the problem to be solved. It formally defines the part of the world that constitutes the speech universe between the user and the system designer, that is, it comprises the shared knowledge on the application domain. Typically, such shared knowledge are represented by the specification of an ontology (USLANDER, 2010).

The reference model is composed by an architectural framework, responsible for orientation and rules on how to specify the system, and by a conceptual model. The SERVUS conceptual model is a metamodel in accordance with the precepts of Model Driven Architecture (MDA) and Meta-Object Facility (MOF). The main metaclasses are feature, interface, service and resource. The conceptual model is composed by three subsets of metaclasses. The first one is associated to the service view, the second one to the information view and the third one to the resource view. The project model may be seen as a composition of the three models: requirements (REQ's), capabilities (CAP's) and mapping between the two first models (REQ2CAP). Figure 3 displays a general table of models and activities of the project stage.

Below there is a description of the main activities in the development process.

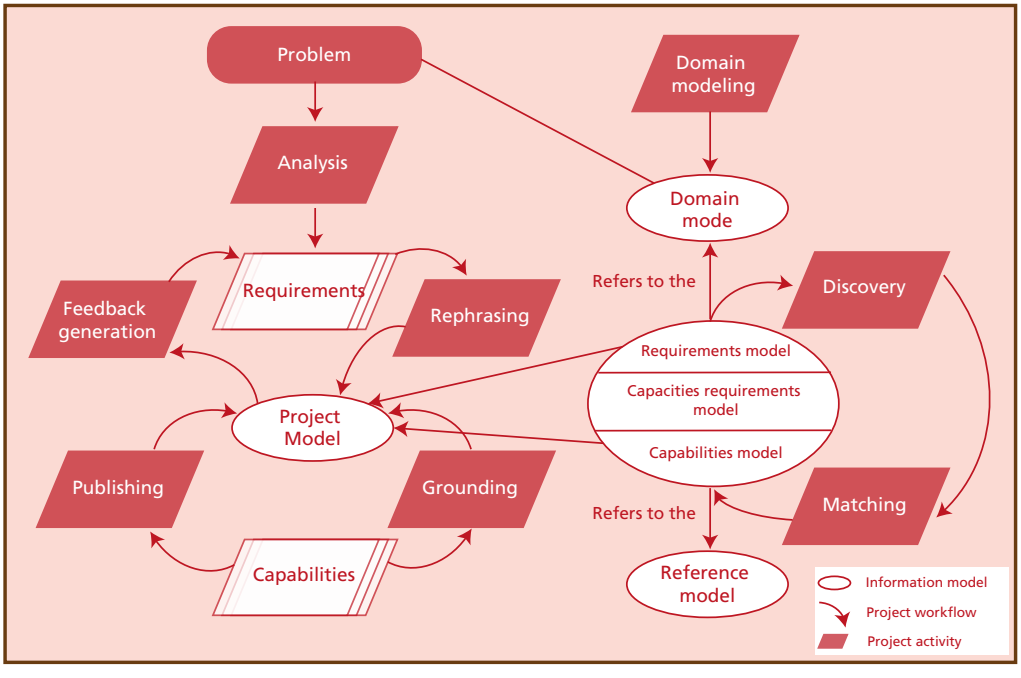
a. ***Publishing***: manual or automatic activity of searching for capabilities in the geospatial services network and of incorporation to the capabilities model. Capabilities are converted into the offered resources, that is, those available in the geoservices network;

b. ***Rephrasing***: activity responsible for converting requirements to extract a set of necessary resources for holding the use cases. Such resources may be associated to the concepts of ontology and then uploaded in the semantic resources network, composing a requirement model;

c. ***Discovery***: activity responsible for selecting, from the package of capabilities arisen in the publishing stage, a set which meets the required resources arisen in the rephrasing stage. For each required resource there may be a set of resources offered which meets the needs;

d. ***Matching***: activity responsible for associating requirements to capabilities, that is, required resources to offered resources. It evaluates among the candidate capabilities, catalogued in the discovery stage, the most appropriate ones to the requirements. The final result of

Figure 3:
Models and
Project activities



Source: Adapted from Usländer (2010, p. 99)

this stage is the requirements/capacities mapping model;

e. ***Grounding***: activity which provides a new capability in the geospatial services network.

SERVUS proposes the creation of a semantic catalogue, which composes an implementation architecture and a project environment. The semantic catalogue role is to communicate with the network geoservices and to process searches through resources. It works as a semantic extension that allows searches based on ontologies as well as the evaluation of the semantic proximity to the achieved results. Regarding the support to the methodology, the semantic catalogue accounts for activities of harvesting, which comprises the collection of services metadata, and of publishing, the publication of resources available to a network of semantic resources.

4. PROPOSED MODEL

This paper was developed in stages, as follows: literature review with studies on the SDIs, SOA development methodology, SERVUS methodology, graphs-oriented data banks, geocatalogue and semantic search. Another stage was a survey of requirements based on the geocatalogues studied - mainly the one proposed by the SERVUS methodology - on the needs of the environment stakeholders and on the current technological infrastructure of the body. The next stages were definition of high-level architecture, considering, among other aspects, data sources, SDI architectures and NSDI-BRA geoservices and development of prototype, including the definition of development languages, components, software libraries and the development environment. Then the stages of implementation with definition of environment and prototype installation; elaboration of a scenery to use the geocatalogue typical of environment auditing, prototype validation and finally evaluation of results (SANTOS, 2016).

The architecture chosen to implement the geocatalogue prototype was a web application in three layers. The presentation layer was developed in HTML5 by using JavaScript libraries: Bootstrap 3.3.5, jQuery 2.1.0, Google Maps 3.0 and the bootstrap-slider.js component. The business layer was implemented in Java version 1.7 and the application wrapped in web archive format – WAR. The requests of the presentation layer are made through the http protocol and they access a REST (Representational State Transfer) standard server, which returns documents in a JSON (JavaScript Object Notation) format. The business layer was hosted in an Amazon EC2 cloud service. The persistence layer has been implemented through an

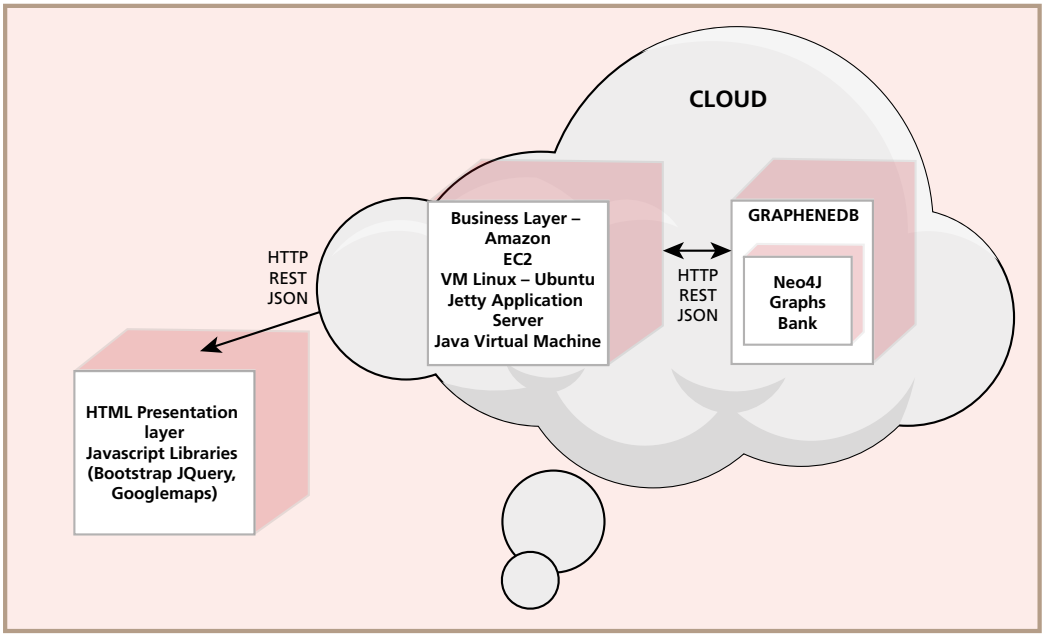
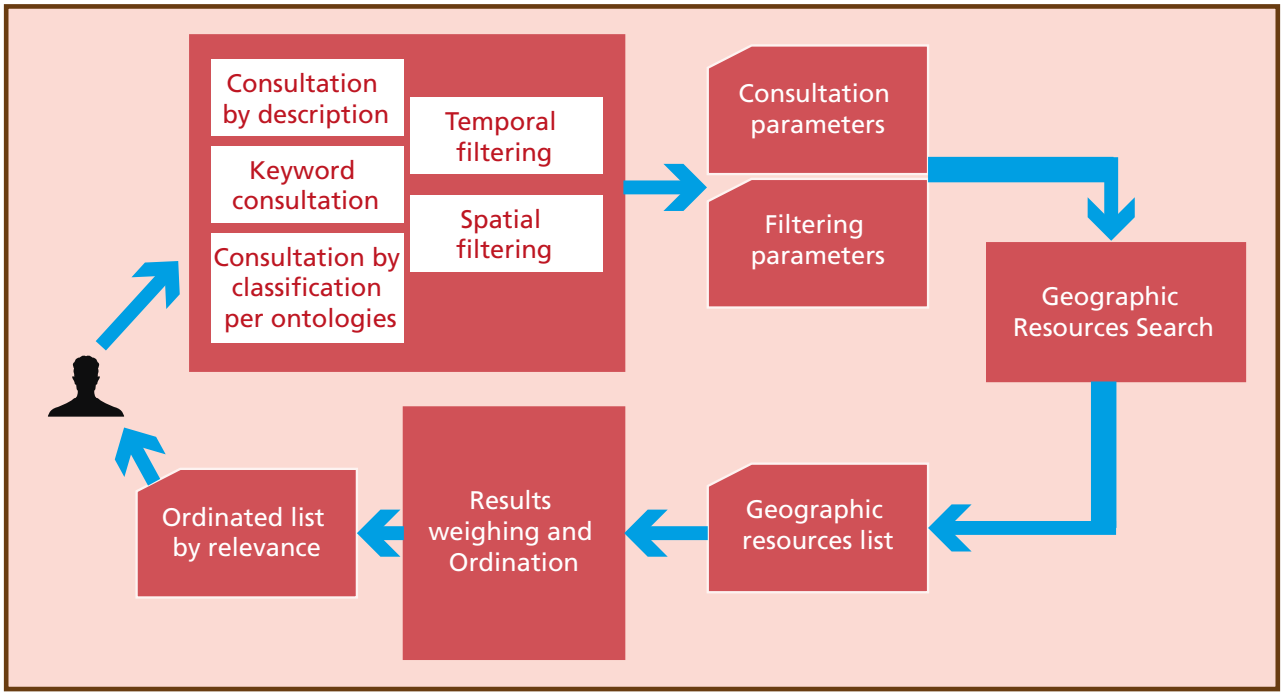


Figure 4: Implementation architecture

Figure 5:
Catalogue search model



NoSQL graphs data bank, the Neo4J 2.3.1. Hosting of the persistence layer was held in the GrapheneDB web page, which implements the Neo4J banks in a cloud environment. The communication between the business and persistence layers also takes place through REST interfaces and in JSON format. Figure 4 displays the implemented architecture.

Consultations can be made through description (free text), keywords and also through ontologies terms. For each one of the previous items, sentences or terms are divided into tokens which will be the terms used in the search. Filtering may be held by geographic coordinates (spatial filter) and by year (temporal). A weight is attributed to each ordered pair, composed by the term and the searched field, to be the base for procedure to generate a ranking. Figure 5 displays the geocatalogue search model.

The effective result of a search depends on the metadata quality and on the business value attributed to each field. To deal with this aspect

Table 1:
Catalogue search model

Search terms	Field with localized term	Weight	Maximum weight
Description: "vegetation" – (weight 8)	Name – (weight 8)	64	216
VCE ontology (root term): "vegetation" – (weight 6)	Description – (weight 6)	36	162
Keyword: "agricultural" – (weight 5)	Abstract – (weight 5)	25	135
	Total weight	125	513
	Relevance index calculated (125/513)	0,24	

a weighting scheme of the search fields and of the metadata fields was elaborated and a relevance index that acts as the basis for the process of generating a results ranking was established. Table 1 displays a sample of the relevance index calculation.

The records of resources are recovered through REST requests to the remote data bank, composed by the Cypher language commands. The main types of interaction are consultations of ontologies, creation of relationships between the resources and the generation geographic resources ranking. Figure 6 displays the ranking in a list format with the corresponding links to the resource preview, download and metadata preview.

For validation of the use of the geocatalogue, stages of the SERVUS methodology were used. We used as a baseline a problem related to an environmental audit that we present as follows: how to obtain the necessary data, in Brazil, to raise the locations of environmental protection areas, by using the available geoservices as a basis in the network

Figure 6:
Geocatalogue search screen

Figure 7:
Environmental protection areas imported to QGIS tool

of national spatial data infrastructure of public institutions and bodies?

The result was the location and preview of those areas. Figure 7 presents the environmental protection areas located by the geocatalogue, whose data were downloaded from WFS services and imported to the Quantum GIS tool.

5. CONCLUSION

This paper approached aspects of the implementation of a geocatalogue that allows the use of ontologies as research parameters. Consultations are carried out in the metadata of geoservices of spatial data infrastructures. We addressed implementation aspects such as architecture and data models. Regarding architecture, the geocatalogue was implemented as a three-layer web application that can be implemented in cloud environments. The data model was implemented in a graphs-oriented data bank.

The geocatalogue proved to be feasible as a tool to search for geographic resources and to support the SERVUS methodology, allowing operation in the publishing and matching stages. We can consider the integrated use of Cypher language in the resources network, ontologies and metamodel manipulation in non-relational graphs-oriented database, as contributions. The establishment of parametrization mechanisms for consultations, through parameters weighting and metadata fields, and the establishment of a relevance index, are contributions to the search field in metadata bases. For future works, we can consider automation of the harvesting process and integration with the corporate spatial data infrastructure, in such a way that the geocatalogue can perform together with the geoportals.

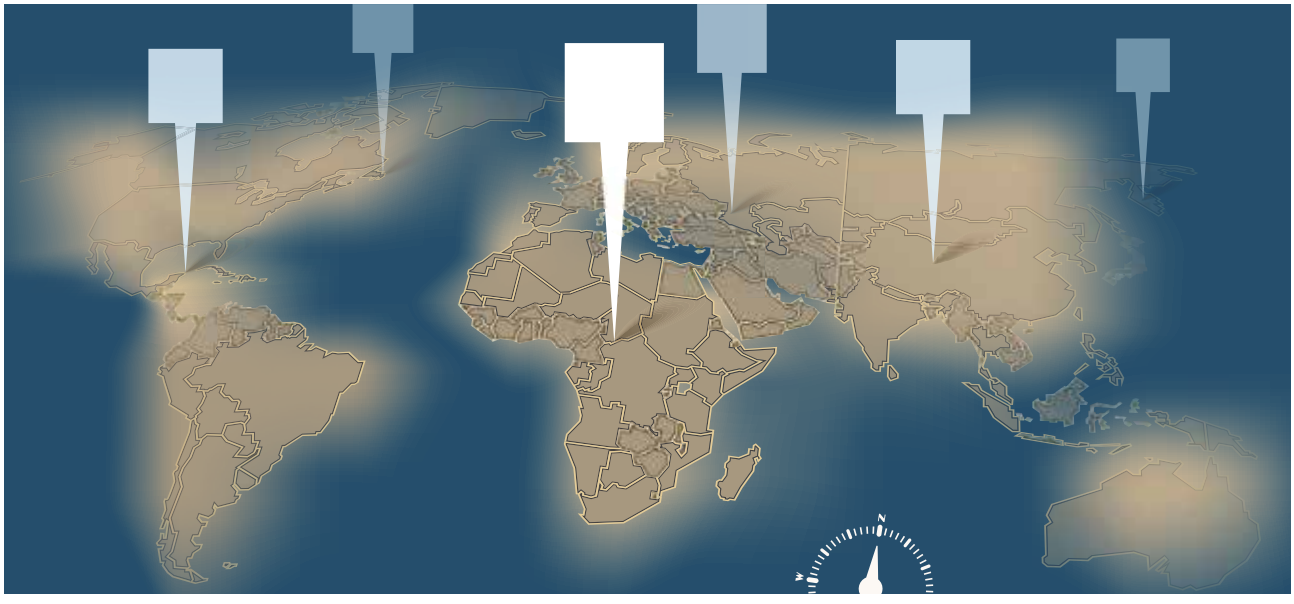
REFERENCES

ANDRADE, F. G.; BAPTISTA, C. S. Using semantic similarity to improve information discovery in spatial data infrastructures. Journal of Information and Data Management, v. 2, n. 2, p. 181, 2011.

BERNABÉ-POVEDA, M. Á.; LÓPEZ-VÁZQUEZ, C. M. (Eds.). Fundamentos de las infraestructuras de datos espaciales (IDE). Madrid: Universidad Politécnica de Madrid, 2012. 593p.

BIEBERSTEIN, N. Service-oriented architecture compass: business value, planning, and enterprise roadmap. Indiana: FT Press, 2006.

BRASIL. Decreto n. 6.666, de 27 de novembro de 2008. Institui, no âmbito do Poder Executivo Federal, a Infraestrutura Nacional de Dados Espaciais – INDE, e dá outras providências. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2008/Decreto/D6666.htm>. Acesso em: 22 out. 2016.



CAMBOIM, S. F. Arquitetura para integração de dados interligados abertos à INDE-BR. 2013. 140f. Tese (Doutorado em Ciências Geodésicas) – Universidade Federal do Paraná, Curitiba, 2013.

DALTIO, J.; CARVALHO, C. A. Um framework para recuperação semântica de dados espaciais. In: BRAZILIAN SYMPOSIUM ON GEOINFORMATICS, 13, 2012, Campos do Jordão. Proceedings... São José dos Campos: MCTI/INPE, 2012. p. 60-65.

GIMENEZ, P. J. A.; TANAKA, A. K; BAIÃO, F. Uma proposta de integração semântica para a Infraestrutura Nacional de Dados Espaciais usando geo-ontologias. In: WORKSHOP DE COMPUTAÇÃO APLICADA AO GOVERNO ELETRÔNICO (WCGE), 2013, João Pessoa. Anais... João Pessoa: UFPB, 2013. v. 5. p. 25-32.

GREENWOOD, J.; WHITESIDE, A. OGC Web Services Common Standard. 2010. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.248.6351&rep=rep1&type=pdf>. Acesso em: 22 out. 2016.

OGC. Catalogue Services 3.0 Specification: HTTP Protocol Binding. 2010. Disponível em: <http://docs.opengeospatial.org/is/12-176r7/12-176r7.html>. Acesso em: 22 out. 2016.

SANTOS, D. Implementação de um geocatálogo utilizando banco de dados orientado a grafos para apoio à metodologia SERVUS. 2016. 82f. Dissertação (Mestrado em Computação Aplicada) – Universidade de Brasília, Brasília, 2016. Disponível em: <http://alexandre.zaghetto.com/wp-content/uploads/2016/07/Disserta%C3%A7%C3%A3o_Vers%C3%A3oFinalParaEdi%C3%A7%C3%A3o.pdf >. Acesso em: 22 out. 2016.

USLÄNDER, T. Service-oriented design of environmental information systems. Vol. 5. Karlsruhe: KIT Scientific, 2010.

The purpose of data governance in external control organizations

Ricardo Dantas Stumpf
is an employee of the Federal Court of Accounts – Brazil. He works at the Department of IT Solutions in the area of Technical Management of Data. He has a B.A. in Computer Science from the University of Brasilia (UnB).

ABSTRACT

There are different levels of governance in an organization, whether dealing with strategies, IT, data in general or open data. Data are manageable assets, whether for the healthy operation of the institution or for sophisticated data analysis. This paper aims at defining the utility of data governance in this context, as well as the basic elements for its implementation with a bias for External Control organizations.

Keywords: Data governance. Data management. Data Science.

1. INTRODUCTION

What is important in an organization? People, budget, equipments and real estate, data... It is difficult even to think of the operation of organizations if any of these elements is missing, which proves the crucial importance of all of them. Several of these assets are essential to most entities and, consequently, they must be well **managed and governed** in order to maximize their value. In summary, managing means to guarantee that we do things right, while governing means to guarantee that we do the right things (WODZINSKI et al, 2015).

Now that the Federal Court of Accounts (TCU) has been investing heavily in data analysis, taking care

of the data became even more critical, as we can infer from the excerpt from the quarterly report below:

> A consensus has been reached on the three main challenges that the court must overcome in order to obtain good results with data analysis: the technical challenge, especially related to data quality; the regulatory challenge, regarding legal and normative restrictions; and the cultural one, concerning behavioral aspects of the people involved (TCU QUARTERLY REPORT - 4th QUARTER/2014, p. 101).

The second edition of the TCU Governance Guide (2015) brings a summary of interaction between management and governance:

"The functions of governance are:

a. To define strategic direction;

b. To oversee management;

c. To involve stakeholders;

Figure 1:
TCU – Basic Governance Reference Guide applicable to the Public Administration bodies and entities, 2nd edition, p. 32

d. To manage strategic risks;

e. To manage internal conflicts;

f. To audit and evaluate the management and audit system; and

g. To promote accountability and transparency" (TCU, BASIC GOVERNANCE REFERENCE GUIDE APPLICABLE TO THE PUBLIC ADMINISTRATION BODIES AND ENTITIES, 2nd edition, p. 31).

This logic applies both to the strategic level and to the different levels and more focused facets of the institution. Governance and information technology management, among others, gravitate towards corporative governance and management. Governance and management of produced and custodial data management gravitate towards the latter. Even data governance and management subdivide in the most complex organizations.

2. DATA AND INFORMATION

It is common to use data and information as synonyms, as well as knowledge and skill. However, strictly speaking, they represent diferent concepts according to professor Valdemar Setzer (2015), from the Universidade de São Paulo.

Data is "a sequence of quantified or quantifiable symbols", while information means "an abstraction on one's mind".

As a simple example, the sequence of symbols **880**. It is a **data**, either digital or not. If one recognizes 880 in a list of salaries, we then have the **piece of information** of a numerical salary with 880 units.

On the oher hand, **knowledge** demands a personal abstraction and experimentation (SETZER, 2015). A company's employee knows, according to their experience, that it means BRL 880,00, minimum wage in October 2016 attributed to a just-hired employee. The number 880 changed from data into information and then into knowledge. Finally, an auditor has the **skill** to interact and act on such knowledge during an audit.

This process takes place every day in professional environments. "Process" is a document, a repeatable workflow, a legal instrument, a computer code in execution or a protrusion of bones? It depends.

That is when **Data Governance** comes in, by using people's knowledge and skill to define policies, responsibilities, glossaries, metadata, workflows of data

Figure 2:
Different levels of governance "gravitating" simultaneously

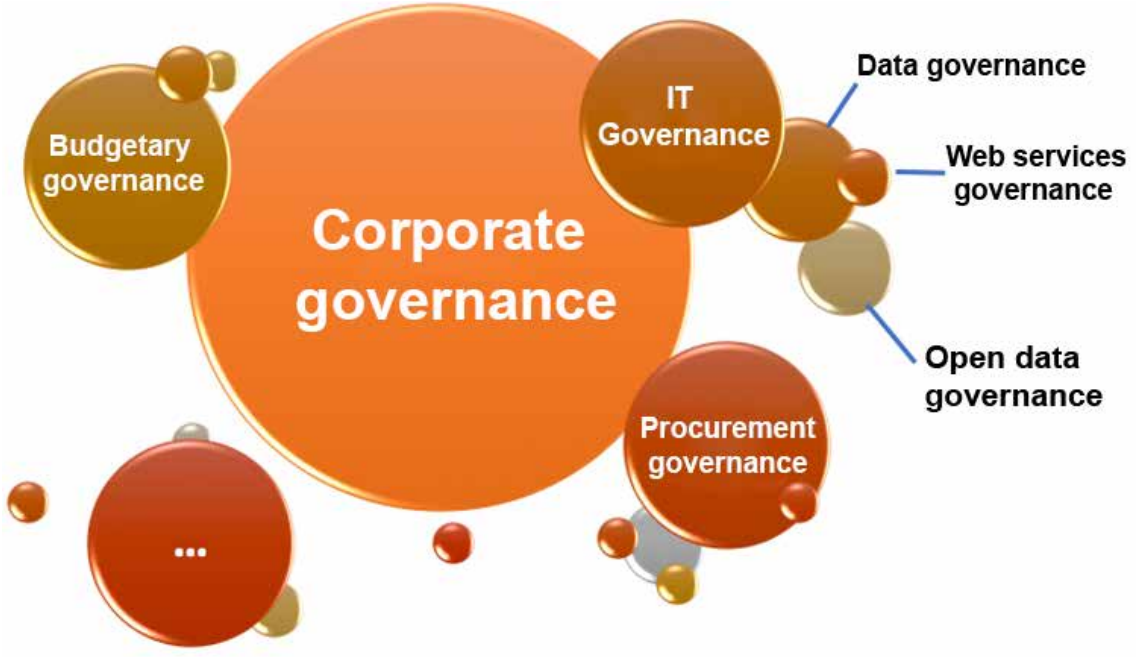
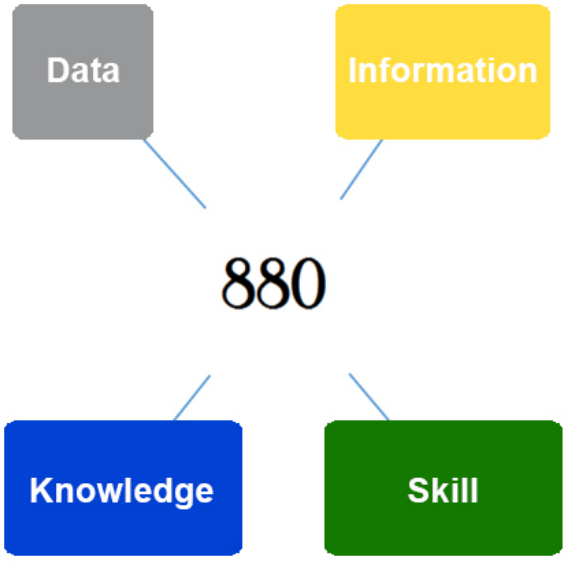


Figure 3:
Example of how symbols relate to several concepts



in movement, data monitoring and lineage, which allow the organization to improve their quality, discovery and understanding in order to simplify information and knowledge extraction, leading to better analysis and business decisions.

3. NEW OIL IS DIGITAL

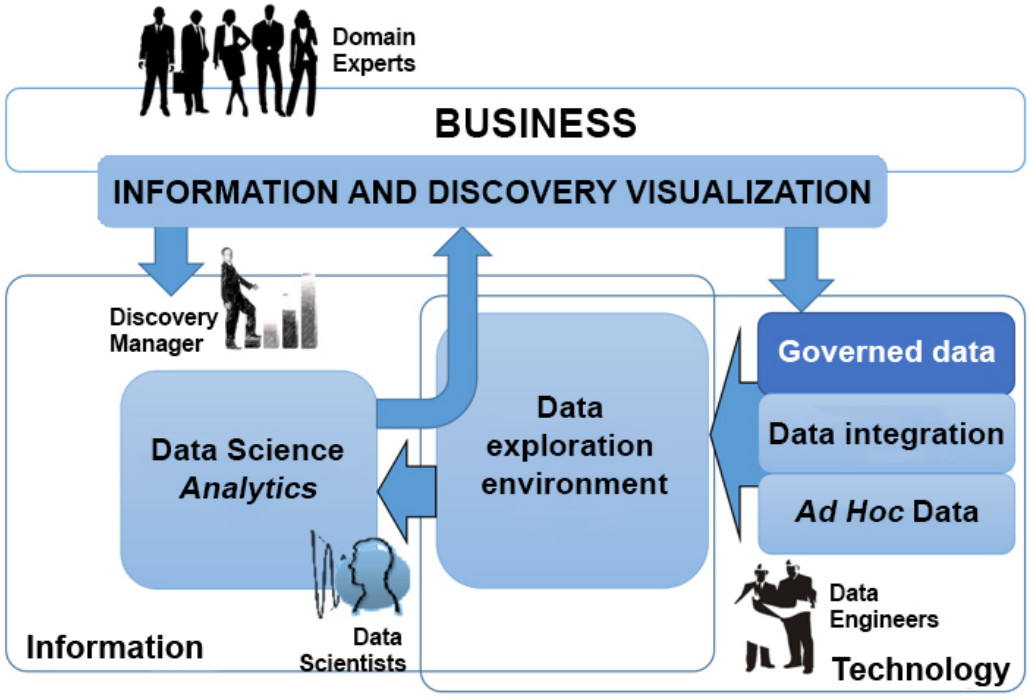
Data are being called "new oil" (VANIAN, 2016), but they have always been important and present, even before the digital format. However, our focus is on digital data, which have been growing in such a rythm and variety that they have been associated to "tsunamis", "landslides" and "storms", or Big Data. Internal data, usually the most aligned to the corporate internal work processes are called Small Data (HENDERYCKX, 2016).

A trend among companies is to invest more and more in data analysis and visualization, in order to envision more business opportunities and to increase operational efficiency. Likewise, public bodies, especially oversight ones, have been collecting great quantities of data from their partners and auditees, in addition to the feedback from society in the social networks.

Jan Henderyckx, DAMA Benelux, Chairman, presented a generic model for the data cycle:

In this model, notice that the data coming from business areas must be well governed. That does not mean that custodial external data do not require any governance, but primary worries may be others. For example, the TCU has an environemnt used by the External Control as an exploration and external and internal data-crossing laboratory.

Figure 4:
Data cycle samples, adapted (HENDERYCKX, 2016)



When necessary, the area responsible for the lab adopts the "Opportunity Principle", which justifies the ad hoc inclusion of new bases, even before there are quality treatments and more detailed documentation. The reason is to avoid the window of opportunity to pass. From then on, however, such base becomes eligible for a more complete treatment process.

In addition, there is the **contextual** issue. Low quality external bases may be useful to discover audit findings in information systems, while good quality derived bases, with full documentation of metadata, facilitate the analysis of the effective use of transfers of financial funds, for instance.

When the model presented is low in quality and organization, problems occur in three moments (DYKES, 2016):

a. *a priori* people do not trust the data;

b. there are so many sources that people do not know which one to use; and

c. there is analysis paralysis, when it is difficult to know when the analysis is good enough.

That said, we can turn to data management and governance.

4. DATA MANAGEMENT

Organizations usually already manage their data, especially the structured and semistrcutured data. Regarding the way it is done, it is common for management to follow good practices in their niches by force of necessity and knowledge shared by the internal peers and partners, such as the ones from the IT-Audit community, with around 20 oversight bodies that discuss better practices in IT. The systematization of the performance niches is displayed in frameworks as proposed in the DMBoK (MOSLEY et al., 2012), of DAMA International.

In the new version of this framework, even techniques that are more modern, such as integration by web services and publish-subscribe, have been addressed as ways of data to circulate locally or in clouds among systems or systems modules (BRADLEY, 2016).

Nevertheless, the topic Data Governance (which is a kind of "glue" for all the processes related to the subject) is only now gaining energy, since companies and organizations that depend on those data want to extract the maximum value from this more and more widespread and more complex asset. It is worth remembering that there might already be areas that govern their data well. In such cases, according to the English politician Lucius Cary, "where it is not necessary to change, it is necessary not to change" (In: SEINER, 2014).

Information is the raw material for the work of any oversight body. Their final products (determinations, recommendations) are primary sources of information to all of their auditees.

5. DATA GOVERNANCE

We finally reached a formal definition, according to John Ladley (2012):

> Data governance is the organization and implementation of policies, procedures, structure, roles and responsibilities which outline and enforce the rules of engagement, decisions, rights and accountabilities for the effective management of information assets.

Figure 5:
Synthetic view of the data framework, according to DMBoK

Regarding the definition, we will briefly discuss the need for principles, policies, data stewards and data classification, glossaries and metadata, leaderships and implementation approaches.

It is important to know the limits. Very bureaucratic data governance is an invitation to disobedience; on the other hand, excessive flexibility can lead to disgovernance, to a less efficient data management. That is why we should start by defining **principles**.

5.1 PRINCIPLES

Very specific, logical and feasible principles facilitate all the next steps. For instance, if the principle "data are corporate assets" is adopted by the organization, we must treat the entity "data" as we treat other assets: continuous improvement, definition of responsibilities etc. (LADLEY, 2012).

Each organization has its own principles. The following list is illustrative and is not exhaustive.

5.2 POLICIES

As John Ladley (2012) said, policy is what "give teeth to the principles". It is the formal document that makes the organization adopt the discussed principles. It is also common for policy to define the responsibility for the data (data stewardship), as well as the organizational structure which will conduct and monitor the efforts of Data Governance.

As a good example, we suggest reading the Information Governance Policy of the Central Bank of Brazil, published under Ordinance 47, of February 20, 2013, updated in 2016.

5.3 DATA STEWARDSHIP AND CLASSIFICATION

Data stewards are people or groups of people who have the responsibility of **taking care of data under their business sphere**. This is a fundamental change since it shares with IT the mission of taking care of the corporate data.

Suppose that the quality principle has been adopted. That means that business areas have to worry about the quality of data they generate for themselves and for others, and they must state quality problems in reports and any other sources that come from other business areas. The **Data Governance Office**, or its equivalent, is the department that will support data stewards in this duty.

Law 12.527/2011, known as the Information Access Law, formalized a guideline for the Public Administration, which says "publicity [of information] as a general rule and of secrecy as an exception". Therefore, the data stewards must also actively classify information they produce and guard, in order to accelerate the availability of public data or to restrict them according to the law.

5.4 GLOSSARIES AND METADATA

Glossaries, or vocabularies, define and eliminate ambiguities regarding business terms, such as the term "process", mentioned in section 2. The TCU

Table 1:
List of some of the principles for implementation of Data Governance

Principles	Description
Golden rule	All data are treated as corporative assets
Federation	There are defined patterns for data structures
Efficiency	Relevant data must be available to authorized users at the right moment, at the right place and in the right format
Quality	Corporative data are measured and managed to provide quality
Risk management	Comply with legislation, policies and internal rules regarding data
Collaboration	Corporative data are shared and disclosed resources
Contextualization	The context of use of data changes the way it is stored, treated and used
Innovation	New techniques are encouraged, following the other principles

Source: adapted from LADLEY (2012)

keeps the Vocabulary of External Control (VCE), created, according to the Minister-President Aroldo Cedraz, "to standardize the treatment of specialized information and to grant greater agility and precision to the recovery of contents in the TCU information systems" (VCE, p.5).

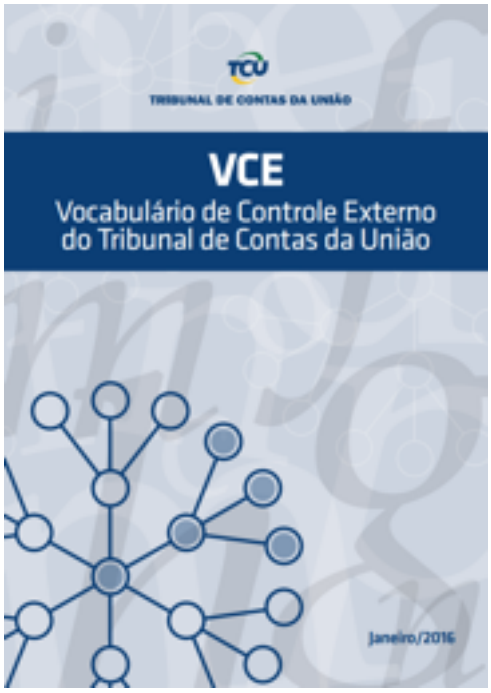
The classic definition of metadata is "data about the data". However, we can extend such concept to two aspects: business and technical metadata.

Business metadata are those that contextualize data that is stored or in movement. For instance, consider the WZV table of TTNN (hypothetical) database. Since the business area and the domain experts know this subject, they can **enrich the** existing **metadata** with specific information. Now, data scientists will know that the WZV table contains suppliers' registration data and that the TTNN base brings audited organizations that, among others, have purchased from those suppliers.

On the other hand, technical metadata are those which bring IT information about the data, such as tables and columns names, load dates, version, sizes, types etc.

The "magic" happens when technical and business metadata are **integrated to discovery** and use for analysis and change management areas. At TCU, there are some actions in this direction. The VCE data are being slowly integrated to other information technology solutions, in order to optimize and standardize the use of concepts. Another one is a tool specialized in metadata that shows, in an integrated way, information originated from data modeling, integrations through ETL (*Extract Transform Load*) and corporate databases.

Figure 6:
VCE front cover of the January/2016 issue



5.5 STRCUTURES

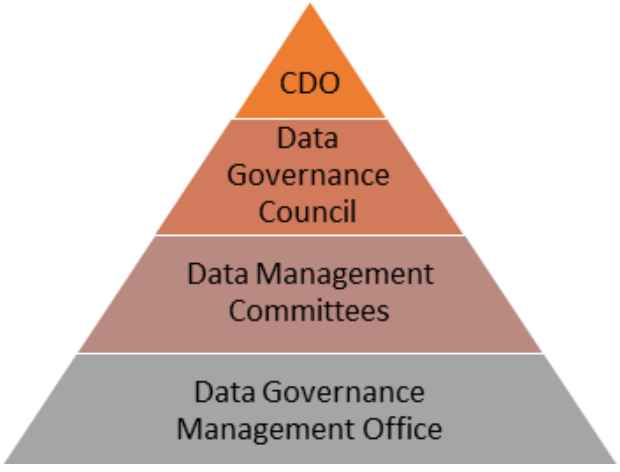
Each Corporation has its own culture and needs, and it would not make sense to have only one type of organizational structure to govern data. The CDO (Chief Data Officer) figure is increasingly common. The following are some of the ways used, according to Bergson Rêgo (2013):

Figure 7:
Edited sample of integrated technical metadata

a. CIO heading IT centralized team, which interacts with the business;

b. CDO interacting with IT and a central data management area of the business;

Figure 8:
Simple hierarchic structure (RÊGO, 2013)



c. CDO interacting with IT and several data managemnt areas of the business; or

d. a CDO for each great line of business interacting with IT (hybrid).

The hybrid structure is particularly interesting and used in practice in a great multinational bank (IPPOLITI, 2016). Due to the wide internal variety of available data, they have decided to divide them in great lines of business.

5.6 IMPLEMENTATION APPROACHES

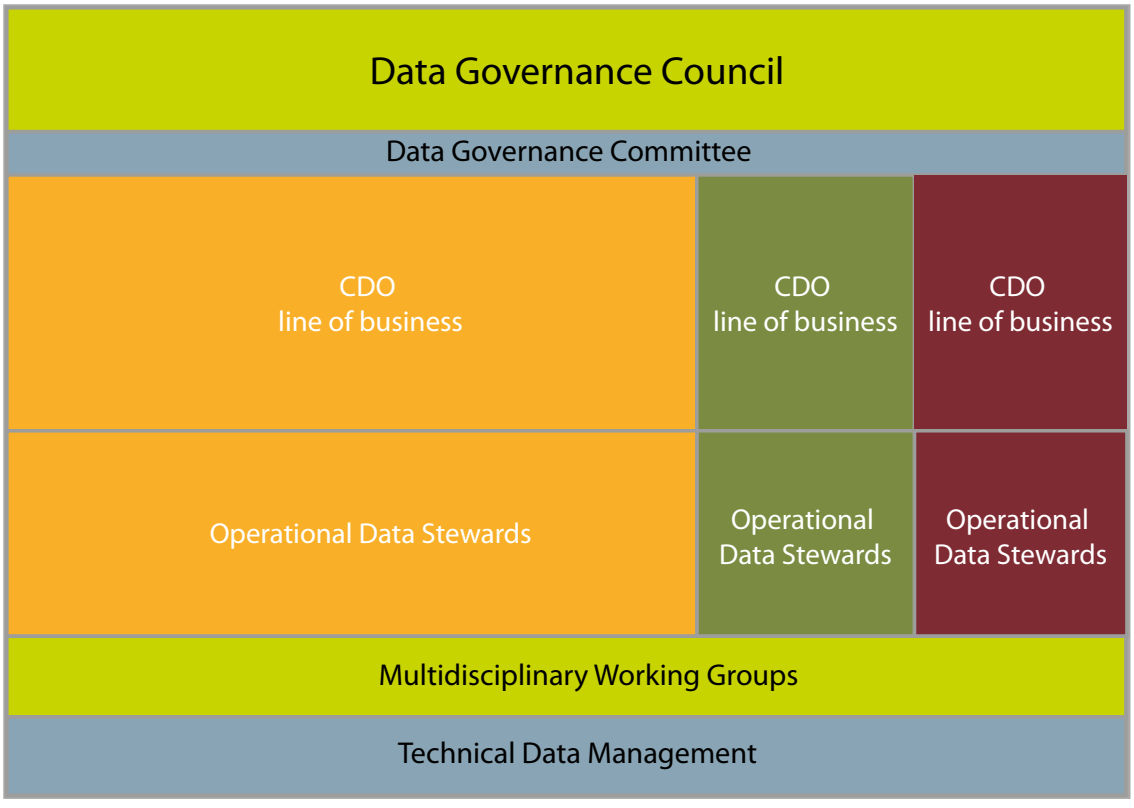
There are three more common structured ways of implementing Data Governance (SEINER, 2016):

a. Command and control;

b. Traditional; and

c. Non-invasive.

Figure 9:
Hybrid structure, adapted (IPPOLITI, 2016)



On the command and control approach, the highest-level committee defines the rules, determines data stewards, purchases tools and demands results.

The traditional one identifies the data stewards and guides them with more generic processes and existing tools, and measures results by analysing the involved data.

Finally, the non-invasive one formalizes the tacit data stewardship, focuses on applying the existing processes, encourages building of tools and the use of the current ones and measures results through the perceived increase in efficiency and effectiveness of analysis capabilities.

As we have already highlighted, none of them is necessarily better than the other one. It all depends on the culture and needs of each organization.

4. CONCLUSION

Despite the current importance of the theme Data Governance, it will be even greater in the future, when there will be great automation, even of intelectual activities, as long as they are repeatable. Data, whether originated from informatized systems or from sensors in day-by-day objects, will be a basic instrument for such automation.

That means that organizations should capture data, store them in an effective way and automate analysis with cognitive algorithms (HENDERYKCX, 2016). According to Ian Rowlands (2016), business metadata will have their semantics, and not only their syntax, inferred by algorithms and crowdsourcing (big groups thinking and collaborating).

It is also an increasing challenge to identify and reduce dark data, which represent loss of opportunities, waste of resources and risk (ROWLANDS, 2016). Rowlands reiterates an excerpt from a statement of the Auditor General of Canada in which this is a problem to be fought:

> One of the topics that unite many of our audits is that data collected from many governamental organizations are either unusable or are not usable, or are not used (Auditor General of Canada in ROWLANDS, 2016).

To make such reduction possible, the need to govern such data is pressing – in the right dosage for each case and context, always **focusing on the mission and vision of the institution**.

REFERENCES

BRADLEY, C. The new DMBOK 2 discipline of Data Integration. Disponível em: <https://www.brighttalk.com/webcast/12405/186095/the-new-dmbok-2-discipline-of-data-integration>. Acesso em: 12 abr. 2016.

BRASIL, BANCO CENTRAL DO BRASIL – BCB. Portaria nº 47, de 20 de fevereiro de 2013. Diário Oficial da União – DOU, 21 de fevereiro de 2013. Seção 1, p. 24.

BRASIL, Comunidade de Tecnologia da Informação Aplicada ao Controle – TI Controle. Sobre a Comunidade. Disponível em: <http://www.ticontrole.gov.br/ticontrole/sobre-a-comunidade/sobre.htm>. Acesso em: 20 out. 2016.

BRASIL. LEI nº 12.527, de 18 de novembro de 2011. Lei de Acesso à Informação. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 20 out. 2016. 2011.

BRASIL, TCU – TRIBUNAL DE CONTAS DA UNIÃO. Referencial básico de governança aplicável a órgãos e entidades da administração pública. Versão 2 – Brasília: Tribunal de Contas da União. Secretaria de Planejamento, Governança e Gestão. 80p. Disponível em: <http://www.tcu.gov.br/governanca>. Acesso em: 20 out. 2016. 2014.

_____. Relatório de Atividades: 4º trimestre de 2014 – Brasília: Tribunal de Contas da União. Secretaria-Geral da Presidência, p.101. Disponível em: <http://portal.tcu.gov.br/publicacoes-institucionais/relatorios/relatorios-de-atividades>. Acesso em: 20 out. 2016. 2015.

_____. Vocabulário de Controle Externo do Tribunal de Contas da União – Brasília: Tribunal de Contas da União. Instituto Serzedêllo Corrêa. Disponível em: <http://portal.tcu.gov.br/vocabulario-de-controle-externo/>. Acesso em: 20 out. 2016. 2015.

CANADÁ, Office of the Auditor General of Canada – OAG. Spring Reports of the Auditor General of Canada. Auditor General's Opening Statement. Disponível em: <http://www.oag-bvg.gc.ca/internet/English/osm_20160503_e_41358.html>. Acesso em: 20 out. 2016. 2016.

DMBOK. MOSLEY, M.; BRACKETT, M.; EARLEY, S. HENDERSON, D. DAMA Guia para o corpo de conhecimento em gerenciamento de dados. Technics Publications, versão brasileira 2012.

DYKES, B. Big Data Paralyzing Your Business? Avoid These 3 Common Traps. Forbes, Entrepreneurs. Disponível em: <http://www.forbes.com/sites/brentdykes/2016/09/28/big-data-paralyzing-your-business-avoid-these-3-common-traps/>. Acesso em: 20 out. 2016.

HENDERYCKX, J. Sustaining Data Driven Innovation. In: DATA MANAGEMENT CONFERENCE LATIN AMERICA, 2016, São Paulo. Anais eletrônicos… São Paulo: DAMA Brasil, 2016. 42 slides. Disponível em: <http://www.dmc-latam.com/palestrantes/jan-henderyckx/>. Acesso em: 19 set. 2016.

IPPOLITI, J. Implementing Data Governance Strategies. In: ENTERPRISE DATA GOVERNANCE ONLINE, 2016. Anais eletrônicos… 20 slides. Disponível em: <http://video.dataversity.net/video/keynote-implementing-data-governance-strategies/>. Acesso em: 12 fev. 2016.

LADLEY J. Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program. The Morgan Kaufmann Series on Business Intelligence. Morgan Kaufmann. 2012.

RÊGO, B. L. Gestão e Governança de Dados: Promovendo Dados Como Ativo de Valor Nas Empresas. Rio de Janeiro: Brasport, p. 60-74. 2013.

ROWLANDS, I. Data Management in Motion. In: DATA MANAGEMENT CONFERENCE LATIN AMERICA, 2016, São Paulo. Anais eletrônicos. São Paulo. DAMA Brasil, 2016. 25 slides. Disponível em: <http://www.dmc-latam.com/palestrantes/ian-rownlands/>. Acesso em: 19 set. 2016.

SEINER, R. Non-Invasive Data Governance: The Path of Least Resistance and Greatest Success. Technics Publications. 2014.

_____. Comparing Approaches to Data Governance. Disponível em: <http://tdan.com/comparing-approaches-to-data-governance/20386>. Acesso em: 5 out. 2016.

SETZER, V. W. Dado, informação, conhecimento e competência. Journal Data & Information 1(1), DAMA Brasil: p. 38-55. 2015.

VANIAN, J. Why Data Is The New Oil. Fortune, Brainstorm Tech. Disponível em: <http://fortune.com/2016/07/11/data-oil-brainstorm-tech/>. Acesso em: 20 out. 2016. Julho de 2016.

WODZINSKI, M. et al. Building an Impactful Data Governance Program One Step at a Time. Disponível em: <https://www.brighttalk.com/webcast/10477/142503/building-an-impactful-data-governance-program-one-step-at-a-time>. Acesso em: 26 fev. 2015.

The georeferencing of public real estate in the Brazilian geodetic system for the purpose of incorporation into the multipurpose technical registry: building real estate regularization in municipalities

Davi Lopes Silva
has a B.A. in Geography from the State University of the Ceará – (UECE), a B.A. in Computing from the State University of Ceará, and a specialist degree in Geoprocessing and Georeferencing from the Cândido Mendes University (UCAM). Currently, he is Manager of the Real Estate Assets Division in Fortaleza.

ABSTRACT

Public assets are, by nature and origin, a property of the cities, but they were left out of Brazil's technical registries, since the purpose was tax revenue. This resulted in a sparse updating of the information regarding the boundaries and perimeters of public assets throughout history. This paper highlights the importance of using georeferencing of real estate in the Brazilian geodetic system - Sirgas 2000 – to survey the dimensions of the real estate, with the goal of integrating the information into the Multipurpose Technical Registry (*Cadastro Técnico Multifinalitário* – CTM) that Brazilian cities are implementing. This article will also present the relevance of registering the updates performed in the public areas in the Real Estate Registration Notary Offices (*Cartório de Registro de Imóveis* – CRI). The methodology employed was the existing Brazilian bibliography about the CTMs and the CRIs, as well as the case studies on the subject from several authors. Finally, it must be emphasized that there is the need to include the information on real estate assets into the CTM technical registries, aiming for centralization, control and transparency of public expenditure.

Keywords: Georeferencing. Public assets. Brazilian geodetic system. Multipurpose Technical Registry.

1. INTRODUCTION

Based on the refinement of the geodetic techniques, the new technologies, and the importance of preservation of public property, the need arose to locate and delimit the public assets of cities in a more accurate way, using, to that end, the georeferencing of real estate properties.

By applying topographical knowledge and the planimetric survey techniques, the relevance of spatially representing objects by creating georeferenced polygons is made clear. In a simple and concise way, the goal is to demonstrate the importance of accurately knowing, delimiting and locating public real estate properties, seeking to promote the incorporation of this information into the cities' Multipurpose Technical Registry (CTM).

In the last decades, public agents have invested in equipment and real estate properties to meet the population's needs and to provide services. This has significantly increased the cities' real estate property, thus generating new assets that do not always have real estate registrations that are assertive or updated regarding their boundaries and location.

The Real Estate Registration Notary Offices (CRI) have, in some cases, received and stored imprecise information, such as distance measurements in "steps or spans", which, throughout Brazilian history have been replaced by the measurement unit in meters, of the International System of Units. It also happens that many registries were updated without a very precise assessment by the municipal bodies, which may lead in the future to the overlapping of private real estate property registrations and public lands, possibly causing discussions and disputes.

We recommend adoption of a CTM that involves a single registry of public and private assets making it possible to verify if there is overlapping of real estate property registration or irregular occupation, thus opening the possibility of filing administrative appeals to challenge registrations at the CRIs or finding quicker legal interventions.

This paper is supported by the Brazilian Legislation which addresses the georeferencing of real estate property in the Resolution nº 01/2005 of the Brazilian Institute for Geography and Statistics (*Instituto Brasileiro de Geografia e Estatística* – IBGE). This resolution altered and defined the new Brazilian geodetic system. This research also borrows from The Civil Code of Brazil, which details assets considered public. However, there is no norm or rule that also determines the application of georeferencing to state-owned lands and properties. To study this theme we also take into account Ordinance nº 511/2009 of the Ministry of Cities - which addresses the guidelines for creating, establishing and updating the CTM in Brazilian municipalities -, as well as the works of scholars who address the issue at

hand, such as Carneiro (2003), Pimentel (2012), Erba (2005) and others.

2. DEVELOPMENT

To promote public management, the municipality must know its territory beyond its geographical boundaries. Moreover, for the optimization of resources, it must consider surveying of the public properties, such as squares, rivers, public addresses, hospitals, government buildings, daycare centers, health centers or any property that is registered by it or in its possession.

In the past, to locate a property, it was necessary to read its property registration or descriptive memorandum to get to know its dimensions and areas. In the end, this would lead to a field survey to assess the measurements, which do not always match the expectation. Furthermore, in these registration documents, there is also the description of the adjoining properties. If the registration is very old, those may not describe the property or the adjacent land, but rather mention the name of the adjoining landowners or reference the local geography, like mentioning rivers, lakes, seas, hillsides and mountains.

In addition, the use of addresses, description of the terrain and adjoining landowners names does not contribute in the long run to the correct and accurate localization in case there are building removals, new constructions, or change of ownership, not to mention the fact that the public properties may be the target of unlawful occupations and third party property damage.

Georeferencing arose to solve this problem, reduce disputes and increase the accuracy in property localization. The term can be defined as the act of "describing and determining the location of a property through topographical survey, with precision equipment used by specialized professionals" (GEORREFERENCIAR, 2015, our translation).

This technique consists in representing, by means of polygons in a given standard reference system, the perimeter of a property based on its location coordinates, with the purpose of incorporation into the CTM. Afterwards, this information can be object of overlapping and comparison with the other public records or serve for rectification of property documents in the CRIs.

According to Zocolotti (2005), a polygonal represents a series of consecutive lines with known lengths and directions, which are obtained through field survey. The researcher also notes that "The survey of a polygonal is performed through the traverse method, which involves going through a path's outline defined by a series of points and measuring all angles and sides and an initial orientation" (p. 10, our translation). It is recommended to use closed traverses, since the starting and ending points should be at the same coordinate to calculate the area of public properties.

From February 2005 to February 2015, it was allowed to use simultaneously the Geocentric Reference System for the Americas (Sirgas 2000), the South American Datum 1969 (SAD 69) and Córrego Alegre, and there was support from the country's legislation for technical works using property georeferencing.

With the growth in the use of this tool in Brazil, these reference systems were used for the performance of several surveys, besides being taken into consideration for the CRIs registrations. In the future, this may lead to disputes if the property registrations do not have the information about the reference system in the previous property title documents. Not knowing the reference system employed implies the incorrect localization or polygonal displacement.

Since February 25, 2015, Sirgas 2000 is recognized and officially employed for georeferencing. This is set forth in a Resolution of the IBGE Presidency, which also elucidates:

> For the development of geodetic activities, it is necessary to establish a geodetic system that serves as reference for positioning in national territory. The materialization of this reference system, through geodetic stations properly spread across the country, is established in the reference infrastructure from which the new positionings are performed. (BRASIL, 2005, our translation).

According to Erba (2005, p. 25, our translation), "The problems derived from the relative localization disappear when the absolute positioning of the properties is employed. In this system, each detail surveyed gets a coordinate matching a unique reference system, whether municipal or national". Thus, the registry survey measured by coordinates and a unique reference system contributes to the property localization, besides comparing existing property registration with the actual reality of things.

According to Carneiro (2003), information surveying and the resulting registry have, in Brazil, a focus on taxation that mostly disregards the party with no tax interest, such as the public areas in this case study.

Considering that this author sees the municipal registry as a tool that should go beyond the tax aspect and that must participate in planning, urban control and spatial planning, we highlight:

> The multipurpose registry is defined by Dale & McLaughlin (1990) as a spatial information system designed to serve both public and private organizations, as well as serving the citizens. It differs from the other spatial information systems because it is based on plots. It serves as a basis for the other types of registry (legal, fiscal etc...) (CARNEIRO, 2003, p. 24, our translation).

Erba (2005, p. 21) is in line with the idea of a land registry in the cities and demonstrates this in a statement by the International Federation of Surveyors as follows:

> FIG's Statement on the Registry is in accordance with this last affirmation when it asserts that the Registry is a land information system, normally plot based, that records interests in land, such as rights, restrictions and responsibilities. It also adds that the Registry may be established for fiscal or legal purposes and/or to support planning, always seeking economic and social development. It highlights, however, that there is no need to think about a uniform Registry for all countries or jurisdictions.

The registry of public areas should follow the pattern of other Brazilian technical registries. Considering the guideline from the Infrastructure for Spatial Information in Europe (Inspire), it must possess at least the following information: "plot geometry, sole identifier, geodetic reference and index of the plots for printing/publishing" (PIMENTEL; CARNEIRO, 2012, p. 210, our translation). In this aspect, by creating a registry of public areas, those should be able to be integrated into the other municipal registries that also possess these data.

To promote understanding of the matter, the concept of registry plots defined in Ordinance 511/2009 is adopted. This Ordinance, which addresses public assets, is transcribed below:

> Art. 2. A registry plot is the smallest unit of the registry. It is defined as a contiguous portion of the surface of the land with a single legal system.
>
> Paragraph 1. Registry plot is all and every portion of the municipality surface that will be registered.
>
> Paragraph 2. The other units, such as lots, public roads, squares, lakes, rivers and others, are modeled by one or

> more plots mentioned by the head of this article, identified by their respective codes.
>
> Paragraph 3. A single and stable code shall be assigned to every plot.

In its third article, this abovementioned Ministry of Cities Ordinance explains that "all and every portion of land surface must be registered in plots". This refers us to start from the legislation regarding composition of a CTM that, depending on the situation, may require public area plots to be registered through one or more plots - or whatever provides more transparency and suits the preference of the party performing the registry –, and these plots must be identified by unique codes.

According to Pimentel and Carneiro (2012, p. 205, our translation), the International Federation of Surveyors clearly depicts the matter of composing the registry information:

> According to the International Federation of Surveyors (FIG, 1995), the plot is the smallest unit of land of the registry, and it can be defined in several ways depending on their purpose for the registry. For example, an area with a specific land use, an area of exclusive control, or a property owned by an individual or group of individuals. The boundaries can be formal or informal and, for the identification of the polygons, a unique code is used.

This author also recommends that any registry of land plots must be multipurpose and use standardization, as he states in the following words:

> It is suggested that plot surveying should be georeferenced to the Brazilian Geodetic System for unequivocal identification of its boundaries. The UTM conformal projection is recommended until a specific projection is determined for large scale cartography. (loc. cit., our translation).

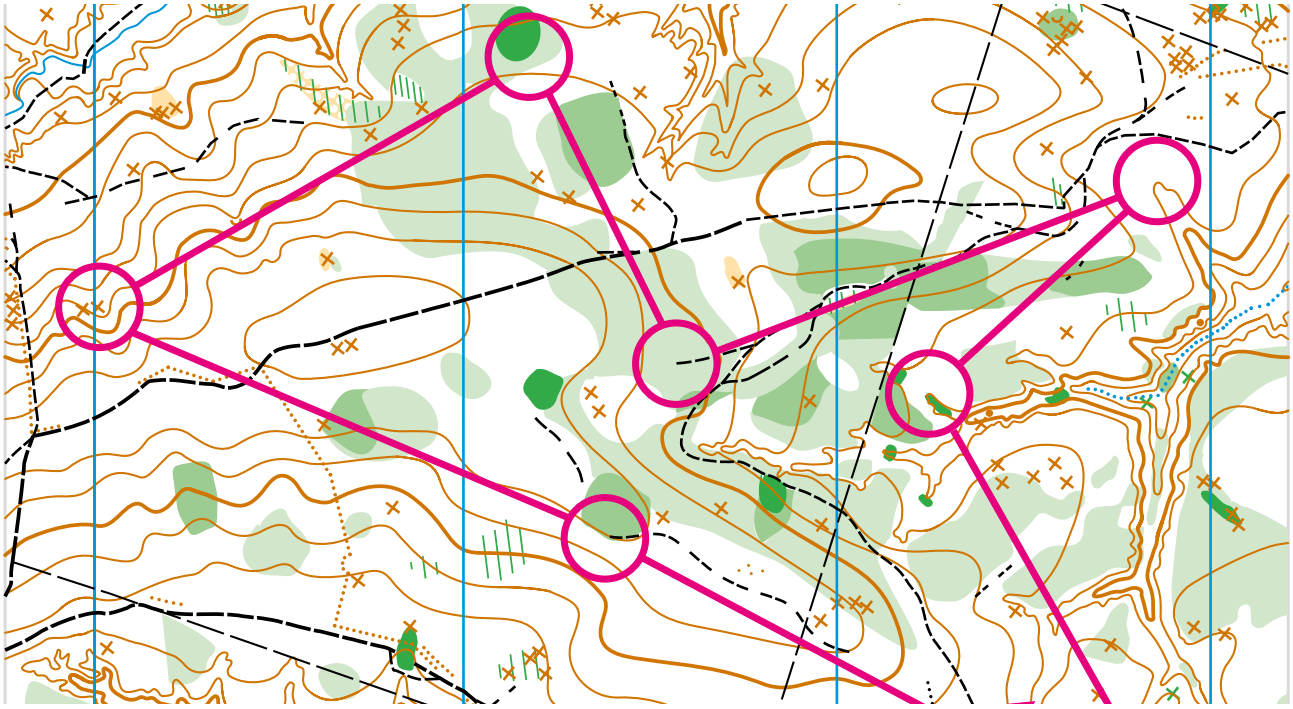
In Fonseca's (2010, p. 14) view, one of the goals of creating a CTM in a municipal context is also developing a public patrimony registry, which he describes as "an inventory of the properties that belong to public patrimony".

This same author highlights that, for a registry to be multipurpose, it must fulfill a social role, because "Nowadays, there is already a wider and more diverse meaning to registry, which showcases its important social role due to the various information it may contain, lacking only the taxation role, thus being considered multipurpose" (loc. cit., our translation).

In his thesis, Galdino (2006, p. 56) also emphasizes a concern about the registry in plot form, since this format must preserve other characteristics typical of the Brazilian legislative legal system, which implies the need for adaptation of the public surveying of public properties in this regard. The author also addresses the responsibility of territorial authorizing officers regarding the observation of property rights and restrictions, whether in a rural or urban environment, which he points out as follows:

> Here in our country, a plot-based land registry that takes observes the urban or rural property rights, restrictions and responsibilities - in management, economic, legal and geometric aspects – is only currently being interactively discussed by institutions, the public authority, and the scientific community, particularly according to the FIG models and adapted to the country's reality. However, it is necessary that the institutions and scholars explicitly persevere in their inclusion in specific legislation, appointing the institution in charge, rules, and regulations. (loc. cit., our translation).

When delimiting and registering a city's real estate properties, it is also possible to work on land re-



gularization, as Carneiro (2003, p. 25, our translation) points out:

> The work of land regularization consists in a series of technical, legal and administrative procedures (topographic registry and survey, analysis of original ownership of properties, discriminatory actions, demarcations, title legitimization etc.), which have the goal of ending ownership uncertainty by separating the unoccupied lands from the private ones and legitimizing the ownership and use of public lands.

In the case of public areas, other important information should be collected to construct said registry: type of public property, legislation of use and occupation of the property, lot and building dimension, identification of correspondence of the physical boundaries of property to real estate, acquisition cost and property valuation cost, in case there is any.

Public properties have been classified by The Civil Code of Brazil as proprietary, of common use by the people and of special use. Still according to the mentioned Law, in article 99:

> Article 99. Public properties are:
>
> I – those of common use by the people, such as rivers, seas, roads, streets and squares;
>
> II – those of special use, such as buildings or lots destined for service or establishment of federal, state, territorial or municipal administration, including those of their autarchy;
>
> III – the proprietary assets, which form the patrimony of legal entities of public law, as an object of personal or effective right of each one of these entities. (BRASIL, 2002, our translation).

Public properties have a purpose from their acquisition or receipt and should be classified based on their destination. According to Di Pietro (2000 apud CARNEIRO, 2003, p. 99), there is also the need for a separation between own real estate properties, defined in the legislation as special and proprietary, and properties of common use by the people, since the first promote equipment installation and service provision, while the second integrates the patrimony with the purpose of preservation and conservation by the public authority.

According to Carneiro (2003), because updating property registrations is not compulsory and there is lack of interest from the property owners, a dichotomy has been installed between what has been registered in the CRIs and the reality of Brazilian properties, whether public or private. In the case of public properties, they are frequently susceptible to expansion, reduction and incorporation, and there are interventions made

by the municipalities that are often not informed at the CRIs, which leads to the abyss between reality and the property's legal boundary.

According to Haar (1992 apud ERBA, 2005, p. 24), the difference between real boundary and ownership boundary often results in disputes and administrative and legal expenses for the property's defense and reintegration, which is transcribed as follows:

> Regarding the registry, there are two plot boundaries: the legal boundary, defined by Haar (1992) as an imaginary line that cannot be located in the land without an indicator to materialize it, thus requiring the study of the documents of the plot in question and those of the surrounding properties for the boundary's definition; and the ownership boundary, which is determined by the use of the property, materialized by natural or anthropological entities.

Regarding the previous topic, it is wise to report and inform in the registry incorporation act the difference between the legal and ownership boundaries, since these data are not always immediately added in the CRI, for there are rules, expenses and need of approval for the entry of these updates and corrections.

Building on the difference between legal and ownership boundaries, as highlighted by Carneiro (2003) and Erba (2005), it is also important to note that there is no good practice regarding updating information in the CRIs, which require, the payment of fees and emoluments, not to mention the technical responsibility.

In part, this dichotomy between the municipal registry, the property registrations and the properties' reality has created countless consequences for Brazil's real estate historical path, as Erba (2005, p. 25, our translation) explains in a portion of his research, as follows:

> In Brazil, a large portion of measurements carried out by professionals aims only to survey existing facts, thus determining the ownership boundaries of the properties and lacking the knowledge about the legal causes related to the effective property right. This fact ends up causing a generalized and known situation of confusion about boundaries and overlapping of property documents. A weak point of this system is the lack of accuracy caused by the existing subjectivity in the moment in which the aforementioned starting point is determined when the plot is tied to the urban grid. The use of this kind of reference has caused big problems in territorial advertising in many countries, generating overlapping of titles and conflicts of boundaries.

To facilitate the accounting balance of the cities, it is recommended to include, at the time of the registry, the acquisition cost of the property, with the aim of providing convenience and agility in the accountability, as well as making it possible to associate the investments to each patrimony, thus facilitating transparency in the public expenditure.

This need receives the contribution of the Brazilian Accounting Standards Applied to the Public Sec-

tor (Normas Brasileiras de Contabilidade Aplicadas ao Setor Público - NBCASP), incorporated by the Federal Council of Accounting (Conselho Federal de Contabilidade – CFC), which requires from the public spheres an investment accountability, aiming to represent the real value of every public property, which should also be associated to the municipalities' CTMs.

3. CONCLUSION

Brazil is developing its registry culture based on the Multipurpose Technical Registries (CTM), but it still lingers its focus in the tax aspect. In this aspect, the inclusion of georeferencing of public property in the CTMs must take into consideration not only its boundaries and outlines, but also the addition of information regarding the type of property being registered, aiming not only for its delimitation, but adding to the transparency in public expenditure as well.

This paper presented the opportunity to modernize and adapt public property surveys and descriptions in the Brazilian geodetic system, so that it becomes possible to avoid the loss of patrimony with overlapping of third party property registrations and to provide more integration with the Brazilian development platforms.

It is worth highlighting that, in order to protect the public areas, it the possibility of registering them in the same environment as the private areas should be taken into consideration, so that overlapping or incorporation into private patrimony can be avoided. By using this methodology, the municipalities will be able to add other information, such as current investments and use, thus avoiding the waste of State resources.

Brazilian history demonstrates the dichotomy between the CRI's real and legal boundaries. However, property georeferencing must contribute to attest the reality of the properties and serve to update this information in the CRIs, which possess a very important role in property defense.

REFERENCES

BRASIL. Casa Civil. Lei nº 10.406, de 10 de janeiro de 2002. Institui o Código Civil. Subchefia para Assuntos Jurídicos. Disponível em: <http://bit.ly/1hBawae>.Acesso em: 21 nov. 2016.

_____. Instituto Brasileiro de Geografia e Estatística. Resolução do Presidente nº 1, de 25 de julho de 2005. Altera a caracterização do Sistema Geodésico Brasileiro. Disponível em: <http://bit.ly/2glKejh>.Acesso em: 22 nov. 2016.

_____. Ministério das Cidades. Portaria nº 511, de 7 de dezembro de 2009. Diretrizes para a criação, instituição e atualização do Cadastro Territorial Multifinalitário (CTM) nos municípios brasileiros. Diário Oficial da República Federativa do Brasil, Brasília, DF, 8 dez. 2009, seção 1, p. 75.

CARNEIRO, Andrea Flávia Tenório. Cadastro imobiliário e registro de imóveis. Porto Alegre: Instituto de Registro Imobiliário do Brasil, 2003.

CONSELHO FEDERAL DE CONTABILIDADE (CFC). Contabilidade aplicada ao setor público. 2008. Disponível em <http://bit.ly/1xdNMmH>. Acesso em: 25 set. 2015.

ERBA, Diego Afonso, O cadastro territorial: passado, presente e futuro. In: ERBA, Diego Afonso; OLIVEIRA, Fabrício Leal de; LIMA JUNIOR, Pedro de Novais (Orgs.). Cadastro multifinalitário como instrumento de política fiscal e urbana. Rio de Janeiro: Studium, 2005 p. 13-40.

FONSECA, Cláudio Eduardo. A importância do cadastro tributário na arrecadação municipal e na auditoria de tributos: estudo de caso do município de Belo Horizonte. 2010. 32 p. Trabalho de Conclusão de Curso (Especialização em Auditoria em Tributos Municipais) – Faculdade de Direito da Universidade Gama Filho, Belo Horizonte, 2010. Disponível em: <http://bit.ly/2gkA92a>. Acesso em: 22 nov. 2016.

GALDINO, Carlos Alberto Pessoa Mello. Cadastro de plotas territoriais vinculado ao sistema de referência geocêntrico: Sirgas 2000. 2006. 255 f. Tese (Doutorado em Engenharia Civil) – Universidade Federal de Santa Catarina, Florianópolis, 2006.

GEORREFERENCIAR. In: MICHAELIS Dicionário Brasileiro da Língua Portuguesa. 2015. Disponível em: <http://bit.ly/2fJMzRj>. Acesso em: 20 nov. 2016.

PIMENTEL, José; CARNEIRO, Andréa Flávia. Cadastro territorial multifinalitário em município de pequeno porte de acordo com os conceitos da portaria n. 511 do ministério das cidades. Revista Brasileira de Cartografia, Rio de Janeiro, v. 64, n. 1, p. 202-212, 2012. Edição especial.

ZOCOLOTTI FILHO, Carlos Alberto. Utilização de técnicas de poligonação de precisão para o monitoramento de pontos localizados em galerias de inspeção: estudo de caso da U. H. de Salto Caxias. 2005. 112 f. Dissertação (Mestrado em Ciências Geodésicas) – Universidade Federal do Paraná, Curitiba, 2005.

The use of artificial intelligence techniques to support control activities

Luís André Dutra e Silva is an employee of the Federal Court of Accounts – Brazil. He has a B.A. in Computer Science from the Brasilia University Center (UniCeub). He has certifications in Software Engineering from IEEE and Project Management from Stanford University.

ABSTRACT

Cognitive services are an alternative based on artificial intelligence, with the purpose of obtaining solutions that are capable of detecting any kind of patterns in texts, images or any other source of data. This article describes the experiments on the use of artificial intelligence with unstructured data carried out by the TCU. The project stages are detailed and linked to the future applications and possibilities. The results of these experiments were very promising and we intend to adopt the techniques used during this period in new technological initiatives for the TCU and for the Public Administration in general.

Keywords: Cognitive services. NLP. Artificial Intelligence. Machine Learning. Text mining.

1. INTRODUCTION

A big part of the information consulted and produced by the TCU (Federal Court of Accounts) is received from bodies under its jurisdiction, registered in reports, votes, rulings, orders and other documents. These records are textual and complex, and demand sophisticated interpretation resources to obtain linguistically represented knowledge, especially because the data is unstructured. This characteristic demands countless analyses and combinations to explore and

add value to the information and to the decision-making process.

This fact entails considerable effort from the Federal Court's employees in order to structure these data and to systematize knowledge for use and decision-making. An example of this is the effort required for the performance of certain management and control activities, such as monitoring of TCU's deliberations and the classification and sorting initiatives of the Special Rendering of Accounts (*Tomada de Contas Especial* – TCE) cases.

In this context, the use of tools and algorithms supported by machine learning models for automation of document interpretation turns out to be essential and strategic for classifying and automatically extracting information contained within unstructured data sources.

The artificial intelligence techniques explored and systematized in this context allow the machine to learn more complex characteristics of concepts present in different documents. In view of this, the intention is to structure information that was initially scattered around different documents and formats and make it available and useful.

For information purposes, the development of initial experiments using techniques of Deep Learning has shown great promise. The prototype was executed during the period from July 1st to December 31st, 2015, and 257 thousand rulings from 1993 to 2013 were used as a training basis in the performed concept test, which classified deliberations contained in the text of 5,300 rulings delivered between 2014 and 2015. The result obtained revealed an average accuracy of more than 96%.

2. DEVELOPMENT OF WORK

Given the exploratory and pioneering nature of utilization of artificial intelligence techniques in the TCU context, the development of the work was divided into stages of experimentation with initial focus on cases of Special Rendering of Accounts. Given their experimental nature, many of the analyses performed were dismissed so that the best solution could turn up. With this in mind, the idea of this work has always been to replicate it in other control objects and business challenges.

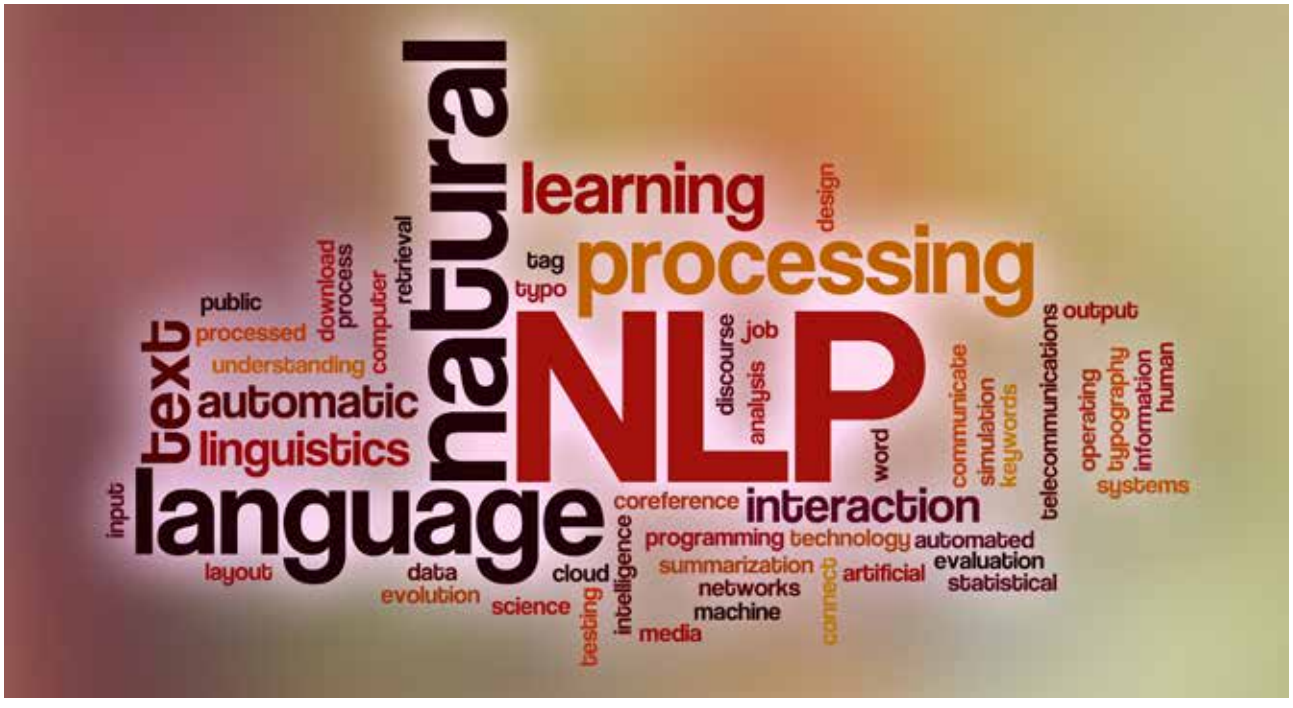
The first stage consisted in obtaining the documents of Special Rendering of Accounts cases directly from the EDM (Electronic Document Management) repository. The purpose of this was to establish a local base for future mining of the TCE case documents. The number of case documents used in this stage was around 680 thousand. To extract them, an algorithm in Phyton language was used to access the development database and the EDM in production, with the purpose of downloading all items into folders related to each TCE case. This stage was carried out between February 20 to 24, 2016, and the main challenge faced was the frequent interruption in the download of the case documents, which occurred due to connection issues. In addition, at first it was not possible to download all

documents. In the end, however, all documents of all TCE cases were downloaded.

In the second stage, the TCE's Registration database in Apex and the CGU's (Comptroller General of the Union) TCE Registration database were obtained, as well as the CGU's report documents in Microsoft Word file format. This was done with the purpose of composing the database associated with the TCE cases to perform text mining on the documents obtained. The number of documents used in this stage was around two thousand items, and an SQL (Structured Query Language) script was used for the databases dumps and for access to the CGU's shared data through local network sharing. The actions of this stage were performed between March 2 and 3, 2016, and the main challenge faced was that not all of the CGU's documents in Word format had matching metadata in their registration system. The system's registration data in Apex matched a small fraction of the liability present in the TCE cases. Due to these actions, the downloaded data became unavailable for further analysis.

In the third stage, the documents in TCE cases were grouped according to the government function and the types of irregularities found in the registration system in Apex. The purpose here was to attempt to classify automatically the non-annotated case documents, which comprise most of the cases, aiming to speed up sorting and distribution of the cases. The amount of case documents used was around 680 thousand items, analyzed through neural networks of the Self-Organizing Maps or Kohonen Maps type, using the free software Somoclu in a Linux environment. These actions were performed between March 4 and 10, 2016, and the main challenge faced was the fact that a large portion of the analyzed documents demanded several consecutive days of processing. Furthermore, the proportion between the number of documents that had previous classification and the number of documents without classification would be too divergent to obtain a useful result. Notwithstanding the difficulties encountered, it was possible to classify a great part of the documents by government function based on the registration obtained from the Apex system, which was used, until then, for the TCEs.

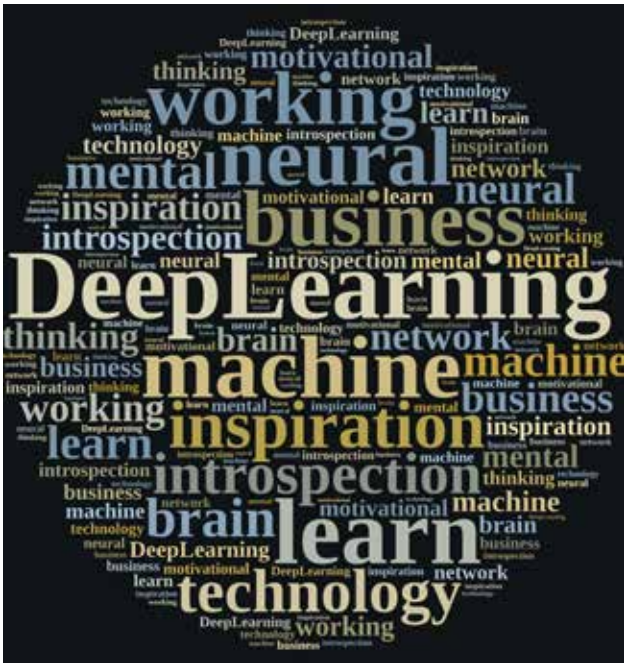
The fourth stage consisted in analyzing the irregularities in the cases of the Healthcare government function, aiming to classify the cases, sent in Excel spreadsheet format, that were pending distribution. Around 3,146 case documents were analyzed through the clustering technique using Self-Organizing Maps. This stage was carried out between March 11 and 21, 2016. The main challenge faced was the fact that, due to the small number of cases previously classified in comparison to the existing total and to the diffuse configuration of the irregularity classifications, the result would probably not be very accurate. Thus, of these 3,146 cases identi-



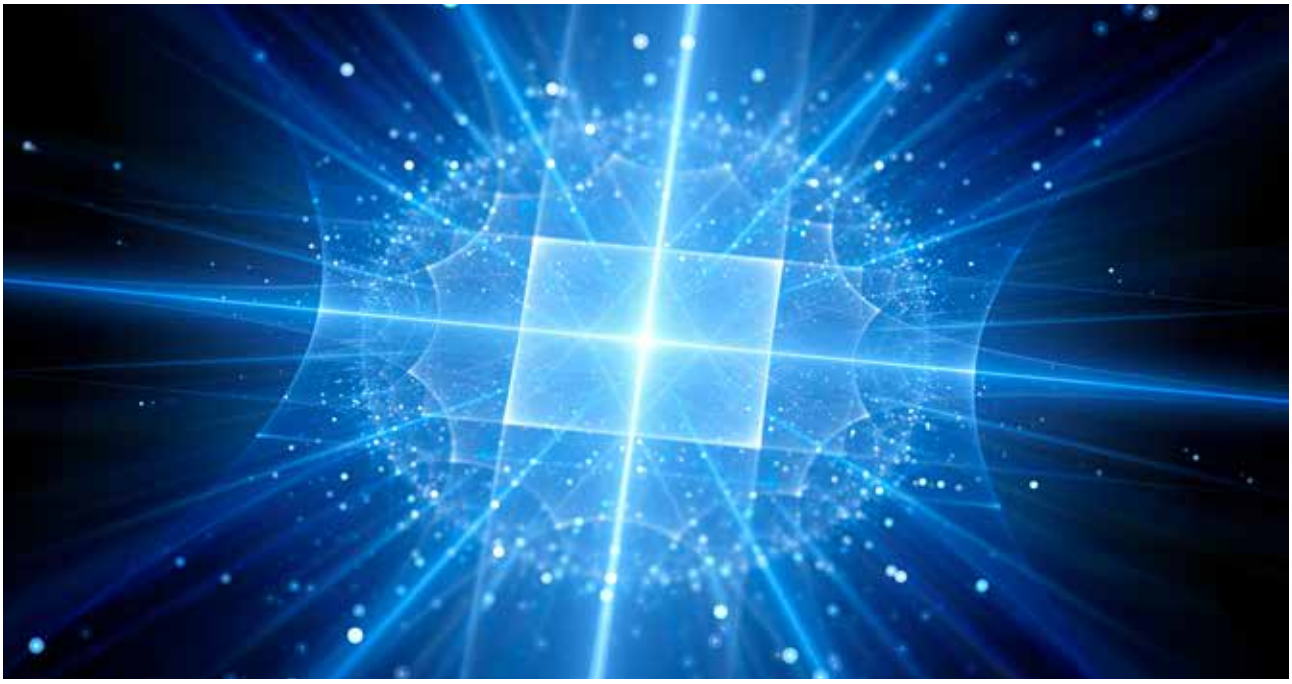
fied in the Healthcare government function, all of them would have irregularities types II, III, and IV, and most of them would have type V. None of these 3,146 items would have the type I irregularity, which, according to the specialists, should be detected in at least 10% of the cases.

The fifth stage consisted in creating the deliberations extraction service. The objective was to make available a web service to extract deliberations from the texts of court rulings, aiming at performing the semi-automatic insertion of the Determination to Agency/Body deliberation types into the Radar system, using the Apex interface,. Around 200 rulings are delivered per week and, to analyze them, Deep Learning in Python was used along with the Theano/Keras framework, POS tagging using the Aelius free library, regular expressions and the Flask framework in Python with the Werkzeug infrastructure. This stage was carried out from March 22 to 28, 2016. During its execution, the challenge was the existence of too many linguistic ambiguities to be dealt with through specific rules when extracting attributes of types of ruling deliberations. A few of these ambiguities need some kind of text comprehension that is at a level close to that of human intelligence. Nonetheless, the largest part of the Determination to Agency/Body deliberation items is identified with over 96% of accuracy and, according to the specialists, their feature are also extracted with high precision. Therefore, the type of deliberation chosen to demonstrate the usefulness of this service is considered the most adequate to provide a high level of success in the initiative to automatize the extraction process.

The sixth stage was dedicated to establishing a correlation between the texts of opinions given by the TCU contained in the TCEs and the 37 standard paragraphs, corresponding to each theme, available as aid on the website. The number of documents used was around 5,997 TCE opinions, and the techniques used in this approach were the ones of lexical and semantic similarity, using the Levenshtein edit distance algorithm and the cosine between the Bag-of-Words vectors. These actions occurred between April 1st and 7, 2016, and the main challenge faced was the high level of similarity between each theme and the opinions regardless of the chosen theme. Thus, it is too difficult to distinguish precisely the presence of standard paragraphs in each analyzed opinion. Therefore, the result was the fact that the percentage of similarity found for each opinion/paragraph, as indicated by the TCE specialists, was unsatisfactory because the similarities were too close to each other. Therefore, the experiment was considered concluded and this approach was hence dismissed.



The seventh stage was an experiment whose purpose was to analyze the structure of the CGU reports, and determine which information could be extracted and structured and apply text correction algorithms with a high level of noise derived from OCR (Optical Character Recognition). Other purposes include elaborating an information extraction algorithm similar to the one used for the Deliberations Extractor; providing a validation interface for the TCE specialists; defining classification and routing criteria based on the extracted information; classifying and suggesting routing through the validation interface; and validating the classification and routing by the specialists. The number of documents involved was around 3,800 reports, and the technique used was information extraction using regular expressions and crossing of structured data based on the entities extracted from the documents. This stage was carried out from April 8 to 26, 2016, and the main challenge faced was the fact that the quality of the scanned documents was not sufficient to ensure extractions that matched the same level of reliability achieved by the reading performed by the auditors that analyze TCE cases. In the end, this stage rendered several results. With the help of the TCE specialists, we determined what information would be structured and extracted from the CGU reports in Word format and extracted information from these reports making it possible to fill in with precision an opinion draft for the irregularity "Omission in



the duty of rendering an account". In addition, the crossing of extracted information with structured systems was performed, such as the Sisobi (Death Control System), the Sincov (Administrative Agreement and Transfer Contract Management System), and the Siafi (Federal Government's Financial Administration System). In addition, the same extractions performed in Word format documents started being used for the PDF scanned reports, which have less quality due to the OCR process; and a validation interface of the extractor was provided to the TCE specialists. The extractions performed by this experiment were considered useful for the TCE inventory management. According to the specialists, the task that is performed manually to this end can be partially automated, thus reducing the time needed to distribute/sort the TCE cases. However, the specialists considered automatic generation of drafts troublesome because it presented data that do not completely fulfill the needs of the auditors that analyze TCEs.

In the eighth stage, carried out between April 27 and May 25, 2016, the first deliveries had the purpose of providing the first versions of the services stipulated in the project's schedule. According to the schedule of the cognitive services creation project, the following products, in their first version, are available for validation and access: REST service for extracting deliberations of rulings; Deliberation Extractor validation interface; REST service for extracting TCE reports of entities; TCE Content Extractor validation interface.

The ninth stage, carried out between July 8 and 15, 2016, had the purpose of improving the scanned documents by using outsourced services. A sample of ten reports with major scanning issues was used, and the Google Vision API was employed for this attempt. The main challenge was to improve the quality of the text extracted from the TCE scanned documents without resorting to a new scanning process. However, in terms of the amount of errors, the achieved result was no better than the OCR software used in the TCU context (Adobe).

The tenth stage, performed between July 16 and 25, 2016, had the purpose of developing the first version of the recognition service of mentioned entities, NER (Named Entity Recognition), which is capable of extractingfrom any text the names of natural persons, legal entities, CPF (National Identification Number), CNPJ (National Registry of Legal Entities), and normative references. The number of documents used for the training of the entity recognition neural models was around 58 thousand rulings and the Amazon base of the University of Lisbon, which contained the manual annotations of each type of entity mentioned. To that end, the machine learning framework Apache OpenNLP, in Phyton and Java, was used with manual and automatic annotations in texts of rulings. With the use of machine learning algorithms instead of preset rules, the accuracy and the other associated metrics may not be ideal if there is not a large amount of manually annotated texts. The first

version of the generic entity extractor was exposed as web service in the production infrastructure. Seeing that an untrained validation base of approximately 10 thousand sentences was set apart to assess the quality of the NER web service, a FI score of around 81% was obtained for natural persons, which reflects the state of the art in terms of extraction of mentioned entities. The F1 score is the most adequate to demonstrate the balance between accuracy and recall, which should be the pursued goal.

In the eleventh stage, carried out between August 1 and September 14, 2016, there was the development of the second version of the generic entity extractor in pure Java, locally using the OpenNLP library and the previously trained models for the first NER version. The main challenge was the fact that portability from code Phyton to Java is not always possible in many customizations. Therefore, this initiative may not have been very successful if one or more frameworks used in Phyton did not similarly exist in Java. Nevertheless, the NER service, according to the standard TCU service architecture, Reference Architecture 8, was developed and is currently in production in the JBoss 6 EAP environment.

The twelfth stage, which occurred from September 16 to October 11, 2016, aimed to construct a service capable of detecting possible material errors in rulings delivered by the TCU before they were formalized. Around 600 documents were used to test the tool chosen from among one thousand rulings - which an inspection by the TCU's Internal Affairs Office pointed out as having material errors in their elaboration. The result achieved by the cognitive service was a capture of 40% of the material errors present in the rulings, after verifying the correct spelling of the names of the responsible parties, non-existent CPF and CNPJ or ones belonging to deceased natural persons or to inactive legal entities.

The thirteenth stage, which is ongoing since October 18, 2016, aims to provide a service of automatic extraction of ontologies in OWL format, of text documents, using machine learning algorithms. Around 280 thousand TCU precedent documents are being used. The Latent Semantic Indexing technique is being used by decomposing the matrix W of documents x concepts, which is obtained by the Bag of Words technique and the TF-IDF normalization, using Singular Value Decomposition. Latent Semantic Indexing is a statistical method that connects terms in a useful semantic structure, with no syntactic or semantic analysis and no manual intervention. By using this method, each document is represented not by terms, but by concepts that are truly and statistically independent in a way that the terms are not.

3. CONCLUSION

The cognitive services that were developed can be used in a successful way by other systems, with the purpose of improving the TCU and the Public Administration work processes. The need to structure texts that are produced continuously and depend on the classification and extraction of the information contained in these unstructured bases is inherent to the work of the TCU and of other bodies.

Therefore, making these services available may give more accuracy to the work of analyzing cases and make available to providing External Control professionals contexts that are relevant to every work developed in textual form.

For example, while producing a certain report all precedents on the topic could be automatically related to the text in creation, thus simplifying the process of searching for information relevant to the content being elaborated. Another example of use would be the automatic elaboration of summaries of texts received from bodies under TCU jurisdiction, which would speed up the processes of analysis of these external documents.

Moreover, by means of cognitive services, customized clippings can be elaborated about each subject addressed in a certain work. Thus, the auditors would always have the most updated information to support the elaboration of audit reports and all the other necessary documents.

REFERENCES

NOVELLI, Andreia, OLIVEIRA, José. Simple Method for Ontology Automatic Extraction from Documents. International Journal of Advanced Computer Science and Applications. Vol 5. No. 12. 2012

MADDI, Govind, VELVADAPU, Chakravarthi. Ontology Extraction from text documents by Singular Value Decomposition. ADMI. 2001

BERRY, Michael, DUMAIS, Susan, O'BRIEN, Gavin. Using Linear Algebra for Intelligent Information Retrieval. SIAM Review. Vol 37. No.4, pp-573-595, December 1995.