Abstracts

Panel 1: State of the art and future perspectives for scientific data management

Principles and Policies for International Coordination of Research Data Networks - Mustapha Mokrane (Executive Director, International Programme Office ICSU-WDS)

International data networks enable the sharing of data within and between scientific disciplines and countries and thus provide the foundation for Open Science. Developing effective and sustainable international research data networks is critical for progress in many areas of research and for science to address complex global societal challenges. However, the development and maintenance of effective networks is not always easy, particularly in a context where public resources for science are limited and international cooperation is not a priority for many countries.

The global landscape for data sharing in science is complex; many international data networks already exist and have highly variable structures. Some are linked to large intergovernmental research infrastructures, have highly developed centralized services and deal mainly with the data needs of single disciplines. Some are highly distributed, have much less rigid governance structures and provide access to data from many different domains. Most are somewhere between these two extremes and they cover different geographic regions, from regional to global. All provide a mix of data and associated data services which meets the needs of the research community to various extents and this provision depends on a mix of hardware, software, standards and protocols and human skills. These come together, working across national boundaries, in technical and social networks. In all of this, what makes a network function effectively or not is unclear. This means that there is also no simple answer to what can usefully be done at the policy level to promote the development of effective and sustainable data networks. Hence the rational for the present project - to study a variety of currently successful networks, explore the challenges that they are facing and the lessons that can be learned from confronting these challenges, and, where applicable, to translate this analysis into potential policy actions.

Detailed descriptive, operational and reflective information was collected on a total of 31 international data networks including several in the geosciences domain. This presentation will summarize the lessons learned and overall conclusions and recommendations from the project.

The Research Data Alliance: potential value for the Latin American and Caribbean Region Leslie McIntosh Borrelli, Ingrid Dillo (RDA-US)

The Research Data Alliance (RDA) was launched as a community-driven organization in 2013 with the goal of building the social and technical infrastructure to enable open sharing of data and data interoperability. Since its launch, RDA has become a highly-leveraged, community-driven, quickly-growing, output-focused, organization with over 90 groups designing, developing, adopting, and using the open-source tools, standards, practices and models needed to remove the infrastructure roadblocks on the path to data-driven innovation. Today RDA has over 6500 members from over 130 countries, covering many scientific domains and data sharing challenges.

Although RDA is a global enterprise, the organisation relies heavily on regional activities; they are crucial to the success of RDA in carrying out its mission and principles. In Europe, Australian, Canada,

and the United States, the RDA is relatively well known; this is not yet the case in the LAC region.
The Research Data Alliance Region of North America (RDA-RNA) is a concept that aligns the work of leaders in this region to improve the implementation and dissemination of RDA work. The concept is to increase the benefits to members while minimizing the burden of coordinated work through developing, operationalizing, and strengthening the RDA-RNA center. We will build resources for continued global RDA support and growing opportunities for a regional community.
In this presentation, the RDA organization and its groups and activities will be introduced as well as the RDA-Region of North America plans and activities. The discussion will focus on the potential mutual benefit of more engagement between the LAC region and the RDA.

Panel 2: Challenges for data management projects in Latin America & the Caribbean

Data, Open Data and Big Data: results of the LEARN project - Wouter Schallier (Hernán Santa Cruz Library ECLAC/CEPAL, Chile)

The Hernán Santa Cruz Library of UN/ECLAC participated in the 2 year LEARN project (http://www.learn-rdm.eu/) on Research Data Management, financed by the European Commission. The main objectives of this project were: to disseminate good practices and develop a model policy for Research Data Management. In October 2016, UN/ECLAC organized the first regional meeting on Research Data Management in Latin America and the Caribbean. Recently, the Hernán Santa Cruz Library published a free access guide on the Internet to disseminate good practices in research data management (http://biblioguias.cepal.org/gestion-de-datos-de -investigation/). In this talk, Wouter Schallier will present the results of the LEARN project and analyse the relevance, challenges and opportunities of data, open data and big data for libraries in Latin America and the Caribbean.

Global to Regional: Global Biodiversity Information Facility and the LAC Region - Anabela Plos (Museo Argentino de Ciencias Naturales MACN-CONICET, Argentina)

GBIF is an open-data research infrastructure funded by the world's governments and aimed at providing anyone, anywhere access to data about all types of life on Earth. In 2011, the GBIF Governing Board approved the establishment of the Nodes Steering Group, which includes representatives from each of the six regions: Africa, Asia, Europe, Latin America and the Caribbean, North America and Oceania.
The GBIF network draws all the sources together through the use of the Darwin Core standard, which forms the basis of GBIF.org's index of hundreds of millions of species occurrence records. Publishers provide open access to their datasets using machine-readable Creative Commons licence designations, allowing scientists, researchers and others to apply the data in hundreds of peer-reviewed publications and policy papers each year. Many of these analyses would not be possible without this.

GBIF provides different options for interaction between the region's members: a) Regional meetings (funded by GBIF), b) Biodiversity Information for Development -BID- meetings and workshop (funded by the European Union), c) Capacity enhancement support program -CESP- mentoring activities, support for regional events, GBIF advocacy actions, documentation and promotion of data use (funded by GBIF).
All the GBIF's members are encouraged to collaborate within their assigned or selected regions upon joining the network. Regional meetings, held either annually or biannually, allow Participants to address common tasks and issues, set regional priorities and stimulate specialization within the region.

Public goods for scientific data policies in Latin America -
Bianca Amaro de Melo, Paola Azrilevich, Alberto Cabezas, Patricia Palma Muñoz, Silvia Nakano (LA Referencia, Chile)

This paper presents LA Referencia's vision and actions regarding open science in general and scientific open data in particular for Latin America. LA Referencia is the network of open access repositories from nine Latin American countries that supports national open access strategies and policies through shared standards, metadata guidelines, and a single discovery platform. It harvests metadata of scholarly articles, theses and dissertations from national nodes. This federated platform came out of technical and political agreements between public science and technology organizations (National Ministries and Science & Technology Departments) with support of RedCLARA. It involves national nodes from Argentina, Brasil, Colombia, Costa Rica, Chile, Ecuador, El Salvador, México, Perú.

LA Referencia also fosters an articulation of policies and actions in open science, in order to create an ecosystem of open scientific information in the region as public good, led by the organisms of S & T. The service and actions in Open Science are based on policy agreements; guidelines and technology. This paper describes the regional context, priorities and vision on research data. For example, research data that validate the publications; data collected or generated with public funds or based on Data Management Plans. In addition, it includes a description of recommendations for metadata standards (Datacite), technology (open source and transferable), licenses (creative commons); among others. It presents the actions regarding internationalization in the context of Coar and OpenAIRE. Finally, it includes a reflection on alignment with F.A.I.R Principles and concludes with a summary of the key elements that characterize LA Referencia's approach to research data as a public good.

Session 1: Case Studies in Agriculture and Soil Science

A Data Repository for Agricultural Science & Technology -
Debora Pignatari Drucker, Inácio Henrique Yano, Roberto Hiroshi Higa (Embrapa, Brazil)

To address the complex and intertwined challenges facing humanity such as climate change, invasive species and biodiversity loss, scientists, educators, decision-makers and citizens need open, persistent and secure access to well-described and easily discoverable scientific data. We aim to

present our effort on building a Data Repository for Agricultural Science & Technology to enable new science and knowledge creation through universal access to data about agriculture, environment and sustainable development. Our approach encompasses strategic, technological, sociocultural and innovation aspects within Embrapa, the Brazilian Agricultural Research Corporation, a networked public company composed by 46 research centers distributed throughout the country. Our repository runs at Embrapa Agricultural Informatics Center and is based on Dspace, a free and open source repository software package. It was established in order to promote efficient data management and facilitate access to agricultural research data in accordance with current Brazilian policy implementation guidelines and aligned with internationally widely adopted standards. Researchers are encouraged and trained to store data and metadata according to the FAIR principles (findable, accessible, interoperable and reusable). A curation process was defined to assure data quality and usually there are requests to improve data documentation accurateness. We are currently building a sustainability and governance strategic plan to ensure long term data preservation and we are exploring business models to foster data driven innovation. Our results provide insights on how to overcome sociocultural barriers regarding data management practices, mainly data sharing and reuse, and can be useful for other organizations.

The Free Brazilian Repository for Open Soil Data -
Alessandro Samuel-Rosa (Universidade Federal de Santa Maria, Brazil)

Brazilian soil science has produced a great deal of data. Most of the information is published as a single paper, and the primary data is unavailable to other researchers. Lately, soil scientists have increased their concerns with data discoverability and reusability, and reproducible research. To address this issue, Brazilian soil scientists have recently created a data repository using community-built standards and following open data policies. The Free Brazilian Repository for Open Soil Data – febr, www.ufsm.br/febr – is a centralized repository targeted at storing open soil data and serving it in a standardized and harmonized format. The repository infrastructure was built using open source and/or free (of cost) software, and was primarily designed for the individual management of datasets. This is accomplished by storing each dataset using a collection of Google spreadsheets accessible online. Spreadsheets are familiar to any soil scientist, the reason why it is easier to enter, manipulate and visualize soil data in febr. Soil scientists can help in the definition of standards and data management choices through a public discussion forum, febr-forum@googlegroups.com. A comprehensive documentation is available to guide febr maintainers and data contributors. A detailed catalog gives access to the 14 477 soil observations from 232 datasets contained in febr. Global and dataset-specific visualization and search tools and multiple download facilities are available. The latter includes standard file formats and connections with R and QGIS through the febr package. Various products can be derived from data in febr: specialized databases, pedotransfer functions, fertilizer recommendation guides, classification systems, and detailed soil maps.

Making Available Digital Soil Maps from Brazil in an Interactive WebGIS -
Ricardo de Oliveira Dart (Embrapa Solos, Brazil)

Soil maps at appropriate scales are essential information for land use planning. However, the Brazilian soil information is scattered in several institutions and in several formats, as well as the interruption of the systematic soil survey program. In order to organize the spatial data produced at the Brazilian Agricultural Research Corporation (Embrapa), a spatial data infrastructure was developed (IDE-Embrapa) where thematic collections related to soil have been gathered and publish in a web environment. The objective of this work is to present the initiative of organizing the spatial data of Embrapa related to soil information through the development of the IDE-Embrapa. Initially the spatial data that were loaded in a previous geoinformation infrastructure developed by Embrapa Soils were shifted to the IDE-Embrapa infrastructure. The implementation of the IDE-Embrapa was performed using open source software, based on the Open Geospatial Consortium standards. The IDE-Embrapa infrastructure uses GeoNode platform, which integrates a geospatial database (PostGis) with a map server (GeoServer) and a metadata catalog (PyCSW), and is controlled by a Content Management System in the Web environment. Currently, 100 information layers and 60 documents were catalogued in the IDE-Embrapa (geoinfo.cnps.embrapa.br). These data and metadata are already available for download. Maps represented by various territorial boundaries and scales were registered. Currently, the IDE-Embrapa infrastructure is making available the Brazilian soil information available to any external user. This work is under construction and we hope soon to have all maps prepared by Embrapa Soils catalogued and available, in order to safeguard data and metadata, for ready use of these by society.

InfoSoilsBr: The Brazilian Soils Database -
Marcos Bacis Ceddia, Renan Miranda, Gabriel Santiago Cardoso Rizzo, Sabrina Oliveira Cruz, Pedro Vieira Cruz, Sergio Manuel Serra da Cruz (Universidade Federal Rural do Rio de Janeiro, Brazil)

Providing accurate and updated information on the soil profiles of Brazil is an enormous task and continuous challenge that requires an interdisciplinary approach. InfoSoilsBr is a novel databease that aid researchers to collect, store, compute, harmonize and publish Brazilian soils profiles data and geographic information. The InfoSoilsBr aims to serve high quality-assessed, georeferenced soils profiles database to the Brazilian and international communities upon their standardization and harmonization. The focus of this work has been on developing the computational framework and a database that can store not only new soils profiles data (collected by mobile devices like cell phones and tablets during field works) but also to import/export legacy data from/to the other soil databases. Besides that, the database was conceived to be fully compatible with ISRIC, SSURGO, BDSolos and FAO databases. The database was designed according to big data, data provenance, open data, data integration requirements, taking advantage of the latest information communication technologies. The database can store large amounts of raw and curated soils data. Each profile description recorded in the database has a set of key attributes (e.g., mineralogical, morphological, chemical, physical and environmental data). Furthermore, the database stores georeferenced data as

text and images about each profile and analytical data from soil samples that were analyzed in wet laboratories. (Financial Support: UFRRJ; PET-MEC/FNDE, red CYTED-BigDSSAgro)

Toward Integration of Data-Centric Agronomic Experiments with Data Provenance - Sergio Manuel Serra da Cruz (Universidade Federal Rural do Rio de Janeiro, Brazil)

With improvements in high-performance computing, sensors, and communications, the amount of scientific data in agriculture has been exploding. Thus, researchers must rely on computational simulations to model the data-centric in silico agronomic experiments. Reproducibility, transparency, independent verification are major features of Science. However, even agricultural research of exemplary quality may have irreproducible empirical findings because of random or systematic error. Funding agencies, researchers, and reviewers are demanding improved processes and the use of open data to increase reproducibility of those experiments. Currently, there are no scientific criteria to evaluate the integration of data-centric agronomic experiments with data provenance. We propose RFlow, a framework that aid researchers to manage, share, and enact the scientific in silico experiments of research projects that use reusable R scripts. The framework uses open data and W3C provenance standards and transparently captures provenance of the agronomic experiments. RFlow is non-intrusive, can be connected to workflow systems and does not require researchers to change their working way. Our computational experiments show that the framework can collect provenance metadata and enrich a scientific project. This study shows how RFlow can serve as the primary integration platform for statistical systems, like R, with implications for other data and compute-intensive agronomic projects. As a proof of concept, we show the concrete effectiveness and expressive power of the RFlow which was evaluated through a set of data-driven agronomic in silico experiments and provenance queries that exemplifies what kind of information was gathered.

Lightning Talks for posters

Repositorio Institucional de la Universidad Nacional de Asunción, Una Iniciativa Para el Acceso Abierto - Marta Barrios de Alvarez, Emilce Sena Correa (Universidad Nacional de Asunción, Paraguay)

En la actualidad el mundo se encuentra enfocado en los procesos de acceso abierto a fin de lograr la visibilidad institucional donde el acceso a la información juega un rol fundamental, utilizando como uno de sus principales mecanismos los Repositorios Institucionales (RI).
En ese contexto a nivel mundial se puede hacer mención al directorio mundial de repositorios académicos de acceso abierto Open Doar con aproximadamente 3.500 repositorios, de los cuales casi el 90% corresponde a RI, distribuidos de la siguiente manera: Europa 45%, Asia 20%, Norteamérica el 18%, y Suramérica, con el 9%, distribuyéndose el porcentaje restante entre los demás continentes, imponiéndose Brasil, con aproximadamente 100 repositorios, ocupando el 6% de la distribución mundial. Esta distribución coincide en cuanto a porcentajes a otro importante referente del tema, el Registro de Repositorio de Acceso Abierto.

De esta manera se observa que Paraguay, no figura en el inventario mundial de los RI.
A este respecto, nuestro trabajo pretende crear un repositorio, para la UNA, proponiendo un modelo de repositorio funcional que implique el registro y difusión de toda su producción intelectual, que actualmente, están inéditas en muchos casos, y sin visibilidad alguna. La universidad, es la que cuenta con mayor cantidad de investigadores, categorizados por el Sistema Nacional de Investigadores del CONACYT, y por ende, la que tiene mayor productividad, sin embargo, no se cuenta con una sistematización de la producción científica.
El trabajo estará dividido en dos etapas: la primera de investigación: Revisión de la literatura y relevamiento de los recursos disponibles y la segunda de preparación: la elaboración de la propuesta y definición de la estructura funcional.

In search of a repository of scientific data in an adverse climate: the Salvadoran experience -

Willian Heriberto Carballo Sánchez (Universidad Escuela de Comunicación Mónica Herrera, El Salvador)

State support for the operation and maintenance of repositories of scientific databases in El Salvador has decreased over the years. From 2013 to 2015, the Ministry of Education allocated $ 90,000 to the program called ""Annual maintenance of subscriptions to electronic resources"". However, in 2016 there were no funds for that purpose. And in 2017 there were different bureaucratic hurdles that were close to leave the program without resources again. After intense negotiations led by universities, support was finally achieved; however, it was lower than the amount delivered in previous years: only $ 72,000. As for 2018, there is still no resolution. This happens while, on the other hand, the Ministry requires the universities more and better research policies and their teachers more projects and results in this area.
Faced with this reality, the main universities of El Salvador have decided to work together. This is how the Consortium of University Libraries of El Salvador (CBUES) emerged, which brings together 12 of the most prestigious educational institutions in the country and whose work was vital in negotiating the resources of the repository for 2017. This consortium aims to support the development of library policies that allow greater and better access to collections and create a central repository for the consultation of thesis and other research documents.
This paper aims to expose this Salvadoran experience by creating this database in adverse situations, in a country that only allocates 3 percent of the Higher Education budget to scientific research and in which access to scientific information is still limited. The experience to report will include limitations, well done tasks and lessons learned from teamwork and organization.

Exploring Costa Rica scholarly records by interactive knowledge networks: case of implementation of VIVO software in Universidad Nacional, Costa Rica -

Andrea Mora-Campos, David Hine (Universidad Nacional de Costa Rica)

VIVO - is software from Cornell University, to develop knowledge networks with interactive elements to visualize the intellectual production of the academic community, to development Open Science.

The implementation of this platform in the Universidad Nacional (UNA), Costa Rica, allowed incorporating science areas according to Frascati, used for indicators and to visualize areas of impact in research, research collaborations with other universities and institutions, amount of publications by academic faculties, production typology, comparisons since there was no tool in which this information could be displayed. The Academic Network of the UNA was fed from the valid internal information systems, later they were unified in a single information matrix, which was carefully chosen. A process of retrieval of publications was carried out through APIs of Scopus , Web of Science, internal repositories and a crawler developed for the recovery of metadata.The initial information import was done in blocks using the Karma application to generate the models, applying the ontologies to the data sets according to the needs of the UNA.This is the first initiative that implements this type of software for the visualization of this type of information, with a distinctly Costa Rican effort and the first university in Costa Rica to do so.

Towards a Brazilian Geoinformation Suite for Rational Use of Fertilizers in Agriculture - Ronaldo Pereira de Oliveira (Embrapa Solos, Brazil)

Soil fertility directly influences the quality and quantity of food that can be produced, as the rational use of fertilizers provides essential macro- and micronutrients and attenuate negative environmental impacts. However, soil fertility information in Brazil lacks on integrated data management and detailed soil attribute mapping, which are essential to several agricultural and environmental issues. In addition, the national consumption of Nitrogen, Phosphorus, and Potassium (NPK) based fertilizers increases 6% per year, in a production sector that is already 79% dependent on imports. Therefore, Embrapa Soils and other governmental institutions in Minas Gerais State have proposed a cooperative initiative to develop a geoinformation suite supporting the rational use of soil fertilizers from regional to local farm scales. A preliminary model has been introduced aiming to provide information tools and decision support applications as means of territorial infrastructure planning for fertilizer distribution and sustainable use. Main challenges are the automation and integration of operational procedures, scientific methods and strategic planning workflows using IoT technologies to match tacit and academic knowledge. The FertilizaMG geoinformation suite is a pioneer initiative in Brazil, which is designed as a prototype solution to be expanded from state to national territory and customized to main agricultural production systems.

The panorama of open government data to research in Brazil - Neide De Sordi (Open Knowledge Brasil, Brazil)

This presentation focuses on the panorama of open government data available to research institutions in Brazil, especially those who are members of institutional portals for the free use, reuse and redistribution of data to society, with emphasis on the data available in the Brazilian Open Data Portal.The subject of open government data as a public policy established by the Law on Access to Information (LAI) is also addressed in this presentation. It addresses the Brazilian commitments included in the National Action Plans of the Open Government Partnership (OGP) aimed at opening data and open data use.This presentation also addresses issues related to the use of open data in

civic applications and digital public services, aimed at facilitating citizens' lives, expanding social control and the changes occurred in terms of public transparency, digital democracy and social transformation. Topics that have become of interest to civil society organizations, hackers, digital activists and other players who are aware of the possibilities that have arisen with the internet and with open data.

Linked Administrative Data Management for Research and Public Policymaking Purposes in Brazilian Public Health -

Bethania de Araujo Almeida, Paula Xavier, Mauricio Barreto (Fiocruz, Brazil)

The deployment of large-scale administrative databases for research purposes holds great potential. Administrative data becomes even more useful when linked to other datasets, making it possible to elucidate the effects and impacts of combined factors that could potentially impact the health of individuals and populations. Access, use and reuse of these administrative datasets, principally ones containing personally identifiable information, are all topics being widely discussed nowadays. In fact, as many countries are in the incipient stages of adopting access policies, we believe that our experience at the Fiocruz Center for Data and Knowledge Integration in Health (CIDACS) can contribute significantly, in both local and international contexts. Our institution has implemented a complex data management system that implements linkage practices between large volumes of data from multiple administrative systems for public health purposes. Our efforts involve the areas of data science, data curation and statistics to manage data in an attempt to assure data management best practices, as well as high levels of linkage accuracy. In addition, we are also in the process of defining institutional policies related to open data. For this reason we are conducting studies to better understand how open science is been implemented around the world. We seek to prioritize the protection of individual and collective rights and interests in our practical experience in the establishment of proper and effective ways to researchers access linked anonymized administrative datasets in attempt to answer complex questions, seek to make new discoveries and support policy making decisions.

Data management for health care and research: An experience of a leprosy reference centre in Rio de Janeiro -

Ximena illarramendi (Fiocruz, Brazil)

The Souza Araújo Outpatient Clinic (ASA) and the Leprosy Laboratory, Oswaldo Cruz Foundation are a Leprosy Reference Centre, credited by the Joint Commission International. It has a multidisciplinary team in charge of the health care of people affected by leprosy and the development of clinical, basic and translational research.

Both health care and research activities have generated a considerable amount of data produced during more than two decades that were, for a long period, stored independently in numerous fractioned and unrelated files. Several researchers and clinicians created the archives according to personal interest along the years, using various programs such as ACCESS™ and the freely available EpiInfo 6.0 (Center for Disease Control).

There was a clear need of having an integrated view of the patients' health condition and care, of gathering all of the Centre's clients data in a single database, and to improve the quality of data for use in scientific research and care. Thus, a system for data management, Sistema ASA, was developed in order to have a strategic platform to integrate, share and allow the use of all information captured in the clinical, laboratory, demographic and epidemiologic data repository.
The system was developed in mysql and is hosted in the local institutional server. It allows the use by various operators with different access profiles. The development and implementation of Sistema ASA was a challenge. It required intense involvement and learning by several of the ASA professionals and the contract of an IT firm. The data quality and administration, as well as the efficiency of information production significantly improved with the use of Sistema ASA, which has required extensive training of personnel.

Machine learning techniques to find correlations between open data and confidential data
Laci Mary Barbosa Manhães (Universidade Federal Fluminense – UFF, Brazil)

The Brazilian Education System generates large amount of semi-structured data. The INEP and the Ministry of Education (MEC) have a key role in education. INEP collects and maintains massive datasets from all public and private educational institutions, ranging from basic to higher education. Such as: School census for the calculation of the Basic Education Development Index – IDEB; National Student Performance Examination – ENADE; National Exam for Certification of Young and Adult Skills - ENCEJA and others. This vast amount of data is available, but there are still few institutions that analyze these data to turn them into useful information to direct the work of teachers and academic managers. Educational data, when correctly analyzed can bring benefits to institutions, students and teachers, such as: reduction of school drop-out rates, identification of students who cannot complete the course within the average time to complete the course, policies of inclusion in the Brazilian educational system, etc. Currently school and academic managers act as mere tools to feed these databases. Little or almost no information is returned back to facilitate school or academic management. Data Science applied to educational data brings together the contributions of several areas of study (Mathematics, Statistics, Machine Learning, Artificial Intelligence, Database, Information Retrieval, Visualization of Information among others) to properly explore the data, aiming to useful information. Our work uses machine learning techniques to identify the correlations between available public data and confidential access data to school and academic managers, this correlation of public and confidential databases could create an individualized profile for each student.

Opportunity and challenges to deal with public sector information management in science
Renato Cerceau, Luis Alfredo Vidal de Carvalho (ANS, Brazil)

The open data strategy has been worldwide spread as an instrument to promote the society benefice and the integration between actors to cooperation. Since 2011, Brazil has a legal instrument to promote open access to public sector information management, the Act 12.527. In some cases, this public agent act only as a depositary, an information custody agent. This study evaluates the

trajectory of a federal regulatory agency to open data. Since 2013, for various moments, some data was requested using the prerogatives of this law, with several denials of supply by custodian govern. The requirements for database access were intensified in the years 2016 and 2017, and finally several contents were granted only after receiving the support by the Union General Comptroller, in the fourth administrative instance. On average, the databases have taken eight months to be made available, requiring repeated and repeated requests for rectification of the databases provided to the citizen. Finally, in 2017, the institution starts its process of providing open data, also by determination derived from federal regulations for the production and delivered to the society. As a proof of work, we show that dealing with open data remains a challenge to public sector information management.

Cross-border epidemiological data integration and harmonization Application to malaria in the cross-border area between French Guiana and Brazil -

Emmanuel Roux, Raphael Saldanha, Christovam Barcellos, Théophile Mandon, Margarete do Socorro Mendonça Gomes, Emilie Mosnier, Basma Guarmit, Jean-Christophe Desconnets (ICICT - Fiocruz, Brazil)

The number of malaria cases drastically dropped worldwide between 2000 and 2015, with 58% and 37% decreases of the mortality and incidence rates, respectively. Such a success conducted, in 2015, the statement of an ambitious target in the framework of the Sustainable Development Goals of the United Nations: “By 2030, end the epidemics of [...] malaria [...]”. However, several obstacles make actually such a target difficult to reach. First, malaria is still highly present worldwide, with 216 million cases and 445 thousand deaths worldwide in 2016 (WHO, 2017). Secondly, local recrudescence of cases have been noticed in 2016 and 2017, especially in the Americas. Particularly, Brazil notified 174,522 cases between January and November 2017, i.e. 56690 cases more than for the same period in 2016, representing a 48% increase (PAHO, 2018). Moreover, specific contexts favour the maintenance of foci of intense transmission. Particularly, “cross-border malaria” is considered as a major obstacle for elimination. In fact, on either side of the border, different public policies and different strategies and means of surveillance, prevention and control of diseases coexist, without regular exchange of comparable data and information, within socio-demographical and environmental contexts that already vulnerabilize the local populations. Such a situation prevents to have a shared and unified vision of the epidemiological cross-border situations and, consequently, to define concerted (between the border countries), targeted and effective control actions.

“Transform malaria surveillance into a core intervention” is one of the three pillars on which the Global Technical Strategy of the World Health Organization (WHO) bases the actions towards the disease elimination. However, surveillance remains a challenge in cross-border contexts due to above mentioned issues, making necessary the harmonization of data types and formats, nomenclatures and concepts used by each country.

The border between French Guiana and Brazil is typical of the above described cross-border contexts. Systematic, regular and perennial information and data exchange related to malaria cases does not exist yet. Consequently, a cross-border malaria surveillance system has been built, in the framework of the Brazilian Climate and Health Observatory. It is based on the existing country-specific surveillance systems: the Epidemiological Surveillance System for Malaria (SIVEP-Malária) in Brazil

and the epidemiological surveillance system of the department of Delocalized Centres for Prevention and Care (CDPS) of the Cayenne hospital, French Guiana. It relies on expert knowledge in parasitology, epidemiology, national surveillance systems and informatics in order to specify data transformation rules leading to cross-border harmonized information. Nomenclatures chosen satisfy as far as possible international standards. Extraction, Transform and Load (ETL) tools were used in order to implement the rules and build a database containing harmonized information and shared by the two countries. Data visualisation tools implemented within the R-Shiny programming environment permit to disseminate the harmonized epidemiological indicators online, both temporally (time-series) and spatially (maps). Eventually, this system aims at being operationally implemented, for use for research works, epidemiological surveillance and general public information.

The poster will present the societal and scientific issues involved in establishing a permanent data flow. The technological choices made to implement it, as well as illustrative examples of harmonized data and indicators will be presented and discussed. Eventually, some perspectives will be drawn.

Acknowledgements: 1) Project “Fighting malaria: from ‘global war’ to ‘local guerillas’ at international borders”, Grand Challenges Explorations Round 18 (GCE18) of the Bill and Melinda Gates Foundation (Investment ID OPP1171795). 2) Project GAPAM-Sentinela (Guyamazon program: IRD, CIRAD, French Guiana territorial collectivity, France Embassy in Brazil, Foundations for research support of the Brazilian states Amapá, Maranhão, Amazonas).

Hydrophysical Database For Brazilian Soils -

Marta Vasconcelos Ottoni, Theophilo Ottoni Filho, Marcel Schaap, Maria Leonor Lopes-Assad, Otto Rotunno-Filho (Geological Survey of Brazil)

Soil water retention and hydraulic conductivity data are fundamental in soil modeling. Direct measurement of this information demands high costs and laborious fieldwork. The development of pedotransfer functions (PTFs) has been considered a powerful tool to predict these two hydraulic variables from soil attributes more easily available. The elaboration of PTFs requires representative databases containing information on potential predictors and corresponding hydraulic properties to be estimated. The objective of this study was to develop a hydrophysical database for Brazilian soils (HYBRAS) aiming at the establishment of PTFs suitable for Brazil. HYBRAS currently has 14 related tables, with information on the sampling site and with the description of the adopted analysis methods of physico-chemical attributes and hydraulic properties. Soil granulometric fractions, bulk density, organic matter content, saturated hydraulic conductivity and water retention in a wide suction range, as well as derived data, were included in HYBRAS. Adjusted parameters of the van Genuchten equation for water retention were also included. Data from 1075 soil samples were recorded in HYBRAS. The states with the highest data expression were Rio Grande do Sul, São Paulo and Rio de Janeiro, and the predominant soils were Ferralsols and Acrisols. The soil variables recorded in HYBRAS covered a wide range of values, obtained through consistent and well defined determination methods. Soils with a high content of organic matter (> 60 g Kg-1) and low bulk density (<0.8 kg dm-3) were poorly represented, as were those with high silt content (> 500 g Kg-1). HYBRAS

involved a wide, varied and consistent set of soil hydrophysical data that can be considered useful for the development of Brazilian PTFs.

Session 2: Case Studies in Biodiversity, and Climate

PPBio's Metacat Data Repository –
Timothy Lee Vincent (Instituto Nacional de Pesquisas da Amazônia – INPA, Brazil)

The Western Amazon Biodiversity Research Program (PPBioAmOc) maintains an online data repository used by researchers from several regional programs. It is member-node of the Earth Data Observation Network (DataONE). The repository's data are also harvested and made available through the Brazilian Biodiversity Information System (SiBBr).
PPBio elected to use Metacat because it is flexible and open-source and can be linked to the DataOne system. Unlike worksheet style data repositories, Metacat allows a wide variety of information to be uploaded; maps, scans of raw data, photographs and so on.
Morpho is used for preparing and uploading the dataset and to update or add information and make controlled access data publicly available. This standalone tool allows researchers to prepare their metadata and data without the need for an internet connection.
Datasets prepared by researchers are sent to the data manager who checks that metadata has been entered correctly and that there are no issues with the data before uploading to the server. This human intervention in the process is very important for quality control.
Correctly entered metadata influences how easily the data-set can be discovered in the future. Metadata may describe data using an ad-hoc set of descriptive terms and there may be subsequent issues with recall. Morpho enables the data to be described using terms that facilitate retrieval, but this must be correctly done by the person entering the data.
The PPBio website provides detailed instructions on how to use Morpho, but person-to-person training is very useful since there are some small hacks that are necessary in order to resolve issues with Morpho's user interface and the terminology used for defining and describing the data.

Big Data Landscape to Improve LBA Scientific Data Management –
José Laurindo Campos dos Santos, Andréa Corrêa Flôres Albuquerque, Kleberson Junio do Amaral Serique, Daniel Lins Silva (Instituto Nacional de Pesquisas da Amazônia – INPA, Brazil)

The term Big Data began to appear in moderation in the early 1990s, and its prevalence and importance have increased exponentially as the years went by. Today, Big Data are often seen as an integral part of an organization's data strategy. Ten important data characteristics and properties, known as 10 V's, are consolidated to prepare for both the challenges and the benefits of large data initiatives. The characteristics include: volume; velocity; variety; veracity; value; vagueness; vocabulary; venue; variability; and validity.

Due to open scientific questions and complex scenarios regarding the Amazon, the Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA Scientific Program) faces huge data management challenges.
The current Big Data landscape has shown to be very effective in providing well suited technology, allowing us to consider adopting in LBA in order to move towards better solutions that can overcome limitations in infrastructure (nosql/newsql databases, cloud, governance, etc.) and analytics (data analyst/science platforms), preferable across different computer platforms. Additionally, the wide range of open source solutions and free access to a large number of data resources, is very appealing for the LBA scientific data management environment.

A unified South American Paleohidrological Database: LOTRED SA –
Juliana de Sousa Nogueira (Universidade Federal Fluminense)

Recent paleoclimate advances, especially the creation of high-standard regional proxy record databases, allow describing South American climate from a new perspective. However, large areas of tropical South America are still underrepresented in those databases. The PAGES/LOTRED-SA (Long-Term climate REconstruction and Dynamics of South America) group aims to collate existing and new multi-proxy paleoclimate data sets for the last ca. 1000 - 2000 years available for South America in order to create an unified database to be used as a source for climate works. This project started in 2006 and in this last phase, new precipitation data were added to its database in the context of this revision, summing up more than 360 metadata entries. Upon availability, we gather the proxy precipitation data and selected the ones used in this paper from various matrix (tree rings, ice cores, instrumental river quota, lake sediments, documents based index, coastal marine sediments, bog sediments and speleothems). The dataset was divided between pre-industrial period (0-1750 AD) and industrial period (1750-2016 AD). As we used different proxies from many matrix, the data was normalized and than analyzed for its cyclity (wavelet analysis) and behavior (empirical orthogonal function - EOF) for the main historical periods, such as Roman Warm Period – RWP, Dark Age Cold Period – DACP, Medieval Climate Anomaly – MCA, Little Ice Age – LIA, Modern Warm Period – MWP. The aim was to fill the existing gaps, allowing to recognize priority areas to produce new data and also to produce a South American hydro-climate reconstruction.

A snapshot of glacier monitoring in South America —
Isabelle Gärtner-Roer (World Glacier Monitoring Service - WGMS)

Glacier observation data are key to improving our knowledge of glacier changes: they deliver fundamental baseline information for the understanding of climatological and hydrological processes. Glacier monitoring enhances the awareness of the populations that depend on water resources from glacierized mountains or that are affected by hazards related to glacier changes. It has therefore been suggested to include it in the development of sustainable adaptation strategies in regions with glaciated mountains.
We present a standardized assessment of the monitoring status of all South American countries with glaciers (Argentina, Bolivia, Chile, Columbia, Ecuador, and Peru) with the aim to evaluate the national

implementation of the Global Hierarchical Observing Strategy (GHOST), an internationally agreed monitoring strategy elaborated on behalf of the United Nations Framework on Climate Change Convention (UNFCCC). The country profiles are established based on glacier fluctuations series as well as glacier inventory data. The country profiles summarize the present glacier monitoring state in the respective countries and contribute to the programme on Sustainable Mountain Development for Global Change (SMD4GC) that was initiated to support mountain populations to increase their resilience in the context of global climate change.
The assessment is expected to increase the visibility of existing glacier datasets and ongoing monitoring activities, to define potential deficiencies and needs for sustainable glacier monitoring, and to elaborate tailored recommendations under special consideration of possible future impacts of glacier changes. Related issues such as hydrological or ecological measures within climate-change adaptation strategies might also profit from the assessment.

Initiatives on Pollinator and Pollination Data Digitization and Sharing and Data Quality – Antonio Mauro Saraiva (Universidade de São Paulo, Brazil)

Pollinator have a special relevance in the biodiversity domain because of their role on both natural and agricultural systems. We have been working on the standardization, digitization, sharing and aggregation of other types of data besides species occurrence data: pollen and plant-pollinator interaction data. The Online Pollen Catalogs Network – RCPol (www.rcpol.org.br) addresses pollen grain collections, which are important as indicators for conservation, for agriculture and for forensics. RCPol includes digitization of pollen collections from countries in the Americas and Europe, interactive keys for the identification of plants based on pollen and plant descriptors, and pollinators-plant interaction data based on pollen data. Biological interaction data is also very important to the understating of the relations between species and among communities, and although a lot of data has been collected they are hardly found online due to the lack of digitization, and sharing and aggregating them is hindered by the lack of a data standard. To tackle that problem two initiatives are in place currently. At the Biodiversity Information Standards (TDWG), an interest group (IG) has been created to try and develop a data standard for biological interaction data. In parallel the Brazilian Network on Plant-Pollinators Interactions (www.rebipp.org.br) is working to develop a proposal of a data standard for the specific case of plant-pollinator interaction data. Finally, in order to harmonize the way Data Quality is handled in biodiversity informatics, we have developed a framework that encompasses user's needs, tools and reports on data quality, under the concept of fitness-for-use of data, also hosted as an IG at TDWG (www.tdwg.org) and (https://github.com/tdwg/bdq).

Special Session: Accreditation of Scientific Data Repositories

- Why and how promote the certification of Scientific Data Repositories –
  Ingrid Dillo (Vice-chair of ICSU-WDS and Data Archiving and Networked Services: DANS, Netherlands)

- The accreditation process of ICSU World Data System —
  Rorie Edmunds (ICSU World Data System International Programme Office, Japan)

Session 3: Case Studies in Astronomy, Space, and Private Sector

The LIneA Science Portal: Handling the Large Volumes of Data from Modern Astronomical Surveys –
Luiz Nicolaci da Costa (LIneA/Observatorio Nacional, Brazil)

In order to handle the large volumes and variety of data being generated by modern astronomical surveys LIneA's IT team has developed over the past 10 years an innovative web-based framework integrated to a relational database. The main objective is to enable astronomers to carry out comprehensive and timely analyses of the data in a collaborative environment addressing the data management (transfer, storage, visualization and data mining) and processing needs associated. In this presentation some of the tools developed are presented and their use in ongoing and future surveys described.

The implementation experience of the Chilean Virtual Observatory –
Mauricio Solar (Universidad Técnica Federico Santa María, Chile)

The Chilean Virtual Observatory (ChiVO) seeks to offer, under standards and protocols of the International Virtual Observatory Alliance (IVOA), the public data of the observatories installed in Chile openly to the community, both to download and process in our datacenter as a cloud service. ChiVO is running in a scientific datacenter with a high storage capacity (1 PB), in addition to the capacity to perform HPC. An infrastructure for the process of ingestion, curation, storage and processing of astronomical data has been implemented on this hardware architecture, which considers the complexity inherent with this data, i.e. the cubes of ALMA data can reach tens of GB, in addition to the problem of the dimensionality of cubes axes. The pipeline of ingestion includes the parallel processing of the FITS using multithreading, necessary to parsering the large amount of metadata present in the header of these files. Currently, ChiVO is in full operation with ALMA public data of cycles 0, 1, 2 and 3 available. In addition to offering fully interoperable search services with IVOA networks, as part of our curation policies we have created the prototype of a library based on Tensor Decomposition that aims first at achieving supercompression rates, reaching space savings of 90% but maintaining the multidimensional geometric structure. In the scope of the processing taks, new libraries with different purposes have been developed, such as Astropy, ACALib and CASAC libraries to the Python kernel, and we are currently porting CUPID (Starlink), MPICASA and ADMIT through suitable wrappers, all through an interface for the interactive analysis of data through Jupyter notebook.

The United Nations Open Universe initiative and the Brazilian Science Data Center –
Ulisses Barres de Almeida (Centro Brasileiro de Pesquisas Físicas - CBPF)

Our times are characterised by a confluence between unprecedented rates of data production and a similarly exceptional capacity for exchanging information, which demand new paradigms for data management in all areas. Space Science is no strange to these challenges. Although it has always been at the forefront of data science technology, much still needs be accomplished to meet the expectations and demands of the near future. Many initiatives across the globe, such as the Virtual Observatory, have developed essential technology for data management, exchange and interoperability, which form the base for facing future challenges. While such frontier initiatives are mostly circumscribed to the sphere of specific or specialised applications in space research, additional steps are required towards answering the global demand for openness and growing accessibility to knowledge. In fact, Space Science has the potential to be a powerful driver of development and education across the Planet, bridging gaps between developed centres and the peripheries. With this purpose, the United Nations is launching the "Open Universe" Initiative. Initially proposed by Italy, and with direct participation of Brazil since its early stages, through the Brazilian Science Data Center (BSDC) at CBPF, it aims to increase the reach of space science data worldwide and across all spheres of society, harnessing the full potential of space as a driver for democracy, equality and capacity-building for the future. The Initiative will coordinate players across the globe to develop and foster new data science technologies, paradigms and standards. In this talk I present in detail the "Open Universe" Initiative, focusing on the BSDC activities, and on the potential benefits and opportunities for Latin America.

Research Data Management in Nuclear: the experience of the Nuclear Engineering Institute Area –
Luana Farias Sales (Comissão Nacional de Energia Nuclear -CNEN-IEN / IBICT)

The Institute of Nuclear Engineering (IEN) of the Brazilian Nuclear Energy Commission (CNEN), has put together technology and managerial resources in the development of a digital repository that supported the custody and preservation of nuclear knowledge. The main goal of the repository project was to share and reuse data and information from the nuclear area at a national level, making this data the basis for new research activities. This repository was initially created to jointly archive publications and scientific data produced by IEN research activities. However, the repository has also proved to be an important laboratory for studies in Information Science, serving as a basis for research focusing on issues belong the knowledge organization. By means of this research laboratory in information science it was possible to empirically observe problems that have arisen over the time regarding management, organization and recovery of this data, for example, the growth in the number of metadata required to indexing data and publications.To solve this problem, was proposed in partnership with the IBICT to separate the repository into two distinct platforms: One managed by the Dspace where the publications will continue to be archived and a second one in Dataverse designed to archive the research data. Currently the data sets are in processing of migration from Dspace to Dataverse. The future perspectives include integrating, through participation in the Cariniana Network, the dataverse into the Archivematica system. With these integration, the

research data repository will become trusted for the knowledge preservation, which is of great relevance for the continuation of research in the field of Nuclear Sciences in Brazil.

Research Data Management in Exploration and Production context: opportunities and challenges —

Dean Pereira de Melo, Marcelo Fagundes de Rezende (Petrobras)

Data management function is to support business needs proactively through processes, people, and technology focusing on planning, controlling and delivering data and information. Exploration and Production companies (E&P) need data management because of tough regulations, their decisions rely on data, new reserves are a need, and exploit these effective and safely is vital. Several complex data domains in E&P require appropriate structures to control and maximise data use to meet business goals. The framework adopted by E&P Companies varies, but includes crucial elements as why data management is necessary, what problems should be solved, and who has accountability to make data decisions. Rules and definitions clarify how to solve these problems. In general, authority under few people is risky, so data stewardships and governance offices share the responsibility. Lastly, the data governance process controls when each step is done. There are challenges to the governance framework, such transferring full process responsibility to IT teams. Usually, this approach fails, as only business experts know how concepts connect to data issues. Also, data managers get burden by the data maturity level and structure, leading to customized projects to preserve data or its products. Petrobras E&P to solve these problems keeps three business data management offices. Besides these, IT offices support daily activities as custodians. The technological strategy is an integrated database, ensuring data quality and conceptual unicity. The database has environments to retain data and to retain files (i.e. models and reports). Therefore, sharing data, interoperability among databases and access grant are challenges that collaborations between Petrobras and other institutions have to overcome.

Session 4: Case Studies in Health and Humanities

The Brazilian Initiative on Precision Medicine (Bipmed): The First Publicly Available Genomic Database in Latin America –

Iscia Lopes-Cendes (University of Campinas – UNICAMP, Brazil)

BIPMed is an initiative of five Research Innovation and Dissemination Centers in Brazil.The BIPMed genomic database is the first product released to fulfill the increasing need for publicly available genetic and genomic information. BIPMed is also part of the Beacon Project (the Global Alliance for Genomics and Health), it is the Brazil node of the Human Variome Project, as well as a founding partner of LATINGEN, the Latin American Database of Genetic Variation.(www.latingen.org). BIPMed genomic database is based on a software platform, the Leiden Open Variation Database and it was implemented as a fully web-based gene sequence variation database. The design of the database follows the recommendations of the Human Genome Variation Society and the principals and

guidelines of the GA4GH for the ethical and responsible sharing of genomic and clinical information. Currently, the database (www.bipmed.org) contains variants detected using whole exome sequence as well as SNP-genotyping from reference population ascertained based on their geographic origin in Brazil. In this dataset we identified over ten million variants in 20842 genes. There were 209 variants which were unique to the population studied. In addition, our genomic database has attracted world-wide attention and it has been accessed by an average of 150 users daily from countries all over the globe. We expect the database to grow fast and include data-sets from disease cohorts as well as other types of –omics data. This platform, the first of its kind in Latin America, is intended to be used by clinicians and scientists worldwide, to obtain and share information about various aspects of genomic medicine and human health, as well as to support dissemination and training. Support: FAPESP.

The Center for Data and Knowledge Integration for Health (CIDACS-Fiocruz) — Mauricio L. Barreto (CIDACS, Fiocruz, Brazil)

The humanity faces huge health challenges (p.ex Non-communicable Diseases, new epidemics) and the present health research and development system has been lengthy and inefficient. By the other side the health system produces a huge and growing amount of data. Has been produced some preliminary evidence that the use of such data can be an advancement regarding finding solutions to some of the questions faced. Some countries have developed policies to use the data produced by the health system. The Center for Data and Knowledge Integration for Health (CIDACS-Fiocruz) is an experience based in Salvador-Ba, build up as a resource to put together different Brazilian social(p.ex. Cadastro Único) and health (births, deaths, infectious diseases, etc.) identified databases. By linking different databases it is plan to design large studies aimed to answer relevant health research and managerial questions. An example is the 100 million cohort designed to evaluate the impact of different social protection programs on health.

Setting up a data center of this magnitude requires careful preparation. The space needs to be carefully designed according to rules to give full physical protection for the data as well as to manage access to the relevant personnel. The computing capacity for data with identifies and for anonymized dataset, as well as providing access needs to be carefully defined. The software must be appropriate for the magnitude of the data base and the functions of the center, ranging from receiving the data, managing it, cleaning, linking, updating, maintenance, and provision of anonymized datasets and carefully documenting meta-data and recording access to the dataset. All activities must be carefully planned and codified in standard operating procedures.

MaoBD - Open-access data in forensics anthropometry of Brazil – Maria Elizete Kunkel, Thiago Nunes, Falipe Granado, Flávia Cristina Mariano (Universidade Federal de São Paulo, Brazil)

In all fields of science, the development of open-access data sharing resources has increased the amount of available data and quantitative statistical analyses. Forensic anthropology provides human

identification as the biological profile of an individual (age, sex, stature, ancestry). Anthropometry is the basic tool of anthropology in forensic sciences. Unlike other countries, Brazil does not have reliable and comprehensive anthropometric databases of its population. Therefore, the national experts still use methods and techniques of human identification based on regression equations developed with population data from other countries. Since the Brazilian population presents a differentiated anatomical constitution, due to the miscegenation of different ethnicities, the use of methodologies based on other populations may lead to incorrect results. MaoBD is the first significant collection of multimedia information of anthropometrical data, especially hands (linear measurements and images) and stature of 1.000 volunteers of Sao Paulo. As Brazil is characterized by different groups of populations in different regions, to create a national anthropometric database, the acquisition of measures from volunteers from Brazil will be carried out with the collaboration of several universities. A methodology was developed to favor the standardized acquisition of hand measurements with low resources and reduction of the time and associated error. The MaoBD data will be available in an open platform for statistical forensic analysis. In addition, the data from MaoBD have been applied in other areas as automatic acquisition of hand parameters by artificial intelligence and development of hand prostheses and orthoses. MaoBD provisionally is available at www.biomecanicaeforense.com

Music in eighteenth-century periodicals —
Martha Tupinambá de Ulhôa (Universidade Federal do Estado do Rio de Janeiro, Brazil)

Music in eighteenth-century periodicals is an online database containing news and commentary on music in nineteenth-century periodicals microfilmed by the National Library (BN), Rio de Janeiro. It has served as support for research on musical genres such as modinha, lundu and the waltz, as well as on musicians, the music press, musical theater and opera in the 1800s. At the beginning of the research, in 2002, microfilm data collection was slow because of the small number of microfilm reading machines or the restriction of the library opening hours. The ability to feed the database remotely via the internet was an improvement over handwritten transcription and typing on the computer. With the launch of the Brazilian Digital Newspaper Library (HDB) by BN, in July 2012, the possibilities of research were expanded exponentially. The query on the approximately 2,000 HDB titles can be done by title, period, edition, place of publication and keyword. Several researchers have used it as a source, enabling the writing of several dissertations and theses. The pity is that this enormous amount of information collected could be being studied by another angle (s), which does not happen, since they are not available to the public. Therefore, the proposal to open the bank to other Brazilian and Latin American researchers to insert their primary data collection on music in the nineteenth century, so that other researchers in any part of the world can have access to this data to reuse them in their investigations. In order for the database to have an innovative profile, it is necessary to align its structure and operation with the criteria and standards of network data sharing - the so-called e-Science or open science.

Intelligo: Exploring Large Science and Technology Data Collections Using Online Semantic Maps –

Lautaro Matas, Rodolfo Barrere (Observatorio Iberoamericano de la Ciencia, la Tecnología y la Sociedad, Argentina)

Intelligo is a technology for exploration of scientific and technological data that offers a different way of navigating and accessing data collections. The technology developed by the "CTS-OEI Observatory" allows the generation of real-time maps of relevant topics based on user queries, allowing the exploration of large volumes of data.

Intelligo analyzes document texts using natural language processing techniques, which automatically extracts the most relevant concepts and normalizes the metadata available for each item in the collection.

For a given repository or source of data a latent semantics model is trained, this model allows to infer semantic relations based on the contexts shared directly or transitively in the whole corpus.

Using a web browser, the users can perform queries as in a standard search engine, the results are visualized as graphs (concept maps) generated in real time using a clustering algorithm that group and organize the topics. In addition, frequency graphs are generated on the metadata, which allows descriptive analysis of the authors, disciplines, institutions, years and among other metadata fields. It also allows access to documents in their original source, working as a content aggregator also.

Currently Intelligo technology feeds two free access portals: 1) Intelligo Patents (http://patentes.explora-intelligo.info/): focused on WIPO PCT patent documents (2007-2016) that allows analyzing 1.7+ million documents; 2) Intelligo Repositories (http://repos.explora-intelligo.info/): which analyzes documents from the most relevant open access portals in the region, such as SciELO.org, REDALYC, LAReferencia, CSIC (Spain), among others aggregating 2+ million documents.

A short video demonstration (in Spanish) can be found at https://vimeo.com/99931286

The Intelligo development team is currently working on a new version of the technology that will be presented at the workshop. The exploration of communities of people and their relationships for large collections of metadata is being developed. The pilot project is being developed using the contents of the "Brazilian Digital Library of Tesses and Dissertations (BDTD) of the Brazilian Institute of Information in Ciência e Tecnologia (IBICT).

Panel 3: Infrastructure, training and funding of Scientific Data initiatives

RNP support to data-driven research —

Leandro Ciuffo (RNP, Brazil)

Brazil's National Research and Education Network (RNP) has been actively engaged in supporting Brazilian institutions to participate in scientific collaborations, especially addressing the need of transferring large amount of data over long distance networks. In the past few years, RNP's R&D division developed, in cooperation with academia, a set of tools to optimize the network usage for

transferring large datasets.
In addition to supporting researchers moving data, recently RNP started an R&D project, in cooperation with 2 universities (UFRGS and FURG) and the National Institute of Information in Science and Technology (IBICT), to seek technological solutions in order to deploy a research data repository. It also envisions bringing together individuals and institutions with needs to share knowledge and practices on Open Access to Research Data (AADP, in Portuguese).
This project just started and initially will last 1 year. The two initial activities are: a) identifying AADP practices in Brazilian institutions; b) mapping AADP users and their needs. To meet these objectives, a survey was designed to collect data directly from the Brazilian researchers, from all areas of knowledge.
The survey aims at identifying the researcher's perceptions and practices in AADP, as well as unveiling alternative solutions and practices in management, storage and dissemination of research data. It also intends to verify the demands for the incorporation of data in research data repositories and the existence of potential interest in participating in a nationwide initiative.
The proposed study aims at presenting some of RNP's solutions to assist data-driven research as well as the results of the survey. The presentation will also discuss RNP's future plans to support e-science projects.

Scientific Data Analysis at LNCC, Towards a Generic Platform – Fabio Porto (LNCC, Brazil)

At LNCC Data Science Research Group, we have been facing challenges in scientific data management and analysis applied to different scientific domains, such as: astronomy, systems biology, health science, bio-diversity, sport science, oil & gas etc... Despite their particular domain bias, from a pure data viewpoint we observe that managing and analyzing large amounts of scientific data guards some basic principles and requirements that can be explored when developing computational tools and techniques.
As an example, the analyses of observation and simulation results of natural phenomena take advantage of spatial-temporal techniques that can be better explored using multi-dimensional matrices as the fundamental data representation, irrespectively of the scientific domain in which it is applied. Similarly, when employing a scientific in-silico experiment, scientific dataflows become the generic, de-facto, tool for integrating processes, while keeping provenance data and supporting reproducibility and human-in-the-loop techniques. Moreover, whenever the research integrates different pieces of data autonomously produced trying to put together a global integrated view, different co-relation algorithms producing new knowledge can be evaluated, leading to Knowledge Construction systems.
Identifying the fundamental traces that should guide the design of such a generic platform is, however, not obvious and requires experience, experimentation and study. Most probably, there should be some functional intersection among the members of the platform, which by no means jeopardizes the main intention of designing it.
Thus, in this paper we aim at presenting the current view on the scientific data management and analysis platform in development at LNCC and collaborator institutions. We expect this work to offer

scientists from different domains a generic platform from which to pick the fittest solution to each problem.

WDS China Data Centers Activities and the Common Clearing House — Juanle Wang (Institute of Geographic Sciences and Natural Resources Research, China)

There are 9 regular member data centers in China under the umbrella of the International Council for Science-World Data System (ICSU-WDS). They are Chinese Astronomical Data Center, WDC-Renewable Resources and Environment, WDC – Oceanography at Tianjin, World Data Centre for Microorganisms, Chinese Space Science Data Center, Cold and Arid Regions Science Data Center at Lanzhou (CARD), WDC for Geophysics at Beijing, Global Change Research Data Publishing and Repository (GCdataPR), and Fish Database of Taiwan. The presentation will give a whole picture of 8 data centers in China mainland, including the host agency, development history, main field, core data resources, special services and its domestic and international profile of each centers. Their main contributions for the international communities are summarized and highlighted. Although WDS China data centers have made obviously progress and achievements in these years, due to the differences between subject areas and the lack of associative mechanism, WDS-China data portal and metadata exchange system has not built yet. In order to solve this problem, in this presentation, the frame of WDS China Common Clearing House was preliminary put forward and the prototype system was built. A separate metadata server was built based on pycsw open standards. The metadata capture module was built based on data harvesting. Through the above development WDS China Common Clearing House has had functions including document releasing and updating, Wiki knowledge releasing and updating, user right management, website full-text retrieval, back-end information statistics, system of online presentation, metadata releasing/push interface, etc. At present the established prototype system has already preliminarily applied in the Renewable Resource and Environment data center of World Data System (WDS-RRE) and will be released for all WDS members in China soon.

Posters

Renewable Energy Generated by the Impacts of Natural and Accidental Disasters — Fátima Antonethe Castaneda Mena (UNESCO CON E ECT, Guatemala)

The worldwide use of fossil fuels as the major energy source, increases the impacts of climate change. Fossil fuel industries own a critical role in the climatic alterations, global temperatures and sea level rises [8, 9]. For that reason, the outcome of the Paris Agreement within the United Nations Framework Convention on Climate Change clearly warns all the nations that the global average temperature rises should never increase more than 2°C. Failing to comply with this direction will result in an on-going environmental catastrophe [10]. In order to prevent this scenario, the most realistic approach is to restructure the world energy usage; specifically, to reduce the fossil fuel production and maximize the potential use of renewables [11, 12]. Hydropower that is currently the largest renewable energy source could be used to achieve a successful change.

The many applications of hydropower and its potential in disaster management have not yet been exploited, mainly due to the disproportionate cost to benefit relationship. The Renewable Energy Generated by the Impacts of Natural and Accidental Disasters (REGINA) research proposes an alternative idea for the partial utilization of the potentially lost energy during water-based disasters. More specifically, it introduces the conceptual model of a mini (or smaller-scaled) hydropower generator that includes early-warning and alarm systems appropriately designed at the local level, as well as a water purification and storage space. This unit is designed to provide support during both normal and extreme cases to all the vulnerable populations that reside in remote areas with poor disaster resilience and energy insufficiency.

Database Crimes May 2006: towards the establishment of forensic anthropology and transitional justice in human rights in Brazil –

Maria Elizete Kunkel, Javier Amadeo, Cláudia R. Plens, Raiane S. A., Bruno Comparato, Camila D. Souza, Thabata Ganga, Natália A. Santos, Marina Figueiredo, Rebeca Padrão, Juliana M Carrapeiro, Edson B da Rocha, Débora M da Silva, Aline L. G. Rocco, Valéria A. de Oliveira, Delphine D. Lacroix, Lorrane Rodrigues, Bruno Rocha and Leigh Payne (Universidade Federal de São Paulo, Brazil)

The Database Crimes_May2006 is an open-access data collection created from a research project in collaboration between the Center for Anthropology and Forensic Archeology (CAAF) from the Federal University of Sao Paulo in Brazil and the Latin American Center from the University of Oxford in England. The transition justice is a field of knowledge under construction and very important to discuss questions related to de violence state and to improve the area of human rights. The goal of this project was to reanalyze the May 2006 crimes, when hundreds of civilians were killed in the Baixada Santista Region in Sao Paulo State in Brazil. For the creation of the database many original documents provide from official institutions and documents provided from de relatives of 60 fatal victims were analyzed. The largest number of victims were men (92%) with low economic income and 29% were between 20 and 24 years old. Earlier studies indicate that state agents are potential perpetrators of the crimes. The database allows an analysis of the circumstances of the crimes and brings to the fore the question of State accountability in investigating crimes and condemning their executors, such as the voices of the relatives of the victims of the Maes de Maio de 2006 Movement. The purpose of the Crimes_May2006 database is to facilitate the visualization and statistical analysis of these crimes. In addition, the analysis of the data allows future studies about improvement of the democracy in Brazil and new perspectives posed by transitional justice, forensic anthropology and investigation of violence. Making these data available can greatly reduce the cost of future scientific researches. Database Crimes_May2006 provisionally is available at www.unifesp.br/reitoria/caaf/

Regional WDS-Oriented Activities in the Asia-Oceania Area –

Watanabe Takashi, Iyemori Toshihiko, Murayama Yasuhiro, Li Guoqing (ICSU-WDS International Programme Office, Japan)

The first WDS Asia-Oceania Conference (http://wdc2.kugi.kyoto-u.ac.jp/wds2017/) was held at Kyoto University, Kyoto, Japan, on 27-29 September 2017. A cooperative event "Data-Analysis Workshop on Solar-terrestrial Environment" was held also in Kyoto University before the start of the conference under the promotion of the SCOSTEP community of Japan. The total number of participants is 107, including 68 domestic participants. The Conference tried to brought together data practitioners, data repositories managers and researchers to reinforce the data stewardship community in the region and help establish a collaborative system for access to and dissemination of research data. In this conference, data activities in Japan, China, Australia, India, Indonesia, and Thailand were reported. In the course of preparation of this conference, it was recognized also that international collaboration is inevitable to encourage long-term preservation and provision of data by their own hands because of insufficient manpower and infrastructures. Basing on reports and discussions in the conference, we reached to a common understanding on our further activity: (1) Establish a cooperative network of WDS-oriented data centers in the Asia-Oceania area to reinforce collaborations to improve current problems in data management; (2) Involve governmental data centers to our activity; (3) Collaborate with Future Earth and other scientific programs in this area; (4) Expand aerial activities on data to the world-wide community. The continuation of our activities has been recognized to be essential. We decided to have a next conference in China, 2019

More Information at: http://lacworkshop.icsu-wds.org/