CONCEPTUAL VIEW REPRESENTATION OF THE BRAZILIAN INFORMATION SYSTEM ON ANTARCTIC ENVIRONMENTAL RESEARCH

R Zorrilla[1]*, M Poltosi[1], L Gadelha[1], F Porto[1], A Moura[1], A Dalto[2], H P Lavrado[2], Y Valentin[2], M Tenório[2], and E Xavier[2]

[1]Extreme Data Laboratory, National Laboratory for Scientific Computing, 25651-075 Petrópolis, Brazil
**Email:* romizc@lncc.br
Emails:{maira,lgadelha,fporto}@lncc.br, anamaria.moura@gmail.com
[2]Instituto de Biologia, Universidade Federal do Rio de Janeiro, 21941-902 Rio de Janeiro, Brazil
Email: yocie@biologia.ufrj.br

ABSTRACT

Data generated by environmental research in Antarctica are essential in evaluating how its biodiversity and environment are affected by global-scale changes triggered by ever-increasing human activities. In this work, we describe BrAntIS, the Brazilian Information System on Antarctic Environmental Research, which enables the acquiring, storing, and querying of research data generated by the Brazilian National Institute for Science and Technology on Antarctic Environmental Research. BrAntIS' data model reflects data acquisition and analysis conducted by scientists and organized around field expeditions. We describe future functionalities, such as the use of linked data techniques and support for scientific workflows.

Keywords: Antarctic environmental research, Ecosystem informatics, Biodiversity informatics, Antarctic data management, Long-term preservation

1 INTRODUCTION

Increased availability of high-capacity sensors in various scientific domains is causing exponential growth in the amount of scientific data generated (Bell, Hey, & Szalay, 2009). Consequently, the acquisition, storage, querying, and analysis of such vast data demands the introduction of new data management techniques (Ailamaki, Verena, & Debabrata, 2010).

Biodiversity and Ecosystem Informatics data has shown a similar pattern of growth. In particular, humans have extensively changed global environments, affecting their biodiversity. Antarctica is no exception to this trend (Cook, Fox, Vaughan, & Ferrigno, 2005; Ingels, Vanreusel, Brandt, Catarino, David, De Ridder, et al., 2012) and has seen increases in air temperature and reduction of its glaciers. To precisely determine the extent and rate of biodiversity change, it is essential to gather, archive, and analyze data on spatial and temporal distributions of species as well as information about their surrounding environment (Michener, Porter, Servilla, & Vanderbilt, 2011; Hardisty & Roberts, 2013).

The use of integration techniques is extremely important in facilitating the discoverability and querying of these data, which can be generated in different locations and by different institutions. Data quality evaluation and improvement techniques can transform raw data collected during field observations into fit-for-use data that can be input to statistical analysis tools or biological system models for synthesis studies or generating predictions (Chapman, 2005). These analysis and synthesis routines should also be supported by scientific workflow management systems that automate many of the tasks involved in managing a computational scientific experiment (Deelman, Gannon, Shields, & Taylor, 2009), thus providing scientists the opportunity to dedicate a greater share of their time to actual scientific problems.

In this work, we present BrAntIS (**Br**azilian **Ant**arctic Environmental Research **I**nformation **S**ystem), an information system that enables the acquiring, storing, and querying of research data generated by the Brazilian National Institute for Science and Technology on Antarctic Environmental Research (INCT-APA; Valentin, Dalto, & Lavrado, 2012). INCT-APA is a collaborative research network consisting of 21 universities and

research institutes, and about 70 researchers, from Brazil, and research focuses on four thematic areas: atmosphere, terrestrial environment, marine environment, and environmental management.

This article is organized as follows. In Section 2, we describe the requirements analysis we performed, the resulting scope definition of the system, and the current implementation of BrAntIS, which consists of a web application for uploading and querying field observation data, along with a relational database for storing those data. In Section 3, we describe additional components planned for the system. Finally, in Section 4, we make some concluding remarks.

2 BRANTIS: SCOPE, CONCEPTUAL VIEW, AND IMPLEMENTATION

To define the scope of BrAntIS, we determined the demanded requirements by surveying research routines of scientists affiliated with INCT-APA, from data gathering to analysis. Scientific data in INCT-APA are generated by automated sensors or are the result of both biotic and abiotic analysis of material samples gathered during field expeditions. Such field expeditions are organized and grouped into an Antarctic Operations (or OPERANTARs).

INCT-APA scientists wish to trace the publications resulting from biotic and abiotic analyses. Therefore, one of the primary requirements of BrAntIS was to provide a data model that adequately captures (1) the gathering and generation workflow of data, (2) any publications that might be associated with these data, and (3) the tools that facilitate their uploading and querying. These data are subsequently analyzed using, for instance, statistical tools or species distribution models, and BrAntIS should supply web-accessible tools for supporting these activities, such as scientific workflow management systems and statistical libraries.

We also considered several other functionalities commonly recommended for information systems that support biodiversity and ecosystem research (Hobern, Apostolico, Arnaud, Bello, Canhos, Dubois, et al., 2013). To ensure data quality, for example, species identifications should be validated against various existing accurate taxonomic databases, such as the Integrated Taxonomic Information System (ITIS, 2013) and the World Register of Marine Species (WoRMS, 2013). Furthermore, the vast body of knowledge spread across the network of experts in those domains forming INCT-APA's research activities should be leveraged. Specifically, it should be utilized to annotate data with identified errors, validations, or details. A history of annotations to each data record should also be kept, along with proper attribution.

Figure 1 presents a layered overview of the BrAntIS architecture. The *Application* layer contains the logic for rendering the *User Interface*, in this case using HyperText Markup Language and JavaServer Pages. This layer consists of five interface modules. The *Login* interface is responsible for main access to the system. The *Administration* interface is used for user management. The *Data Sample* and *Analysis* interfaces generate data input formats corresponding to those data collected during the sampling stage of each OPERANTAR. The *Publication* interface lists the scientific publications associated to the analysis results.

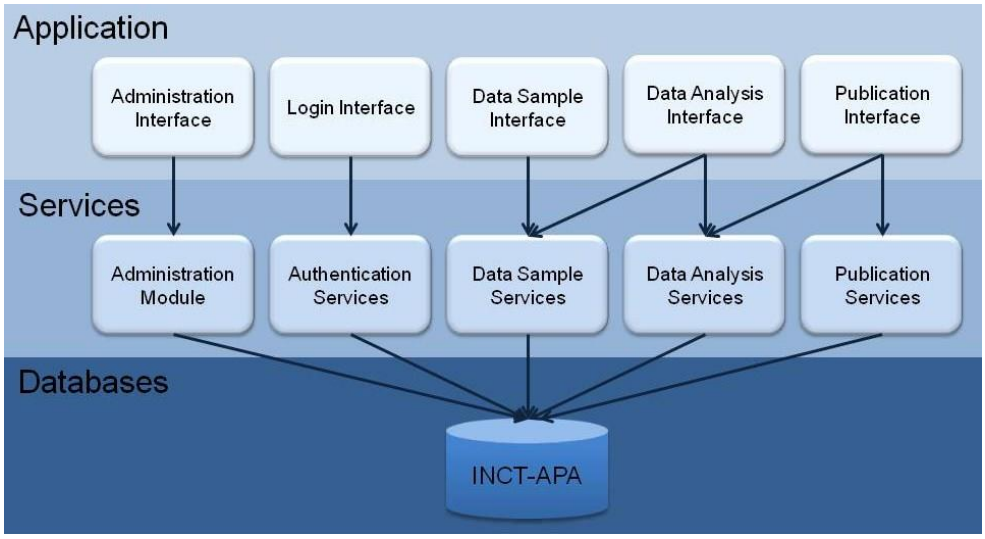

Figure 1. Layered view of BrAntIS architecture

The *Services* layer is responsible for production and submission of transactions related to the application domain and is also composed of five modules. The *Administration* module handles administrative tasks, such as user creation and role assignment. The *Authentication* module verifies whether the user is registered on the system. The *Data Sample*, *Data Analysis,* and *Publication* modules perform three common tasks, described as follows. For each request, these modules first verify if the user is authorized to make that request. The modules then validate the data received from the respective interfaces. Finally, to store the data, each module is responsible for the *create*, *read*, *update,* and *delete* operations necessary to make them persistent. A relational database is then used in the *Databases* layer to ensure this persistence.

Figure 2 shows a simplified view of the proposed data model for the application. An *OPERANTAR* represents the beginning of an annual expedition consisting of several collections in the Antarctic region. Each collection takes place along several stations in a geographical region with fixed sites, from which sampling for every thematic area is carried out. Various analyses are performed on the collected samples using a determinate method of analysis, classified according to the thematic area. The results of these analyses are then recorded and are classified into two types: biotic or abiotic. Biotic results are stored following the structure of a known taxonomic database whereas abiotic results are stored as a set of descriptors and values. When results produced by an analysis lead to a scientific publication, information about the publication, such as the author(s), type of publication, title, and so on, should be registered in the system. In addition, the data model includes constraints on certain data values that require validation: (a) the geographic coordinates are formatted in grades, minutes, and seconds; (b) sites must be contained in a determined region; (c) date intervals related to a task must be contained within the date interval of the activity that includes the task; and (d) the analysis timestamp must be later than the timestamp related to when the sample was collected.

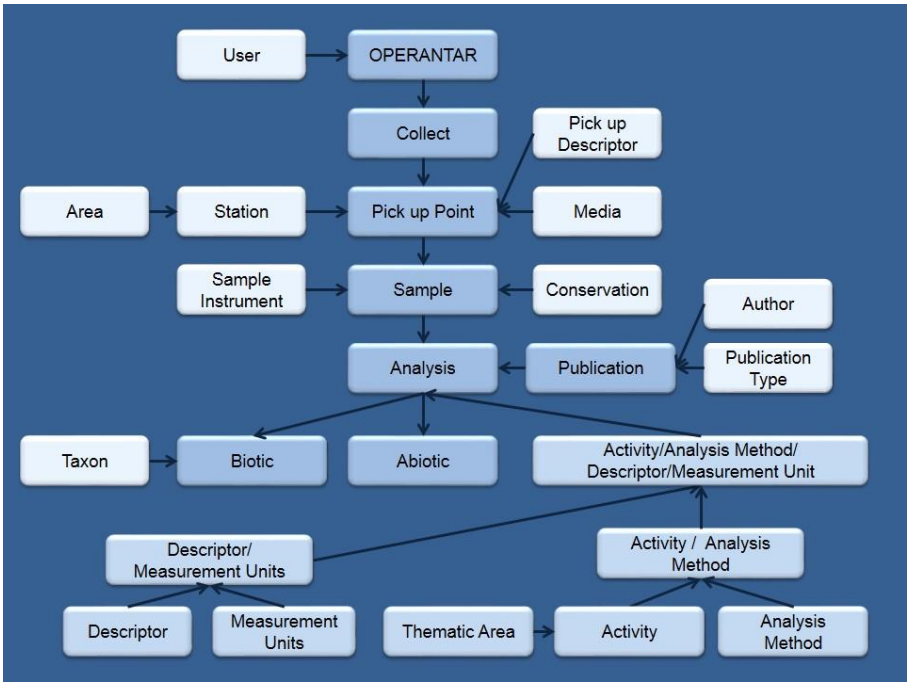

Figure 2. Simplified view of BrAntIS database model

Data integration techniques are essential tools for discovering, querying, and retrieving biodiversity and ecological data. These tasks are currently achieved mainly through employment of metadata standards and data publishing tools, where standard sets of terms are defined to describe datasets and are used during their packaging, formatting, and dissemination. Darwin Core (DwC; Wieczorek, Bloom, Guralnick, Blum, Döring, Giovanni, et al., 2012) is a data management standard that facilitates the sharing of biodiversity data, its core schema describing the occurrence of a species both geographically and temporally. It was produced within the Biodiversity Information Standards and contains a set of well-defined expressions that enable data published using DwC to be automatically extracted. The standard does not enforce a particular physical format for representing data, and adopters use various formats, such as comma-separated value files and Extensible Markup Language. Ecological Metadata Language (EML; Fegraus, Andelman, Jones, & Schildhauer, 2005) is used for describing ecological and environmental data, which are more complex and heterogeneous than data

typically described by DwC because they may include, for instance, environmental observations, used techniques, and measurement units.

Global data infrastructures have also been implemented to collect and disseminate biodiversity and ecological data. The Global Biodiversity Information Facility (GBIF, 2013; Yesson, Brewer, Sutton, Caithness, Pahwa, Burgess, et al., 2007) consists of a worldwide group of biodiversity information nodes, usually representing countries, serving data using the DwC standard. The Integrated Publishing Toolkit (IPT) it has developed translates biodiversity information from a data publisher, which may be in various formats, such as relational databases or spreadsheet files, into DwC. IPT installations are remotely accessible and are catalogued by the GBIF-hosted Global Biodiversity Resource Discovery System (GBRDS); they are constructed of biodiversity information provider catalogues as well as resources endorsed by specific node managers. The datasets available in the resources catalogued by GBRDS are harvested by the central GBIF data portal, where they can be queried and downloaded by users.

Other, specialized data portals might harvest datasets related to a specific theme or geographic region. For instance, the Antarctic Biodiversity Information Facility (AntaBIF, 2013) harvests datasets about the Antarctic region and makes them available on the Marine Biodiversity Information Network portal of the Scientific Committee on Arctic Research (Griffiths, Danis, & Clarke, 2011). Moreover, the Data Observation Network for Earth (DataONE, 2013) performs a service for ecological and environmental data that is analogous to that of GBIF, forming a federation of nodes that publish data about long-term ecological research initiatives by using the EML standard.

In contrast, BrAntIS stores data about both biodiversity and ecological and environmental observations within the context of INCT-APA. By extracting subsets of data from its database regarding each of these areas and by formatting them according to their respective data standards, it has been relatively straightforward for BrAntIS to contribute toward the aforementioned global data infrastructures. Similarly, BrAntIS can publish its data in the Brazilian Biodiversity Information System (SiBBr, 2013).

With INCT-APA divided into four thematic areas, the system must manage users according this division. Data belonging to a thematic area should only be manipulated (undergo create, update, and delete operations) by users who have permission to do so. Such restrictions are achieved in the system by using role-based access control (RBAC; Ferraiolo & Kuhn, 1992) to limit certain services to authorized users only. RBAC is founded on three concepts: users, roles, and permissions. A user can log into the system and perform a set of operations consistent with the role assigned to them. This role defines the user's permissions, namely, authorizations that approve or deny the performing of a specific operation. Figure 3 shows how the data model is extended to support the RBAC model.

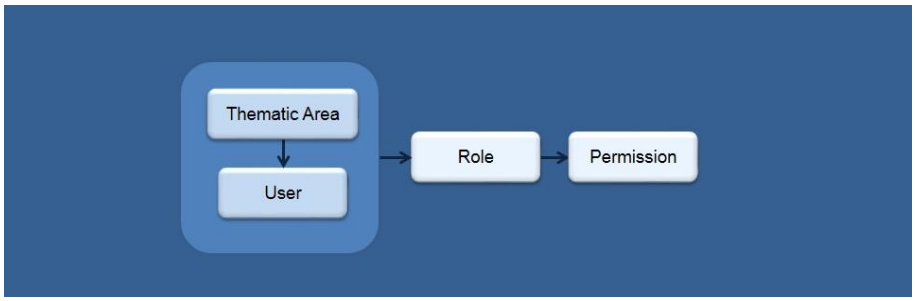

Figure 3. RBAC model

3 FUTURE WORK

Collaborative, large-scale synthesis studies in ecology require integration of data from many disparate studies and disciplines, for example, population studies, hydrology, and meteorology (Michener & Jones, 2012). A new technological architecture derived from the World Wide Web, known as Linked Data, has been proposed to realize data sharing and reuse on a massive scale (Heath & Bizer, 2011). The uptake of this technology in the Life Sciences has been considerable (Heath & Bizer, 2011), enabling the connection of a large number of datasets from highly diverse scientific domains (Linked Data, 2013). In previous data integration scenarios, each data source depended on a particular code or on a data integration workflow definition in an Extract-Transform-Load environment. Conversely, in the Web-of-data scenario, data publishers may contribute

toward simplifying integration for consumers by: reusing terms from widely used vocabularies and publishing mappings between terms from different vocabularies as well as setting Resource Description Format (RDF, 2013) links pointing at related resources and at identifiers used by other data sources to refer to the same real world. It is worth observing that when data publishers describe their data well, it becomes much easier to integrate them (Heath & Bizer, 2011).

The data stored in BrAntIS goes through a series of statistical analyses and can be consumed by, for instance, biological system models. These analyses can be assembled as a *scientific workflow* (Deelman, et al., 2009), in which a large number of analytical activities are efficiently performed by means of data exchanges (i.e., data produced by one activity can be consumed by other activities). Scientific workflow management systems provide features, such as fault tolerance, scalable execution, scalable data management, data dependency tracking, and provenance recording, that greatly reduce the complexity of managing the lifecycle of these analytical activities. Provenance information, in particular, can document the parameters used, and the data derivations that took place, during the execution of a scientific workflow (Freire, Koop, Santos, & Silva, 2008; Gadelha, Wilde, Mattoso, & Foster, 2012). As a future development, we plan to incorporate provenance-enabled scientific workflow management tools into BrAntIS to support analytical activities. Because many of these activities are computationally demanding, the computational resources of the Brazilian National System for High Performance (SINAPAD, 2013) will also be used in their execution. We also plan to include a visualization module in BrAntIS for displaying georeferenced data in maps and for generating charts from tabulated data to identify trends and make predictions.

Finally, an annotation system will be developed to enable comments and corrections created by users to be given as feedback on a per-record basis. A log record of these annotations and their authors will be kept to document the derivation history of a dataset such that users can better assess its data quality. BrAntIS will thus leverage existing knowledge available through the network of domain experts spread across the research activities of INCT-APA.

4 CONCLUSION

In its current version, BrAntIS facilitates data acquisition, storage, and querying, providing a valuable tool to the Brazilian scientific community focused on Antarctic environmental research. Its data model was created to reflect the research routines of the scientists affiliated with INCT-APA. Data are thus easier to explore because they are organized around the same conceptual framework that scientists use during sample collection and analysis. BrAntIS also simplifies tracking of analyses used in articles published by members of INCT-APA. Furthermore, it employs data quality techniques to improve data accuracy and consistency, both geospatially and taxonomically.

BrAntIS' data model is straightforward to map to the Darwin Core and EML data standards, which enables integration between those data available in BrAntIS and those available in regional, national, and global biodiversity and ecosystem data infrastructures, such as SiBBr, GBIF, AntaBIF, and DataONE. BrAntIS also uses RBAC to ensure that each data record can be: (1) manipulated only by users with appropriate credentials and authorization (2) kept track of to ensure correct authorship attribution.

Additional functionalities presently under development include: data integration applying Linked Data techniques; a data visualization and analysis module, where data can be visualized in maps or through charts; and a scientific workflow module such that scientists can automate their analysis routines. These planned features are, in part, inspired by research documenting challenges and best practices for biodiversity and ecosystem informatics (Hobern, et al., 2013).

5 ACKNOWLEDGEMENTS

This work is funded by the National Institute of Science and Technology Antarctic Environmental Research (INCT-APA), which receives scientific and financial support from the National Council for Research and Development (CNPq process: n° 574018/2008-5) and the Carlos Chagas Research Support Foundation of the State of Rio de Janeiro (FAPERJ n° E-16/170.023/2008). The authors also acknowledge the support of the Brazilian Ministry of Science, Technology, and Innovation (MCTI), the Ministry of Environment (MMA), and the Interministry Commission for Sea Resources (CIRM).

6 REFERENCES

Ailamaki, A., Verena, K., & Debabrata, D. (2010) Managing Scientific Data. *Communications of the ACM 53*(6), pp 68–78.

Bell, G., Hey, T., & Szalay, A. (2009) Beyond the Data Deluge. *Science 323*(5919), pp 297–1298.

ANTABIF (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.biodiversity.aq

Chapman, A. (2005) *Principles of Data Quality*, Copenhagen: GBIF Secretariat.

Cook, A.J., Fox, A.J., Vaughan, D.G., & Ferrigno, J.G. (2005) Retreating Glacier Fronts on the Antarctic Peninsula over the Past Half-Century. *Science 308*(5721), pp 541–544.

DataONE (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.dataone.org

Deelman, E., Gannon, D., Shields, M., & Taylor, I. (2009) Workflows and e-Science: An Overview of Workflow System Features and Capabilities. *Future Generation Computer Systems 25*(5), pp 528–540.

Fegraus, E.H., Andelman, S., Jones, M.B., & Schildhauer, M. (2005) Maximizing the value of ecological data with structured metadata: An introduction to ecological metadata language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America 86*(3), pp 158–168.

Ferraiolo, D., & Kuhn, R. (1992) Role-Based Access Controls. *Conference Proceedings 15th National Computer Security Conference.* Retrieved September 10, 2014 from the World Wide Web: http://csrc.nist.gov/groups/SNS/rbac/

Freire, J., Koop, D., Santos, E., & Silva, C.T. (2008) Provenance for Computational Tasks: A Survey. *Computing in Science & Engineering 10*(3), pp 11–21.

Gadelha, L., Wilde, M., Mattoso, M., & Foster, I. (2012) MTCProv: A Practical Provenance Query Framework for Many-Task Scientific Computing. *Distributed and Parallel Databases 30*(5–6), pp 1–370.

GBIF (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.gbif.org

Griffiths, H.J., Danis, B., & Clarke, A. (2011) Quantifying Antarctic Marine Biodiversity: The SCAR-MarBIN Data Portal. *Deep Sea Research Part II: Topical Studies in Oceanography 58*(1–2), pp 18–29.

Hardisty, A., & Roberts, D. (2013) A Decadal View of Biodiversity Informatics: Challenges and Priorities. *BMC Ecology 13*:16.

Heath, T., & Bizer, C. (2011) *Linked Data: evolving the Web into a global data space (1st edition)*, Bonita Springs, FL: Morgan & Claypool.

Hobern, D., Apostolico, A., Arnaud, E., Bello, J.C., Canhos, D., Dubois, G., et al. (2013) Global Biodiversity Information Outlook—Delivering Biodiversity Knowledge in the Information Age, Copenhagen: GBIF Secretariat.

Ingels, J., Vanreusel, A., Brandt, A., Catarino, A.I., David, B., De Ridder, C., et al. (2012) Possible Effects of Global Environmental Changes on Antarctic Benthos: A Synthesis across Five Major Taxa. *Ecology and Evolution* 2(2), pp 453–485.

ITIS (2013) Retrieved in August 15, 2013 from the World Wide Web: http://www.itis.gov

Linked Data (2012) Retrieved June 21, 2012 from the World Wide Web: http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets

Michener, W.K., Porter, J., Servilla, M., & Vanderbilt, K. (2011) Long Term Ecological Research and Information Management. *Ecological Informatics 6*(1), pp 13–24.

Michener, W.K., & Jones, M.B. (2012) Ecoinformatics: Supporting Ecology as a Data-intensive Science. *Trends in Ecology & Evolution 27*(2), pp 85–93.

RDF (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.w3.org/TR/rdf-primer

SiBBr (2013) Retrieved November 6, 2013 from the World Wide Web: http://www.sibbr.gov.br

SINAPAD (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.sinapad.lncc.br

Valentin, Y.Y., Dalto, A.G., & Lavrado, H. P. (Eds.) (2011) *INCT-APA Annual Activity Report 2011*, São Carlos: Editora Cubo 2012.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE 7*(1), p e29715.

WoRMS (2013) Retrieved August 15, 2013 from the World Wide Web: http://www.marinespecies.org

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., et al. (2007) How Global Is the Global Biodiversity Information Facility? *PLoS ONE 2*(11), p e1124.

(Article history:Available online 23 September 2014)