CRIS 2006

Current Research Information Systems

Anne Gams Steine Asserson /
Eduard J. Simons (eds.)

Enabling Interaction and Quality: Beyond the Hanseatic League

8th International Conference on Current Research Information Systems

promoted by euroCRIS
Current Research Information Systems

Leuven University Press

Toward CERIF-ScienTI cooperation and interoperability

Roberto C.S. Pacheco[1,2,3], Vinícius Medina Kern[1,3],
José Francisco Salm Jr. [1,3],
Abel Laerte Packer[4], Renato Murasaki[4], Luís Amaral[5],
Leonel Duarte dos Santos [5], Alberto R. Cabezas Bullemore[6]
[1] Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (EGC),
Universidade Federal de Santa Catarina, Brasil
[2] Departamento de Informática e Estatística (INE),
Universidade Federal de Santa Catarina, Brasil
[3] Instituto Stela, Brasil
[4] Latin American and Caribbean Center on
Health Sciences Information (BIREME), Brasil
[5] Departamento de Sistemas de Informação, Universidade do Minho, Portugal
[6] Departamento de Información, Comisión Nacional
de Investigación Científica y Tecnológica (CONICYT), Chile

Summary

The more acknowledged the role of knowledge as wealth creation factor, the more urgent becomes making S&T information compatible, useful, integrated and accessible to all innovation players. In this paper we discuss two projects that bring some answers to this issue: the Iberian-Latin-American ScienTI and the European CERIF models. ScienTI was established during the last five years as a network of S&T information, currently involving eleven countries from Latin America and Portugal. CERIF has been developed and proposed to European research information systems since its first version CERIF2000. In this study we conclude that there is a favorable scenario to promote interoperability between existent CERIF and ScienTI CRIS.
Index words: ScienTI Network; CERIF; research information systems; interoperability; online information services; information systems modeling

1 Introduction

The standardization and interoperability of science and technology (S&T) information created and used by players of national innovation systems (NIS) have improved in the last few years. Nonetheless, the systematic management of NIS S&T information remains a challenge.

In a NIS, information results from highly decentralized processes, with players working with different timetables, requirements, and, most importantly, distinct world views. As a result, there is a proliferation of data models, information projects and data sources in all levels of the innovation chain (government, S&T institutions, S&T community, and firms). Such variety makes the information sharing and combination expensive or even impossible. On the other hand, the lack

of effective access to information is an obstacle to wealth creation and a source of overhead costs.

In this scenario, there has been an increasing call for workflow management, standards, interoperability, and information sharing. Such demand has been pushing IT personnel and public decision makers to find ways of fulfilling the several NIS workflows. At the same time IT people and public authorities are demanded to make the resultant S&T information available and useful in all levels.

This challenge brings the following research issues to the knowledge and information system community:

- How can the combination between workflow management and information modeling approaches foster S&T information sharing?
- How to make such combination a resource for cooperation among the several innovation players?

We address these issues considering two continental projects: the ScienTI Network and the CERIF model. In Europe, CERIF is a reference model for research information applied in several projects (e.g., IWETO, NARCIS, METIS, FRIDA). In Latin-America, ScienTI Network established an e-gov architecture for managing information on a NIS. It has been applied by national research councils from Latin-American and Caribbean countries (Brazil/Lattes, Chile/SICTI, Colombia/ScienTI Colombia, Peru/ScienTI Perú, Venezuela/ScienTI Venezuela, Ecuador/CvLAC, Argentina/SiCyTAR, and Mexico/SIICYT) and by Portugal (DeGóis).

In this paper, we examine CERIF (Jeffery et al. 1989, Asserson, Jeffery, and Lopatenko 2002) and ScienTI (ScienTI 2002, Mugnaini 2003) as complementary approaches to the issue of how to accomplish S&T information sharing on an international scale. We introduce CERIF and ScienTI models and discuss their complementariness, respectively, as references to information modeling and information system platform. We also discuss the opportunities for cooperation between ScienTI and CERIF for S&T information sharing on a cross-continental way.

2 The ScienTI model

ScienTI Network[1] genesis goes back to June 2000, when the Brazilian Research National Council (CNPq) and the Pan-American Health Organization (PAHO) agreed to collaborate on combining two complementary research information projects: the Virtual Health Library (PAHO) and the Lattes Platform (CNPq 2002, Pacheco 2005). The combination of these two research information views (i.e., the bibliometrics and informetrics) led to a model based on a four-layer e-gov architecture: the ScienTI Information Architecture (Pacheco and Kern, 2003).

[1] http://www.scienti.net – ScienTI Portal, with access to documentation, search for competences and other ScienTI services.

This architecture is the technological basis for the collaboration between National Research Councils from Portugal and eleven Latin-American and Caribbean countries, seven international organizations related to ST&I (PAHO, BIREME, OEA, UNESCO, RICYT, I-Research and SHARED), and four research and development groups on information and knowledge systems (from Brazil, Portugal, Peru and Colombia).

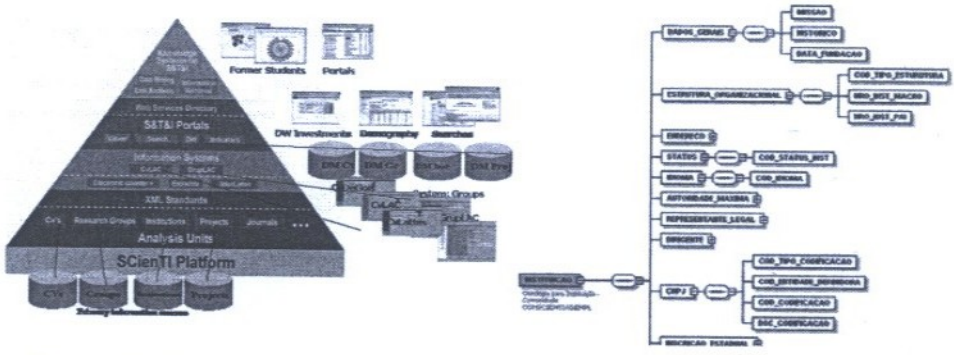

Figure 12: ScienTI technological architecture (left) and section of the ScienTI information model (right)

In the ScienTI Architecture (Figure 1, left), the first layer comprises the information units of a national research information platform (i.e., curricula, research groups, institutions, projects, etc.). All units are standardized based on the agreement among information management and research personnel from national councils and universities (e.g., the XML model showed in Figure 1, right). The second layer is composed by the information systems used to update the primary sources. Once updated, the primary information is migrated to the secondary sources composed by as many data marts as necessary to support analyses. The secondary information sources feed the ScienTI portals, search, and knowledge systems. This architecture is replicated in all countries and the information search and sharing is accomplished by web services available in the ScienTI International Portal.

ScienTI Model is based on standardized information units with three data models: the relational, multidimensional, and XML formats. In the client environment (i.e., researcher's desktop), the primary sources are relational. The connections between units are treated with lookup tables (updated by downloading from the server environment – national research council's server). In the server environment, the primary sources are also relational at each information unit level as well as in the cross models (bi-relational models that connect the information units). The secondary information sources in the server environment are also in three formats:

multidimensional (data marts), index files (for searches), and XML files (for transferring). The XML formats are also used in the client environment to import and export data from and to other information systems.
With this approach, ScienTI systems can deal with decentralized and independent workflows by maintaining the unit connections as rigid as possible in the client environment (by using lookup tables) but allowing the user to include new controlled records whenever necessary (e.g., firms in which the user has worked but are still not registered in the lookup table). In some cases, the XML format implies semantic losses (specially regarding binary connections). Such losses can be managed by duplicating related information. In return, with XML, ScienTI systems and sources can be easily linked outside ScienTI mainstream (e.g., by university administrative systems, by digital libraries).
Besides the methodological and technological aspects, ScienTI has dealt with several other issues related to NIS. Some of the most controversial are: (a) independent information flows versus information redundancy prevention; and (b) user information trusting versus institutional accreditation;
The first aspect is related to the assumption that information redundancy is inevitable at the capturing level. In theory, information on team articles and projects, for instance, could not be asked directly from the people involved (project team or authors), since none is a full owner of the record. However, since the curriculum holds personal information and is, in most countries, one of the main criteria for funding analysis, the process of gathering information cannot rely exclusively on external sources. The redundancy is taken at the subunit level and can be treated at the secondary information sources (data warehouses). Doing otherwise, the process of funding and the individual information rights are not covered.
Another polemic issue is related to information validation. Some decision makers are led to believe that the only information that should be available is the information validated by external sources. Some have even suggested manual validation. At the ratio of more than 3 thousand curriculum updates a day, it is impossible to check them (and recheck at every new updating). The trusting process consists in making public the data as reliable information based on the fact that it was provided by its owner. Of course, this does not prevent other players to make their own decision for mandatory accreditation to fulfill their ends (as it happens with research groups certified by the institutional research authority, or with institutional authorities that assess institutional scientific production).

3 The CERIF model

The CERIF standard has been developed as a multi-institutional and multinational project managed and maintained by euroCRIS. It is recommended to European Union Member States. It has been applied in several projects as a referential basis not only to develop S&T information systems but also to make them interoperable.

In Figure 2 there are two CERIF components: the schematic view of how CERIF model is related with other CRIS models and the main CERIF units (business objects). The CERIF model is conceptually relational, establishes basic entities (person, organization, and project), and secondary entities (funding program, result publication, facilities and equipment). To cope with multiple languages CERIF separates the textual fields and associate them with the idiom and kind of translation (human or automatic). To deal with multiple classifications (e.g., kinds of organizations, knowledge fields) CERIF allows including different schemes.
As a result of CERIF developments and applications, the model became a reference in the European Research Information System. It is used in practical applications (by universities and federal governments in Europe) or as a best-practice reference.

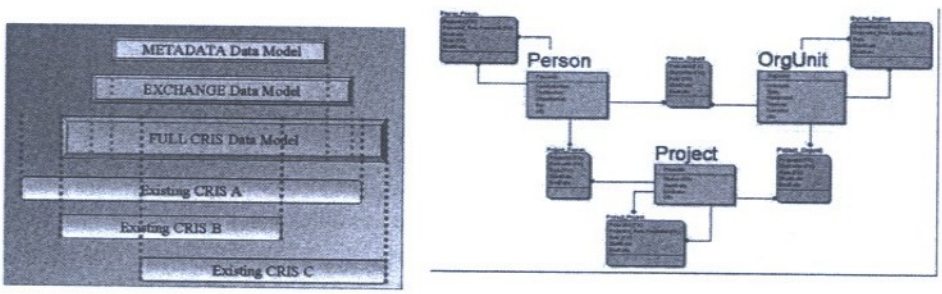

Figure 13: Schematic view (left) and main business objects of the CERIF model (right)

4 On differences, strengths, and opportunities for CERIF-ScienTI cooperation

CERIF and ScienTI have followed different paths. In Europe, CERIF is established as a mature reference model to deal with a variety of complex issues related to CRIS. Its goal is making the information generated by European CRIS available and effective to political decision makers, funding organizations, entrepreneurs, researchers, innovators, media, and general public. In Latin America and Portugal, ScienTI offers a set of CRIS to capture, manage and publish scientific and technological information at every NIS user level. With ScienTI, databases, national portals, and interoperable systems can be developed (as the Innovation Portal and Institutional CV version of Lattes Platform in Brazil, for instance).
Regarding information modeling, CERIF and ScienTI differ when it comes to dealing with record redundancy. Contrary to the single CERIF normalized approach, ScienTI uses three kinds of formats: relational and XML, in the client's environment, and relational and multidimensional, in the server's environment. For

instance, in the ScienTI Curriculum System, the individuality of each curriculum leads to redundant publication records (since each co-author has its own CV).
In order to solve such redundancy, ScienTI applies knowledge systems to identify similar records at the server environment. It also uses institutional certification processes when applicable (e.g., publication catalogs in the universities). This the ScienTI approach to cope with the heterogeneous requirements from all NIS players. The decision in favor of redundancy comes from the principle of individual ownership of the curriculum, complemented by the trust in each researcher's word – until proved otherwise, the national platform takes each curriculum as trustworthy.
Still at the information modeling level, it is important to note that CERIF recommendations to classification and idioms schema bring answers to current discussions in ScienTI regarding how to cope with the variety of particular needs in ScienTI countries (e.g., the need of Portugal of making DeGóis compatible to the European standards).
CERIF, from its strong foundation in information modeling – normalized and mature –, could serve as design pattern to be applied to open problems in ScienTI. For instance, the lack of identity of co-author and production item objects has been approached in ScienTI with semi-automatic (supervised) record matching, using similarity thresholds. A high similarity index, however, doesn't guarantee that two items are the same. CERIF establishes the non-redundant data structures to which the redundant records should be mapped to, in order to create the missing identity.
ScienTI technological and methodological approaches might bring, in return, a complementary view to some CERIF issues. An example is how to cope with the semantic loss that an XML representation implies to the relational CERIF model. ScienTI has dealt with this issue by using duplicated fields in the XML formats (to maintain references) and decoding such duplications whenever an XML file is imported to the relational model. This is currently an issue in proposal extensions and ontologies for CERIF (e.g., Lopatenko 2001).
The conception of ScienTI architecture takes into account that, in the information market, supply drives demand (contrary to most markets). The shortage of information supply doesn't raise its price – it may put an end to business, in fact. Gibbons et al. (1994) point out that the strategy of large players in the information and knowledge markets is to try to make their products the standard. The strategy to achieve that may be to impose a standard upon the citizens, or to persuade them to use it. The Lattes Platform, ScienTI's Brazilian predecessor, offered the user tools and other resources such as a personal CV-website, online submission, translation from previous file formats, automatic reports, and knowledge-based profiles (automatic résumés and profile's evaluation).
The strategy proved to be successful even when the view from decision makers changed (with the end of a political term and subsequent change in administration), ceasing to consider information as *funds*, or *supply* for research as responsibility of

the funding agency that started the platform. Even with the end of new developments in 2004, the Lattes Platform kept growing in number of curricula, showing a clear evidence that there is a strong principle behind the success. Figure 3 plots the numbers of Brazilian curricula since the launch of the Lattes Platform, in 1999, with about 35,000 curricula imported from previous formats.

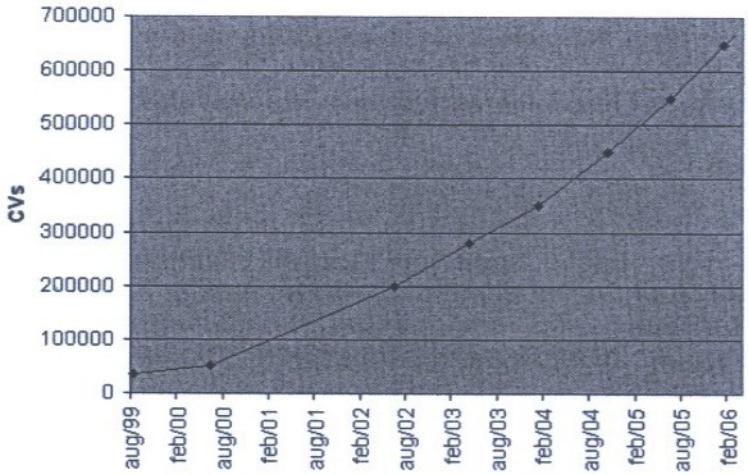

Figure 14: Evolution of number of curricula (Brazil, Plataforma Lattes)

De Los Ríos and Santana (2001) describe the evolution of contacts and meetings that led from the Brazilian platform to the continental CvLAC, precursor to ScienTI. Pacheco and Kern (2003) summarize what the assumptions are and emphasize what is not required from the adopters of the ScienTI principles and technological infrastructure. The adoption requires:

- considering the information needs of all stakeholders of the NIS,
- working collaboratively and to conduct network projects, and
- adopting international standards established collaboratively.

Adherence to ScienTI principles and technology doesn't imply, however:

- adopting the same hardware or software platforms for all participants, or
- changing fundamental workflows of any actor in the system.

Elements of that strategy might very well be used for CERIF's purposes.

5 Conclusion

In terms of interoperability, both CERIF and ScienTI have been proved as open models that can foster connectivity, information sharing, and compatibility in CRIS developments. This is actually the most important goal for both models.

Both initiatives have strengths that may serve the purposes of one another. CERIF has a mature and normalized model that eliminates redundancy and may play the role of design pattern for future ScienTI developments. ScienTI has reached a wide public and uses strategies that proved to be successful to make the database grow and remain constantly updated.

In conclusion, it is relevant to emphasize that a collaboration between CERIF and ScienTI could bring significant contributions to both projects and to all players in an innovation chain. Portugal is a participant of both initiatives and may contribute to accelerate that collaboration. Most importantly, European and Latin American NISs would benefit from information sharing and cooperation.

References

Asserson, A.; Jeffery, K. G.; Lopatenko, A. (2002): CERIF: Past, Present and Future: an Overview. Gaining Insight from Research Information. *6th International Conference on Current Reseach Information Systems*, August 29-31, 2002, Kassel, Germany.

CNPq – Brazilian National Research Council (2002): *The Lattes Platform*. Brasília: CNPq. Available at http://www.cnpq.br/english/aboutcnpq/pub_material/beyond_lattes.htm.

De Los Ríos, R.; Santana, P. (2001): El Espacio Virtual de Intercambio de Información sobre Recursos Humanos en Ciencia y Tecnología de América Latina y el Caribe: Del CV Lattes al CvLAC. *Ciência da Informação*, Vol. 30, No. 3, p. 42-47.

Gibbons, M.; Nowotny, H.; Limoges, C.; Trow, M.; Schwartzman, S.; Scott, P. (1994): *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. London: Sage.

Jeffery, KG.; Lay, J.O.; Miquel, J.; Zardan, S.; Naldi, F.; Vannini-Parenti, I. (1989): IDEAS: A System for International Data Exchange and Access for Science. *Information Processing and Managemen*, Vol. 25, No. 6, pp. 703-711.

Lopatenko, A. (2001): Information Retrieval in Current Research Information Systems. In: *Workshop on Knowledge Markup and Semantic Annotation at K-CAP'2001*.

Mugnaini, R. (2003): *La Platteforme Lattes*. Dossier CenDoTeC, octobre 2003. São Paulo: Centro Franco-Brasileiro de Documentação Técnica e Científica. Available at http://www.cendotec.org.br/dossier/cendotec/lattesfr.pdf.

Pacheco, R.C.S. (2005): *Lattes Platform: the methodological steps*. Presentation at the euroCRIS members meeting. Lisbon, November 10th, 2005. Available at http://www.eurocris.org/en/meetings/lisbon_november_2005_portugal/presentations/lattes_platform_the_brazilian_system_of_researchers_cvs/M%3A%5CHARRIE%5CDOCUMENT%5CLattesPlatformInBrazil.ppt.

Pacheco, R.C.S.; Kern, V.M. (2003): Arquitetura conceitual e resultados da

integração de sistemas de informação e gestão da ciência e tecnologia. *DataGramaZero*, Vol. 4, No. 2 (online). Available at http://www.dgz.org.br/abr03/Art_03.htm.

ScienTI – Red Internacional de Fuentes de Información y Conocimiento para la Gestión de la Ciencia, Tecnología e Innovación (2002): Red ScienTI: Metodología de Gestión de Fuentes de Información en Red. In: *Propuesta de Metodología de Gestión de Fuentes de Información en Red.* Available at http://scienti.bvsalud.org/doc/DocReferenciaScienTI2004.doc.

Contact information

Roberto Pacheco, Vinícius Kern, and José Salm Jr.
Instituto Stela
Rua Prof. Ayrton Roberto de Oliveira, 32, 7° andar
88034-050 Florianópolis-SC, Brasil

e-mail: {pacheco, kern, salm}@stela.org.br
Homepage: http://www.stela.org.br/

Abel Packer and Renato Murasaki
Bireme
Rua Botucatu, 862
04023-901 São Paulo-SP, Brasil

e-mail: abel@brm.bireme.br, murasaki@bireme.ops-oms.org
Homepage: http://www.bireme.br/

Luís Amaral and Leonel Duarte dos Santos
Universidade do Minho, Escola de Engenharia
Departamento de Sistemas de Informação
Campus de Azurém
4800-058 Guimarães, Portugal

e-mail: {amaral, leonel}@dsi.uminho.pt
Homepage: http://www.dsi.uminho.pt/

Alberto Cabezas
CONICYT
Departamento de Información, Comisión Nacional de Investigación Científica y Tecnológica
Canada 308, 2° Piso
Providencia
Santiago, Chile

e-mail: acabezas@conicyt.cl
Homepage: http://www.conicyt.cl