CRIS2002
Current Research Information Systems

Wolfgang Adamczak / Annemarie Nase (eds.)

Gaining Insight from Research Information

6th International Conference on Current Research Information Systems

promoted by euroCRIS
Current Research Information Systems

kassel university press

Gaining Insight from Research Information

Programme Committee

Wolfgang Adamczak (University of Kassel, Germany)

Leonel Duarte dos Santos (University of Minho, Portugal)

Jürgen Güdler (Deutsche Forschungsgemeinschaft, Germany)

Jostein Hauge (Bergen University Library, Norway)

Keith G. Jeffery (Rutherford Appleton Laboratory, United Kingdom)

Sinikka Koskiala (Helsinki University of Technology, Finland)

Irmgard Lankenau (University of Koblenz-Landau, Germany)

Dietmar Pfähler (SAP, Germany)

Marga van Meel (Netherlands Institute for Scientific Information Services (NIWI), The Netherlands)

Organising Committee

Wolfgang Adamczak (University of Kassel, Germany)

Klaus Horn (University of Kassel, Germany)

Annemarie Nase (Social Science Information Centre, Bonn, Germany)

Christian Spath (University of Mainz, Germany)

Organiser

UNIKASSEL VERSITÄT

IZ Social Science Information Centre

Sponsors

BMBF, Bundesministerium für Bildung und Forschung, Germany

Stifterverband für die Deutsche Wissenschaft, Germany

SAP Germany

Technologie Transfer Netzwerk Hessen, Germany

SIKA, Systemtechnik GmbH, Germany

Wolfgang Adamczak / Annemarie Nase (eds.)

Gaining Insight from Research Information

Proceedings of the 6th International Conference on Current Research Information Systems, University of Kassel, August 29 - 31, 2002

promoted by euroCRIS

kassel university press

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Gaining Insight from Research Information / Ed. by Wolfgang Adamczak ...-
Kassel : kassel university press, 2002
ISBN 3-933146-844

Edited by: Wolfgang Adamczak, Annemarie Nase

Printed by: kassel university press

Contents

Workshops

Poster

Preface

Current Research Information Systems is the focus of series of international conferences, initiated since 1991 by euroCRIS, a non-profit association. Screening the topics of the former conferences you can see that CRISs always picked up both the social challenges and the technological developments. Topics of former conferences were

- establishment of research databases in single countries
- standardisation and harmonisation
- data exchange formats
- marketing, promotion and dissemination of research information

CRIS 2002 is reflecting the real situation, influenced by social pressure on scientists to provide answers to problems of society and on transparency of research and research results in a world of unbounded (electronic) information networks. In the times of electronic networks it is not the problem to spread information, but the question is how to find the needed information at the right moment within appropriate time. Especially looking to WWW we not only have an enormous amount of information but different kinds of information systems (single systems and those embedded in a university research information management system), structured and non-structured information offers, data of different quality. The most exciting questions that will be discussed at CRIS 2002 are:

- How can we solve the heterogeneity problem searching in distributed CRISs?
- How efficient are search engines? Do we need CRISs in times of search engines like ”google”?
- What achieve new flexible and extensible data- and information-exchange models in addition?
- And last but not least: Is their any user for CRIS? Who are the users? Do they only want to find interesting information about research or do they want to use this information for analysis and evaluation of (local, regional, national, transnational) research systems and appropriate funding of research?

We set our hopes on fruitful lectures and discussions and long lasting effects of CRIS 2002.

This conference was organised by the University of Kassel and the Social Science Information Centre, Bonn, which give all the support needed to organise an international conference. But without funding and sponsoring of the conference by research funding institutions and by private firms of Germany the conference could not have taken place. We want to express our gratitude. We want to thank members of programme and organisation committees working for a good conference too.

Kassel/Bonn June 2002

Wolfgang Adamczak Annemarie Nase

CRIS-Cross: Current Research Information Systems at a Crossroads

Eric H. Zimmerman (Keynote Speaker)
The Research Authority of Bar-Ilan University, Ramat Gan, Israel

Abstract

CRIS: Current Research Information Systems must take center stage as the mechanism of scientific information provision. In the ever-growing complex scientific world, we are faced with mounting challenges. Increasingly we are witness to pressures on universities and scientists to provide answers to the problems of society, while working in multi-disciplinary and cross-border teams, with increased funding competition. The need for quality scientific and technological information is apparent. Content management (with taxonomy) and sound information architecture is key. My thesis is that there is a strong need for intelligent CRISs, and that the Web search engines, powerful as they are, are not a replacement for a good CRIS. CRISs should be used for decision-making at all levels, for the management of research activities, and for the dissemination of results. CRIS, in this respect, is key for facilitating the processes of knowledge creation and management, and hence economic growth.

1 Introduction

With the ERA, the European Research Area, taking shape, and the FP6, the Sixth Framework Programme of the European Commission (significantly different from previous programmes), months away from its inauguration, the European – indeed, global - research community is at a crossroads. In recent years, science systems worldwide have been undergoing radical change. This change includes increased fiscal responsibility, limited research funds, greater numbers of students, increased competition, the changing nature of science in its relationship to society and global economies, and the growing internationalization of science. These changes have occurred just as new and emerging information and communication technologies have been impacting on organizations.

CRIS must take center stage as the mechanism of scientific information provision. In the ever-growing complex scientific world, we are faced with mounting challenges. Increasingly we are witness to pressures on universities and scientists to provide answers to the problems of society, while working in multi-disciplinary and cross-border teams, with increased funding competition. The need for quality scientific and technological information is apparent. Content management (with taxonomy) and sound information architecture is key.

This paper addresses the following points:

1. The nature of CRIS.
2. Why do we need CRIS systems in an age of super-powerful Web search engines?
3. What can CRISs do for us?

My contention is that there is a strong need for intelligent CRISs, and that the Web search engines, powerful as they are, are not a replacement for a good CRIS. CRISs should be used for decision-making at all levels, for the management of research activities, and for the dissemination of results. CRIS, in this respect, is key for facilitating the processes of knowledge creation and management, and hence economic growth.

The world of higher education, science and technology, research and innovation is facing increased complexity. The universities of late, as an example of the changes taking place, have become tremendous "knowledge factories" due to their overwhelming contribution to the production of new knowledge through research endeavors. This has not always been the case. As we increasingly recognize the importance of knowledge, it becomes apparent that the average knowledge worker will be in need of perpetual life-long education and re-training.

We are witnessing a paradigm shift on the global science scene. Science has been shifting from discipline-oriented to cross-disciplinary research. Gibbons (1994) called this Mode Two. This shift in turn is leading us from data through information to knowledge, and from knowledge to wisdom, and from wisdom to insight. This process is spurring increased quality of life and wealth creation, as industrial countries, worldwide, attribute ever-greater importance to R&D because of their acknowledged role as a stimulus for economic progress.

Until recently, much of research has been discipline-oriented, curiosity-led, and motivated and executed by an individual following observation, hypothesis, experiment, or proven method. However, the complex problems of today are such that they require teams with each member having a specialization contributing to the whole. These collaborative teams are often geographically dispersed and of differing disciplines. Researchers today must be proficient in their chosen discipline but also multidisciplinary in their approach to science. This is to enable them to see the „breadth" of problems, but have the requisite „depth" in order to solve the problems.

Increased knowledge, the paradigm shift, recognition of economic stimulus and collaborative interdisciplinary science lead inexorably to the need for systems to assist researchers, administrators, strategists, opinion-formers, entrepreneurs and innovators and also the general public. Systems are needed to provide both information for decision-making and support to the process of knowledge-creation. As Keith Jeffery writes, "The end-user requirement is for the relevant information (relevance, recall), at the right place (wherever worldwide), at the right time (when required), in the appropriate form (optimal presentation, integrated for further use in electronic information / office environments) (Jeffery, 1999, p.5)."

CRIS provide access to and dissemination of research information. This includes People, Projects, Organizations, Results (publications, patents and products), Facilities, and Equipment. The Common European Research Information Format (CERIF), developed by the European Commission[1], is a set of guidelines designed for everyone dealing with research information systems. It is intended to help in the development of new research information systems; to assist existing CRIS systems considering extensions; and to guide CRIS systems on how to structure and index their data.

The advantages of Using CERIF include usability; the common presentation of data providing homogenous view for people accessing data; and the defined format and database structure, simplifying CRIS development and information sharing. The *Added Value* is the guaranteed interoperability allowing for data exchange or homogenous access to heterogeneous information. Further advantages of using CERIF include: nurturing cross-border and cross-disciplinary research; a common model with defined meaning and structure of data, facilitating collaboration between researchers, industry, and policy-makers; common vocabularies making data searchable in a distributed environment; and XML and RDF solutions promoting distributed data retrieval based on leading-edge technologies.

Though the need for CRISs is apparent, the take-up has been slow, and questions have been raised as to their need, given the strength of current Internet search facilities. Commercial success stories are few. Failures are many.[2]

1 In 1991 and again in 1999.

2 INDARD, The Israel National Database of Academic Research and Development, is an example.

2 Best Practice: Developing a CRIS

(Research Information Systems, developed using a Common Information Format, according to Good Practice)

Current Research Information Systems (CRIS) and Common European Research Information Format (CERIF) efforts are not new concepts. As early as in the 1970s, serious efforts were being made among research information systems, in the field of international co-operation.

The problems of subject indexing were already tackled long before the Internet era. In the late 60s and the early 70s, both Smithsonian Science Information Exchange and UNESCO attempted to come towards an international standard taxonomy for fields of science and technology. A standard taxonomy is seen as an important tool in the management of scientific and technological affairs, especially in fields related to science policy and science statistics.

2.1 CRIS

Access to information on current research activities throughout Europe is an essential requirement for the success of EU innovation policy. The key asset in European R&D consists of ideas, technical reports, publications, patents, prototypes, products and know-how – leading to technology transfer and wealth creation, and to the generation of new R&D ideas. The key added value to be achieved is the pan-European approach to the generation of and exploitation of R&D. There is a need for information on currently relevant R&D information to be made available widely to encourage both innovation and new, improved R&D.

The innovators in industry and services, the academics pushing the frontiers of R&D, the decision makers in governments and R&D funding agencies all require easy-to-use desktop access to R&D information. The raw data sources are the R&D information held by funding agencies and other information providers in the EU. These are held for the particular purposes of the agencies and the particular clients of the information providers. They are heterogeneous and unconnected. The potential for European wide exchange is not being exploited enough.

There is a need for the consumer of research information to be able to access this data through a uniform familiar interface and to be able to integrate and compare the information between data sources. This “common interface” must not only address the content (what must be exchanged) but also the format of such information (how it should be presented). This information must be presented in a uniform way, at least at summary level. Classification should be consistent for all the research information sources. Subject indexing is required and a controlled terminology should have the same meaning in all languages. To this end, the Common European Research Information Format was proposed in 1991 and revisited in 1999.

The definition of this uniform information description platform requires:

- the definition a full CRIS data model which will cover the database structures of the majority of existing CRIS;
- the definition a set of data models which could provide examples for data exchange (since there are an infinity of possible exchange data models between CRIS;
- the definition of a metadata data model to provide a uniform summary-level view over heterogeneous information sources.

Easy access to information must address not only the availability of information with a common definition and format but also how the consumer could retrieve that information. The consumers need to be able to search, European-wide, for information on a particular research topic or theme. Subject indexing of the information is the obvious key in this respect. Classification should be consistent for all the research information sources; otherwise people will not get consistent results when they retrieve information. Since consumers also use different languages, the controlled indexing terminology proposed should have the same meaning in all languages.

2.2 CERIF

The European Union's Innovation policy aims at improving and strengthening the generation and exploitation of current and new Research and Development (R&D) projects as well as technology transfer. To this end, access to information on current research activities throughout Europe is an essential requirement. New R&D ideas can emerge thanks to a pan-European approach for information sharing and exchange. There is thus a need for a convenient tool to spread relevant R&D information widely to encourage innovation and improved R&D as well as wealth creation.CERIF: Common European Research Information Format provides a practical common standard for information contents and for subject indexing. Further, controlled value lists ease the collection and exchange of data. To provide easily searchable information implies adapting common rules. To this end, it is important to use standard controlled vocabularies. The use of standard controlled vocabularies and standardised data structures, as described further in the CERIF guidelines, should be as wide spread as possible, in order to make both data providers and end-users familiar with common CRIS characteristics.

CERIF is this common language that fosters the diffusion of information across Europe. The EU Commission's Green Paper on Public Sector Information[3] emphasizes the importance of access for European citizens to publicly funded information and equally the opportunities for economic growth and employment that it provides. All Member States are taking initiatives with regards to public R&D information – at an uneven pace. European Union policy should therefore aim to have all Member States arrive at the same point, and as quickly as possible. CERIF in addressing Public Sector R&D information is dealing with an area of economic activity with a high growth potential, and is crucial to the ongoing competitiveness of European industry.

2.3 The Code of Good Practice (CGP)[4]

There are many organizations, large and small, participating in research projects throughout both Europe and the rest of the world. The continuing evolution of the supporting technology (in particular the Internet and Computer-Mediated Communication tools) is making communication available to even the smallest organization. The Code of Good Practice was drafted to establish a framework for encouraging interoperation and harmonization between European and all research institutions worldwide. The intent was that this document be regarded as a set of good practices for existing research institutions.

The CGP was developed as a guide for both new and existing producers of CRISs. The intention was to focus clearly on the reasons for having a CRIS and on the main components of the system. The CGP is not a comprehensive guide to building and developing information systems.

As a stand-alone system each CRIS plays an important role for the host institution or organization. Together, a collection of CRISs is potentially a very powerful information tool, the true value of which can only be harnessed if interoperability can be achieved. Universal adoption of the CGP by CRIS producers for both new and existing CRISs will be a significant step towards realizing this goal and will provide CRIS users and data providers alike with a framework for knowledge transfer. Likewise, there will be greater scope for the CRIS institutions to exchange data for mutual benefit or commercial gain.

The cornerstone of the CGP is consistency. This is the single most important benefit that will result from the adoption of the CGP. Consistency will lead to:

- increased usability of data and value of CRISs to the users (often with varied demands and requirements);

3 Public Sector Information: a key resource for Europe", Green Paper on Public Sector Information in the Information Society, European Commission, COM (1998) 585 http://europa.eu.int/ISPO/docs/policy/docs/COM(98)585/ <accessed 2 May 2002>

4 Developed by euroCRIS <http://www.eurocris.org> and the European Commission (CORDIS), 1998.

- increased interoperability between CRISs;
- lower operating costs for CRISs;
- reduced effort for information exchange.

In order to realize the overall benefits offered by the CGP, it is first necessary for all relevant parties to adopt the CGP and then ensure its implementation within their working environment as a recommended standard or norm to be used. The increasing accessibility of information through developments in technology further emphasizes the need for consistency (with regard to information exchange) to ensure that the wealth (and potential diversity) of information available locally is accessible globally.

3 Why CRIS and Not a Web Search Engine?

The system should work harder so that the user need not. Rather than force the user to choose the best query term, the system should be able to perform intelligent searches in a multi-lingual environment, with the assistance of a controlled vocabulary (ontology / taxonomy). Below we discuss the problems associated with jargon and language through the collaboration of multinational research teams.

"Power" users, relying on CRIS for decision-making, must be able to find all relevant information in a homogenous fashion, with a high degree of precision, and with minimum noise. This is possible only with a dedicated CRIS. Internet search engines cannot provide users with all this – yet. CRIS, based on CERIF, offers the combination needed to aggregate information from distributed heterogeneous systems and allow the user to access the information via a homogenous interface.

As we have demonstrated above, developing research information systems according to agreed-upon conventions of good practice, and according to uniform data standards, offers a powerful tool, unmatched by the most powerful of today's Internet search facilities (engines and directories alike).

3.1 What can CRISs do for us?

The functions of CRIS are as varied as there are information systems. They can be used, among other things, to detect and map trends in science, identify discipline specialists, find specialized equipment or facilities, recognize innovations and results (to avoid duplication of effort), manage the grant process, produce statistics and reports, evaluate projects and assess science, promote science in society, and to locate funding sources. In these lines we examine but a few of the major applications of CRISs.

3.1.1 Information-Sharing (Knowledge-Sharing) and Communication Between Scientists

As some have pointed out, CRIS can be a facilitator of cross-discipline research. Jewitt and Görgens (2000, p. 410) point out that "multidisciplinary communication is one of the missing links of science." Koku et al writes (2000) that increased specialization has added urgency to the need for communication. CRIS can be used to power the communication networks of 21st century science.

An intelligently designed CRIS may provide context to science. With ever-increasing cross-disciplinary science being performed across borders, the importance of preserving context cannot be overstated. As Gibbons has stated (1994), Mode 2 knowledge is produced within a context of application. The involvement of different perspectives, languages (and discipline-oriented

jargon[5]) and cultural (national) settings, by teams of people with heterogeneous skills, mandates that a system to preserve context exists.

Gloria Mark (2000) writes that to build a culture of true information sharing on the Web, we must begin to think in terms of communal information repositories. No longer would we view the Internet as a set of individual knowledge warehouses.

CRIS can also be used as the engine of *Collaboratories* and collaborative white-boards. Access rights could be carefully monitored.

With the spread of electronic mail, the world witnessed a growth of invisible colleges. With the Web we are witness to the flourishing of "communities of practice". euroCRIS is a well-defined community of practice, and with many members in academia, we may be considered a college as well, as we promote the free sharing of ideas, information and insight.

3.1.2 Decision-Support and Statistical Analysis

CRIS can be used for Budgeting and Reporting at all levels. Also, strategic and operational decision-making at all organizations is based on good information. This information comes from multiple varied sources (i.e. experts), yet a lone individual often makes the decisions. CRIS data can fuel the support systems necessary to guide decision-makers.

R&D Unit and Researcher Assessment at the local (university) and national level will benefit from a system of comparable indicators. CRIS, taking into effect (and recording) global diversity, can develop a system of indicators for the international comparison of science.

It has been shown that science policy making worldwide has lacked the ability to compare data sets, especially in the social sciences.[6] CRIS could prove instrumental in integrating analogous data from different countries. Standardization is key to the success here, and the CERIF recommendation would play an instrumental role. This will prove to be of immense importance with respect to the ERA (see below).

3.1.3 Research Administration

CRIS can be used (and indeed is used) to support the management of institutional expertise, the administration of the proposal development process, the submission and peer-review process, the supervision of grants, and the publication of results. In this sense, when working as an integrated whole, CRIS is a full life-cycle research knowledge management system. In the United States this system is known as Electronic Research Administration (ERA – not to be confused with the European Research Area, also known by the same acronym). Hunt Williams (2000) termed this the "research process" – from the conceptualization of a research idea through the commercialization and exploitation of results. Jeffrey (1999) lists several of the life-cycle components, from a "products" angle. They are publications, patents, products, results, know-how and IPR, education and training, and publicity. Each stage or product must be managed.

Having the right information on funding opportunities is crucial for scientists, in today's increasingly competitive funding environment. Traditionally, argues Richard Tomlin (2000), funding opportunities information systems have been discipline oriented. With increasing interdisciplinary research and new research paradigms, Tomlin claims we might be doing harm by not allowing (helping) the individual scientist to see the bigger picture. By targeting information based on discipline-specific keywords, we might be reducing information overload, but we might also be inhibiting novelty, serendipity and innovation.

Barriers to grant-getting among junior faculty can be attributed to a lack of mentoring and to a need for faculty development programs (Boyer and Cockriel, 1998). CRIS can be used to fill this need by cataloguing writing guides, proven techniques, and the like.

---

5 See, for example, Jewitt and Görgens, p. 411.

6 http://www.oecd.org/dsti/sti/prod/intro-24.htm

3.1.4 Publishing and Access to Research Information

Gibbons (1990) writes that in Mode 2 science, results are disseminated among the producers in the first place. The diffusion of the results is part of the knowledge production process. Later, results are published via the normal academic channels. As stated above, it is imperative that the context – ever evolving, as new problems are tackled – not be lost.

CRIS can be used to publish research results in peer-reviewed electronic journals, making full use of multi-media and hypertext capabilities. Likewise, preprints can also be published via a CRIS-based system.

CRIS can be used as the backbone of digital repositories (digital libraries). Similarly, CRIS can be used to drive Web-based scientific sites.

3.1.5 Knowledge Management (KM)

KM has, it can be argued, two main functions: "community of practice" building and adding structure to a rich body of knowledge so that it might be used and reused. CRIS can serve both functions effectively.

The management of "communities of practice" necessitates the development of expertise databases (i.e. Community of Science) or "competency dictionaries".

Mode 2 Knowledge Production, as postulated by Gibbons (1994), is performed in transient group settings. Networks are developed on an ad-hoc basis. When a given problem is solved, these networks are oftentimes dissolved. It is vital that the experiences of these groups be captured. This knowledge management is crucial if new knowledge is to be transferred to solving a new problem.

CRIS can be the natural bridge between scientific information and innovation information; two different yet complementary types of information, necessary for the presentation of the "complete picture" of science. Whereas the general aim of science is to understand nature and society, innovation seeks to create new products, processes and services.[7] In recent years we are witness to the growing dependency of innovation on science. This rapidly expanding body of knowledge could be captured by CRIS.

With ever increasing amounts of information available it is becoming more important to be able to find the relevant information needed. Productivity cannot be jeopardized by the time needed to sift through the vast amount of data and information, of varying quality. Efficient retrieval can be accomplished by using CERIF to control the "language of science". The key here is to develop and maintain a global taxonomy of science. This will help all actors drive science forward, by filtering out the "noise" associated with high recall and low precision (relevance) of retrieval. Though some stages of the information coding can be automated, the intervention of human information specialists would be crucial in order to ensure the proper monitoring and control of language, concepts and rules.

Gray Literature is increasingly recognized as an important area of research information study, because "in a R&D environment [it] represents the cutting edge of this knowledge and so its management is of utmost importance (Jeffery, 1999, p. 1)". As Jeffery writes, an organization documents and stores its knowledge assets within gray literature.

3.2 European Added Value: CRIS as the Lynchpin of the European Research Area (ERA)

In its tender for the ERIS study[8], the European Commission stressed the importance of a European research information system as "… co-ordination of national and European research

7 The Management of Science Systems, OECD, 1999, p. 5.

8 Contract N° COPO-CT -2000-00002

programmes, including the mutual opening-up of national programmes, mapping of excellence, definition of a European approach to research infrastructure, better use of instruments of indirect support to research, benchmarking of national research policies…" The Commission cited that while initiatives exist (i.e. CORDIS, ERGO and euroCRIS), no comparative data of national research policies is readily available, under "one roof".

CRIS can be used as a Facilitator of Integration in that access could be given equally to less prominent researchers, less prestigious institutions, and less favored regions. This would have a democratizing effect on science.

It is known (Oxbrow and Abell, 2002, p. 26) that Europe suffers from a low entrepreneurial spirit (as compared to the United States for example) and less professional mobility than the US. Barriers to this no doubt include language and culture. CRIS can be instrumental in bridging these barriers. Europe's strength (as compared to the US) might be said to be in its networking capabilities. Here euroCRIS will prove instrumental in spearheading the drive for increased CRIS awareness.

The European Commission recently conducted its most recent "Eurobarometer" survey of attitudes towards science in Europe. As the survey shows, there is work to be done in the promotion of science in Europe. CRIS can play a role. The report demonstrates that scientists have a strong image in Europe, though the knowledge they possess affords them much power (European Commission, 2002, p. 3). The survey goes on to report that most Europeans do not believe science will help to "eradicate" poverty and famine (European Commission, 2002, p. 8), though a majority do believe a cure for AIDS and Cancer will come, and that science and technology will help pave the way for a better tomorrow. Basic science is also highly regarded. Finally, Europeans do not know much about European excellence in science (European Commission, 2002, p. 14), though they would like to see Europe more heavily involved. However, they think better organization is key here, not increased funding. This is where the role of CRIS can be pivotal.

4 Barriers to Wider Use of CRIS Systems on Global Scale

There are several issues that have yet to be settled, owing to the dynamic nature of the Internet and new forms of collaborative work. These are listed here, but left for a more critical review in a future study. The issues include information quality and timeliness, language and the ability to convey social cues and non-verbal communication, information presentation (maximizing the Web's capabilities), intellectual property rights, publication and copyright, liability and other legal issues, privacy, national security, and foreign access to sensitive data.

A major barrier to greater use of CRIS is strengthening awareness of CRIS among scientists, administrators and policy-makers. CRISs have generally been considered a starting point for locating information, not the repository of what is actually being sought.[9] Well-established researchers have long felt that they do not need the services of such systems to locate collaborators or information, as the formal and informal networks serve this purpose better. This may no longer be the case with the growing cross-disciplinary nature of science.

5 How are CRISs currently used around the globe?

During April-May of this year we are conducting a global survey of CRISs. The results will be presented at CRIS2002. It is expected that we shall learn from this study of the uses of CRIS, the problems facing CRIS developers, directions for future development, and how global CRISs can cooperate to enhance science. A catalog of CRISs is available on the Web, sponsored by NIWI,

9 Jostein Helland Hauge, Director of the Research Documentation Section, Bergen University Library, Norway, personal email, 8 April 2002.

the Netherlands Institute for Scientific Information Services. DRIS, the Directory of Research Information Systems, may be accessed at
http://www.niwi.knaw.nl/cgi-bin/nph-dris_search.pl?language=us.

6 Summation

CRIS can and must take center stage as the mechanism of information provision in the world of science. In the ever-growing complex scientific world, we are faced with mounting challenges. Increasingly we are witness to pressures on universities and scientists to provide answers to the problems of society, while working in multi-disciplinary and cross-border teams, with increased funding competition. The need for quality scientific and technological information, available to all, is apparent.

In our knowledge-based society, where so much is dependent on scientific discovery, we must be positioned to intelligently make use of information and apply it to address the problems and issues of the day.

7 Acknowledgements

The author wishes to thank his many colleagues at euroCRIS and Bar-Ilan University for years of stimulating conversations, both face-to-face and digitally – especially Keith Jeffery, Anne Asserson, Jostein Hauge, Bernd Niessen, Peter Finch, Richard Tomlin, Hunt Williams, Natalie Lapidot, and Andrei Lopatenko.

8 References

Boyer, Patricia and Cockriel, Irv. (1998): Factors Influencing Grant Writing: Perceptions of Tenured and Nontenured Faculty. In: *SRA Journal*, Spring, pp. 61-68, USA.

euroCRIS and the European Commission. (1999): *CERIF 2000 Final Report*, Luxembourg.

euroCRIS and the European Commission. (1998): *ERGO Pilot Project Final Report*, Luxembourg.

euroCRIS and the European Commission. (1998): *Code for Good Practice*, UK.

European Commission. (2002): Europeans, Science and Technology: Survey Findings. In: *RTDinfo*, Special Edition, March, Luxembourg.

Gibbons, Michael (1994): *The New Production of Knowledge*. London: Sage.

Jeffery, Keith G. (1998): The Future of CRIS. In: *CRIS98: The Nutcracker*. <Accessed 14 April 2002>.

Jeffery, Keith G. (1999): *An Architecture for Grey Literature in a R&D Context*, unpublished.

Jewitt, G.P.W and Gorgens, A.H.M. (2000): Facilitation of interdisciplinary collaboration in research: lessons from a Kruger National Park Rivers Research Programme project. In: *South African Journal of Science*, Volume 96, August.

Koku, Emmanuel, Nazer, Nancy and Wellman, Barry. (2000): Netting Scholars: Online and Offline. In: *American Behavioral Scientist*, Volume 43 (preprint of article), USA.

Mark, Gloria. (2000): Social Foundations for Collaboration in Virtual Environments. In: *Access to Knowledge: New Information Technologies and the Emergence of the Virtual University*, Tschang, Ted and Senta, Tarcisio Della, Pergamon Press, UK.

OECD. (1999): *The Management of Science Systems*. <Accessed 1 December 2001>.

Oxbrow, Nigel and Abell, Angela. (2002): Is There Life After Knowledge Management? In: *Information Outlook*, 6:4, April, Special Libraries Association, USA.

Tomlin, Richard. (2000): *CRIS and the Challenge of New Research Paradigms*. In: CRIS2000, June, Helsinki. <Accessed 14 April 2002>.

Williams, Huntington. (2000): Research Partnerships in the Internet Age. In: *Community of Science*, USA.

9 Contact Information

Eric H. Zimmerman
Bar-Ilan University
The Research Authority
The Begin Building
Ramat Gan 52900
Israel

e-mail: zimmee@mail.biu.ac.il

Current Research Information as Part of Digital Libraries and the Heterogeneity Problem Integrated searches in the context of databases with different content analyses

Jürgen Krause (Keynote Speaker)
University of Koblenz-Landau and Social Science Information Centre (IZ Bonn), Germany

Abstract

Users of scientific information are now faced with a highly decentralized, heterogeneous document base with varied content analysis methods. Traditional providers of information such as libraries or information centers have been increasingly joined by scientists themselves, who are developing independent services of varying scope, relevance and type of development in the WWW. Theoretically, groups that have gathered current research information (CRI), literature or factual information on specialized subjects can emerge anywhere in the world. One consequence of this is the presence of various inconsistencies:

- Relevant, quality-controlled data can be found right next to irrelevant and perhaps demonstrably erroneous data.
- In a system of this kind, descriptor A can assume the most disparate meanings. Even in the narrower context of specialized information, descriptor A, which has been extracted in an intellectually and qualitatively correct manner, and with much care and attention, from a highly relevant document, is not to be compared with a term A that has been provided by automatic indexing in some peripheral area.

Thus, the main problem to be solved is as follows: users must be supplied with heterogeneous data from different sources, modalities and content analysis processes via a visual user interface without inconsistencies in content analysis, for example, seriously impairing the quality of the search results. A scientist who, for example, is looking for social science information on subject X does not first want to search the social science literature database SOLIS and the current research database FORIS, and then the library catalogues of the special compilation area of social sciences at the library catalogues and in the WWW – each time using different search strategies. He wants to phrase his search query only once in the terminology to which he is accustomed without dealing with the remaining problems.

Closer analysis of this problems shows that narrow technological concepts, even if they are undoubtedly necessary, are not sufficient on their own. They must be supplemented by new conceptual considerations relating to the treatment of breaks in consistency between the different processes of content analysis. Acceptable solutions are only obtained when both aspects are combined. The IZ research group (Bonn, Germany) is working on this aspect in four different projects: Carmen, ViBSoz, ELVIRA and the ETB project. Initial solutions for transfer modules are available now and will be discussed.

Keywords:
virtual library, current research information, content analysis, metadata, heterogeneity, text-fact integration, multimodality, ELVIRA, CARMEN, ViBSoz, ETB

1 Current research information as part of digital library integration

The objective of a specialized virtual library is to enable users to gain integrated access, within a complete system and irrespective of the location and time, to all relevant information in their spe-

cial scientific field – from metadata at the individual document level through to full text which can be called up online. Currently existing media breaks in the acquisition of literature and in searching must be overcome. An integrated user interface has to ensure user-friendliness and, compared with existing solutions, makes it easier for the users to get information of different data types and sources.

Digital libraries combine scientific information from traditional libraries, information and documentation centres with their specialized databases, and from the WWW. They are hybrid libraries containing electronic and printed texts, as well as information with completely different modalities and medialities which, in addition to literature, include current research information, factual databases, photos, graphics, films and teaching materials. They therefore combine rules and standards which are each valid for the different worlds that have to be integrated.

Current research data are today treated like text data and are therefore accessed by descriptors. However, they are rather a mixed form consisting of factual as well as textual information. From the aspect of digital libraries, they are one subset which must interact with all other subsets. Users looking for text documents concerning a special topic might want to retrieve all the data available in different databases. They might be looking for time series, survey results, current research data or a list of experts. Therefore, future innovative research information systems should be embedded in the context of extensive integration of different data types, modalities and medialities. The concept describing this form of integration used here is the digital library concept.

The CRI problem of integration belongs to the same class of problems as that of integrating texts combined with different indexing methods. At the first level a CRI system will face documents which were made accessible through various different methods of content analysis. The use of detailed thesaurus for special purpose collections are one example. At the second level the differences between CRI content analysis and those of literature databases (and others) have to be handled.

1.1 Libraries and specialized information centers

Libraries use the traditional method of standardization to provide users with integrated access to their collections. In Germany alone, there are a large number of relevant standards for libraries. In addition to DIN and ISO standards, there are library regulations and exchange formats based on these standards[1]. With regard to bibliographical description (non-subject indexing), DIN 1505 "Titles of literature" form the basis for the "Alphabetic Cataloging Rules, Scientific Libraries (RAK-WB)". The related German exchange format is the MAB format. In spite of this implemented standardization, the results of non-subject indexing are often heterogeneous. If German collections are also offered together with Anglo-American collections, this produces the problem of comparison between RAK-WB and AACR2 ("Anglo-American Cataloging Rules") and the related exchange format MARC21 (Machine Readable Cataloging Records). The existing conversion software provides less satisfactory results, which is why libraries often do not convert the already recorded documents in AACR2 and MARC21, and instead re-record them.

Differences in bibliographical descriptions were not so pronounced and of little consequence in using printed catalogues or index cards. During an integrated search in digital libraries, however, these types of heterogeneity already result in the loss of relevant documents.

Much more critical is the situation with respect to content analysis by means of classifications or thesauri (verbal content analysis). In this case Germany has no longer been able to implement uniform standards. Although the RSW rules ("Keyword Catalogue Rules"), which were used to create a standard keyword file, the SWD, have become widely accepted in subject indexing dur-

1 The following comments are based on Gömpel/Niggemann 2002, Krause/Niggemann/Schwänzl 2002, Geisselmann 1999

ing the past years, special libraries and specialized information systems normally use their in-house-developed special thesauri (the social sciences, for example, with the SOLIS thesaurus). The changeover, for example from SWD or the SOLIS thesaurus, which is also available in English, to the world's most widespread American "Library of Congress Subject Headings" (LCSH) has not taken place and is also not trivial.

German scientific libraries have been unable to agree on a system for classifications. Although the Regensburg composite classification, the GHB listing system and the basis classification are all fairly widespread, it is difficult to convert them to the world's most commonly used system, i. e. the "Dewey Decimal Classification" (DDC), or to the Library of Congress Classification.

Again, specialized libraries and information centres frequently use their own individual special classifications.

The contents of a large part of library collections have also not even been indexed. Geisselmann 1999: 46 mentions a quota of 40 % to 60 %. According to Krause/Niggemann/Schwänzl 2002, the percentage of verbal subject indexing ranges between around 12 % in the south-west German library network and around 46 % in the Bavarian library network. This means that only the terms of the title, and possibly the full text, are available for searching.

Heterogeneity of this kind is also typical for the data of specialized information centres. Since they only handle a few specialized disciplines, different rules are required for more extensive subject indexing than for general libraries. Every discipline has in turn its own rules which are not matched to associated disciplines, a situation that leads to unavoidable interoperability problems in intersecting areas.

The same is true of CRI, also if connected with specialized libraries and information centres. To give some examples: CERIF, the Common European Research Information Format, developed by members of EUROCRIS, recommends for subject indexing the „ORTELIUS" thesaurus. The Information Centre for Social Science Research IZ in Bonn (Germany) uses the SOLIS thesaurus and the SOLIS classification (both German, English, French and Russian) for the CRI system FORIS (Germany, Austria and Switzerland), which are also used for the literature database SOLIS. CRI data at the WWW homepages of the universities in Germany have no standardized thesaurus or classification at all. With respect to classification in Austria most accepted is „OESTAT" which is based on the „Fields of Science and Technology" of the UNESCO (proposed in the OECD Frascati Manual 1980; http://www.statistik.at/). This classification is used in the Austrian database AURIS which contains all university projects of the country (www.auris.ac.at).

The Belgium CRI database IWETO: does not have a thesaurus or authority list for keywords at all (like the Danish National Research Database). The keywords are freely chosen by the researchers when providing the information. As a subject classification IWETO uses a specific version of the CERIF-disciplines and an older version of the NABS.

It is interesting that at the end of 2001 German libraries decided to at least partially eliminate the heterogeneity between German and Anglo-American rules by trying to replace RAK-WB by AACR2 ("Anglo-American Cataloging Rules") and the associated exchange format MAB by MARC21 (Machine Readable Cataloging Records) (see Gömpel/Niggemann 2002 for discussion). This decision of the German "Standardization Committee" was approved by the Library Committee of the German Research Association. Thanks to a joint arrangement and agreement, the traditional form of standardization therefore applies once again to rules for an important area. No such attempts have been made with regard to interoperability between libraries and specialized information centers.

1.2 WWW

For many years, CRI and every type of textual documents have been found to an increasing extent in the Internet websites of university institutes and research institutions. This information

will also be accessed in a digital library. The Internet is therefore extending the previously described system of a wide range of different standards for scientific information by adding an increasing number of new formats[2]. Uniform Resources Identifiers are used as a common organization system, but they regulate little more than the use of characters. Access mechanisms to WWW objects use standardized protocols such as tcp/ip, http, ftp, telnet, mail, news, etc. Subject indexing is not one of the standardization objectives of these techniques – bibliographical description in the conventional sense does not exist.

Parallels to the bibliographical description and subject indexing of information and documentation centres and libraries are found in the WWW under the term "metadata" (as a starting-point for knowledge representation techniques through to the "semantic web"). The DublinCore Meta Data Initiative (DCMI), which probably has the broadest basis in vocabulary development, pursues a similar strategy as the HTML standard with slogans such as "everything optional, everything repeatable". The HTML standard contained the following statement: "WWW parsers should ignore tags which they do not understand, and ignore attributes of tags they do not understand."

This basic attitude, which is not found in the traditional standardization of libraries and specialized information centres, stems from the fact that the WWW has hardly any means of bringing pressure to bear to ensure that standards are implemented. Non-consideration of the proposed rules must therefore also be modeled right from the very beginning. (http://www.w3.org/History/199921103-hypertext/WWW/MarkUp/MarkUp.html).

1.3 General Trends

This review of standardization activities in digital libraries shows:

All attempts, especially in connection with the content analysis of data sets, follow the traditional standardization philosophy which is also closely attached to the theoretical principles of content analysis in information and library sciences. Documents are recorded uniformly based on a standardized, intellectually controlled method, which is developed and implemented by a central organization. In this concept maximum priority is attached to data consistency, so that the user (idealiter) is always faced with a homogenized data set. The widest possible regulation will ensure attainment of the consistency that is regarded as necessary for user questions. In subsets such as library catalogues and specialized databases, this model certainly turned out to be a feasible method which has proved its worth over the past twenty years. However, the general conditions have changed. The technological, economic, political and social changes in recent years have produced trends and opinions which contradicted this models in some aspects.

In spite of all plausibility of the associated advantages, the demand for standardization has only actually been implemented in some subsets. At the latest since the introduction of specialized databases, whose central new content were metadata to newspaper articles, users have had to work with different content analysis concepts. A large number of subject indexing thesauri and classifications for the aspired-to subcomponents in digital libraries up to automatic indexing solutions are now represented in all disciplines.

The discussion in Germany concerning the conversion to American standards does not contradict this general trend. It involves an important subsegment where it remains clear that consistency is increased, but will also not be sufficient in a best case scenario for the interoperability required by virtual libraries.

2 See Krause/Niggemann/Schwänzl 2002: Section 1.3

1.4 Consequences and potential solutions

The above observations harbor serious consequences for digital libraries: Narrow technological concepts, even if they are undoubtedly necessary, are not sufficient on their own (see Krause 1996 for more details). They must be supplemented by new conceptual considerations relating to the treatment of breaks in consistency between the different processes of content analysis. Acceptable solutions are only obtained when both aspects are combined.

The existing extensive heterogeneous databases have been developed in very different ways. Examples of the now attained technical integration of heterogeneous databases in Germany[3] are the KOBV library network (see Lügger 2000 and http://www.kobv.de/se/cont.html), the virtual library network in North Rhine-Westphalia or the virtual library catalogue DVK (http://www.ifs.tu-darmstadt.de/dvk.html[4]).

What the KOBV and comparable projects have not been able to do so far is take sufficient account of the different content analysis processes of the document subsets. The existing conceptual gap mainly concerns the differences in meaning between the distributed resources tied together in a virtual library:

The wide range of breaks in consistency are shown very clearly, especially through the technological merger at several points of libraries, information and documentation centers and WWW sources. A descriptor A may assume the widest possible range of meanings in such a system. In the narrow field of specialized information a descriptor A, which was determined with a great deal of expert intellectual effort from a highly relevant document set, cannot, for example, be equated with term A which provides automatic indexing from a marginal area.

Nowadays, hardly anyone still believes that the document area of digital libraries could be homogenized through global standardization agreements across all subsets, reduced again in organizational terms to a few players or designed by means of an hierarchically organized cooperation model. On the contrary, current concepts start from even greater decentralization in the creation, bibliographical description, subject indexing and distribution of documents, which means that "anarchistic tendencies" will increase still further.

Standardization attempts such as the connection of German-speaking countries to AACR2 are an important step towards provider-overlapping search processes in the heterogeneous data area. However, they do not produce any continuous homogeneity of data, they improve them to some extent. The remaining and unavoidable heterogeneity must therefore be countered by means of different strategies. New problem solutions and further developments are therefore necessary in two areas:

- Metadata
- and the methods of dealing with the remaining heterogeneity. Matching transfer modules must be specified between the individual data types (e. g. literature databases and Internet sources). These modules must counter the lack of standardization by means of automatic methods.

The demands in both areas are closely connected. First of all, the lost consistency will be partially created through the continued development of metadata. Secondly, heterogeneity treatment methods will be used to interrelate documents with different levels of data relevance and subject indexing. The general premise formulated in Section 1 therefore applies to digital libraries: from the aspect of the remaining heterogeneity, their standardization attempts are showing the first visible signs of success.

---

3 See also the Stanford Digital Libraries Project http://www-diglib.stanford.edu/diglib/WP/PUBLIC/DOC104.html and Roscheisen et al. 1997.

4 In ViBSoz "Virtual Social Science Library" the DVK is combined with the SOLIS and FORIS databases of the IZ, Bonn, the library catalogues of the University of Cologne and the database of the Friedrich Ebert Foundation. This takes place with consideration of the heterogeneity problems as discussed here.

2 Handling the Remaining Heterogeneity

We can assume now that standardization efforts such as Dublin Core (DC) or the change from the German RAK standard to AACR2 (see Chapter 1.1) are a useful precondition for comprehensive search processes in the heterogeneous data pool, but despite voluntary consultation by everyone participating in the information process, there is still no sufficient homogeneity in the creation of documents and never will be. The remaining and unavoidable heterogeneity must therefore be treated adequately. The IZ group is working on this aspect in Carmen[5]. It is also one central theme of other current projects of the research department of the IZ: ViBSoz[6], and ELVIRA[7].

The model outlined below represents a general framework in which certain categories of documents with different content analysis are analyzed and referred to one another. The central features are transfer components between the different forms of content analysis, which take account of semantic-pragmatic differences. They interpret the integration between the individual document sets with different content analysis processes (including automatic indexing) on a conceptual basis by cross-referencing the conceptual world of specialist and general thesauri, classifications, etc. The system must know, for example, what it means when term A was used from a specialist classification or a thesaurus for intellectual indexing of a magazine article, but the WWW source could only be automatically indexed. The classification term A could probably only be found by accident in the article, and yet there are conceptual references between both which have to be evaluated.

It is therefore necessary to develop transfer modules between data sets. These modules will permit transfer both in technical and conceptual terms. In these considerations it does not matter, in principle, whether the semantic-pragmatic differences between two texts, between a text and current research information, between a text and the results of a survey in tabular form or a video archive have to be bridged or alternatively between multilingual sources.

In principle, there are three methods which have to be checked and implemented in relation to their effectiveness in individual cases. None of the methods is solely responsible for transfer. They are interlinked and interact with one another.

- The development of an intellectually gained crosswalk of various indexing and classification systems ("cross-concordances").
  Cross-concordances have already been used and implemented in the current projects of IZ and will not be described here (see Krause/Plümer/Schwänzl 2000).
- Statistical methods and neural networks.
- Deductive methods. They can be seen in parallel with artificial intelligence methods in expert systems; and will also not be discussed here.

Using cross-concordances, a simple rule transforms one term into another term suitable for mapping. Although this method works well for some terms, empirical studies have shown that classifications and thesauri are often so different that such simple mappings will be restricted to a sub-

5 CARMEN investigates the conceptual problem of heterogeneous data stores on the basis of an exemplary data pool from the contents of major publishers' servers linked to electronic information primarily in the fields of mathematics, physics and social sciences (see Krause/Schwänzl/Plümer 2000).

6 The "Virtual Social Science Library Project" concentrates on the connection of different library catalogues with the documents of the SOLIS literature database (see Kluck et al. 2000). It is part of the digital library research program of the German Research Association (DFG).

7 "Elektronisches Verbandsinformations- Recherche- und Analyse-System = Electronic Retrieval and Analysis System for Industry Associations" (funded by the German Federal Ministry of Economic Affairs). ELVIRA started as an online system for time series used by major German industry associations to provide information to their member companies. Text retrieval functionality was added in 1998 to search text collections mainly relating to foreign trade and to address fact–text integration. The system is now used by companies in more than 350 installations (see Krause/Stempfhuber 2001).

set of the terms in question. Therefore, a second strategy was integrated in combination with the first transformations which rely on vague methods already successfully applied in information retrieval (IR).

In ELVIRA the texts are indexed automatically using a probabilistic standard algorithm from FULCRUM. Contrary to this, the time series are indexed intellectually using a hierarchical nomenclature (fact thesaurus). The statistical-based crosswalk has to transform these meaning differences of the different content analysis procedures of textual and factual data. In ViBSoz the statistical crosswalk is used between two different thesauri (SWD as a universal thesaurus and SOLIS as a special thesaurus). Unlike cross-concordances, the statistical transformation is not based on the general semantics of the intellectually acquired term-to-term link. Instead, words are transformed into a weighted vector of terms representing the use of the term in the document pool.

In the following the treatment of heterogeneity by means of transfer modules will be described in more detail.

2.1 Treatment of vagueness in information retrieval

Vagueness in IR is normally countered by using statistical approaches in which vagueness is modeled between the user query and the document set, whereby the document level is regarded as the uniform modeling basis. This is shown most clearly when all documents were automatically indexed. Even if additionally intellectually determined descriptors of a controlled vocabulary were produced, modeling follows this homogeneity demand in principle. The user can either carry out his/her search using one of the two term groups and then base his/her search strategy on the controlled vocabulary or alternatively via the free text terms of automatic indexing using another strategy. If he/she chooses to perform a search via both term areas, differentiation in the match is not distinguished, i. e. it is treated as if the semantic differences in both content analysis methods do not exist.

The problems become even clearer when we use digital libraries and link, for example, a social science literature database such as SOLIS with its own controlled vocabulary (IZ thesaurus and IZ classification) with library catalogues, for instance in the ViBSoz Project with the documents of Cologne University Library (USB Thesaurus), which is being developed intellectually according to the keyword list of the German Library (DDB) in Frankfurt. A comparison of two such thesauri or classifications reveals that vagueness already occurs at the semantic description level of two document sets to be integrated in the search, and not just in the user query. Terms in a library classification with its controlled vocabulary form a separate description level. It cannot simply be translated on a 1:1 basis into the terms of another classification, e. g. from the area of specialist information systems. The meaning of a descriptor A in the library classification is different from the meaning of the same term in another classification or even in a thesaurus such as the IZ Thesaurus of the SOLIS social science literature database, even if a simply “vague” connection exists. These vagueness relations can be integrated in a non-differentiated way into the modeling of the IR process. In this case vagueness between the user query and documents will be modeled without explicitly treating differences at the document level between two heterogeneously developed document sets separately by means of transformation modules. Therefore, we speak here about a “one-step” method.

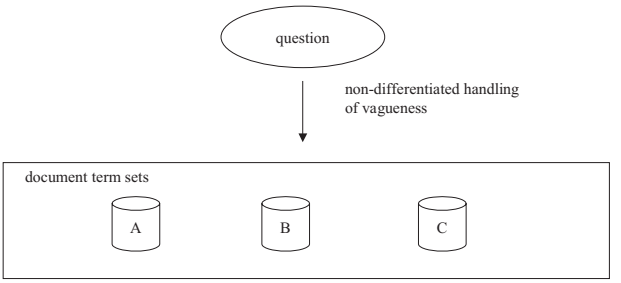


Fig. 1: One-step method

2.2 "Two-step" method and transfer modules

An alternative "two-step" method plays an important role in the projects of the IZ (ELVIRA, ViBSoz and CARMEN). It is based on the thesis that heterogeneous document sets should first be interlinked through transfer modules (vagueness modeling at the document level) before they are integrated in the superordinate process of vagueness treatment between documents and the query (the traditional IR problem). If, for example, three heterogeneous document sets have to be integrated, transfer modules bilaterally treat the vagueness between the different content analysis methods. The aim behind this form of vagueness treatment, which differs considerably from the procedure used traditionally in IR, is to produce greater flexibility and target accuracy of the overall procedure through separation of the vagueness problem. Different forms of vagueness do not flow uncontrolled into one another, but can be treated close to the causal interface (e. g. the differences between two different thesauri). This firstly appears more plausible in cognitive terms and secondly permits the combination of a wide range of modules for treatment of vagueness. Together, these modules can become effective during retrieval using heterogeneous data sets.

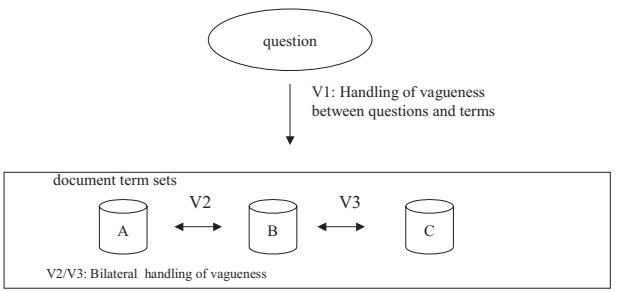


Fig. 2: Two-step method

This procedure appears to be highly promising, especially in view of the broad background of empirical knowledge that IR processes differ more in the volume of results than in the quality of the evaluation parameters such as recall and precision. For example, a probabilistic method can be combined with neural transformation modules in the match between the user query and documents. Alternatively, neural transformation modules can also be combined with IR processes based on Boolean algebra. But even if the transformation modules between the heterogeneous

document sets use the same similarity function as at the IR level between the user query and documents, the results between the "one-step" and "two-step" methods will probably differ.

2.3 Integration of the results of vagueness treatment through transfer modules

The architectures which were introduced as the "one-step" method and which model vagueness in an unspecified manner between query terms and document terms have no integration problems whatsoever since the query terms occupy the input layer of the neural network and the documents the target structure. However, vagueness relations occur in several places in the two-step method. With just two heterogeneous document sets, it is necessary to answer the question of how the vagueness treatment of "query → document set" can be combined with that between document sets.

2.3.1 ELVIRA: directed transfer

ELVIRA permits access to data on production, foreign trade, the economy and economic structures. The time-series fact tables are not addressed via their cell values and the table names, but indirectly via intellectually assigned descriptors relating to three categories, i. e. the subject in question (e. g. export), industry/product (e. g. microwave ovens) and the country. User tests quickly showed that association customers demand both time series and textual information sources to solve their problems. The transfer problem arises because different indexing methods are used in time series and texts. Time series are indexed intellectually and the descriptors used are assigned hierarchically (e. g. by nomenclatures of the Federal Statistical Office). Some texts are indexed automatically, other ones additionally with the same nomenclatures as the fact tables. In this case not only standardized thesaurus terms, but also words which occur in a text are valid search terms. Texts also appear to refer less to individual products, as is the case, for example, with time series of deeply structured production statistics. Texts can often only be found at higher aggregation stages (e. g. for drive systems or electric motors as a whole) and not for special types of motor such as direct-current motors.

In the case of text-fact integration like in ELVIRA, simple directed transfer appears to be the rule. The user first looks for facts and then wants to obtain evidence of the associated texts (and vice versa) in an iterative search step (or right from the beginning). In a "two-step" architecture simple directed transfer does not place any special demands on the integration of both vagueness treatment modules. However, the difference in the function of both architectural concepts becomes clear in this case. Traditional term extension strategies therefore resemble the transfer modules discussed here. But since they work within the meaning of one-step methods, they have different impacts, even if the same mathematical processes were used. This becomes even clearer and - in the impacts - more serious when several heterogeneous text sets are integrated.

2.3.2 Transfer between heterogeneous text sets

The parallel with 2.3.1 is produced in text retrieval when, for example, an IZ user, who has precise knowledge of the IZ Thesaurus and gears his/her search strategy according to this thesaurus, also wants to obtain the texts of Cologne University Library without knowing anything about the thesaurus used there. This technique was implemented as one of the first transfer modules of ViBSoz. However, this will probably be the exception rather than the rule.

More than two document sets with different indexing methods can firstly be expected in heterogeneous text searches. Users can also be expected, who are not incorporated in any of the used content analysis methods in such a way that they can direct their search strategies according to a special standard.

This does not matter in the one-step method as no distinction is made between bilateral transfers. However, the two-step method must ensure that the advantages of measuring vagueness differentiated between two indexing systems are not lost as a result of the way in which these values are integrated in the internal search. The simple method of non-specific extension of the query term is therefore no longer sufficient. If the query is supplemented, for example, by a term from the vagueness definition B→ C, this term will only become effective in C, but not in A, for instance. Instead of a globally effective term extension strategy, this makes it necessary to extend the query terms on a differentiated basis and operate on different subquantities of the heterogeneous document sets.

3 The consistency problems of digital libraries as a general standardization problem

A general discussion concerning the limits of current standardization problems in industry and administration[8] is being presided over by the responsible national authorities (DIN in Germany), European bodies (EUN) and international organizations (ISO). This discussion shows that the problems described here for digital libraries are of a very general nature. However, there are clear indications that the traditional methods of standardization are coming up against limiting factors in an increasing number of areas (not only in digital libraries) – in spite of all the detailed improvements in the methods. On the one hand they appear indispensable and substantially improve quality and cost-effectiveness in subsets. On the other hand they can still only be partially implemented, with increasing costs, within the framework of global provider structures and changed general conditions. The current standardization concepts must therefore be revised. The remaining and unavoidable heterogeneity must be countered by different intellectual and automatic methods of retrospective conceptual integration. A new way of thinking regarding the existing remaining demand for consistency retention and interoperability is necessary. It can be described by means of the following premise: Standardization must be considered from the aspect of the remaining heterogeneity. A solution strategy, which takes account of general current technical, political and social conditions, can only be obtained through joint interaction between intellectual and automatic methods relating to heterogeneity treatment and standardization. In specific terms, this means the following for standardization projects: If standardizations cannot be implemented or can only be implemented to a partial extent in subsets in a reasonable amount of time, every remaining detail must be analyzed specifically to determine the consequences of the lack of standardization and how the remaining heterogeneity can be at least roughly countered by means of automatic or intellectual methods. Any resulting costs and losses of quality must be compared with the time expenditure and success prospects of other intensive attempts to attain standardization.

4 Literature

Geißelmann, Friedrich (1999): Zur dritten Auflage der RSWK, Bibliotheksdienst, 33. Jg., H. 1

Gömpel, Renate, Niggemann, Elisabeth (2002): RAK und MAB oder AACR und MARC? Strategische Überlegungen zu einer aktuellen Diskussion, ZfBB 49.1.

Kluck, Michael; Krause, Jürgen; Müller, Matthias; in Kooperation mit Schmiede, R.; Wenzel, H.; Winkler, S.; Meier, W. (2000): Virtuelle Fachbibliothek Sozialwissenschaften: IZ-Arbeitsbericht 19, IZ Sozialwissenschaften, Bonn.
http://www.bonn.iz-soz.de/publications/series/working-papers/index.htm#Virtuelle

Krause, Jürgen (1996): Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung. (IZ-Arbeitsbericht Nr. 6) Bonn 1996.

8 to appear in http://www.din.de

Krause, Jürgen; Niggemann, Elisabeth; Schwänzl, Roland (2002): DIN-SICT Papier „Strategie für die Standardisierung der Informations- und Kommunikationstechnik (ICT)“ (to appear)

Krause, Jürgen; Plümer, Judith; Schwänzl, Roland (2000): Content Analysis, Retrieval and Metadata: Effective Networking for Mathematics, Physics and Social Sciences, RC33-Session “New Conceptual Developments in Information Systems and the WWW”. Proceedings der Fifth International Conference on Social Science Methodology, October 3 - 6, 2000. Köln (CD-ROM).

Lügger, Joachim (2000): Über Suchmaschinen, Verbünde und die Integration von Informationsangeboten, Teil 1: KOBV-Suchmaschinen und Math-Net. ABI-Technik 20, Nr. 2, S. 132 - 156.

Roscheisen, Martin; Baldonado, Michelle; Chang, Kevin; Gravano, Luis; Ketchpel, Steven; Paepcke, Andreas (1997): The Stanford Infobus and its Service Layers. Augmenting the Internet with Higher-Level Information Management Protocols
http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1997-0065
[eingesehen: 26.05.98]

5 Contact Information

Jürgen Krause
Social Science Information Centre (IZ Bonn)
Lennéstr. 30
D-53113 Bonn
Germany

e-mail: krause@bonn.iz-soz.de
http://www.uni-koblenz.de/~krause/

CERIF: Past, Present and Future: An Overview

Anne Asserson, Keith G Jeffery, Andrei Lopatenko
UiB, CLRC, MU

Summary

CERIF (Common European Research Information Format) provides a canonical reference data model at both data and metadata levels. As such it is a model for the development of new CRISs (Current Research Information Systems) and a template both for data exchange between CRISs and for mediating access to multiple heterogeneous distributed CRISs. CERIF originated in 1988 but was based on earlier work in several European countries. The CERIF91 standard had some defects which became apparent in use. In 1997 a working group of the EC was set up to produce CERIF2000. This is a formal datamodel and thus unambiguously implementable. The CERIF Task Group of euroCRIS is working actively on implementations, lessons learned and improvements.

1 Introduction

This paper is organised as follows. CERIF (Common European Research Information Format) has a history described in Section 2. Section 3 describes how it is used currently as a datamodel for CRISs (Current Research Information Systems) in several variants in several countries and raises some issues. A Task Group of the euroCRIS organisation (www.eurocris.org) is now the custodian of CERIF ensuring its integrity, flexibility and usability. Section 4 outlines some of the developmental directions for CERIF and discusses the relationship to the original aims and objectives. Section 5 concludes.

2 CERIF Past

2.1 CERIF91

CERIF (Official Journal 1991) has its origins in the late eighties arising from the Liaison Committee of Rectors' Conferences of Member States and parallel, independent, work by several national Research Councils especially in projects IDEAS (Jeffery et.al 1989) and EXIRPTS (Naldi et al. 1992). A Group was formed to formulate CERIF91 (vanWoensel 1988a); (vanWoensel 1988b). Experience with CERIF91 led, in 1997, to the requirement for a new CERIF standard. The major aspects were:

a) the original CERIF covered only research projects as entities with persons, organisations and other information represented as attributes. Users of CRIS wanted to extend it to data on persons, organisations and other entities;
b) the "research subject classification scheme" recommended in CERIF 1991 had not been updated since 1988 and needed to be extended to cover the new data areas plus give enhanced coverage of existing ones;
c) new technologies, in particular, the introduction of the Internet and World Wide Web, had changed the nature of basic CRIS activities and opened new ways to serve various CRIS user groups.

The (CERIF2000) standard was created in late 1999 following two years of work by the Group formed to undertake this task. There were strong interactions with the ERGO (European Research Gateway Online) Group which was implementing a prototype portal system (ERGO) based on (a slightly extended) CERIF91.

2.2 CERIF2000

2.2.1 Problems with CERIF91

One of the major problems with CERIF91 – and operational CRISs from the eighties and nineties – was that they tended to have a single-entity focus. There were three main classes of systems:

(a) those focused on projects, with other information as attributes e.g. ASCENDA (UK);
(b) those focused on persons, with other information as attributes e.g. BEST (UK) or COS (USA);
(c) those focused on organisational unit, with other information as attributes e.g. LABO (FR);

the characteristics of all of them included:

(1) Problems of repeating groups. For example, in the case of CRISs focused on projects, it was not possible to record accurately the relationship between a person and this project. In fact usually only the project leader was recorded. Thus there were problems:
    (i) being unable to record repeating groups (multiple instances of groups of attributes representing e.g. person repeating against one project);
    (ii) having attributes with the same value (e.g. a group of attributes representing person) occurring multiply in the database – where the same person was associated with more than one project

    Similar problems occurred with repeating of organisational unit, publication, equipment, facility etc etc. This is known in the database theory literature as a problem of functional dependency.
(2) Problems of relationships. For example, in the case of CRISs focused on projects:
    (i) it was not possible to record that project A was a subproject of project B,
    (ii) nor a follow-on project from project B.
    (iii) Similarly, it was not possible to record that person M was project leader, person N was the designer, person O was the analytical chemist for project A.
    (iv) It was not possible to indicate that project A was a cooperation between two organisational units.

This all pointed to deficiencies in the data model. Specifically, it indicated that it was necessary to define more entities (rather than attributes of an entity) and that it was necessary to represent relationships between those entities that included 1:n, n:m and recursion (self-referencing).

2.2.2 CERIF2000 Design

(CERIF2000) has a particular feature of three major entities {project, person, organisational unit}interlinked through n:m relationships with {role} and {date / time} attributes and each capable of recursive reference (e.g. the relationship between one organisational unit and another, one project and another, one person and another). This provides great flexibility and robustness because not only can complex role and date-limited relationships between the three major entities be expressed but also other entities can be linked by role/date relationships to any or all of these three major entities. The following example facts can all be recorded accurately by CERIF2000:

(1) person *a* works for organisational unit *j* which is a sub-unit of organisational unit *k* which is a sub-unit of organisational unit *l*

(2) person *b* works for organisational unit *m,* a sub-unit of organisational unit *n*
(3) result-publication *x* came from project *p* which is a sub-project of project *q*
(4) person *c* is a reviewer of result-publication *x*
(5) person *d* is the editor of the journal or proceedings containing result-publication *x*
(6) organisational unit *h* (a publisher) claims copyright on result-publication *x*
(7) person *a* transferred copyright to organisational unit *h*
(8) for result-publication *x*, person *a* transferred copyright

and since all these statements include roles (eg author) and also date/time stamping {<start date/time><end date/time>}it is possible using, for example, date range intersection, to induce from (7) and (8) into person *a* transferred copyright to organisational unit *h* for result-publication *x*.

It is clear from the example that CERIF has both tremendous expressive power yet has the flexibility to allow simplified instances of the data model – for example in an academic environment the <date/time> attribute could be, simply, academic year thus allowing easy retrieval of result-publications of academic year yyyy with authors and organisational units (and projects if desired). However, it does not even end there. Additional unique features of (CERIF2000) which provide even greater flexibility are:

(a) all contact information is stored in one entity which has relationships (with role and datestamping) to person and to organisational unit. Thus a person may have different contact information instances for different roles;
(b) all attributes with enumerated lists of valid values have those values stored in an entity with a relationship to the entities including the attribute thus providing flexibility and extensibility;
(c) all textual attributes have subordinate entities with language variants to allow a clean, flexible and extensible implementation of multilinguality;
(d) CERIF is delineated by key reference links to databases known to be pre-existing with more detailed information on certain entities e.g. publications, patents

2.2.3 Three Data Models

(CERIF2000) also proposes three data models:

(a) 'full CRIS' datamodel which defines entities, attributes and relationships for a 'greenfield' CRIS implementation;
(b) export CRIS datamodel which provides a set of proposed subsets of (a) for data exchange between CRISs capable of exporting / importing CERIF;
(c) CRIS metadata datamodel, a proper subset of (b), which provides a succinct description of the contents of a CRIS in a form readable by any CRIS capable of importing / exporting CERIF and also forms the key to the development of portal systems wishing to provide a homogeneous view over heterogeneous CRISs.

3 CERIF Present

3.1 Introduction

CERIF is established. EC (European Commission) tenders in the area of ERIS (European Research Information System) emphasise CERIF. It is - with the CRIS Conference Series - a major raison d'etre of euroCRIS. It is used either in practice or as a best-practice reference. Utilisation of CERIF in practice has advanced our knowledge.

A list of current known CRISs which have CERIF compatibility is given:

- SICRIS http://sicris.izum.si/default.asp?lang=eng: a CRIS providing access to total university research in Slovenia. It is highly CERIF-2000 compatible, based on MS SQL installation of CERIF-2000. Uses CERIF schema and CERIF vocabularies
- AURIS-MM: The CRIS developed to provide access to Austrian university research. It is highly CERIF-2000 compatible, based on Oracle installation of CERIF-2000. It uses CERIF schema and CERIF vocabularies. It extends CERIF to serve better Austrian users (uses Austrian vocabularies), to deal with other information (multimedia, web sites) and to serve for better information retrieval (full-text indexes, views)
- CRIS-MER http://www.ercomer.org/research/ReSchools/Re_plans.html: under development for research information on Migration and Ethnic Relations. Highly CERIF-2000 compatible, RDBMS implementation, uses CERIF schema and vocabularies. It extends CERIF for humantitarian information (new vocabularies and relations)
- Scottish Research Information System http://www.scottishresearch.com: is a CRIS for public research in Scotland. It is CERIF-2000 compatible. The data schema and metadata schemas to describe data are based on CERIF-2000.
- ARAMIS http://www.aramis-research.ch: a CRIS Intended to provide information to interested parties about research which is financed or carried out by the Federal Government in Switzerland. It has CERIF-2000 compatible data structures.
- INTACCOMP http://www.intaccomp.ro/: is a network of key data about Central Europe research projects sponsored with either national or international funds. The data schema, vocabularies and metadata schema for data exchange are based on CERIF-2000 recommendation (http://www.man.poznan.pl/ist/isthmus/programme/slides/goczyla/ISThmus-goczyla.PPT)
- SAFARI http://safari.vr.se/: a CRIS to provide information to Swedish academic research already available on the Internet. It is based on metadata technologies. The metadata schema is based Dublin Core and utilizes CERIF vocabularies to classify subjects.

Joint Electronic Submission (Je-S): a proposed system for electronic submission of grant applications (and hence into databases) of the 6 UK Research Councils has specified CERIF compatibility.

Experience has shown that CRIS developed for public research in Universities are commonly very compatible with CERIF-2000, even if they were developed without knowledge of CERIF-2000. Austrian examples are University of Salzburg, Technical University of Graz, University of Linz but similar examples are found in all countries. This is not surprising as CERIF2000 was defined utilising the experience of CRIS managers from all over Europe.

Particular implementations at UiB in Bergen (emphasising research results-publications) and in CLRC near Oxford (as a corporate data model for a R&D enterprise to drive business processes and provide R&D management information for decision support) have stress-tested the CERIF model and led to proposals for some extensions.

In parallel, detailed technical work on the EC-provided variants of CERIF schemas at the website (www.cordis.lu/cerif) by Andrei Lopatenko at TUW (Vienna University of Technology) has indicated some deficiencies, and – interestingly - some variations from the model defined by the CERIF2000 Group. Andrei has also implemented a CERIF-compatible CRIS at TUW named AURIS-MM and has provided a 'clean' version of CERIF for euroCRIS.

Thus we have several kinds of CERIF-developments today:

(1) corrections to the EC-provided datamodel and schemas;
(2) extensions to CERIF for research results-publications;
(3) extensions to CERIF for corporate data model usage;
(4) precision of CERIF dictionaries (lists of valid terms) associated with attributes that have the property of an enumerated list of possible valid values;

These developments all aim either:

(a) to make precise and formal the definition of CERIF and to formalise its change control processes to ensure clarity;
(b) to extend the capability and usability of CERIF for supporting CRIS in the widest sense, with extensions both in depth (detail) and in breadth (business requirements areas supported);

One interesting feature is that of all the extensions to CERIF proposed very few actually require extensions to CERIF – the original datamodel had the capability to represent the requirement. This is a testament to the skill and ability of the CERIF2000 Group.

3.2 Extensions

3.2.1 UiB

UiB had a particular need to relate {result_publication} to {person a} who at the time was working for {orgunit n} and to {person b} who was working for {orgunit m} In other words, they wished to relate a particular {result_publication} to the intersection of {person} and {orgunit}.

This can be expressed in CERIF2000. However, it requires the induction that if the date range (with appropriate role) in the relationship {person-result_publication} intersects the date range in the relationship {person-orgunit} then the person was working for that {orgunit} at the time of publication. Of course, the {person} could have been working for more than one {orgunit} and more than one {person} could have been working for the same {orgunit} and on the same {result_publication}.

For reasons of efficiency UiB decided to implement this as ternary relation {person-orgunit-result_publication}, without role and dates and so constructing a specialised fixed ternary relationship. This has the advantage that induction is not required, and that there are fewer join operations during selection (search). It has the disadvantage that it is difficult to represent the role and time relationship of a {person} to either a {result_publication} or to an {orgunit}, and it also makes it more difficult to handle multi-author publications (because of repetition of the other two key attributes in the ternary relation). In practice this has proved inefficient in implementation.

As an aside during this work it was noted that there is no (recursive or non-recursive) link table {result_publication-result_publication}. Such a link-table could be useful for handling semantics such as ‘paper x in proceedings y’ where the proceedings is clearly a separate publication or the relationship ‘paper x is an extended journal paper from paper y given at conference z’. In (CERIF2000) the original idea as that publications were recorded outside of CERIF, and that CERIF should hold only a pointer (e.g. URI) to the publication. It is now clear that this is insufficient and thus we now propose that this feature is added to the CERIF2000 standard. It is completely in line with the philosophy of CERIF and is analogous to the recursive relationship of {project} or {orgunit}, or the non-recursive relationship between one {project} and another or one {orgunit} and another.

3.2.2 CLRC-RAL

CLRC has decided that it requires a corporate data model to underpin the drive to make all its business processes electronically supported. Work started on this independently of CERIF but after a relatively short time the model proposed was observed to be close to CERIF and so they were compared formally. The result was the adoption of CERIF but with extensions for this CLRC purpose. As CLRC is an organisation for the purpose of R&D it is perhaps not surprising that a CRIS data model should form a suitable basis for a corporate data model. However, CERIF was aimed originally at recording R&D information and not at supporting the business of R&D.

The major extensions required in the CLRC datamodel are as follows:

- {project}: considerably more information including project plan, costs, milestones, deliverables;
- {person}: considerably more information including annual performance assessment which itself includes work objectives and their achievement and learning and development needs and their achievement. The record of past positions within the organisation can be recorded in CERIF, as can an employee's manager and senior managers. Records of travel need to be added, related to project. The authority of one person over another can be recorded in CERIF but not the financial authority of a person (authorising expenditure by project). Although the CV of a person (as recorded within CERIF) can record skills or competencies, it is not necessarily in a form suitable for processing within the business of an organisation.
- {orgunit}: CERIF does not provide for information on the mission or objectives of an orgunit, nor its terms of reference (e.g. for a committee). It does not provide for financial information of an orgunit (e.g. annual budget, invoices, orders) nor human resource aspects of an orgunit (how many staff-years of effort does it control).

Furthermore, the extension – a linking relation - to allow {result_publication-result_publication} noted above is required by CLRC. Current work is evaluating how much extension is required to {equipment} and {facilities} from the CERIF model to accommodate CLRC needs.

CLRC staff are still working on the details of the data model and expect to provide a full proposal for CERIF extensions for consideration by the CERIF Task Group of EuroCRIS (www.eurocris.org) in due course.

3.3 Precision and Formalisation

3.3.1 DataModel Corrections

Work while at the Technical University of Vienna by Andre Lopatenko has discovered errors in (CERIF2000), particularly in the EC-provided schemas driven from the extended entity-relation diagrams. A few inconsistencies were also discovered in the spreadsheet tables in the appendix of (CERIF2000). The current correct version of the datamodel is available within the documentation of the CERIF Task Group of EuroCRIS (www.eurocris.org/cerif).

3.3.2 Dictionaries, Thesauri and Ontologies

Given a formally correct datamodel (syntax) the next step to permit effective use of CRIS and effective data exchange or homogeneous access is converging the semantics (meaning). This requires agreed terms in dictionaries, thesauri or ontologies. Some work was done in this area using classification schemes (codes and meanings) and is documented (CERIF2000). However, many attributes were not subjected to rigorous content definition. Recent work by Andrei Lopatenko has provided a XML-encoded RDF description of CERIF which provides the basis for a definition of formal semantics, now being attempted using DAML + OIL. (W3C)

4 CERIF Future

4.1 Introduction

CERIF clearly still has much to offer the CRIS designer, or the systems engineer providing import / export from a legacy CRIS to other systems. It provides a formalised reliable model. It appears – with the formal definition of the dictionaries, thesauri and ontologies - to be complete for CRIS requirements. Furthermore, it is clearly extensible for other purposes including a corporate business data model.

4.2 Data Exchange

CERIF can be used for data exchange with data from CRIS A being converted from CRIS A format to CERIF, transmitted to CRIS B, received as CERIF and stored in the format of CRIS B ready for use by users of CRIS B. Although it can be shown theoretically that this is accomplished easily, demonstration of this capability is a target.

4.3 Data Access

However, CERIF can also be used for access – that is provision of a portal to all CRISs for an end-user either attached to a particular CRIS or free-standing. The portal allows query expression in one expressive language and translation of that query to the target CRISs. They export their results as CERIF to be integrated at the portal for the end-user. The end-user then receives an answer consisting of the union of the results from the different target CRISs in CERIF format, ready for storing in the user's local CRIS or independently. Such a system is subject to access rights, copyright, IPR and other restrictions. Such a portal system has yet to be constructed, but the ERGO pilot demonstrated feasibility.

Here CERIF intersects with the (W3C) concepts of the 'semantic web' and the 'web of trust', both areas of active research by the authors among others.

5 Conclusion

CERIF has demonstrated the basic soundness of the datamodel both in formal correctness and in its designed-in flexibility. This provides optimism for its success in the future. Of the proposed extensions few have required changes to CERIF. Some extensions (UiB) provide, arguably, efficiency but at the expense of program maintenance effort. Others (CLRC) are required in order to utilise CERIF in a much wider environment (corporate business processes and information systems) than originally intended (CRIS).

6 Acknowledgements

The excellent work of the CERIF2000 Group is there for all to see. The authors acknowledge the work of colleagues particularly Johanne Revheim at UiB and Stuart Robinson at CLRC.

7 References

CERIF2000 www.cordis.lu/cerif

ERGO www.cordis.lu/ergo

Jeffery,K; Lay,J; Miquel,J-F; Zardan,S; Naldi,F; Vannini-Parenti,I (1989) IDEAS: A System for International Data Exchange and Access for Science *Information Processing and Management* Volume 25 No 6 pp703-711, 1989.

Naldi, F; Jeffery, K; Bordogna, G; Lay, J; Vannini-Parenti, I A Distributed Architecture to Provide Uniform Access to Pre-Existing Independent, Heterogeneous Information Systems *RAL Report 92-003*

Official Journal (1991) Recommendation to the Member States to use the CERIF format In Official Journal of the European Communities, OJ L 189, 13th July 1991.

van Woensel, L (1988a) 'CERIF Manual' October 1988

van Woensel, L (1988b) Towards harmonisation of databases on research in progress – Final report of the European Working Group on Research Databases November 1988. Published by the Liaison Committee of Rectors' Conferences of Member States of the European Communities and Directorate General for Science, Research and Development of the Commission of the European Communities; financed by the Commission of the E.C., contract PSS*0058/B, compiled by Dr. L. Van Woensel.

W3C www.w3.org

8 Contact Information

Anne Asserson
Research Documentation Unit
University Library
University of Bergen
N-5020 Bergen
Norway

e-mail: anne.asserson@ub.uib.no

Treatment of Semantic Heterogeneity using Meta-Data Extraction and Query Translation

Robert Strötgen
Social Science Information Centre, Bonn

Summary

The project CARMEN[1] ("Content Analysis, Retrieval and Metadata: Effective Networking") aimed – among other goals – at improving the expansion of searches in bibliographic databases into Internet searches. We pursued a set of different approaches to the treatment of semantic heterogeneity (meta-data extraction, query translation using statistic relations and cross-concordances). This paper describes the concepts and implementation of these approaches and the evaluation of the impact for the retrieval result.

1 Treatment of Semantic Heterogeneity

Nowadays, users of information services are faced with highly decentralised, heterogeneous document sources with different kinds of subject indexing. Semantic heterogeneity occurs e.g. when resources using different systems for content description are searched by using a single query system. It is much harder to deal with this semantic heterogeneity than the technological one. Standardization efforts such as the Dublin Core Metadata Initiative (DC) are a useful precondition for comprehensive search processes, but they assume a hierarchical model of cooperation, accepted by all players.

Because of the diverse interests of the different partners, such a strict model can hardly be realised. Projects should consider even stronger differences in document creation, indexing and distribution with increasing „anarchic tendencies" (cf. Krause 1996, Krause/Marx 2000). To solve this problem, or at least to moderate it, we introduce a system consisting of an automatic meta-data generator and a couple of transformation modules between different document description languages. These special agents are able to map between different thesauri and classifications.

The first step in handling semantic heterogeneity should be the attempt to enrich the semantic information about documents, i.e. to fill up the gaps in the documents meta-data automatically. Section describes a set of cascading deductive and heuristic extraction rules, which were developed in the project CARMEN for the domain of Social Sciences.

The mapping between different terminologies can be done by using intellectual, statistical or neural network transfer modules. Intellectual transfers use cross-concordances between different classification schemes or thesauri. Section describes the creation, storage and handling of such transfers.

Statistical transfer modules can be used to supplement or replace cross-concordances. They allow a statistical crosswalk between two different thesauri or even between a thesaurus and the terms of automatically indexed documents. The algorithm is based on the analysis of co-occurrence of terms within two sets of comparable documents. The main principles of this approach are discussed in section .

1 Funded by the German Federal Ministry of Education and Research in the context of the programme "Global Info", FKZ 08SFC08 3.

We used intellectually and statistically created semantic relations between terms for query translation (cf. section) and evaluated the impact of this translation on the query result (cf. section).

2 Meta-Data Extraction

2.1 Approach

The goal of extracting meta-data during the gathering process is to enrich poorly indexed documents with meta-data, e.g. author, title, keywords or abstract – while the other methods described in the following section run during the retrieval process. This meta-data should be available for retrieval along with certain intellectually added meta-data, but with a lower weight.

The actual algorithms for meta-data extraction depend on file formats, domain properties, site properties and style properties (layout). No stable and domain independent approach is known so far. Until conventions for creating HTML documents change and become standardized towards a semantic web only temporary and limited solutions can be found.

Relevant documents from the Mathematics exist mostly in PostScript format. These documents (thesis papers) stored in an unstructured file format contain some meta-data of high relevance, which are marked only by layout information (e.g. font size) or special keywords. By using and modifying some tools as prescript from the New Zealand Digital Library, PostScript documents have been transformed and analysed. Abstract, keywords and classification terms can be extracted from these documents with good success.

Internet documents from the Social Sciences are mostly structured HTML files, but they use html features mainly for layout reasons, not as mark-up for content. Meta tags are infrequently used, and their syntax is often not correct. Different institutions use different ways of creating their Internet documents and a large number of documents contain no information about author or institution at all. This makes extraction of meta-data very difficult.

Nevertheless we developed a set of heuristics for identifying some meta-data in this heterogeneous set of documents. Because operating on (frequently incorrect) HTML files is not very comfortable, we transformed the documents into (corrected) XHTML and implemented our heuristics on these XML trees using XPath queries. (cf. Strötgen & Kokkelink 2001)

2.2 Evaluation

The test corpus for the Social Sciences contains 3.661 HTML documents collected from Web servers from different research and education institutions. The documents are of different types (e.g. university calendars, project descriptions, conference proceedings, bibliographies). We analysed these documents and used them as basis for the creation and improvement of the extraction heuristics.

From these documents 96% contain a correctly coded title; 17.7% of the rest contain an incorrectly marked title, the other documents contain no title at all. Only 25.5% contain keywords, all of them are marked properly. Not more than 21% contain a correctly coded abstract, 39.4% of the rest contain a differently marked abstract. This survey is the base for the evaluation of the meta-data extraction. Of course meta-data can only be extracted, if it is present in the document at all – we did try to implement automatic classification or automatic abstracting.

For the evaluation of the extracted meta-data we created a random sample containing every tenth document (360). We intellectually rated the relevance and correctness of the extracted data in four grades (accurate and complete; accurate in detail, but not complete or inaccurate parts; not accurate; not rateable).

We found that 80% of the extracted titles are of medium or high quality; almost 100% of the extracted keywords are of high quality; and about 90% of the extracted abstracts are of medium or high quality. (cf. Binder et al. 2002)

3 Query Translations using Semantic Relations

Intellectual and statistical semantic relations between terms or notations expand or modify the query during retrieval. The translation of the user's query, which was formulated for one of the document collections, leads to specific queries for each target document collection considering the different systems for content analysis.

3.1 Intellectual Transfer Relations (Cross-Concordances)

Intellectual transfers use cross-concordances between different controlled documentation languages like thesauri and classifications. These languages have a limited number of indexing terms or classes used to describe document subjects. A thesaurus is a natural language based documentation language with different kinds of vocabulary control and terminological control (i.e. synonym control and homonym control). Every thesaurus term is unique. Relations like equivalence, hierarchy and association are defined between the descriptors (cf. Burkart 1997). A classification is usually an artificial (alphabetical, numeric or alphanumerical) documentation language. A classification has classes or notions, which are systematically ordered with hierarchical relations, the classification structure. The class and its concept have a verbal class description (cf. Manecke 1997).

To build cross-concordances documentary and professional experts created semantic relations between thesaurus terms or classes with similar meanings.

Different kinds of inter-thesaurus or inter-classification relations are defined: "exact equivalence", "inexact equivalence", "narrower equivalence" and "broader equivalence". These relations can be annotated with weight information ("high", "medium", "low") reflecting the relevance of the relations for retrieval quality.

In the project CARMEN cross-concordances are provided between universal or special classifications and thesauri in the domains involved (mathematics, physics, social sciences). These cross-concordances allow safe transfers between different documentation languages and the document sets using them.

Problems may arise if there are insufficient resources (time, money, domain experts) to create and maintain such cross-concordances; furthermore not all documents – particularly Internet documents – are indexed with a controlled vocabulary. Therefore additional approaches like statistical transfers based on the analysis of co-occurrence of terms (see section) are necessary.

The software tool SIS-TMS[2] proved useable for creation of cross-concordances between different thesauri. CarmenX[3] has proved to be equally useful for creating relations between different classifications.

3.2 Statistical Transfer Relations

Quantitative statistical methods offer a general, automatic way to create transfer relations on the basis of actual bibliographic data. Co-occurrence analysis is one of those methods. It takes advantage of the fact that the content analysis from two different libraries on a single document held in both collections will represent the same semantic content in different content analysis systems. The terms from content analysis system A which occur together with terms from con-

2 http://www.ics.forth.gr/proj/isst/Systems/sis-tms.html

3 http://www.bibliothek.uni-regensburg.de/projects/carmen12/index.html.en

tent analysis system B can be computed. The assumption is that the terms from A have (nearly) the same semantics as the related ones from B, and thus the relation can be used as a transfer relation. The prerequisite for such computations is a parallel corpus, where each document is indexed by two different content analysis systems.

The classical parallel corpus is based on two different sets of documents, e.g. two different library catalogues. Each is indexed with a specific thesaurus/classification. To be able to create co-occurrence relations between the terms of these thesauri, the indexations of the documents have to be brought into relation. This is done by finding identical (or at least equivalent) documents in both catalogues. Considering print media, the problem of identity can be solved quite easy. An identical ISBN in combination with at least one identical Author should be a sufficient criterion for the identity of two documents. But the situation is worse, if the underlying data sets are not bibliographic ones, e.g. if text data should be combined with fact data or if Internet texts are considered.

In dealing with the World Wide Web we are concerned with many Web pages that are not indexed by a specific (given) thesaurus or classification system. The only terms we can rely on are the full-text terms supplied by a full-text indexing machine.

Taking into account that a user might start his search with controlled vocabulary obtained from a thesaurus, relevant Internet documents should be retrieved as well as well-indexed documents stored in a database. In order to facilitate this, we have to realize a transfer from classification terms, thesaurus terms respectively, to full-text terms, and vice versa. As long as we cannot fall back to any standards of classifying Internet documents, we have to use a weaker strategy of combining keywords and full-text terms. Note that intellectual indexing would result in enormous costs. This weaker strategy is simulating parallel corpora for supplying semantic relations between keywords and Internet full-text terms.

First of all, in order to provide a simulated parallel corpus, we have to simulate intellectual keyword indexing on the basis of a given thesaurus. Simulating intellectual indexing implies that a method is used that produces *vague* keyword-document-relationships, i.e. unlike intellectual indexing, where each assignment is (usually) un-weighted (weighted 1 respectively) simulated keyword-document-ties are weighted on a [0,1]-scale. This yields a situation as indicated in Figure 1: Unlike the situation in public databases (like the German social science literature database SOLIS), where we have exact assignments of keywords and documents, we produce vague keyword indexing as well as vague full-term indexing.

Parallel corpora simulation via vague keyword and full-text term assignments is described for the CARMEN test corpus. The CARMEN test corpus is a collection of about 6.000 social science Internet documents that have no keyword assignments.

For the assignment of controlled vocabulary (keywords) to non-classified Internet documents a given thesaurus is used, i.e., in the case of the CARMEN corpus, the thesaurus for Social Sciences (IZ 1999). As a basic method assigning thesaurus keywords to Internet documents we consider each single keyword of the thesaurus as a query that is "fired" against a full-text indexing and retrieval machine maintaining the CARMEN corpus. The full-text retrieval engine Fulcrum SearchServer 3.7 is used to provide ranking for values the documents in the corpus according to their similarity to the query vector. Each document in the result set that has a ranking value greater then a certain minimum threshold is then indexed by the keyword requested. The ranking values supplied by Fulcrum are considered the weights for the keyword assignments. This basic method has been improved to consider the relevance of a keyword for the document retrieved (cf. Hellweg et al. 2001).

Full-text terms are obtained by tokenising the full-text of an Internet document, eliminating stop words, and stemming the remaining terms using a Porter stemming algorithm for German. For weighting the terms the *inverse document frequency* (cf. Salton 1987) is used. The full-text is then indexed with full-text terms having a weight greater than a certain minimum threshold.

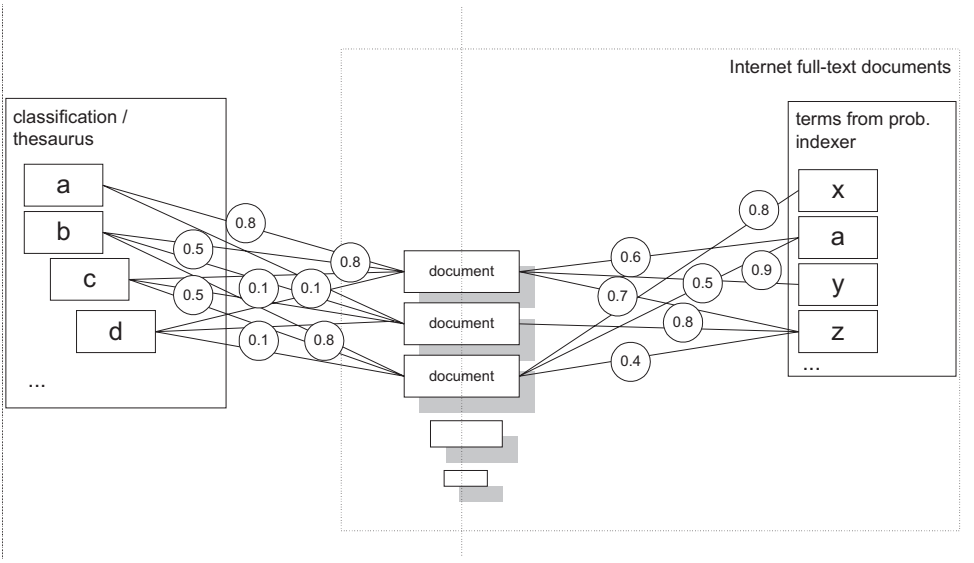


Figure 1: Parallel Corpus Simulation with vague keywords and full-text terms

In the context of the project ELVIRA, a tool for generating statistical correlation relations based on parallel corpora was implemented. JESTER (the Java Environment for Statistical Transformations) is a general workbench that allows for the interactive selection of parameters for optimising the transfer relation between a pair of classification systems. JESTER also employs a number of heuristics for the elimination of systematic errors, introduced by the simulation of an actual parallel corpus as described before.

In particular, the graphical representation of the term-document frequencies permits the eliminations of documents and/or terms from the following steps, based on their occurrence. In the case of a simulation of a parallel corpus, some documents got too many terms assigned. This happens, when the probabilistic search engine erroneously returns the same document on almost every query because some domain specific phrase appears verbatim.

3.3 Query Translation

Once the transfer relations have been realized, the question remains how to incorporate them into information retrieval systems. As a query term manipulating methodology they have to be placed somewhere between the user interface and the databases. Transfer relations, or – to be precise – transfer modules, become necessary, if data sets have to be combined, which are indexed by different content analysis systems. Usually those data sets reside in different databases. In a classical coordination layer the user query is simply distributed – unchanged – to the different databases and the results are combined to an overall result set. Most of the meta search engines in the WWW work this way.

But this procedure is not applicable to data with heterogeneous indexing systems. To send one and the same query to all databases would lead to a lot of zero results in the connected databases. This is due to the fact that e.g. a queried classification is available in only one database, but not in the others. At this point the transfer relations come into play. Through their ability to transform controlled vocabulary, they are able to adapt the users query to the different requirements of the

database. Therefore the query the user has issued will be transformed into different queries fitting the different content description systems (cf. Figure 2).

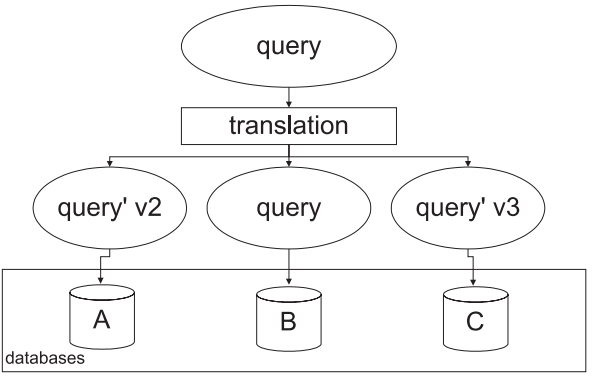


Figure 2: Query Translation Process

During the actual transformation the relevant part of the users query (e.g. the classification terms from system A) is separated from the other parts. The terms of this separated part act as the input for the different transformation modules. After the transformation, the resulting output forms the new, transformed query part. This new part consists of terms from system B, and is combined with the rest of the users query (e.g. author/title query) to form the new, transformed query. Afterwards the query is sent to the corresponding database.

This procedure follows the so-called "two-step" model (cf. Krause 2000) developed by the Social Science Information Centre (IZ) in the context of different projects.[4]

For CARMEN this approach was implemented in the project's architecture. The retrieval system HyRex[5] is part of a package developed at the University of Dortmund (cf. Fuhr et al. 2000). This search engine uses transfer services running remotely on servers at the Social Science Information Centre. Relevant parts of the complete query (e.g. author is no transferable query type) are sent to the transfer service as XIRQL (XML Information Retrieval Query Language) statements by http request and answered with a new transferred XIRQL statement that represents the transferred query (cf. Figure 3).

4 ELVIRA, CARMEN, ViBSoz and ETB (cf. Hellweg et al. 2001).
5 http://ls6-www.informatik.uni-dortmund.de/~goevert/HyREX.html

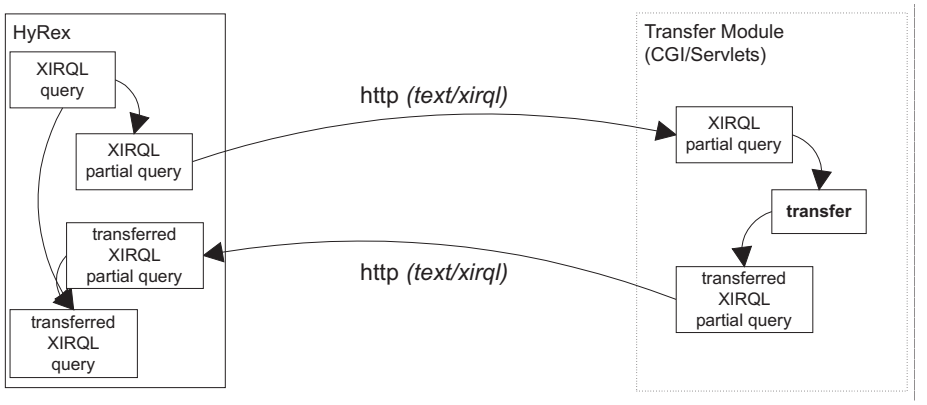


Figure 3: Query Transfer Architecture

3.4 Evaluation

In order to evaluate the impact of query translation with statistically created transfer relation we performed retrieval tests. We indexed about 10.000 HTML documents using Fulcrum Search Server (HyRex was not available for the retrieval tests in time). Using Fulcrum we did not make use of the weight information attached to semantic relations; this feature is implemented in HyRex, but the effect is lost for the test.

Our scenario consisted of a search starting from the bibliographic database SOLIS using the "Thesaurus Sozialwissenschaften" as documentation language for query formulation. This query was supposed to be expanded to Internet documents. For this purpose the query is translated from the controlled thesaurus term to free terms. In this special case no translation of one controlled language to another leads to a replacement of terms; the uncontrolled free terms as translation target allow the addition of semantically related terms.

We executed the search in two ways: We sent both the original SOLIS query and the translated (expanded) query to the retrieval engine and compared the results. We have not been able to perform representative tests but exemplary spot tests. For each of three domains from the Social Sciences (women studies, migration, sociology of industry) we carried out two searches. Two of them are described exemplarily:

One query used the SOLIS keyword "Dominanz" ("dominance") and returned 16 relevant documents from the test corpus. Using the query translation 9 additional terms were found: "Messen", "Mongolei", "Nichtregierungsorganisation", "Flugzeug", "Datenaustausch", "Kommunikationsraum", "Kommunikationstechnologie", "Medienpädagogik", "Wüste").

You have to be very careful interpreting the statistically created relations in a semantic way. In this example the new terms "Mongolei" ("Mongolia") and "Wüste" ("desert") result of documents describing an excursion of a women activist group to China with a stop in the Mongolian desert. One should not deduce a "dominance problem" in Mongolia or in desert regions from this relation, but in other cases you will find semantic relations not complying semantic equivalence but a problem field.

With this expanded query 14 new documents were found; 7 of them were relevant (50%, a gain of 44%). The precision of this search is 77%. In this case without too much noise a significant number of new documents could be reached with the query translation.

Another, less successful example: A query using the keyword “Leiharbeit” (“temporary employment”) returned 10 relevant documents. The query translation added 3 new terms: “Arbeitsphysiologie”, “Organisationsmodell”, “Risikoabschätzung” and produced a result of 10 new documents, but only 2 of them were relevant (20%, a gain of 20%). The precision of this search is 60%. In this example the translated query results in very little gain of new relevant documents but 80% noise.

Summerising all examples we can state that we always found new relevant documents using the translated query compared with the original query. The precision of the additionally found documents ranges between 13% and 55%. Without being already able to find systematic conditions we find rather successful and weaker query results.

4 Outlook

The modules for meta-data extraction proved to be satisfactory. Meta-data was extracted tolerably from Web documents. They have been integrated in the gathering system (“CARA”) and can be used for other projects. It seems promising to transfer the heuristics for HTML documents to other domains than the Social Sciences. Probably the weighting component will need some adjustment.

The modules and heuristics are in general functioning; some improvement is conceivable by tuning the heuristics. Because of the transient Web standards and the fast changes in Web style very high effort for maintenance seems necessary to keep the heuristics up to date. It seems questionable if this effort can be raised.

The query transfer modules using statistically created semantic relations proved able to improve the query result in retrieval test. New relevant documents were found using the transfer from a thesaurus to free terms for an Internet search, but some queries produce more noise then useful documents.

Some aspects remain unanswered and require more research and tests.

How can the document corpus, used for computing the semantic relations, be improved; e.g. what kind of documents create bad artefacts or which properties does a corpus need to be representive? Probably the statistical methods need some refinement.

In the project’s context intellectually created cross-concordances have been created and evaluated separately. The tested transfer modules can handle both kinds of semantic relations, and both should be compared directly. Also methods of combining both ways should be implemented and evaluated.

Of course the user interaction remains an important topic. By now the transfer process is a black box for the user. An user interface is needed that allows the user to understand and to influence this process and its parameters without handling incomprehensible numbers like statistical thresholds. The outcome of an interactive retrieval using transfer modules must be evaluated with real user tests.

An output of the project are services and software modules for query translation, which are offered to interested users. We already integrated them into existing services like “ViBSoz” (Virtuelle Fachbibliothek Sozialwissenschaften); other new services like “ETB” (European Schools Treasury Browser) and “Informationsverbund Bildung – Sozialwissenschaften – Psychologie” will follow.

5 References

Binder, G.; Marx, J.; Mutschke, P.; Strötgen, R.; Plümer, J.; Kokkelink, S. (2002): Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht Nr. 24).

Burkart, M. (1997): Thesaurus. In: Buder, M.; Rehfeld, W.; Seeger, T.; Strauch, D. (Eds.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Dokumentationsarbeit. München, p. 160 - 179.

Fuhr, N.; Großjohann, K.; Kokkelink, S. (2000): CAP7: Searching and Browsing in Distributed Document Collections. In: Borbinha, José; Baker, Thomas (Eds.): Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL 2000; Lisbon, Portugal, September 18-20, 2000; Proceedings. Berlin: Springer 2000. (Lecture notes in computer science ; Vol. 1923). p.364 - 367.

Hellweg, H. (2002): Einsatz von statistisch erstellten Transferbeziehungen zur Anfrage-Transformation in ELVIRA. In: Krause, Jürgen; Stempfhuber, Maximilian (Hrsg.). Integriertes Retrieval in heterogenen Daten. Text-Fakten-Integration am Beispiel des Verbandinformationssystems ELVIRA. (Forschungsberichte des IZ Sozialwissenschaften Band 4) (to appear 2002).

Hellweg, H.; Krause, J.; Mandl, T.; Marx, J.; Müller, M.; Mutschke, P.; Strötgen, R. (2001): Treatment of Semantic Heterogeneity in Information Retrieval. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht Nr. 23).

Krause, J. (1996): Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung; Schalenmodell. Bonn (IZ-Arbeitsbericht Nr. 6).

Krause, J. (2001): Virtual libraries, library content analysis, metadata and the remaining heterogeneity. In: ICADL 2000: Challenging to Knowledge Exploring for New Millennium: the Proceedings of the 3rd International Conference of Asian Digital Library and the 3rd Conference on Digital Libraries, Korea, December 6 - 8, 2000, Seoul, p. 209 - 214.

Krause, J.; Marx, J. (2000): Vocabulary Switching and Automatic Metadata Extraction or How to Get Useful Information from a Digital Library. In: Information Seeking, Searching and Querying in Digital Libraries. Proceedings of the First DELOS Net-work of Excellence Workshop. Zurich, Switzerland, December 11-12, 2000. Zurich. p. 133 - 134.

Manecke, H-J. (1997): Klassifikation. In: Buder, M.; Rehfeld, W.; Seeger, T.; Strauch, D. (Eds.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Dokumentationsarbeit. München, p. 141 - 159.

Salton, G. (1987): Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg – New York.

Strötgen, R.; Kokkelink, S. (2001): Metadatenextraktion aus Internetquellen: Heterogenitätsbehandlung im Projekt CARMEN. In: Schmidt, Ralph (Ed.): Information Research & Content Management: Orientierung, Ordnung und Organisation im Wissensmarkt; 23. Online-Tagung der DGI und 53. Jahrestagung der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V., DGI, Frankfurt am Main, 8. bis 10. Mai 2001; Proceedings. Frankfurt am Main: DGI 2001. (Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis; 4), p. 56 - 66.

6 Contact Information

Robert Strötgen
Informationszentrum Sozialwissenschaften
Lennéstr. 30
53113 Bonn
Germany

e-mail: stroetgen@bonn.iz-soz.de
www.gesis.org/iz

Proposals for a new flexible and extensible XML-model for exchange of research information

Jens Vindvad, Erlend Øverby
National Office for Research Documentation, Academic and Special Libraries, Oslo
Conduct AS, Oslo

Summary

In this paper a new flexible extensible XML-model for the exchange of research documentation is proposed, and a working XML-exchange model is described. The working model is limited to documentation produced by researchers. The ideas, XML-model and construction proposed and used in the working model are extensible, and can be expanded to the whole field of research documentation, and to other fields as well.

The paper is partly based on a report to be completed in May 2002. That report will provide a full description of the model, which is not possible in a short paper presented at the CRIS2002 Conference. The report aims to set out the groundwork and facts, to document the proposed new XML-exchange model, and to contribute to the further discussion in euroCRIS in respect of data exchange between different systems.

1 Introduction

To be able to share information between different systems, a well-defined protocol for information exchange must be in place. XML (Bray et al. 2000) has emerged as a new protocol used in information systems for exchanging information between different systems.

This introduces a new problem of how to specify the information structure in XML. Our proposal addresses this problem. When exchanging information between different systems, the system needs to know what type of information it receives, and the structure of this information, so that the system is able to transform the information into the specified system. Normally when information is exchanged XML, a DTD (ISO 8879:1986) or an XML-Schema (Biron & Paoli 2001) is specified to describing the structure of the exchanged information. However, creating an XML document from existing information that is valid with the XML-schema could be as difficult as transforming your information into a specified protocol.

One example of this situation is taken from Scotland (Woolman 2001) where a large-scale program is under way, designed to develop an XML schema for electronic clinical communications. Another example is reported from the chemical community (Murray-Rust et al. 2001) of an operational system for managing complex chemical content entirely by means of an XML-based markup language called “Chemical Markup Language (CML)” where an XML schema has been developed.

Two alternatives exist to describing the defined structure of information; the first is a DTD and the second is an XML-schema. Both these approaches currently have the disadvantages that in order to validate and check the structure of the information, a description of the whole structure and all its possibilities and constraints must be in existence. This makes the exchange model large and inflexible, which makes it harder to establish an efficient exchange of information between different systems.

Normally validation is sacrificed for well-formed structures. The disadvantage of well-formed structures is that they could include almost any element, and there is no control of what the ele-

ment names are and what their semantic meaning is. In our proposal we try to address and solve this problem by introducing the concept of *micro-schemas*.

2 Fundamental concepts

2.1 XML-exchange model

In many of the structured systems based on SGML (ISO 8879:1986) - and to some extent XML - the structure of the information is vital where the structure of the information is described in one DTD. And for the documents to be valid with the DTD - all possible information within the document/information has to be described in that DTD. This means all the information in all contexts has to be described in that context in the DTD. As a consequence of this complexity the DTDs are hard to understand and to use.

To solve this problem we have introduced a new way of looking at complex information structures - described in XML - using a much more flexible approach to the possible ways of describing information structures, and to ease the exchange of information between information systems.

In a recently published paper (McClelland et al. 2002), the exchange of metadata between digital libraries is discussed. The final conclusion is: *"This article outlines some of those challenges and underscores the point that not all the problems will be solved merely by adopting a common metadata element schema."*

This would seem to support the view that there is a need for a more flexible approach than which is currently used, e.g. by using a micro-schema.

2.2 Architectures

Since the information domain is known - Current Research Information Systems (CRIS) - the nature of the information is somewhat similar. Different systems describe the information in different ways - but since the information is in the same information domain, the information model is to some extent the same and is often based on CERIF2000 (Alexandraki et al. 1999).

The architecture forms the building blocks to describe the information models, used when specifying the exchange model.

To ease the exchange of research information, the information should be compliant to an exchange architecture that is flexible, but where all the information elements are well defined, and with a vocabulary agreed upon.

To be able to exchange information, the existing data model needs to be transformed into the exchange architecture model. Similarly, to receive information, the exchange architecture model must be transformed into one's own data model.

By using this idea of an architecture specified in micro-schemas, it will be easy to adopt and modify the model for other types of information, and perhaps this model could ease the exchange of information between other systems as well. The micro-schemas will be the building blocks of the information exchange structure.

2.3 Vocabularies

To make the exchange model work it is mandatory that the vocabulary be agreed upon, in order to achieve a unique and well-defined semantic meaning of the information element. This view is supported in a recently published article about metadata (Duval et al. 2002) where one of the conclusions is: *"For these opportunities to be realized, some convergence of encoding formats and commonly agreed semantics will be necessary."*

2.4 Micro-schema

Our idea is to address the specification of the smaller information elements, in order that the logic information can be transformed to these smaller information elements, with the system able to check the validity of these smaller information elements.

The idea is that the export and import system knows the structure of these information elements, and can then easily encode the information using these smaller specifications, without taking into account the overall information model. The information elements are then mapped to the specified data structure. Since the information is within the same information domain, the natural structure of the information should be easily identifiable, even if the overall information model is different.

The import system will then validate the smaller information models against the micro-schema, and transform the information into the structure of the import system. This ensures a flexible and open method of exchanging information between systems using XML as the information medium.

This proposal also specifies how to parse and validate information using these micro-schemas, and how to enhance the model by specifying new micro-schemas.

Addressing of the micro-schema

To be able to address the micro-schema used, namespaces are employed. At the given namespace URI, it is expected that a schema will be found, which will be parsed by the export/import system. If there is no such schema, the information will only be well formed, and it cannot be ensured that the exchange information conforms to the exchange model. Each micro-schema may address other micro-schemas by means of a new proposed syntax in the schemas by using their namespace URI. The system must then look up this schema and parse it to check the validity of these structures.

3 Step-by-step description of how to achieve the new XML-exchange model

To achieve the new XML data exchange model the following eight steps are proposed:

3.1 Establish an extensible framework

The working model is limited to information produced by researchers. Information produced by researchers can be described in the following framework:

3.1.1 Outputs

The collection of all types of information produced by the researchers is called output. Outputs are divided into four subgroups: - results, - communication, - documentation and - art.

3.1.2 Results

Results are taken to mean the results of the research produced by the researcher in person. Examples of results are: publications, patents and products.

3.1.3 Communication

By communication we want to label the forms of communication that researchers use in their work. Researchers often need to or wish to discuss their ideas and views. This form of communication is not a result of their work, but represents interesting and important steps in the process of producing results. Such communication has two main audiences: the professional community

and the general public. Examples of communication are: conference presentations, workshops, broadcasting and interviews in the press.

3.1.4 Documentation

A researcher has to carry out administrative tasks and produce documentation, which cannot be classified as results or forms of communication. This can be pure administration or high level of professional work. In connection with CRIS systems documentation is not always the most interesting part, but it is a necessary element to give a complete picture of the information produced by a researcher. Examples are: reports to funding institution, application for funding, documentation of a laboratory upset, administrative tasks of a research projects, and computer programs.

3.1.5 Art

Art is not a necessary output of a researcher's work but it can be. Art can be seen as a result in itself, a form of communication or type of documentation or all of these. Art needs and deserves a classification based on standards used and accepted in the art community. Examples of art are: works of art, exhibition and performance.

3.2 Establish an internal information structure

It is necessary to establish an internal information structure. The internal information structure will be used to define micro-schemas and to build information containers. This step also has to be seen in relation to the next step, wording and vocabulary, and the first step of framework. The internal information structure is the basis the new XML-model will rest on.

3.3 Propose wording and establish a vocabulary

Successful communication requires common wording and definitions. For an exchange data model to be established, a definition has to be agreed upon. In the report, suggestions for definitions of all used terms are given. A lot of the terms employed are commonly used words. Precise definitions are required, and for some terms a taxonomical approach is used, since this permits testing to see if a concept or term belongs to a particular definition.

Space does not permit definition of all terms given in this paper. The following is an example of how it is proposed to provide a definition of the term "publication".

3.3.1 Publications

Publication is a commonly used word, which in daily use does not have a precise definition. To establish a vocabulary and a namespace, we need the word "publication" and have to give it a precise and distinct definition. To do this we establish the following five tests, which must be complied with before we call an information unit a publication. The five tests involve: - *addressee*, - *copies*, - *location*, - *readability* and - *time*.

Addressee: For a work to be a publication the general public has to be the addressee of the publication. A publication cannot have an explicit addressee. If the publication is addressed to a limited group or single identity it is not a publication.

Copies: For a work to be a publication it must be available in several equal copies. It should not be possible to change history by changing or destroying the source of a publication or a few copies.

Location: For a work to be a publication it should in principle be possible to acquire a copy of the publication all over the world from any library that has access to the international library community network.

Readability: A publication should be readable without the use of technical equipment. Either the original publication must be readable without the use of technical equipment or a copy must be obtainable by transforming a publication into a format that permits reading without the use of technical equipment.

Time: For a work to be a publication the user should have the option to access and read the publication at any time the user wishes.

3.3.2 Test for addressee, copies, location, readability and time

The five tests which involve: - *addressee*, - *copies,* - *location*, - *readability* and – *time* are also used against the terms communication and documentation, but with other requirements to pass the test.

3.4 Propose a namespace

Based on the two previous steps, the internal information structure and wordings and vocabulary, a namespace is proposed. Establishing a namespace is an important step in establishing an exchange model. This allows others to make programs and style sheets using the data exchange model.

3.5 Define reusable microschemas using the internal information structure

The idea of a micro-schema is that it should only describe a very small piece of information, and only such information as is relevant to the specific description. Information that is not relevant to the specific context is described in another schema.

To be able to express the relevance and the connection between the micro-schemas we need to develop a standard method of enhancing the schema specification in order to address the valid elements in the specific context. Using namespaces, introducing the term „Allow-schema-namespaces“, will do this. The system then scans for possible new schema structures allowed at the specific point.

Each information model in the exchange model will be defined in its own schema definition. The system needs four different types of schema definitions: templates, context, phrases and inline.

3.6 Build information containers, which fits into the framework

Based on the defined micro-schemas, information containers are built. Such information containers must conform to the internal information structure and fit into the framework.

3.7 Testing

It is important that the concepts and a specific model are tested. Establishing test beds will make it easier for one community to exchange information with other communities. The creating of a working model should make it possible to prove the concepts. By testing the exchange between specific information systems, a check can be made to ensure that all information elements and the information structure are taken care of.

So far we have successfully imported real data form one of the main CRIS systems in Norway into the XML-exchange model. The work has demonstrated that properly structured information can easily be exported into the XML-exchange model. Tests have also been carried out using data from a library system. Our results have shown that certain minor information details are difficult to extract from a library system due to the cataloguing rules.

To carry out these tests, the data was transformed to the XML-exchange format with the aid of XSL (Adler et al. 2001) style sheets processed by an XSLT (Clark 1999) processor.

3.8 Extension rules

So far the model is not easily extensible. To make the model extensible the following rules are established:

- Rules for defining new micro-schemas.
- Rules for defining new information containers.

4 Operation of the XML-exchange model

When first using the exchange model it is necessary to set up a mapping from each system's information model onto the exchange system's architecture. The vocabulary specified in the exchange model is intended to give guidance in this mapping.

The mapping is best described using XSLT, which transforms your XML-data model into the XML-exchange data model.

At the conference a live demonstration of this model will be given.

5 Final section

Conclusion and advise for further work

A more flexible approach is needed to the exchange of data between different data systems. To solve this need, the concept of micro-schema is introduced.

A new flexible and extensible XML-model for exchange of research information is proposed, using micro-schema. The new XML-model has been tested against existing CRIS-systems, and data has been successfully imported into the model.

The model has also been tested with success against ordinary library catalogue data.

The new model needs more testing against other systems. Furthermore, the model requires further examination and discussion. If the ideas and direction outlined in the proposal are accepted, a way forward for further development and change has to be agreed upon.

6 References

Adler, S.; Berglund, A.; Caruso, J.; Deach, S.; Graham, T.; Grosso, P.; Gutentag, E.; Milowski, A.; Parnell, S.; Richman, J.; Zilles, S. (2001): *Extensible Stylesheet Language (XSL) Version 1.0.* W3C. http://www.w3c.org/TR/xsl

Alexandraki, M. et al. (1999): *CERIF 2000 Guidelines Final Report of the CERIF Revision Working Group.* DG XIII-D.4, European Commission. http://www.cordis.lu/cerif

Biron, P.V.; Malhotra, A. (Eds.) (2001): *XML Schema Part 2: Datatypes.* W3C. http://www.w3c.org/TR/xmlschema-2

Bray, T.; Paoli, J.; Sperberg-McQueen, C.M.; Maler, J. (Eds.) (2000): *Extensible Markup Language (XML) 1.0.* 2nd edition. W3C. http://www.w3c.org/TR/2000/REC-xml-20001006

Clark, J. (Ed.) (1999): *XSL Transformations (XSLT) Ver. 1.0.* W3C. http://www.w3c.org/TR/xslt

Duval, E.; Hodgins, W.; Sutton, S; Weibel, S. (2002): Metadata Principles and Practicalities. In: *D-lib magazine,* Vol. 8, No. 4

ISO 8879:1986. (1986): *Information processing – Text and office systems – Standard Generalized Markup Language (SGML)*

McClelland, M.; McArthur, D.; Giersch, S.; Geisler, G. (2002): Challenges for Service Providers When Importing metadata in Digital Libraries. In: *D-Lib Magazine*, Vol. 8, No. 4

Murray-Rust, P.; Rzepa, H.S.; Wright, M. (2001): Development of chemical markup language (CML) as a system for handling complex chemical content. In: *New journal of chemistry*, Vol. 25, No. 4, p. 618-634
Woolman, P.S. (2001): XML for electronic clinical communications in Scotland. In: *International journal of medical informatics,* Vol. 64, No. 2-3, p. 379-383.

7 Contact information

Jens Vindvad
Riksbibliotektjenesten
Tel. +47 23 11 89 00
Fax. + 47 23 11 89 01
e-mail: jens.vinvad@rbt.no
www.rbt.no

Postal Address:
P.O.B 8046 Dep
N-0030 OSLO
Norway

Visiting Address:
Kronprinsensgt. 9
Oslo, Norway

Erlend Øverby
Conduct AS
Tel. + 47 90 12 96 42
Fax. + 47 22 33 60 24
e-mail: erlend.overby@conduct.no
www.conduct.no

Postal Address:

P.O.B. 805 Sentrum
N-0104 OSLO
Norway

Visiting Address:

Biskop Gunnerus gate 2
Oslo, Norway

CERIF - Information Retrieval of Research Information in a Distributed Heterogeneous Environment

Andrei Lopatenko, Anne Asserson, Keith G Jeffery
UM, UiB, CLRC

Summary

User demands to have access to complete and actual information about research may require integration of data from different CRISs. CRISs are rarely homogenous systems and problems of CRISs integration must be addressed from technological point of view. Implementation of CRIS providing access to heterogeneous data distributed among a number of CRISs is described. A few technologies – distributed databases, web services, semantic web are used for distributed CRIS to address different user requirements. Distributed databases serve to implement very efficient integration of homogenous systems, web services - to provide open access to research information, semantic web – to solve problems of integration semantically and structurally heterogeneous data sources and provide intelligent data retrieval interfaces. The problems of data completeness in distributed systems are addressed and CRIS-adequate solution for data completeness is suggested.

1 INTRODUCTION

One of the challenges of CRIS development is to provide access to research data which are scattered on the web pages, stored in different research information systems. The informational needs of a researcher or policy-maker are very seldom limited to information from one CRIS, which usually represents research from a particular region or sector of science. There is a strong need to integrate information from different sources and to provide access to all information to users, enabling them to utilize a wide range of sources. One of the problems of such system development is heterogeneity of research information systems. The CERIF Task Group develops the CERIF system to solve problems of providing transparent access to disparate research sources.

The paper is organized as follows. Section 2 outlines requirement for research data integration and describes results already achieved in this area. Section 3 describes which approaches are being developed by CERIF Task Group. Section 4 described user demands for quality of data and suggests solutions for distributed data queries to achieve data completeness. Section 5 concludes.

2 REQUIREMENTS FOR RESEARCH DATA INTEGRATION

Typical needs of research information consumers are not limited to information from one research information system. The information about research in the same area, about different stages of one research, about different research relevant information are scattered among different information systems. Getting the information requires knowledge of where which information is stored and efforts to visit all systems, learn them and use them to find required information. Of course, when the systems are not integrated, information from one system cannot be reused in another or by the user system automatically. Furthermore, the information consumer

must know all differences in description of research information, vocabularies used - which makes retrieving of information often an intractable task.

Full-text search engines, like Google, AskJeeves, widely used to search information on the web:
- do not go inside many information systems
- do not provide attributed search which is important for policy-makers and researchers search
- do not filter information by its quality, actuality
- do not provide facilities to use terms, thesauri specific for science and which are very important for research informational retrieval
- do not manage the semantics of query and information searched

The strong requirement to have the ability to search research information from all European sources led to the development of the prototype project ERGO – European Research Gateways Online. This project, organized and sponsored by European Commission, aimed to collect metadata information about research projects from all European countries into one central database and to provide information retrieval access to that database to users. The ERGO project, in which the euroCRIS group strongly participated, created a data exchange approach to collecting information based on SGML technologies.

The need to integrate distributed research data and provide unified access was recognized in the Santa Fe Convention (VanDeSompel, 1999). This led to the development of the Open Archives Initiative – an approach which consists of a metadata set and protocol to integrate distributed metadata on the web, and which is already used by several communities. The need to provide access to data from distributed information systems of scientific libraries led to the creation a number of distributed systems such as MathNet (MathNet) PrePrint (PrePrints) network, Networked Digital Library of Thesis and Dissertation mentioned(NDLTD, NDLTD-01, NDLTD-01), Research Papers in Economics (RePEC), Clinical Medicine NetPrints (NetPrints) . A number of other examples show a strong need for distribution information retrieval in science. The requirement for informational retrieval from multiple data sources demands resolution of semantic, structural and system heterogeneities.

A lot of research and development were done in development of distributed systems for information retrieval but not all maybe applied to CRISs.

Requirements to distributed solution for CRIS:

1. **Easiness to implement**. The technology used should be well-known, the solution must be implemented with minimum demand for financial, human and time resources. Due to this reasons it was decided to implement free toolkit, which maybe installed and configured. The toolkit is implemented on Java and well-documented.
2. **Flexibility**. CRISs usually are not static systems, they are very changeable. The system should accommodate changes. The solution should be easily configured for new CRISs. A few solutions for distributed CRIS are suggested. Each solutions is address different types of requirements for distributed CRIS and different type of data sources. We tried to develop software architecture making configuration of the system easy.
3. **Support of open standards**. Very probably many CRISs will be embedded into enterprise or university infrastructures, research information should be possible to integrate into established data flows. Also reuse of team experience and software tools makes CRIS development and maintains cheaper. Now some of the most popular standards or activities related to data access in software development are JDBC, XML, RDF, SOAP, Java Beans (+EJB), CORBA. Current version of distributed CERIF may use JDBC, XML, RDF, SOAP.
4. **Effectiveness**. Seeking of research information is data and reasoning intensive operations. Users expect getting data on their request very fast. A solution for fast data search based on distributed database technologies is suggested.

5. Ability to solve problems of **semantic and structural interoperability**. The CRISs are very heterogeneous. Despite their heterogeneity uniform access to data should be provided. Semantic Web solution for CRIS address a wide range of semantic diversity problems. The meaning of data may be included into the systems and used for search

3 CERIF DISTRIBUTED INFORMATIONAL RETRIEVAL SOLUTION

The CERIF Task Group decided to develop different solutions, because no one solution can be enough to solve all problems of transparent access to disparate data sources. Different requirements of a distributed solution, and the technology used, necessitate different approaches. Different approaches also are diverse in efficiency, compatibility with legacy technologies, ability to solve problems and so real systems might require use of all of them.

The new solutions being explored and developed are in addition to existing solutions utilising CERIF such as:

- simple portal pointing to CRISs (e.g. DRIS);
- central catalogue (possibly replicated-distributed) and contact information for individual CRISs e.g. ERGO Pilot
- central catalogue and automated retrieval from CRISs (ERGO 2++ proposal);
- central catalogue and advanced knowledge-assisted retrieval from CRISs (ERGO3 proposal)

To nurture the CRIS community, help CRIS developers and finally give better services to researchers and others, the principal policy of the CERIF Task Group is to provide free source code for solutions and develop them based on open standards and technologies. The source codes and documentation with demonstration are available at (CERIF TG).

3.1 Distributed Database Approach (CERIF-DD)

The aim of this approach is to provide very efficient search of multiple databases, when semantic and structural heterogeneity is minimal, the systems use common thesauri, or thesauri mappings are provided. This approach does not solve automatically semantic or structural heterogeneity problems, but is characterized by efficiency and is easy to develop and deploy. The most likely application of this approach is unification of research databases belonging to different branches of one organization.

Each database must export a set of views conforming to the requirements of CERIF (signature of view and semantics of attributes). JDBC access to database views for select operations must be provided. For each network one central server repository exists which registers all databases and provides a search over all databases to the end user. When views are exported for a database, the database must be registered at the central server of network, with description of source of data, quality, actuality, sector of research and access information. After the database is registered it will be used by the central search facility to answer queries of users. The data source should be registered at the central server through SOAP web service interface just sending RDF description of data source.

The set of view definition for CERIF-DD is precisely defined (CERIF 4). The signature of view and semantics can be clearly understood from CERIF-DD view definitions and they are described at CERIF TG pages. Currently the search for mostly used entities like persons, projects, organization units, publications and links between them maybe performed in CERIF-DD.

Data source should be described in RDF according CERIF source description DAML Schema (CERIF 3).

This solution makes possible to solve manually very simple semantic and structural heterogeneity problems, just adequate definition of views for data source must be provided which maps

local tables and columns into right views and view attributes. Investigation of some custom CERIF independent university CRIS has shown that such mapping is possible in most cases what says about high CERIF compatibility with CRIS (Lopatenko 2001.1, Lopatenko 2001.2).

Also this solution makes possible to solve problems of usage of different vocabularies in different CRISs. When the local data source use different from CRISs network vocabularies for describing data only table definition of vocabulary mapping should be provided and this mapping should be included into view definition. So if such mapping exists, then search operation will find adequate data even if they described by different vocabulary.

Another advantage of this solutions is due to relational access to data and use of JDBC tools for data analysis, warehousing, decision making based on this technologies could be used.

Disadvantages: possible security problems with direct access to database (must be guarded), high demand for manual configuration (view definitions, vocabulary mapping look-up tables), solving only simple semantic interoperability problems, incompatibility with web infrastructure

3.2 Semantic Web Approach (CERIF-SW)

The Semantic Web (SemWeb, TBL 2001) approach aims to solve problems of distributed information retrieval when:

- structure and semantic are different in different data sources and search must take into account differences of data sources
- sophisticated informational retrieval operations are demanded
- the schema of existing and new data sources can be changed
- already published Semantic Web data must be used
- direct access to database cannot be provided
- compatibility with other SW based architectures is important

The Semantic Web approach is based on CERIF as the core ontology (CERIF 1) to describe the meaning of research data. CERIF is the base vocabulary which provides a common set of terms and which must be understood by any system conforming to CERIF-SW approach.

One of main features of Semantic Web solutions is ability to easy integrate very heterogonous data sources – digital libraries, databases, repositories, legacy sources. To add new CERIF based CRIS to network it would be enough to install CERIF-SW application, configure it to get data from local database and publish on the web.

When systems with different structure and meaning of data from CERIF must be integrated into the network, the ontology of systems must be described in CERIF terms if that is possible. It allows integration also over search of not-CERIF but CERIF-compatible data sources (Dublin Core, MathNet, etc). Each data source must publish its ontology on the web and express data in RDF according to its ontology. The ontology must conform DAML(DAML) format and to make data searchable by CERIF queries, must explain local terms in CERIF terms like, for example, IST project is a Project, IST project is part of EU programme, financed by CORDIS, done in Information Technologies.

To make data accessible for queries they must be transformed into RDF (use of CERIF toolkit is possible). The RDF data may be published on the web to directly access through http (for harvesting), may be accessed through web services (see next, very integrated with CERIF-WS approach), or may be published on RDF Networked Query Facility – a Network Query Engine for RDF data developed for CERIF, based on HP Jena toolkit. Then data source should register its data and ontology at search servers, which provide transparent access to data to end users.

Another one of main features of Semantic Web solutions of CERIF is ability to use ontology driven **intelligent information retrieval.**

The same information can be represented in information system in different ways, described from different point of view due to diversity of data sources, policy-restrictions, or even the posi-

tion of the author / compiler. The views of the same information can be very different for different categories of information consumers. The ontology-driven information retrieval is intended to

1. solve problems of discrepancies between the viewpoints of information consumers and description of information in the system;
2. provide more powerful and sophisticated retrieval facilities, allowing to information consumers to utilize domain knowledge; make it possible for information consumers to investigate in which terms information is described in the system, what is the meaning of
3. those terms and if it is needed using this knowledge to create new more focused queries

Example

The following example shows how ontological descriptions of the terms of an information system can be used to find out relations between those terms to create queries which satisfies user needs (completeness and relevance of returned information).

This ontology is an example ontology, not real one, some facts maybe wrong from political point of view or some types of ontology design. It is just a demonstration.

The demonstration is a database of European projects, programme, organizations. It stores information about objects - such as programmes, projects - each objects is described as belonging to some of the classes.

Base ontology

There are two types of classes - defined and primitive.

If it is said that the FundingOrganization is an organization, which finances some activities and it is also said that class FundingOrganization is primitive class then it means each Funding Organization finances some activities, but if we know that X is an organization and X finances activities it is not enough to say that X is a FundingOrganization

But if it was said that FundingOrganization is a defined class then each object which is known as organization and which also is known to be financing some activities, it would become FundingOrganization in information system

Table 1. Ontology of application for searching projects

Classes	Definition
Activity	
Programme	
EUProgramme	Programme is **finaced-by** EU or by any EU-country
ISTProgramme	of EU Programmes
Location	Any location can be a **part-of** another location
City	
Continent	
Europe	
Country	
EuropeanCountry	A country is a **part-of** continent Europe
EUCountry	A country is a **part-of** EU
Union	Geopolitical or economical union of countries
EU	of unions, but which is **a part** of Europe
Organization	
FundingOrganization	organization which finances some Activities or Projects
EUFundingOrganization	Any funding organization which is **situated-in** EUCountry

EuropeanFundingOrganization	Funding organization is **situated-in** Europe
Proiect	
FinancedProject	Project which is **financed-by of FundingOganization**
EUFinancedProject	Project is **financed-by** EUFundingOrganization
EuropeanFinancedProject	Project is **financed-by** EuropeanFundingOrganization
ISTProject	Project is a **part-of** IST Programme
Properties	
financed-by	A financed-by Y reverse relation finances When A financed-by Y then Y finances A
Finances	
part-of	transitive relation. X is a part of Y, and Y is a part of Z, then X is a part of Z
situated-in	Geographical inclusion. transitive relation
Axioms	**Descriptions**
Project is a **part-of** any of EUProgrammes is **financed-by** EU	it is known that a project is a part of EUProgramme then we are sure that this project is financed by one of EUFundingOrganizations

After ontology verification and classifying its terms, the verifier asserted new statements about relations between classes. Some of them:

Table 2. Automatically inferred conclusions from ontology

Statement	Proof
ISTProject is an EUProject	1. fact: ISTProject is a part of ISTProgramme, 2. fact: ISTProgramme is a EUProgramme 3. statement: ISTProject is a part of EUProgramme 4. 3 + Axiom -> ISTProject is financed by EU 5. 4 + definition of EUFinancedProject -> IST is an EUFinancedProject
EUFinancedProject is an EuropeanFinancedPtroject	1. EUFinancedProject is financed by FundingOrganization which situated in EU 2. EU situated in Europe 3. 1 + 2 + transitiveness of situated-in EUFinancedProject is financed by organization which is situated in Europe 4. 3 + definition of EuropeanFinancedProject -> EUFinancedProject is an EuropeanFinancedProject

Such reasoning performed by Description Logic can

1. increase knowledge of the data of users of the information system, helping them understand more precisely the data stored in the system and their own needs (user gets information that IST project is EU project, despite it was not directly asserted in the system - this can help a user in search of IST projects)
2. provide mechanism for implementing more correct query facility which utilizes new knowledge to find more complete answers on users' requests - eg. on request of IST projects, the databases of EU projects will also be searched
3. provide facility to describe results of answer to user - eg. why IST project was returned on EU projects request

As reasoning facility CERIF-SW uses FaCT system (FaCT) – Description logic classifier developed in Manchester University. As user friendly tool to develop CERIF ontology and other ontologies for CERIF solutions, ontology editor OilED(OilEd), developed in Manchester University was used.

So, main advantages of Semantic Web solution for CERIF
- ability automatically deal with heterogeneous data sources
- ontology-driven intelligent information retrieval
- compatibility with a set of emerging projects, activities
- ability to implement knowledge management solutions
- no security problems of direct access to database
- maybe build over any, not only database system allowing to utilize legacy systems, knowledge management systems, workflow support systems and others

Main disadvantages of CERIF-SW
- very inefficient now, search operation on database over tens megabytes and requiring transfer of a lot of data may not satisfy fast access requirement
- effective use of technology requires from developers knowledge of Semantic Web, ontologies and some tools – additional resources needed to implement
- warehouse, analysis and other reporting tools developed for relational databases can not use data from SW directly.

3.3 Web Services Approach (CERIF-WS)

The Web Services (W3 Web Services, IBM Web Services, Glass-2000) approach aims to solve problems of distributed informational retrieval
- important to make research information sources compatible with emerging standards of corporate networking and put research data into enterprise information flows
- demands to provide Web Services to research data, which are becoming popular for corporate internetworking
- direct database access is impossible for security or other reasons
- XML/SGML is already used in the organization and any new solution must be compatible with those already developed. Also experience of the team is a great asset and new solution must be based on such experience
- to provide efficient transport level protocol for CERIF-SW
- searched CRIS are CERIF-compatible (at intensional level) but have very different database schemas

CERIF-WS consists of
- an XML Schema based on the CERIF ontology;
- a SOAP (Simple Object Access Protocol) implementation of wrappers to each data source which implements functions to search and retrieve data;
- the CERIF-WS search facility, which searches registered SOAP services and provides access to research data from all systems to the user;

CRISs with a schema different from CERIF can easily participate in the CERIF network by publishing (through a wrapper) their data in CERIF XML. CERIF-WS has been built in a such way that it is very easy for developers to create or change the XML encoding of their data in the wrapper. Thus the technique allows CRISs with non-CERIF encoding to be accessed as if they were CERIF-encoded at the expense of building and maintaining the wrapper.

The main advantages of Web Services are
1. implementation of WS layer for CRIS maybe very easy. The tools to support WS for a lot of languages, platform are already available.

2. it is possible that in near future Web Services will dominate as a standard for interoperability of systems in business applications (Gartner 2002) – research information maybe access by other business applications
3. Web Services solution for CERIF may utilize Semantic Web for sophisticated queries, integration of heterogeneous data sources

Disadvantages of Web Services for CERIF
1. now solutions based on them are not very efficient comparing to distributed data search
2. resources to teach team use Web Services are needed
3. Web Services are not yet very mature, some tools are not efficient or have a lot bugs
4. current CERIF implementation of Web Services allows to use only very simple queries, hard to implement sophisticated queries which requires to investigate relations between objects

4 Data Quality

It is expected to have high demand for quality of scientific data in some CRISs. We believe that main parameters of quality of scientific data are completeness, actuality and correctness. Completeness of data in CRIS is presence of data about all entities which are subject of given CRIS. For example, CRIS about European projects since 1991 in chemical research in Norway will be complete, if it contains data about all European projects in chemical research since 1991 in Norway.

Actuality of data in CRIS is presence of newest information about CRIS subject. For example, CRIS about research projects in university will store actual data, if data in CRIS will be updated on new project developments immediately.

Usually correctness, actuality and completeness of data in given CRIS are results of administrative efforts for data harvesting, analysis, input.

But in distributed case, when data on request are returned from a few different independent information system, for quality-critical applications reasoning procedures to judge quality of data are required.

A few approaches to judge about quality of answers in distributed systems or to get complete answers for some queries are investigated (Levy 96, Motro 89, Duschka 97, Enzioni 94, Minok 99). It was decided to accommodate in CERIF-2000 distributed architecture (Duschka 96) because it satisfies information needs of CRIS systems, well-described and feasible to implement. (Minok 99), for example, approach requires existence of universal relation what is not a case of CERIF-based CRIS.

To make possible reasoning if query answers are complete metadata about completeness (local completeness) of sourcers must be provided according to CERIF registry metadata schema (CERIF 3).

The actuality and correctness of data are not yet addressed in distributed CERIF project

5 CONCLUSION

CERIF has demonstrated the basic soundness of the datamodel both in formal correctness and in its designed-in flexibility, as well as compatibility with developed and emerging CRISs. The authors' investigation of formats for scientific data shows high compatibility of CERIF with new systems and networks especially in the various W3C (World Wide Web Consortium) development groups. Extensions and solutions being built for CERIF by the CERIF Task Group provide implementations of importance for the research community information services, like distributed information retrieval and semantic informational retrieval.

6 References

CERIF TG
http://www.eurocris.org/cerif
CERIF 1 DAML ontology for research information
http://www.eurocris.org/cerif.daml
CERIF 2 XML Schema for research information
http://www.eurocris.org/cerif/cerif.xsd
CERIF 3 DAML Schema for data source metadata
http://www.eurocris.org/cerif-sources.daml
CERIF 4 Set of view definitions for distributed database solution
http://www.eurocris.org/cerif/cerif-dd.sql
DAML
http://www.daml.org
FaCT
http://www.cs.man.ac.uk/~horrocks/FaCT
Gartner 2002 Gartner Says Web Services Will Dominate Deployment of New Application Solutions for Fortune 2000 Companies by 2004-Reports Examine The Future of Web Services And Vendors Driving The Industry, Jan. 2002
http://industry.java.sun.com/javanews/stories/story2/0,1072,41782,00.html
Duschka 97 Duschka O., Query Optimization Using Local Completeness, In Proc. of 14th AAAI National Conference on Artificial Intelligence, AAAI-97., July 1997
Enzioni 94 Oren Enzioni, K. Golden, D. Weld, Tractable closed world reasoning with updates, In. Proc. 4 Knowledge Representation
Glass-2000 Glass G.; The Web services (r)evolution, Part 1,
http://www-106.ibm.com/developerworks/webservices/library/ws-peer1.html
IBM Web Services
http://www-106.ibm.com/developerworks/webservices/
Levy 96 Levy A., Obtaining Complete Answers from Incomplete Databases Proceedings of the 22nd VLDB Conference, Bombay, India. 1996
Lopatenko 2001.1 Comparison of CERIF-2000 and Salzburg University research information system schemas,
http://derpi.tuwien.ac.at/~andrei/documents/SzbCERIF.htm
Lopatenko 2001.2 Comparison of CERIF-2000 and AURIS (AUstrian Research Information System) schemas,
http://derpi.tuwien.ac.at/~andrei/documents/AURIS-CERIF.htm
MathNet
http://www.math-net.de
Minok 99 Minock M; Rusinkiewicz M.; Perry B; The Identification of Missing Information Resources by using the Query Difference Operator, Proc. of the 4th International Conference on Cooperative Information Systems
Motro 89 Motro M.; Integrity = Validity + Completeness. ACM Transaction on Database Systems, (4): 480-502 (1989)
NetPrints
http://clinmed.netprints.org/
NDLTD
http://www.ndltd.org/
NDLTD-01 Suleman H.; Atkins A.; Goncalves M., France R.; Fox E.; Chachra V.; Crowder M.; Young J. Networked Digital Library of Theses and Dissertations, Bridging the Gaps for Global Access - Part 1: Mission and Progress, D-Lib Magazine, Vol. &, Num. 9, Sep. 2001
NDLTD-01.1 Suleman H.; Atkins A.; Goncalves M., France R.; Fox E.; Chachra V.; Crowder M.; Young J. Networked Digital Library of Theses and Dissertations, Bridging the Gaps for Global Access - Part 2: Services and Research, D-Lib Magazine, Vol. &, Num. 9, Sep. 2001
OilEd
http://oiled.man.ac.uk/

Preprints
http://mathnet.preprints.org
RePEc
http://www.repec.org/
SemWeb
http://www.w3.org/2001/sw/
TBL 2001 Berners-Lee T., Hendler J.; Connoly D.; Swick R.; The Semantic Web, Scientific American, May 2001
Van de Sompel H.; Lagoze C.; The Santa Fe Convention of the Open Archives Initiative, D-Lib Magazine., Feb 2000, Vol. 6., Num. 2
W3 Web Services
http://www.w3.org/2002/ws

7 Contact Information

Andrei Lopatenko
The University of Manchester
Oxford Rd., Kilburn Building, r.2.112
Manchester M13 9PL
UK

e-mail: alopatenko@cs.man.ac.uk

Effectiveness of tagging laboratory data using Dublin Core in an electronic scientific notebook

Laura M. Bartolo[1], Cathy S. Lowe[2], Austin C. Melton[3,4], Monica Strah[5], Louis Feng[3], Christopher J. Woolverton[5]

[1]College of Arts & Sciences, [2] School of Library and Information Sciences, [3]Department of Computer Science, [4]Department of Mathematics, [5]Department of Biological Sciences, Kent State University, Kent, Ohio, USA

Summary

As a form of grey literature, scientific laboratory notebooks are intended to meet two broad functions: to record daily in-house activities as well as to manage research results. A major goal of this scientific electronic notebook project is to provide high quality resource discovery and retrieval capabilities for primary data objects produced in a multidisciplinary, biotechnology research laboratory study. This paper discusses a prototype modified relational database that incorporates Dublin Core metadata to organize and describe the laboratory data early in the scientific process. The study investigates the effectiveness of this approach to support daily in-house tasks as well as to capture, integrate, and exchange research results.

1 Introduction

Interdisciplinary scientific research efforts in academic, industrial and public settings need novel technologies to organize, access, and communicate research results, especially for a mobile society. This paper reports on one stage of a multi-stage project to construct an electronic scientific notebook for recording, storing, and manipulating multidisciplinary and multi-institutional scientific information from raw data to finished research papers. Long-term goals of this project include: 1) learning how to organize and store biotechnology information in formats which will encourage multidisciplinary use of the information; 2) applying the organizing knowledge gained and tools developed in storing biotechnology information to the storage of other scientific information; 3) developing an environment in which scientific information from different disciplines can be made more easily accessible by and meaningful to multidisciplinary research teams; and 4) constructing electronic scientific notebooks for the storage, retrieval, and dissemination of multidisciplinary scientific information.

Grey literature, such as data generated within scientific laboratories, has been recognized as an important area for innovation, new knowledge, and industrial enterprise (Jeffery 2000). In order to provide electronic access to laboratory data and to replace the standard paper notebook which scientists currently use in their laboratories, an electronic notebook needs to be able to perform all the functions of paper notebooks as well as facilitate greater operational flexibility. This paper presents a practical application of the software architecture utilizing laboratory data from an ongoing multidisciplinary, multi-organizational research program. The discovery process requires the recording of ideas, the identification of individual efforts, data acquisition, data analysis and presentation of data in scientific, lay and summary outputs (Buckland 1997). The investigation presented here used the traditional laboratory notebooks of the principal investigator to demonstrate the utility of the electronic architecture and Dublin Core (DC) in a scientific laboratory setting. Long term goals of this research include interfacing scientific notebooks with large data-

bases, facilitating the exchange of data in scientific notebooks among researchers, and making scientific notebooks central tools for presentations of results gained from laboratory research.

2 Interdisciplinary research and user needs

Scientific discovery can result in various final products such as published manuscripts, monographs, and patents. Our primary goal has been to create a software architecture that facilitates the discovery process across institutions and between investigators by creating an environment that mirrors the traditional laboratory notebook. Expanded features seek to enhance user access, data sharing and data mining while reducing or eliminating data integrity issues and redundant functions. A second goal of the project is focused on tagging data utilizing Dublin Core metadata so as to link individual data, analogous data types and non-chronological data entries across users and institutions. Tagging data by such elements as date, format, creator, relation, and rights is useful, if not necessary, to enable the retrieval of laboratory data. Data acquisition and analysis are facilitated and shared more readily electronically. Furthermore, data sharing enhances the team approach resulting in better quality control through enhanced data integrity and better data analysis.

The software architecture seeks to integrate all the features necessary to support and augment the scientific discovery process. We demonstrate the utility of the software with a current biotechnology example uniting the disciplines of microbiology, chemical physics, and medicine to promote basic research as well as to develop new diagnostic tools. This interdisciplinary, multi-institutional research team, comprised of research faculty and personnel from the Department of Biological Sciences and the Liquid Crystal Institute at Kent State University, the Northeastern Ohio Universities College of Medicine, and Summa City Hospital, invented a novel microbial biosensor. Their needs for data acquisition, analysis and sharing are numerous. The team needs to collectively conceive new ideas, prevent redundant experimentation, swap results and write manuscripts while performing other daily activities in different physical locations. Thus, each team member needs to identify and view raw data, summary data, photographs, reactions, analog and digital equipment output, manuscripts, and visual presentations produced by other team members. The linear retrieval format of the traditional, chronological entry lab notebook hinders efficient data sharing between the team members. Optimal data sharing begged for an electronic database system that tagged data in multiple ways beyond the time/date format. Furthermore, discipline-based investigators tend to be restricted by domain-specific database systems that do not readily permit identification of relevant literature across multiple disciplines. The software we present uses Dublin Core as the common data cataloging format by which multidisciplinary, multi-institutional research teams can describe, identify and exchange relevant data in early stages of the scientific process.

3 Advantages of Dublin Core with Laboratory Data

A number of characteristics associated with DC make it a promising choice for describing scientific laboratory data (Duval et al. 2002, Weibel 1995, Weibel 2000). DC is relatively concise, simple, and easy to learn, increasing the likelihood that busy scientists, technicians, and students have the time to create and maintain metadata records for laboratory data as it is generated (Baker 2000, Greenberg et al. 2001). DC supports multiple formats including text, still images, video, audio, and datasets generated within scientific labs. Because DC is designed to facilitate Internet resource discovery, DC facilitates making scientific data rapidly and widely available at appropriate times. Plans for a DC metadata registry promise to ensure consistency of its application. Attention to international concerns increases the likelihood that DC will be used worldwide. While the current paper focuses solely on using only DC elements, in the future, possible

extensions of DC will be integral to this project to address the multidisciplinary, multi-institutional scientific research (Forsberg 2000). Elements from other metadata schemas will be added to enhance desired functions (e.g., administration) or provide more specific information for certain groups, e.g., interdisciplinary team members representing different domains. Finally, the American National Standards Institute in conjunction with the National Information Standards Organization approved the Dublin Core Metadata Element Set as a national standard (ANSI/NISO Z39.85-2001) on September 10, 2001 (NISO 2002). This approval and efforts toward the acceptance of DC as an international standard are expected to strengthen confidence in the value of DC as a tool for improving resource discovery and information exchange leading to increased use of the metadata standard in the U.S. and abroad (Dekkers and Weibel 2002).

4 Database Design of the Scientific Notebook

A modified relational database has been constructed which provides all of the information typically incorporated in a print scientific laboratory notebook including raw data collected from numerous experiments; intermediate documents, such as procedures, memos, and technical documentation; and final products, such as completed research papers and multi-media presentations. The information that is most central to a scientific laboratory notebook is contained in the Main Notebook section of the database. The Main Notebook includes a number of tables which capture general descriptions of past, present and planned projects (Topic); experiment design concepts specific to a given project (Experiment Goals); procedures and materials used in actual experiments (Materials & Methods); and specific procedural components (Steps). Results would include drafts and finished papers (Topic Results), data tables and graphs (Experiment Results), or images and datasets (Materials & Methods Results). This section is organized hierarchically (see Figure 1). Higher level tables may be associated with an unlimited number of lower level table items (e.g., each Topic may be associated with an unlimited number of Topic Results and an unlimited number of Experiment Goals. An entry form has been created to facilitate recording of laboratory notebook data and to test the initial design of the Main Notebook. It is estimated that approximately 200 entries will be available to be described by DC from the paper laboratory notebook, including 1 Topic, 8 Experiment Goals, 61 Materials & Methods, and 127 Materials & Methods Results. Supplementary Tables support the scientific investigation by archiving additional information related to the Main Notebook that support the scientific investigation (See Figure 2). Dublin Core Records are being associated with items contained in each table of the Main Notebook except for Steps. See Appendix 2 for an example of a DC record associated with a Materials & Methods Results entry. Materials collects detailed information such as MSDS (material safety data sheets), Specification Sheets, and Materials Lot Analyses about organisms and liquid crystal substances involved with experiments. User includes basic contact information about individual researchers involved with the scientific investigation and specifies authentication and access rights. Memos includes entities such as correspondence, equipment issues, and notes for future experiments.

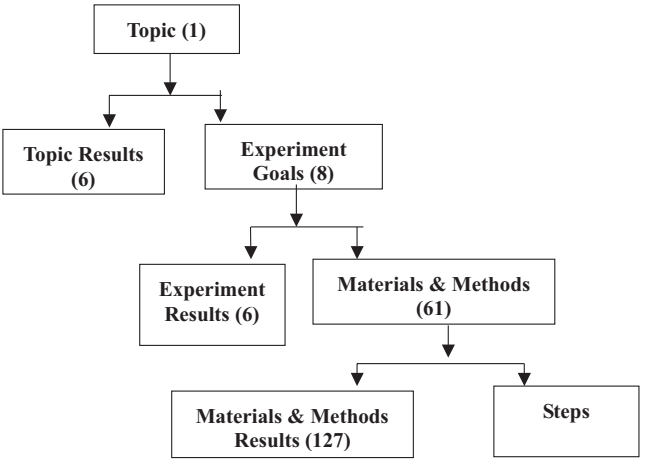


Figure 1: Representation of hierarchical database design of the Main Notebook and estimated number of entries available for Dublin Core description in parentheses

All tables within the Main Notebook will be associated with Dublin Core records except for Steps.

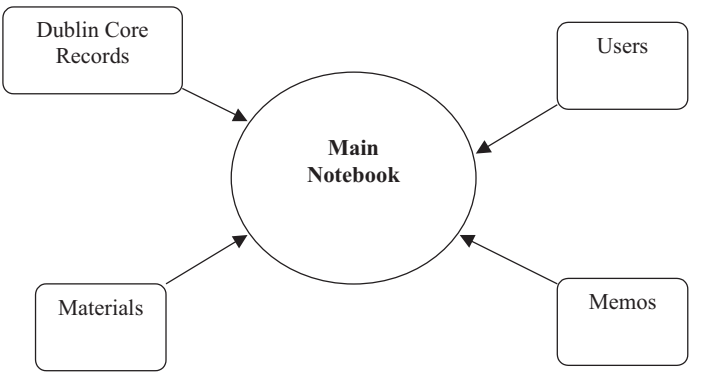


Figure 2: Representation of Main Notebook and Supplementary Tables

Supplementary tables are linked to the Main Notebook to provide additional information about the data.

5 Methodology

DC is being applied to a database representation of the laboratory data recorded in a paper notebook for a single project by one investigator over the course of one year. Microbiological experimental design, implementation, and result components are being described using DC. A content analysis will be used to investigate whether the DC schema is a feasible method of describing laboratory data. An expert user study will investigate the utility of DC for data description. Ex-

amples of the DC records with their associated laboratory data and results from the feasibility testing will be presented at the CRIS 2002 conference.

5.1 Feasibility Testing

During the DC application phase, any difficulties encountered in describing different types or aspects of the data are being noted (see Appendix 1). After application is completed and noted difficulties are addressed, if possible, a computerized content analysis (Murray 1998) of the DC records will be conducted in order to determine the frequency with which each element is used overall as well as for different types of information objects. Since the sample used in this study is much smaller and more heterogeneous than the sample used by Murray, much less content variability may be obtained for a given element. If results show that an element is used consistently but its content seldom or never varies, only unique instances of the element will be included in the frequency count in order to avoid artificially inflating results. The DC records will also be visually analyzed to identify any nonstandard DC usages. These measures provide an indication of the ease with which the DC element set can be applied to laboratory data as information objects. If most DC elements are used in accordance with established standards, then a good fit between the metadata schema and the data being described exists. If many unusable elements or inappropriate element assignments are identified, extensions or alternatives to the DC element set will be considered.

Following the content analysis, the functionality of the DC element set will be examined by using four specific types of metadata classes that provide information about digital objects. The four metadata classes are: discovery, use, authentication and administration (Greenberg 2001). Metadata classes group element level metadata by the function(s) each element supports. For example, the DC elements *creator* and *subject* both contribute to the function of discovery. In applying Greenberg's definitions of the classes to the laboratory environment, it is assumed that discovery supports research needs of users while use, authentication, and administration support functional needs of the workplace. The DC schema's ability to sustain required information functions will be assessed by aggregating the content analysis element frequencies for each class. Information regarding average frequency of use for each metadata class will be compared across data types to identify any marked differences in DC's effectiveness regarding specific functions among data types. Results of the feasibility testing will be presented at the CRIS 2002 conference.

5.2 Utility Testing

After feasibility has been determined, the next stage of the investigation will be to conduct usability tests with experts to assess the effectiveness of the prototype database and DC in handling laboratory routines as well as scientific research. User subjects will comprise in-house researchers from a Kent State University biotechnology lab, including graduate assistants, technical assistants, and scientists. The subjects will query the database using two separate interfaces to complete predetermined tasks, representative of both laboratory workplace needs and scientific research needs. One interface will access only the data and the second interface will access the DC enhanced data. The tasks will correspond to discovery, one of Greenberg's four metadata classes. Database transaction logs, subject interviews, and analyses of result sets will be compared for the discovery metadata class to record user evaluations of DC's effectiveness in handling laboratory data.

Future studies will explore the utility of the prototype database and DC with scientific laboratory data through additional user studies. One type of user study will include researchers from other laboratories and in other disciplines. Such research will focus on the multidisciplinary, multi-institutional, collaborative components of the biotechnology project. The second type of

user study will extend the utility study to include Greenberg's remaining metadata classes: use, administration, and authentication. Such an investigation would provide quantitative analysis of DC's functionality in areas outside of discovery.

6 Discussion

This project is not limited to the scientific community but also seeks to address the growing need in all segments of our mobile society for tools where people can transform information into usable knowledge and share this knowledge effectively. The prototype scientific notebook would support distributed work environments, tying together people from multiple organizations to collaborate on complex projects from different locations and at different times.

Future developments for the scientific notebook project would enable a scientist using his or her notebook as an interface to upload data to national and international data repositories or to search literature databases relevant to his or her research questions. Whenever the researcher would wish to review results, compare them to the results of others, and read relevant, similar experiments and studies, he or she would use the notebook to access this information. In addition, the scientist would be able to use the notebook to interface with stored results in order to prepare papers and presentations for scientific journals and to give presentations at conferences. Further, in time, the scientist would be able to set up and run an experiment, using the electronic notebook to control a virtual laboratory. Such capability would allow for the refinement of already running wet lab experiments and the efficient planning of future wet lab experiments.

Though the scientific arena is a specialized one, the ideas, methods, and tools used to gather, organize, and present scientific information can also be used in other settings. In fact, what makes the ideas presented in this paper especially appealing is that the gathering, organizing, and presenting are done in a multidiscipline setting. Thus, these methods and ideas are potentially useful in almost any situation where data need to be gathered, assimilated, and presented or used.

7 References

Baker, T., 2000. A grammar of Dublin core. *D-Lib Magazine*. 6 (10). [On-line]. Available at: http://www.dlib.org/dlib/october00/baker/10baker.html

Buckland, M.K. 1997. What is a "document"? *Journal of the American Society for Information Science*, 48 (9): 804-809.

Dekkers, M. and Weibel, S. 2002. Dublin Core Metadata Initiative Progress Report and Workplan for 2002. *D-Lib Magazine*, Volume 8 (2) [On-line]. Available at: http://www.dlib.org/dlib/february02/weibel/02weibel.html

Duval, E. Hodgins, W. Sutton, S. Weibel, S.L. 2002. Metadata Principles and Practicalities. *D-Lib Magazine*, 8 (2) [On-line]. Available at: http://www.dlib.org/dlib/april02/weibel/04weibel.html

Forsberg, K., 2000. Extensible use of RDF in a business context. *Computer Networks*, 33:347-364.

Greenberg, J. 2001. A quantitative categorical analysis of metadata elements in image-applicable metadata schemas. *Journal of the American Society for Information Science*. 52 (11): 917-924.

Greenberg, J., Pattuelli, M., Parsia, B. Robertson, W. 2001. *Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization*, Journal of Digital Information, 2 (2) [On-line]. Available at: http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg

Jeffery, K. G. 2000 An architecture for grey literature in a R&D context. *The International Journal on Grey Literature*. 1 (2): 64-72.

Murray, K. 1998. CIMI DC Simple Testbed *Record Content Analysis*. [On-line]. Available: http://www.cimi.org/old_site/documents/CIMI_DC_Simple_RCA.html; accessed January 25, 2002.

National Information Standards Organization (NISO). 2002. NISO standard: Dublin Core Metadata Element Set (ANSI/NISO Z39.85-2001). [On-line]. Available: http://www.niso.org/standards/standard_detail.cfm?std_id=725

Weibel, S. 1995. Metadata: The foundations of resource description. *D-Lib Magazine*. 6 (07). [On-line]. Available at: http://www.dlib.org/dlib/July95/07weibel.html
Weibel, S. 2000. The Dublin core metadata initiative: Mission, current activities, and future directions. *D-Lib Magazine*. 6 (12). Available at: http://www.dlib.org/dlib/december00/weibel/12weibel.html

8 Contact Information

Laura Bartolo
Kent State University
College of Arts & Sciences
Kent, Ohio 44242-0001
USA

e-mail: lbartolo@kent.edu

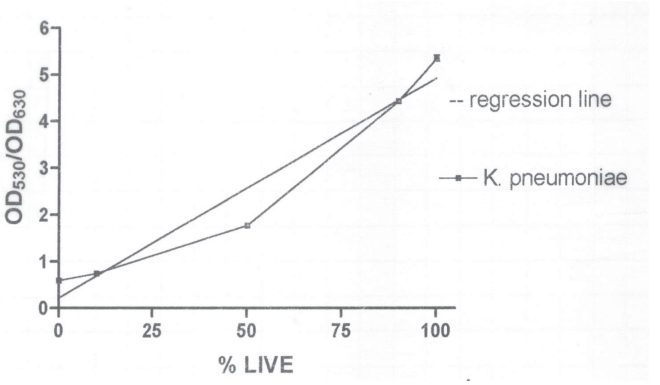
Appendix

APPENDIX: 1 Problems with DC Application

Description of Difficulty	*Item Identification*
	*Area: **Info Object: Record #:
	Area: Info Object: Record #:
	Area: Info Object: Record #:
	Area: Info Object: Record #:
	Area: Info Object: Record #:
	Area: Info Object: Record #:

*Area: Goal, Procedure, Goal Result, Procedure Result
**Info Object: Text, Table, Graph, Image, Dataset, Video, Audio

APPENDIX: 2 Example of Laboratory Data and Associated Dublin Core Record



Title = “Bacterial Toxicity Assay of CPCl treated Klebsiella pneumoniae”

Creator = “Woolverton, Christopher J.”

Subject= “Bacterial Toxins—analysis” (MeSH)

Subject= “Klebsiella Infections—immunology” (MeSH)

Subject= “Klebsiella pneumoniae” (MeSH)

Description = “Graph of Bacterial Toxicity Assay of CPCl treated Klebsiella pneumoniae. % live standard curve used to evaluate CPCl effects.”

Date = “2000-09-06”

Type = “image”

Format = “image/jpg 183 KB”

Identifier = “CJW2_043001.JPG”

Identifier = “Materials and Methods Result #39”

Language = “en-us”

Relation = “IsPartOf Materials and Methods #18”

Comparative Study of Metadata for Scientific Information: The place of CERIF in CRISs and Scientific Repositories

Keith G Jeffery, Andrei Lopatenko, Anne Asserson
CLRC, MU, UiB

Summary

Metadata provides the human- and machine-accessible gateway to data, improves data to information, and provides the semantic context within which knowledge can be induced from information. Metadata is the means for using together scientific data from heterogeneous sources. A CRIS commonly holds data which, while useful in itself, commonly is also metadata describing more detailed data and information on projects, persons, organizations, products of R&D (patents, products, publications) equipment used for R&D and R&D funding. It is important, therefore, to classify the metadata formats used in various scientific repositories in order to understand their scope and interoperability, and their relationship to CERIF representing CRISs. Metadata formats are reviewed according to intention, abstraction level and technology criteria. The place of CERIF in CRISs in this wider sense (including scientific repositories) is considered and compared with other metadata models and formats. The superiority of CERIF (in formalism and flexibility) is demonstrated.

1 Introduction

1.1 Data, Information and Knowledge

Europe is encouraged to evolve to 'the knowledge society'. In order to reach this state it is necessary to take data and convert it to information by structuring within context. For example '06-03-2002' would mean to a USA user 3 June 2002 but to a European 6 March 2002. Indicating which digits represent day, month and year by use of metadata (e.g. mm-dd-yyyy or dd-mm-yyyy) structures the data and makes it comprehensible. Of course, the unambiguous 20020306 is preferred, and is reliable information. Then knowledge (defined as commonly accepted belief) can be produced by induction over the information – sometimes described as generating the intension from the extension. This process is analogous to the classical scientific mind-process of building hypotheses based on observed patterns in information. As an example, if one observed the timetable for planes from London to Frankfurt to be :

Start-Time	Departureairport	Flight	Arrivalairport	End-Time
0800	LHR	BA123	FRA	1000
0900	LHR	BA125	FRA	1100
1000	LHR	BA127	FRA	1200
1100	LHR	BA129	FRA	1300
etc				
etc				
1800	LHR	BA137	FRA	2000

one might induce that, during normal working hours, a BA flight leaves LHR every hour on the hour for Frankfurt and takes 2 hours (including any time-change).

Knowledge applied to a situation provides insight, and this fuels the process of technology transfer, of inspiration to do further research and of education.

1.2 The Data Deluge or Information Explosion

The exponential growth of scientists, and especially of the power of data collection devices they use, has led to problems in dealing with data and information, and finding the needed knowledge within information. Earth observation satellites provide megabytes per minute. The LHC (Large Hadron Collider) at CERN will generate approximately 5 petabytes per year. Scientists publish vast numbers of peer-reviewed papers and even more grey literature material. One of the possible approaches to solve these problems is classification and structuring of information. Metadata provides interpretation or meaning of data and facilitates information and knowledge management.

The question facing us concerns the relationship of all this scientific repository data to conventional CRISs and how to manage the data together. In general the end user requires lateral browsing over a CRIS and in-depth searching for the detailed scientific repository data associated with a record in the CRIS. For example, the user might search for all projects investigating the link between smoking and lung cancer, and then wish to dive deeply into the analysis software used, the original raw scientific data and the publications from one particular project. The answer lies in the use of metadata to provide an integrated information space and to provide access: in this way scientific repositories and bibliographic repositories (among others) - while continuing to be used for their normal purposes - come within the sphere of access of CRISs.

1.3 Data and Metadata

Metadata is the key to achieving interoperation of CRISs and of CRISs with scientific and bibliographic repositories. Metadata has been divided into three main types (Jeffery 1998):

1. schema metadata is an intensional description of extensional instances. Typically a schema consists of: database {name, size, security authorisations}, attributes {name, type, constraints}. Some of the constraints concern the attribute domain, some are inter-attribute and as such may express relationships. The schema (intension) has a formal logic relationship to the data instances (extension). This is important in ensuring data quality. It also provides a formal basis for systems. In short, schema metadata constrains the data it describes to ensure its integrity.
2. navigational metadata provides information on how to get to an information source. Mechanisms include: filename, DB name + navigational algorithm, DB name + predicate (query), URL (Uniform Resource Locator), URL + predicate (query) or various combinations of them. Navigational metadata has no formal logic relationship to the data instances, however attributes of the navigational metadata may describe the name of the collection of data instances (e.g. filename).
3. associative metadata provides additional information for application assistance. The assistance may improve performance, accuracy or precision of the system and / or provide assistance to the end-user through a domain aware supportive user interface. The main kinds of associative metadata are:
    a. descriptive: catalogue record (eg (Dublin Core))
    b. restrictive: content rating (eg PICS) or security, privacy (cryptography, digital signatures) (W3C) or rights usage
    c. supportive: dictionaries, thesauri, hyperglossaries (VHG), domain ontologies eg (PROTÉGÉ)

Associative metadata usually does not have a formal logic relationship to data instances although there may be systematic association relationships.

2 Usage of metadata in CRISs

2.1 Introduction

All types of metadata are used in CRISs and also in the extended sense of CRISs including scientific repositories. To assess the different metadata formats, we evaluate them on three main dimensions – (1) kind (schema, navigational, associative), (2) intended use (3) CERIF compatibility. The compatibility with CERIF is subdivided into two parts 1) if a metadata format is compatible and an interoperable solution can be developed; 2) if a metadata format describes entities different from CERIF and those can be used (extending CERIF) to provide additional services for CRIS.

2.2 Schema metadata and CRIS

The schema metadata in CRIS can be used in implementing distributed information access solutions, integration of data, building up directories of CRISs (systems assisting user to find CRIS relevant to his data requirements) and describing web resources generated by CRIS for data collecting by agents.

CERIF RDBMS is defined in SQL DDL. The EuroCRIS CERIF Task Group (ECTG) developed an (XML) Schema for basic CERIF entities (projects, persons, orgunits and relations between them). Due to a weakness of XML Schema, integrity rules like uniqueness, referential integrity are not describable. Universal Modeling Language (UML) is powerful enough to describe CERIF. In terms of schema metadata, CERIF is as fomally defined as it can be. This means it can be co-utilised with schemas from scientific repositories and bibliographic repositories; the quality of the co-utilisation depending on how formally their schemas are defined.

2.3 Navigational metadata and CRIS

The possible uses of navigation metadata for CRIS include description of data location for distributed systems, providing persistent data access for information sources, and resolving location of data by its characteristics. Examples include the University of Bergen index of CRISs (graphical interface to URLs) (BergenCRIS), DOI (Document Object Identifier) and its use in the publishing industry (a number of URLs for the same resource for different aims). CERIF uses URIs to provide navigational metadata referencing (pointing to) more detailed sources of information. In terms of navigational metadata, CERIF is as complete as it can be. This mechanism provides the primary extension of conventional CRISs to include scientific and bibliographic repository data.

2.4 Associative-descriptive metadata and CRIS

CERIF provides – when used as metadata – rather complete associative-descriptive metadata for CRISs. This CERIF metadata provides a formal information context for consideration of data from scientific or bibliographic repositories, by relating the detailed repository data to a canonical model describing research information.

2.4.1 Metadata for scientific repositories

These metadata describe scientific data or scientific data collections and represent scientific objects or results of scientific experiments, allowing them to be exchanged between applications,

stored in databases or published using WWW. In CRISs they can: 1) assist in the seamless extension of the user search space from the conventional CRIS information to detailed scientific and technical information; 2) serve for precise and formal description of subjects of research information to improve search effectiveness. Some examples are:

(1) Spatial Data Standards for Facilities, Infrastructure and Environment (SDSFIE): can be used in CRIS which needs geospatial description of research information entities or scientific results
CERIF compatibility: the standard maybe used to extend the location entity in CERIF. No technologically compatible version now.
(2) Content Standard for Digital Geospatial Metadata (CSDGM): CERIF compatibility: CERIF is able describe location of objects. CERIF location consists of country (2-letters code), region code, city and mail address. CSDGM maybe used as extension to CERIF notion of location.
(3) OpenMath is a standard for representing mathematical objects with their semantics. CERIF compatibility: OpenMath can be used to describe formulae or used in mathematics research information and then used to search information.
(4) CSCM (Content Standard for Computational Model (Hill et. al. 2001) CERIF compatibility: CSCM can be used to describe computational models and then used to search information. Also it may be used to describe in details product results in CRIS oriented on computational models
(5) GILS (Guideline Interchange Standard) CERIF compatibility: GILS can be used to extend result product description in CERIF for special purpose CRISs.

2.4.2 Metadata for research information

These metadata describe sources (such as bibliographic repositories) with similar entities to those described by CERIF e.g. Digital Libraries. Translating CERIF metadata into other metadata schemas allows re-use of CERIF data by other systems: 1) to attract new users of research information; 2) to provide additional services to work with research information; 3) to organize multi-system information processing. Translating research information expressed in other metadata formats to CERIF allows: 1) one to load new data into CERIF CRIS; 2) to interoperate by exchange between CRISs through the canonical CERIF definition; 3) to interoperate by access to heterogeneous distributed CRISs using CERIF as the canonical model to drive the portal. Such metadata formats are compared:

(1) CERIF ontology is a formalized CERIF-based metadata format. CERIF2000 specification defines metadata format for CERIF in full-text and diagrams and formal RDBMS schema. CERIF ontology is a formal specification in (DAML) of what CERIF metadata means and how they maybe encoded into (RDF). CERIF ontology was developed by the euroCRIS CERIF Task Group. (RDF) maybe used to encode metadata. CERIF TG provides a toolset to import/export data from CERIF RDBMS, use them for integration of distributed sources, provide intelligent information search. CERIF compatibility: CERIF ontology is formalized CERIF metadata definition. The final aim is full CERIF implementation..
(2) Dublin Core (DC) was adopted and applied for a national distributed CRIS in the Safari project in Sweden (SAFARI). DC is machine readable but not machine understandable (Jeffery 1999) and a formalized DC was developed. The current thinking is that DC provides a high-level associative-descriptive metadata and that more detailed, formal and domain specific associative-descriptive metadata sets are required for real interoperability or advanced information processing including query improvement and results explanation (Lagoze 2000).
CERIF compatibility: The mapping from CERIF entities and attributes to Dublin core is provided by the ECTG.

(3) Math-Net application profile : CERIF compatibility: in description of CERIF entities Math-Net mostly is a subset of CERIF. Mapping between CERIF and Math-Net is provided by CERIF TG.
(4) Grey Literature metadata (GL) includes adaptation of CERIF and DC metadata formats to describe grey literature. The importance of Grey Literature for science (Wentraub 2000) and CRIS (Jeffery 1999) has been identified. The proposed Grey Literature metadata format is developed as extended formalized DC based on Extended E-R-A modeling and may be coded as RDBMS DDL. It may also be encoded in (RDF) and (XML). CERIF compatibility: the format is developed with CERIF compatibility as a key requirement. It provides for extended description of publication entities. A formalized model of attribute encoding to preserve the referential integrity of CERIF data is proposed.
(5) SWRC (Semantic Web Research Community) SWRC is a formal (DAML) ontology. CERIF compatibility: highly intersected with CERIF.
(6) Common European format for curricula vitae (CV) CERIF compatibility: describe person expertise mostly like CERIF.

IN addition there are exist a large number of other formats to describe scientific publications: ETD-ms: an Interoperability Metadata Standard for Electronic Theses and Dissertations, Euler's metadata for electronic journals etc.

2.5 Associative-restrictive metadata and CRIS

These metadata restrict the access of users (or software systems) to sources, modes of use or alternatively provide security services like preserving intellectual rights, or encryption. CERIF handles associative-restrictive metadata by placing constraints (mapped as attribute values together with temporal constraints) in the linking relations which represent the relationship between, for example, an author and a publication or a user and a publication. This provides much more flexibility than storing associative-restrictive attributes with the publication or with the person (in role author or user) since it is easily changeable and extensible.

Intellectual rights and secure policies for Research Information

The intellectual rights issues for CRIS were investigated in (Seipel 2000); (Losano 1995); (Beren & Rubert 1995). Such restrictions require a metadata to describe information access/ copy/use policies, for users, and for information. A general description of the application of Digital Rights technologies in CRIS-like applications can be found in (Ianella 2001); (Erickson 2001).

(1) XrML (eXtensible rights Markup Language) is a Markup Language to describe intellectual rights and conditions associated with digital content, resources, and services. XrML was developed at PARC (Xerox Palo Alto Research Center) and governed by ContentGuard, Inc. CERIF compatibility: CERIF entities can be described as XrML resources.
(2) ODRL (Open Digital Rights Language) CERIF compatibility: ODRL can describe intellectual right for research information. CERIF entities can be modeled as ODRL asset elements.

2.6 Associative-supportive metadata and CRIS

These types of metadata are being used to create domain values for descriptive metadata and other support services and include dictionaries, thesauri, hyperglossaries, and domain ontologies. CERIF provides a framework whereby almost any associative-supportive metadata can be used. The use of separate relations (lookup tables) to hold permitted values of an attribute together with a term description provides maximum flexibility in both classifying and reclassifying an entity by its attributes. CERIF has built-in multilinguality features. Furthermore, terms in CERIF can be referenced to and from externally established thesauri and / or domain ontologies.

2.6.1 High-level Thesauri

The objective is to increase search effectiveness (Buckland 1999); the use of controlled terms improves both precision and recall, and if the attribute of the term is also a foreign key it can improve greatly referential integrity (and therefore relevance and recall of deep searches):

(1) Library of Congress Subject Headings () is a high level thesaurus used by many libraries.
(2) Cyc is a formal ontology, covering a large set of terms with formal description of their meaning. The best use of Cyc for CRIS is likely to be development of CRIS for knowledge management, and as a terminology server for CERIF to extend such existing CERIF vocabularies as the role of persons in projects and organizations.
(3) WordNet is a terminological ontology and could be useful for CRIS as a metadata for assisting in query building or information search. Example: "deontic logic" is a hyponym of "modal logic" - and for a particular kind of search the hyponym term may give improved precision and recall.

2.6.2 Thesauri for Science

Thesauri for science are thesauri of scientific terms and research areas which can be used for indexing research information about projects or expertise to improve search effectiveness

(1) CERIF2000 / ORTELIUS, has wide multilinguality, but only very basic terms The CERIF Task Group provides a distributed version of ORTELIUS as SQL DDL for immediate installation into any CERIF-compatible database and also the (XML) encoding of ORTELIUS according to VocML.
(2) (COS Keywords) is a controlled vocabulary of terms for science which is used to describe research expertise and funding opportunities.
(3) AMS MSC – The Mathematical Subject Classification of American Mathematical Society categorises items in mathematics; there are other ontologies and thesauri in this field
(4) Research Methods Glossary – an index of terms to describe research methods of a wide set of sciences.
(5) Physics and Astronomy Classification Scheme® (PACS) is a thesaurus developed by the American Institute of Physics, and has been used in Physical Review since 1975.

2.6.3 User-Oriented Presentation

The following metadata types are used to support additional value-added services which could make CRIS more usable by providing higher-level entry-points into CRIS systems.

(1) Metadata for collection descriptions serves for description of collections. There are several developing standards in this area :
    (a) (RSLP) (Research Support Libraries Programme) Collection Description is based on modeling of collections and their catalogues. Has a metadata schema and associated syntax using (RDF)
    (b) (SCD) Simple Collection Description. A format of eLib Collection Description working Group.
    (c) (ISAD(G)) (General International Standard Archival Description) – provided a general framework for description of archival collections
    (d) (DCMICD) A standard for collection description as a set of elements as extension to Dublin Core to describe and share information about collections

CERIF, because of its naturally recursive nature, can describe collections and reference the individual items within the collection. The proposed extension to CERIF2000 to handle publications in more detail within a formal Dublin Core framework, would ensure that CERIF provides a compatible superset capability over the majority of these proposed standards.

(2) Metadata for quality rating serve to describe the quality of information resources to help users in filtering qualitative resources and evaluation of resources. Description of the quality of research information resources is created by the information provider or an expert. Value-added services can help information consumers in getting only high-relevant and qualitative information according their needs. A review of quality control procedures for research information in the medical field is provided in (Eysenbach & Diepgen 1998).
   (a) PICS – Platform for Internet Content Selection, is a recommendation developed by W3C. Although developed initially to provide parental (or other) control over access by minors to unsuitable material on the WWW, an example of PICS use in the CRIS field is "Critical Appraisal of Medical Information on the Internet" (CCAMI) – a collaborative project of physicians dedicated to medical quality on the Internet.
   (b) There are attempts (led by ACM) to improve quality of scientific publications in the IT field by having online digital review, where known and trusted reviewers are encouraged to provide commentary attached to any digital publication. The aim is to assist readers to judge the quality. It is not unlike the peer review process but is not anonymous. The commentary can be regarded as 'annotation metadata'.

2.6.4 Workflow and data processing

The following metadata types described workflow data: description of routines for data and document processing which maybe required for CRISs of policy-maker organizations or CRISs with a high demand for data quality, in which research information is processed by experts according to some corporate policies. Some examples of using workflow processes for research information are described in (Lindgren 2000), (Shyu-2000). Very often workflow routines are used in digital libraries providing access to theses, publications, and preprints with research information. The Workflow Management Coalition metadata [WFMC] - a metadata set allowing representation of many facets of workflow, is one of the most popular among workflow standards. It has specified a reference model, vocabulary and XML Schema to describe workflows

A particular Schema for SGML ETD Workflow Record[SGMLETD] exists and is a SGML Schema to describe the workflow processes of submitting, approving, publishing electronic thesis and dissertations. CERIF2000 provides classification schemes related to results (products, patents, publications) by linking relations such that different classification schemes may be used to apply to the same result (e.g. a publication). However, a more general treatment using annotation metadata may be more applicable.

3 Metadata and CERIF

The current (CERIF2000) definition includes metadata of the three types:
- Schema metadata to describe information sources, CRIS data models, metadata formats;
- Navigational Metadata with URLs to more detailed sources
- Associative metadata of the kinds:
  - Associate-descriptive metadata – CERIF metadata model, and CERIF database to describe entities of research information;
  - Associative-restrictive metadata - CERIF metadata model, and CERIF database to describe entities of research information;
  - Associate-supportive metadata – CERIF vocabularies to classify entities or values.

Looking forward, potential solutions to future CRIS requirements (including scientific and bibliographic repositories) are considered to lie in different architectures, depending on the then user requirements, but utilising (an extended) CERIF.

(a) Distributed information access solutions aim to provide transparent access to data in a distributed heterogeneous environment where there is a strong need to have a semantic and structural description of the data model in repositories for CRIS. CERIF2000 is an example.
(b) Directories of CRISs are systems which assist the user or machine mediator in choosing the information source from which to obtain data. To reason about whether an information source satisfies user requirements, the data of the information source should be described by their content, actuality and other characteristics – i.e. metadata. Examples include NIWI (DRIS), California Digital Library – Directory of Collections, OBSERVER (UGA) (and scientific publications data), TSIMMIS.
(c) Describing web resources requires that research information can be collected from metadata scattered on the web. As different organizations or communities use different formats to describe research, the metadata describing the meaning of attributes, objects, properties and constraints for valid data should be defined. ECTG defined a mapping between ontologies.

One of the key requirements for CERIF was the compatibility with other metadata formats on the basis of the layered information model (Jeffery et. al. 1994) with distinct semantic, conceptual, intensional, logical and physical levels. This is to simplify the technologies to implement solutions and allow better technologies to be utilized independently of each other by using clean interfaces between the layers. To satisfy the requirement that CERIF is based on an open, widely-used standard or formats (RDBMS, (XML), (DAML + OIL), (RDF)), the CERIF2002 model is developed from the CERIF2000 model by agreement of experts from different countries after a long analysis stage. The CERIF2002 model thus extends CERIF2000 and contains formal machine-understandable definitions of entities, properties and constraints to make the compatibility study of CERIF with other metadata formats more automatic and formally provable. This will allow also automatically to integrate CERIF based CRISs with other, non-CERIF-compatible, CRISs. Furthermore it provides the foundation for integrating scientific repositories and bibliographic respositories seamlessly into the information space of a CRIS. The vision of CERIF – in addition to its canonical data content and structure for exchange and access - is for it to provide the common vocabulary for research which can be used to create global schemas, to be a kernel (or reference) ontology for building CRIS ontologies, and to be a shared vocabulary for federation systems or systems based on harvesting of metadata.

4 Conclusions and recommendations

CERIF2000 provides an excellent basis and is demonstrably mappable to formal metadata classifications of other CRISs. CERIF2002 as proposed by the ECTG goes further and provides a basis for open interoperability providing seamless acces for users across CRISs and associated scientific and bibliographic repositories.

5 References

BergenCRIS
http://www.nsd.uib.no/english/research/eucris/

Beren, M; Rubert, C; (1995) The new Spanish data protection act, EuroCRIS-1995, Milan, Italy.

Buckland, M (1999) Vocabulary as a Central Concept in Library and Information Science in: Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science (CoLIS3, Dubrovnik, Croatia, 23-26 May 1999)

CCAMI Collaboration for Critical Appraisal of Medical Information on the Internet,
http://www.dermis.net/medpics/

CORBA
http://www.corba.org/
COS Keywords
http://keywords.cos.com/
DAML
http://www.daml.org/
DAML+OIL see [W3C]
DC
http://purl.oclc.org/metadata/dublin_core/
DCMIC]
http://dublincore.org/groups/collections/
DRIS
http://www.niwi.knaw.nl/cgi-bin/nph-dris_search.pl?language=en
Dublin Core
http://purl.oclc.org/metadata/dublin_core/
Erickson, J; (2001) A Digital Object Approach to Interoperable Rights Management, D-Lib Magazine, Vol. 7 Num. 6
http://www.dlib.org/dlib/june01/erickson/06erickson.html
Eysenbach, G; Diepgen, T; (1998) Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information. BMJ 1998;317:1496-1502 (28 November 1998)
Hill, L; Crosier, S; Smith, T; Goodchild, M (2001) A Content Standard for Computational Models, D-Lib Magazine, Vol 7 Num. 6., Jun. 2001.
http://www.dlib.org/dlib/june01/hill/06hill.html
Ianella R. (2001) Digital Rights Management (DRM) Architectures, D-Lib Magazine, Vol. 7 Num. 6
http://www.dlib.org/dlib/june01/iannella/06iannella.html
ISAD(G)
http://www.ica.org/biblio/com/cds/isadg_2e.pdf
Jeffery, K; Hutchinson, E; Kalmus, J; Wilson, M; Behrendt,W; Macnee, C (1994) A Model for Heterogeneous Distributed Databases Proceedings BNCOD12 July 1994; LNCS 826 pp 221-234 Springer-Verlag 1994
Jeffery, K (1998) The Future of CRIS Invited Paper CRIS98 Conference, March 1998, Luxembourg.
http://www.cordis.lu/cybercafe/src/jeffery.htm
Jeffery, K (1999) An Architecture for Grey Literature in a R&D Context Proceedings of Grey Literature Conference, 1999, Washington, US
Jeffery K.; Asserson A.; Revheim J; (2000) CRIS, Grey Literature and the Knowledge Society, Proceedings CRIS-2000, Helsinki
ftp://ftp.cordis.lu/pub/cris2000/docs/jeffery_fulltext.pdf
Lagoze, C (2000) pers.comm.
Lindgren N.; Rautamki A.; Managing Strategic Aspects of Research. Proceedings CRIS-2000, Helsinki,
ftp://ftp.cordis.lu/pub/cris2000/docs/rautamdki_fulltext.pdf
Losano, M (1995) The Schengen Treaty and Italy - Some problems in transnational data flow, Proceedings CRIS-95, Milan, Italy
MARC
http://minos.bl.uk/services/bsds/nbs/marc/commarcm.html
PROTÉGÉ
http://smi-web.stanford.edu/projects/protege/Hpkb-web/MusenWestKickoff/tsld001.htm
(RDF) see (W3C)
RSLP
http://www.ukoln.ac.uk/metadata/rslp/
SCD
http://www.ukoln.ac.uk/metadata/cld/simple/
Seipel P., (2000) Copyright, Information Technology and the Edifice of Knowledge, Proceedings CRIS2000, Helsinki
ftp://ftp.cordis.lu/pub/cris2000/docs/seipel_fulltext.pdf

Shyu Y.-M.; Kao C.-F.; The Integrated Research Information System: Government Research Bulletin (GRB), Proceedings CRIS2000, Helsinki,
ftp://ftp.cordis.lu/pub/cris2000/docs/shyu_fulltext.pdf
SAFARI
http://www.lub.lu.se/~colm/safari/safari.html
SGMLETD
http://www.ndltd.org/workflow/workflow.htm
UKOLN
http://www.ukoln.ac.uk/metadata/interoperability/
UML
http://www.omg.org/technology/uml/
van Woensel, L: (1988) CERIF Manual October 1988
VHG
http://www.venus.co.uk/vhg/
W3C
www.w3.org
W3Cmetadata
http://www.w3.org/Metadata/
Wentraub I., (2000) The Role of Grey Literature in the Sciences,
http://academic.brooklyn.cuny.edu/library/access/greyliter.htm
WFMC
http://www.wfmc.org
(XML) see (W3C)

6 Contact Information

Keith G. Jeffery
IT Department
Rutherford Appleton Laboratory
Chilton, Didcot
Oxfordshire OX11 UKCLRC-RAL
UK

e-mail: keith.g.jeffery@rl.ac.uk

Metasearch engine for Austrian research information

Marek Andricík
Vienna University of Technology

Summary

Majority of Austrian research relevant information available on the Web these days can be indexed by web full-text search engines. But there are still several sources of valuable information, which cannot be indexed directly. One of effective ways of getting this information to end-users is using metasearch technique. For better understanding it is important to say that metasearch engine does not use its own index. It collects search results provided by other search engines, and builds a common hit list for end users. Our prototype provides access to five sources of research relevant information available on the Austrian web.
Keywords: search engine, metasearch, ranking, transformation rules, Perl, CGI, parsing, forward.

1 Search engines

Basically, web full-text search engines can be divided into two main categories:

1. General, big well-known search engines, which try to index the „whole Internet“ (e.g., Google, Altavista, Yahoo, Excite, Lycos). They utilize their own special software, which, in most cases, is not open to public and details about implementation and interface are available very seldom. It is not rare case that even among these engines exists significant incompatibilities in query languages.
2. Specialized, usually smaller, topic-based or area-restricted engines. Small engines do not have to have the „military grade“ level like their general counterparts - they do not handle very big load and even middle-sized computer can host them. Also, software requirements and mainly data storage optimization are not so high. Several commercial and free solutions are available on the Internet these days.

Note: On the homepages of search engines one can usually find also the directory service. „Cheap“ implementation, often found on engines aimed at advertisement, is done as a gateway to the search service itself (e.g., Epilot). Professional engines maintain directories by hand. Yahoo provides these days most credited directory.

According to Sergey Brin and Lawrence Page [1], first search engines started appearing in 1994 with hundreds of thousands of indexed documents. Three years later they reached tens of millions of indexed documents. Nowadays, indexes of big well-known full-text search engines reach milliards ($=10^9$) of indexed documents. Together with the growth of indexes also the average number of queries raised from thousands per day at the beginning to hundreds of millions per day now. Such a big amount of data and traffic has to have corresponding processing power and storage capacity behind (e.g., Google consists of 10 000 computers in cluster, has several tens of terabytes of storage capacity and has over 2 milliard documents indexed [2]).

Major, well-established metasearch service providers claim that no single search engine can have all available documents indexed [5]. Having on mind the fact, that Internet is decentralized heterogeneous network, practically none single search engine can have up-to-date index of all available documents. There are several factors, which can turn search results unsatisfactory:

1. The engine does not know about the existence of some particular.
2. Documents on the web may change anytime. Search engine will not get any notification about changes. The only practical way of being up-to-date is polling every document in the index and check for changes. It is clear that it cannot be done immediately for all indexed documents - they have to be checked step by step. Depending on the nature of the document, its attributes or configuration of the search engine, checking period vary. Simple consequence of this fact is that search engines sometimes return links which do not conform to the query (document has changed) or even points „nowhere“ (document was deleted). In both cases, search results do not reflect reality.
3. There still exist sources of research relevant data, which cannot be directly crawled and indexed, but they can be searched through proprietary search interface. Mostly, they are dynamically on-the-fly generated documents from databases, which could have static URL but not every single document is listed anywhere and the only other way to get it would be „guessing“ its URL. One is thus forced to use proprietary search interface.

The existence of several search engines working concurrently, means that some of them already could index documents which others did not. It also effectively raises the probability, that the right and up-to-date document is found when several search engines are queried together. This is the time when the metasearch comes onto the scene to help end user overcome some problems and simplify querying process.

2 Metasearch engines

First metasearch engines appeared early, just one year later after regular search engines did. There are several ideas behind the metasearch, one of the most important is bringing „the best of all worlds“, which in terms of searching, means collection of relevant documents. When user decides to search for documents using several search engines, metasearch service comes handy. Not only it provides much more comfortable way how to get list of documents but it also saves the time. It has to be said, that metasearch engine does not have its own index nor it combines indexes of other search engines nor even it has direct access to them.

In the previous section some bottlenecks of search engines were identified. Let us examine how metasearch deals with problems:

1. Finding document. Metasearch directly will not help. Query is submitted to several engines at the same time, theoretically, bigger area is covered, which raises the chance of finding documents.
2. Tracking document changes. Applies the same as above.
3. Accessing documents behind proprietary search interface. Previous two cases could be considered supplemental but this one is the case where metasearch brings significant improvement over regular search. The way how general search engine crawl the web and collect documents fails in the situation when the is no complete list of all available documents - simply, search engine only follows links but does not submit forms, which are usually the only way how to get to those documents. On the other hand, metasearch engines have natural support for submitting forms. It is their main task to „simulate“ user.

Every coin has two sides, so does the metasearch. Major problem is that query languages used by different engines differ. It is usual problem of finding „common ground“. Some support full range of boolean operators (and, or, not), suffixes and grouping of terms with parenthesis. It is rare, but one can sometimes find search engines supporting even prefixes and infixes.

Each query has its syntax and semantics. When metasearch engine accepts primary query from end user it has to submit semantically the same query to every engine participating at the search. Depending on supported syntax of each particular engine submitted secondary queries would

very likely syntactically differ. When the richness of query languages differ, it is possible that some complex queries cannot be expressed in all languages. Developers of metasearch engines has two possibilities (not counting the case when they ignore the problem):

1. Define reduced common grammar of metasearch engine, so that any primary query can be transformed to all secondary queries.
2. Define simplifying rules, which will be applied when primary query cannot be transformed to secondary queries.

In the first case it is guarantied that results obtained by metasearch will be the same as results obtained by manual searches (assuming the same conditions on the search engines). But, it must be noted, that more complex queries (those, which cannot be, expressed in common grammar) are not allowed. In second case, it can happen that search results differ. Very likely it occurs when transformation rules are applied. Awkward consequence is that in case of more complex queries, metasearch engine must really submit modified or even simplified query, possibly with different meaning. Results are very likely to be different when compared to results obtained from search engines queried directly. [4] Since the process is quite complex, it is very hard to tell in advance what will happen. Somewhere between these two extremes lies our prototype.

3 Our metasearch prototype

All what was said in previous sections naturally applies also to Austrian research relevant information. Some of them can be indexed directly - and they are indexed on the regular basis by our instance of mnoGoSearch, web full-text search engine. For the rest (data, which cannot be indexed directly) there is metasearch engine. Current prototype is implemented as stateless CGI gateway written in the Perl language. As it is usual with the CGI technology, the program processes only request at the given time. In case of several parallel requests, each of them will have its own independent instance of running program. Program operates in several steps:

1. Accepts primary query and using transformation rules transforms it into set of secondary queries. The table of features of each search engine supports user's decision about the level of complexity of the query.
2. Requests are in parallel dispatched and metasearch engine then waits until all search engines deliver results or until timeout occur.
3. Responses are serially parsed; title and URLs are extracted and put into the final list of links.
4. If requested, the list is sorted according to ranking.
5. List of documents is displayed together with number of hits and time spent for each engine.

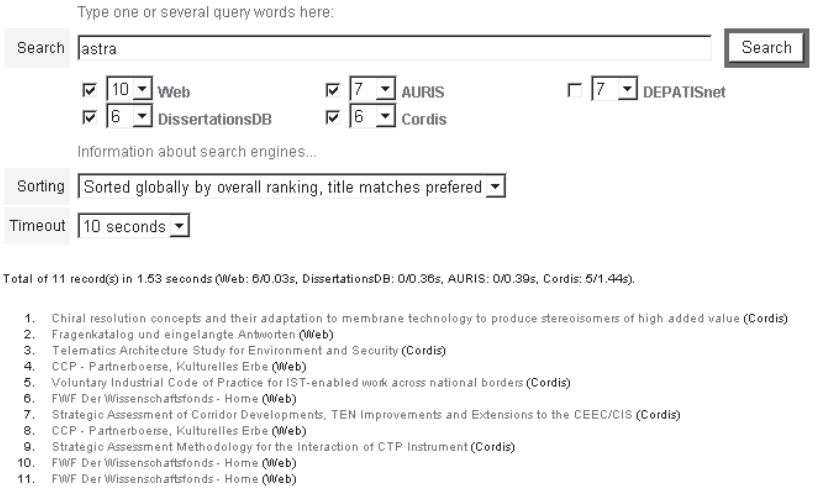


Figure 1: Metasearch results

3.1 Transformation rules

Current prototype has support for five search engines:
- Our mnoGoSearch search engine - indexes web documents
- Österreichische Dissertationsdatenbank - database of dissertations
- AURIS - Österreichische Forschungsdatenbank - old version of research portal
- Cordis - Community Research & Development Information Service
- DEPATISnet - German Patent and Trade Mark Office

Each of them has varying support for complex queries. All of them support the AND and OR operators but, one does not support the NOT operator and only minority has full boolean, phrase and support for the * and ( ) operators.

According to the content of the document there are four categories defined:
- Persons - researchers, project leaders, ...
- Institutes - research units, ...
- Projects
- Results - dissertations, thesis, patents, ...

Table on the page lists either level of support of operators and content coverage for each particular engine (see) . Using transformation rules, each primary query is converted to set of secondary queries. Why information retrieval standards as (e.g. Z39.50) are not used? Such standards can be utilized only when both sides participating at search support them. Vast majority of search engines provides only web interface accessible only through HTTP.

Name	AND	OR	NOT	( )	Full boolean	Phrase	*	Persons	Institues	Projects	Results	Comment
Web	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	Mnogosearch base full–text search engine
AURIS	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	AURIS – Osterreichische Forschungsdatenbank
DEPATISnet	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	German Patent and Trade Mark Office
DissertationsDB	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	Austrian Research Centers – Dissertationsdatenbank
Cordis	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	Community Research & Development Information Service

Figure 2: Feature table

3.2 Ranking

engine has very small amount of information, which can be used during the process of ranking (metainformation). Moreover, very few engines list numerical rating, which can be used by metasearch. Algorithm used by the prototype assumes, that partial results from each engine come already ordered by relevance and preserves that order. Furthermore, links whose title contains some of the searched words are pushed towards the top of the list. Search engines have also so-called overall ranking number assigned. Not only it is the way to measure quality of the sources but, it also allows each user to select his own preferred engines and rank it higher.

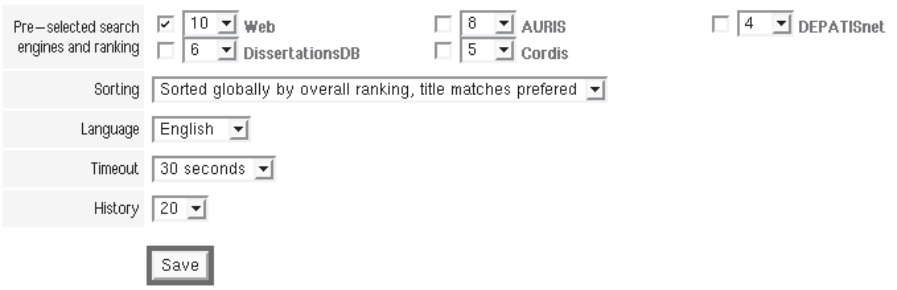


Figure 3: Customization page

3.3 User interface

User interface is designed to be highly customizable. One can work anonymously or can choose his login and password. Since then the system can keep user's preferences (pre-selected engines, their overall ranking, language and sorting criteria). User do not have to login every time, until he manually logs off his session remains valid. Moreover, user can use his login from several computers - he will have the same preferences selected. Logged users have one more advantage: reusable history of their queries. It is question of time to add more features (e.g., statistical information).

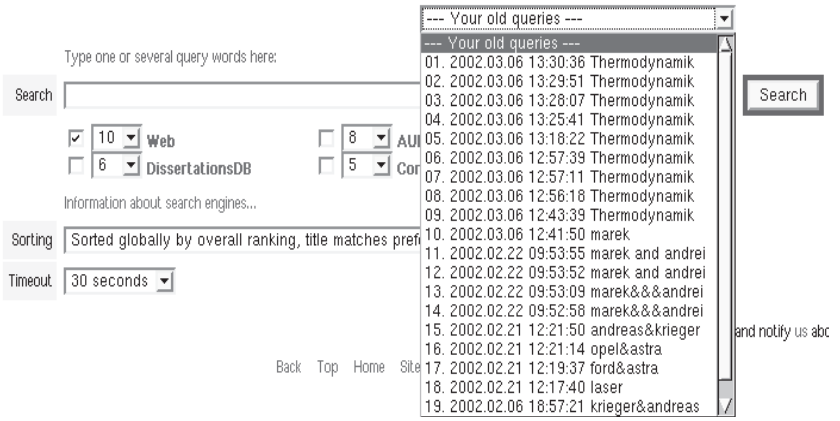


Figure 4: History of queries

4 Conclusion

CRIS users already noticed that nowadays there is a big information overload and vast majority of documents does not follow any structural or formatting standards. Full-text search becomes more and more important over structural search and the importance of search engines raises every day. Metaseach technology can help to work around limitations of search engines.

Although current version of the prototype is fully functional there still exist areas for improvement. Alternative to stateless CGI-based implementation is standalone server application, which can provide query-cashing, smooth, breaking of long listing and slightly faster responses.

The system already provides several benefits:

- Give end user comfortable way to access several research relevant data sources in one run.
- Through the simple yet powerful forward feature of the user interface overcome major drawback of metasearch engines. There are direct links to the search pages of every listed engine. After the search they are changed to links which simulate querying process on particular engine.
- Allows end user to customize behavior of the engine.
- It is general enough to be installed elsewhere and tuned for different set of search engines.

Major drawbacks of metasearch are manual search engine addition by skilled person and requirement for maintenance. This problem will be faced during the next development phrase. Another weakness which has to be noted is that user should be aware of limitations both of search and metasearch.

5 References

Brin, Sergey; Page, Lawrence: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Science Department, Stanford University.
http: //www-db.stanford.edu/pub/papers/google.pdf

Google Features page.
http: //www.google.com/help/features.html

3.Search Engine Watch - the authoritative guide to searching at Internet search engines.
http: //searchenginewatch.com
The Seven Habits of Highly Effective Web Searchers. Peachpit Press, 2001.
http: //beta.peachpit.com/vqs/73401/excerpt.html
MetaCrawler History. Metacrawler Press Center. InfoSpace, 2000.
http: //www.metacrawler.com/press/bg.html
mnoGoSearch - Full Featured Free Web site Open Source Search Engine.
http: //www.mnogosearch.org

6 Contact Information

Marek Andricík
Vienna University of Technology
Gusshausstrasse 28 / E015
A-1040 Vienna, Austria

e-mail: andricik@derpi.tuwien.ac.at

SEAL - a SEmantic portAL with content management functionality

Steffen Staab[1,2](Keynote Speaker), Rudi Studer[1,2,3], York Sure[1], Raphael Volz[1,3]
[1]Institute AIFB, University of Karlsruhe, [2]Ontoprise GmbH, Karlsruhe
[3]Research Group Knowledge Management
FZI - Research Center for Information Technologies at the University of Karlsruhe

Abstract

"OntoWeb" is an European Union IST-funded thematic network for "Ontology-based information exchange for knowledge management and electronic commerce". The corresponding OntoWeb portal constitutes a Web-based research information system that is driven by some of the technologies which it reports about.

In this paper, we present the core methodology underlying the OntoWeb portal, viz. SEAL (SEmantic portAL). In particular, we describe some of the core challenges that SEAL must meet. Because of the distributed nature of research information, SEAL has been developed as a methodology that integrates heterogeneous information from distributed resources. Because of the complexity of the application domain, SEAL is based on ontologies about research information that greatly contribute to the combined goals of low-effort information integration and user-friendly information presentation. Because of the high quality requirements obliged onto the OntoWeb portal, SEAL has been extended with content management functionality supporting portal editors in their process to rule out undesirable content.

1 Introduction

By its very nature, information about scientific research on the Web tends to be distributed, heterogenous, volatile, interrelated, and focused around topics, persons, projects, and organizations. There are plenty of structures on the Web that host research information, e.g. conference web sites, homepages of researchers, project web sites and web information providers (e.g. free providers like http://www.ceur-ws.org or providers-by-fee like http://www.acm.org). Typically, however, these existing structures do not make the context explicit under which their research information may be found. Thus, the provisioning of research information mostly remains idiosyncratic and access to it is at most as good as the information retrieval mechanisms that let you find research information by keyword search.

In order to overcome some of the difficulties associated with accessing research information by keyword search, we have developed an ontology-based approach. An ontology is an explicit specification of shared conceptualizations for a domain of interest. I.e. ontologies make assumptions explicit that a community of people shares about a particular subject. The use of ontologies for information exchange may help to put the many different pieces of research information available into a coherent, re-usable and re-configurable picture. Therefore, the core idea of our SEmantic portAL methodology (SEAL) consists of exploiting ontology structures for specifying the context of particular pieces of research information within one research community - for the purpose of information integration as well as for information presentation.

"OntoWeb" is an European Union IST-funded thematic network that propagates research related to ontology technologies and that, of course, has similar knowledge sharing needs as other

research communities. Therefore, we are currently developing the *OntoWeb Portal* as part of the effort to develop ontology technology and nourish ourselves on it, too.

Developing the portal, we have seen the needs for ontology-based integration of information that we have also met when dealing with developing a Web presentation of our institute (just a small research community; cf. [8]). In addition, new challenges showed up calling for the combination of managing a portal for a highly-distributed community (about 100 partners spread all over Europe and beyond) at low costs and high quality. Thus, SEAL has been extended with content management functionality supporting portal editors in their process to rule out undesirable content.

In the following, we will first sketch the core SEAL approach that we had developed before the OntoWeb portal (Section 2). Then, we describe the scenario of the OntoWeb portal and some of its new requirements (Section 3) - including a revision of the existing architecture (compare with [8]). Thereafter, we describe the process model employed in the OntoWeb portal.

2 SEAL - The core approach

The recent decade has seen a tremendous progress in managing semantically heterogeneous data sources. Core to the semantic reconcilation between the different sources is a rich conceptual model that the various stakeholders agree on, an ontology [4]. The conceptual architecture developed for this purpose now generally consists of a three layer architecture comprising (cf. [12]) (i) heterogeneous *data sources* (e.g., databases, XML, but also data found in HTML tables), (ii) *wrappers* that lift these data sources onto a common data model (e.g. OEM [10] or RDF [7]), (iii) integration modules (*mediators* in the dynamic case) that reconcile the varying semantics of the different data sources. Thus, the complexity of the integration/mediation task could be greatly reduced.

Similarly, in recent years the information system community has successfully strived to reduce the effort for managing complex web sites [1, 2, 5, 9]). Previously ill-structured web site management has been structured with process models, redundancy of data has been avoided by generating it from database systems and web site generation (including management, authoring, business logic and design) has profited from recent, also commercially viable, successes [1]. Again we may recognize that core to these different web site management approaches is a rich conceptual model that allows for accurate and flexible access to data. Similarly, in the hypertext community conceptual models have been explored that im- or explicitly exploit ontologies as underlying structures for hypertext generation and use (e.g. [3]).

SEAL (SEmantic PortAL)[1], our framework to building community web sites, has been developed to use ontologies as key elements for managing community web sites and web portals. The ontology supports queries to multiple sources (a task also supported by semi-structured data models [5]), but beyond that it also includes the intensive use of the schema information itself allowing for automatic generation of navigational views[2] and mixed ontology and content-based presentation. The core idea of SEAL is that Semantic Portals for a community of users that contribute and consume information [11] require web site management and web information integration. In order to reduce engineering and maintenance efforts SEAL uses an ontology for semantic integration of existing data sources as well as for web site management and presentation to the outside world. SEAL exploits the ontology to offer mechanisms for acquiring, structuring and sharing information between human and/or machine agents. Thus, SEAL combines the advantages of the two worlds briefly sketched above.

1 Cf. [8] on the history of SEAL.

2 Examples are navigation hierarchies that appear as has-part-trees or has-subtopic trees in the ontology.

3 OntoWeb Scenario

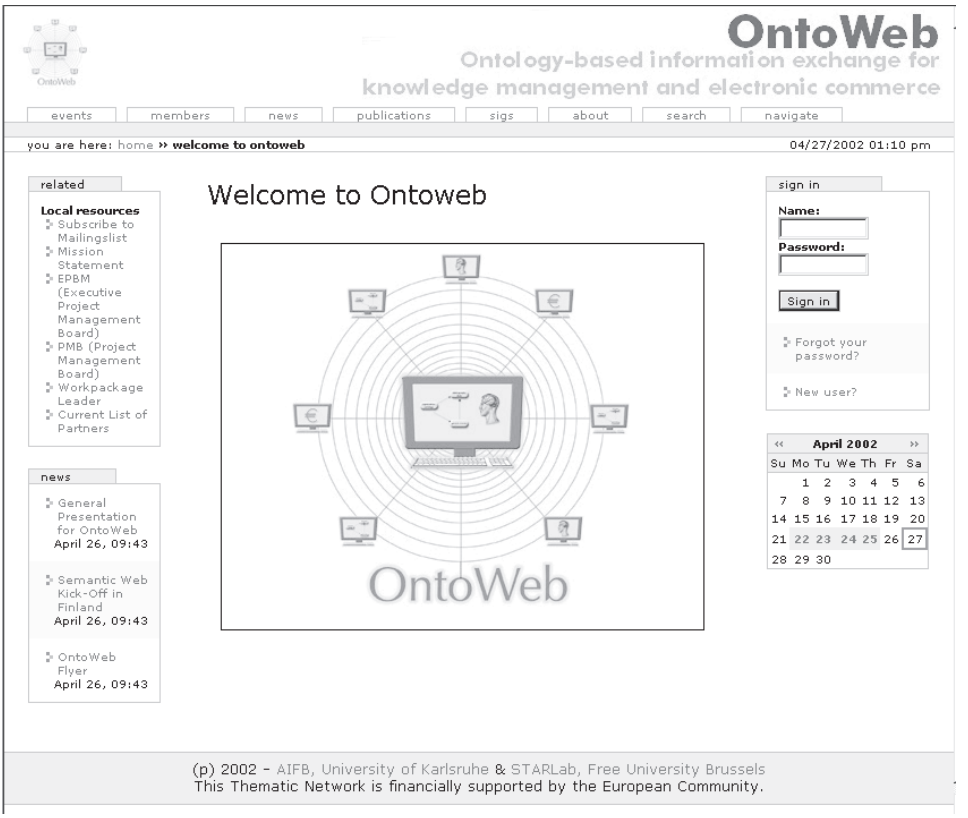


Figure 1: www.ontoweb.org - The OntoWeb portal

The EU thematic network “OntoWeb - Ontology-based information exchange for knowledge management and electronic commerce” aims at bringing together researcher and industrials to “enable the full power ontologies may have to improve information exchange in areas such as: information retrieval, knowledge management, electronic commerce, and bioinformatics. It will also strengthen the European influence on standardization efforts in areas such as web languages (RDF, XML), upper-layer ontologies, and content standards such as catalogues in electronic commerce” (cf. [13]). One of the tasks of the OntoWeb partners is to create a portal for this community serving as a platform for communication between partners and also between partners and other members of the Word Wide Web.

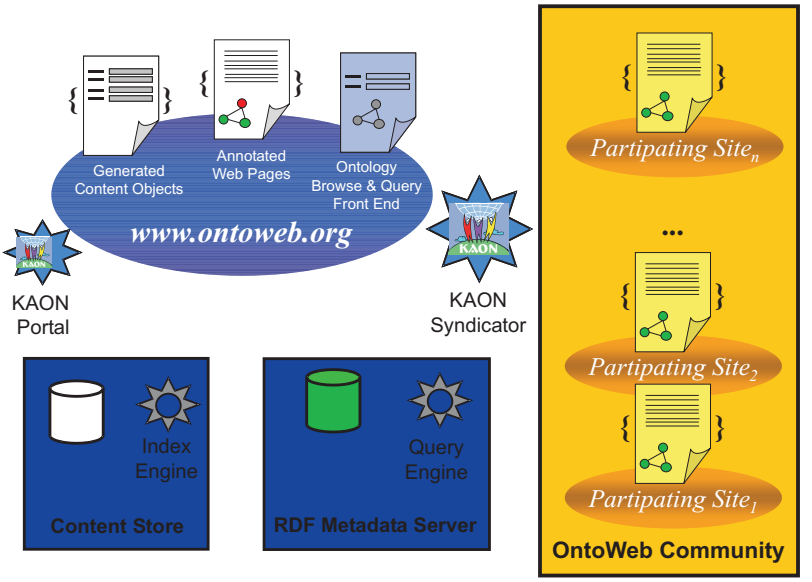


Figure 2: OntoWeb architecture

Portal approach. The OntoWeb portal (cf. Figure 1) is structured according to an ontology which serves as a shared basis for supporting communication between humans and machines. The general goal of our approach is the semi-automatical construction of a community portal using the community's metadata to enable information provision, querying and browsing of the portal. For this purpose we could reuse the framework as explained in Section 2, but we also had to provide new modules for content management resulting in the extended architecture depicted in Figures 2 and 3. The use of core SEAL modules is explained in the following, new ones follow subsequently. The process model is introduced in Section 4.

Use of core SEAL modules

Integration. One of the core challenges when building a data-intensive web site is the integration of heterogeneous information on the WWW. The recent decade has seen a tremendous progress in managing semantically heterogeneous data sources [12, 5]. The general approach we pursue is to "lift" all the different input sources onto a common data model, in our case RDF. Additionally, an ontology acts as a semantic model for the heterogeneous input sources. As mentioned earlier and visualized in our conceptual architecture in Figure 3, we consider different kinds of *Web data sources* as input. However, to a large part the Web consists of static HTML pages, often semi-structured, including tables, lists, etc..

Presentation. Based on the integrated data in the warehouse we define user-dependent *presentation views*. First, we render HTML pages for human agents. Typically *queries for content* of the warehouse define presentation views by selecting content, but also *queries for schema* might be used, e.g. to label table headers. Second, as a contribution to the Semantic Web, our architecture is dedicated to satisfy the needs of software agents and produces machine understandable RDF.

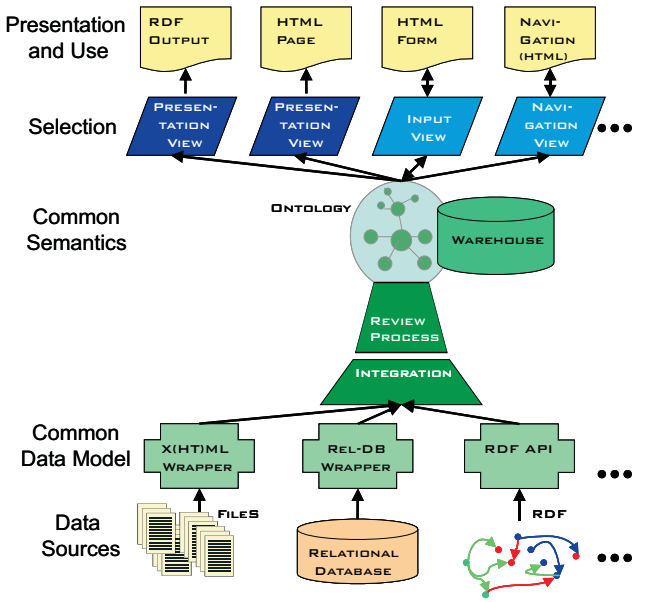


Figure 3: Extended conceptual SEAL architecture

To maintain a portal and keep it alive its content needs to be updated frequently not only by information integration of different sources but also by additional inputs from human experts. The *input view* is defined by *queries to the schema*, i.e. queries to the ontology itself. Similar to [6] we support the knowledge acquisition task by generating forms out of the ontology. The forms capture data according to the ontology in a consistent way which are stored afterwards in the warehouse.

To navigate and browse the warehouse we automatically generate navigational structures, i.e. *navigation views*, by using *combined queries for schema and content*. First, we offer different user views on the ontology by using different types of hierarchies (e.g. *is-a, part-of*) for the creation of top level navigational structures. Second, for each shown part of the ontology the corresponding content in the warehouse is presented. For non-typed content such as documents we take several heuristics to offer navigation: First, all other objects that have the same physical location (folder on the web server) are assumed to be related, as the user put it at that exact location for a certain reason. Second, we use the metadata of the document to find similar objects using the objects' metadata, e.g. objects having the same subject, keywords or author. This provides a simpler way of exploring the content for users that are unfamiliar with the portal.

Implementation

In a nutshell, the upper two levels in the conceptual architecture of SEAL (cf. Figure 3) are implemented as KAON Portal (cf. Figure 2). It generates content objects and provides browsing as well as a query frontend. The replication of distributed knowledge into the metadata server, i.e. the RDF Server, is done by the KAON Syndicator. Please note that only structured data is replicated and not, e.g., documents. The storage consists of (i) a content mangagement system (CMS) that allows for creation and management of documents (but not annotations), (ii) the RDF management system that stores ontologies and associated instance base annotations of the content management system. The OntoWeb Community provides metadata on their websites which are

syndicated with the KAON Syndicator. The workflow component described in the next section is provided by the CMF framework[3], an extension of the Zope web application server[4].

4 Process Model

As mention in Section 3 OntoWeb is an open community. Open communities pose additional constraints since data that is (re)published through the portal could be provided by arbitrary people. In order to guarantee quality of data in such an environment an additional model regulating the publishing process is required, which prevents foreseeable misuses. To support this requirement the established SEAL architecture was extended with a workflow component which regulates the publishing process. In the following we will begin with introducing the concept of a publishing workflow in general. Afterwards we explain how we instantiated this generic component in OntoWeb.

Publishing workflows

A publishing workflow is the series of interactions that should happen to complete the task of publishing data. Business organizations have many kinds of workflow. Our notion of workflow is centered around tasks. Workflows consist of several tasks and several transitions between these tasks. Additionally workflows have the following characteristics: (i) they might involve several people, (ii) they might take a long time, (iii) they vary significantly in organizations and in the computer applications supporting these organizations respectively, (iv) sometimes information must be kept across states, and last but not least, (v) the communication between people must be supported in order to facilitate decision making.

Thus, a workflow component must be customizable. It must support the assignment of tasks to (possibly multiple) individual users. In our architecture these users are grouped into roles. Tasks are represented within a workflow as a set of transitions which cause state changes. Each object in the system is assigned a state, which corresponds to the current position within the workflow and can be used to determine the possible transitions that can validly be applied to the object. This state is persistent supporting the second characteristic mentioned above.

Due to the individuality of workflows within organizations and applications we propose a generic component that supports the creation and customization of several workflows. In fact, each concept in the ontology, which - as you might recall - is used to capture structured data within a portal, can be assigned a different workflow with different states, transitions and task assignments.

As mentioned above, sometimes data is required to be kept across states. For example, envision the process of passing bills in legislature, a bill might be allowed to be revised and resubmitted once it is vetoed, but only if it has been vetoed once. If it is vetoed a second time, it is rejected forever. To model this behavior, the state machine underlying our workflow model needs to keep information that “remembers” the past veto. Thus, variables are attached to objects and used to provide persistent information that transcends states. Within our approach variables also serve the purpose of establishing a simple form of communication between the involved parties. Thus, each transition can attach comments to support the decision made by future actors. Also metadata like the time and initiator of a transition is kept within the system.

3 http://cmf.zope.org/
4 http://www.zope.org/

Workflows in OntoWeb

Figure 4 depicts the default workflow within OntoWeb. There are three states: private, pending, and published. If a user creates a new object[5] the object is in private state. If the user has either a reviewer or a manager role the published state is immediately available through the publish transition. For normal users such a transition is not available, instead the object can only be send for a review leading to the pending state. In the pending state either managers or reviewers can do the transition to the published state (by applying the transition "publish") or retract the object leading back to the private state. The reject transition deletes the object completely. When an object is in the private state, only the user who created it and users with manager roles can view and change it. Once an object is in published state the modification by the user who created it resets the object into pending state, thus the modification must be reviewed again. This does not apply to modifications by site managers.

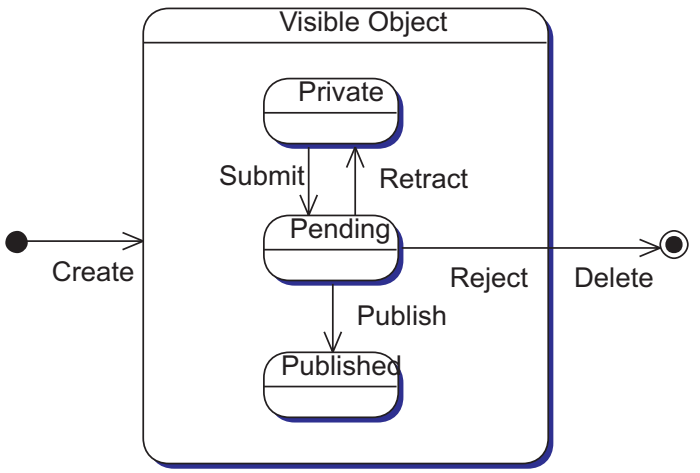


Figure 4: SEAL Publishing workflow

5 Related work

Given aforementioned difficulties with managing complex Web content, several papers tried to facilitate database technology to simplify the creation and maintenance of data-intensive web-sites. Systems, such as ARANEUS [9] and AutoWeb [2], also take a declarative approach. In contrast to SEAL that relies on standard Semantic Web technologies these systems introduce their own data models and query languages, although all approaches share the idea to provide high-level descriptions of web-sites by distinct orthogonal dimensions.

The idea of leveraging mediation technologies for the acquisition of data is also found in approaches like Strudel [5] and Tiramisu [1], they propose a separation according to the aforementioned task profiles as well. Strudel does not concern the aspects of site maintenance and personalization. It is actually only an implementation tool, not a management system.

5 (currently only within the portal, the content syndicated from other OntoWeb member web sites and within the databases is "trusted". We assume that this kind of data already went through some kind of review.

From our point of view the SEAL framework and it's application as the OntoWeb portal is rather unique with respect to the collection of methods used and the functionality provided.

6 Conclusion

In this paper we have shown the application of our comprehensive framework SEAL for building "SEmantic portALs". In particular, we have focused on three issues. First, we have described the general architecture of the SEAL framework. Second, we have presented our real world case study, the OntoWeb portal. Third, to meet the requirements of the OntoWeb portal, we extended our initial conceptual architecture SEAL by publishing workflows to make user focussed access to the OntoWeb portal maintainable.

For the future, we see a number of new important topics appearing on the horizon. For instance, we consider approaches for ontology learning in order to semi-automatically adapt to changes in the world and to facilitate the engineering of ontologies. Currently, we work on providing intelligent means for providing semantic information, i.e. we elaborate on a semantic annotation framework that balances between manual provisioning from legacy texts (e.g. web pages) and information extraction.

Finally, we envision that once semantic web sites are widely available, their automatic exploitation may be brought to new levels. Semantic web mining considers the level of mining web site structures, web site content, and web site usage on a semantic rather than at a syntactic level yielding new possibilities, e.g. for intelligent navigation, personalization, or summarization, to name but a few objectives for semantic web sites.

7 Acknowledgements

We thank our colleagues at StarLab, VU Brussels headed by of Robert Meersmann and at Institute AIFB, University of Karlsruhe, in particular Daniel Oberle, for fruitful discussions on the work reported here. This work has been funded under the EU IST-2001-29243 project "OntoWeb" and the EU IST-2001-33052 project "WonderWeb".

8 References

[1] C. R. Anderson, A. Y. Levy, and D. S. Weld. Declarative web site management with tiramisu. In ACM SIGMOD Workshop on the Web and Databases - WebDB99, pages 19-24, 1999.

[2] S. Ceri, P. Fraternali, and A. Bongio. Web modeling language (WebML): a modeling language for designing web sites. In WWW9 Conference, Amsterdam, May 2000, 2000.

[3] M. Crampes and S. Ranwez. Ontology-supported and ontology-driven conceptual navigation on the world wide web. In Proceedings of the 11th ACM Conference on Hypertext and Hypermedia, May 30 - June 3, 2000, San Antonio, TX, USA, pages 191-199. ACM Press, 2000.

[4] D. Fensel, J. Angele, S. Decker, M. Erdmann, H.-P. Schnurr, R. Studer, and A. Witt. Lessons learned from applying AI to the web. International Journal of Cooperative Information Systems, 9(4):361-382, 2000.

[5] M. F. Fernandez, D. Florescu, A. Y. Levy, and D. Suciu. Declarative specification of web sites with Strudel. VLDB Journal, 9(1):38-55, 2000.

[6] E. Grosso, H. Eriksson, R. W. Fergerson, S. W. Tu, and M. M. Musen. Knowledge modeling at the millennium: the design and evolution of PROTEGE-2000. In Proceedings of the 12th International Workshop on Knowledge Acquisition, Modeling and Mangement (KAW-99), Banff, Canada, October 1999.

[7] O. Lassila and R. Swick. Resource Description Framework (RDF). Model and syntax specification. Technical report, W3C, 1999. http://www.w3.org/TR/REC-rdf-syntax.

[8] A. Maedche, S. Staab, R. Studer, Y. Sure, and R. Volz. Seal - tying up information integration and web site management by ontologies. IEEE Data Engineering Bulletin, 25(1):10-17, March 2002.

[9] G. Mecca, P. Merialdo, P. Atzeni, and V. Crescenzi. The (short) Araneus guide to web-site development. In Second Intern. Workshop on the Web and Databases (WebDB'99) in conjunction with SIGMOD'99, May 1999.
[10] Y. Papakonstantinou, H. Garcia-Molina, and J. Widom. Object exchange across heterogeneous information sources. In Proceedings of the IEEE International Conference on Data Engineering, Taipei, Taiwan, March 1995, pages 251-260, 1995.
[11] S. Staab, J. Angele, S. Decker, M. Erdmann, A. Hotho, A. Maedche, H.-P. Schnurr, R. Studer, and Y. Sure. Semantic community web portals. In WWW9 / Computer Networks (Special Issue: WWW9 - Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, May, 15-19, 2000), volume 33, pages 473-491. Elsevier, 2000.
[12] G. Wiederhold and M. Genesereth. The conceptual basis for mediation services. IEEE Expert, 12(5):38-47, Sep.-Oct. 1997.
[13] Thematic Network EU IST-2000-25056 OntoWeb: Annex 1 - "Description of Work". Technical report, Information Societies Technology (IST) Programme, February 11 2001.

9 Contact Information

Steffen Staab
Institute AIFB
University of Karlsruhe
76128 Karlsruhe
Germany

e-mail: staab@aifb.uni-karlsruhe.de
http://www.aifb.uni-karlsruhe.de/WBS

Discovery of patterns of scientific and technological development and knowledge transfer

Anthony F.J. van Raan[1] (Keynote Speaker), Ed C.M. Noyons
Centre for Science and Technology Studies (CWTS)
University of Leiden

Abstract

This paper addresses a bibliometric methodology to discover the structure of the scientific 'landscape' in order to gain detailed insight into the development of R&D fields, their interaction, and the transfer of knowledge between them. This methodology is appropriate to visualize the position of R&D activities in relation to interdisciplinary R&D developments, and particularly in relation to socio-economic problems. Furthermore, it allows the identification of the major actors. It even provides the possibility of foresight. We describe a first approach to apply bibliometric mapping as an instrument to investigate characteristics of knowledge transfer.

1 Introduction

In this paper we discuss the creation of 'maps of science' with help of advanced bibliometric methods. This 'bibliometric cartography' can be seen as a specific type of data-mining, applied to large amounts of scientific publications. As an example we describe the mapping of the field neuroscience, one of the largest and fast growing fields in the life sciences. The number of publications covered by this database is about 80,000 per year, the period covered is 1995-1998. Current research is going on to update the mapping for the years 1999-2002. This paper addresses the main lines of the methodology and its application in the study of knowledge transfer.

2 Basic Principles of Bibliometric Mapping

Each year about a million scientific articles are published. How to keep track of all these developments, particularly the relations with other fields? Are there *cognitive structures* 'hidden' in this mass of published knowledge, at a 'meta-level'?

A research field can be defined ('delineated') by various approaches: on the basis of classification codes, selected terms in a (discipline-) specific database, selected sets of journals, a database of field-specific publications, or any combination of these approaches. In this paper we take *neuroscience*, a large field within the life sciences, as an example.

We delineate this field with the Neuroscience Citation Index of the Institute for Scientific Information (ISI)[2]. This is an appropriate database that covers over 80,000 publications annually. We collected the titles and abstracts of all these publications, for a series of successive years (1995-1998, we are currently adding 1999-2002), thus operating on several hundreds of thousands publications. With a specific computer-linguistic algorithm we parse the abstracts *of all*

1 Corresponding author, vanraan@cwts.leidenuniv.nl

2 The Institute for Scientific Information in Philadelphia, the publisher of the Science Citation Index (SCI) and all other related citation indexes.

these publications. These completely automated grammatical procedures yield all nouns and noun-phrases (standardized) that are present in the entire set of publication abstracts.

An additional algorithm creates a frequency-list of these many thousands of parsed nouns and noun-phrases while filtering out general, trivial words (Noyons 1999). We consider the most frequent nouns/noun phrases as the most characteristic concepts of the field (this can be 100 to 1,000 concepts, say N concepts).

The next step is to *encode* each of the yearly 80,000 publications with these concepts. In fact this code is a binary string (yes/no) indicating which of the N concepts is present in title or abstract. This encoding is as it were the 'genetic code' of a publication. Like in genetic algorithms, we now compare the encoding of each publication with that of any other publication. So we calculate 'genetic code similarity' (here: *concept-similarity*) of all 80,000 publications pair-wise. The more concepts two publications have in common, the more these publications are related on the basis of concept-similarity and thus can be regarded as belonging to the same subfield, research theme or research specialty. In a biological metaphor: the more specific DNA-elements two living beings have in common, the more they are related. Above a certain similarity threshold, they will belong to a particular species.

The above procedure allows clustering of *information carriers* -the publications- on the basis of similarity in *information elements* - the concepts ('co-publication' analysis). Alternatively, the more specific concepts are mentioned together in different publications, the more these concepts are related. Thus, information elements are clustered ('co-concept' analysis). Both approaches, the co-publication and the co-concept analysis are related by simple matrix algebra rules. In practice, the co-concept approach (Callon et al 1983; Noyons and Van Raan 1998) is most suited for science mapping, i.e., the 'organization of science according to concepts'.

Intermezzo: For a super market 'client similarity' on the basis of shopping lists can be translated into a clustering of either the clients (information carriers, where the information elements are the products on their shopping lists) or of the products. Both approaches are important: the first gives insight into groups of clients (young, old, male, female, different ethnic groups, etc.), and the second is important in the organization of the super market.

In main lines the clustering procedure is as follows. We first construct a matrix composed by co-occurrences of the N concepts in the set of publications for a specific period of time, e.g., 1997-1998. We normalize this 'raw co-occurrence' matrix in such a way that the similarity of concepts is no longer based on the pair-wise co-occurrences, but on the co-occurrence 'profiles' of the two concepts in relation to all other concepts.

This similarity matrix is input for a cluster analysis. In most cases, we use a standard hierarchical agglomerative cluster algorithm including statistical criteria to find an optimal number of clusters. The identified clusters of concepts represent 'subfields'. These subfields are labeled with the four most frequent concepts in a cluster.

The clusters resulting from the mapping procedure are tested for internal coherence. We calculated the average linkage between all concept-pairs within a cluster, and the standard deviation. This internal coherence measure indicates the robustness of the identified cluster. We refer to Noyons and Van Raan (2002).

Each subfield represents a sub-set of publications on the basis of the above discussed concept-similarity profiles. If any of the concepts is in a publication, this publication will be attached to the relevant subfield. Thus, publications may be attached to more than one subfield. The overlap between subfields in terms of joint publications is used to compile a further co-occurrence matrix, now based on subfield publication overlap. This matrix is used to calculate a similarity measure of subfields by comparing their co-occurrence profile with others.

To construct a map of the field, the subfields (clusters) are positioned by multidimensional scaling. Thus, subfields with a high similarity (with a similar 'cognitive orientation' within the field) are positioned in each other's vicinity, and subfields with low similarity are distant from

each other. The size of a subfield (represented by the surface of a circle) indicates the share of publications in relation to the field as a whole. Particularly strong relations between two individual subfields are indicated by a connecting line (see discussion in Section 4).

A similar mapping procedure can be applied to documents other than publications, for instance patents. Thus, maps of technology can be constructed. In this paper we confine ourselves to the mapping of neuroscience.

3 Maps as Analytical Instrument

The above procedure generates the *cognitive structure* of the field neuroscience. As discussed above, it is entirely based on the total of relations between all publications. The fascinating point is that the discovered structure is not the result of any pre-arranged classification system or whatsoever. Nobody prescribes this structure. It emerges solely from the internal relations through concept-similarities of the whole ensemble of publications together. In other words, what we make visible by our mathematical methods, is a *self-organized cognitive ecology of science*.

A detailed discussion of science maps is given by Noyons (1999). Our mapping procedure depends partly on expert input. Special internet-facilities enable experts to comment on the concepts used to generate the structure of the field.

The maps are put in a digital form on a (protected, in cases of confidentiality) part of the CWTS website[3]. Thus we make the maps easily accessible for users in order to explore the field or to validate the results[4]. We also provide information 'behind' the map (actors, and their output and impact indicators) by an interface that can be used via standard graphical internet browsers (e.g., MS Internet Explorer and Netscape Communicator).

This advanced bibliometric mapping has many interesting analytical potentials. First, it visualizes the landscape of a scientific field 'embedded in its surroundings', i.e., in its interdisciplinary relations. We found that a major part of the landscape relates to socio-economic problems (Van Raan, 2001). For neuroscience obvious examples are: Altzheimer disease, Parkinson's disease, aging, stroke. Second, by making these maps for a series of years, we are able to observe trends and changes in structure (see our website for examples). Extrapolation of these trends enables foresight of near-future developments.

Third, bibliometric maps allow localization of major actors. Thus we are creating a strategic map: who is where in science, and, more precisely, what is the position of these actors in terms of interdisciplinary relations of the different fields? In addition to that, we may assess an actor's scientific influence ('impact') in the field by applying standard CWTS bibliometric analysis. In this way, the two major pillars of bibliometric methodology, concept-based mapping and citation-based impact analysis, are combined. This combined approach is very useful in the identification of scientific 'centers of excellence'. We will not further address this search for excellence and refer to Van Raan (2000). However, citation analysis is crucial in this paper, not as an instrument to assess impact, but to identify communication patterns.

In this paper we focus on the application of bibliometric mapping in the analysis of knowledge transfer. This is a first and still experimental approach, to begin with an analysis on a not-too-large scale (science as a whole), but within a major field, in this case neuroscience.

In order to develop this map-based knowledge transfer analysis more systematically, we distinguish two types of intra-field (i.e., between subfields) relations: (1) *conceptual linkages* (which is the basis if the map structure, as explained above), and (2) *communication linkages*,

3 http://www.cwts.leidenuniv.nl

4 An important aspect of the mapping methodology is the retrieval rate: which part of the publications covered by the dataset used as a starting point for the mapping procedure, can be found back in the map? For this field, neuroscience, we reach a retrieval rate of at least 80%.

based on the extent to which publications in a specific subfield cite publications in other subfields.

We hypothesize that these two linkage modalities are basic elements of scientific development and that the bibliometric mapping allows the visualization and further analysis of the patterns involved. Furthermore, we regard conceptual linkage as the source of potential knowledge transfer, and communication linkages as the realization of knowledge transfer. Thus, comparison of these two types of linkage may reveal differences in potential versus 'already existing' knowledge transfer.

4 Results

The result of the neuroscience mapping is shown in Figure 1a. The clusters represent subfields and research themes according to List 1 in which the (at most) four most frequent concepts of the cluster are given to label the cluster. In addition, we have indicated as an example the relatively strongest *conceptual linkages* between cluster 10, brain infarction research (stroke) with other research subfields of neuroscience, particularly subfields 3 (Etiology), 11 (Subarachnoid hemorrhage), 15 (Magnetic resonance imaging, MRI), and 21 (Ischemia).

In Figure 1b we present the relatively strongest *communication linkages* (citation-based) between brain infarction research and other research subfields of neurosciences, and now these linkages are particularly with subfield 3 (Etiology) and 20 (Animal model).

We observe that brain infarction research is an example of reasonably similar but still different conceptual (words) and communication (citations) linkages, as is illustrated by comparison of Fig. 1a and Fig. 1b. For instance, we see more conceptual linkages than communication linkages. A first step to explain these findings is the analysis of the research fields involved in the different subfields. Although we deal with subfields of neuroscience, publications in these neuroscience subfields may belong to other fields than neuroscience only. For instance, it is clear that brain infarction research will involve the field of cardiovascular system. This means, that publications on brain infarction research, may appear in cardiovascular journals.

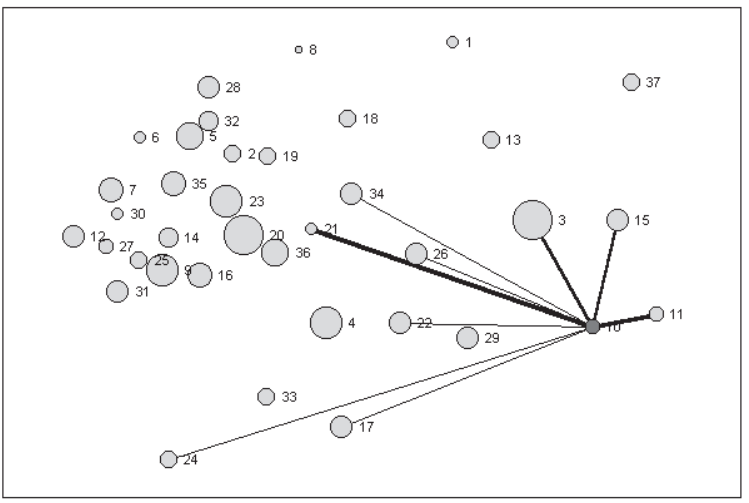


Figure 1a:Conceptual linkages between brain infarction research with other subfields of neuroscience. Two-dimensional representation based on the similarities between identified clusters of concepts (subfields). For the list of subfields with corresponding number we refer to List 1. The size of the subfields represents the number of publications in a specific subfield.

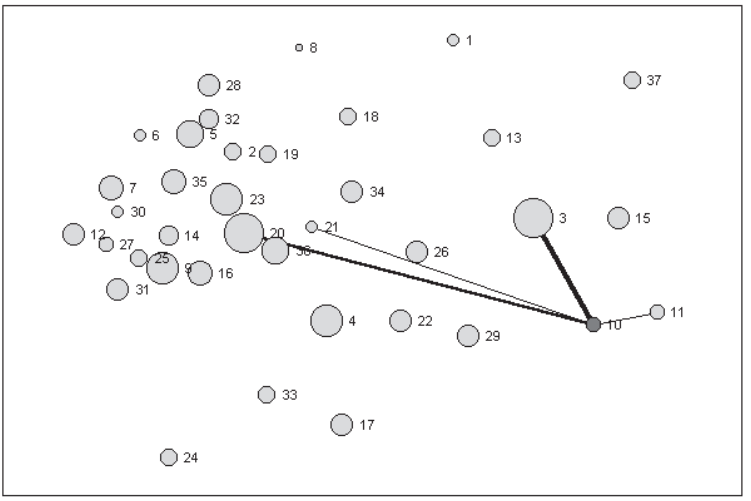


Figure 1b: Communication linkages between brain infarction research and other subfields of neuroscience.

List 1: Most frequent concepts in the neuroscience subfields

1 Multiple sclerosis / myelin basic protein / experimental autoimmune encephalomyelitis / lewis rat
2 Astrocytes / Glial cell / TNF Alpha / acidic protein
3 Etiology / differential diagnosis / neurological deficit / spinal cord injury
4 Schizophrenia / Ethanol / Alcohol / normal control
5 Retina / skeletal muscle / neuronal cell / molecular mechanism
6 NGF / nerve growth / neurotrophic factor / pc12 cell
7 Ca2+ / inhibitory effect/ protein kinase
8 Amyotrophic lateral sclerosis / motor neuron disease
9 h 3 / Dopamine / Antagonist / Agonist
10 Stroke / ischemic stroke / stroke patient / cerebral infarction
11 subarachnoid hemorrhage / middle cerebral artery / internal carotid artery
12 Peptide / Hormone / Secretion / Male Rat
13 CSF / HIV / AIDS / human immunodeficiency virus
14 Glutamate / NMDA / NMDA Receptor / glutamate receptor
15 MRI/ computed tomography / Functional MRI
16 Acetylcholine / Neurotransmitter / Uptake / Norepinephrine
17 Depression / Placebo / Anxiety
18 Alzheimers Disease / a beta / amyloid precursor protein / beta amyloid
19 Apoptosis / cell death / neuronal death / neurodegenerative disease
20 Animal model / electrical stimulation / Fiber / Pathophysiology
21 Ischemia / cerebral ischemia / neuronal damage / neuroprotective effect
22 Dementia / Aging / cognitive function / cognitive impairment
23 Axon / Immunoreactivity / Immunohistochemistry / Adult Rat
24 Hart rate / blood pressure / sympathetic nervous system / heart rate variability
25 Gaba / synaptic transmission / gamma aminobutyric acid / synaptic plasticity
26 PET / cerebral blood flow / white matter
27 Hypothalamus / c fos / paraventricular nucleus / locus coeruleus
28 Gene/CDNA / polymerase chain reaction / expression pattern
29 Seizure / EEG / Epilepsy / temporal lobe
30 nitric oxide synthase / l arginine / neuronal nitric oxide synthase
31 Stress / substance p / neuropeptide y / tyrosine hydroxylase

32 spinal cord / Peripheral nerve / sensory neuron / dorsal root ganglion
33 Memory / Learning / working memory / memory impairment
34 Pathogenesis / Parkinsons Disease / basal ganglion / oxidative stress
35 MRNA / rat brain / gene expression / olfactory bulb
36 Hippocampus / Cortex / Cerebellum / Striatum
37 Tumor / Brain Tumor / radiation therapy / primitive neuroectodermal tumor

For the classification of journals into fields we use the ISI classification system. Thus, on the basis of the journals in which the publications of the different subfields have been published, we make a frequency list of the research fields involved.

Below we present the ranking of the first ten research fields involved in brain infarction research with the number of publications in 1997-1998. These results clearly show the interdisciplinary 'make up' of the different subfields. The infarction focus of subfield 10 is clearly visible in the fields immediately following both general neuro-fields (neuroscience and clinical neurology): cardiovascular system and vascular diseases.

Sf 10	*Brain infarction research*
1305	NEUROSCIENCE
1042	CLINICAL NEUROLOGY
710	CARDIOVASCULAR SYSTEM
575	VASCULAR DISEASES
206	MEDICINE, GEN. & INTERNAL
184	SURGERY
111	PHARMACOLOGY & PHARMACY
104	RADIOLOGY & NUCL MEDICINE
101	PSYCHIATRY
86	HEMATOLOGY

As discussed above, the strongest *conceptual* relations of subfield (Sf) 10 (Brain infarction research) are with subfields 3 (Etiology), 11 (Subarachnoid hemorrhage), 15 (MRI) and 21 (Ischemia). An analysis of the research fields involved in these subfields gives the following results:

Sf 3	*Etiology*	*Sf 11*	*Subarachnoid hemorrhage*
5963	NEUROSCIENCE	1491	NEUROSCIENCE
3577	CLINICAL NEUROLOGY	1056	CLINICAL NEUROLOGY
1616	SURGERY	650	SURGERY
1082	OPHTHALMOLOGY	382	CARDIOVASCULAR SYSTEM
865	PEDIATRICS	307	VASCULAR DISEASES
856	MEDICINE, GEN. & INTERNAL	305	RADIOLOGY & NUCLEAR
786	PSYCHIATRY	91	MEDICINE, GEN. & INTERNAL
653	OTORHINOLARYNGOLOGY	78	ANESTHESIOLOGY
627	RADIOLOGY & NUCLEAR MEDICINE	62	PEDIATRICS
575	ANESTHESIOLOGY	58	PHARMACOLOGY & PHARMACY

Sf 15	*Magnetic resonance imaging (MRI)*	*Sf 21*	*Ischemia*
2404	NEUROSCIENCE	1409	NEUROSCIENCE
1925	CLINICAL NEUROLOGY	394	CLINICAL NEUROLOGY
886	RADIOLOGY & NUCLEAR MEDICINE	214	CARDIOVASCULAR SYSTEM
683	SURGERY	175	PHARMACOLOGY & PHARMACY
329	PSYCHIATRY	157	VASCULAR DISEASES
328	PEDIATRICS	155	BIOCHEMISTRY & MOLEC BIOLOGY
213	MEDICINE, GEN. & INTERNAL	153	ENDOCRINOLOGY & METABOLISM
134	CARDIOVASCULAR SYSTEM	147	HEMATOLOGY
121	ONCOLOGY	130	SURGERY
107	VASCULAR DISEASES	63	ANESTHESIOLOGY

The strongest *communication* linkages of subfield 10 are with subfields 3 and 20. The fields involved in subfield 3 (Etiology) are already presented above, for subfield 20 (Animal model) we find:

Sf 20	*Animal model*
9000	NEUROSCIENCE
2815	PHARMACOLOGY & PHARMACY
1655	CLINICAL NEUROLOGY
1437	BIOCHEMISTRY & MOL. BIOLOGY
1254	PHYSIOLOGY
763	PSYCHIATRY
488	CELL BIOLOGY
427	PSYCHOLOGY
422	ENDOCRINOLOGY & METABOLISM
393	SURGERY

We observe a preference for the conceptual linkages in both (cardio)vascular research (subfields 11, 15, 21) as well as surgery (subfields 3, 11, 15), whereas at the communication side we see less cardiovascular research and more orientation toward surgery (subfield 3) and toward the quite general subfield 20, Animal model.

This first observation suggests that formal communication (in terms of citation patterns) is more inclined to clinical practice, and less to the more scientifically based interaction with other medical fields such as cardiovascular research. It is more ‘general’, and less specific.

Similar analysis of the conceptual and communication linkages of other subfields tends to confirm that the communication (citation-based) linkages are more clinically and more generally oriented than the more scientifically oriented conceptual linkages. Such ‘mismatches’ of the two types of knowledge linkages are very interesting as they may point to (significant) differences in ‘potential knowledge transfer’ and ‘realized knowledge transfer’.

But our above observations represent first and preliminary results, we are currently investigating this phenomenon more systematically and will report on further results during the conference.

5 Concluding Remarks

Bibliometric mapping is a powerful methodology to visualize the cognitive landscape of a research field. In this paper we present our current work on a method to analyse knowledge transfer by distinguishing two types of intra-field research relations, conceptual linkages and communication linkages. On the basis of first results of this approach we are convinced that further systematic work along these lines will lead to a better understanding of the process of knowledge transfer.

6 References

Callon, M., J.-P. Courtial, W.A. Turner, and S. Bauin (1983). From translations to problematic networks: an introduction to co-word analysis. *Social Science Information* 22, 191-235.

Noyons, E.C.M. and A.F.J. van Raan (1998). Monitoring Scientific Developments from a Dynamic Perspective: Self-Organized Structuring to Map Neural Network Research. *Journal of the American Society for Information Science (JASIS),* 49, 68-81.

Noyons, E.C.M. (1999), *Bibliometric mapping as a science policy and research management tool*. Thesis Leiden University. Leiden: DSWO Press.

Noyons, E.C.M., M. Luwel and H.F. Moed (1999). Combining Mapping and Citation Analysis for Evaluative Bibliometric Purpose. A Bibliometric Study on Recent Development in Micro-Electronics. *J. of the American Society for Information Science (JASIS),* 50, 115-131.

Noyons, E.C.M. and A.F.J. van Raan (2002). Science mapping from publications. In: *Dealing with the data flood*, J. Meij (ed.), The Hague: STT/Beweton, ISBN 90-804496-6-0 (also available on CD-Rom), p. 64-72.

Van Raan, A.F.J. (2000). The Pandora's Box of Citation Analysis: Measuring Scientific Excellence, the Last Evil? In: B. Cronin and H. Barsky Atkins (eds.), *The Web of Knowledge. A Festschrift in Honor of Eugene Garfield*, p. 301-319. Medford (New Jersey): ASIS Monograph Series, 2000. ISBN 1-57387-099 -4.

Van , A.F.J. (2001). Mapping R&D related to socio-economic problems. In: *Proceedings of the Quality of Life Impact Workshop of the European Commission*, June 2000. Brussels: European Commission.

7 Contact Information

Anthony F.J. van Raan
Centre for Science and Technology Studies (CWTS)
University of Leiden
Wassenaarseweg 52
P.O. Box 9555
2300 RB Leiden

e-mail: vanraan@cwts.leidenuniv.nl

Development of a central Knowledge Transfer Platform in a highly decentralised environment

Dominik Ulmer, Beat Birkenmeier
Rat der Eidgenössischen Technischen Hochschulen, Zürich
Business Results GmbH, Zürich

Summary

This paper explains the development of the new "Knowledge Information and Sharing System" (KISS) for Swiss polytechnics, research institutions and universities initiated by the ETH-Board (Rat der Eidgenössischen Technischen Hochschulen). The system aims a simple presentation of information about exploitable knowledge on a common electronic user interface for all institutions and, at the same time, quick and user-friendly access to this information by the interested public. Apart from the description of the development process the application is presented. Contrary to conventional research databases it offers not only plain research data but also usage-oriented information such as the benefit of possible applications of the knowledge in products or services, patent data and links to websites containing additional information. KISS is accessible on the Internet via www.knowledgetransfer.ch.

Keywords: Knowledge Management, Knowledge Transfer, Knowledge Database, Research Database, Knowledge Information System

1 Introduction

1.1 Initiation

The ETH-Board is co-ordinating the activities of the six autonomous Swiss research and education institutions affiliated to the ETH-Domain:

- Eidgenössische Technische Hochschule Zürich (ETHZ),
- Ecole Politechnique Fédérale de Lausanne (EPFL)
- Paul Scherrer Institut (PSI)
- Eidg. Forschungsanstalt für Wald, Schnee und Landschaft (WSL)
- Eidg. Materialprüfungs- und Forschungsanstalt (EMPA)
- Eidg. Anstalt für Wasserversorgung, Abwasserreinigung und Gewässerschutz (EAWAG)

In each of these six institutions new knowledge is constantly developed. Due to new objectives given by the Swiss government, the transfer of this knowledge into industry and society should increase (Schweizerischer Bundesrat, 1999, p.6).

For this reason the ETH-Board has initiated the development of a new interactive platform.

1.2 Project aim

The project was given the working title KISS, which stands for "**K**nowledge **I**nformation and **S**haring **S**ystem". KISS is based on the following intentions:

To establish a platform which allows

- a simple presentation of information about exploitable knowledge on a common electronic user interface for all institutions and, at the same time

- quick and user-friendly access to this information by the interested public.

With respect to the conception of the system a set of basic conditions had to be taken into account:

- a high degree of de-centralisation
- the diversity of research and teaching
- the variety of the IT environments in the different institutions.

2 Project organisation

2.1 Project phases

It was chosen a process organisation according to a six-stage phase concept for the development of application software similar to Becker et al. (see ; Becker et. al., 1997, p. 246).

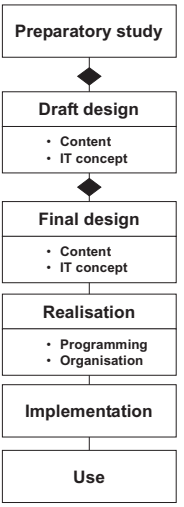


Figure 1: Overview project phases

Even if the project aim was quite clear at the beginning, a preparatory study was made in order to get an even better understanding of the problem. It included an analysis of existing systems or parts of systems already used by the institutions and supposed to contain information about the knowledge available at the respective institution. This analysis made quite clear that there were two main gaps:

1. Existing information media like research databases and websites of research teams qualify only very limited for the advancement of corresponding contacts between the research institutions and the public because of restricted access as well as lack of lucidity of the information.
2. Due to the lack of a common base information comes in different format and quality which makes it nearly impossible to get a concise and comparable overview.

Based on these insights it was possible to give a rough sketch of what elements the future system should consist. This resulted in a decision of the steering committee to go on and to elaborate a draft design.

This draft design was divided into two parts: On the one side there was a content-oriented part containing elements like a more exact definition of the characteristics of the data and an estimation of the number of records. On the other hand there was an IT-oriented part defining the key elements of the system's architecture. Consequently, both parts were compiled into a specification which made it possible to make a cost-benefit equation of different concepts.

After another "go"-decision by the steering committee, the final design was worked out. Therefore detailed definitions of data organisation, data editing and processing, search functionalities and information output were set.

In the phase of realisation, the applications were programmed in the development environment of the IT-partners. By choosing one institution as a "test bed", it was possible to test it with real data and organisational structures.

The phase of implementation was defined as the process of migrating the applications from the development environment to the equipment of the organsation hosting the productive system.

For the use of the systems by the institutions was chosen a "one-after-the-other-system" in order to guarantee a proper education of the staff

2.2 Organisation structure

Within the evaluation of a suitable organisation structure it had to be considered that the interests of the single institutions were duly accommodated during the development of the common platform. A dynamic project structure depending on phases was chosen. The preparatory study and the first draft were elaborated by an external project team, whereas the interests of the institutions were taken care of by representatives in the steering committee. Beginning with the phase of the final design a completely new project organisation structure was established (see):

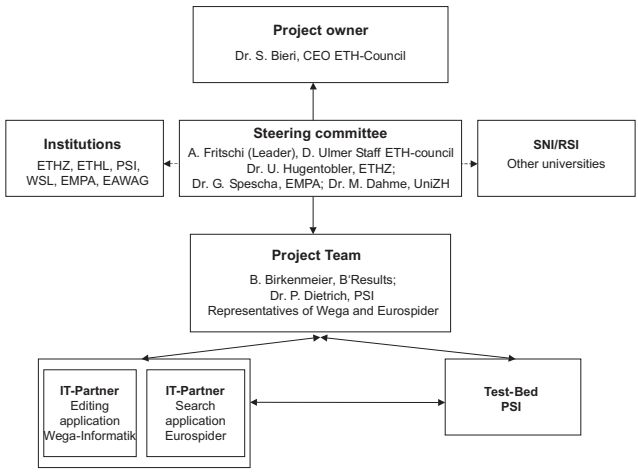


Figure 2: Project organisation structure

It was built up a core project team consisting of an external project manager, a principal's representative as well as a delegate of the institutions and of particular representatives of the companies charged with the development of the application during the different stages of processing. One of the six institutions served as a sample environment in terms of a prototype. During each stage of development all other institutions were constantly informed about the work's progress and consulted before important decisions.

3 Characteristics of the system

Within the conception of the solution the high level of organisational decentralisation, the different contents of research and education as well as the peripheral responsibilities and the involved heterogeneity of the existing IT infrastructure had to be taken into account.

3.1 Applications

Based on that a concept was elaborated intending that the data model and the application as well as the preparation of resources would be realised and run centrally. The inputs into the system will be made directly by the peripheral research groups whereby, with regard to the connection of the contents, it is possible to assign different user rights. In the overview, the solution consists of the following applications and interfaces (see):

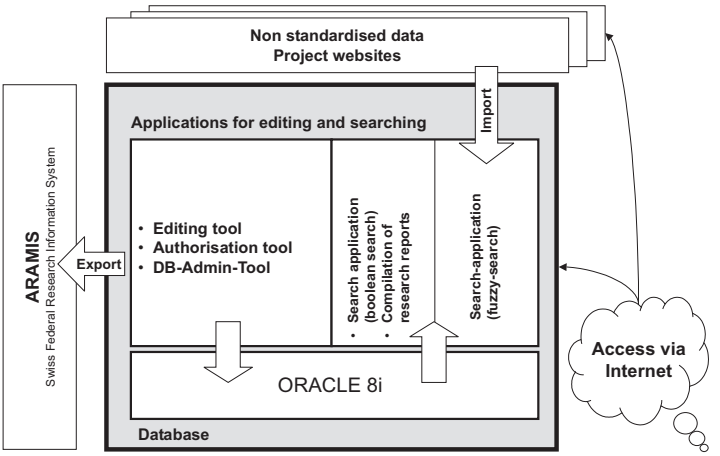


Figure 3: Application overview

Only the meta-data about the projects will be stored in the central database, whereas for complete information there will be links to the more detailed full-text documentations on websites of institutes and research groups. Information about existing knowledge is based on data to research projects such as: Project title, -number

- Date of project start/project end
- Abstract
- Keywords
- Research field
- Scientific target
- Benefit of possible applications
- Patent data/Product data
- Links to URLs
- Personal data of researchers involved in the project
- Possibilities for participation

Contrary to conventional research databases there are also relevant data such as information about the benefit of knowledge in possible applications or patent/product information.

The core of the search application is a full-text search system which checks both the content of the input fields and the linked websites. Besides there is a Boolean standard search available.

As federal institutions are obliged to enter research data to the Swiss Research Information System ARAMIS an interface was created, which allows it to easily export relevant data from KISS to ARAMIS.

3.2 Architecture

The system is based on a so called Three-Tier-Architecture:

- A relational database (ORACLE 8i) with the structured meta-data and links to unstructured data as the first tier.
- An application server which provides the Java based application and represents the intermediate layer.
- The user's standard browser as the last tier.

In order to communicate with other programs and data management systems there are PDF and XML interfaces available. Both the data recording and the search for information take place on the internet using a standard web browser so that no further installation on the user's terminal will have to be taken up. There are neither restrictions regarding the type or the settings of the client.

In the preliminary stages of the project it was checked if there should be utilised standard or individual application software for the platform being developed. As possible standard solutions the groupware "Lotus Notes" as well as the knowledge database "Community of Science" were taken into account. However, the evaluation made it clear that these solutions would result in technical, organisational or financial disadvantages compared with the development of an individual solution.

4 Operation

The operation is based on the so-called "KISS community" concept (see):

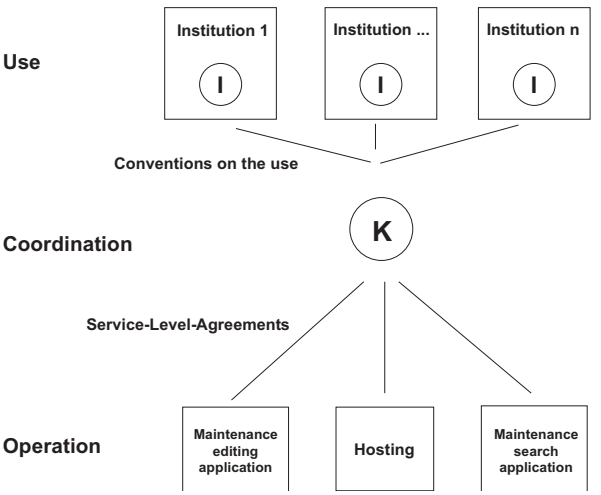


Figure : The KISS-Community

The members of the KISS-Community are located on three different levels: the level of Co-ordination, the level of Operation and the Level of Use. The assignment of duties and the relations between each other can be described as follows:

4.1 The level of co-ordination

The level of Co-ordination consists of the KISS co-ordinator. It is the virtual lynchpin of the KISS-Community as all information is bundled and spread here. The job of the Co-ordinator is done by a staff-member of the ETH-Board. Furthermore the KISS-Co-ordinator has the following duties:

- Preparation and Conclusion of Service-Level-Agreements for Hosting and Maintenance of the applications.
- Evaluation of needs for the further development of the applications.
- Contact point for external institutions out of academia and industry seeking information on possibilities to participate or use the system.

4.2 The level of operation

On the level of Operation the accurate function of all applications is provided. Hence the organisations responsible for the hosting as well as for the development and maintenance of the applications are located on this level.

As a hosting partner the IT-services department of the ETH Zurich was chosen, where the applications can be run on an existing server using Solaris system software. The companies which developed the applications keep the responsibility for the maintenance of their respective part of the system.

The compliance of distinct operational requirements is regulated in a support-concept comprising three levels of support and additionally in bi-lateral Service-Level-Agreements between the ETH-Council and the three organisations.

4.3 The level of use

The level of use is composed of the institutions which make use of KISS by entering their data. Every university and research institution can join the KISS community. For the use of the service the institutions sign a ”convention of use” with the ETH-Board and commit therewith particularly observing the code of conduct included in the so called ”KISS-Charta” in terms of a quality assurance:

- They encourage the researchers to aliment the platform actively. This concerns not only the initial launch but also regular updates of the data.
- They establish the organisational preconditions to assure that no data are published that are not related to the research done at the institution or data that may be offending to the public.

The further development of the common core application is also co-ordinated by the ETH-Board. It is based on the KISS community member’s common needs. Furthermore, the members of the KISS community are allowed to develop supplementary applications individually provided that the function of the common core application is not affected. Contrary to the operation of the application the full responsibility for the data is allocated to the institutions.

5 Conclusions

5.1 Perspectives

The first institutions start with the entry of data in May 2002. The data search will be available from June 2002 on www.knowledgetransfer.ch.

Apart from the institutions of the ETH-Domain several cantonal universities and polytechnics have in the meantime expressed their interests in www.knowledgetransfer.ch. They will be integrated in the near or intermediate future so that KISS middle-term is expected to contain data about around 15'000 research projects.

5.2 Perceptions

Different perceptions can be derived from the experiences in the development and the operation of this central knowledge platform in a very distributed environment:

- The Operation of a web based three-tier-architecture and the concentration on content instead of workflow has proven itself in a peripherally organised environment without a homogeneous IT-infrastructure.
- The concept of entering only a few meta-data with links to further information increases the user acceptance and therewith the chances of success of the system.
- Practically all institutions intend to use the collected data not only within the scope of the publication on www.knowledgetransfer.ch but also simultaneously:
  - in order to support internal organisational processes (such as applications, statistical evaluations, reporting)
  - to serve superior positions with research information (e.g. feed of the database ARAMIS which contains data on research projects financed by the Swiss government)

The realisation of the corresponding technical interfaces will occasionally decide about the employment of KISS in the individual institutions.

6 References

Becker, M.; Haberfellner, R.; Liebetrau, G. (1997): *EDV-Wissen für Anwender*. Zürich: io Verlag

Schweizerischer Bundesrat (1999): *Leistungsauftrag des Schweizerischen Bundesrates an den ETH-Rat für die Jahre 2000-2003.*

7 Contact Information

Dominik Ulmer
Stab ETH-Rat
ETH Zentrum
CH-8092 Zürich

Beat Birkenmeier
Business Results GmbH
Siewerdtstr. 105
CH-8050 Zürich

e-mail: birkenmeier@bresults.ch

International Research Information System: Support to Science Management

Barend Mons[2,4], Renée van Kessel[1], Albert Mons[3], Ruud Strijp[1], Bob Schijvenaars[2,4], Erik van Mulligen[2,4]
[1]Netherlands Organization for Scientific Research, The Netherlands
[2]Collexis B.V. Geldermalsen, The Netherlands, [3]Collexis Solutions, USA,
[4]Erasmus University Medical Center, The Netherlands

Summary

In response to the ever increasing complexity of international scientific networking, the Dutch Government through NWO, The Netherlands Organization for Scientific Research, has taken the initiative to develop a global information system for Research Councils with the working title IRIS (*International Research Information System).* Most Research Councils consider finding referees a frustrating and time-consuming process. Too often major resources are spent on maintaining websites, indexing research proposals, trying to find the perfect referee, and evaluating researchers and research institutes. Science Managers should ideally have access to information about referees from any country or discipline in the most effective way possible. This paper describes the pilot phase of the IRIS project, aimed at the construction and set up of an International System to share reviewers.

1 Introduction

1.1 Preamble

This paper will focus on the basic principles of a very specific form of knowledge management, being the optimal use of validated explicit knowledge worldwide to support the upward knowledge spiral through high quality scientific research. It will first give a minimal background description of the basic philosophy behind the system and the underlying technology, followed by a description of the plans to set up an International Research Information System (working title IRIS) that is designed to support the management of the increasingly multidisciplinary and international research arena.

Many previous attempts to set up networked knowledge resources have failed, and yet another one is therefore likely to meet with a fair deal of skepticism. However, the newest generation of Information Mediation (IM) technology and the advent of a growing validated body of information available via the Internet have created an unprecedented challenge, as well as a unique historical opportunity to make it *work* this time.

1.2 How is knowledge generated?

Before we embark on the description of a system to organize optimal exploitation of existing knowledge word-wide we will briefly reflect on some basic aspects of knowledge generation and try to draw some conclusions and lessons from that reflection.

Many failures of "Knowledge Management" initiatives have been blamed on either a fundamental lack of distinction between the various levels of knowledge components, namely *data*, *information*, *knowledge* and finally *expertise* and *competence* (Sveiby 1997).

Even the basic building blocks of information (data) are not free from human interpretation (Heisenberg 1976) as they can only be observed, measured and stored by organisms with at least a basic form of knowledge. This implies that knowledge generation is in essence a *cyclical process* (Tuomi 1999).

Knowledge as it exists in the heads of people can only be effectively communicated to other human beings after being made explicit in written or spoken language or in visual form. This is a crucially important basic assumption for knowledge management. After being made explicit in communicable form, *knowledge* has in fact been reduced to *information* with the loss of one level of complexity (human creative and associative power). It is therefore crucial to deduce that Information Mediation is not the same as Knowledge Mediation; if knowledge is indeed confined to people, effective Knowledge Management is therefore more complex than simple Information Management. An international system aiming at optimal exploitation of knowledge should therefore take into account the enabling environment for *direct human interaction*.

It should also fully encapsulate the notion that information as captured in natural language, introduces multiple variations in expression of the same concept. Different national languages can be used, but in addition jargon can introduce multiple synonyms for the same concept within one and the same national language. In addition, language introduces the homonym problem (multiple meanings of the same expression).

The human drive to share knowledge has also been assumed in a rather naïve way in some systems. Before success of a networked knowledge system is even a viable concept, one has to realize that the Immediate Return on Investment (mainly time) for all distributed content owners should be obvious. Both scales on the balance should therefore be addressed:

On the "entry side", the investment needed to make data and information "exchangeable" via the network should be absolutely minimized and duplication (filling in the same data repeatedly for different initiatives) should be banned. Such a reduction in time investment has both a technical (no forms) aspect and a networking (political) aspect: It requires an upfront collaboration between the major players in the field to be networked to avoid duplicative submissions wherever possible.

On the "output side" the focus should be on *immediate benefits* for the users who decide to share information. Scientists have been intensively trained to disagree *by default*, but there is probably at least one aspect on which they all agree: they hate the time investment is searching for relevant information, partners, meetings, (re-) writing applications in horribly complicated forms and filling in their registration details for meetings, applications and surveys over and over again.

Scientific policy makers and science managers on the other side of the table have similar time-consuming problems in finding the right referees, distributing calls for proposals to the right people and institutions, keeping their public information updated and to optimally inform scientists as well as the society about ongoing research.

If the output of a system where people register by sharing their information, would be the *immediate* return of relevant and validated information on people, literature, projects, meetings and other relevant issues based on the information provided by a registered user, including a subsequent alert function for "more like this" the incentive to join the network by sharing information would be optimal.

With the enormous explosion of data and information going on at present and the resulting drive towards international and multidisciplinary scientific networks all these aspects become orders of magnitude more complex than they have been for the past few decades. One major constraint is introduced by the advent of the World Wide Web as an essentially uncontrolled infor-

mation carrier with unprecedented, low cost publication possibilities. As a result, the user is confronted with massive amounts of information, of which much may be irrelevant and even wrong and thus counterproductive. *Validation* of information is therefore crucial and scientific publishers see their role in this area for the future as a crucial part of their core business.

Last but not least, players that have an immediate interest in keeping parts of the system validated and updated should be placed at the core of the network as they will have a radiating quality effect in their specific sector. In the case of the IRIS project national research Councils and comparable Research Funding institutions will be forming a core backbone.

In summary, the critical elements of an International Knowledge-driven Communication System should include:

- A clear basic understanding of the difference between data, information and knowledge and their specific interaction (knowledge being essentially confined to people)
- A technology to deal with natural language variation and jargon issues
- A clear and immediate incentive for content providers and evaluators/editors to join and remain active in the system
- A minimal need for "forms to be filled"
- A clear distinction between validated and doubtful information
- A strong network of prominent partners in the area from the very start.

2 The Technology

2.1 The Matching on concepts

The selected technology for IRIS was originally developed to match across jargon and languages in large, distributed text corpora. Collexis® is based on proprietary technology, originally developed in the public sector (Van Mulligen *et. al* 2000, http://www.collexis.com)

The basic theory behind the core technology is that, although humans communicate in explicit language, including many variations and ambiguities, the final aim of communication is to share "concepts". Concepts are in this context, the "real life entities" that constitute the reference framework of human knowledge. An effective Information Mediation technology should therefore search and match at the *concept* level rather than at the *word* or *term* level to enable cross-jargon and cross-language communication. Collexis® has realized just that. It is a web-based technology, essentially driven by pre-existing, validated knowledge as made explicit in thesauri and ontologies. A process of "pre-training" enables the system to accumulate all existing human knowledge in a given field based on the validated knowledge resources in that field before starting to analyze large amounts of information. After being "made smart" by incorporating validated human knowledge into its text analysis component, the abstraction engine is equipped with the most unique feature of Collexis®, namely the *immediate* "normalizing" of all textual variations of words or terms referring to the same concept to a unique concept ID, without the need of extensive cycles of machine learning. The abstraction component thus creates a *Conceptual Finger Print* (CFP), a numerical representation of the real content of full text through a list of approximately 50 concepts listed in order of their relative importance.

Queries can be either pre-computed CFP's of existing text (the "*more like this*" function) or natural language queries typed by users, which are abstracted in the same way as the content in which the search is done and again, all natural language variations of a concept will lead to the normalized unique ID of the concepts searched for. In addition the underlying thesauri allow conceptual searches with definition support as well as dynamic categorisation leading to "dynamic portals"

Accumulated CFP's from multiple publications or projects form dynamic *interest, activity* or *publication* profiles of scientists and experts (Interactive Scientist Information Card [*ISIC*]).

Hundreds of millions of CFP's can be stored on one server and can be compared to each other by vector matching in a matter of milliseconds. This matching process is language and jargon independent and can be used in conjunction with any existing local platform or database used by networked partner institutes. There is no need to change existing Web Based Information Systems, Document Management Systems, or local databases. Web access to the pages containing the actual information is the final step. Entire data collections can be fingerprinted at a speed of 250.000 pages per day per PC independent of the local data system used to present the original information on the user's screen. This new web based approach thus allows unprecedented search and networking properties across distributed and non-standardized data sets.

Collexis® also contains drivers to automatically detect words and terms in texts that are neither identified as irrelevant (classical and user-defined stop words) nor incorporated in the word or concept lists of the thesauri used to train the Collexis® application. Such words and terms (word combinations) can be presented to the application administrators or authorized users as lists of suggested concepts. The system has recently been expanded with homonym detection and disambiguation based on context (semantic laterality) and is currently adapted for the life sciences by the collaborating teams of Collexis and the University of Rotterdam. An ambiguous term like BSE, which can either mean *Bovine Spongiform Encephalopathy* or *Breast Self examination* in a medical context, will normally lead to two possible, distinct concepts. However even in a very short query (bse scrapie, or bse behavior) the system will already make the distinction, based on the contextual concepts used by the corrective technology (figure 1).

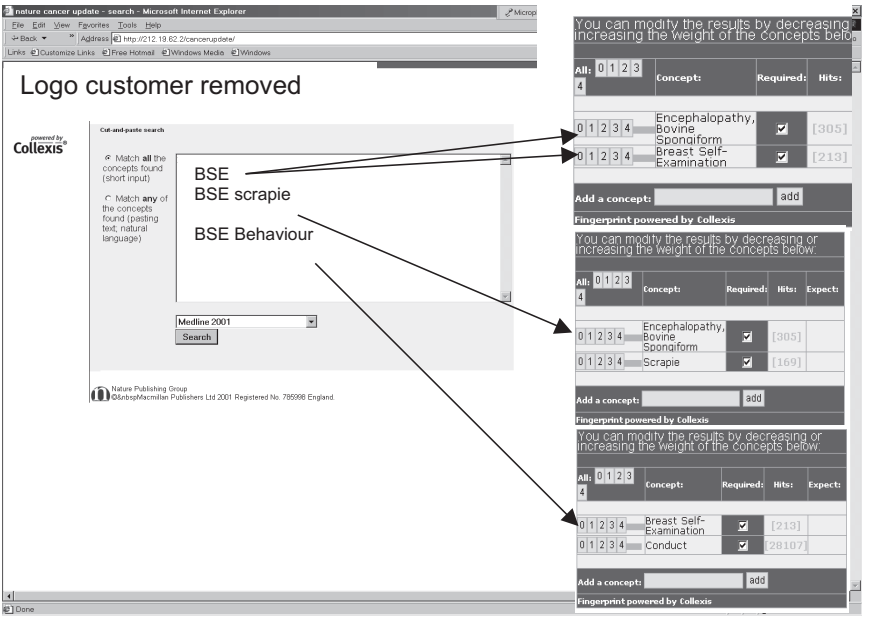


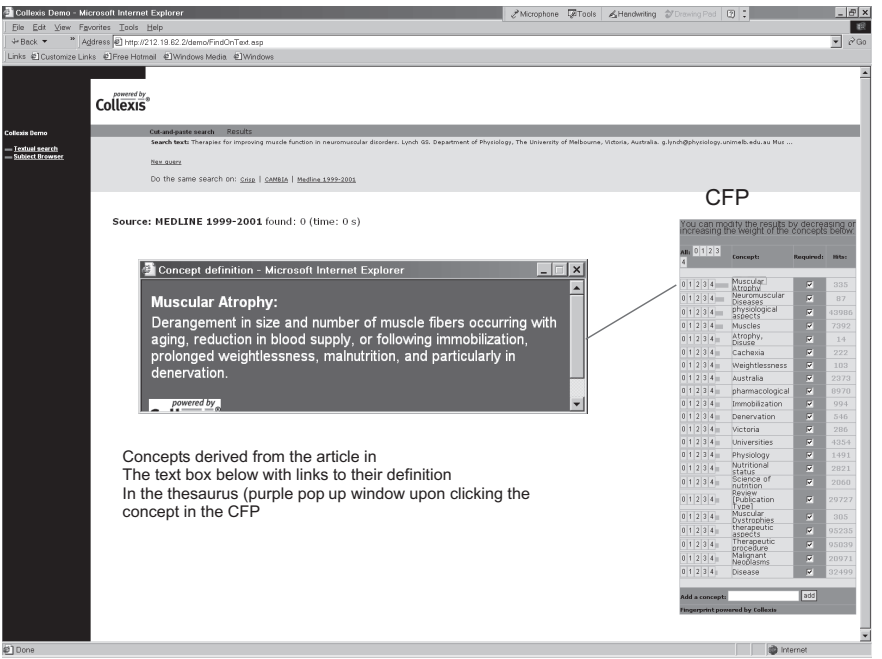
Fig. 1: BSE, either alone or with one context word attached leading to suppression

An additional value in Collexis® technology is that background information on concepts found in text can be used to inform the user directly about existing background knowledge about the concept. Figure 2. depicts what happens in the current public life sciences demo of Collexis when the user clicks on the textual expression of a concept in the CFP (in this case Muscular At-

rophy): A window opens which depicts the description of the concept as provided by the Medical Subject Headings (MeSH)

This example is also illustrative for the difference between Collexis and manual concept or keyword assignment: Many words and terms in the abstract depicted below are almost literally mapped to the corresponding concepts (blue) and some are associated and "interpreted" by Collexis (orange) in the CFP shown in figure 2. The coverage of correct concepts in the example is very high, which is the general trend. The precision and recall aspects of Collexis can be influenced by tuning the system to the need of individual user communities and applications.

Therapies for improving muscle function in neuromuscular disorders.



Lynch GS

Department of Physiology, The University of Melbourne, Victoria, Australia.
g.lynch@physiology.unimelb.edu.au

Muscle atrophy or wasting is a loss of muscle tissue resulting from disease or lack of use. This review examines recent pharmacologic or nutrition interventions for ameliorating wasting and improving muscle function in neuromuscular disorders. The information has application for treating the muscular dystrophies, cancer cachexia, weightlessness, immobilization, denervation, and disuse atrophy

Fig. 2: Definition support directly from the Conceptual Fingerprint (CFP).

The CFP based connection of "similar" information in different and distributed databases can be achieved with minimal effort of the content providers. Simply providing access to their data will allow the system to create CFP's of all desired content fragments and when these fragments (CV's projects, publications etc.) are linked to contact data of people and organizations, the CFP's of people and organisations will automatically accumulate into activity profiles and

knowledge profiles. The matching process is now carried out completely at the (language independent) concept level and can be started from any text fragment of interest. Most systems exploit a simple full text Boolean search engine in addition to the Collexis® matching technology, in order to be able to find highly specific and infrequent words that are not covered by the thesauri used in Collexis®.

3 The International Research Information System (IRIS)

3.1 Background

The Internet, although contributing significantly to the almost unmanageable information explosion itself, also offers unprecedented opportunities to reduce the burden of science management. For the first time in the relatively short history of Information Mediation, the newest technologies provide for tools to handle information in a revolutionary different way. Matching proposals across languages with the best referees for example, and the efficiency of Information Exchange can be improved significantly through the method of matching information based on knowledge-driven indexing described in chapter 2.

Effective International Research management will contribute significantly to the increasing societal demand for accountability, transparency, effectiveness, accessibility and productivity of public and private resources spent on Scientific Research. IRIS is intended to enable all research councils to utilize the IRIS technology, which will provide for inter-council exchange of information and international selection of referees and much more. The Netherlands National Research Council (NWO), the Ministry of Education, Culture and Science (OC&W) and the National Institute for Scientific Information (NIWI) have taken the initiative to develop a prototype of a comprehensive international research management system tailored to the needs of Research Councils and other institutions dealing with research management.

3.2 Scope and aims

IRIS intends to create an International network of National Research Councils (NRC's), sharing minimal information on both national research and scientists on a global scale, while allowing each participating organization to keep using its existing Information Management system. IRIS is meant to function as a decentralized system with national responsibility for the quality of all data. The basic principle is a collective back up of local data through an ASP based approach. Each participant will manage its own domain in the system, which allows for an organization's own "look-and-feel".

The project will result in a working prototype of a Web based information system that allows for:

(a) Web-based, global and interactive selection of referees.
(b) On-line submission of proposals in a variety of formats.
(c) On-line reporting and updating.
(d) Personalized support for scientists and science managers.
(e) Real time evaluation of scientific performance.

The ultimate goal is to make the human and institutional aspects of scientific networking and management much more meaningful, efficient and thus less time consuming. In addition, a decrease in monetary overhead can be achieved. If ideally all (or the vast majority) of active scientists would be represented in the system, the search for (international) referees on any given proposal or paper would be reduced to seconds rather than hours and collective contacting of the referee panel could be performed directly from within the system.

3.3 Status of the IRIS project in May 2002

NWO is in an advanced stage of discussion with a number of research councils in Europe, The European Commission, The United States, Asia and Latin America to join the prototype phase. Several councils and major scientific publishers have expressed their keen interest in joining the project, which is crucial for its success.

NWO has found financial and moral support for its initiative to launch the IRIS project. In The Netherlands the participating organizations are convinced of the need for IRIS and have decided to fund and implement it. At the international level NWO has already found major interest in the development of an internationally applicable system. To NWO's knowledge no similar initiative has been launched with the same scope IRIS.

Since the first stage of development of the necessary applications has been completed, NWO intends to test all features implemented in the Netherlands with several international partners, including but not limited to a selection of National Research Institutes.

IRIS will become operational for all fields of expertise. The Medical and Heath Sciences field plays a major role in the pilot phase based on the exceptional quality of the thesauri available for these fields. Other fields will be subsequently included in the project as more thesauri become available.

NWO intends to honor the general understanding that validation of knowledge and management of content and personal data is foremost the responsibility of NRC's. Therefore the choice has been made to implement a fully de-centralized system with optimal opportunity for NRC's to control their own data and give them autonomous authority of quality control through National Focal Points. Based upon the technology used it will still be possible to provide global access to all data entered through the CFP's stored on an ASP server.

In order to ascertain the feasibility of the proposed approach at an international level NWO intends to have 5 major NRC's committed to participation in the pilot phase, prior to launching the formal try-out. At a formal meeting in September 2002, the project will be presented and a broad representation of NRC's will be invited to participate in this meeting. Up to 10 additional Councils can participate in IRIS during the Pilot Phase at *no cost* other than local overhead and training of personnel.

The final goal of IRIS is to develop a functional global network of NRC's and related content providers for the benefit of international scientific networking and management. As many Institutes as possible should have joined the initiative by then. After successful completion of the pilot phase, it is intended to spin off IRIS as an independent organization. The legal structure will be discussed in detail with all IRIS partners in the evaluation phase (2003).

The term for completion of the pilot phase of IRIS is two years. The project was launched in October 2001 and should be completed in October 2003.

Participation in IRIS has several immediate benefits for IRIS partners. It provides the ability to find referees online based upon available knowledge profiles of all scientists available. In addition each Partner will have the ability to search all fingerprinted scientific data and match them with queries based upon Search Fingerprints created from full text sources.

Another obvious advantage is the ability to compare grant-applications with other already funded projects on a global scale. Lastly, all scientists working with or for the Partner will have the ability to utilize the features of IRIS for their own projects.

3.4 Technical Scope of subsequently developed tools and steps in IRIS

Note: The steps and tools are described following the chronological order of the science management process: Calls for proposals-submission-review-reporting and evaluation. The actual technical development of the tools does not necessarily follow the same sequence. It has been de-

cided to take *Step 2*, the "referee finder" as a major first step as it contains the greatest challenges, both technically and in terms of networking.

Moreover, several potential partners may have excellent systems in place for on-line submission and/or reporting. In such cases the interactive connection of the existing systems to the central referee selection system would be a first step and that makes the referee module the only default centralized tool in IRIS.

Several partners may decide to implement only the referee tool (2) and this would not in any way jeopardize the successful implementation of the IRIS concept.

Step 1. On-line submission of proposals in a variety of formats

IRIS will design and implement an extremely user-friendly on-line submission tool for scientists. It can be used by organizations that do not have such a tool available. For those councils that already have on-line submission of proposals operational, only additional functionality may be imported, which will render existing submission procedures intact

Step 2. Web based, global and interactive selection of referees

Most research agencies or scientific content management organizations have a major burden during the selection of referees for submitted proposals. Not only the selection of the correct reviewer in terms of expertise and knowledge, but also the practicalities of finding these people, composing mailings and organizing the review process are a serious management issue. The IRIS Referee Finder Tool will include a feature that allows semi-automated collection of publication profiles of referees based on literature available in the public domain.

The Referee Finder will allow for an efficient decentralized tool with global access to all data available in IRIS through the shared back up of all National data. As the first step in the process, the agency approves a submitted proposal for the review process and the following steps are "interactively automated". Selected text fields in the on-line application will be fingerprinted and will generate a "proposal profile", which can be reviewed for inconsistencies by the project manager. The council manager will be able to see publications, projects of other research councils including knowledge validated anywhere in the world. The fingerprint of the proposal will be used in the matching engine to find the most suited reviewers. With this application, the organization can find the best referees for a particular project proposal in the most efficient and expedient manner possible.

Step 3. On-line reporting and updating

Research councils usually have a contractual relationship with the scientists whose research was funded, which includes regular reporting. This is at present a cumbersome process, requiring regular reminders, approval of the reports and updates of existing databases at the Councils' offices. IRIS' technology allows far-reaching automation of many elements of these processes.

Step 4. Personalized support for scientists and science managers

All the scientific output of a researcher represented in the IRIS knowledge or activity profiles (ISIC's) is constantly updated and used for search by institutes and individuals. IRIS will provide for personalized "interest rooms" for individual scientists. Since scientist can receive constant updates on information matching their interest profiles, this feature could create a critical incentive for reviewers and scientists to keep their profiles updated in the system.

Step 5. Real time evaluation of scientific performance

In close collaboration with national and international partners involved in bibliometric analysis of scientific performance it is proposed to develop dedicated tools for internal and external evaluation of scientific institutes. Based on the cross-referencing technology used, such tools can eas-

ily be conceptualized and developed. However, serious discussion between the players is still needed to define the parameters and consequently the (fine-tuning of the) technology and to work out schemes that respect full privacy security of sensitive data related to persons or institutes. Therefore, these tools will be heavily protected and only made available to officials dealing with science policy.

3.5 Organizational Scope

National Research Councils and Science Management Organizations can participate in IRIS as co-owners of the initiative. For the initial phase, a small international administrative office (through NWO) will provide the necessary administrative support including help desk and provisions for training employees of local Institutes. NWO is prepared to assume responsibility for effective execution of IRIS during the first two years of its existence. An International Advisory Board will be established after the Pilot Phase has been evaluated and a decision to continue IRIS on a permanent basis will be made by all collaborating IRIS Partners. For further details, please contact the corresponding author via: iris@nwo.nl

4 References

Heisenberg, W. (1976) The uncertainty principle. Zs. F. Phys. 43, 172-98

Sveiby, K.E. (1997) The new organizational wealth: Managing and measuring Knowledge based assets. http://hallinternet.com/net_history_trends/71.shtml

Tuomi I, 1999 Data is more than knowledge, Implications of the Reversed KnowledgeHierarchy for Knowledge management and Organizational Memory, Proceedings of the Thirty-second Annual Hawaii International Conference on System Sciences. http://www.computer.org/proceedings/hicss/0001/00011/00011071abs.htm

Van Mulligen EM, Diwersy M, Schmidt M, Buurman H, Mons B. (2000) Facilitating networks of information, Proc AMIA Symp 2000: 868-72

5 Contact Information

Barend Mons
Nederlandse Organisatie voor Wetenschappelijk Onderzoek
WOTRO
P.O. Box 93120
2509 AC Den Haag
The Netherlands

e-mail: barend.mons@inter.nl.net

DBClear: A Generic System for Clearinghouses

H. Hellweg[1], B. Hermes[1], M. Stempfhuber[1], W. Enderle[2], T. Fischer[2]
[1] Social Science Information Centre (IZ), Bonn
[2] Lower Saxony State and University Library, Göttingen

Summary

Clearinghouses – or subject gateways – are domain-specific collections of links to resources on the Internet. The links are described with metadata and structured according to a domain-specific subject hierarchy. Users access the information by searching in the metadata or by browsing the subject hierarchy.

The standards for metadata vary across existing clearinghouses and different technologies for storing and accessing the metadata are used. This makes it difficult to distribute the editorial or administrative work involved in maintaining a clearinghouse, or to exchange information with other systems.

DBClear is a generic, platform-independent clearinghouse system, whose metadata schema can be adapted to different standards. The data is stored in a relational database. It includes a workflow component to support distributed maintenance and automation modules for link checking and metadata extraction. The presentation of the clearinghouse on the Web can be modified to allow seamless integration into existing web sites.

1 Introduction

Clearinghouses – also called subject gateways – are domain-specific collections of links to high quality resources on the Internet. Experts judge the relevance of the resources, describe them according to a predefined metadata schema and assign them to a subject hierarchy or classification. Clearinghouses try to give users orientation on the fast changing World Wide Web (WWW) and efficient access to relevant online information. Well-known examples are CetusLinks[1], Geo-Guide[2] (Enderle 1999) and SocioGuide[3] (Hellweg 2000).

In the first years of the WWW, clearinghouses were maintained as lists of bookmarks to Internet sites, often collected, maintained and published by a single person and stored as static HTML files on a web server. Understood as vertical (thematically limited) directories, they were relatively small, structured by a single subject hierarchy, and described the resources only by a very limited set of metadata elements. When growing past a certain size, it became difficult and time consuming to manage and maintain these bookmark lists, especially if the gateway's domain was subject to rapid change. Furthermore, file-based bookmark lists do normally not support to structure their content according to different subject hierarchies at the same time and the describing metadata can not efficiently be used for searching. New content was acquired mostly by browsing the hypertext structure of the Internet for relevant resources, by using search engines to find topically related sites, by exchanging links with fellow researchers, or even by chance.

But in most cases, a single person can not have a complete overview on a particular subject or domain. As a consequence, finding new sites, judging their relevance, describing resources and

1 http://www.cetus-link.org
2 http://www.geo.giude.de
3 http://www.gesis.org/SocioGuide/

maintaining them must be spatially and temporally distributed to several persons in order to keep the link collection at the same high level of quality and relevance. Automation techniques are needed to support the editors in keeping the collection as complete and up-to-date as possible. Furthermore, a single shared database must be maintained to ensure the consistency of the link collection and to allow the flexible generation of views on the data with current web technologies.

2 Main features of DBClear

To solve the problems listed above, the Social Science Information Centre (IZ), Bonn develops –together with the Lower Saxony State and University Library Göttingen (SUB Göttingen) – a generic, multilingual clearinghouse system, DBClear[4]. The project is funded by the Deutsche Forschungsgemeinschaft (DFG)[5]. DBClear is based on a relational database management system and allows distributed maintenance and administration of its content. Its main features are:

- Support of different, user-definable metadata schemas and bibliographic standards.
- Storage of the clearinghouse's content in a relational database management system with a JDBC (Java DataBase Connectivity) compliant interface.
- A Workflow system to route the clearinghouse's content from the initial quality check all the way to the release on the Internet.
- Automatic metadata extraction from HTML pages.
- Modules to automate frequently recurring tasks, like checking if links are still reachable or have been updated.
- Support for multiple languages at the metadata and the user interface level.
- Separation of content and presentation, allowing flexible adaptation and branding of content for different usage scenarios.
- Gateways (e.g. Z39.50) to integrate DBClear into larger contexts, like Virtual Libraries.
- Import and export of data with support for mapping DBClear metadata to and from other metadata standards (e.g. Dublin Core).
- Generic and platform-independent Java-based system with plug-in architecture for future extensions (e.g. metadata generated on-the-fly from external systems).

3 Target scenarios for using DBClear

When designing the overall architecture of DBClear, a number of scenarios were evaluated where DBClear and its content could be integrated with existing systems. This includes integration in web presentations of institutes, connecting DBClear with legacy systems, or searching its content as part of a larger Virtual Library. The technologies and standards used in developing the system where determined by their support for these target scenarios.

3.1 DBClear as a standalone portal system

The most common scenario for using DBClear will be as a standalone portal system which is part of an organization's web site. Here it is necessary to adapt the look and feel of DBClear to the surrounding web pages. Through strict separation of content and presentation, it is possible to design and implement different user interfaces. DBClear generates an intermediate XML format from database content and then uses XML/XSL (eXtensible Style sheet Language) transformations to generate views for browsing and searching the clearinghouse and for displaying information. The XSL style sheets allow to include a wide range of Internet technologies in the generated

4 http://www.gesis.org/research/information_technology/DBClear.htm

5 DBClear is funded under grant no. Gz: III N – 554 922(1)00

tions to generate views for browsing and searching the clearinghouse and for displaying information. The XSL style sheets allow to include a wide range of Internet technologies in the generated HTML pages (e.g. advanced functions in JavaScript) and to transform data into different formats. Values stored in the system's database in one format (e.g. a numerical value which describes a resource's relevance) can be rendered as text (e.g. "high", "medium", "low") or as graphics (e.g. a number of "stars" representing the relevance).

This flexibility in generating different output formats can also be used in settings where multiple organizations collaborate in creating and maintaining a subject gateway and where every participant "donates" and maintains its existing sub-collection (figure 1). A single instance of DBClear – running at one institute – can be integrated in every institute's own web site and use different style sheets. This allows every organization to adapt DBClear to its corporate layout and to brand content with the contributing organization's logo.

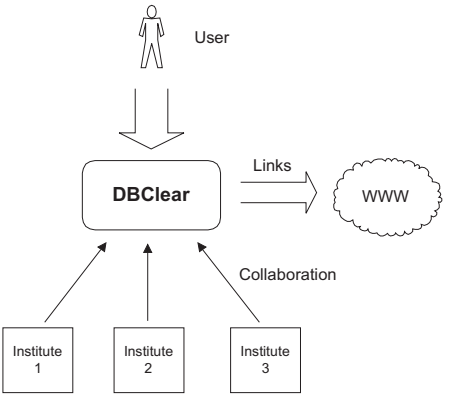


Figure 1: DBClear as a standalone portal

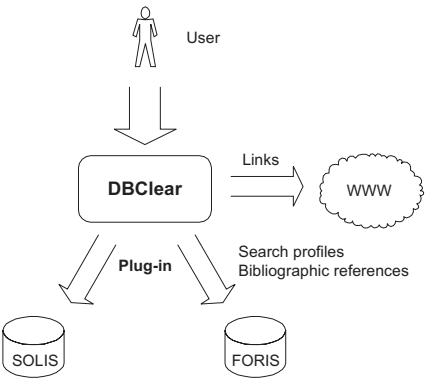


Figure 2: Using external data sources

3.2 Using external data sources

A subject gateway is often maintained as one of a larger number of topically related information services. The Social Science Information Centre (IZ) for example offers a reference database for social science literature (SOLIS), a database with research projects (FORIS) and smaller databases with institutions and conferences. To achieve the maximum benefit for the user, these databases should be integrated with DBClear.

To access external data sources, the DBClear architecture contains a plug-in interface to extend the system with additional features in the form of Java classes, which are dynamically loadad. New features could be "live" attributes, whose values are not stored in the DBClear database, but are retrieved from external data sources and generated on-the-fly (figure 2). By defining an attribute which contains identifiers of bibliographic records in SOLIS or FORIS, Internet resources could be linked with highly relevant external data. In a similar way, search profiles could be stored for single or groups of resources and executed every time a user accesses these resources. This technology is not limited to locally available sources, also external search engines or harvesters – specialized in searching well defined parts of the Internet – could be integrated.

3.3 Integrating DBClear with library catalogues

In other contexts, e.g. ViBSoz, the Social Science Virtual Library[6] (Meier et al. 2000), the resources contained in a clearinghouse could enrich the information contained in other information systems, e.g. library catalogues. DBClear therefore contains an interface which allows third party systems to search its content (figure 3). Currently, the Z39.50 protocol is supported, which is widely used for connecting universities' library catalogues for integrated searches. In ViBSoz, it connects library catalogues with the collection of the Friedrich Ebert Stiftung[7] and the IZ's database SOLIS[8]. The Z39.50 interface of DBClear uses a mapping between the bib-1 attribute set of Z39.50 and the DBClear metadata schema. This mapping can be freely defined and adapted.

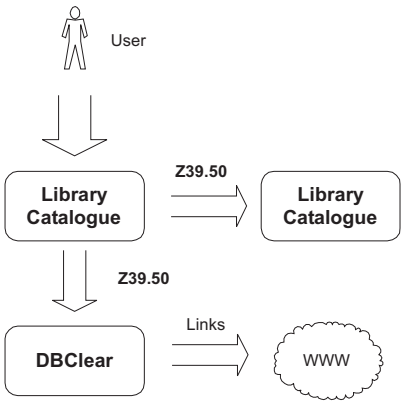


Figure 3: Integration into other systems

4 Storing metadata in DBClear

One of the primary goals of DBClear was to develop a system whose metadata schema can be adapted to nearly every standard. This should not only include official standards (e.g. Dublin Core, see DCMI 1999), but also standards which are currently defined in projects like RENARDUS[9], where the metadata schemas of 12 gateways are mapped onto a common schema suitable for cross searching with a broker architecture.

The following sections describe the fundamentals of the DBClear metadata schema and illustrate its flexibility. In addition, an overview of the overall system architecture is given.

4.1 Metadata

To design a generic, multilingual clearinghouse system, several existing clearinghouses and Internet portals were analyzed. They used different sets of metadata to describe the Internet resources and various subject hierarchies (e.g. classifications) to group and structure them thematically. These differences were not only domain-dependent – clearinghouses on the same subject or domain differed as well.

6 ViBSoz is funded by the Deutsche Forschungsgemeinschaft (DFG), http://www.vibsoz.de/
7 http://www.fes.de/
8 http://www.gesis.org/Information/SOLIS/
9 http://www.renardus.org/

4.1.1 Metadata types

DBClear has a generic and flexible model for representing metadata, which allows every clearinghouse to define its own set of metadata elements. The metadata elements are divided into *facets* and *attributes* (figure 4). For both, the cardinality (i.e. how many values must be entered at minimum and how many may be entered at maximum) can be defined by the administrator.

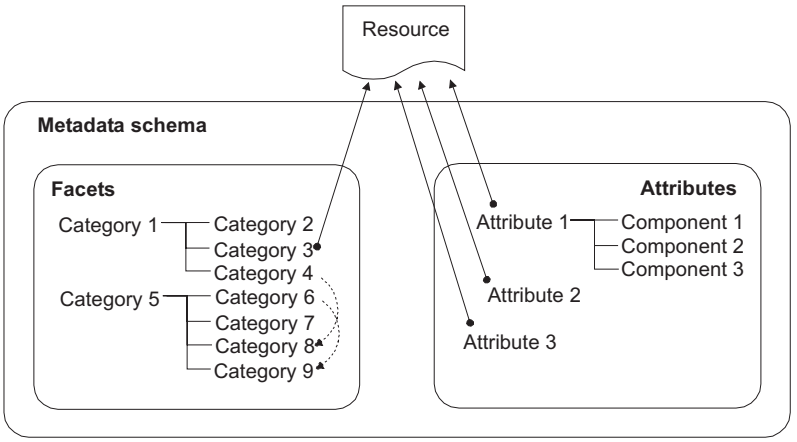


Figure 4: Metadata schema with facets and attributes

Facets are used for metadata elements which consist of a controlled vocabulary whose values – called *categories* – do not change frequently (e.g. a classification, thesaurus, or a list of index terms or country names). The categories may be semantically related to each other (e.g. broader/narrower term relations in a thesaurus) or arranged in a mono-hierarchic list. These relationships can be used to browse the clearinghouse's content.

Attributes, in contrast to facets, are used to model sets of values which are not limited in size and where the single values may not be limited by any formal restriction (e.g. title, author, postal address). Attributes can consist of components (e.g. the attribute *author* may consist of the components *first name*, *last name* and *e-mail address*), which allow more detailed searching and a more flexible definition of output formats. For some metadata elements whose values are naturally limited (e.g. the names of cities), a controlled vocabulary may not always be available. This makes it necessary to use an attribute instead of a facet. In this situation, the reuse of previously entered values is helpful to limit homonyms and to reduce spelling errors. DBClear allows to mark attributes as "reusable" and then presents a list of existing values as needed.

In a clearinghouse with multilingual content, some metadata elements may be language dependant (e.g. keywords or the abstract), whereas others are not (e.g. publisher). In DBClear, a list of languages that defines which languages are optional or mandatory can be assigned to every metadata element. For facets, the vocabulary itself is stored in multiple languages. By associating a category of a facet with a resource, every translation of the category is automatically associated with the resource, too. In contrast, values for language dependant attributes have to be entered separately for every language.

4.1.2 Organizing metadata with stocks

In a clearinghouse, different types of resources may be collected (e.g. links to institutions' home pages, online dissertations or a calendar of events), which can even belong to different domains. The set of metadata elements describing a resource may vary depending on its type. To describe

for instance homepages of institutions, the elements *country* and *city* may be useful, which are not applicable to resources like online dissertations.

A *stock* in DBClear defines a group of semantically related resources, together with its set of metadata elements (figure 5) and serves as a blueprint to automatically generate entry forms for the maintenance of the content. Stocks can be used to filter data during searching and browsing, so that only resources of the specified type are visible to the user. Resources are normally assigned to only one stock, taking into account that stocks will have different metadata elements. Cross-searches over multiple stocks are possible, if they share at least one attribute.

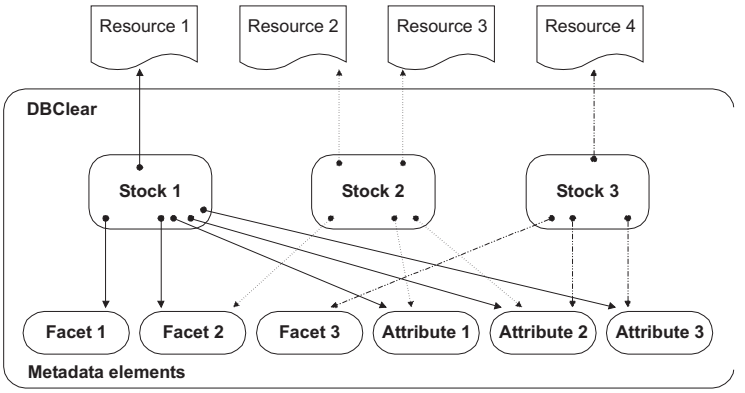


Figure 5: Usage of stocks to organize resources and metadata

4.2 System architecture

DBClear is designed as a multi-tiered information system (figure 6), consisting of:

- A presentation layer, which presents data to the user or editor and permits data manipulation and data entry.
- An application layer, which contains the business logic of the clearinghouse system.
- A data abstraction layer, which stores/retrieves the data in/from a relational database.

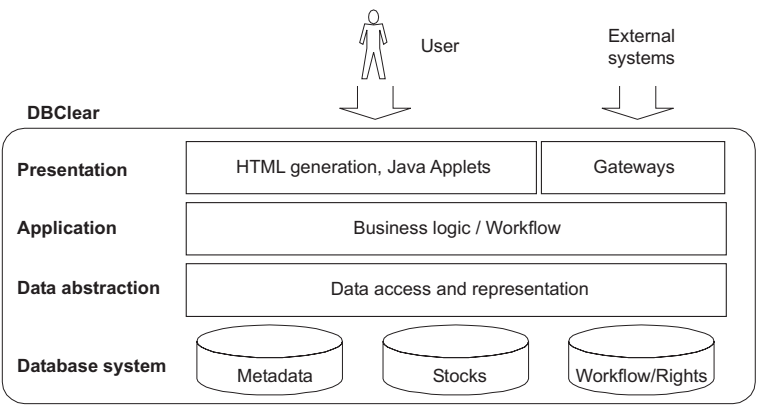


Figure 6: Multi-layered architecture

The application layer receives requests from clients (e.g. Java Servlets), manipulates data according to workflow definitions, and retrieves or stores data using the data abstraction layer. During processing, data is represented in an intermediary XML format. This representation is transformed using XSL style sheets and the resulting output is delivered back to the client.

Using an XML representation of the data and transforming it with XSL into viewable content allows the separation of content (XML) from presentation (e.g. HTML). The presentation can be adapted to different clients according to their display features (e.g. different browsers, Java Applets or WAP-enabled cell phones). It is also possible to tailor the presentation to the corporate design of different institutions collaborating in the same clearing house by simply adjusting the style sheets.

5 Automation and workflow

To facilitate the creation and continuous maintenance of a data collection that exceeds a certain size, tools for automation of tasks are required, as well as support for the distribution of work among several cooperating editors. DBClear provides a number of modules to automate recurring tasks and a workflow system to route information between the people involved in a clearinghouse.

5.1 Automation

Several aspects of clearinghouse maintenance can be automated, the most obvious are regular checks if an Internet resource can still be reached or was modified since the last time an editor had a look at it.

Interviews with people involved in running clearinghouses (e.g. editors and administrators) showed that a number of recurring processes or tasks are not very complex and can either be automated completely – or at least be supported – by the clearinghouse software. This is especially the case with data extraction or classification tasks, which occur mostly in the process of analysing and describing new resources. Here, the system can automatically extract data from the HTML pages by looking for special mark-up elements, e.g. the META elements describing the document's content, or a sequence of elements that suggest a certain document structure. As results from the CARMEN project show (Strötgen 2002), paragraphs of a document can be classified as e.g. abstract, author or keywords with high precision by using heuristics. Additional information can be extracted from the protocol information (HTTP header) a web server sends with each page or from the URL, which can contain country codes or institution names.

The extracted information can be used to produce suggestions for DBClear metadata elements which are then presented to the editors in the normal course of describing a resource and its content. Other values which can be computed automatically are the number of "backlinks" and the document language. The number of backlinks shows how many external sites refer (link) to the resource in question. Its value is usually determined automatically by querying search engines like Google or AltaVista, and therefore can even be updated regularly without user interaction. The document language can be guessed by statistical means, taking the language-specific frequency of certain characters or combinations of characters into account.

DBClear provides a general framework for the implementation of these metadata extraction modules and allows each of its metadata elements to be linked to such a module. Modules for text or header extraction are provided as part of the core system. They can be configured to extract elements like the date or title of web pages and also to retrieve backlinks information.

To maintain the clearinghouse's content and keep it as complete and up-to-date as possible, regular searches for new and relevant content on the Web are necessary. This is usually accomplished by accessing one or more search engines with a predefined query which covers the rele-

vant aspects of the domain, and comparing the results to the resources that are already known (either stored in the clearinghouse's database or in a list of rejected resources).

DBClear allows each clearinghouse editor to define and store any number of queries to external search engines. These queries are executed regularly, and their results are compared to the global list of known resources, as well as to the editor's personal rejection list, which is maintained as part of his personal configuration. New resources are inserted into the editor's personal collection of resources, which he can review and then decide, which resource to reject or to put on his or another editor's work list.

5.2 Workflow

DBClear supports distributed collaboration among clearinghouse editors by allowing to assign a sequence of tasks (workflow), that have to be performed on a resource, to different persons. Each task holds information on the resource and the activities that have to be performed on it. To coordinate the tasks, the system maintains information on the state of a resource together with a work list for each editor. A work list contains the tasks assigned to a person or a group.

Tasks are based on the type of a resource (its metadata) and the (group of) people who carry out the task. Each task consists of activities which can be tied to metadata elements. Rules evaluate the values of those metadata elements and decide which activity has to be performed next, or how the state of a resource changes. This activity can either be associated with an automation module, or with a (group of) person(s). In the latter case, a new item consisting of the resource and information on the activity is entered into the respective person's work list. The person to which a task is assigned may manually change the resource's state or forward the task to someone else in case the rules didn't apply or an exception from the assumed sequence of tasks occurred.

Actions, like checking if a resource is reachable or has been modified, act as initiators of workflow processes, like suggestions of new resources by users of the clearinghouse do. For already catalogued resources, the system can determine the responsible persons and create an entry in the appropriate worklists. For new resources, especially if sufficient information was not provided by the user, the content analysis modules can be used to generate enough context for the workflow rules to determine the initial person responsible for this resource.

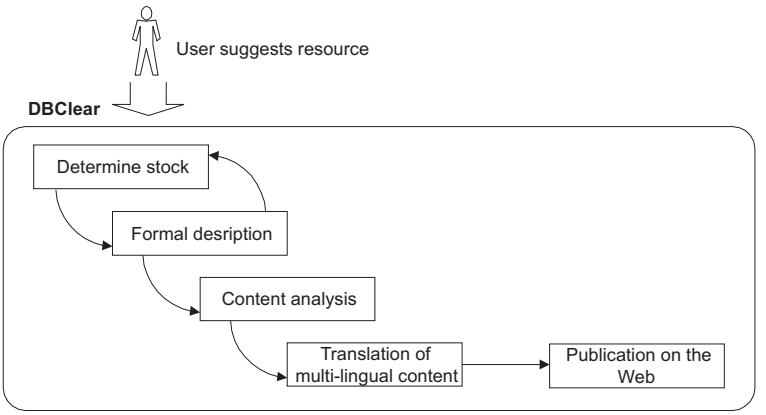


Figure 7: Example workflow sequence in DBClear

Figure 7 shows a simple workflow sequence, in which a user's suggestion is categorized and assigned to a stock, based on the information supplied. By using this initial information, an editor

(or a group of editors) is selected for entering the formal description of the resource, like country of origin, language or resource type. If this editor decides that the automatic assignment to the stock was incorrect, he is free to reassign the resource to some other stock. Once the formal information is entered, the resource is forwarded to the next person responsible for content analysis (e.g. writing an abstract) and indexing. The following translation step has to be performed by an editor with knowledge of the required language. The final publication of the resource on the Web is performed by the editor responsible for the consistency of the collection.

6 Conclusion

With the migration of GeoGuide and SocioGuide – two large clearinghouses with different metadata schemes – DBClear has already proven its flexibility as data storage and system features are concerned. Both clearinghouses are currently prepared for release on the Internet. Besides that, additional metadata schemes and workflow requirements are analysed to make sure that DBClear can also be adapted to them.

After project completion in September 2002, the DBClear system will be open for use at other institutes and organizations who want to offer a clearinghouse on the Internet.

7 References

DCMI (1999): Dublin Core Metadata Element Set, Version 1.1: Reference Description. Dublin Core Metadata Initiative, available at .

Enderle, W. et al. (1999): Das Sondersammelgebiets-Fachinformationsprojekt (SSG-FI) der Niedersächsischen Staats- und Universitätsbibliothek Göttingen: GeoGuide, MathGuide, Anglo-American History Guide und Anglo-American Literature Guide. dbi-materialien 185, DBI Berlin.

Hellweg, H. (2000): Der GESIS Socio-Guide: ein kooperatives Link-Verwaltungs-System. In: Ohly, P.; Rahmstorf, G.; Sigel, A. (Eds.). Proceedings der 6. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO), 23.–25. September 1999, Hamburg, S. 291-298.

Meier, W.; Müller, M. N. O.; Winkler, S. (2000): Virtuelle Fachbibliothek Sozialwissenschaften. Problembereich und Konzeption. In: Bibliotheksdienst, 34, 2000, Nr. 7/8, S.1236-1244.

Strötgen, R. (2002): Treatment of Semantic Heterogeneity using MetaData Extraction and Query Translation. In: Proceedings of CRIS 2002: Gaining Insight from Research Information, Kassel, 29. - 31. August 2002 (to appear).

8 Contact Information

Maximilian Stempfhuber
Informationszentrum Sozialwissenschaften
Lennéstr. 30
D-53113 Bonn
Germany

e-mail: st@bonn.iz-soz.de

Research Information and Strategic Decision Making

Richard Tomlin
Community of Science, UK

Abstract

Research management is popularly described as being like herding cats. Researchers themselves show no instinctive desire to be managed, rather the opposite, and the process of managing creativity is notoriously problematic. However, decisions that directly impact on research have to be made at many levels, from the multi-national all the way down to the personal. The process by which such decisions are made is, however loosely interpreted, management.

Some decisions are of a narrowly technical nature where the researchers' own expertise is sufficient. Other decisions take place on a wider horizon and often involve people not directly concerned with the research itself. Among such decisions are those concerning the allocation of resources, the future direction and coordination of research efforts, and the evaluation of research outcomes. As the competition for research resources intensifies, the quality of such decisions takes on even greater importance. Rational decision making requires that the decision should be made in the light of timely, relevant, and accurate information, yet it is often difficult to find such information efficiently and use it effectively.

This paper will explore some of the key situations in which decisions affecting research are made and how research information systems could be deployed to support the making of those decisions. The examples to be considered include mapping research capabilities from a variety of perspectives as a basis for investment-type decisions; portfolio analysis as a basis for managing research collaborations and other relationships; and the need for less intensive methods of research evaluation.

Contact Information

Richard Tomlin
Community of Science
c/o 24 Bluebell Close
Wylam, Northumberland, NE41 8EU
UK

e-mail: rtomlin@cos.com

AARLIN: Seamless information delivery to researchers

Doreen Parker; Earle Gow, Edward Lim
Victoria University of Technology, Melbourne; La Trobe University, Melbourne

Summary

The Australian Academic Research Library Information Network (AARLIN) aims to provide seamless access to Australian and international information resources for researchers via their personal computers through a personally customisable portal. The project has funding from the Australian Government. AARLIN commenced in the year 2000 with a pilot project and will develop into a fully operational service in Australian universities over the next three years. During the pilot project Ex Libris' Metalib and SFX software have been used to trial the AARLIN portal concept with a group of researchers. The results of a survey of the researchers are presented. It is concluded that the portal has the potential to enhance the work of researchers by improving their success in information searching.

Keywords = portals, research, information resources, libraries

1 Background

The Australian Academic Research Information Library Network (AARLIN) project emerged from a Council of Australian University Librarians (CAUL) planning workshop in 1999 and was adopted as one of the strategic objectives of CAUL. Twenty of the thirty eight Australian universities and the National Library of Australia are active participants in the project and have contributed funds. The AARLIN project has established cooperative arrangements between the institutions involved for the direction and management of the project. The project is based at La Trobe University where a Project Officer has been employed. The project is guided by a Steering Committee which includes representatives of CAUL and the Council of Australian University Information Technology Directors (CAUDIT). The Project Officer is supported by designated staff in each of the participating universities.

2 Aims

AARLIN aims to develop a major network infrastructure to support research in Australian universities and other research organisations. The AARLIN vision is a national virtual research library and information system that will provide unmediated, personalised and seamless end user access to the collections of Australian libraries, to research databases and to document delivery services from the work stations of research staff and students. To achieve this vision the AARLIN project is using portal technology. The first stage in the creation of AARLIN is the development and testing of the AARLIN portal prototype. Subsequent developments will involve the establishment of an administrative structure or entity; a legal framework to ensure compliance with agreed performance standards and quality assurance; and a business plan to ensure sustainability and financial viability of AARLIN as the national vehicle for discovery of, and access to, research information resources for the Australian research community.

It is intended that the national portal will have context sensitive and open reference linking software which will permit researchers once authenticated to:

- access a context-sensitive and "standardised" search interface and undertake concurrent searches of electronic databases, web sites, online library catalogues and other electronic information resources;
- pass appropriate metadata for an unmediated document delivery request and generate a document delivery request, if required;
- access a range of appropriate or extended services (including deeplinking to full-text where available) using context sensitive reference or OpenURL linking software;
- personalise their search "environment", including access to the information resources which are relevant to their research interests, the capacity for them to suppress and expand various resources presented to them as a default, and the capacity for them to add their own bookmarks;
- have pushed to them the relevant "information landscape" or suite of information resources as determined by their authenticated user profile;
- establish or modify profiles for, and receive literature alerts informing them of newly available material matching the criteria specified.

It is envisaged that, in due course, the services offered through the AARLIN portal will incorporate a payments system and a rights management system.

It is intended that the software which is selected for the operational system will have the capacity to integrate with local authentication and profiling systems and services and will comply with industry standards.

Diagram 1 outlines the components of the proposed national system.

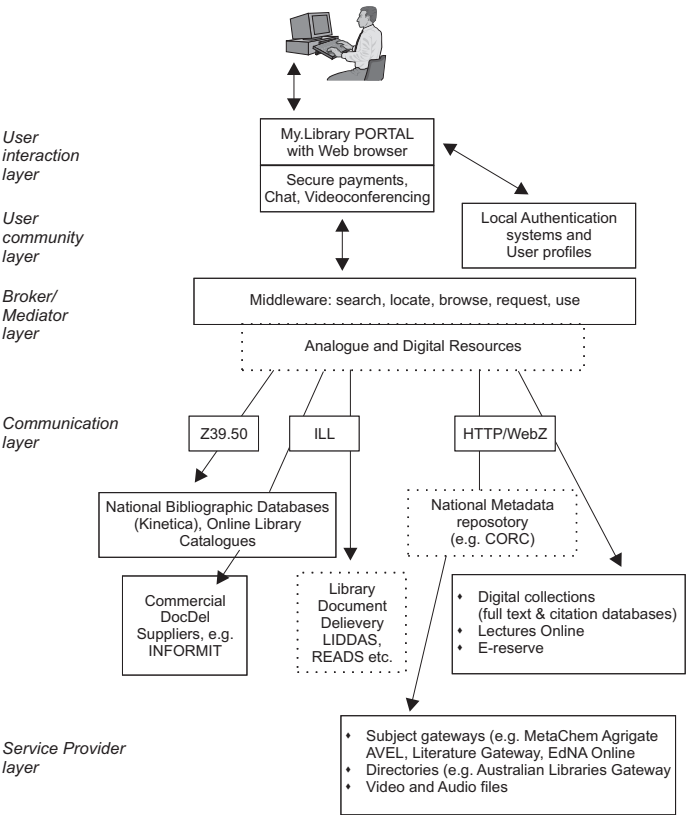


Diagram 1: Australian Library Research Network Information Infrastructure

3 Related portal developments

Portals are being used by a number of universities, university libraries and by library consortium.

In Australia ten of the thirty eight universities had operational portals and another ten or more were actively planning the implementation of portals in the year 2000 (CAUDIT, 2000). Monash University has incorporated library portal functionality in the design of its university portal. The "My Monash" personal portal includes "My Library" and "My Digital Library". (http://my.monash.edu.au/). Curtin University of Technology library's portal "My Library@Curtin" draws on the experience of the Digital Library Initiatives Department of the NC State University in the USA (http://john.curtin.edu.au/mylibrary/). Victoria University of Technology library has implemented the Innovative Interface Innopac system catalogue portal "Your Library" which allows searches to be saved and updated results to be emailed periodically (http://w2.vu.edu.au/library/cat/yourlibrary.html).

Collaborative portal developments have the potential to complement individual library portals by providing access to a wider range of national and international information resources and document delivery services. They also reduce duplication of effort through use of common hardware, software and systems administration. The Denmark electronic research library's deff.dk service provides web page links to portals, databases and research library resources (http://www.deff.dk/). The Association of Research Libraries in the USA has developed specifications for a Scholars Portal which will aggregate, integrate and delivery a licensed and openly available digital content across a broad range of subject fields and from multiple institutions.

The authenticated network graded environment for learning (Angel) project in the UK is creating middleware services to integrate learning environments with digital library developments. Angel plans to address the issue of appropriate end user authentication and access management (Angel, 2002). The development of an authentication and access management system will be a major part of the wider implementation of AARLIN over the next two years. AARNet is not at this stage providing access to undergraduate students or linking to online learning resources however these are future possibilities.

Although libraries have been using portals for about four years and they are installed in more than thirty-six U.S. libraries, the number of clients using portal services is relatively small at around 5%. Attention will need to be paid to the protection of user confidentiality and the design of the portal if use is to be increased (Crawford, 2002).

Portal services are also provided by commercial organisations directly to researchers. TheScientificWorld (www.thescientificworld.com/) is an integrated scholarly web portal to services, resources and products intended to enhance and accelerate the research efforts of science professionals. Services which are provided on a fee for service basis include e-publishing, personalised e-mail alerts, forthcoming information, the ability to prepare and submit funding applications online and procurement of scientific equipment and supplies (McKiernan 2002).

AARLIN aims to use "push" technology to provide current awareness services to researchers. The challenge will be to provide this service in a way that does not create information overload for the user. In the longer term it may be possible to utilise an intelligent agent system to assist in pushing information to users in the portal environment (Martin and Metcalfe, 2001).

4 Implementation of the pilot project

Planning for and implementation of the pilot project were undertaken during 2001. This involved selection and installation of hardware and software; training of library staff and researchers, programming and configuration of databases. Researchers who attended the training expressed enthusiasm for the service.

The AARLIN pilot project aims to establish proof of the AARLIN portal concept. Issues being explored during the pilot project include the availability of suitable software for the components of the portal and the usefulness of the portal to researchers.

The pilot project is the precursor to a fully operational service which will be implemented on a staged basis in Australian university libraries. A tender process for the selection of software for the operational service commenced in early 2002.

The software and hardware required for the system was installed during 2001. Available portal software was evaluated for its functionality, cost and other features. Ex Libris' Metalib and SFX software was selected for the pilot project. IBM equipment was installed at La Trobe University running on Linux. The Australian Academic Research Network (AARNet) provides communication facilities for the project.

The Metalib software is used to search resources in a range of subject areas. Searches can be made of multiple resources and searches can be saved. Users can set up a personal profile. Once search results are obtained, the SFX software is used to access the selected items online in full text format or via library catalogues or document delivery services.

For the purposes of the pilot project a centralised authentication system was used. It is planned to develop "handshaking software" using. XML which can communicate with the local authentication systems of participating universities for the operational service. In connection with this, a survey of the existing authentication systems is being undertaken jointly by CAUL and CAUDIT. One of the objectives of the survey is also to ascertain what additional profiling metadata may need to be added to the directory services used by the local authentication systems. Thus it is intended that the authentication systems will also provide the profiling data that can be used to "push" relevant research resources to the users.

It was decided to develop the portal, for the purposes of the pilot, in the major research areas of health sciences/medicine, engineering and humanities. Six of the twenty participating universities were selected for participation in the pilot project on the basis of their ability to provide library liaison staff and researchers in these research areas. The selected universities were La Trobe University, Swinburne University of Technology, Victoria University of Technology, Murdoch University, Flinders University and the University of Canberra.

A number of resources have been configured for access via the Metalib/SFX software. These include full text serial databases, indexing and abstracting databases, library catalogues, research information databases, web subject gateways and document delivery services.

Several research information databases are already accessible via AARLIN and it is planned to add more when the portal is fully operational. Research Finder is a database of Australian research, including university research which is maintained by the Australian Commonwealth Government (http://www.industry.gov.au/science/ResearchFinder). The Australian Digital Theses (ADT) service provides access to completed Australian university higher degree theses (http://adt.caul.edu.au/). VOCED is maintained by the National Council for Vocational Education Research (NCVER) and provides access to information about research projects in the Australian Technical and Further Education (TAFE) sector (http://ww.voced.edu.au).

5 Survey

A survey was conducted of researchers who were participants in the pilot project in order to ascertain their satisfaction with the service and to assist in identifying potential areas for improvement when the operational system is implemented. The survey was conducted at the start of the project. A further survey will be conducted at the end of the project in order to compare researchers' anticipated use of the service with their actual use.

Seventy-four researchers from the six pilot project institutions responded to the prepilot survey. Respondents were asked to indicate their discipline area; 32.9% indicated that they were

from medicine or health, 23.3% from humanities; 11% from engineering and 32% were from a variety of other disciplines.

Most respondents (86.1%) indicated that their main reasons for using the portal would include research; 33% included teaching preparation and support and 21% included current awareness.

62% of respondents indicated they expected to use the portal weekly, 70% monthly, 16% daily and 4% other frequencies.

5.1 Frequency

Respondents were asked the frequency with which they expected to use the various portal resources and services. The responses are shown in Table 1.

Table 1: Within AARLIN, how often do you expect that you will use /search the following?

		1 Never	2 Once or twice	3 Fort-nightly	4 Weekly	5 Daily	No response
a	Indexing and abstracting databases	2.9% (2)	30.9% (21)	17.6% (12)	41.2% (28)	7.4% (5)	
b	Searchable e-journal collections		9.7% (7)	31.9% (23)	45.8% (33)	12.5% (9)	
c	Subject Gateways (EEVL, OMNI, BIOME, AVEL)	13.3% (10)	28% (21)	24% (18)	21.3% (16)		13.3% (10)
d	Search engines (Yahoo, Alta Vista	12.5% (9)	15.3% (11)	31.9% (23)	25% (18)	15.3% (11)	
e	Recommended websites (Library web pages and other websites "recommended" by the library)	7% (5)	31% (22)	29.6% (21)	28.2% (20)	4.2% (3)	
f	Own University catalogue	2.8% (2)	9.7% (7)	34.7% (25)	38.9% (28)	13.9% (10)	
g	Other Library catalogues	2.8% (2)	33.3% (24)	47.2% (34)	12.5% (9)	4.2% (3)	
h	Other "combined" library catalogues (e.g Coolcat, Kinetica, Serials in Australian Libraries)	8% (6)	40% (30)	26.7% (20)	13.3% (10)	4% (3)	8% (6)
i	Interlibrary loans/document delivery requesting	11.1% (87)	37.5% (27)	34.7% (25)	13.9% (10)	2.8% (2)	
j	SFX, to find out other services available	13.3% (10)	30.7% (23)	26.7% (20)	14.7% (11)	1.3% (1)	12.3% (10)
k	SFX, to access e-journal articles	8% (5)	16% (12)	32% (24)	30.7% (23)	2.7% (2)	10.7% (8)
l	SFX, to check availability of cited item, in local library catalogue	6.7% (5)	22.7% (17)	32% (24)	26.7% (20)	1.3% (1)	10.7% (8)
m	Other usage (give details)	14.7% (11)	2.7% (2)	10.7% (8)	2.7% (2)	1.3% (1)	68% (51)

The resources and services for the highest anticipated weekly use included searchable e-journal collections (33), indexing and abstracting databases (28) and participants own libraries' catalogues (28). Highest anticipated fortnightly use included use of other libraries' catalogues (34), use of own libraries' catalogue (25), interlibrary loans/document delivery requesting (25), use of SFX to access e-journal articles (24) and use of SFX to

check availability of cited items in local library catalogues (24). Overall, use on a daily basis was expected to be lower than weekly or fortnightly use.

Of the categories of resources and services available, respondents indicated that they expected the indexing and abstracting databases to be most useful (28) followed by e-searchable e-journal collections (47). Resources and services nominated as likely to be the least useful were search engines (Yahoo, Alta Vista) and recommended websites.

Respondents suggested a number of resources and facilities that they would like to see added to the portal. These included improved access to library databases and document delivery services. The addition of other research services, such as links to company research information, news or emails groups for scientists and an interface to Endnote were also suggested.

5.2 Functionality

Researchers were asked how useful they expected to find various aspects of the portal's functionality. The results are given in Table 2.

Table 2: How useful do you expect to find the following functionality?

		1 Useful	2 No opinion	3 Not useful	4 Never used
5a	Can search multiple databases (targets) in parallel (e.g. can search 5 databases at the same time)	94.6% (70)	4.1% (3)	1.4% (1)	
5b	Can search different types of targets in parallel (e.g. 2 Subject Gateways at the same time as 4 abstracting databases)	71.6% (53)	27% (20)	1.4% (1)	
5c	Can save searches between sessions, and re-run searches (using History)	85.1% (63)	9.5% (7)	4.1% (3)	1.4% (1)
5d	Can mark records and add to e-shelf, (and contents of e-shelf are retained between sessions)	83.8% (62)	13.5% (10)	1.4% (1)	1.4% (1)
5e	Can save records in original format, to import into Endnote or Procite	74.3% (55)	12.2% (9)	1.4% (1)	12.2% (9)
5f	Can use the Locate resources function, to search for "resources" (targets) by "type", discipline, etc	56.2 (41)	38.4% (28)	1.4% (1)	4.1% (3)
5g	Can use the Locate resources function, to add "resources" (searchable or link-to targets) to "My (fave) Resource List"	56.2% (41)	32.9% (24)	5.5% (4)	5.5% (4)
5h	From any "full-record"/citation viewed in Metalib, can link (using SFX) to a list of optional services, such as check local opac for this item, link to full-text of item, request item through Interlending and Document Delivery, etc.	75% (54)	16.7% (12)	4.2% (3)	4.2% (3)
5i	Interlending and Document Delivery requesting form – is populated with details of item to be requested.	67.6% (50)	23% (17)	1.4% (1)	8.1% (6)
5j	Metalib navigation buttons at the top provide easy navigation	66.2% (47)	28.2% (20)	1.4% (1)	4.2% (3)
5k	Metalib provides on-line "HELP"	56.9% (41)	34.7% (25)	2.8% (2)	5.6% (4)

Almost all researchers (70) expected searching multiple targets in parallel to be useful; 63 expected saving searches between sessions to be useful; and 62 expected the ability to mark records and add to an e-shelf to be useful.

The majority of respondents expected all functions to be useful. A relatively large number of respondents stated that they had no opinion on the potential usefulness of aspects of the functionality, probably because of lack of familiarity with these particular functions.

In response to a subsequent question which asked respondents to indicate which function they expected to find most useful, 47 respondents stated that they expected to find the ability to search multiple databases most useful.

5.3 New features

Researchers were asked which of a list of new features they would like to have incorporated into AARLIN. The responses are given in Table 3.

Table 3: We are considering incorporating the following features into the AARLIN portal in the future. Please indicate the extent to which you would expect to find these aspects useful:

		1 Very Useful	2 Useful	3 No opinion	4 Negligible usefulness	5 Never used
a	Capacity for you to add / modify your favourite "bookmarks" (as compared with library-selected resources) within AARLIN	31.1% (23)	54.1% (40)	12.2% (9)	1.4% (1)	1.4% (1)
b	Capacity for you to "personalise" your AARLIN environment	40.5% (30)	41.9% (31)	10.8% (8)	4.1% (3)	2.7% (2)
c	Capacity for you to see an indicator of the "local availability" of items without having to link to local catalogue first.	62.2% (46)	21.6% (16)	13.5% (10)	2.7% (2)	
d	Provide an indicator, while you are viewing citation details, of whether full-text article is available electronically.	81.1% (60)	16.2% (12)	1.4% (1)	1.4% (1)	
e	Capacity for you to use the same authentication, both in your University/library environment and in the AARLIN portal (e.g. removal of need to authenticate more than once)	62.5% (45)	22.2% (16)	12.5% (9)	1.4% (1)	1.4% (1)
f	Capacity for you to view the status of your Interlibrary Loans quota (where relevant)	21.6% (16)	48.6% (36)	16.2% (12)	13.5% (10)	
g	Capacity for you (using e-commerce) to purchase services beyond standard services (e.g. where relevant…to purchase document-delivery when quota has been exceeded, or to request fast-track delivery of items)	17.8% (13)	27.4% (20)	31.5% (23)	16.4% (12)	6.8% (5)
h	Capacity for you to link easily between AARLIN portal and the teaching/learning environment of your institution.	29.7% (22)	37.8% (28)	21.6% (16)	10.8% (8)	
i	Capacity for the library to "push" news to users about new resources and so on (focussing messages so that they are presented only to relevant user-groups), at the time that users log in to the portal.	14.9% (11)	36.5% (27)	28.4% (21)	14.9% (11)	5.4% (4)
j	Capacity for each user to set up "auto-alerts" within the AARLIN portal (Auto-alerts are searches predefined by you, and run regularly against selected databases, with the results emailed to you)	45.9% (34)	41.9% (31)	9.5% (7)	2.7% (2)	
k	Capacity for you to browse and search thesauri of databases from within the AARLIN environment (where the native interface of a database currently offers thesaurus searching/ browsing)	20.3% (15)	44.6% (33)	28.4% (21)	6.8% (5)	
l	Additional options for field searching (e.g. capacity to search within the Abstract field, the Notes field, etc; or capacity to search within a specific timespan, such as "after 1997")	44.6% (33)	43.2% (32)	8.1% (6)	4.1% (3)	

m	Greater context sensitivity in search interface (eg choice of searchable fields alters when only ONE database has been selected, and when that database can offer additional fields to be searched)	27% (20)	40.5% (30)	25.7% (19)	5.4% (4)	1.4% (1)
n	Greater context sensitivity in SFX-like services offered (eg Interlending and document requesting is only offered when item is not available locally)	28.4% (21)	36.5% (27)	28.4% (21)	4.1% (3)	2.7% (2)

An indication of the availability of full text while using citation details was the most highly ranked possible new feature; 60 responses ranked this as very useful. Other possible new features ranked as very useful included renewal of the need to authenticate more than once and capacity to see the local availability of this without having to link to the local catalogue first (both 62%). The next most important were the ability to receive automatic emails about predefined searches (45%), additional field searching options (44%) and greater capacity to personalise the AARLIN environment (40%).

The ability to "push services to clients was regarded as "very useful" by only 11 respondents, although a further 28 noted it as "useful".

The capacity to link only between AARLIN and the teaching and learning environment of the institution was rated as "very useful" by 22 respondents and "useful" by 28.

5.4 Value

Respondents were asked whether they thought that their information searching and client access processes would be enhanced by AARLIN services. 59 respondents strongly agreed that they expected their research to be enhanced by the AARLIN portal making it easy to access full text; 47 by the AARLIN portal providing a one stop shop; 39 by the AARLIN portal enabling results to be viewed in a standardised format and 34 with the portal offering a list of personal resources. Very few respondents disagreed with the statements and none strongly disagreed; several respondents reported that they had no opinion.

Researchers were asked to describe their current information searching and document access processes. 9.3% reported them as highly satisfactory, 60% is acceptable and 26.7% is unsatisfactory.

Researchers were asked whether their research endeavors would be enhanced by increased awareness of delivery of resources. The majority of respondents strongly agreed that their research would be enhanced. Researchers were then asked whether the AARLIN portal would increase their awareness. Responses to this were mixed with a substantial number having no opinion; comments made under this response suggest that respondents were not yet familiar enough with the AARLIN functionality to respond.

Participants were asked whether they currently used any other portal sites. Eleven reported use of Google, 8 of Yahoo and 3 of Alta Vista.

6 Conclusions

The results of the AARLIN pilot project participant researchers survey suggest that the concept of a library portal for researchers is a viable one. Researchers were generally positive about the portal service and saw it as having significant potential for improving their information searching. The survey results also provide some areas for consideration in the future development of the portals.

Development of improved database search and delivery features should be a high priority. The survey results suggest that the main value of the portal to researchers will be in improved searching of library subscription databases and library catalogues.

Particular attention should be paid to delivery of full text direct to the desk top with transparency between service providers.

The survey results suggest that an authentication system which is based on participants' university authentication systems is desirable.

Although AARLIN was designed as a research portal, there could be value in developing links to teaching and learning systems. A significant number of participants indicated that they expected to use the portal to assist in teaching as well as in research. A significant number also viewed improvements in this area to be of value.

It would appear undesirable to commit substantial resources into the development of e-commerce services for library research portals without further investigation of the market. Respondents expressed little interest in this, possibly because access to information resources of this type is currently charged library funds. The potential to develop specialised services for charging to departmental or research centre funds could be considered.

Consideration could also be given to expanding the resources in the portal beyond traditional library information. Comments included in the survey suggest that it could be useful to have links to other types of tools such as news and email groups and to more non-bibliographic research databases.

The grouping of resources by subject area needs to be reviewed. When AARLIN was first conceived it was planned to start with a few key subject areas. The researchers who are participating in the pilot project are from a wider range of areas than originally planned and a larger list of subject areas has been developed. Given the cross disciplinary-nature of much research and of many information resources it may be appropriate to give further consideration to the most useful grouping of resources.

The findings of the AARLIN pilot project must be regarded as tentative at this stage, pending the results of the second participant survey which will be carried out at the end of the pilot project. It will be important to compare the results of the two surveys and to ascertain the impact of familiarity with the service on participants' views.

7 References

ANGEL (2002) Authenticated Network Guided Environment for Learning. Available at *http://www.angel.ac.uk*

ARL (2001) Access and Technology Program. Scholars Portal. Available at *http://www.arl.org/access/scholarsportal/index*

CAUDIT (2000) Survey on portal software at Australasian universities. Available at *http://www.caudit.edu/caudit/surveys/00portals.html*

Crawford, Walter (2002) 'Talking 'bout My Library' American Libraries; The Crawford Files 4 Available at *http://www.ala.org/alonline/crawford/cf402.html*

Martin, Paul and Metcalfe, Michael (2001) Informing the Knowledge Workers Reference Services Review 29 (4) pp 267-275.

McKiernan, Gerry 'E-profile': The Scientific World: an integrated scholarly knowledge network Library Hi Tech News 2 2002 pp 21-29

8 Contact Information

Doreen Parker
University Librarian
Victoria University of Technology
PO Box 14428 MCMC
Melbourne, Victoria, 8001
Australia

Telephone 61 3 9688 5097
Facsimile 61 3 9688 4117
e-mail: doreen.parker@vu.edu.au

Information Retrieval in Distributed Environments Based on Context-Aware, Proactive Documents[1]

Michael Friedrich, Ralf-Dieter Schimkat, Wolfgang Küchlin
Wilhelm-Schickard-Institute for Computer Science, University of Tübingen, Germany

Summary

In this position paper we propose a document-centric middleware component called *Living Documents* to support context-aware information retrieval in distributed communities. A *Living Document* acts as a micro server for a document which contains computational services, a semi-structured knowledge repository to uniformly store and access context-related information, and finally the document's digital content. Our initial prototype of *Living Documents* is based on the concept of mobile agents and implemented in Java and XML.

1 Introduction: *Living Documents* and Information Retrieval

Considering multiple user perspectives, a conceptual model of information retrieval dealing with text or multi-media documents should integrate different views on documents. In addition it should allow queries addressing each of these document views separately, as well as queries for combinations. Each view is described in terms of meta data and therefore associated with the document.

Generally, meta data is data that describes other data to enhance its usefulness (Marshall 1998). We see context information as meta data describing different orthogonal aspects of the respective document. This meta data is characterized by its diversity and continuous evolvement. Incorporating such kind of meta data into the retrieval process allows to improve the overall precision because more relevant and accurate information can be comprised into the overall retrieval process.

The major problems in querying flexible attributes for meta data are the
- proprietary encoding and accessing schema for meta data of each document view.
- continuous creation and updating of meta data related to document views (temporal aspects).
- decentralization and distribution of documents' view meta data.

In decentralized communities documents should be shared, cloned and downloaded for use without a network connection. With these requirements several problems arise, like:
- Where should context information be stored?
- How can documents be processed on only temporarily connected clients?
- How can the evolution of contexts be tracked in such environments?

Our approach is to provide a general middleware component which accompanies documents during their entire document life cycle. The middleware component provides facilities to uniformly access the context information which is represented in a flexible XML-repository. Furthermore, the concept which we call *Living Documents* supports several retrieval paradigms (namely: reactive, proactive and cooperative retrieval).

1 Supported by the *Ministerium für Wissenschaft, Forschung und Kunst* of the state Baden-Württemberg, Germany. Project: Verbund Virtuelles Labor (www.vvl.de)

2 Design of *Living Documents*

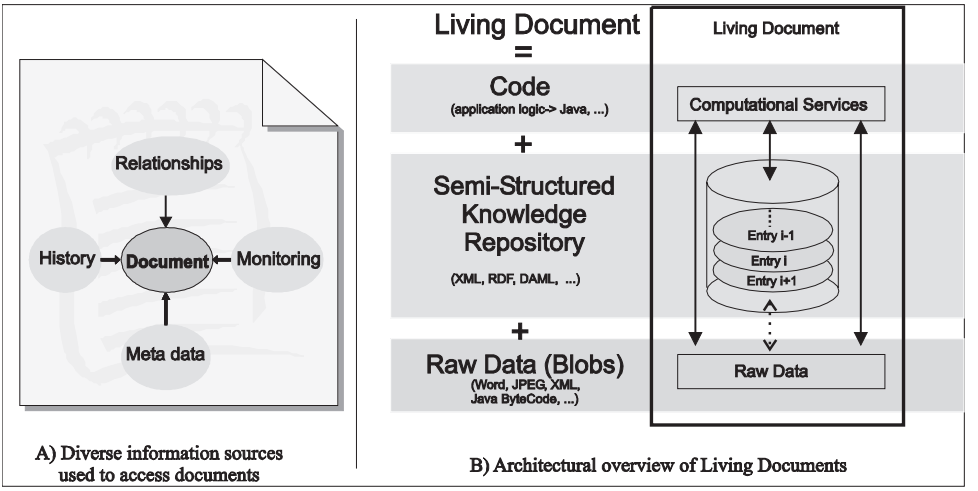


Figure 1: A) Examples for various different types of information sources which improve access to and search of documents in digital libraries. B) Living Documents are divided into three sections: Raw Data carries the documents to manage, Semi-Structured Data contains all meta data about the managed documents and the Code section keeps the computational services for accessing a Living Document and processing incoming requests (i.e. queries).

2.1 Towards a micro server architecture

A Living Document (*LD*) is a logical and physical unit consisting of three parts, as depicted in Figure 1B:

1. code or computational services
2. semi-structured knowledge repository
3. raw data

2.1.1 Computational Services

The *Computational Services* are essentially *code* fragments which provide several facilities, such as access and query capabilities or general application services. An example for such an application service is a viewing component for a document which is encoded in XML (World Wide Web Consortium (2001) . The code fragments determine the degree of activeness of a *LD* ranging from passive documents which are enriched with some arbitrary application logic to proactive documents.

The two other parts of a *LD* are accessed through the code section to ensure the integrity of the whole component. The creation and modification of the knowledge repository and the raw data and even the code itself is controlled by this layer which inhibits malicious attempts from outside to temper with the data.

2.1.2 Knowledge Repository

The *Knowledge Repository* of a *LD* provides facilities to store and retrieve the document's meta data or other related context information as depicted in Figure 1A. Each document has its own knowledge repository describing the content of the raw data section. This information is referred to as a document state information. A set of document state information builds a so-called document state report (DSR) which contains information about document behaviour, its history and reflects the contexts of this document. The context of a document reflects who uses it when, where and with which goal in mind. We present more details about contexts of *LD*s in Section .

Each DSR is encoded as a semi-structured XML document according to the *SpectoML* (Schimkat et al. 2000). Following a XML-based implementation, the generation of DSR is accomplished in an uniform way which neither does favour a particular data format nor the use of special programming or scripting languages. The use of XML as the primary data format for document state information enables a DSR with query capabilities, such as the execution of structured queries to each document state information. Therefore, a DSR builds a XML-based knowledge repository which holds all context information about the entire document life cycle.

2.1.3 Raw Data

The *Raw Data* part contains any information encoded as a digital document such as a word processing document, a music file or even serialized application code. This data is the actual information to be stored and which is described by the knowledge repository.

2.1.4 Putting it all together

The whole entity named *LD* is an atomic and self-sufficient unit which serves as a micro server for documents since it carries their logic, state and data along themselves. It can be seen as a micro edition of a huge client-server document management system for only one or a few documents. The contained meta data and logic allows the *LD* to operate in a mobile environment because every part of it is inextricably glued together. Therefore, a *LD* is equipped for only temporarily and highly distributed environments.

Finally, why is a *LD* called *living*? The state including the context of a *LD* which is represented in the knowledge repository, is characterized by its dynamic structure. Its data is growing and changing as the *LD* is used and the context of the embedded document changes. Apart from that the raw data section and even the code section of the *LD* can evolve over time. This constant change in representation and behaviour leads to the name *Living Documents*.

2.2 Implementation path

We are currently experimenting with a component architecture based on mobile agents. The aspect of mobility contributes to the notion of „living" in *Living Documents as well*. Every *LD* is represented by an agent which can be mobile in principle. Mobile documents support the asynchronous nature of communities based on only partly connected clients like PDAs or notebooks. It allows sharing and keeping track of documents without the need of a network connection. Moreover, the agent paradigm provides a basis for real distributed communities without centralized servers and services.

We chose the mobile agent paradigm to be most flexible to distribute the *LD*s within a computer network. However, we like to emphasize that using mobile agents as basic implementation components does not restrict the concept of *LDs* to this domain. Though unlikely, other component models can be considered as more suitable in the future.

3 Implications of design

3.1 Uniform access

Living Documents provide uniform access to any kind of context-related information. Therefore, *LDs* serve as a middleware layer (Bernstein 1996, Geihs 2001) for accessing context-related information. This behaviour is essential for middleware services in general and from a rather general point of view similar to accessing relational databases uniformly by using a communication middleware component such as JDBC (Java Database Connectivity).

The uniform access of meta data is essential for *LD*s to interoperate and collaborate in groups or communities. This does neither imply that all knowledge repositories are structured the same way nor that all share the same set of access functions in the code section. The structured knowledge repository enables other *LD*s to query the stored context information in a determined way.

3.2 Context

Obviously, a document can be shared between more than one community. So the different views of a *LD* are twofold: First, each user has its view on the documents, and secondly, each community defines its own view on its subset of all documents. Consequential, *LD*s are versioned documents[2] as they change their behaviour with respect to their current context. For one person a *LD* might look as a picture document (i.e. a JPEG in the raw data section), another one might only be interested in the meta data, for instance in the author of that document. An example for such a deployment by distinct communities is different access rights. Thinking of a staff administration, everybody might see the telephone numbers but only few people are allowed to see and change the salary.

The context of a *LD* depends also on its physical location (”On which host am I?”) and on its logical location (”Which other *LD*s are nearby?”). The physical location determines the available resources like network bandwidth, memory and computing power, for example whether it resides on a PDA or a Workstation permanently connected to the Internet. The logical location affects the functionality of an *LD* if it relies on others to perform a task. As collaboration between *LD*s can improve their capabilities, one *LD* might use some specialized functions of another. This conforms to our earlier definition of a LD as a self-contained unit, because only fundamental functions for accessing LDs have to be provided. Building more complex systems out of these components goes well with the design and is strongly encouraged.

The items above affect the behaviour of a *LD* by means of general environmental conditions. The knowledge repository keeps document related information. Generally, a knowledge repository is a collection of sentences in a representation language that entails a certain picture of the world presented (Levesque & Lakemeyer 2000). For the domain of *LD*s the world is meant to be the documents’ world and the representation language is *SpectoML*. Having a knowledge repository entails being in a certain state of knowledge where a number of other properties hold. Assigning a knowledge repository to each *LD* provides several benefits for managing a DSR such as (a) easy adding of new, context-related document state information by making them dependent on the previous knowledge contained in the repository, (b) extending the existing DSR by adding new beliefs and document artifacts, (c) the possibility to explain and justify precisely the current document’s state. In our previous work (Heumesser & Schimkat 2001) we have shown how to deploy logical inference mechanisms on top of an XML markup language such as *SpectoML*.

2 Versioned documents refers to the capability of *Living Documents* to provide different and complex views on documents dynamically. It does not mean that there are different versions of one document.

3.3 Proactive *LD*s

With *LD*s acting as micro servers for documents, information retrieval can either be accomplished *reactive* in a Client-Server based way, *proactive*, as the document itself triggers a query, or *cooperative* manner, where multiple *LD*s work together to achieve a common goal. The latter paradigm can make use of a distributed environment, where the retrieval process is done in parallel by multiple components.

By deploying *LD*s the distinction between documents and applications blurs, because documents can contain (and mostly do contain) application logic which defines their behaviour. Therefore, the location of an application is hard to define during runtime. This spreading of functional components is issue of current research (Zambonelli & Parunak 2002). Application development using *LD*s follows this trend by distributing the application logic over interconnected nodes.

A proactive *LD* initiates complex tasks, discovers new services and is more than just a reactive component. For example, a *LD* can gather information over a network (e.g. web pages), process this data and continuously check the original source for updates. Another example is a persistent query for an information system, which is performed when the data source of this system changes.

3.4 Flexibility

Our document centric approach trades flexibility for performance. The basic idea is to provide a knowledge repository which is flexible and open because the set of tasks built upon this paradigm might not be predictable in advance due to the complexity and non-determinism of the documents' world and the surrounding environment. Therefore it is feasible to make the documents knowledge explicit since we might not know in advance how the knowledge will be used. Furthermore, there might be a wide variety of different context-relevant application and information sources interfacing to the documents knowledge repository. Each of them might have its proprietary application logic for dealing with the externalized *LD*s knowledge. It is viable therefore, to have a flexible, open architecture at hand when it comes to building real world applications.

4 Future Work

Initial experiments with our prototype implementation of *Living Documents* based on mobile agents show promising results. However, there are still several open issues to explore more deeply. Among these are the aspects document usage policies like sharing, cloning and distributing documents, and a thorough performance and implementation analysis.

Since the described concept of *Living Documents* is related to intensional documents as discussed by Schraefel et al. (2000), we currently explore appropriate facilities to create an active hypertext of interlinked *Living Documents* which would build a basis for further improvements of the information retrieval process within such distributed document spaces.

5 References

Bernstein P. (1996): Middleware: A Model for Distributed System Services. In: *Communications of the ACM*, Vol. 39, No. 2: p. 86 – 98.

Geihs, K. (2001): Middleware challenges ahead. In: *IEEE Computer*, Vol. 34, No. 6, p. 24—31.

Heumesser,B. D.; Schimkat, R.-D. (2001): Deduction on XML documents: A case study. In: *Proceedings of the 14th International Conference of Applications of Prolog (INAP 2001) - Stream Content Management*, p. 20 – 29.

Levesque, H. J.; Lakemeyer, G. (2000): *The logic of knowledge bases*. Cambridge, Massachusetts: MIT Press.

Marshall, C. C. (1998): Making metadata: a study of metadata creation for a mixed physical-digital collection. In: *Proceedings of the third ACM Conference on Digital libraries*, ACM Press. p. 162 - 171.

Schimkat, R.-D.; Häusser, M; Küchlin, W; Krautter, R. (2000): Web application middleware to support XML-based monitoring in distributed systems. In: Debnath, N. (Ed.): *Proceedings of 13th International Conference on Computer and Applications in Industry and Engineering (CAINE 2000)*. International Society for Computers and Their Applications. p. 203-207.

Schraefel, M. C.; Mancilla, B.; Plaice, J. (2000): Intensional hypertext. In: Gergatsoulis,M.; Rondogiannis, P. (Eds.): *Intensional Programming II*. Singapore: World-Scientific. P. 40 – 54.

World Wide Web Consortium (2001), http://www.w3.org/TR/REC-xml. *Extensible Markup Language (XML) 1.0*.

Zambonelli, F.; Parunak, H. V. D. (2002): From Design to Intention: Signs of a Revolution. IN: Proceedings of Autonomous Agents and Multi-Agent Systems (AAMAS 2002). To appear.

6 Contact Information

Michael Friedrich, Ralf-Dieter Schimkat, Wolfgang Küchlin
Wilhelm-Schickard-Institute for Computer Science
University of Tübingen
Sand 13
D-72076 Tübingen, Germany

e-mail: {friedrich, schimkat, kuechlin}@informatik.uni-tuebingen.de
http://www-sr.informatik.uni-tuebingen.de

Integration via Meaning: Using the Semantic Web to deliver Web Services

Brian M Matthews
CLRC, UK

Summary

The major developments of the World-Wide Web (WWW) in the last two years have been *Web Services* and the *Semantic Web*. The former allows the construction of distributed systems across the WWW by providing a lightweight middleware architecture. The latter provides an infrastructure for accessing resources on the WWW via their relationships with respect to conceptual descriptions. In this paper, I shall review the progress undertaken in each of these two areas. Further, I shall argue that in order for the aims of both the Semantic Web and the Web Services activities to be successful, then the Web Service architecture needs to be augmented by concepts and tools of the Semantic Web. This infrastructure will allow resource discovery, brokering and access to be enabled in a standardised, integrated and interoperable manner. Finally, I survey the CLRC Information Technology R&D programme to show how it is contributing to the development of this future infrastructure.

1 Introduction

The major initiatives in the development in the support of information system across the World-Wide Web (WWW) in the last two years have been *Web Services* and the *Semantic Web*, both of which have been taken up as major activities of the World-Wide Web Consortium (W3C), the leading standards body on the WWW.

Web Services have been one of the most talked about emerging technologies with major initiatives from the large manufacturers, such as IBM and Microsoft. With this major commercial backing it is seen as the next big step of the development and exploitation of the Intranet and Web infrastructure already in place so that businesses and other communities, such as science, government and education, can communicate and interact together in a flexible and automated manner.

The Semantic Web on the other hand has had a much longer development, and while it still has a degree of hype surrounding it, there has been much less commercial support. It seems to be met with as much mystification as excitement, as people do not understand what it seeks to achieve, and when they do understand, they cannot see how it can be used in practice.

However, the aims of these two initiatives are complementary, and in the long term very similar; to evolve the World-Wide Web into a seamless interactive, and automatic system of the future. In a sense, they both aim to make the Web both ubiquitous and invisible.

In this paper, we shall discuss the current state of the development of Web Services and the Semantic Web. We shall go on to discuss how they are complementary, and what areas can be supported through their combination, and then discuss the current programme of work at CLRC to move towards this combination.

2 Web Services

There is a definite buzz in the air at the moment about Web Services. This method of loosely integrating different application on the Web into a unified service is being proposed as a universal mechanism for seamlessly integrating services offered across the Web to provide integrated distributed systems. They have become the cornerstone of the enterprise integration platforms of major vendors such as Microsoft (.NET) and IBM (IBM Web Services). Indeed, it was the submission of the SOAP proposal to W3C by an industrial consortium in May 2000 which kick started the whole area.

Others who have been developing different approaches to developing open integrated distributed systems have been forced to change direction to accommodate the emerging consensus on Web Services. For example, Sun Microsystems has introduced classes supporting Web Services to its Java Enterprise Infrastructure (Sun Microsystems 2001). The ebXML initiative which was developing an alternative architecture for message passing via XML has changed course to use Web Services as a base technology underlying its customised approach (ebXML, Irani 2001). The Grid concept, an approach to developing large-scale distributed computing infrastructures to support high-performance computing, which has created particular interest within scientific applications (Foster & Kesselman 1999), has been seen as the natural successor to the Web. Nevertheless, the Grid community has recently recognised the usefulness of Web services as an underpinning of the further development of the Grid (Foster et. al. 2002); it is likely that in the future there will be a convergence Web and Grid technology around Web Services. A recent paper has also noted the complementary nature of Web Services and the well-established middleware architecture CORBA (Gokhale et. al. 2002). That Web Services is the method of choice for providing web-enabled integration of computing resources has become the consensus in the community is further supported by the UK government's *Electronic Government Interoperability Framework* recent recommendation of Web Services as the preferred integration method for the electronic delivery of public services (eGif 2002).

Nevertheless, despite this large amount of activity and publicity, there are few public examples of the use of Web Services in action, and there is remarkably little consensus of what Web Services are and what they offer! However, most definition seem to agree that Web Services require three basic components:

- **A Messaging Service**: a standard protocol for communicating between resources on the Web. The most common means of supporting this is currently SOAP (SOAP 1.2 Parts 1 & 2), although the W3C is now developing a standard recommendation for this purpose (XML Protocol).
- **An Interface Description Language:** a mechanism for services to describe their interfaces in a standard fashion, so client applications know what form (types) of messages to send to them, and what form the client should expect the response to be. The most common method for this to date is the Web Service Description Language (WSDL).
- **A Registration Service:** a mechanism of registering web services so that clients looking for a service to satisfy their requirements can find them easily. The Universal Discovery Description and Integration format (UDDI) is the most common mechanism proposed for this purpose.

In a sense, there is not very much new here. What these together provide is a *loosely coupled middleware solution.* It allows a relatively simple mechanism to publish and combine services together to provide integrated yet distributed systems, with out the need for expensive proprietary middleware components. It is based on open standards (TCP/IP, XML), and allows the integration of systems based on different platforms, which are independently provided and maintained. It is these advantages that make it more attractive than existing middleware solutions such as CORBA or DCE. Nevertheless, the basic infrastructure is relatively weak, and there are

many proposals as to what to add to the basic infrastructure. In particular, the aspects of *security, resource description, discovery, brokering and negotiation,* are all missing, whilst the general problem of how to compose web services into new ones has hardly be approached (see for example Florescu et al 2002).

3 Semantic Web

The Semantic Web on the other hand has had a lot of publicity, but less enthusiastic commercial support. It has been driven by a more academic approach, and has yet to prove itself within widespread application. Nevertheless, it has an ambitious long-term aim; nothing less than imbuing the Web itself with meaning. That is, providing meaning to the network of *resources* available on the web and, perhaps more importantly, meaning to the *links* that connect them (Koivunen & Miller 2001). Once the web has a mechanism for defining semantics for resources and links, then the possibility for *automatic processing* of the Web by software agents, rather than the constant mediation by people providing meaning to the Web.

This ambitious aim has been there from the beginning (Berners-Lee 1989) as Tim Berners-Lee's original description of the WWW included types for objects and links. However, it was not until the Semantic Web Roadmap (Berners-Lee 1998) that the initiative became fully underway.

The Semantic Web has been developing a layered architecture:

- **Resource Description Framework:** A basic knowledge representation language, describing a graph model and XML format for describing relationships between resources (Lassila & Swick 1999).
- **RDF Schema Language (RDF Schema):** a basic type modelling language for describing classes of resources and properties between them in the basic RDF model (Brickley & Guha 2002).
- **Ontologies:** a richer type modelling language for providing more complex constraints on the types of resources and their properties. The current most complete work is DAML+OIL (Connolly et.al. 2001).
- **Logic and Reasoning:** an (automatic) reasoning system provided on top of the ontology structure to *make new inferences.* Thus using such a system, a software agent can make deductions as to whether a particular resource satisfies it requirements (and vice versa) (e.g. RQL (Karvounarakis et. al. 2002).

Thus the Semantic Web initiative has an ambitious programme to bring existing work on knowledge representation and reasoning to bear on the largest information resource of all. Nevertheless, the work on all areas has been slow, with the process returning to its beginnings on several occasions. This has given the impression that the activity is of largely academic interest, whilst of course the application and potential of this work is enormous.

Applications of RDF have emerged however, notably, Dublin Core (Miller et al 1999), RDF Site Summary (Beged-Dov et. al. 2000), Composite Capability/Preferences Profiles (Klyne et. al. 2001), and most recently proposals for the Protocol for Internet Content Selection (Brickley & R.Swick 2000) and Protocol for Privacy Preferences Project (McBride et. al. 2002). These are all applications that are ideal for the Semantic Web as they describe properties of web based resources. Nevertheless, each individually could be described using some domain specific method, and possibly in a more succinct manner. What has yet to be demonstrated is the benefits to be gained from expressing such applications within a single framework, one that allows semantics based interoperability. Web Services provides an ideal environment where the advantages of a joined up semantic approach could be demonstrated.

4 Integration via Meaning

In a briefing that has subsequently become well-known Keith Jeffery proposed a three layer architecture for considering a future distributed information systems architecture (Jeffery 2000). These three general levels were:

- **Computation/Data layer**: the basic protocols for accessing, invoking and scheduling the use of computation and data resources. This includes the fundamental networking protocols, the addressing of resources, and the low-level data formats and remote method invocation to access the raw power. However, to use this layer effectively, the user has to know in advance the location, the data formats and the functionality of resources available.
- **Information Layer**: information on resources available on the distributed system is available via *descriptions of those resources* (commonly known as *Metadata*). This will allow the discovery and negotiation of resources within known domains of practice: that is the *meaning* of the metadata is agreed within a particular community.
- **Knowledge Layer:** access to resources is negotiated through the semantics of metadata encapsulated within the system. This layer will provide *contextualised* access to information, utilising semantic knowledge and reasoning. Processes within this layer include reclassifying information against new ontologies, to enable interoperability between different semantics, and knowledge discovery in databases (including data mining).

Whilst originally promoted within the domain of distributed Grid systems supporting an infrastructure for scientific applications, this analysis of distributed systems applies equally well to the Web. Indeed, the distinction is likely to be nonsensical as the whole world converges on a single architecture for wide-area distributed systems.

Within the Web, the basic existing infrastructure (for addressing Data) augmented with the basic Web Services architecture (for accessing computational resources) forms the *data and computation* layer, with the familiar tool of the Web Server as the defining tool; one which dumbly responds to requests for resources, possibly passing control to other systems in a peer-to-peer manner.

Existing community efforts to standardise on particular XML Schemas for both data and Web Services, particularly some infrastructure efforts including those mentioned above (P3P, PICS, CC/PP, RSS) - which can be seen as augmenting the basic Web Service architecture, distinguish the *information layer of* the Web. In this layer the concept of the *Portal* is the defining tool. This a tool that uses metadata defined within a known domain to access resources which, although they are unknown to the user, have a previously agreed semantics and can be used by the user. Thus the intelligence still resides with the human user.

The Knowledge layer within the Web will be supported by the Semantic Web. We shall discuss in more detail what this entails.

Jeffery identifies *control* as a connecting feature between the layers: knowledge about the relationships between resources controls access to information about resource, which in turn controls access to the resources themselves. We could equally well say that another distinguishing feature of the layers is the *decrease* in the necessity for *prior knowledge* as we go up the layers, and the increase in *delegation,* as more functions are delegated from the user to the system, ultimately to a system of intelligent agents. Thus in the data/computation layer, users exercise direct control upon resources which they know about in advance. In the information layer, users delegate some tasks, such as resource discovery and access, to portals, and may not have knowledge of the location of resources, but they have to have prior knowledge of the portal and the nature of the information that that portal processes.

In the Knowledge layer, delegation should increase, prior knowledge decrease, so the user will be able to delegate the task of discovering appropriate information sources, and also not need to know the meaning or existence of that information in advance. The user should be able to specify

the task that he or she desires to perform and then delegate it to the system. Thus, the defining software component for the knowledge layer is the *Intelligent Agent.*

Three types of agent will typically be present within this system:

- **User Proxy agent:** An agent acting on the users behalf. It will initiates and coordinates user actions and queries to the web, seek out and offer to the user relevant resources, acts as user proxy to react when user is absent, and automatically responds to requests on the user according to user preferences and security settings.
- **Resource agent:** An agent acting on the behalf of a resource. It will respond to requests for access to resources, coordinate queries with other resources, and control and monitor access to the resource.
- **Broker agent:** These agents are not connected to any resource, but provide a discovery and negotiation service for other agents, searching for appropriate resources, negotiating access and monitoring usage.

Agents will thus negotiate with each other on a basis of attempting to determine the meaning of resources - thus the main integration mechanism of a Semantic Web enabled Web Service architecture will be meaning itself. We will have *Integration via Meaning.* Meanwhile, the portals and other tools which are at the lower levels will *disappear into the infrastructure of the Web itself;* the user will need no prior knowledge of the portals or the semantics they support; the interaction with the portal will be mediated by the user's agent which will attempt to resolve the semantics provided by the portal with the semantics of the user.

5 Particular functions of the Semantic Web

Particular functions will be provided via this Semantic Web layer on top of Web Services, as processed via the actor's agents.

- **Resource organisation, searching and discovery.** By expressing the semantics of resources in terms of Ontologies, together with interoperability and reasoning, agents can discover relevant resources, and present them back to the user in the user's own terms. This would include access to web service descriptions; currently these are expressed in terms of either interface descriptions, or domain specific criteria. In the Semantic Web, web service descriptions would include a formal expression of the functionality of the service; a proof could determine whether the service satisfies the user's requirements.
- **Brokering and negotiation.** Once suitable resources have been identifies, intermediary brokers can reconcile the requirements of users and resources, in terms of their meanings, to determine whether a suitable deal or contract can be established for usage of the resource.
- **Trust management.** Establishing trust between agents that have no prior knowledge of each other is a major problem within the Web, and one which could potentially prevent the establishment of a universal Web Service infrastructure. Agents need to be able to negotiate access to resources, through a process of negotiation. Again Semantic Web techniques can enable this. Services could publish policies, user agent could present their credentials, possibly with reference to a trusted third party, and once trust has been established, negotiate suitable access rights and obligations (Dimitrakos 2002).
- **Quality of service.** Similarly, users and resources may have conditions with respect to the quality of service they require (e.g. in terms of response time, accuracy of data, level of confidentiality). These properties of resources can be expressed in Semantic Web terms and negotiated via agents.
- **Auditing and monitoring.** Monitoring agents can track the usage of the web, and provide audit trails. Included in this mechanism would be functionality to track expenditure and perform billing.

- **Personalisation.** User agent will be able to represent and enforce the preferences of the user. This would include how the user would prefer their own information to be used, and also what requirements they might have on the information they would like to access and how it is presented back to them. This infrastructure is already emerging with CC/PP, P3P, and PICS. Through negotiation with RSS, and other specifications such as XHTML modules, a negotiation of the form of information that a user requires can be performed.

6 Research programme at BITD CLRC

The Business and Information Technology Department of CLRC has a research and development programme focused on realising the various components of the emerging Web combining Web Services with the Semantic Web. A collection of new projects has recently started which are developing fundamental components of this architecture, and also applying the architecture in practice. We consider some of these projects.

- **Data and Information Portals: Universal access to information.** These two projects, one within the e-Science programme (Ashby et. al. 2001a, 2001b) developing a portal for accessing scientific data across a wide variety of domains, and the Information portal, providing single point of access to business information provide fundamental components of the Information Web which can be used to build the semantic layers above.
- **SWAD-Europe: bringing the Semantic Web to reality.** The SWAD-Europe project intends to deepen the experience with the current generation of Semantic Web technologies, providing tools, demonstrators and applications to show the practicality of using the semantic web to deliver added value now. Of particular interest is the application to trust management, developing mechanisms to negotiate access to resources through presenting credentials to policies and using a process of negotiation; and the use of controlled vocabularies as restricted ontologies to provide focussed access to information (SWAD-Europe; Brickley et. al. 2002). Thus these applications will provide evidence that the Semantic Web can augment a Web Service architecture.
- **GRASP: Bringing the Grid to business.** The GRASP project aims to explore a new advanced system infrastructure for Application Service Provision based on GRID technologies. An ASP, realized using Grid technologies, will be characterized by a high level of scalability and reliability, innovative solutions for the security, for the accounting, the Quality of Service (QoS) and for resource management (GRASP). Thus GRASP will develop semantic techniques to augment global computing infrastructure (as realised in the Grid) and demonstrate its application within business applications.
- **PELLUCID: using Agents to coordinate knowledge.** The objective of the Pellucid project is to develop an adaptable platform for assisting organisationally mobile employees by providing a suitable knowledge of the task in hand to the employee. It is taking the approach of *agent supported knowledge management.* Thus intelligent agents, both user proxies and task agents, acting in consort with a workflow system will use domain knowledge to reason what information is relevant to a particular employee at any instant. (PELLUCID). Thus within this project, we are developing intelligent agent systems negotiating with user profiles to provide domain specific information through reasoning on the current context of that user - important features of the merger of the Semantic Web and Web Services.
- **I-Trust: establishing trust between agents.** This working group is looking at trust management at a high level to establishing security and confidence in the global computing infrastructure (I-TRUST)
- **E-Lege: introducing the Grid into e-Learning.** This working group is considering how to introduce the global computing infrastructure into the field of e-Learning.

7 Conclusions

With the advent of Web Services, and the long development of the Semantic Web coming to maturity, it is inevitable that they should be considered as complementary. Indeed, others have discussed similar ideas (e.g. De Roure 2001; Houstis et. al. 2002; Ryssevik 1999). In this paper, however, I have set the combination within the context of a structure of the three layered architecture and a notion of *decreasing prior knowledge,* and *increasing delegation.* Together they offer the opportunity for a new global computing infrastructure, and I consider some specific areas where significant gains could be found through this combination.

Ultimately, will the combination of Semantic Web and Web Services take off and become commonplace in the development of information systems? This is hard to say - there have been many attempts at providing „universal computing solutions", which have had only limited success. Perhaps we can say that this one has a better chance than many because of the momentum behind it, and the existing much higher level of basic infrastructure (in networks, protocols and software platforms) than there has ever been before.

However, there is still a need for a „killer app" to get this emerging architecture onto everybodies computer and mobile phone, in the way the current web is now. To determine what that might look like, we should consider what has made the WWW so successful - it was not the most advanced information system available at the time. However, the key features were all things which made it accessible to people: it was easy to set up, both as user and, crucially, as content provider; it provided an immediate feedback so the user could how they gained by using it (especially when made more visually appealing through the inclusion of images in Mosaic); and it is was very tolerant to all to human errors, not collapsing when faced with poorly produced pages, or pages which were not there at all! With the Web Service infrastructure offering the potential for ever more increasingly complex applications, and more and more of the functions of the system becoming opaque to the user, the intelligent use of the Semantic Web offers the opportunity to recover accessibility for humans.

8 References

J V Ashby, J C Bicarregui, DR S Boyd, K Kleese van Dam, S C Lambert, B M Matthews, K D O'Neill. (2001a): The CLRC Data Portal *British National Conference on Databases*

J V Ashby, J C Bicarregui, D R S Boyd, K Kleese van Dam, S C Lambert, B M Matthews, K D O'Neill (2001b): A Multidisciplinary Scientific Data Portal *HPCN 2001: International Conference on High Performance and Networking Europe* Amsterdam

G. Beged-Dov et. al. (2000): RDF Site Summary (RSS) 1.0 http://purl.org/rss/1.0/spec

T. Berners-Lee (1989): Information Management: A Proposal, *CERN* http://www.w3.org/History/1989/proposal.html

T. Berners-Lee (1998): Semantic Web Road Map http://www.w3.org/DesignIssues/Semantic.html

D. Brickley and R.V. Guha.(2002): PICS Ratings Vocabularies in XML/RDF *W3C Note* http://www.w3.org/TR/rdf-pics

D. Brickley and R.Swick.(2000): RDF Vocabulary Description Language 1.0: RDF Schema *W3C Working Draft* http://www.w3.org/TR/rdf-schema/

D. Brickley, S Buswell, B Matthews, L Miller, D Reynolds, M Wilson (2002):SWAD-Europe: Semantic Web Advanced Development in Europe *Presented at the International Semantic Web Conference.*

D. De Roure, N Jennings, N Shadbolt (2001): Research Agenda for the Semantic Grid: A Furture e-Science Infrastructure http://www.semanticgrid.org/html/semgrid.html

CERIF: the Common European Research Information Format www.cordis.lu/cerif/

CLRC Data Portal Project www.escience.clrc.ac.uk/Activity/ACTIVITY=DataPortal

D. Connolly, F. van Harmelen, I. Horrocks, Deborah L. McGuinness P. F. Patel-Schneider, L. A Stein (2001): DAML+OIL Reference Description W3C Note 18 December 2001

T. Dimitrakos, I. Djordjevic, B. Matthews, J. Bicarregui, C. Phillips (2002): Policy-Driven Access Control over a Distributed Firewall Architecture Policy 2002: IEEE 3rd International Workshop on Policies for Distributed Systems and Networks IEEE-CS, California, U.S.A.,

Dublin Core Metadata Initiative
www.dublincore.org/

EbXML: (Electronic Business using eXtensible Markup Language) *Home page*
http://www.ebxml.org

E-LEGE: E-Learning within a Grid environment
http://www.bitd.clrc.ac.uk/Activity/lege-wg/

eGIF (2002) e-Government Interoperability Framework Part Two: Technical Policies and Specifications *Version 4, Office of the e-Envoy,*
http://www.govtalk.gov.uk/documents/e-GIF4Pt2_2002-04-25.doc

D. Florescu, A. Grünhagen, D. Kossmann (2002): XL: An XML Programming Language for Web Service Specification and Composition *WWW2002,*
http://www2002.org/CDROM/refereed/481/index.html

I. Foster , C. Kesselman (eds.) (1999): The GRID: Blueprint for a New Computing Infrastructure, *Morgan-Kaufmann.*

A. Gokhale, B Kumar, A Sahuguet (2002): Reinventing the Wheel? CORBA vs. Web Services. *WWW2002, Web Services and Metadata track,*
http://www2002.org/CDROM/alternate/395/index.html

GRASP: Grid architecture for Application Service Provision
http://www.bitd.clrc.ac.uk/Activity/GRASP

C. Houstis, S. Lalis, V.Christophides, D. Plexousakis, M. Vavalis, M. Pitikakis, K. Kritikos, A. Smardas, X. Gikas, (2002): A Service infrastructure for e-Science: the case of the ARION system, *submitted to the E-Services and the Semantic Web workshop (ESSW2002).*
http://dlforum.external.forth.gr:8080/papers/ARION-paper_v52.pdf

IBM Web Services Home Page. http://www-3.ibm.com/software/solutions/webservices/

R. Irani (2001): ebXML and Web Services: The Way to Do Business
http://www.webservicesarchitect.com/content/articles/irani03.asp

I-TRUST: Trust Management in Dynamic Open Systems http://www.bitd.clrc.ac.uk/Activity/itrust/

K G Jeffery (2000): Knowledge, Information and Data. *A briefing for OST*
http://www.bitd.clrc.ac.uk/Publications/1272/KnowledgeInformationData20000124.htm

G. Karvounarakis, S Alexaki, V Christophides, D Plexousaki M Scholl (2002) RQL: A Declarative Query Language for RDF. *WWW2002*
http://www2002.org/CDROM/refereed/329/index.html

G. Klyne, F. Reynolds, C. Woodrow, H. Ohto,(2001): Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies *W3C Working Draft 15 March 2001*
http://www.w3.org/TR/2001/WD-CCPP-struct-vocab-20010315/

M-R Koivunen, E. Miller (2001): W3C Semantic Web Activity, Proceedings of the Semantic Web Kick-off Seminar in Finland,
http://www.w3.org/2001/12/semweb-fin/w3csw

O. Lassila R. Swick (1999): Resource Description Framework (RDF) Model and Syntax Specification *W3C Recommendation 22 February 1999*
http://www.w3.org/TR/1999/REC-rdf-syntax-19990222

B. McBride, R. Wenning, L. Cranor, (2002): An RDF Schema for P3P *W3C Note 25 January 2002*
http://www.w3.org/TR/2002/NOTE-p3p-rdfschema-20020125

E. Miller, P Miller, D Brickley (1999): Guidance on expressing the Dublin Core within the Resource Description Framework (RDF) *Dublin Core Metadata Initiative Draft Proposal*
http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/

.NET Microsoft .NET Homepage
http://www.microsoft.com/net/

PELLUCID: An Agent Based Platform for Organisationally Mobile Public Employees
http://www.bitd.clrc.ac.uk/Activity/PELLUCID
http://www2002.org/CDROM/refereed/329/ - note1

J. Ryssevik, S. Musgrave (1999): The Social Science Dream Machine: Resource discovery, analysis and delivery on the Web, *the IASSIST Conference, Toronto,*
www.nesstar.org/papers/iassist_0599.html
SOAP Version 1.2 Part 1: Messaging Framework *W3C Working Draft 17 December 2001*
http://www.w3.org/TR/2001/WD-soap12-part1/
SOAP Version 1.2 Part 2: Adjuncts *W3C Working Draft 17 December 2001*
http://www.w3.org/TR/2001/WD-soap12-part2
Sun Microsystems (2001): Web Services Made Easier: the Java APIs and Architectures for XML *Sun Microsystems white paper*
http://java.sun.com/xml/webservices.pdf
SWAD-Europe: Semantic Web Advanced Development for Europe
http://www.bitd.clrc.ac.uk/Activity/SWAD-Europe
UDDI: Universal Discovery Description and Integration, *Home Page*
http://www.uddi.org
W3C: World-Wide Web Consortium
www.w3.org
WSDL: (2001) Web Services Description Language (WSDL) 1.1, *W3C Note*
http://www.w3.org/TR/wsdl
X Protocol: XML Protocol Working Group Home Page
http://www.w3.org/2000/xp/Group/

9 Acknowledgements and Contact Information

I would like to thank my colleagues in CLRC Business and Information Technology Department for their help and advice over many discussions on this topic.

Brian Matthews
CLRC
Rutherford Appleton Laboratory
Didcot
OX11 0QX
UK

e-mail: b.m.matthews@rl.ac.uk

Weaving the Web of European social science

Jostein Ryssevik (Keynote Speaker)
Director of Technology and Development, Nesstar Ltd.

In the late 1950s Dr. J.C.R Licklider observed that most of his time as a researcher was spent on getting into a position to think, and not on creative thinking as such. "Much more time went into finding or obtaining information than into digesting it." (see Howard Reingold: "Tools for Thought – The History and Future of Mind-Expanding Technology", The MIT Press, Cambridge Massachusetts 2000, p133). A few years later Licklieder became the director of ARPA, the research organization that initiated the forerunner to todays Internet, the ARPAnet.

Licklieder's observation might be seen as a general justification for the development of any research infrastructure, including the Internet. The overriding goal of a research infrastructure is to facilitate the maximization of the time spent on digesting and thinking over the time spent on finding and accessing.

However, even today nearly 50 years after Licklider's observation and about 10 years after the invention of the World Wide Web, comparative social science research in Europe is hampered by the fragmentation of the scientific information space. Data, information and knowledge are scattered in space and divided by language and institutional barriers. As a consequence too much of the research are based on data from a single nation, carried out by a single-nation team of researcher and communicated to a single-nation audience. The state of affairs is preventing the development of a comparative and cumulative research process integrating and nurturing the entire European Research Area.

Yesterday's answers to these challenges would probably have been formulated in terms of centralization and establishment of large-scale European-wide institutions. Today's answers should rather focus on the power of emerging information technologies to encourage communication, sharing and collaboration across spatially dispersed but scientifically related communities.

What we would like to see is the research infrastructures that could facilitate the following scenarios:

Scenario 1

A user is looking for data on political trust from three different regions of Europe. He uses the geographical interface to circle in the relevant regions on the map and enters the additional search criteria. From the returned hit list he is able to browse layers of increasingly detailed metadata describing potential sources. He is even allowed to perform simple statistical analysis and visualizations on-line to make sure that the data fulfills his requirements. Several datasets can be brought to the desktop at the same time to ease the comparison. As soon as a decision has been made, the chosen datasets can be downloaded and automatically converted to the format of his favorite statistical package. All relevant metadata travels along with the data to assist the researchers in his analysis.

Scenario 2

A researcher analyzing a group of variables in dataset X would like to know if there are similar datasets from other countries that could be used for a comparative study. By hitting the "get com-

parable dataset" button, a list of potential datasets is immediately returned. By fine-tuning the search criteria to make sure that all datasets fulfills her methodological requirements she is able to circle in on a handful of sources that might be used.

The researchers would also like to have an overview of knowledge products (papers, articles etc.) that are based on these studies and even to browse these objects if they are available on-line. As references and links to derived knowledge products are an integrated part of the metadata of each single dataset a sorted list can be displayed by a single keystroke. Some of these references are also including e-mail and website-addresses to the relevant researchers.

Finding a problem with one of the variables, the researcher writes a note and appends it to the "user experience-section" of the metadata to alert future users about the quality of the data (she also leaves her e-mail address to allow them to contact her). And when the research paper is ready and published in an on-line journal, links to the dataset is added to allow future users to re-visit her analysis

Scenario 3

A researcher who is reading an article in an on-line journal finds a link that connects him to the data that was used by the author to underpin the arguments. By following the link, he is able to load the dataset and to rerun the original analysis. He is even allowed to dig deeper into the same data-source, testing alternative indicators or models. The system is also making him aware of several other comparable data sources published after the article was written and he uses these sources to challenge the conclusion of the author. Links to knowledge products based on these newer data sources is also available. From one of the sources he is even brought to a mail-list that discusses the phenomena in further detail

Scenario 4

A user is looking at a table showing variation in nationalistic attitudes among different educational groups in country X. Through the systems integrated multilingual thesaurus service he is able to pick up the relevant key-words describing this table and to automatically create a multilingual query for datasets that might be used to create comparable tables. He is also leaving the query with the system's "digital research assistant" (an active agent), to make sure that he is alerted by e-mail if a new dataset meeting his requirements is published somewhere in Europe at a later stage. He even ask the agent to look for knowledge products addressing the same topics.

The paper will outline the visions, requirements and architecture of a virtual research infrastructure that would allow these dreams to become true. It will also describe a real life implementation of these visions, the NESSTAR (Networked European Social Science Tools and Resources) system. All of the functionality described in these scenarios are technically feasible using the Nesstar software platform.

Nesstar has been developed according to the following "dream-list":

- all existing empirical data available on-line
- an integrated resource discovery gateway in order to identify and locate relevant resources (including the relevant tools to overcome the language barriers)
- extensive amounts of metadata available (multi-media and integrated with the data)
- the ability to carry out simple browsing, visualisation and analysis of the data on-line
- the ability to subset and download the data and metadata to a favourite analysis too
- „active research agents" mining the net and informing user when new resources are available
- efficient hyperlinks from the metadata to every relevant report and publication
- current e-mail/web addresses to all relevant researchers, support staff, departments etc.
- an efficient feedback system to the body of metadata allowing the user to add to the collective memory of a data source

This technology is currently being used to develop web-based data services by a variety of data archives, libraries and providers world wide. In Europe, the Council of European Social Science Data Archives (CESSDA) has decided to use the platform to develop a portal (integrated catalogue) to the data resources of all the European data archives. An EC-funded project working toward this goal (MADIERA) will be started up this autumn.

Nesstar might be seen as an example of a new type of application providing the building blocks of what has lately become known as the Semantic Web. These are metadata rich applications where all relevant information is given well-defined meaning, making it easier for computers and people to work in cooperation. Agreed metedata standards are the glue of the Semantic Web, and the most important message form this paper is the need to develop flexible and extensible metadata standards, which will facilitate smooth interoperation across systems and domains.

Contact Information

Jostein Ryssevik
Director of Technology and Development
NESSTAR Ltd.
Norwegian Social Science Data Services (NSD)
Hans Holmboesgt. 22
N-5007 Bergen
Norway

Tel: +47 5558 2654
Fax: +47 5558 9650
e-mail: Jostein.Ryssevik@nsd.uib.no
http://www.nesstar.org

Is there any user for this CRIS?

Benedetto Lepori, Lorenzo Cantoni
Facoltà di Scienze della comunicazione, Università della Svizzera italiana, Lugano

Summary

In this paper, we will analyse the issue of how research information services build their relationship with their public and, in particular, if (and how) they try to target specific user groups, rather than publishing information for everybody. An analysis of some Swiss experiences brings us to define two basic models, which we call corporate-oriented and market-oriented CRIS. The first category includes services where publishing information on the Internet is mainly guided by the visibility needs and the corporate image of the organisation, e.g. to present their own research or funding opportunities. On the other side, the second category includes a small number of services which are clearly targeted towards specific user groups and where contents and information layout are specifically designed to match the needs of these users.

We also argue that it is very difficult and conflict-laden to try to reconcile these two models in a single information service, since the organisational and economic logic behind are clearly different. We then conclude that the issue of targeting users could be at best addressed by a two-layer structure of research information services, one layer composed by (mostly public) organisations which produce information according to their corporate orientations and for a "generic" public and a second layer which exploits, rewrites and regroups this information according to the needs of very specific user groups.

1 Introduction

The aim of this paper is to analyse the implications of some recent developments in Internet communication for the development of Research Information Services. In fact, recent research shows that Internet communication is rapidly evolving from a model where information is published for a generic public (for everybody, i.e. for nobody in particular) and where the amount of information and its quality are considered the most important success factors (i.e., in the realisation of Web sites) to a much more structured model where publishers increasingly target the needs of specific publics and their success is measured by the capability of building stable relationships with a group of customers (which, possibly, are ready to pay for the delivered services).

This is of course a consequence of the growing amount of information available on the Internet (i.e., increasing competition for visibility among information providers), but also of greater cleverness among users, who are increasingly able to distinguish between different information sources and their quality. Since in many areas Internet has become a professional tool to support work, users measure the quality of a given web site in terms of the relevance of the information for their specific activity (its pragmatic function; see section 2). The major implication is that to be a successful provider on Internet is not any more sufficient to publish large amount of information, even of very good quality, but one needs to draw from this information very specific services tailored to the needs of specific user groups; moreover, these services have to be better than those which can be obtained through other communication tools.

Our thesis is that many of the problems which were found in the exploitation of data contained in CRIS are due to many CRIS overlooking this structure of Internet communication. In fact, a

look to available literature on CRIS[1] shows that most of the efforts in this area were (and still are) devoted to improve the quality of the collected information (e.g, through standards like CERIF) and to find good semantic descriptors (e.g., thesauri or ontology) and efficient ways to generate and implement them on large sets of data (e.g., through metadata standards or XML). It then seems that the ideology behind the development of many CRIS, especially at the national level, was that a very large and good-quality database of information on research activity, once collected, would automatically find its users, either at the level of the science policies (to monitor use of funds and to avoid duplications) and of the researchers themselves (to find out other projects in the same area).

It is not of course our intention to underestimate the importance of building good-quality data sets with a wide coverage of research activities and the complexity of the (technical and semantic) problems of retrieving and managing these data. But we wish to stress that these data are going to be useful only if they are exploited to deliver services to the main target groups for CRIS and that this activity entails a good deal of expertise and knowledge of the users' needs. For instance: the mapping of the quality of research activities in Europe, being one of the major support instruments for the European Research Area, relies on soft methods like questionnaires and panels of experts; one may wonder if all existing databases on European research are of any use for this exercise, since they are not at all mentioned in the methodological document of the Commission (European Commission 2001). Thus, as we will discuss in our paper, user-centeredness is a crucial issue for CRIS also from the point of view of their political legitimacy and thus of the support they can gain from decision-makers[2].

The rest of the paper will be divided into three sections.

(1) First, we briefly discuss the concepts of user-centeredness and user-targeting, introducing some conceptual tools for the analysis of the Internet communication.
(2) Secondly, we analyse how existing Research Information Systems in Switzerland address this issue and which structural elements determine their ability to respond to the user's needs.
(3) Finally, we draw some conclusions on different possible architectures of CRIS and we show that a multi-layer model seems to be much more efficient in responding to users' needs.

2 User targeting and Internet communication

2.1 Syntax, semantic and pragmatics

According to a famous distinction done by Morris (1957), human languages can be studied from three main points of view: *syntactic*, *semantic* and *pragmatic*. While syntax takes into account the message in itself, semantics looks at the relation between text and the world, and pragmatic at the relation between language and its users.

Over the Internet many studies have focussed on the syntactic/semantic aspects: electronic text is completely available for automatic analysis, allows for new (hypertextual) organization and (hypermedia) representations of the world (Cantoni & Paolini 2000). Also in the area of document retrieving the syntactic/semantic point of view offers many new and challenging research

1 See for instance the reference documents available on the Eurocris Web site www.eurocris.org and the presentation at the European Conferences on Research Information Services CRIS '98 in Luxembourg (http://www.cordis.lu/cris98/) and CRIS2000 in Helsinki (http://www.cordis.lu/cris2000/); an area of major concern of both conferences was how to exploit the information contained into existing CRIS. All on-line references have been checked April 30, 2002.

2 The decision of the Commission to abandon the ERGO project (http://www.cordis.lu/ergo/), aiming to build a comprehensive database of research projects in Europe, shows that failing to prove a real usefulness for policy-making can result in lack of support to such services.

opportunities: let's think, for instance, of xml for data interchange or metadata in semantic mapping.

Recently, also the role of the user, hence the pragmatic facet, is attracting the attention of people who study the net (and operate in it). An example taken from the field of Internet search engines can help to explain this sort of paradigm shift. While at first search engines relied, in order to answer user queries, on ranking algorithms based mainly on the analysis of the html pages themselves, more recently, they are moving to take into consideration more and more their users. This is done in different ways: considering people who produce the websites (the senders) – almost all search engines allow for or require the payment of a fee by a website either for being analysed, or spidered or to get a better ranking –, or taking into consideration people who make queries, integrating their usages of responses into the ranking algorithm itself. Moreover, both Internet Explorer and Netscape have integrated in their interface a service offered by Alexa, which provides – when surfing a given website – information about related websites. These data are not calculated on the basis of computational linguistic algorithms; instead, they are taken from the database of actual navigation paths of Alexa subscribers (something like the "Customers who bought this book also bought…" service offered by amazon.com).

Google takes into consideration the actual communication paths over the Internet considering links as being sort of votes, so that the more links there are toward a website, the more important it is considered, something similar to the citation impact factor used in the analysis of scientific journals (Brin & Page 1998).

The consideration of people involved in communication acts pushes Internet actors towards a careful analysis of the *relevance* (Sperber & Wilson 1995) of information for given, selected publics – *pragmatic focus* – and also of the quality of what is published – *semantic focus*.

2.2 Towards a social shared hierarchy of Internet sources

The public perception of what is published over the net is parallelling in some way what already happened in the history of the radio. When the radio was born, there was one tool, the radio apparatus, which was the single gate towards very different information sources – as now the computer connected to the Internet is an open window on an infinite number of websites – and it seemed impossible to assess information quality: everything seemed to be condemned to remain in a flat world without any difference. But in some years, due also to technological evolutions and governmental regulations, people became able to reconstruct and internalise a new hierarchy of information sources, including radio (Gackenbach & Ellerman 1998).

So users are constructing, according to their needs and experiences, as well as relying on others' accounts, a hierarchy of Internet sources that suit their information needs.

2.3 Two opposing forces

In human languages, two opposing forces are at work when communicating (Uspenskij 1996). The first one is toward simplification and standardisation, and pushes to avoid any un-necessary redundancy; this force suits mainly the need of the sender: she knows exactly what she wants to say and tries to save resources in term of time (oral communication) and space (written communication). On the contrary, the second opposing force pushes toward redundancy: it is in defence of the receiver, who does not know in advance which meaning the sender wants to convey; the receiver needs hence more time and more clues to better get and interpret the message.

Human languages are shaped by these conflicting forces; let's take a very simple example: the –s ending in the third person of English verbal conjugation. Of course, it is not necessary in order to convey the meaning of a sentence (anyway, not more than other possible endings for all the other persons, as it happens in different languages), if I say **she study English* the meaning would be quite clear. Anyway, and in particular for oral language, every redundant element helps

to better reconstruct the sentence (to test the meaning hypothesis the listener does when understanding), and gives more time for the interpretation process. All grammars are shaped by these opposing forces.

The same can be said about the Internet communication we are studying here. On one side, communication senders try to reduce and standardize their outputs – it is one of the main efforts of database designers – while on the other side receivers need more pieces of information to get the proper meaning. For example: if the European Union stores in its database the information about a new call, it could write something like:

8th IST Call published on 16.11.2001

An information broker could translate this text – in order to meet her clients' information needs, thus to make the same information relevant to them – as follows:

The European Commission launched on the 16th November 2001 a call for new research projects in the field of communication and information technologies.

3 A case study for Switzerland

In this section we shortly review some existing research information services in Switzerland and examine how they address the needs of specific user groups.

The first important feature of Swiss CRIS is that there is no central repository for information on research activities, like the national information services present in other European countries.[3]. Research information is available on the following web sites:

- the Federal Office of Education and Science (FOES; www.admin.ch/bbw): this site contains, besides some general information on research funding, all the data on the participation of researchers to European projects;
- the Swiss National Science Foundation (SNF; http://www.snf.ch) holds a database of all the projects it funds (about 1000 new projects every year);
- the web sites of the 10 cantonal universities and of the two federal institutes of technology (http://www.switch.ch/edu/educ_orgs.html); most of them have on-line research reports with description of individual research projects.

Other information is to be found on the web-sites of the federal offices financing research in specific areas (like energy: http://www.energieforschung.ch or environment: http://www.umwelt-schweiz.ch/buwal/de/fachgebiete/forsch/index.html).

To improve this situation, the Swiss government launched in 1996 a project for a central database of all research activities funded by the state; the major objective of the ARAMIS project (http://www.bbw.admin.ch/f/forschnat/aramis/aramis.html) was to give a complete overview of the publicly funded research in each sector to allow for better coordination; a secondary objective was to produce better statistical information of public research funding[4]. Until now, ARAMIS has succeeded in collecting many of the data on research financed by the federal administration, but not on research financed by the SNF or by the Swiss federal institutes of technology; the cantonal universities are also not obliged to provide their data to ARAMIS.

Looking at these services from the point of view of a user interested to find more information on research activities one could get disappointing results; let's give some examples:

3 In the Eurocris map (http://www.ub.uib.no/avdeling/fdok/cris/EuroCRIS/map.htm) there is simply a link to a web page where all Swiss research institutions are presented and a link to their websites is provided.

4 Existing data are produced by the Federal Office for statistics only every second year through a questionnaire distributed to all public institutions financing research activities (http://www.statistik.admin.ch/).

- SNF: the project database gives some very basic possibilities of searching for different project (according to the project leader, the institution, the scientific area); it offers also a full text search, bu the website itself discourages the uset because it is quite slow; the description of each project contains very little information, basically of administrative type;
- FOES: information on European projects with Swiss participation is to be found (with some difficulties) under "publications", where an Internet version of a CD_ROM published each year by the FOES is available; search tools are quite limited (only full-text search and by programme); information on the projects is more complete than in the SNF case (including a project abstract), but unfortunately no contact address for each project is given.

The conclusion is that these two services are basically Internet versions of the annual report of the institutions and that no effort has been made to adapt and to organise this information to the need of (possible) users; they mostly respond to the (legal) obligation to document which projects were funded or to the wish to publish the funding activity of the institution[5].

The situation is partially different with ARAMIS; the Web search interface (http://www.aramis-research.ch/) is not very sophisticated – being impossible to search in specific fields, like project responsible, or through disciplines -, but the available information on the projects is quite complete and also classified according to the CERIF standards. The major flaw of the system is of course its incomplete coverage, since the major funding organisations of research activities in Switzerland are not yet covered; in fact, ARAMIS, is by now essentially an information service on research financed directly by the public administration. It seems also that most of the effort has been devoted until now to develop tools to administer research projects, rather than to deliver information to the general public (this probably explains also the quite limited search interface on the Web).

The basic move behind ARAMIS seems then to be that of better managing publicly funding research, a rationale that may partly explain the resistances of institutions like the SNF to deliver their data.

User-centred information systems

However, in our review we could find a small number of services which are built according to a quite different purpose; these include:

- The Swiss portal on science Swiss-science (www.swiss-science.org). Swiss-science has been developed by a private company (Science Com AG), which is specialised in the publication of information products on science, its main product being the monthly magazine Vision. The web site offers a news service on science, as well as a series of dossiers on specific subjects; it hosts also other specialised services, like the web sites of different research programmes. Information is gathered from many official sites like CORDIS, SNF, FOES, but also from (specialised or general) newspapers, and then edited by a team of scientific journalists. Swiss-science is financed by sponsors (e.g. the ticker of the Union Bank of Switzerland on the homepage) and by selling specialised services (e.g., the publication of a dossier on technology transfer in Switzerland financed by the Federal Office for Professional Training and Technology).
- The Swiss Information Network on European Programmes Euresearch (www.euresearch.ch). Euresearch has the mandate from the Swiss State to promote the Swiss participation to European programmes and to inform the researchers on the opportunities for participation. The web site offers information on European programmes, including announcements of events, news, frequently asked questions; a new information platform is planned for 2003: among other features, it will offer the possibility to subscribe to a push service, delivering information

5 For example, the SNF web site has a section where successful projects are presented in a journalistic style.

according to specific user-profiles. Information is mostly drawn from CORDIS, but edited and tailored to the specific interests of Swiss researchers.
- The research information service for the Italian-speaking researchers in Switzerland (www.ticinoricerca.ch). This service, managed by the university of Lugano, offers to researchers a quick access in Italian language to research funding opportunities in Switzerland and in Europe; available services include news, events, calls for papers and for research programs, as well as a push service that delivers weekly updates to subscribers (http://www.ticinoricerca.ch/swisscast/). Information is mostly gathered from different web sites, including CORDIS, SNF, FOES through an automatic gatherer module and then edited by the service responsible (see Lepori 2000 fur full details).

All these services target explicitly a specific user group and, thus, they define accordingly their information content, that is:
- Researchers and other people interested on a "general" view of the Swiss research system for Swiss-science;
- Swiss researchers interested in participating to European programmes for Euresearch;
- Italian-speaking researchers wishing to participate to Swiss or international research programmes and who need first information in their mother language for ticinoricerca.

As we will discuss more in depth in the next section, their organisation and business model is also quite different from the more generic services discussed previously.

4 Structural consequences for CRIS

We may interpret the observed patterns for Swiss CRIS in the light of the model presented in section two. Thus, the development of Swiss CRIS can be interpreted to be subject to two opposing forces:
- The push from the information providers, i.e. institutions active in science policy, which have a need to diffuse information on research activities either as a legal obligation or simply to promote their activity and their image.
- The pull from (potential) information users, who ask for information and services which are tailored to their needs.

The first movement promotes the development of CRIS which are strictly bound to strategies and the needs of their father institution; this means also that, for the sake of simplicity, they tend to contain generic information (without reference to a specific user groups) and in a very standardized form (e.g., records in databases). Limitation of resources means also that the information content is defined more from its availability (e.g., having this data in the corporate database) than from the assessment of user needs. The information services of the Swiss National Science Foundation and of the Federal Office for Education and Science correspond to this model; also the public part of the ARAMIS service show the same pattern[6]. These services don't have to care very much for their actual impact (e.g., if potential users are really exploiting the data); they simply fulfil the mandate to publish information on how public money is spent. We call this model *corporate-oriented* CRIS to emphasize that these services are mostly tools at the service of the objectives of their father institutions, like to promote their image and functions in the research policy or activities like research funding for researchers.

On the contrary, the second movement gives rise to services which are explicitly tailored to the needs of specific user groups, like researchers interested in European programmes for www.

6 For instance, the actual coverage of research activities in ARAMIS depends very much on its institutional structure (being a project of the Federal Office for Education and Science) rather than on considerations on the relevance of the data for the public.

euresearch.ch or people interested in general information about the Swiss research system for www.swiss-science.org. This means that the structure and the content of each information item is defined according to its users' needs (e.g. to participate to European programmes); using the categories presented in section 2, we may say that the pragmatic function of the information determines its semantic and syntactic structure. This implies also that the contents and the structure of the same information could vary widely, because it is focussed on a different public[7]; the ideal of a central database on research information fails thus to take into account the heterogeneity of its potential publics.

We call this model *user-oriented* CRIS, to stress that their function is to deliver information which is useful for specific user groups and that their business model is built according to this function; this means that these services, in order to receive funds, have to demonstrate their ability to improve the information status of their users and that this improvement has practical consequences (e.g., increasing the Swiss participations to the European programmes for Euresearch). For Science Com AG, being the proprietary of the portal www.swiss-science.org, the selling of information services is in fact the major business and source of revenue.

However, these two types of CRIS are strictly linked, so that we can describe them as two different layers of a chain bridging the gap between the information providers (whishing to promote their activities and their corporate image) and the information users (seeking for relevant and targeted information). Corporate oriented-services tend to be linked to a specific provider and to define the scope of the information and its format in terms of its needs. This is the case of CORDIS for the European Union or of the ARAMIS service for the Federal Office for Education and Science. On the contrary, user-oriented CRIS are defined in terms of a specific user group. To this aim, these services gather the available information from the corporate-oriented CRIS, select it according to the interests of their users and, finally, rewrite the information in order to suit their needs and to make it understandable (see figure 1).

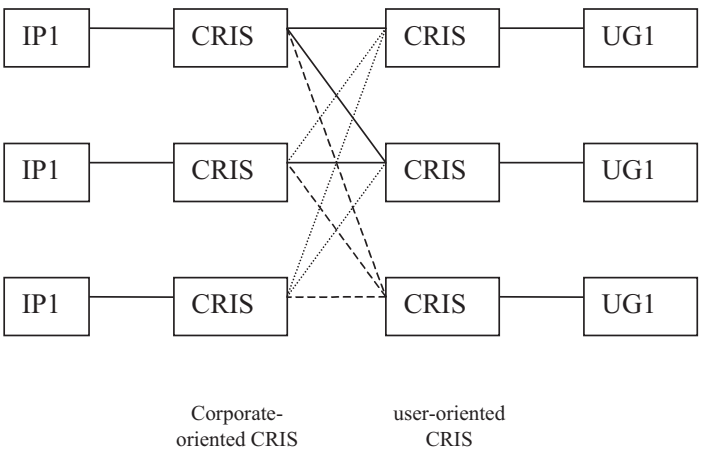


Figure 1. Two different models of CRIS

7 Let's give an example. The information on a call for proposal for European programmes delivered to Swiss researchers must indicate if the call is accessible or not to Swiss researchers (give that Switzerland is not part of the European Union). This information is essential in the Swiss context, but of course secondary in the European context; thus, it will not be available on CORDIS, but it must be added on www.euresearch.ch

5 Concluding remarks

Thus, the major conclusion of this paper is that user-centeredness is not an issue that can be addressed at the level of a single CRIS only, since there are structural factors which may limit the capability to respond to users' needs. The coexistence of services which are different not only in respect to the information content, but also to their mission and institutional binding seems to be a very promising avenue; these services will compete to get information users, but also to get recognition and financing from different information providers.

In our opinion, this institutional diversity opens also very interesting avenues for the future development of CRIS. In fact, as we have documented in this paper, the syntax and the semantics of the information contained in a CRIS are directly linked to its functions (both for the users and for the information providers) and to its organisational structure; this means also that suitable technical solutions to handle information cannot be designed in a generic way, but will be very sensitive to the specific organisation of each CRIS and to this function as a communication and working tool. In other words, the pragmatic function of a CRIS precedes its information contents or technical specifications (see Lepori 2000 and Lepori et. al 2001 for more details and for an example). This means that the design process of CRIS needs the integration of a wider range of competences than in the past, ranging from communication sciences to organisational sciences and to informatics.

An important implication is that CRIS will probably resist each attempt to standardize contents and information formats, because this would destroy their link to specific information providers and/or users and thus destroy their *raison d'être*. If we agree that the future of CRIS will be a system of services linked through Internet (Adamczak 1998), it then becomes clear that the real challenge will be to find suitable communication tools which allow information to circulate between different CRIS, being translated to each specific communication context. A task that in our opinion can be realized only through a careful integration of human competence with technical tools (see Lepori et al. 2001).

6 References

Adamczak W. (1998): The future of CRIS: a „LINK" system, *Conference on European Research Information Systems CRIS98*, Luxembourg, 12th-14th March 1998. Available on-line at: http://www.cordis.lu/cybercafe/src/adamczak.htm).

Brin, S.; Page, L. (1998): The Anatomy of a Large-Scale Hypertextual Web Search Engine, *7th International World Wide Web Conference*, Brisbane, Australia, 14-18 April 1998, available online: http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm).

Cantoni, L.;Paolini, P. (2001): Hypermedia Analysis: Some Insights from Semiotics and Ancient Rhetoric, *Studies in Communication Sciences*, vol. 1, N. 1, p. 33-53.

Commission of the European Communities (2001): *How to map excellence in research and technological development in Europe*, Brussels, march 2001 (available on-line at http://www.cordis.lu/rtd2002/indicators/projects_era.htm).

Gackenbach, J.; Ellerman E. (1998): Introduction to Psychological Aspects of Internet Use, in Gackenbach J.: *Psychology and the Internet: Intrapersonal, Interpersonal, and Transpersonal Implications*, Academic Press, San Diego (Ca) – London.

Lepori B. (2000): , *Conference on European Research Information Systems CRIS2000*, Helsinki 25-27 May 2000 (available on-line at: http://www.cordis.lu/cris2000).

Lepori B., Cantoni L., Mazza R. (2002), Push Communication Services: a Short History, a Concrete Experience and some Critical Reflections, *Studies in Communication Sciences*, vol. 2, N. 1, p. 149-164.

Morris Ch. W. (1957): *Foundations of the Theory of Signs*, The University of Chicago Press, Chicago

Sperber, D.; Wilson D.(1995): *Relevance. Communication and Cognition*, Blackwell, Oxford, 2nd edition.

Thévignot C. (2000): The redesigned CORDIS web service contributes to the Commission's eEurope Initiative, *Conference on European Research Information Systems CRIS2000*, Helsinki 25-27 May 2000 (available on-line at: http://www.cordis.lu/cris2000).

Uspenskij B. A. (1996): Problemi di tipologia linguistica alla luce della differenziazione fra “chi parla” (mittente) e “chi ascolta” (destinatario), in Uspenskij B. A.: *Linguistica, semiotica, storia della cultura*, Il Mulino, Bologna, pp. 39-62.

7 Contact Information

Benedetto Lepori
Facoltà di Scienze della Comunicazione
Università della Svizzera italiana
via G. Buffi 13
6900 Lugano
Switzerland

phone +41 91 912 46 14
e-mail: blepori@unisi.ch; lorenzo.cantoni@lu.unisi.ch

What's your question? The need for research information from the perspective of different user groups

Nieske Iris Koopmans
Netherlands Institute for Scientific Information Services (NIWI)
Royal Netherlands Academy of Arts and Sciences (KNAW), Amsterdam

Summary

In this paper results of a field study into the need for research information of different user groups are presented: scientists, policy makers and policy researchers, industry and media. Main questions of semi-structured interviews were: what kind of research information users need, what kind of research information resources are used and which information resources are missing at the moment. User groups are missing for a diversity of reasons the overview of research, experts and institutes in the different scientific fields. Especially for the accessibility and transparency of the scientific world these overviews are reported to be needed. Neither Google nor any of the research institutes or policy research organisations are able to present surveys for different science fields at the moment. Giving users the possibility to search, browse and navigate through accessible and more specialised layers of research information might give answers to different user groups simultaneously.

1 Introduction

In former days developers of Current Research Information Systems (CRIS) have not always asked questions from the perspective of the user. The first priority was to build a database and to fill it, with due regard to coverage and currency (product orientation). Now time has come to look at systems from the user's point of view (user orientation).

What's your question? Asking this seems to be straightforward and simple, but simple questions can be the most complex ones. In the world of research information, this question is an important one. Why? It brings us back to the user and the information he or she is looking for. Do (potential) users need research information? Who are these users? What kind of research information is the user looking for? For which reason is the user searching for research information? We will try to give first answers to questions like these in section 3. How users want to search is the central question in section 4. In the last section we are looking at how we can fulfil the need for research information of the different user groups for different goals.

The research that has recently been done in the Netherlands for the Dutch Research Database (NOD) will be used as case material. The study has been carried out in order to be able to adjust the NOD to the needs of users of research information. In the months April and May 2002, 32 semi-structured interviews have been held among representatives of four user groups[1]: researchers (10), policy maker's (6) and policy researchers (5), media (7) and industry and services (8).

1 Some of the respondents can be placed in two of these groups.

2 Method

In the study for the NOD, the central subject is the need for research information of different user groups in the Netherlands. Research information is hereby being defined as: "National as well as international scientific research information resources around researchers, scientific institutes and projects. These can include expertise and contact information of researchers and institutes, descriptions of research, as well as overviews and references to products that are the result of research as publications, data sets, patents and tools."

Several user groups need research information in their daily work. The different tasks people have to fulfil at work, give an idea on the reason why someone is looking for research information and the circumstances from which the need for information arose. Borgman states that a lot of theories and research in the past presumed that the human activities involved in access to information could be isolated sufficiently to be studied independently (Borgman 2000, 7 pp). In a lot of these researches only one information system or service is being studied or the gathering of information is studied as an independent task. Here the search for research information is seen as an interwoven part of other activities at somebody's work. We also see that the need for a research information system is dependent on the total information environment with all the other information resources available. This is why we have chosen in first instance to do qualitative explorative research, using the method of semi-structured interviews.

For the preparation of the fieldwork, a literature study has been done and brainstorm sessions have been held, focused on the needs for research information of each user group. As respondents were selected: experts in the field of scientific information, key-representatives like the chairmen of certain user groups, as well as 'ordinary users'. The interviews were being held by two persons of the department of research information of NIWI-KNAW and were 1.5 hours of length. This research has given us a first impression of the needs of different user groups and will be a starting point for further studies in this area.

3 Main questions and the needs for research information of different user groups

What are the main questions different user groups are asking in the field of science? With which aim is the user searching for research information? What are the needs for research information of the different user groups at their work? What kind of research information are the different user groups missing? In this section we will try to give an answer to these questions.

3.1 Researchers

In this part of the study the group of researchers is defined as scientific researchers working at a university or research institute, who are not primarily working for third parties. Policy researchers and researchers who are working for the industry are represented in the other groups. We have spoken with researchers from different disciplines: history, social sciences, technical sciences and humanities.

There are three main reasons given by researchers to search for information:

- As part of doing a specific research
- To stay informed about the own discipline
- For networking and acquisition

Scientists need highly specialised research information resources, which are mostly internationally orientated. Most researchers say that they know their colleagues in their discipline and that they know which research is going on in their field. Contact with colleagues is mentioned as the most important way to get information. At conferences scientists are meeting

colleagues and keep in touch. Researchers are also very interested in conference papers, which are more easily distributed, because they are still without copyrights. Publications are stated to be the most important resource for information to researchers. Internet is being used a lot, most of the time by going directly to research institutes in the field or by searching with engines like Google. Researchers more and more see the need and necessity of databases. In general the trend is that databases are seen as part of the research infrastructure.[2] Other resources named as important are newsletters, data and software, pre-print archives, internal databases of organisations, discussion lists and email.

Descriptions of current research projects are scarcely used by researchers. Reasons are that in this early stage mostly results (publications) are not yet available, that the abstracts are written for the financier, that during the actual research they may already have changed and that they do not contain enough information. Scientists prefer to look at somebody's publications, to find out what somebody's specialisation is. Bibliographic databases are also seen as less interesting to scientists. Researchers are most interested in full-text online publications.

But still a lot of information is missed. The mentioned need for research information, which is not fulfilled by other resources among scientists, is:

- Research information from another discipline or subject, for multidisciplinary or interdisciplinary research.
- Research information for a researcher at the beginning of his career, for example as PhD researcher or Master student, when somebody is still orientating himself in the scientific world.
- Research information that can be used for trend studies and exploratory studies, to see patterns and trends, as well as studying the history of science over the years.
- Underlying research information system for the European Research Area (Networks of Excellence), as well as for electronic conferences and the 'virtual networking' of the future.
- Overview of 'networks' in different areas.
- Finding the more unknown researchers in a certain discipline.
- Research information might be more needed in specific areas. Mentioned is for example the field of Social Sciences, where there is also a need for publications in the own language (Dutch) which is not found in international literature.
- Contract research

Besides users of research information, researchers are also the producers of research information. Researchers say they are willing to supply information to a research information system, but only once and they need an incentive[3]. They don't have time to supply information for all different kinds of internal and external inquiries, visitations and databases. Information supply to the NOD is found especially interesting when research information is distributed from the NOD to other (inter) national databases.

3.2 Policy makers and policy researchers

Policy makers will be defined here as decision-makers and financiers of scientific research and policy research, working at a ministry, province or municipality (3.2.1). Policy researchers are doing contract research on behalf of these ministries, provinces or municipalities (3.2.2).

2 This is also visible in the expenses of the Netherlands Organisation for Scientific Research (NWO).

3 This is why in the Netherlands the research information system METIS is developed for universities, in which each researcher will have his own personal database, where he can among others publish his articles, presentations in different formats for the Internet. This system is also used for management information needed by universities.

3.2.1 Policy makers

The information need of policy makers is very broad. At a ministry the subject, about which information is needed, differs from time to time and is depending on the political agenda. Most of the time policy makers are not asking themselves which information sources to use to fulfil their information needs. Information seeking behaviour is done very subconscious and in an unstructured way. Some respondents were saying that there is not enough time to look for the right information. Above all there is already a lot of information coming automatically at their desks. A couple of respondents have the feeling of an information overload. But are policy makers able to find the research information they really need?

Two areas are seen where policy makers need research information, science policy and policy in general. There are three different sets of activities for which policy makers need research information:

- Making of decisions
- Justification of decisions
- Commissioning of research

Policy makers want to have surveys of current and recently completed research, linked to full-text online reports and publications. Summaries, recommendations and conclusions are the most important parts of publications for policy makers. Also research information is needed to justify policy decisions in a certain policy field. Being able to evaluate and monitor policy is becoming more important for transparency of government expenses and policy[4]. Personal networks and knowledge are dominant information resources. But also Internet and Intranet are used a lot. Among others several state-of-the-art, evaluation, and monitoring studies, are being commissioned to policy researchers.

For decision making in science policy, information is needed at macro-level. Questions dealt with are questions around the volume of research, the available expertise in different science fields. Policy makers want to make comparisons and to be able to see trends and patterns in and between science fields. To justify investments in science, it is important to know which resources have been put in the different science fields and to know the results of these activities. Policy makers want to see if resources that have been put in (input indicators), are leading to certain output, like a growth of publications, patents and so forth (output-indicators). To justify the scientific budget in the Netherlands, a lot of figures are missing. A systematic supply of quantitative information in the field of science is not available. Activities in the field of knowledge transfer are very difficult to measure at this moment. Information at international - particularly European - level is needed to be able to make comparisons between science policy in the Netherlands and other countries.

For commissioning of research knowledge is required to select institutes or experts for certain tasks. An overview of institutions or experts who are specialised to do research in a certain research field is necessary, as well as information about the individual institutes (profiles, contact information) and experts.

Information about individual researchers and research institutes is being found by searching at Internet with search engines like Google or by going directly to the sites of known organisations in the field. Intranet and internal databases are also mentioned as important sources of information. Also personal networks and knowledge are again playing an important role. Publications are mentioned as the most important source to get an indication about somebody's expertise. Guides are used frequently, especially the guide of universities and research institutes in the

4 Especially since from 1999 there is in the Netherlands the obligation to report in May of each year on the results of policy (VBTB), in addition to the government budget in September. Responsibility for the content of policy is being put central in these progress reports.

Netherlands[5]. Structured databases are used less frequently. An overview of the different databases is missing. It is also said to be very difficult to search in all the different databases, if one does not use them frequently.

Policy makers need information about experts and institutes in the first place on a national scale, but this has changed with the increasing internationalisation of research and becomes more manifest with the advent of the European Research Area. For finding referees information on an international scale is needed.

Policy makers have also other aims, which are not connected to their own information needs. In general these aims can be summarised in three main goals:

- Knowledge transfer among scientists
- Knowledge transfer among science and society
- Promotion of science on a national and international scale

Providing information services for the needs of other groups like scientists, business, media and the general public, is being mentioned as very important in all the interviews with policy makers. It is one of the tasks of the government to make publicly funded research available to everyone[6]. In the report of the study on the future of science in the Netherlands by Rand Europe, it is stated that: "Knowledge about knowledge could be enhanced by the existence of a publicly-available, detailed database containing information about scientific research. Such a database would facilitate the co-ordination among financiers, users and producers of research, resulting in the demand for research being more easily met and the supply being more easily accessed" (Kahan et al. 2001, 16). Hermans, the Minister of Education, Culture and Science in the Netherlands, is saying at the conference 'Access to publicly financed research' that he would break a lance for access to scientific knowledge for a broad and general public (Hermans 2001, 22). One of the interesting discussion points, which came up in one of the interviews, was if a country makes al its knowledge freely available on the Internet to its international competitors.

3.2.2 Policy researchers

Policy researchers are doing research for third parties (contract research). The information resources needed are very much depending on the research questions that are coming in from the contract parties. That is why it is very difficult for them to give a good overview of information needs. The contract parties are often policy makers at ministries, provinces or municipalities. The information needs of policy researchers are in this way connected to the information needs of the group of policy researchers (see section above).

The aims of looking for research information in this group are the same as those mentioned by scientific researchers: information needed for doing a specific study; information to stay informed about the own discipline or information for networking and acquisition. The latter aim seems to be more important for policy makers, since policy researchers are more dependent on getting contracts for new projects and studies than scientific researchers.

Differences between scientific researchers and policy researchers are that there is in general less time for literature study. There are also fewer possibilities to go to conferences. Policy researchers seem to find grey literature more important than scientific researchers. In summary it can be said that the need for research information is more specialised than the information need of policy makers, but more general than the information need of scientific researchers.

It is notable that policy researchers often speak about the necessity to work with databases for doing state-of-the-art, evaluation or monitor studies. To do systematic research controlled

5 The NOD is the data resource for this guide.

6 This is also why the NOD and several different thematic databases in for example the field of ICT or Health Care are being financed.

environments are essential. Google cannot be used as an analytical tool, but can be sufficient for doing a simple search e.g. for finding specific information about an expert.

3.3 Media

The user group of media will be defined here as professionals, who are focussed in their work to translate scientific information to the general public, like for example public relations officers of universities and science journalists. The aim of activities in the field of scientific communication is to make scientific information accessible for society.

The aims for using research information of *science journalists* are:

- Inspiration for (actual) subjects for articles
- Finding background information about subjects, among which information about experts

To get inspiration, journals like 'Scientific American', 'New Scientist', 'Science' and 'Nature' are playing an important role. Sometimes a topic is being translated to the Dutch situation. Dutch science magazines and science sections of newspapers are also being read. Press releases are an important source of information. Internet is mentioned as a very important source: searching with Google, looking at websites of universities for press releases or agendas, or news sites of for example the BBC. Also own databases with published articles are being used.

For finding background information the interviewees use the same kind of resources. Homepages of researchers are mentioned as important for finding information about experts, especially publication lists and CV-information. Also the guide of Universities and Research Institutes in the Netherlands is frequently used.

Science journalists attach much value to information about professors as well as PhD-researchers. Professors have an overview of their discipline. The graduation of PhD-researchers is a good moment to write about a research topic. Journalists seem to be looking especially for information about experts in the Netherlands. In general scientists have the reputation to be very open and willing to help journalists, but they are not seen as very pro-active. They do not often write articles or give material to the public media themselves. Mentioned is that information needs to be understandable for science journalists, but can be more specific than information for the general public. Publications are found to be too specific, but references and bibliographic information are being used. Information about current research is seen as important for knowing who is doing which kind of research at this moment.

The main task of *public relation officers* in science is to promote research and expertise of the institute. Public relation officers issue press releases and give advice to scientific journalists about research and expertise in the institute. Internal publications and expertise are the most important information resources. Also information about research and expertise of research available at other institutes is important for giving references. Personal contacts with other public relation officers are used as information resource for this reason, as well as the NOD and various guides. Research information databases are being used to make research and expertise of a university easily accessible and transparent for internal users like other scientists and external users like policy makers and science journalists. One university is using dynamic links to the NOD for this aim and another university has built an expert database making use of a content management system.

Suggestions coming up during the interviews were: to add press releases as information source, to add a field 'media willingness' in the NOD and to make a societal thesaurus.

3.4 Industry and services (Business)

The most important aim of the use of research information in industry and services is innovation. Knowledge is of becoming more and more important to all companies. The character of knowledge needed in companies is always multidisciplinary and applied.

In the user group of industry and services two groups can be distinguished:
- R&D-oriented companies (about 10% of the companies)
- R&D-followers (about 90% of the companies).

In the first category there are multinationals as well as high-tech orientated small and medium-sized enterprises. For the first category of companies, particularly current knowledge is needed. Persons working in the R&D-orientated companies predominantly have a high level of education and the need for and the use of research information is nearly the same as that of scientific researchers. One of the differences is that a lot of these companies provide specialised internal information services. 'Knowledge workers' are often doing literature searches.

The R&D-followers also have a need for research information. Here information intermediaries are needed. The Ministry of Economic Affairs is financing these intermediaries, who will work in regional innovation centres. The idea of developing a thesaurus in co-operation with the users is found to be very interesting. It is stated that information about research projects needs to be accessible, but not too specific and without the use of jargon. Press releases about research could be a resource of information.

In any case there is need for transparency and accessibility, being able to know who is doing which kind of research where. Overviews of current and completed research of a certain theme or subject, are stated as being important. Also the suggestion has been made to take a more proactive role by making analyses about overlaps, where companies and knowledge institutes could work together.

One of the main problems in the field of research information in this area is that companies and innovative research institutes will often not supply research information to public databases for secrecy and competition reasons. Van Raan is indicating that there is a trend that more and more economically relevant knowledge is being kept secret (Van Raan 2001).

3.5 Conclusion

In general three main factors can be distinguished regarding the needs for research information:
- Having knowledge and networks around a certain subject/discipline (experts) versus not yet having much knowledge and networks around a subject/discipline (newcomers).
- Working in a small-demarcated field versus working in a broad field.
- Needing specialised information resources versus needing more general and accessible research information resources.

Especially scientists who are having already a lot of knowledge and networks around a subject/discipline, say that they are able to find the information they need for doing their research by using existing available resources. Newcomers like PhD-researchers or master students have more difficulties to easily find the research information they need.

When a policy or scientific researcher is working in a more inter- or multidisciplinary field, it is more often noticed that it is difficult to have an overview and find the research information from different disciplines. User groups like policy makers, science journalists and business-people are all working in multidisciplinary fields.

Researchers are interested in specialised and more detailed information resources like publications. Information about current research projects is considered to be of minor importance. In other user groups people need less detailed information resources like accessible abstracts of current and completed research and would like to have easy mechanisms to find experts, for ex-

ample with a (societal) thesaurus. But these users also want to judge somebody's expertise by looking at publication lists and references.

4 Ways of searching for and the presenting of research information

The way the interviewees want to search is depending among others on how broad or specific the information need is, the knowledge of the subject somebody already has and the aim why somebody is searching for information. Searching is working well if the question is specific and well defined. Browsing and navigating is commonly used when the information need is not clear at all or very broad (see also Borgman 2000; Feng et al. 2001).

All user groups, particularly science journalists, are searching and want to be able to search in a simple and quick way, by typing in one or two words. Most people are searching with Google. The problem with this type of searching is the long lists with answers obtained; these include sometimes a lot of irrelevant material, and it takes much time to scan. Moreover it is not clear which words one should use in order to get the optimal result.

Besides searching, respondents want to be able to browse and navigate through a system, by using the structure of the database. They want to be guided in a visible and hierarchical way (tree structure) from discipline to sub-discipline, from institutes to faculty to researcher. This is also one of the reasons why people are using often the 'offline' printed guide for research institutes and universities. They want to navigate in a structured way though the keywords with thesauri and taxonomies. An interesting form is of course vector space analysis, in which words can be seen in their context in a visible way[7]. In the chaotic world of Internet users in all different user groups want a guide to easily and quickly find the path to reliable research information.

Finally, users would like to consult a system in a more 'unstructured way', by snowball searching. They want to switch for example easily from an overview of all experts on a certain subject to more information about an expert, the homepage of a researcher or to the publications of a researcher. Or from research information to the publications, to information about the researchers who are working or have worked on a research project. They want to click on the different keywords in the system and go through the system in this way.

A well-defined and controlled environment is essential and the most important difference and advantage over searching in Google. Users do not expect that a database is perfect in the sense of coverage and currency. Databases need not to be complete to succeed. What users expect is to have clear information about what they can find in the database. The resources of the database, the way of getting the information, the quality and currency of information in the database and the distribution of information to other databases has to be clear and transparent. Users also want to have options to exclude information (resources).

When it is possible to define a structural information need, users want to search in a demarcated area, like a thematic database. Users do not want to see all information, they only want to see relevant information.

Most interviewees are finding Intelligent Agents and Personal Information Environments not an interesting option, because information needs are often not clear or specific enough. When the information need is very specific, e.g. for researchers who are working in a mono-disciplinary field, they seem already to be able to find their information in other ways. But for researchers working in an inter- or multidisciplinary research environment, it might be interesting to develop these kinds of options. Especially the option of presenting research information about a subject from different disciplinary perspectives has been mentioned.

7 See for example http://www.inxight.com/.

5 How can questions of the different user groups be answered?

Research information has to be provided to user groups with different levels of knowledge and networks around a certain subject/discipline, who are working in demarcated as well as in a broad multidisciplinary field and who need different levels of detail and accessibility of research information. Providing different information layers with different levels of detail and accessibility might be a solution for tailoring research information to different user groups and different aims.

Besides opportunities for searching, users are finding it especially important to be able to browse through a system. It is particularly important to provide the opportunity to present overviews with taxonomies and thesauri, with which a user can start to browse and navigate further in a quality-controlled environment. These taxonomies and thesauri have to be developed in co-operation with the different user groups. In this way CRISes may become pre-eminently important to function as guides in the scientific world.

If a CRIS enables a researcher to profile his research and expertise in a broader context than his own discipline on a national and international - particularly European - level, it will become attractive to deliver research information to a CRIS. Policy makers, industry and business as well as the media are important groups. They are either funding research or are making research visible in a broader context. A researcher has to be asked to put effort in providing information about his expertise and research also in a more common vocabulary. But with distributed systems, the delivery of information to a research information system should take place only once for internal and external information systems and for different user groups. In this way a CRIS may become a central instrument for science communication and knowledge distribution in and between countries.

The development of research information environments on the level of a group or network is becoming more important in the future with the upcoming 'Networks of Excellence' of the European Research Area, as well as developments in the direction of electronic conferences and 'virtual networking'. For a lot of users personalisation in an environment working with intelligent agents is of less importance. The information need of most user groups, except for researchers, is often not well defined, depending on for example currency and scope (multidisciplinary).

Another important field is the studying of the research process and mechanisms for policy or scientific research. CRISes may be used for trend studies and exploratory studies, to see patterns and trends, as well as studying the history of science over the years. For these aims a centrally controlled database ('source database') is needed and archiving is important. For these studies users want to make counts and ratings, for example how many publications and capacity are available in different science fields. Users are also asking to include financial information.

For a diversity of reasons users are missing the overview of research, experts and institutions in different fields. Especially for the accessibility and transparency of the scientific world these overviews are needed. In general Internet is a good source for finding information about individual researchers and institutions or finding specific publications. But neither search engines like Google nor any of the research institutes or policy research organisations are able to present surveys for different science fields at the moment. The Internet cannot provide research information in a controlled environment: this is why research information systems are needed.

6 References

Borgman, C.L. (2000): *From Gutenberg to the Global Information Infrastructure; access to information in the networked world.* Cambridge/London: The MIT Press.

Feng, L.; Jeusfeld M.A.; Hoppenbrouwers J. (2001): Beyond Information Searching and Browsing: Acquiring Knowledge from Digital Libraries. Tilburg: Infolab (Tilburg University).

Hermans, L.M.L.H.A. (2001): ICT for Better Access to Education, Culture and Science (speech). In: Wouters P.; Schröder P. (Ed.): *The Global Research Village III Amsterdam 2000; Access to publicly financed research, conference report.* Amsterdam: NIWI, p. 22-24.

Kahan, J.P.; Van der Linde, E.J.G.; Van het Loo, M.; Vader J.; De Vries, H. (2001): *Vision on the future of scientific research; Executive Summary.* Leiden: Rand Europe.

Van Raan, A.F.J. (2001): The Journal as Pièce de Résistance in an Electronic Environment (interview). In: Wouters P.; Schröder P. (Ed.): *The Global Research Village III Amsterdam 2000; Access to publicly financed research*, background papers. Amsterdam: NIWI.

7 Contact Information

Nieske Iris Koopmans
Netherlands Institute for Scientific Information Services (NIWI)
Joan Muyskenweg 25
1096 CJ Amsterdam
The Netherlands

e-mail: iris.koopmans@niwi.knaw.nl

Accessing the Outputs of Scientific Projects

Brian M Matthews, Michael D Wilson, Kerstin Kleese-van Dam
CLRC, UK

Summary

We describe a science data portal for generic access to scientific data. This data portal is uses a *generic science metadata format* to catalogue and access science data from a range of disciplines. We describe the metadata format that is used and further discuss how this can be used in combination with library metadata formats, such as the Dublin Core, to access all the outputs of scientific projects, both data and publications.

1 Introduction

The scientific research projects have two major outputs: traditional publications, in journals and other forms of literature; and the data sets generated during the course of observations and experiments. These are then subject to analysis and visualisation to generate the results reported in the literature. Traditionally, science has concentrated on the former output as the major means of disseminating the results of research, whilst access to the latter has been restricted to small groups of individuals closely associated with the original researcher. However, modern distributed information systems offer the opportunity to provide access to both outputs to a wider audience. This allows other researchers to verify the results of the analysis, and also to reuse the data-sets to carry out secondary analysis, possibly in combination with results from elsewhere, to produce new insights without the cost of repeating the original experiment.

These data resources are stored in many file systems and databases physically distributed throughout organisations with, at present, no common way of accessing or searching them to find what data is available. It is often necessary to open and read the actual data files to find out what information they contain. There is little consistency in the information which is recorded for each data-set held and sometimes this information may not even be available on-line, being recorded only in experimenters' logbooks. This situation creates the potential for serious under-utilisation of these data resources or to the wasteful re-generation of data. It also hinders the development of cross-discipline research, as this requires good facilities for locating and combining relevant data across traditional disciplinary boundaries.

To address these problems, the concept of a *data portal* has been developed (Ashby et al. 2001a, 2001b; Houstis & Lalis 2001; NESSTAR; Ryssevik & Musgrave 1999). This offers a single method of browsing and searching the contents of scientific data resources, across a variety of scientific domains. Such a system has potentially a wide spectrum of users, from scientists working in related fields wanting to find information on a topic, through experimenters interesting in accessing and analysing their own data, to the data curators based at the facilities themselves who want to use the portal as a data management tool. In order to construct such tools, including mechanisms for cataloguing, browsing and accessing data resources, a generic metadata model for scientific data is needed. Such a metadata for science has the requirement of being both more specific than general metadata models such as the Dublin Core (Dublin Core), whilst being more general than specific metadata formats for specific domains in science, such as earth observation (Hoeck et. al. 1995). There are many metadata formats usually supporting specific data sources; a mechanism needs to be defined to access such metadata in an interoperable way

from the generic metadata that preserves the meaning, and allows deeper searches into the domain specific metadata. This approach also differs from generic representations of *science data* such as XSIL (XSIL) that has elements to represent arrays and tables, but little capability to represent provenance data and other information *describing* the science data.

A common metadata format for scientific data also allows the possibility of providing a single point of access to both the major outputs of science: data and publications. By using the common or interoperable features of the generic scientific metadata model, we allow the possibility of combined searches across both domains, or alternatively, using the metadata from one domain (say scientific publications) to search and access appropriate information from the other (say retrieve relevant data sets to test the claims of the publication).

We describe a briefly describe a Science Data Portal developed in CLRC. As a major component of this project a metadata model was defined. In the main body of this paper, we describe in some detail the structure of this metadata both in its overall structure, and some of the details. Further we then discuss how this metadata model can be related to metadata formats for cataloguing

2 A Science Data Portal

A pilot system has been developed to test these ideas that enables researchers to access and search metadata about data resources held at the ISIS and SRS facilities within CLRC, and further extended to cover the British Atmospheric Data Centre (BADC). The system being developed has 3 main components: a web-based user interface; a metadata catalogue; and generic data resource interfaces. These are integrated using standard Web protocols. It is anticipated that the system will exploit the emerging Grid Service infrastructure to offer a distributed interface to scientific data resources both inside and outside CLRC.

The data resources accessible through the data portal system may be located on any one of a number of data servers. Interfaces between these existing data resources and the metadata catalogue are being implemented as *wrappers* on web services that will present the relevant metadata about each resource to the catalogue so it appears to the user to be part of the central catalogue. These wrappers are implemented as XML encoding of the specific metadata relating to that resource using the metadata model schema; wrappers are an established technique for providing such interfaces (Baru et. al 1999).

3 A Metadata Catalogue

The logical structure of the metadata in the catalogue is based on the scientific metadata model developed in the project. This model exploits experience gained in developing general metadata models for other domains, such as the Data Documentation Initiative for social science (DDI) and has the overall structure of 6 major areas as shown in Figure 1. This structuring is influenced by the classification of metadata given in (Jeffery 2000). The study metadata corresponds to *associative descriptive metadata,* the access condition to *associative restrictive metadata,* data description to a form of *schema metadata* (describing how the data is laid out in the file structure), data location to *navigational metadata,* and related material to *associative supportive metadata.*

It is necessarily very generic to cater for a large range of differing types of data; specialisations of this metadata format will be used for each domain; generic queries can be then devised to search over the common views on the metadata. The model uses a hierarchical model of the structure of scientific research programmes, projects and studies, and also generic model of the organisation of data sets into collections and files. This allows a flexible structure to be developed, relating different data sets and their components together. For example related sets derived from one another from raw data through data reduction and analysis to a final result; alternative

and failed analyses can also be recorded, as well as calibration data sets, against which results are measured.

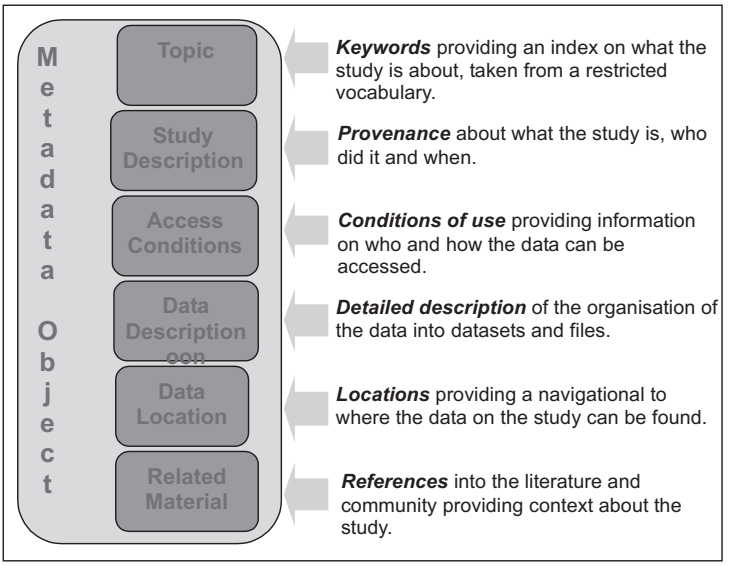


Figure 1: Overall Metadata Structure

The metadata catalogue is implemented using a standard relational database. Once the specific data sets required by the user have been identified using the available metadata, the catalogue provides links to the files holding the actual data. Users can then use these links to access the data with their own applications for analysis as required.

4 The Metadata Structure

The metadata within the general metadata structure is laid in a series of classes and subclasses. We do not describe the whole model in detail for reasons of space, but rather select some areas of particular interest.

4.1 Modelling Scientific Activity

The data model attempts to capture scientific activities at different levels: generically, all activities are called ***Studies***. Each study has an ***Investigator*** that describes who is undertaking the activity, and the ***Study Information*** that captures the details of this particular study. The general structure of the metadata is given as a UML diagram in Figure 2.

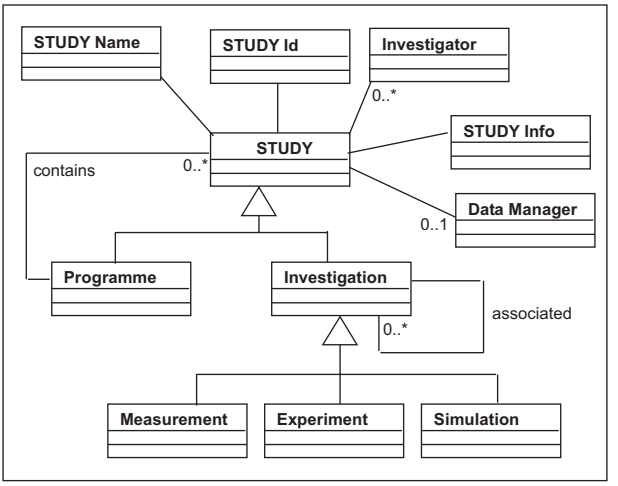


Figure 2: The UML model for the Study

Studies can be of different kinds, as represented by the subclass information in the UML diagram.

- ***Programmes:*** are studies that have a common theme, and usually a common source of funding, instigated by a principal investigator or institution. Programmes can be single projects (such as EPSRC projects, or application for beam time on ISIS), linked sequences of projects; for example an EPSRC Faraday project would have a set of linked projects. Each programme can thus be associated (linked) with a series of sub-investigations. Programmes are not expected to have direct links to data, but rather through the set of investigations within the programmes. ***Investigations:*** are studies that have links directly to data holdings. More specific types of investigations include experiments, measurements or simulations.
- ***Experiments:*** investigations into the physical behaviour of the environment usually to test an hypothesis, typically involving an instrument operating under some instrumental settings and environmental conditions, and generating data sets in files. E.g. the subjection of a material to bombardment by X-Rays of known frequency generated by the Synchrotron Radiation Source with the result diffraction pattern recorded.
- ***Measurements:*** investigations that record the state of some aspect of the environment over a sequence of point in time and space, using some passive detector, e.g. the measurement of temperature at a point on the earth surface taken hourly using a thermometer of known accuracy.
- ***Simulations:*** investigations that test a model of part of the world, and a computer simulation of the state space of that model. This will typically involve a computer program with some initial parameters, and generate a dataset representing the result of the simulation. E.g. a computer simulation of fluid flow over a body using a specific program, with input parameters the shape of the body, and the velocity and viscosity of the fluid, generating a data set of fluid velocities

Each investigation has a particular purpose and uses a particular experimental set up of instruments or computer systems. Experiments may be organised within larger studies or projects, which themselves may be organised into programmes of linked studies.

Classes within the model have several fields. For example, within investigator has a name, address, status, institution and role within the study. For reasons of space we cannot provide a com-

plete description of all the available classes within the metadata model. For illustration, we consider the Study class. Within a Study, there are several fields, as in the following table.

Study Description Class Fields	
Funding	Source of funds of the study, including grant-funding body.
Time	Date, time and duration of study. Can be either a point time and date, or a begin time and end time. We expect it to be in a standard format: dd/mm/yyyy for dates; hh:mm:ss for times.
Purpose	Description of purpose of study, including • Free text abstract of investigation • Keywords categorising subject of investigation – preferably selected from a controlled vocabulary. • Study type: a field that can be used to indicate the type of study being undertaken – such as a calibration run.
Status	Status of study, (*not-started, in progress, complete…*).
Resources	Statement of the resources being used, e.g. which facility.

4.2 Modelling scientific data holdings

The metadata format given here is designed for use on general scientific data holdings. These data holdings have three layers: the experiment, the logical data, and the physical files. The overall structure of the model for scientific data holdings is given in Figure 3.

An investigation is a study that generates raw data. This raw data can then be processed via a set of tools, forming on the way intermediate data sets, which may or may not be held in the data holding. The final processing step generates the final analysed data set. At each stage of the data process stores data in a set of physical files with a physical location. It is possible that there may be different versions of the data sets in the holding. In a general data portal, all stages of the process should be held and available as reviewers of the data holdings may wish to determine the nature of the analysis performed, and other scientist may wish to use the raw data to perform different analyses. Thus each *data holding* takes the form of a hierarchy: one *investigation* generates a *sequence* of logical *data sets*, and each data set is instantiated via a *set* of physical files. The design of the metadata model is tailored to capture such an organisation of data holdings. A single metadata record in this model can provide sufficient metadata to access all the components of the data holding either all together or separately.

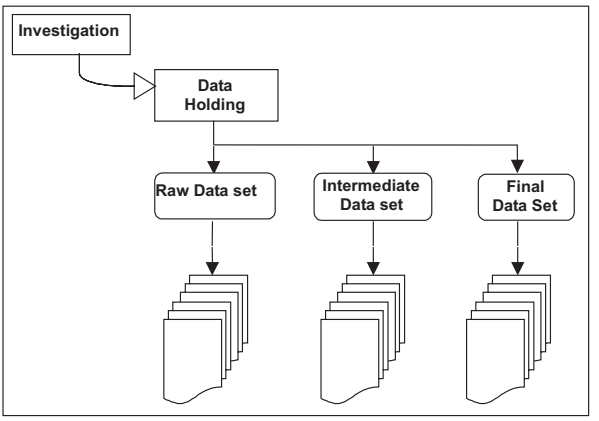


Figure 3: Model of the hierarchy of scientific data holdings

- ***Access control***: Access is controlled by the access entry in the metadata record; how this is actually done is dependent on the data holder. For example, there might be an access type, with settings such as "*open*", "*on application*", "*restricted*", "*commercial in confidence*". This may be given in conjunction with explicit instructions on how to access the data, and who to contact.
- ***Data Location***: The data location provides a mapping between the URI's used in the data definition component of the metadata model, and the actual URL's of the files. This can provide facilities for describing mirror location for the whole structure, and also for individual files.

5 Example

As an example of this scientific metadata model, consider the SXD information from the ISIS neutron spallation source. A *study* in this case is an application for beam-time, uniquely identified with an 'RB number', which covers a programme of investigations, and is described by a description of the purpose in the original study application. This programme is in turn broken down into a series of individual investigations, each of which are experiments on the SXD detector. Each investigation may have a sequence of *runs*, each generating a data set. Each run keeps the major parameters of the experiment the same (e.g. temperature of study), but alter some other parameter (e.g. orientation of the sample in the target).

For example, consider an investigation has name *Benzene*, *variable temperature study: 150K*. It should have a unique ID - this is not necessarily the RB number as that may relate to a programme of investigations, but it might be generated from it. It will have associated with it a set of RAW files, for example: files SXD10091, SXD10092, SXD10093, SXD10094, SXD10095: Benzene, variable temperature study: 150K. There may also be a set of intermediate SXD files, and also a set of processed final files in standard data formats for specific programs, such as .HKL, .INS and .RES files. The system keeps track of the relationships between files, and records which have been processed. We give a small sample of the fields in the metadata. We use *#classname* to represent cross-references between classes.

Experiment	
StudyID	SXD10091
Study Name	Benzene, variable temperature study: 150K
Investigator	*#investigator*
Study Information	*#study-information*
Data holder	*#data-holder*
Instrument	*#instrument*
Environmental Conditions	*#conditions*

The Investigator gives details of the people involved in the study.

Investigator	
Name	Anne X. Perimenter
Institution	University of Somewhere
Status	Lecturer
Role	Principal Investigator
Address	Dept of Organic Chemistry, Univ of Somewhere, UK.

Study information gives the information on this study.

Study Information	
Funding Source	EPSRC
Time	1/11/00, 11.45
Purpose	*#purpose*
Status	Complete
Resources	Beam time on ISIS using the SXD, for 1hr on 1/11/00

The Purpose itself may have several fields.

Purpose	
Abstract	To study the structure of Benzene at a temperature of 150K.
Keywords	Chemistry: organic: benzene: denatured benzene, C6H6

The data holder refers to the institution principally responsible for holding the data - this is not a locator in the sense of a URL.

Data Holder	
Institution	ISIS, CLRC Rutherford Appleton Laboratory

The conditions in this case just record the temperature under which the sample has been studied.

Conditions	
Temperature	150K

Files may also be in several different locations, separating out the identity of data sets from the location. Giving filetype/directory pairs does this:

Data location	
Data holding locations	ftp://ftp.isis.rl.ac.uk/SXD/ SXD1009/http://www.dooc.uos.ac.uk/~perimenter/bezene/
Data set Directories	(RAW, "raw/"), (Intermediate, "SXD/"), (HKL, "HKL/"),…

The data description would break down into a hierarchy of entries. Firstly the top-level entry, which contains references to the data sets of the study.

Data description	
Data Sets	*#raw, #intermediate, #processed*

Then the raw data set would have references to the metadata for each file (not the file itself):

Raw	
Dataset type	RAW
Files	#SXD10091.RAW,#SXD10092.RAW, …

Each file would have an entry, giving its URI:

SXD10091.RAW	
URI	SXD10091.RAW

There will also be a dataset entry for intermediate and processed files.

6 Mapping to Dublin Core

The data portal offers the potential for integrating the outputs of scientific research, thus producing a combined portal for literature and data. Thus a feature of this would be not only the linking of publications to the data set which they depend upon, but the use of literature to guide a more general search for appropriate data in the area, and also from data to appropriate literature which could be used for further analysis. Clearly, to enable this, the metadata formats of the two systems will have to be related to enable searches from one metadata system to be passed to the other. Clearly, there is much commonality between the generic science metadata used in the data portal with generic formats proposed for library systems, especially Dublin Core and CERIF.

The 15 standard elements of the Dublin Core all have their counterparts within the much more details structure used within the Data Portal, and through Dublin Core's "dumbing-down" principle can easily be abstracted, although potentially with little precision.

Mapping between Dublin Core and Data Portal Science Metadata

Dublin Core Element	*Science Metadata Class path and attribute*
Title	Study: Name
Creator	Study: Investigator: Name (Role is principle investigator)
Subject	Topic: Keyword
Description	Study: Study Information: Purpose
Publisher	Investigation: Data Manager
Contributor	Study: Investigator: Name ; Investigation: Data Manager
Date	Study: Study Information: Time
Resource Type	If a data holding is being referenced, this should be set to *Collection*; if a single data-set, then this should be set to *Dataset*.
Format	Data Description: File Format
Resource Identifier	Study: Study Id (for the whole study) Data description: File: URI (for individual data files).
Source	Data description: Data sets: Related Data sets Related Material: Related work
Language	*Not covered in the current metadata format; but an simple extension*
Relation	Related Material: Related work
Coverage	Data description: Logical Description: Coverage
Right Management	Access Conditions

Thus a common search can be set up between the CLRC Data Portal and Dublin Core enabled library catalogues. The more complex model provided by CERIF (CERIF) provides the opportunity for a more precise mapping of the provenance metadata, and a consequentially more better retrieval. For this to be enabled, a mapping would need to be established between the Data Portal's science metadata format's model of the scientific hierarchy (with programmes, projects, participants, studies and experiments) and CERIF's model using People and Project entities.

7 Project Status and Future Plans

This pilot project was completed at the end of March 2001 with the operation of a working prototype system. The longer-term goal is to extend the system to provide a common user interface to metadata for all the scientific data resources held in CLRC. Work in progress is taking the system embedding the system into the facilities and also extending the range of the portal, for example allowing access to earth observation data via the same portal; a new version has been released in April 2002 (CLRC Data Portal) and it is planned to extend the use of the system to materials sci-

ence. In this process, the generic science metadata has proven remarkably robust, with only small changes needed.

Beyond this, the publication of scientific data as "grey literature" in its own right, together with its appropriate metadata affords the opportunity of it being curated as part of the "corporate memory" of the research organisation, treated and available as an important asset in its own right, rather than a disposable, and in the medium term, uninterpretable legacy of past activity.

Acknowledgements and Contacts

The CLRC Data Portal Project is part of the CLRC E-Science programme (http://www.e-science.clrc.ac.uk), within the UK Research Council's e-Science Initiative. We would like to thank the Data Portal team who has contributed extensively to the definition of the Science Data Model.

8 References

J V Ashby, J C Bicarregui, DR S Boyd, K Kleese van Dam, S C Lambert, B M Matthews, K D O'Neill. (2001a): The CLRC Data Portal British National Conference on Databases

J V Ashby, J C Bicarregui, D R S Boyd, K Kleese van Dam, S C Lambert, B M Matthews, K D O'Neill (2001b): A Multidisciplinary Scientific Data Portal HPCN 2001: International Conference on High Performance and Networking Europe Amsterdam

C. Baru, A. Gupta, V. Chu, B.Ludäscher, R. Marciano, Y. Papakonstantinou, P. Velikhov, (1999) XML-Based Information Mediation for Digital Libraries, Digital Libraries '99.
www.npaci.edu/DICE/Pubs/dl99-demo.pdf

CERIF: the Common European Research Information Format
www.cordis.lu/cerif/

CLRC Data Portal Project
www.escience.clrc.ac.uk/Activity/ACTIVITY=DataPortal

The Data Documentation Initiative
www.icpsr.umich.edu/DDI/

Dublin Core Metadata Initiative
www.dublincore.org/

H. Hoeck, H. Thiemann, M. Lautenschlager, I. Jessel, B Marx, M. Reinke (1995): The CERA Metadata Model Technical Report No. 9, DKRZ - German Climate Computer Centre,
www.dkrz.de/forschung/reports/report9/CERA.book.html

C. Houstis, S. Lalis, (2001): ARION: An Advanced Lightweight Software System Architecture for accessing Scientific Collections, Cultivate Interactive, no.4,
www.cultivate-int.org/issue4/arion/

K G Jeffery. (2000): Metadata Information Systems Engineering Sjaak Brinkkemper, Eva Lindencrona, Arne Solvberg (Eds), Lecture Notes in Computer Science, Springer Verlag ISBN 1-85233-317-0.

NESSTAR (Networked European Social Science Tools and Resources)
www.nesstar.org

J. Ryssevik, S. Musgrave (1999): The Social Science Dream Machine: Resource discovery, analysis and delivery on the Web, the IASSIST Conference, Toronto,
www.nesstar.org/papers/iassist_0599.html

XSIL: Extensible Scientific Interchange Language,
www.cacr.caltech.edu/SDA/xsil/

9 Contact Information

Brian Matthews
CLRC
Rutherford Appleton Laboratory
Didcot
OX11 0QX
UK

e-mail: b.m.matthews@rl.ac.uk; .d.wilson@rl.ac.uk; k.kleese@dl.ac.uk

Workshops

Data Collectors meet Data Suppliers on the Internet

Dirk Hennig, Wolfgang Sander-Beuermann
Institute for Computer Networks and Distributed Systems

Abstract

To build up well working research portals data collectors have to follow data suppliers. Almost all data suppliers nowadays do supply their data by themselves to servers on the Internet; they do that in any way, as THEY want to do: no prescribed formats, no prescribed protocols, nothing prescribed ... (we are living in a free world).

Does this mean that nowadays „handmade“ research information databases, like those which are maintained by lots of universities covering their own research projects are dispensable? Certainly not, because they embody high quality information which is not present in any of the searchengines.

Therefore the goal must be to combine these two different worlds of information systems. One way might be http://forschungsportal.net/, various other ways will be presented and discussed too.

Contact Information

Wolfgang Sander-Beuermann
University of Hannover
Computer Center of Lower Saxony
Institute for Computer Networks and Distributed Systems
Schlosswenderstr. 5
30159 Hannover
Germany

e-mail: wsb@rrzn.uni-hannover.de

CERIF-2000
(Common European Research Information Format)

Andrei Lopatenko
University of Manchester, UK

Abstract

CERIF-2000 is a set of guidelines, as well metadata and database formats, developed during teh 1990's. The first version was drafted in 1991; the latest version in 2000. The basic idea behind CERIF is to publish a compendium of common practices of research information publication and use; establish clear guidelines for CRIS (Current Research Information System) development which may be helpful for CRIS developers and hosts; and set forth a metadata format for research information which assists research organizations in information exchange, thanks to its common information structure and use of vocabularies.

The CERIF workshop is organized in order to achieve the following main goals:
- to report on CERIF TG activities and results
- to date (metadata, enterprise portal, semantic web solutions for CERIF)
- to collect feedback and experiences from global CRIS developers
- to organize a CERIF metadata online dialogue between CRIS developers
- and users to plan future CERIF activites

Contact Information

Andrei Lopatenko
The University of Manchester
Oxford Rd., Kilburn Building, r.2.112
Manchester M13 9PL
UK

e-mail: alopatenko@cs.man.ac.uk

Embedding of CRIS in a university research information management system

Jostein Helland Hauge
Bergen University Library, NORWAY

Abstract

Traditionally, a Current Research Information System (CRIS) has been regarded as a portmanteau information system that was supposed to serve the interests of a wide variety of user groups. These would typically include research administration, researchers, students, media organisations, users of research in government and private industry as well as the general public.

In a number of cases it has, however, turned out that the mediation effect of CRISes has been considerably lower than expected. For this reason, we are now witnessing the advent of CRIS variants that more directly targets one of the above-mentioned user groups.

Instead of hold-all systems one can e.g. find CRIS systems that in some way or other are more directly serving administrative and strategic objectives of a university or research organisation. In turn, this will determine the modelling of this kind of CRIS systems and give rise to a suite of challenges concerning e.g. the seamless integration of the research information system with other related parts of the organisation's research management and reporting systems.

In this workshop we invite participants at the CRIS 2002 conference to share information on the above-mentioned topic, demonstrate and comment on their own applications and take part in a general discussion based on the examples presented.

Contact Information

Jostein Helland Hauge
Bergen University Library
University of Bergen
Parkveien 9
5020 Bergen
Norway

e-mail: jostein.hauge@ub.uib.no

A European Research Information System (ERIS): an infrastructure tool in a European research world without boundaries?

M.L.H. Lalieu MSc
NIWI-KNAW, Amsterdam

Abstract

This workshop will address the opportunities of a future ERIS from three perspectives, i.e. policy, content and technology. Aim is to identify all relevant aspects, to develop a coherent view on the subject and to arrive at conclusions and recommendations. This will be the basis for mobilizing joint expertise for the development and implementation of ERIS.

Questions to be discussed are: what kinds of information should be available to the international research community and the policy echelons in order to comply with the goals formulated for the European Research Area? What are the views of national research councils and international science bodies? What would be the technological answer to the challenge to combine a huge variety of distributed and heterogeneous information resources?

Contact Information

Harrie Lalieu
NIWI-KNAW
P.O. Box 95110
1090 HC Amsterdam
The Netherlands

e-mail: harrie.lalieu@niwi.knaw.nl

Poster

CERIF-2000 as a platform for university public research information service

Andrei Lopatenko
University of Manchester, UK

Abstract

The poster describes AURIS-MM (Austrian Research Information System - Multimedia Extended) as CRIS to provide access to information about scientific results of Austrian universities. A set of requirements to university information system is described with description how CERIF-2000 suits those requirements. The advantages and disadvantages of CERIF-2000 are emphasized and solutions to solve possible problems are suggested. The aim of the poster is to provide very practical description of CERIF-2000 as a database for university CRISs.

In details the followings CRIS services are described:

1) full-text search interface for all information in the database with support of sophisticated full-text query operators;
2) attributed search interface for each type of CERIF entity;
3) effective use of keywords and thesaurus in search operations;
4) ability to have some common reasoning about information under condition of a lack of information;
5) sophisticated navigation access for exploration of information.

Contact Information

Andrei Lopatenko
The University of Manchester
Oxford Rd., Kilburn Building, r.2.112
Manchester M13 9PL
UK

e-mail: alopatenko@cs.man.ac.uk

ELFI
ELectronic Research Funding Information System

Andreas Esch
ELFI- Servicestelle für elektronische Forschungsförderinformationen

Abstract

The modern electronic research funding information system ELFI (**EL**ectronic Research **F**unding **I**nformation System) has been developed with the co-operation of the Ruhr-University Bochum and the GMD (Gesellschaft für Mathematik und Datenverarbeitung, Sankt Augustin). One source provides all important information on research such as deadlines, research programmes, application procedures etc. A modern and innovative database-technology and a relevant bibliographic and factual adaptation of the information enables the user to create a personalised web-page adapted to his desires. The most important things are included:

- Funding Organisation
- Funding department
- Procedures
- Conditions and authorisations

ELFI fulfils its aim as a database by being:

- Up to date (fast access to new data by using the web-robots)
- consistent (uniform data format),
- reliable (use of links to original sources of information) and
- precise (selection of personalised profiles/active view).

Contact Information

Andreas Esch
ELFI- Servicestelle für elektronische Forschungsförderinformationen
Ruhr-Universitaet Bochum
Dezernat 2 Forschungsförderung und internationale Angelegenheiten
Universitaetsstr. 150
44780 Bochum
Germany

e-mail: andreas.esch@uv.ruhr-uni-bochum.de

Estonian r&d information system - ERIS

Taavi Tiirik
Archimedes Foundation

Abstract

ERIS (http://www.eris.ee) is a database-based research information system that became operative on the web in September, 2001. It consists of 3 main elements: the database of researchers, projects and research institutions. At the time being, there is data about 2300 research projects, 2150 researchers and 209 research institutions available on ERIS. The research information system reflects the project funding provided by the Estonian Ministry of Education. Key parts of ERIS are proposal entry system and partner search facilities.

The system that is being established should provide an overview of the division of research and development activities, financing and results, and should facilitate the formation and evaluation of the Estonian research and development and innovation policy. To the institutions and scientists involved in research and development the system gives a new opportunity to introduce their activities in Estonia and internationally.

Contact Information

Taavi Tiirik
Archimedes Foundation
Kompanii 2
Tartu 51007
Estonia

e-mail: taavi@ibs.ee

Including a Campus-Wide Publications List System into the existing CRIS of a University

Franz Holzer, Eva Bertha and Franz Haselbacher
(TUG), Austria

Abstract

Graz University of Technology (TUG), which was founded in 1811, has gained some reputation for its competence in research & development. In its previous version, TUG's web-based online CRIS (www.tugraz.at/research) started in 1997, including a freely accessible documentation of ongoing and completed research activities, combined with the lists of scientific publications originating from such activities. It proved to be highly sufficient as long the university was mainly interested, within this framework, in the transfer of knowledge and technology, as well as in the scientific co-operation with partners, both in industry and in academia. Recently, however, an additional aspect has become more important in research policies: the analysis and evaluation of scientific impact, based on bibliometric tools. As a result, the documentation of scientific publications now requires a higher level of data quality than before, when it was sufficient to inform one's peers in the field. In this paper, TUG's newly developed documentation of scientific publications is presented. It satisfies more demanding expectations inside an increasingly competitive academic world. But we had to keep in mind the financial limits of a university which is about to leave the safe harbour of civil service and government administration, and is bound to the challenging seas of academic autonomy.

Contact Information

Franz Holzer
Graz University of Technology
Research and Technology Information Unit (FTI)
Schloegelgasse 9
A-8010 Graz
Austria

e-mail: franz.holzer@tugraz.at

Information Interface for RTD Co-operation between the European Union and Russia

Irina Gaslikova
Centre for Science Research and Statistics, Russia

Abstract

This presentation provides some of the results obtained within the framework of the INCO-Copernicus-2 project „Building of an Information Interface for RTD Co-operation between the European Union and Russia".

The tailor-made web-site http://fp5.csrs.ru developed within this project grants a wide range of information for both Russian and European researchers. The web-site consists of two main parts addressing European and Russian research communities. The English version provides assistance to European RTD units in their search for Russian partners and consortium building. The Russian-language web pages give a detailed description of FP5 structure and opportunities for Russian researchers to participate in FP5 and other EU programmes. As a basis for establishing a national gateway for a two-way flow of FP5-related information the web-site is linked and accessible from CORDIS.

Contact Information

Dr. Irina Gaslikova
Division Head
Centre for Science Research and Statistics
11, Tverskaya str.
Moscow K-9, GSP-9, 101999
Russia

e-mail: gasl@csrs.ru

KM_LINE: Knowledge Management for Local Innovation Networks *e.services* platform

Adriana Agrimi, Giuseppe Bux
Tecnopolis CSATA Scrl, Valenzano (Bari – Italy)

Abtract

The purpose of the KM_LINE platform is to provide local innovation agents, such as Universities, new technology based firms, emerging KIBS (Knowledge Intensive Business Services) agencies, etc., with web services supporting promotion of innovation/research initiatives and dissemination/exploitation of related results. On top of the knowledge being produced and disseminated through the KM_LINE web services, virtual communities of local innovation agents are expected to take off on the Internet. In order to favour such take off, basic issue of the KM_LINE web services is to provide knowledge providers with a common metamodel for describing their knowledge items. CERIF (Common European Research Information Format) 2000 data model expressed in RDF (Resource Description Framework) is assumed as the reference metamodel for knowledge description and manipulation. CERIF 2000 is the reference data model for the current european network of CRIS (Current Research Information Systems). Resource Description Framework (RDF) is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. On top of the CERIF_RDF metamodel, the KM_LINE web services will support local innovation agents to build their own metadata web ontologies, in whose frame their specific knowledge items will be provided and made available for world wide intelligent research and automated manipulation. The KM_LINE platform is currently in the definition phase. Its development occurs in the frame of a two-year (2002-2003) research programme “Il Mezzogiorno Verso la Società dell’Informazione” promoted by the Italian Ministery of University and Research (MIUR).

Contact Information

Adriana Agrimi
Tecnopolis CSATA Scrl
Valenzano
Italy

e-mail: a.agrimi@tno.it

METIS: the Research Information System of the Dutch Universities

Eduard Simons
Nijmegen University, Nijmegen, The Netherlands

Abstract

METIS (previously known as OZIS) is the name of the research information system, currently used by the majority of the Dutch Universities. The system has been (and continuously is) developed by Njimegen University and has, in the last five years, gradually been implemented by other universities in the Netherlands. METIS has been built in ORACLE and is currently available as a Web-based (WWW) information system in both a Dutch and an English version. Part of METIS is a translation module which allows easy conversion of the system to no matter which other language. The translation module also allows an institution to redefine the terms, names, headers, etc… used in METIS, in order to really tailor the system to local needs and habits. The system is fully CERIF compatible and can hold a multitude of data concerning research, e.g.: content description, information on researchers, financial data, publications, cooperation and contacts, etc… Among the most striking features of METIS are: a *personal module* for researchers, allowing them to create their own individual research Web page based on data in METIS and a *management module*, offering a possibility to create almost every table conceivable concerning the in- and output of research, including a graphical bar chart presentation of the table data. From the second half of 2002 on, METIS will be available to interested universities outside the Netherlands.

Contact Information

Eduard Simons
Nijmegen University
The Netherlands
Tel. +31-24-3612343

e-mail: e.simons@dcm.kun.nl / e.simons@icarin.fiuc.org

Research Information System of the University of Tartu

Viktor Muuli
University of Tartu, Head of Research Office

Abstract

The Research Information System of the University of Tartu (RIS UT) has been created in 1999 and introduced to personnel of the university in 2000.

The goal of the RIS UT is to concentrate information about R&D activities and outcomes into one systematic whole to allow quick and complex analysis of R&D activities based on data in RIS UT.

Information about projects and academic staff has lead into RIS UT by university central administration. Entries about research output by registered users from academic structure. There are four different levels of accessibilty in this system: public user, 2 kind of registered users and administrator.

Since year 2002 all scientific information of University of Tartu has to be accessible from RIS UT and all statistical data and reports will base on data in the RIS UT.

Contact Information

Viktor Muuli
University of Tartu
Ülikooli 18-304
50090 Tartu
Estonia

e-mail: Viktor.Muuli@ut.ee

Research Project Database - Forschungsthemen-Datenbank Sachsen-Anhalt (LSAFODB)

Sylvia Springer
Otto-von-Guericke-Universität Magdeburg, Germany

Abstract

Research Project Database of the Region Sachsen-Anhalt with 2,6 Million people
www.forschung- sachsen-anhalt.de

A central organised and distributed actualised, web-based Oracle-database - 1600 projects from 700 project managers from 2 universities, 7 colleges, 4 research institutes.

Developed and managed by Otto-von-Guericke-Universität Magdeburg, Technologie-Transfer-Zentrum,
Head of Office: Dr. Sylvia Springer, www.ttz.uni-magdeburg.de

History of the project:

1997:Analysis of research project data bases at German universities and selection of the database solution of the University of Konstanz

1998: Adaptation of the Konstanz's database solution on the organization structure of the University of Magdeburg. Programming of additional components, how the description of interdisciplinary projects and extended search functions. Beginning of the introduction of the Research Project Database at the University of Magdeburg (MAGFODB) with 400 data records

1999: Use of the MAGFODB by the Universities Greifswald and Potsdam

2000: Extension of the MAGFODB for use by other research institutions of the region Sachsen-Anhalt and creation of the bases of a regional research project data base (LSAFODB). Content wise development of the data base and filling with data.

2002: Presentation of the Data base on fairs and meetings.

Programming of a English-language interface and development of English-language search

Future Aspects: Creation of access prerequisites for companies and single researchers. Preparing components to provide data for the annual research report. Planning links and strategies for integration in overlaid networks.

LSAFODB is supported by the Kultusministerium of Sachsen-Anhalt.

Contact Information

Sylvia Springer
Otto-von-Guericke-Universität Magdeburg
Technologie-Transfer-Zentrum
PF 4120
39016 Magdeburg
Germany

e-mail: sylvia.springer@ttz.uni-magdeburg.de

The Architecture of an Information Portal for Telecommunications

Kerstin Zimmermann
ftw. (Telecommunications Research Center Vienna), Austria

Abstract

Interdisciplinary research in telecommunications is a rapid growing field. The combination of various technologies has created groups with engineers, computer scientists, mathematicians and physicists working together. New problems have arisen within these teams. How do we find and exchange the necessary information produced by the entire field working on the same topic but using different classification schemes?

We would like to have one starting point for one (natural) field. This also helps us to stay in touch with the worldwide community and delivers structured information from around the world suited to our individual needs. A structured information platform is required where the different players can retrieve the data in a particular way. Each discipline uses its own classification scheme such as AMC in computer science, IEEE, IEE in engineering, MSC in mathematics and PACS in physics.

In the online archives you find preprints and documents. The publisher's information portals offer news and edited papers but at both have no additional information about the author or research group. Therefore other types are such as Math-Net, the International Information and Communication System or PhysNet, the Worldwide Physics Departments and Documents Network.

Let us consider the conception of an academic service for telecommunications. At the moment there are nine categories chosen. Some nearly identical to the categories in the portals mentioned above like research institutes, documents / infos and conferences but there are also some different like standards, legal and provider.

You can have a look at the prototype at http://userver.ftw.at/~kerstin/telecomportal/

Contact Information

Kerstin Zimmermann
Researcher for Information Managment
ftw. Forschungszentrum Telekommunikation Wien
Techgate
Donau-City-Str. 1/3
A-1220 Wien
Austria

e-mail: zimmermann@ftw.at
http://userver.ftw.at/~kerstin

Authors alphabetical list of names / addresses

Marek Andricík
Vienna University of Technology, Gusshausstrasse 28 / E015, A-1040 Vienna, Austria
e-mail: andricik@derpi.tuwien.ac.at

Adriana Agrimi
Tecnopolis CSATA Scrl, Valenzano, Italy
e-mail: a.agrimi@tno.it

Anne Asserson
Research Documentation Unit, University Library, University of Bergen, N-5020 Bergen, Norway
e-mail: anne.asserson@ub.uib.no

Laura Bartolo
Kent State University, College of Arts & Sciences, Kent, Ohio 44242-0001, USA
e-mail: lbartolo@kent.edu

Eva Bertha
Graz University of Technology, Research and Technology Information Unit (FTI), Schloegelgasse 9, A-8010 Graz, Austria

Beat Birkenmeier
Business Results GmbH, Siewerdtstr. 105, CH-8050 Zürich, Switzerland
e-mail: birkenmeier@bresults.ch

Giuseppe Bux
Tecnopolis CSATA Scrl, Valenzano, Italy

Lorenzo Cantoni
Facoltà di Scienze della Comunicazione, Università della Svizzera italiana, via G. Buffi 13 6900 Lugano, Switzerland
e-mail: lorenzo.cantoni@lu.unisi.ch

Andreas Esch
ELFI- Servicestelle für elektronische Forschungsförderinformationen, Ruhr-Universitaet Bochum, Dezernat 2 Forschungsförderung und internationale Angelegenheiten, Universitaetsstr. 150, 44780 Bochum, Germany
e-mail: andreas.esch@uv.ruhr-uni-bochum.de

Michael Friedrich
Wilhelm-Schickard-Institute for Computer Science, University of Tübingen, Sand 13 D-72076 Tübingen, Germany
e-mail: friedrich@informatik.uni-tuebingen.de

Irina Gaslikova
Division Head, Centre for Science Research and Statistics, 11, Tverskaya str., Moscow K-9, GSP-9, 101999 , Russia
e-mail: gasl@csrs.ru

Earle Gow
University Librarian, La Trobe University, Bundoora, Victoria, 3086, Australia
e-mail: e.gow@latrobe.edu.au

Franz Haselbacher
Graz University of Technology, Research and Technology Information Unit (FTI), Schloegelgasse 9, A-8010 Graz, Austria

Jostein Helland Hauge
Bergen University Library, University of Bergen, Parkveien 9, 5020 Bergen , Norway
e-mail: jostein.hauge@ub.uib.no

Heiko Hellweg
Informationszentrum Sozialwissenschaften, Lennéstr. 30, D-53113 Bonn, Germany
e-mail: hellweg@bonn.iz-soz.de

Dirk Hennig
University of Hannover, Computer Center of Lower Saxony, Institute for Computer Networks and Distributed Systems, Schlosswenderstr. 5, 30159 Hannover, Germany
e-mail: hennig@rrzn.uni-hannover.de

Bernd Hermes
Informationszentrum Sozialwissenschaften, Lennéstr. 30, D-53113 Bonn, Germany
e-mail: hermes@bonn.iz-soz.de

Franz Holzer
Graz University of Technology, Research and Technology Information Unit (FTI), Schloegelgasse 9, A-8010 Graz, Austria
e-mail: franz.holzer@tugraz.at

Keith G. Jeffery
IT Department, Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire OX11 UKCLRC-RAL, UK
e-mail: keith.g.jeffery@rl.ac.uk

Kerstin Kleese-van Dam
CLRC, Rutherford Appleton Laboratory, Didcot, OX11 0QX, UK
e-mail: k.kleese@rl.ac.uk

Nieske Iris Koopmans
Netherlands Institute for Scientific Information Services (NIWI), Joan Muyskenweg 25, 1096 CJ Amsterdam, The Netherlands
e-mail: iris.koopmans@niwi.knaw.nl

Jürgen Krause
Social Science Information Centre (IZ Bonn), Lennéstr. 30, D-53113 Bonn, Germany
e-mail: krause@bonn.iz-soz.de

Wolfgang Küchlin
Wilhelm-Schickard-Institute for Computer Science, University of Tübingen, Sand 13 D-72076 Tübingen, Germany
e-mail: kuechlin@informatik.uni-tuebingen.de

Harrie Lalieu
NIWI-KNAW, P.O. Box 95110, 1090 HC Amsterdam, The Netherlands
e-mail: harrie.lalieu@niwi.knaw.nl

Benedetto Lepori
Facoltà di Scienze della Comunicazione, Università della Svizzera italiana, via G. Buffi 13
6900 Lugano, Switzerland
e-mail: blepori@unisi.ch

Edward Lim
Assistant Project Director AARLIN, Library, La Trobe University, Bundoora, 3086,
Victoria Australia
e-mail: e.lim@latrobe.edu.au.

Andrei Lopatenko
The University of Manchester, Oxford Rd., Kilburn Building, r.2.112, Manchester M13
9PL, UK
e-mail: alopatenko@cs.man.ac.uk

Brian Matthews
CLRC, Rutherford Appleton Laboratory, Didcot, OX11 0QX, UK
e-mail: b.m.matthews@rl.ac.uk

Barend Mons
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, WOTRO, P.O. Box 93120,
2509 AC Den Haag, The Netherlands
e-mail: barend.mons@inter.nl.net

Viktor Muuli
University of Tartu, Ülikooli 18-304, 50090 Tartu, Estonia
e-mail: Viktor.Muuli@ut.ee

Erlend Øverby
Conduct AS, P.O.B. 805 Sentrum, Biskop Gunnerus gate 2, N-0104 OSLO, Norway
e-mail: erlend.overby@conduct.no

Doreen Parker
University Librarian
Victoria University of Technology, PO Box 14428 MCMC, Melbourne, Victoria, 8001,
Australia
e-mail: doreen.parker@vu.edu.au

Anthony F.J. van Raan
Centre for Science and Technology Studies (CWTS), University of Leiden,
Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands
e-mail: vanraan@cwts.leidenuniv.nl

Jostein Ryssevik
Director of Technology and Development, NESSTAR Ltd.,
Norwegian Social Science Data Services (NSD), Hans Holmboesgt. 22, N-5007 Bergen,
Norway
e-mail: Jostein.Ryssevik@nsd.uib.no

Wolfgang Sander-Beuermann
University of Hannover, Computer Center of Lower Saxony, Institute for Computer
Networks and Distributed Systems, Schlosswenderstr. 5, 30159 Hannover, Germany
e-mail: wsb@rrzn.uni-hannover.de

Ralf-Dieter Schimkat
Wilhelm-Schickard-Institute for Computer Science, University of Tübingen, Sand 13
D-72076 Tübingen, Germany
e-mail: schimkat@informatik.uni-tuebingen.de

Eduard Simons
Nijmegen University, The Netherlands
e-mail: e.simons@dcm.kun.nl / e.simons@icarin.fiuc.org

Sylvia Springer
Otto-von-Guericke-Universität Magdeburg, Technologie-Transfer-Zentrum, PF 4120,
39016 Magdeburg, Germany
e-mail: sylvia.springer@ttz.uni-magdeburg.de

Steffen Staab
Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany
e-mail: staab@aifb.uni-karlsruhe.de

York Sure
Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany
e-mail: sure@aifb.uni-karlsruhe.de

Maximilian Stempfhuber
Informationszentrum Sozialwissenschaften, Lennéstr. 30, D-53113 Bonn, Germany
e-mail: st@bonn.iz-soz.de

Robert Strötgen
Informationszentrum Sozialwissenschaften, Lennéstr. 30, D-53113 Bonn, Germany
e-mail: stroetgen@bonn.iz-soz.de

Rudi Studer
Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany
e-mail: studer@aifb.uni-karlsruhe.de

Taavi Tiirik
Archimedes Foundation, Kompanii 2, Tartu 51007, Estonia
e-mail: taavi@ibs.ee

Richard Tomlin
Community of Science, c/o 24 Bluebell Close, Wylam, Northumberland, NE41 8EU, UK
e-mail: rtomlin@cos.com

Dominik Ulmer
Stab ETH-Rat, ETH Zentrum, CH-8092 Zürich, Switzerland

Jens Vindvad
Riksbibliotektjenesten, P.O.B 8046 Dep, N-0030 OSLO, Norway
e-mail: jens.vinvad@rbt.no

Raphael Volz
Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany
e-mail: volz@aifb.uni-karlsruhe.de

Michael D. Wilson
CLRC, Rutherford Appleton Laboratory, Didcot, OX11 0QX, UK
e-mail: m.d.wilson@rl.ac.uk

Eric H. Zimmerman
Bar-Ilan University, The Research Authority, The Begin Building, Ramat Gan 52900, Israel
e-mail: zimmee@mail.biu.ac.il

Kerstin Zimmermann
Researcher for Information Managment, ftw. Forschungszentrum Telekommunikation Wien, Techgate, Donau-City-Str. 1/3, A-1220 Wien. Austria
e-mail: zimmermann@ftw.at

Editors

Wolfgang Adamczak
University of Kassel, Mönchebergstraße 19, 34109 Kassel, Germany
e-mail: cris2002@uni-kassel.de

Annemarie Nase
Social Science Information Centre, Lennéstr. 30, 53113 Bonn, Germany
e-mail: cris2002@bonn.iz-soz.de

euroCRIS – Current Research Information Systems

euroCRIS is established in Europe to be the internationally recognized point of reference for all matters relating to CRIS: Current Research Information Systems.

euroCRIS serves its membership and the global research community, and advances the field of CRIS through life-long professional development, the sharing of knowledge, and by fostering a sense of true community.

The primary goal of *euroCRIS* is to act as a single forum for all interested individuals and organizations to enter into dialog and resolution of all matters related to the use of information technology in the conduct of all research information system business. *euroCRIS* (through its members) is responsible for maintenance of standards and provision of advice in all matters relating to CRIS. Additionally, *euroCRIS* supports standardized, streamlined information exchange across all aspects of the CRIS lifecycle as follows:

- Apply and maintain internationally accepted standards and guidelines, and develop where necessary (e.g. Common European Research Information Format - CERIF)
- Build and maintain an RTD thesaurus
- Establish a „one-stop shop“ portal/gateway to global CRIS
- Set up a European database of RTD endeavors
- Adhere to principles of best practice
- Include the management of the Code of Good Practice (CGP)
- Provide a forum for exploring and exploiting new and emerging concepts and technologies (incl. data quality, standards, etc…)
- Nurture a community of CRIS practitioners and users
- Addressing the needs of the CRIS user groups

Homepage: http://www.eurocris.org

From digital repository to knowledge management system: the theses and dissertations database of PPGEP/UFSC

José Francisco Salm Jr., Roberto Carlos dos Santos Pacheco, Vinícius Medina Kern
The STELA Group, UFSC/PPGEP, Florianópolis-SC, Brazil

Summary

The Graduate Program in Production Engineering (PPGEP) of the Federal University of Santa Catarina (UFSC) launched, in 1995, its Theses and Dissertations Bank (BTD). From a digital document repository, the BTD was changed into a system that combines bibliometry and informetry features. Besides helping students, faculty, and general public, the BTD assists decision makers. They use it to measure interest and knowledge interchange among PPGEP research areas. This article presents the beginning, main features, theoretical and functional foundations, and perspectives of future development of the BTD project.

1 Introduction

The increment of cooperative organization and the combination of efforts have been suggested as directions for the development of virtual libraries (Mason et al. 2000). This might be achieved through the interoperability of several applications or through the collaboration over a unified, agreed-upon tool.

An important issue in the development of collaborative environments is the consideration of interests and needs of all stakeholders. Virtual library projects should take into account not only librarians' and readers' requirements, but also the interests of people at several managerial levels. While the first are interested primarily in bibliometric indicators and the collection items themselves, the latter are interested in informetric indicators, in a way that regular library users aren't.

Specifically in the graduate school setting, a digital library should be useful not only to students and advisors, but also to managers of the knowledge development process. They are area coordinators, graduate school managers, deans, research group leaders, and others. They are interested, for instance, in the measurement of access frequencies ordered by subject, keyword, advisor, thesis (doctoral level) or dissertation (mastering level), and publication year. This kind of information is important in their decision making process.

The Theses and Dissertations Bank (BTD) of the Graduate Program in Production Engineering (PPGEP) of the Federal University of Santa Catarina (UFSC), Brazil, was conceived taking these considerations into account. Since its first release, in 1995, around 3,000 theses and dissertations approved at PPGEP since 1970 have been made available, more than 900 in full text. It went online in December of 2000 (UFSC/PPGEP 2000), recently reaching 100,000 hits, after one year and a half. Several options of access statistics are offered in addition to access to content and catalog information. These statistics have been used PPGEP managers in their decision making.

This paper reports the development, use, and future of the BTD. Public funding of university digital libraries is pondered, together with an account of information requirements of various stakeholders. BTD's starting points, features, and achievements are described. Current and future developments are discussed.

2 Digital libraries and public funding

Academic workforce qualification depends on public funding to a large extent anywhere in the world. In Brazil, where this work is conducted, the private sector hardly invests in human resources at the academic level. Investment of the scarce public funds must be done wisely. The result of this investment is hard to measure with precision – the wealth generated is mainly in the graduates' minds – but theses and dissertations are concrete results.

Recently Lawrence (2001b) demonstrated that articles that are available online are more cited than those that aren't. The availability of academic works in digital libraries benefits researchers. This benefit, however, depends on the level of availability and ease of access. Well-equipped and content-rich digital libraries reach their full potential only if they offer access at affordable rates.

Building good digital libraries is a means to help improving graduate students' productivity. Better than traditional libraries, they offer ubiquity, flexibility, and advanced search facilities. This is an important investment of public funds, in addition to scholarships and funding for projects. It seems advisable, therefore, to make public (or give back to the taxpayers) the results of their investment – for instance, offering free access to the theses and dissertations produced.

Baggio (2000) states that if the accumulated knowledge is not shared with the society, the abyss that separates rich and poor is deepened. By offering its production to the public through a well-equipped digital library, Brazilian graduate programs take the opportunity to expose their quality and to promote their integration with industry and with society in general. The next section reports the history of BTD, a free digital library and information system maintained by a public university.

3 The history of BTD

This section describes the BTD since its beginning. The building of the Brazilian chemical engineering theses and dissertations bank in 1999 represented a major technological update because it integrated aspects of information integration between an accreditation agency and the digital library. The BTD, version 2001, introduced new services in an online implementation.

3.1 Beginning

The PPGEP strategic planning established in 1985 a ten-year goal: the organization of a distance education program that should allow students to access content without geographic restrictions, based on information technology and telecommunications. In July, 1995, the Laboratory of Distance Education (LED) of PPGEP installed its videoconference infrastructure to be used in a mastering program offered in partnership with other universities in Santa Catarina state (with support from Funcitec, a state foundation for science and technology, and from CAPES, the coordination for the improvement of academic personnel – a federal funding and accreditation agency).

The need for digital libraries and integrated academic and administrative platforms was recognized since LED's inception. BTD was implemented also in 1995, roughly a year after the appearance of Netscape® e Mosaic®. The available Internet infrastructure used a 64 Kbps link. An agreement between PPGEP and the state government allowed for an increment to 2 Mbps.

A team was formed at the Stela Group (a development laboratory at PPGEP) to produce HTML versions of theses and dissertations, usually delivered as .doc files (particular of text editors such as Word® and StarOffice®). The first dissertation was published in digital form in the beginning of 1996. Other 357 works were uploaded to BTD until 1999. By that time several theses and dissertations libraries had gone online, for instance the ones by the Massachusetts Insti-

tute of Technology (MIT 2002), ProQuest-UMI (UMI 2002), Virginia Tech (2002) and IBICT (2002).

BTD was created to give transparency to LED's (and PPGEP's as a whole) production of theses and dissertations. LED's initiatives and activities have inspired universities throughout the country (Wahrhaftig; Ferraza & Raupp 2001).

3.2 The Chemical Engineering bank

The Stela Group also developed, in 1999, the first Brazilian national data bank of theses and dissertations in Chemical Engineering. This bank allows for the capture of data and metadata about theses and dissertations.

The capability of dealing with content, catalog data and metadata is consistent with the requirements of CAPES, providing for information transmission in the context of graduate program accreditation. This approach to information interchange in the Chemical Engineering bank set the foundation for the current BTD implementation.

3.3 BTD version 2001

The theses and dissertations data bank was integrated with the PPGEP academic management platform. The files with the theses and dissertations and their records in the platform were linked. A BTD website was designed, going online in December of 2000.

The main users of BTD are students and faculty. However, since BTD is free and available over the web, it has been useful to strength PPGEP's connection with industry. BTD has received about 23,000 hits in its first six months online, about 60,000 within a year, completing 100,000 in June of 2002, after 18 months in service.

BTD has the records of all theses and dissertations defended at PPGEP since its creation in 1970. There are more than 3,000 works with the following distribution by mid-2002: 74% mastering dissertations, 13% doctoral theses, 13% qualifying monographs. More than 1000 works are available as full text, in HTML or PDF. The access count for the most accessed thesis is about 2,000. The website hit counter is over 100,000. Figure 1 shows the website interface. The next section describes features and innovations of BTD that allowed for such expressive numbers.

Figure 1: The BTD website (UFSC/PPGEP 2000)

4 Features and innovations of BTD

The BTD website is available in Portuguese, organized in six resource areas: general information, recent and coming defenses, access to theses and dissertations, BTD statistics, related links, and access rankings. The features of these resources are detailed next.

4.1 General information

This section describes the objectives of BTD and presents its information sources. The main features are briefly commented, together with some historical background on the digitalization process undertaken by PPGEP's Media and Knowledge Lab (LMC) from 1995 to 1999. From 2000 on, PPGEP requests students' to furnish digital copies in popular formats, to be translated into Adobe®'s portable document format (PDF).

4.2 Recent and coming defenses

This section of the website reports the scheduling of defenses to occur shortly and lists defenses occurred in the last four months. It allows interested parties to search defenses by area or graduate level (doctorate—theses and qualifying monographs, mastering—dissertations) and to contact authors, advisors, and internal (PPGEP's) committee members. It is possible, from each defense record, to find statistics about the advisor, and to access his advisees' theses and dissertations, which also can be done through the specific "Theses and dissertations" link, as described next.

4.3 Access to theses and dissertations

This is the most popular section of the website. It allows for searching theses and dissertations with search filters: by title (complete or incomplete), keywords, author, advisor, committee member, advising tutor (a doctoral student who co-advises a mastering candidate), program area, year, and level (mastering, qualifying, or doctorate). Filters can be combined.

For each record, it is possible to see the hit counting, and to access the advisor's curriculum and statistics of defenses with counting by year and average completion time. The advisor's curriculum is in the Lattes Platform (CNPq 2002b), the Brazilian online platform of information systems and web portals on science and technology.

The steps to access each thesis or dissertation are:
1. Search configuration – instantiate terms and criteria for the search
2. Results – read the match list
3. Choice – click on the thesis or dissertation link
4. Access – read the record and access (where available) the full content

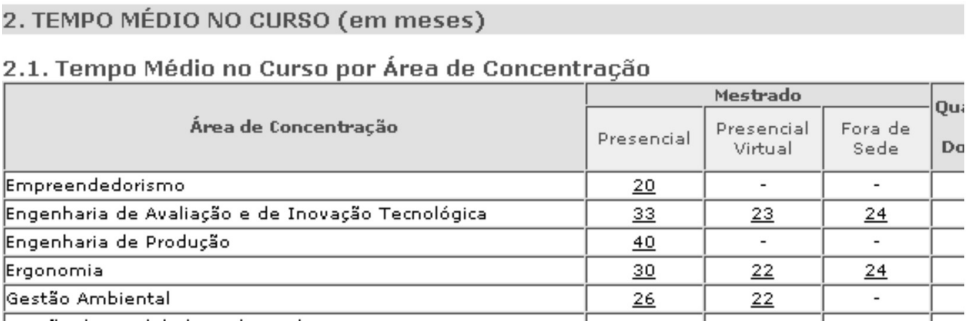
This section is sought mainly by academic users, but also by the community at large. A search for "freight centers", for instance, points to 6 monographs defended between 1993 and 1998. This search allowed PPGEP to demonstrate, online and in real time, its documented know-how in the subject. The demonstration was made during a visit from the president of a telecommunications company who was interested in academic works related to freight center customer service. This illustrates the role of knowledge management tool played by BTD.

4.4 BTD statistics

This link was conceived to assist decision makers and researchers who are interested in measurements of PPGEP's production of theses and dissertations. Figure 2 illustrates one of the statistic tables that can be obtained from BTD: mean completion time for graduates, in months, according

to each program area, type of degree, and program category (presence, distance, or out-of-campus program).

The section displays general statistics on number of defenses, number of full text works, and mean time for completion according to a variety of parameters: area, type of degree, program category, and year of defense. It is also possible to custom-fit statistics using filters for program area, advisor, entry year, and graduation year.



2. TEMPO MÉDIO NO CURSO (em meses)

2.1. Tempo Médio no Curso por Área de Concentração

Área de Concentração	Mestrado			Qu
	Presencial	Presencial Virtual	Fora de Sede	Do
Empreendedorismo	20	-	-	
Engenharia de Avaliação e de Inovação Tecnológica	33	23	24	
Engenharia de Produção	40	-	-	
Ergonomia	30	22	24	
Gestão Ambiental	26	22	-	

Figure 2: BTD's statistics (UFSC/PPGEP 2000)

The generation of statistics is an important contribution of BTD to graduate education management. The detection of a tendency to reduce mean completion time and, especially, the comparison of mean completion times among the three categories of mastering program (presence, distance, and out-of-campus) allow for an appraisal of the results of advising efforts.

The distance (videoconference) mastering program shows the smaller mean completion time among the three categories. Ongoing studies try to assess the reasons for this. It is speculated that two main factors influence the shortest completion time for the distance program: the identity of candidates and topics (usually focused on their companies' problems), and the advising system (in which advisors and tutors teamwork to keep candidates on track).

4.5 Related links

Related links leads the user to other digital libraries, as well as to articles and news. News and articles are presently Portuguese-only. Digital libraries linked include IBICT (2002), MIT (2002), Virginia Tech (2002), ProQuest (UMI 2002), Physics at Unicamp (IFGW 2002), and DiTeD (2002).

This section of the website also offers documents and information on the bureaucratic aspects of graduation. In addition, the Brazilian copyright law (MCT 1998) is presented.

4.6 Access rankings

This section displays access rankings by author, advisor, and area. It is one of the advanced features of BTD, created to allow for the measurement of interest levels.

The ranking by author shows the 100 most accessed works. The ranking by advisor positions them according to the average hit counter for their advisees' available works. It also links each advisor with his curriculum vitae in the Lattes Platform (CNPq 2002b) and statistics. The ranking by area simply ranks program areas according to the number of hits.

5 Present and future

BTD has already made an impact on how PPGEP's candidates research, and on the visibility of graduate monographs. The hit counting suggests the vigor of the introduction of new technologies and their effect on the working style of researchers. The statistics and rankings endow the system with valuable tools for decision support.

New developments are planned for BTD, especially regarding scientific journalism and knowledge management. BTD is planned to offer interviews with graduates and their advisors and committee, and alumni track keeping – forming a talent bank and aiming at the interaction with alumni and the organizations they are associated with.

The next developments include also the verification of methodology, advanced search features (full text), the building of informetrics indexes, an organizational knowledge management module (dealing with the building of a profile of each advisor's areas of interest and frequent topics of the advisees), artificial intelligence tools in theses and dissertations auditing (for instance, for detecting plagiarism), and the application of intelligent auditing systems (for content comparison). Both methodology verification and the application of intelligent auditing systems are current topics of research at PPGEP.

This suite of functionalities, once implemented, tends to accelerate the pace of change that is apparently just beginning in the academic world. The university is changing, with evident impact on libraries. However, we consider that the achievements are faint if we think of the potential of technology usage in libraries. Digital libraries still stumble on inadequate information and software integration. Private information providers offer solutions that are, too, difficult to integrate with the rest of the library environment.

The integration of digital libraries depends on the building of standards that allow for the interchange of information and interoperability of applications. Caplan (2000) observes that the development of standards for digital libraries is hard because of the lack of leadership for a collaboration effort.

Improvements in access to scientific literature have already changed the entire scientific process (Lawrence 2001). Bollacker et al. (2000) maintain that automatic tools to support research will be increasingly important in the future. Nonetheless, there is still much to advance in the development of digital libraries.

Cameron (1997) and Brüggemann-Klein et al. (1997) propose that authors submit, together with their papers, an annotated for database insertion, therefore making it easy to access catalog information on articles. This is only one example of action that requires general collaboration for its implementation. The implications of such actions, however, can have a great impact on digital libraries.

Considering collaboration efforts, Brazil has to offer as example the LMPL (CNPq 2002), The Lattes Platform Markup Language. It is a XML-based ontology designed to allow for the integration of science and technology information systems. LMPL allows for the building of dynamic links among institutional portals – for instance, linking a thesis in a theses portal and its author's curriculum in a funding agency's portal, thus reinforcing the character of public access and transparency.

A similar collaboration in the digital library arena is advisable. A recent initiative led by the Brazilian Institute for Information on Science and Technology (IBICT) works on a national definition for theses and dissertations metadata, based on the Dublin Core metadata initiative (DCMI 2002). Such definition can be used as instrument for harvesting mechanisms for theses and dissertations.

6 Concluding remarks

This article presented the experience of PPGEP/UFSC in the building of its Theses and Dissertations Bank (BTD), emphasizing its character of transparency and knowledge management. The role of digital libraries in public universities was discussed. The history of BTD was reported, together with a description its functions. Perspectives for future development were conjectured, and the future of the digital library was briefly discussed.

The technology and the methodology for the building of BTD have been part of the continued effort for building the Lattes Platform (CNPq 2002b), the Brazilian science and information platform. The Lattes Platform was conceived and operated to support funding, planning, and evaluation of science and technology. Several agencies have worked to integrate science and technology applications into the Lattes Platform, and BTD also has this aim.

The result of these efforts will be the standardization of metadata, allowing for the establishment of links among several institutional web portals. In the future, we expect to have access to theses, dissertations or articles from within a researcher's curriculum Lattes, and from these to the author's curriculum, thus strengthening the character of transparency and public access to public-funded scientific production.

In terms of international trends, the Brazilian efforts are in tune with recent studies on the importance of information visibility (Lawrence 2001b). By making public the production of theses and dissertations, BTD and other Brazilian initiatives contribute to overcome the vicious circle that hides the national scientific production.

7 Acknowledgments

Several groups at PPGEP deserve credit for the realization of BTD. The Media and Knowledge Lab (LMC) and the Applied Intelligence Lab (LIA) were initially responsible for document digitalization. Presently, BTD's growth depends on graduates and advisors who produce and submit theses and dissertations. The Stela Group is in charge of BTD's management and development.

8 References

Baggio, R. (2000): A sociedade da informação e a infoexclusão. *Ciência da Informação*, Vol. 29, No. 2, pp. 16-21.

Bollacker, K. D.; Lawrence, S.; Giles, C. L. (2000): Discovering Relevant Scientific Literature on The Web. *IEEE Intelligent Systems*, Vol. 15, No. 2, pp. 42–47.

Brüggemann-Klein, A.; Klein, R.; and Landgraf, B. (1997): *BibRelEx*: Exploring Bibliographic Databases by Visualization of Annotated Contents-Based Relations.
[Available online at: <http://citeseer.nj.nec.com/352222.html>]

Cameron, R. D. (1997): A Universal Citation Database as a Catalyst for Reform in Scholarly Communication. *First Monday*, Vol. 2, No. 4.
[Available online at: <http://firstmonday.org/issues/issue2 4/cameron/index.html>]

Caplan, P. (2000): Oh What a Tangled Web We Weave: Opportunities and Challenges for Standards Development in the Digital Library Arena. *First Monday*, Vol. 5, No. 6.
[Available online at: <http://firstmonday.org/issues/issue5_6/caplan/index.html>]

CNPq (2002): *Comunidade LMPL - Linguagem de Marcação da Plataforma Lattes*. Brazilian National Research Council (CNPq).
[Available online at: < http://www.stela.ufsc.br/lmpl/>, access in 2002.06.21]

CNPq (2002): *Comunidade LMPL - Linguagem de Marcação da Plataforma Lattes*. Brazilian National Research Council (CNPq). [Available online at: <http://www.stela.ufsc.br/lmpl/>, access in 2002.06.21]

DCMI (2002): *Dublin Core Metadata Initiative*.
[Available online at: <http://www.dublincore.org/>, access in 2002.05.03]

DiTeD (2002): *Dissertações e Teses Digitais*. Biblioteca Nacional, Portugal.
[Available online at: < http://dited.bn.pt/>, access in 2002.06.21]

IBICT (Brazilian Institute for Information on Science and Technology) (2002): *Biblioteca Digital Brasileira*. [Available online at: <http://www.ibict.br/bdb/inicio.htm>, access in 2002.06.21].

IFGW (2002): *Teses Digitais – Instituto de Física Gleb Wataghin*. Unicamp (State University of Campinas-SP, Brazil).
[Available online at: < http://www.ifi.unicamp.br/ccjdr/teses/>, access in 2002.06.21]

Lawrence, S. (2001): Access to Scientific Literature, *The Nature Yearbook of Science and Technology*, edited by Declan Butler, Macmillan, London, England, p. 86–88.

Lawrence, S. (2001b): Online or Invisible? *Nature*, Vol. 411, No. 6837.

Mason, J.; Mitchell, S.; Mooney, M.; Reasoner, L.; Rodriguez, C. (2000): Infomine: Promising Directions in Virtual Library Development. *First Monday*, Vol. 5, No. 6.
[Available online at: <http://firstmonday.org/issues/issue5_6/mason/index.html>]

MCT (1998): Lei Nº 9.610, de 19 de Fevereiro de 1998. *Diário Oficial da União*. MCT (Ministry of Science and Technology), Brasília-DF, Brazil.
[Available online at: <http://www.mct.gov.br/legis/leis/9610_98.htm>, access in 2002.06.21]

MIT (2002): *Digital Library of MIT Theses*.
[Available online at: <http://theses.mit.edu/>, access in 2002.05.03]

UFSC/PPGEP (2000): *Banco de Teses e Dissertações do PPGEP*. Florianópolis-SC, Brazil: Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina. Released in December, 2000. [Available online at: <http://teses.eps.ufsc.br/>, access in 2002.06.21]

UMI (2002): *ProQuest Digital Dissertations*.
[Available online at: <http://wwwlib.umi.com/dissertations/>, access in 2002.05.03]

Virginia Tech (2002): *Digital Library and Archives*.
[Available online at: <http://scholar.lib.vt.edu/theses/>, access in 2002.05.03]

Wahrhaftig, R.; Ferraza, A. M.; & Raupp, M. (2001): *Portas Abertas para a Educação Superior*. Universidade Eletrônica do Paraná.

9 Contact Information

Roberto Carlos dos Santos Pacheco, Vinícius Medina Kern, José Francisco Salm Jr.

Grupo STELA
Programa de Pós-Graduação em Engenharia de Produção (PPGEP)
Universidade Federal de Santa Catarina (UFSC)
Rua Lauro Linhares 2123, torre B, 201-205
88036-002 Florianópolis-SC Brasil

E-mail: {pacheco, kern, salm}@eps.ufsc.br
http://www.eps.ufsc.br/~pacheco, http://www.eps.ufsc.br/~kern, http://www.stela.ufsc.br

A Distributed Portal for Physics

Thomas Severiens
Institute for Science Networking, Oldenburg University, Germany

Many subject specific portals were built during the last year. Most of these are simple user-interfaces to databases of subject specific information added with several lists of links.

This centralised type of portal often looks fine with its consistent facing but is hard to keep up to date and high priced to maintain. Users expect a service be maintained and available 24 hours, 365 days for at least 10 years and this all free of charge. On the one hand, it seams to be impossible to set up a service matching all this demands, on the other hand, many institutions offer information and services which could be parts of a portal, which are maintained frequently and paid by public via these institutions.

The idea is, to collect the existing information and present it in a structured and consistent way. This idea matches in an excellent way with the way knowledge is produced in Physics. Physicists work all over the world often on different continents on the same topic, knowing each others work only from their publications, conferences and online-communication. Information in Physics is published in quite different ways, by journal articles, which can be reviewed, sometimes by peer, or pre-prints. Many information is available in non-textual genres like software sources or datasets or mathematical formula.

Distributed Portals make use of the existing information on the web. In the early days of the web, the very popular link-lists where a kind of portal, linking to (all) pages with information on the specific topic. Indeed, these link lists had many properties of modern portals, offering information in a structured and selected way. But they did not offer the information under a common layout (desktop) and did not offer user-specific views onto the information.

Modern distributed portals combine the advantages of centralised portals (high information structure, common layout, easy navigation through all the information) with the possibilities of distributed portals (up to date information, low budget implementation, good knowledge coverage).

On the technical side, a distributed portal consists of several modules:

1.) A collecting module, which collects the information form the distributed sources. This module is often a simple collection script allowing http-requests to be automated by a scheduling system.
2.) A content extractor, which extracts the content from the downloaded sides. Different ways to implement this are in use in different implementation. The easiest and most used method is to include comment lines into the web-sites to tell the extractor, where the content starts and ends. Also methods basing on artificial knowledge are used sometimes, especially when extracting information form a very huge number of sites in a system with low organisational background.
3.) The portal itself is the same structural part of program, like in every centralised portal.
4.) A mirror system, which allows to mirror the hole portal onto several servers, to enhance the chance of availability of service.

Every portal needs a search-engine, which often is only a simple php-interface to the database containing all the data of the portal, even for distributed portals. The DFN-sponsored project SINN www.isn-oldenburg.de/projects/SINN/ develops and tests methods on distributing even

the search implementation. They use methods developed by W3C's XML-query working group to implement prototypes of total distributed search-engines. To explain the idea, let's use an example: Every Physics department in Germany has a web-server. Most of the web-servers contain a search-engine using one of the approximately ten popular software packages. To implement a search engine, which covers all information on all of the German web-servers, it is a nice idea to use the collected information of all the local search-engines. For this, one needs a common interface, which allows asking queries and collecting answers as well as asking for more administrative information like the age of the offered information, who is responsible for the server, if the information may be offered in other search-engines, information needed to rank the combined result lists and so on. XML-query among other things allows to implement such a protocol in a standardised and open way.

The biggest problem in implementing distributed portals and search-engines is not the technical, but the organisational implementation. One has to bring partners together, which are spread all over the world with all their different cultural background. The partners have to work together in a reliable and continuos way, often without any financial support for all their local work.

The Institute for Science Networking www.isn-oldenburg.de acts as technical nucleus for several distributed systems. The most popular and oldest is the PhysNet www.phyics-network.org a collection of information and link lists. PhysNet has developed from a simple link-collection in 1994 to a portal, which content is maintained in 8 countries on 4 continents. The complete portal PhysNet is mirrored every night onto 10 servers spread over the world. Even the mirroring runs fully automated and installing of a mirror costs only 10 minutes of time for an administrator, it is hard to find computer centres, which are willing to run a mirror, because they do not really understand the idea of mirroring, which is to guarantee availability of service and reduce net-traffic, which is of special importance for the users from the developed countries.

Many national Physical Societies are member of PhysNet, because the EPS developed a charter for the politics of service. This charter was a break-through in acceptance of the service by the users and especially the officials. To write down to basic and common ideas is a good idea, when developing a distributed portal.

In 2000 the German Physical Society DPG asked their members to contribute in building a common portal to Physics in Germany. As a result of this, several new portals were built, some of them are centralised like www.pro-physik.de which is a portal of Wiley VCH the publisher of the Physics-Journal, the journal of the DPG.

The Special Interest Group on Information and Communication the DPG (AKI) www.aki-dpg.de also built a portal "Fachwelt-Physik" www.fachportal-physik.de which is a distributed portal. Members are several scientific libraries, the FIZ Karlsruhe, all Physics Department in Germany and the ISN, which developed the software and collected a starting amount of information and links. The information offered is maintained at many distributed partners. It is collected together onto the server at the ISN and offered into a portal with several views. The portal is bilingual German and English, it offers an optimised view for browsers based on Mozilla (Internet-Explorer, Netscape, Opera, KDE-Browser, …), for robots collecting information for search engines, for users who want to print out the pages and for handicapped users, who ask for a text-only version.

The development and implementation of interface for distributed search-engines will be one of the focuses for future activities. Development of new services as well as continuos operation and maintenance of existing portals will be further emphases.

Contact Information

e-mail: severien@uni-oldenburg.de

ISBN 3-933146-844