Designing the Grey Grid for Information Society

TENTH INTERNATIONAL CONFERENCE ON GREY LITERATURE

Science Park Amsterdam, Netherlands December 8 - 9, 2008

Sponsors:

The needs and demands of Information Society are in constant state of change and flux. Information overload, information loss, information-on-demand are among just a few of the many factors confronting information professionals, practitioners, and net-users on a daily basis. To a great extent, grey literature is the cause of all this. For the past two decades grey literature has grown exponentially in relation to commerci[al] [c]ommunity realizes t[he] [i]nternational Conferenc[e] [h]as been resolved, solutions [a]t scale and magnitud[e] [c]hallenges to grey liter[ature] [e]ven newer challenges [t]o date all. The conference [title] 'Designing the Grey Grid for Information Society' invokes an infrastructure, which must take into account social, political, and organizational factors. For these also impact system-to-system performance when dealing with the scale and diversity of information, data, document types, collections, and subject areas linked to grey literature. As such, interoperability becomes de facto a

CONFERENCE PROCEEDINGS

GL 10

Program Committee

INIST	Institute for Scientific and Technical Information, France
BL	British Library, United Kingdom
CNR	National Research Council, Italy
EuroCRIS	Current Research Information Systems, Norway
GreyNet	Grey Literature Network Service, Netherlands
NYAM	New York Academy of Medicine, USA
OSTI	Office of Scientific and Technical Information, USA
UCI	University of California, Irvine, USA
UNI-LJ	University of Ljubljana, Slovenia

GL10 Program and Conference Bureau

TextRelease

Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands
www.textrelease.com ▪ **conference@textrelease.com**
Tel/Fax +31 (0) 20-331.2420

CIP

GL10 Conference Proceedings
Tenth International Conference on Grey Literature : Designing the Grey Grid for Information Society, 8-9 December 2008, Science Park Amsterdam, The Netherlands / ed. by Dominic J. Farace and Jerry Frantzen ; GreyNet, Grey Literature Network Service. - Amsterdam : TextRelease, February 2009. – 143 p. ; 30 cm. – Author Index. – (GL-Conference series, ISSN 1386-2316 ; No. 10)

The British Library, INIST-CNRS, NYAM, and the University of Ljubljana are corporate authors and associate members of GreyNet. These conference proceedings contain the full text of some fifteen papers presented during the two days of Plenary and Poster Sessions. The papers appear in the same order as in the conference program. Included is a List of Participating Organizations, and Sponsored Advertisements.

ISBN 978-90-77484-11-1

Foreword

Designing the Grey Grid for Information Society

The needs and demands of Information Society are in constant state of change and flux. Information overload, information loss, information-on-demand are among just a few of the many factors confronting information professionals, practitioners, and net-users on a daily basis.

To a great extent, grey literature is the cause of all this. For the past two decades grey literature has grown exponentially in relation to commercially published literature. The grey literature community realizes that while the challenges faced at the First International Conference on Grey Literature in 1993 may not have all been resolved, solutions today lay in a whole new order, on yet another scale and magnitude than ever before. GL10 sought to address the challenges to grey literature that still remain, while dealing with even newer challenges and an infrastructure that can effectively integrate all. The conference title 'Designing the Grey Grid for Information Society' invokes an infrastructure, which must take into account social, political, and organizational factors. For these also impact system-to-system performance when dealing with the scale and diversity of information, data, document types, collections, and subject areas linked to grey literature. As such, interoperability becomes de facto a requirement in the design of the grey grid i.e. an infrastructure that can model and withstand the test of an ever changing Information Society.

On behalf of the Conference Sponsors, the Program Committee and Chairpersons, I would like to thank the authors and co-authors for their content contributions to these proceedings. Likewise, I welcome those reading these conference proceedings to voice their comments and/or recommendations either directly to the authors or to GreyNet, Grey Literature Network Service.

Finally, I would like to bring to your attention the recent GL11 Call for Papers, the Eleventh International Conference on Grey Literature, which will be held in The Library of Congress on December 14-15, 2009.

Dr. Dominic J. Farace Amsterdam,
Program and Conference Director February 2009

GL10 Conference Sponsors

BL, United Kingdom
The British Library

INIST-CNRS, France
Institut de l'Information Scientifique et Technique; Centre National de Recherche Scientifique

EBSCO, USA
EBSCO Information Services

City of Amsterdam, The Netherlands
Co-sponsor to the GL10 Reception

IIA, USA
Information International Associates, Inc.

NYAM, USA
The New York Academy of Medicine

Swets, The Netherlands
Swets Simplifies

Contents

GL10 Program Committee

Dr. Joachim Schöpfel Chair
University of Lille 3,
France

Elizabeth Newbold
British Library,
United Kingdom

Daniela Luzi
CNR, National Research Council
Italy

Anne Asserson
EuroCRIS, Current Research Information Systems, Norway

Dr. Dominic J. Farace
GreyNet, Grey Literature Network Service,
Netherlands

Latrina Keith
NYAM, The New York Academy of Medicine,
USA

Deborah E. Cutler
OSTI, Office of Scientific and Technical Information, USA

Julia Gelfand
UCI, University of California, Irvine
USA

Dr. Primož Južnic
University of Ljubljana
Slovenia

Eleventh International Conference on Grey Literature

Join the Federal Library and Information Center Committee (FLICC) of the Library of Congress in Washington, DC as host for GL11, December 14 and 15, 2009.

FLICC is an organization of U.S. federal agencies dedicated to cooperation and concerted action within the community of federal libraries and information centers. FLICC and FEDLINK, FLICC's purchasing, training and resource-sharing consortium, achieve better utilization of federal information resources and facilities through promotion of common services, coordination and sharing of available resources and professional development. FLICC is also a forum for discussion of federal library and information policies, programs, and procedures to help inform the Congress, federal agencies, and others concerned with libraries and information centers.

For the latest news on GL11 or FLICC/FEDLINK, visit our Web site at http://www.loc.gov/flicc .

FLICC
FEDLINK

OPENING SESSION

Chair, *Dr. Joachim Schöpfel, University of Lille 3, France*

SESSION ONE – INSTITUTIONAL REPOSITORIES AND GREY LITERATURE

Chair, *Anne Asserson, University of Bergen, Norway*

SESSION TWO – GREY LITERATURE IN BIOMEDICAL COMMUNITIES

Chair, *Elizabeth Newbold, The British Library, United Kingdom*

SESSION THREE – LEGAL ASPECTS, INTELLIGENCE, AND TEXT MINING IN GREY LITERATURE

Chair, *Christiane Stock, INIST-CNRS (France)*

SESSION FOUR – GREY LITERATURE IN RESEARCH

Chair, *Daniela Luzi, CNR-IRPPS (Italy)*

POSTER SESSION

Chair, *Dr. Primož Južnic, University of Ljubljana, Slovenia*

Keynote Address

WorldWideScience.org
Bringing Light to Grey

Brian A. Hitson and Lorrie A. Johnson
Office of Scientific and Technical Information
U.S. Department of Energy

Abstract
WorldWideScience.org1 and its governance structure, the WorldWideScience Alliance2, are putting a brighter spotlight on grey literature. Through this new tool, grey literature is getting broader exposure to audiences all over the world. Improved access to and sharing of research information is the key to accelerating progress and breakthroughs in any field, especially science.

WorldWideScience.org has revolutionized access to "deep" web scientific databases. These nationally- and internationally-sponsored databases are comprised of both grey and conventional literature. Consequently, because grey literature is naturally less familiar (and, hence, less accessible) than conventional literature, it receives a disproportionate benefit in terms of usage through its exposure in WorldWideScience.org.

Before expanding on the mechanics and contents of WorldWideScience vis-à-vis grey literature, it is helpful to characterize what is meant by "grey literature." The term "Grey Literature" can be defined in several ways. Wikipedia3, for example, describes grey literature as "…a body of materials that cannot be found easily through conventional channels such as publishers…" The National Library of Australia4 provides a slight variation: "…information that is not searchable or accessible through conventional search engines or subject directories and is not generally produced by commercial publishing organizations." This description goes further to describe electronic grey literature as constituting the "hidden" or "deep" web. Most laypeople, those outside the professional information community, would think of the color "grey" and may be puzzled as to why a color is used to describe literature. To them, the word "grey" likely brings to mind the Webster5 definition, "an achromatic color between the extremes of black and white."

Traditionally, "white" has been equated with conventional, published literature, but perhaps to better illustrate the point, it could be useful to reverse the "achromatic" color spectrum in this case. The extreme of "black," for example, could be thought of as traditional black ink printed on paper. It consists of words that are very clear and easily accessible to everyone, and makes up the conventional literature such as journals, books, and published proceedings. "White," on the other hand, conveys just the meaning of a blank sheet with no words – simply unrecorded ideas, concepts, and thought. So, then, "grey" is between these two extremes. It includes the kinds of literature that information professionals typically associate with "grey," such as preprints, technical reports, theses and dissertations. More recently, grey literature also includes emerging forms of information such as numeric data, multimedia, recorded academic lectures, and Web 2.0-generated information.

Looking back at the National Library of Australia's definition for a moment, though, it also implies that grey literature comprises the "hidden" or "deep" web. "Grey" is synonymous with "deep" when it comes to the Internet; grey literature, more than any other type, is a body of information that resides in the "deep web" and is not easily found.

To put this concept in context, there is a distinction between the "surface web" and the "deep web." Generally, major search engines such as Google6 and Yahoo!7 are searching web pages on the surface web. These are static web pages that are crawled by Google's automatic crawler, where every word on a page is stored in Google's massive index, and the power and sophistication of Google's systems allows it to return millions of hits in milliseconds.

However, the surface web is not where most scientific literature resides. Instead, it resides in databases that typically have their own search interface, and because the contents of those databases do not sit on a static web page, they are not typically indexed by Google. There are ways for databases to expose their contents for Google's crawlers, but by and large, most database owners do not do so. Therefore, this information is firmly planted in the "deep web," only accessible through the database's own search engine. Most experts estimate that the deep web is hundreds of times larger in terms of content than the surface web. Clearly, this situation calls for a solution, which is offered by WorldWideScience.org.

Unfortunately, the perception among a large percentage of internet users is that if it can not be found by one of the big search engines, it must not exist. So, the first challenge of the deep web is a variation on an old cliché, "what you don't know can hurt you, or at least it could help you." For example, if a person with cancer is only searching the surface web to learn about latest clinical trials, she would be missing substantive and possibly helpful information that may reside in key deep web databases. If a scientist wants to explore the latest developments in photovoltaics, he will be missing the most in-depth information if he limits his searches to the surface web. The key challenge here is that most people are unaware of all the rich resources in the deep web.

Making the unrealistic assumption, however, that the world is replete with people who already know about the multitude of deep web databases relevant to their particular field, there is a second key challenge. This challenge is that searching all of these databases individually, one by one, is not physically possible, or at least it will consume precious time needed for actual research and experimentation. Thus, progress will be thwarted.

These challenges can be overcome through the use of federated search technology – essentially becoming a Google or a Yahoo! for the deep web. In a federated search, a single portal is connected to multiple deep web database search engines. A person enters a search query into a single Google-like search box. The query is then sent simultaneously to the many databases that have been previously identified as relevant to the specialty of the federated search engine. These individual search engines receive the query, perform their own searches, and return results to the federated search engine. The combined results are then ranked using a relevance algorithm (just as Google does) with parameters such as where the query terms appear in the title, how often they appear, and other variables.

A search in a federated search engine is not as fast as Google because live searches of the databases are occurring, but results are generally produced within 30 seconds. Working with other federal science agencies in the United States, the Office of Scientific and Technical Information (OSTI)8, first introduced federated searching with Science.gov9, which searches practically all federal science databases.

Building on the successful model of Science.gov, OSTI then used this technology to develop other federated search tools for more niche communities. ScienceAccelerator.gov10 federates searches of all of OSTI's web systems. The E-Print Network11 specializes in federated searches of e-print databases in the U.S. and several other countries. Science Conference Proceedings12 federates the search of several professional societies' conference databases. Lastly, the Federal R&D Project Summaries13 does the same for databases describing ongoing research projects sponsored by the U.S. government.

Science.gov was a major success as a friendlier way to make government-sponsored science information available to the public, and it won significant praise as a "government-to-citizen" model under the President's e-government agenda. The logical extension of Science.gov as a national federated searching model is that there could be a "Science.world" for a global federated search tool. Nations interested in promoting science globally could allow their individual science databases to be searched by a single portal – something that is not possible with major commercial search engines.

Following the success of Science.gov, Dr. Walter Warnick, OSTI's Director, introduced the concept of a Science.world before the public conference of the International Council for Scientific and Technical Information (ICSTI)14 in June 2006. Dr. Warnick invited other national libraries to help OSTI implement the concept. The British Library15, much to its credit and vision, quickly offered a hand of partnership in this effort. In January 2007, the British Library Chief Executive, Dame Lynne Brindley, and the U.S. Under Secretary for Science in the Department of Energy, Dr. Raymond Orbach, signed a statement of intent to partner in the effort, which also invited other nations to join in this partnership.

Between January and June 2007, several other countries participated in offering their databases to demonstrate that federated search could work on an international level. Recognizing that "dot world" was used to simply draw the analogy to Science.gov, a more descriptive and operable web address was needed, and WorldWideScience.org was chose, with the tag line, "The Global Science Gateway." The first prototype of WorldWideScience.org was demonstrated at the ICSTI public conference in Nancy, France. At that time, twelve databases form ten counties were represented in the searches of

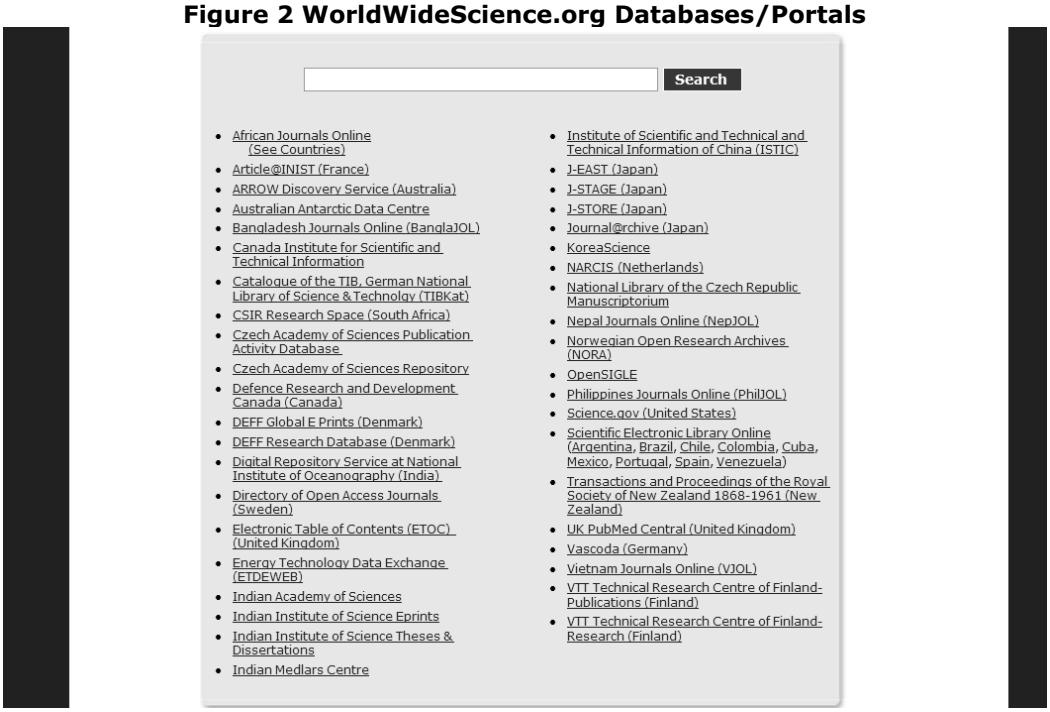
WorldWideScience.org. The successful demonstration of the prototype clearly had the desired effect, as it garnered significant press coverage. In a follow-up ICSTI meeting, it was agreed that ICSTI would play a significant role in helping to form a governance structure for WorldWideScience.org. The formation of the WorldWideScience Alliance was formalized in June 2008 at ICSTI's conference in Seoul. Thirty-eight countries were represented in signing a declaration committing their support to the effort. Completing an international cooperative in a year's time, including terms of reference and governance language, is a reflection of the goodwill and support that this concept received around the world. The Alliance Executive Board is led by Richard Boulderstone of the British Library, who was elected as the Alliance's first Chairperson. A diverse mix of officers from North America, Europe, Asia, and Africa make up the remainder of the Board. The leadership of ICSTI was invaluable in providing a platform to promote this concept to national scientific and technical information officials around the world.

Since the first prototype of twelve databases in ten countries, WorldWideScience.org has now grown to 49 databases in 54 countries (as of December 2008). The scientific content represented in these searches comes from countries accounting for over three-fourths of the world's population. It is estimated, using rough calculations, that these searches cover 375 million pages of science, much of which is obviously grey literature.

Figure 1 WorldWideScience.org

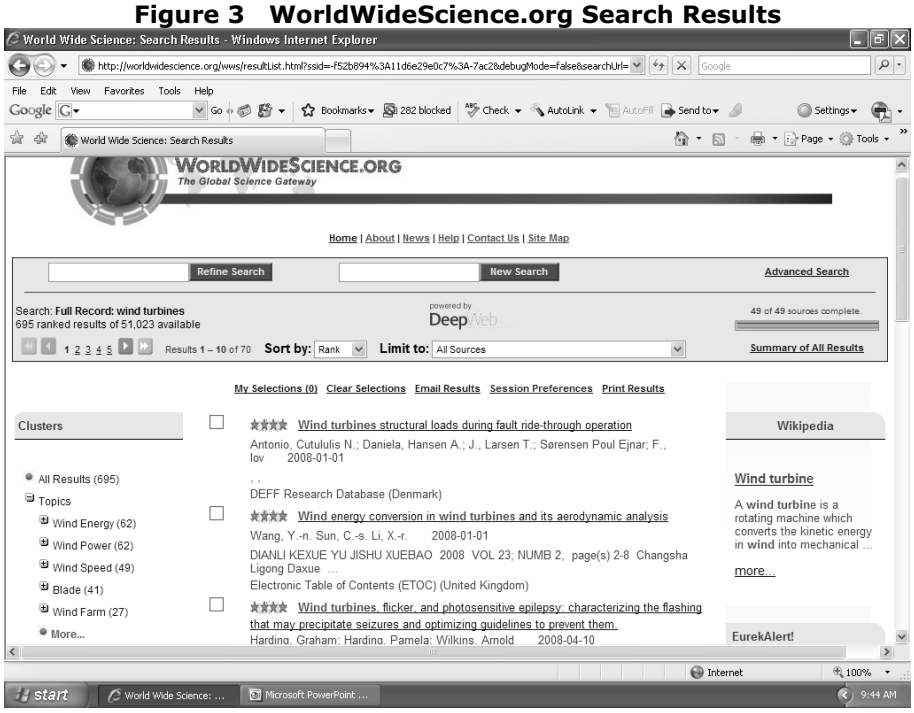
A map of the world (Figure 1) is used to show which countries have databases represented in WorldWideScience.org. At this stage, these are all databases which have some element of national or international sponsorship rather than commercial databases, such as publisher databases. As indicated by the map, sources are covered from practically all of North and South America, Australia, a significant portion of Europe, and major segments of Asia and Africa. Some countries have multiple sources. Japan, for example, has four major databases from the Japan Science and Technology Agency16; India also has four sources. The U.S. source is Science.gov, which is itself a federated search portal of over 30 major databases. In this case, where one federated search engine spawns a search of another federated search engine, it is called nested searching, and it works quite efficiently.

Figure 2 WorldWideScience.org Databases/Portals



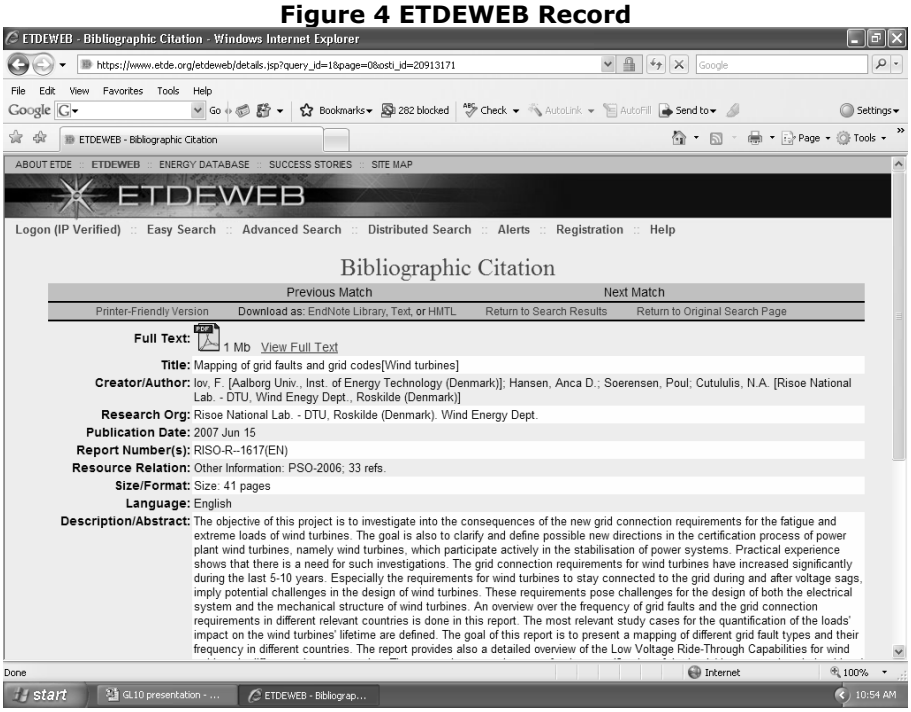
To illustrate with some examples, Figure 3 shows a typical first page of search results from WorldWideScience.org. A search on "wind turbines" has been conducted. All 49 sources were successfully searched, and all together, the sources had over 51,000 records matching this exact phrase. WorldWideScience provides, in this case, the top 695 ranked results. There is a trade-off between showing more results versus the speed of the search; so, typically, the search limits any given source to the top 100 results. A user, if interested in seeing all results, can go to the link "summary of all results" and see which sources have more than 100 results. The user could then go directly to that source for a more in-depth search. Relevance is reflected through the stars (1 through 4) that appear beside each result. Two new enhancements were recently added. On the left side, the user is offered clustering to allow for narrowing results into more refined sets. On the right side, a Wikipedia definition, if one exists, is given for the search term. This is particularly useful for users who simply want to become more familiar with a particular field of science.

Figure 3 WorldWideScience.org Search Results



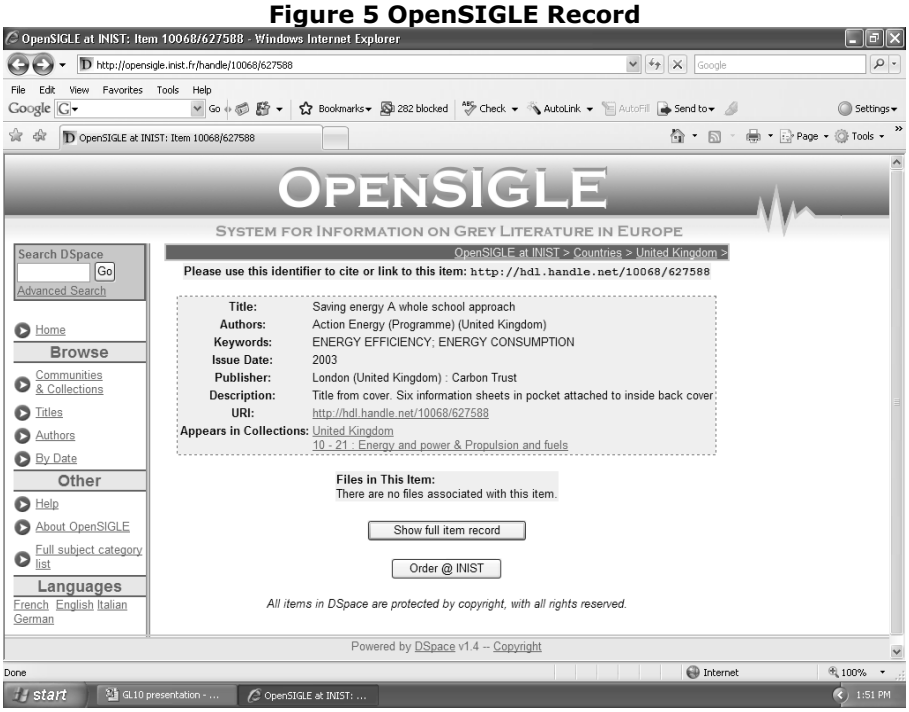
Once the user selects a specific record to view, WorldWideScience.org then takes the user directly to that record within the original database/portal. For example, this record (Figure 4) comes from the Energy Technology Data Exchange (ETDE)17. ETDEWEB is an international database on energy technology governed by an agreement under the auspices of the International Energy Agency. The agreement is comprised of sixteen member countries, who, along with other partners have built a database of 4 million records. As evident by the record, this is clearly a "grey literature" report emanating from Risoe National Laboratory in Denmark. A link to the full text document in PDF format is provided.

Figure 4 ETDEWEB Record



WorldWideScience.org also searches OpenSIGLE18, the system for information on grey literature in Europe. This record (Figure 5) shows how the user could order the full text document from INIST19, the Alliance member from France.

Figure 5 OpenSIGLE Record



Other examples of records include the sub-element of the Norwegian Open Research Archive20, the Bergen University open research archive (Figure 6). The government of South Africa, through its Council for Scientific and Industrial Research21, was one of the earliest supporters of WorldWideScience.org. Its record (Figure 7) also provides the ability to view full text. Working closely with the Alliance member, the International Network for the Availability of Scientific Publications (INASP)22, a number of on-line journal collections from developing countries are available through WorldWideScience.org. These countries include 24 African nations, Bangladesh, Nepal, Philippines, and Viet Nam. A record from Nepal, again providing a link to the full text, is shown in Figure 8. Finally, the last example (Figure 9) shows a record from the Australian open access ARROW23 system, which covers the repository of over half of all universities in Australia. Again, a thesis is a good example of grey literature, and a link to the full text is provided in this case as well.

Figure 6 NORA Record

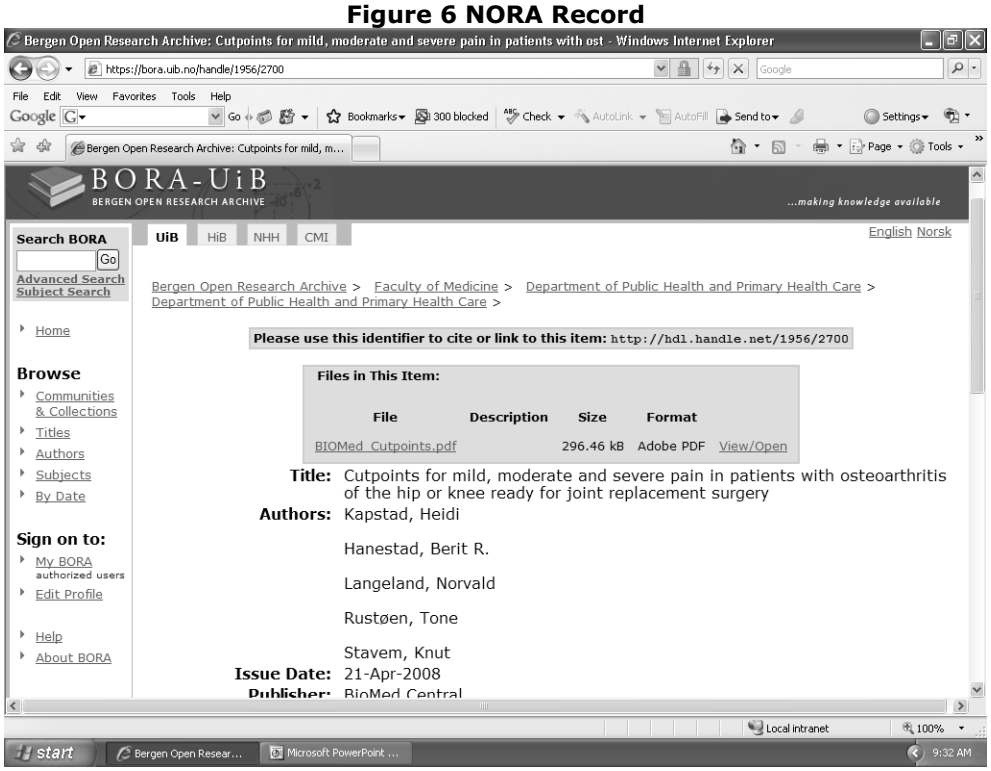


Figure 7 CSIR Record

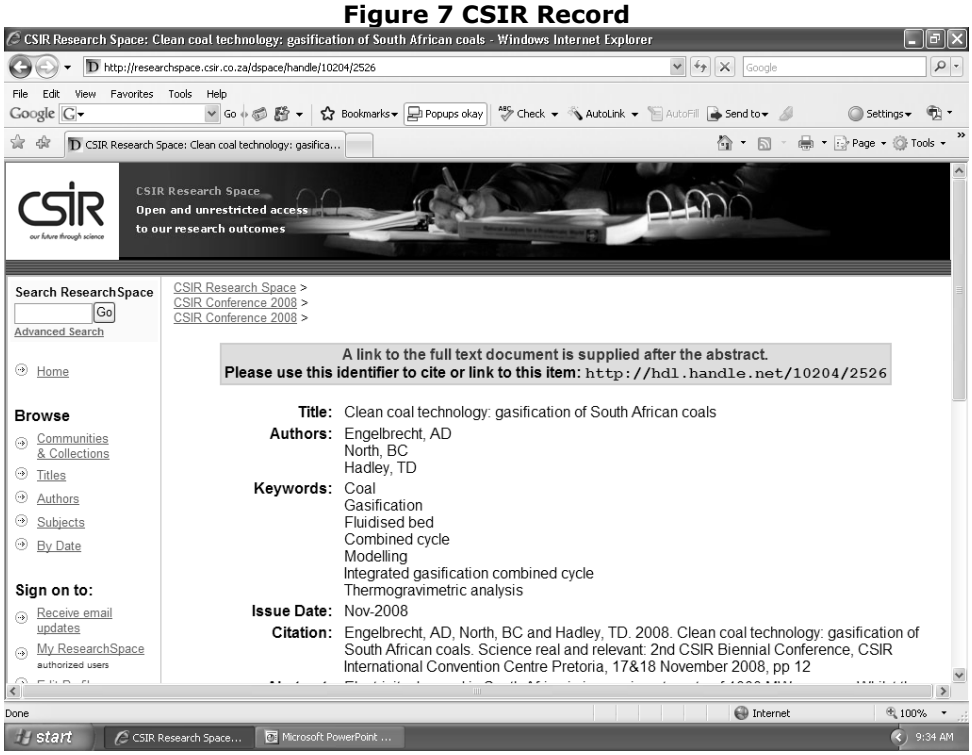


Figure 8 Nepal Journals Online Record

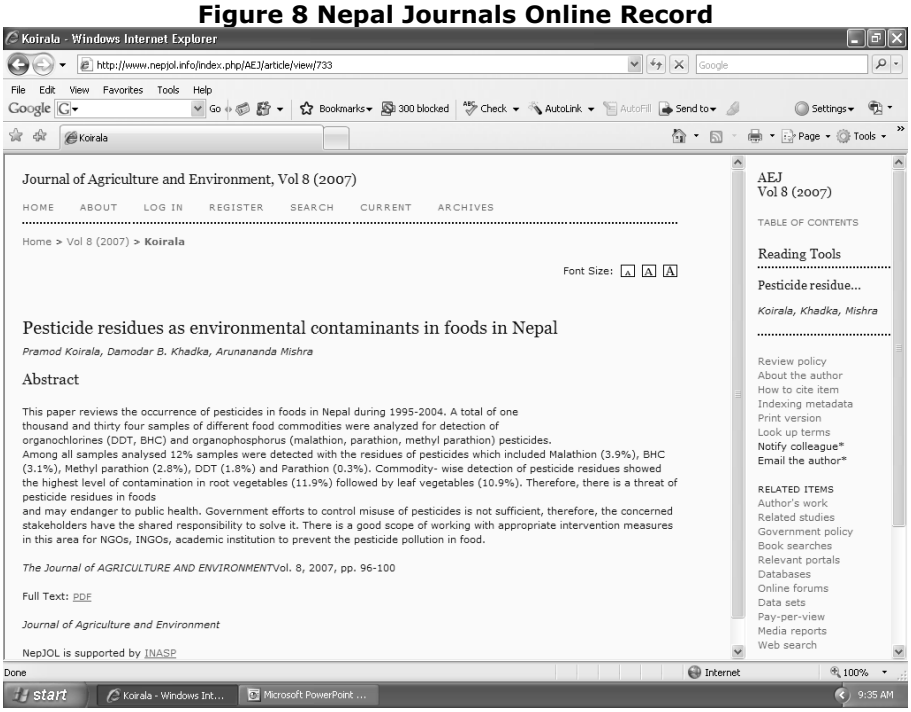
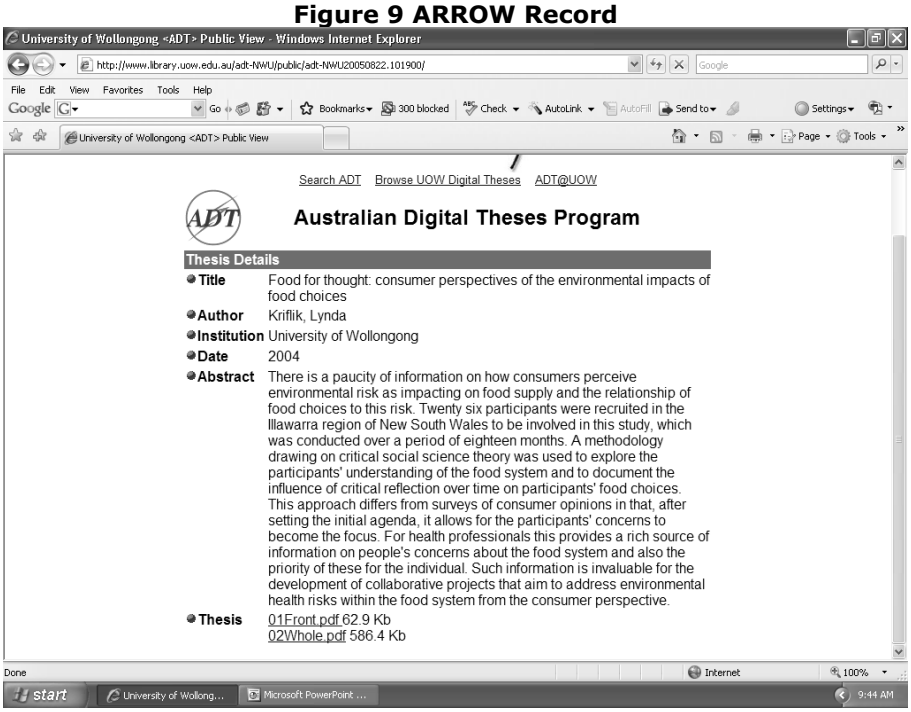


Figure 9 ARROW Record



WorldWideScience.org is continuing to grow consistently both in terms of content and usage. On the content front, the most notable recent addition is an English-language journal source from China. The symbolic significance of opening this access to Chinese science cannot be overstated, and the cooperation of the staff at the Institute of Scientific and Technical Information of China24 was much appreciated.

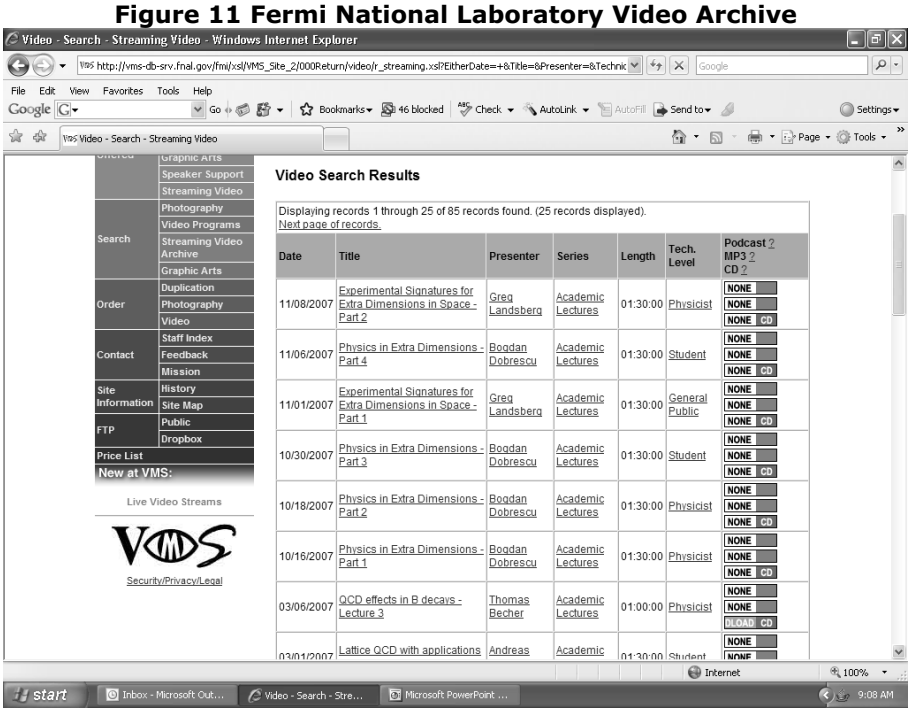
No one really knows ultimately how many sources exist that would make WorldWideScience.org the most comprehensive gateway to nationally- and internationally-sponsored science research, but at around 100 sources, the speed and efficiency of the search engine may start to degrade. A vast amount of computing is involved in processing so many results from so many sources. One of the challenges WorldWideScience.org faces in the future will be overcoming this scalability issue. Strategies have been defined for overcoming this challenge. At a simpler level, one planned enhancement is to offer an alerts service. A user will be able to create a profile and be alerted when any of the WorldWideScience.org sources has added new materials matching that profile.

Another challenge WorldWideScience.org hopes to address in the future is providing access to non-English sources. A few of the Alliance members have experience in this area, particularly INIST in France. The WorldWideScience.org team will begin exploring translation modules that will open access to sources that only exist in a native language, such as the Chinese record in Figure 10.

Figure 10 Chinese Language Record

There are also challenges, not just for WorldWideScience.org, but for all in the grey literature community, with emerging formats such as YouTube videos, podcasts, and other audio and visual sources. There has been a proliferation recently of sites offering access to these types of files. For example, there is a small database of video files of academic lectures from the Fermi National Laboratory25 in the United States (Figure 11), but files such as this are truly in the deep web and are not accessible beyond this interface.

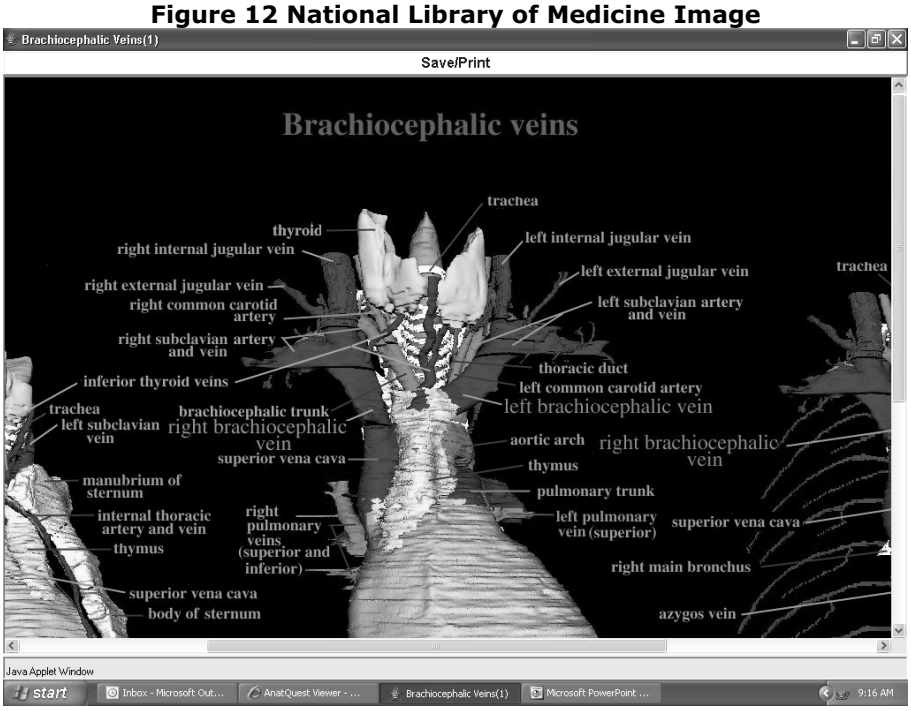
Figure 11 Fermi National Laboratory Video Archive



Beyond sound and video files, there are some fascinating image databases (photographs, drawings, illustrations) that need to be more accessible. This highly-detailed medical illustration (Figure 12) resides in a National Library of Medicine images database26, but the terms on the drawing would not be indexed

by a major search engine, leaving this significant resource potentially under-utilized by the public and medical communities.

Figure 12 National Library of Medicine Image



With the prominence of computational sciences, simulation, and the use of measurements in so many fields, numeric data sets are also critical to advancing science. Yet they are hardly integrated at all into traditional textual search engines, let alone in any meaningful federated way across data sets. This, too, is a rich opportunity for expanding and improving access to valuable information.

ICSTI, who provided invaluable leadership for WorldWideScience.org, is sponsoring a number of technical projects that address some of these challenges. In the area of numeric data, TIB-Hannover in Germany is leading a multinational project to demonstrate the integration of access to numeric data sets from within grey literature textual reports.

On the multimedia front, ICSTI is leading a project to demonstrate how indexing of spoken words in audio and video files can result in profoundly improved search precision. Another ICSTI project is exploring how Web 2.0 technology can be used to improve scientific communication.

Conclusion

Through the demonstration of WorldWideScience.org, it is clear that grey is global, and it has benefited from a global solution. Second, grey is growing, both in traditional formats and media but also in emerging forms, which need to be offered the same level and ease of access as textual literature. Finally, grey is good, and must be treated as an essential commodity for progress in all fields, especially science, medicine, and technology.

References

1. http://worldwidescience.org/
2. http://worldwidescience.org/alliance.html
3. http://en.wikipedia.org/wiki/Gray_literature
4. http://www.nla.gov.au/padi/topics/372.html
5. http://www.websters-online-dictionary.org/definition/grey
6. http://www.google.com/
7. http://www.yahoo.com/
8. http://www.osti.gov/
9. http://science.gov/
10. http://www.scienceaccelerator.gov/
11. http://www.osti.gov/eprints/
12. http://www.osti.gov/scienceconferences/
13. http://www.osti.gov/fedrnd/
14. http://www.icsti.org/
15. http://www.bl.uk/
16. http://www.jst.go.jp/EN/
17. http://etde.org/
18. http://opensigle.inist.fr/
19. http://international.inist.fr/
20. http://www.ub.uio.no/nora/noaister/search.html?siteLanguage=eng
21. http://www.csir.co.za/
22. http://www.inasp.info/
23. http://search.arrow.edu.au/
24. http://www.istic.ac.cn/
25. http://vms-db-srv.fnal.gov/fmi/xsl/VMS_Site_2/000Search/video/f_streaming.xsl
26. http://anatquest.nlm.nih.gov/AnatQuest/ViewerApplet/aqrendered.html

Grey Literature in the Czech Republic

Petra Pejšová and Martina Pfeiferová
State Technical Library, Czech Republic

Abstract

Our paper summarizes and describes activities concerning grey literature in the Czech Republic. The managing organization of the activity is the State Technical Library (henceforth the STL); in the past, it was the STL, which was collecting, publishing and submitting grey literature data into the SIGLE system. The STL was the representative of the Czech Republic in the EAGLE. Now, EAGLE being extinct, there is no coordinated collection of grey literature on the national level since 2005.

The STL complements the role of the Czech National Library, which under the National Digital Library project, aims at accessibility of widespread published documents ("white literature"). On the other hand, the STL intends to deal with literature not acquired through normal bookselling channels (grey literature) and initiated a project for grey literature retrieval.

The project the National Repository of Grey Literature (henceforth the NRGL) is supported by the Ministry of Culture. Its main objective is formation of a digital repository of grey literature in the Czech Republic. The project aims at gathering metadata and possibly full texts of grey documents in the field of education, science and research. The NRGL shall solve the typology of documents collected as well as metadata formats, persistent identifiers, intellectual property issues, SW and HW support, formation of network of collaborating institutions etc.

Close collaboration with representatives of Czech universities has been established. They face the issue of storing university qualification theses, which is one of the segments of typology of documents collected by the NRGL. The National Registry of Theses shall become a component of the NRGL. The STL has also got in touch with further producers of grey literature in the Czech Republic, in particular research institutes of the Academy of Sciences the Czech Republic, the institution covering a major part of production of grey literature in the segment of research and development.

The NRGL should facilitate to research the data on grey literature in the Czech Republic at one place with a single interface, as well as to retrieve the information on the owner of the document and - if possible - the full text of the document, either in electronic form or via the contemporary network of libraries (interlibrary loan, Document Delivery Service etc.).

The NRGL project does not assume retrospective digitalization of grey literature documents. However, we intend to certify the NRLG a trustworthy repository. The aim of the project is to provide services not only to the NRGL contributors, but also to the widest public. The STL has based the project on practice of universities, which had already experience with local repositories. The entire the NRGL project is consulted with the National Library of the Czech Republic as a part of the Czech Digital Library project.

Introduction

Our paper summarizes and describes activities concerning grey literature in the Czech Republic. We present projects addressing grey literature and especially the project the National Repository of Grey Literature (henceforth the NRGL), which solves the State Technical Library[1] (henceforth the STL).

About the State Technical Library

The STL is the central professional library under the governance of the Ministry of Education, Youth and Sports of the Czech Republic. It was founded in 1718; it provides library and information services to corporate and individual public, especially those in higher education, research and development. The STL is also a public library dedicated to science and technology, as such it collects and administers state-funded collections of czech and foreign literature and other information resources and sources pertaining to technology and applied natural and social sciences associated with technology.

The STL collections contain over 1.2 million volumes - books, journals, newspapers, scholarly studies, trade information, electronic documents and other publications and texts from the field of technology and applied natural and technology-related social sciences. Historical collection (books and journals published between 1500 and 1920) contains 22 965 volumes.

The STL is centre of digital document delivery system – the Virtual Polytechnic Library working on the basis of multifunctional centre of information services together with co-operative development of decentralized collection of periodicals specialized by subject and document type. The system has special union catalogue of collection of 50 Czech libraries and facilitates ordering by registered users from the Czech Republic.

The STL is host of the Czech National ISSN Centre (ČNS ISSN), which assigns ISSN numbers and registers continuing resources (serial publications) published on the territory of the Czech Republic.

History

The idea of the NRGL originated in the STL in 2005. The STL aims to collect a grey literature and to complement a role of the National Library Czech Republic[2] (henceforth the NL CR), whose main task is to collect and preserve "white" literature. The NL CR doesn't collect and doesn't plan to collect grey literature. The STL was motivated by the termination of the SIGLE system, formerly implemented by EAGLE, European Association for Grey Literature Exploitation. There were two participating members from the Czech Republic in EAGLE: the STL and the Library of the Academy of Sciences of the Czech Republic. To support the participation and collaboration, in 1994 the STL formed a specialised system, Co-operative System for Grey Literature. The system, based on bilateral agreements with grey literature producers, collected bibliographic records on grey literature, namely dissertation theses, from participating czech universities. Metadata collected were converted into a special SIGLE data format. Besides, the STL was processing grey documents from its own collections into the SIGLE format. Universities had an option to make a preliminary relevance-based choice of data submitted. Hence, this procedure did not cover the entire production of dissertation theses in the Czech Republic. Some universities sent to the STL printed full texts of theses, this collection now contains over 4 000 theses. The task of the Library of the Academy of Sciences of the Czech Republic, i.e. to collect and file grey literature produced by institutes of the Academy of Sciences of the Czech Republic, was not fulfilled during the entire period of existence of the SIGLE. Thus, the STL was the only active contributor and national coordinator of the SIGLE activities in the Czech Republic. After the termination of the SIGLE, the STL began with an initiative to collect grey literature at the national level. The implementation of this initiative has begun in 2008 thanks to the support of the Ministry of Culture of the Czech Republic in the framework of research and development projects.

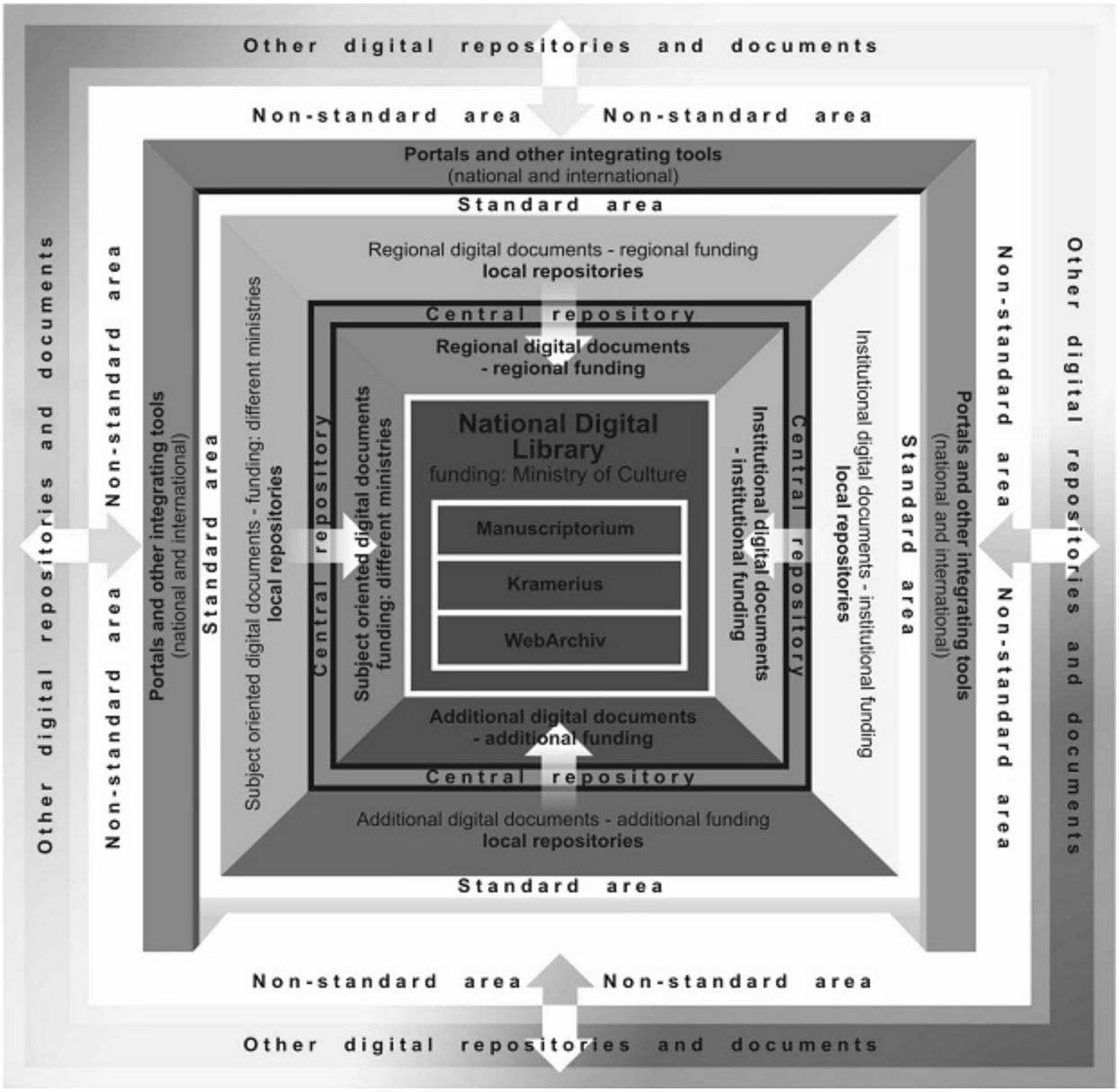
The NRGL Project

The project is supported by the Ministry of Culture of the Czech Republic and its full title is "The Digital Library for Grey Literature –Functional Model and Pilot Implementation" (henceforth the Project). It is planned for period of four years, commencing in 2008 and ending in 2011. There are two participants: the STL and the University of Economics, Prague. The Project shall provide a working pilot application, which shall form the basis of the NRGL. Having evaluated and tested technology and methodology, we shall formulate Standards and Recommendations for institutions establishing their own local grey literature repositories. The Standards and Recommendations shall include our experience gained during the Project as well as rules and methodology; namely recommended metadata format, interchange formats and templates, samples of licence agreements and other relevant legal issues, methodology of protection, archiving and publishing of digital data, and web interface for communication with producers of grey literature. The Standards and Recommendations shall be published both in Czech and English in printed form and on project web pages. The NRGL will hold grey literature records from research and development areas, Civil Service, entrepreneurial sphere, education and open access. We negotiate about co-operation with representatives of systems covering these areas.

Conception of the Czech Digital Library and the NRGL

The NL CR builds a concept of the Czech Digital Library[3] (henceforth the CDK) via a scheme. The core of the CDK is the National Digital Library (henceforth the NDK), which is at the centre of the scheme. The NDL is concerning just for "white" literature in framework of preservation of cultural heritage of the Czech Republic.

Figure 1: Scheme of the Czech Digital Library



The NRGL will comprehend grey literature. In the basic functional scheme of the CDK, the NRGL belongs to the Standard area, which is covering databases and aggregate catalogues of digital documents; the NRGL shall be such a database. The NRGL itself shall collect metadata and full texts from local repositories, both institutional and branch-oriented, and other sources in the Standard area. The responsibility for local repositories (including financial) will be born by particular institutions, and by their authorities, i.e. government departments and agencies.
Within the research project "Development of Mutually Compatible Information Systems for Access to Heterogeneous Information Resources and their Coverage by Uniform Information Gateway" NL CR defined a new sub-objective "Developing Digital Libraries and Repositories in consideration of their possible integration under the Uniform Information Gateway and other national information portals". In order to address this task, a specialised body, Working Group for Coordination of Development and Exploitation of Digital Repositories, originated. The participants in the NRGL project are members of this Group.

The National Registry of Theses and Plagiarism-Tracing System
The National Registry of Theses and Plagiarism-Tracing System is a project involving 20 Czech and Slovak universities. The project has two main parts. The first part of the National Registry of Theses (henceforth the NRT) gathers metadata on university qualification theses (i.e. metadata). The data is accessible to the public. The NRT contains over 30 000 records of theses now. The other part, Plagiarism-Tracing System, serves for detecting plagiarism in scholar texts. The system will help educators to unveil

shady documents and decide, if the document in question (or its part) is or isn't plagiarism. The system can be found on the web page http://theses.cz.
From the outset of the project, the NRT has been considered an important source for the NRGL. Such documents form an important sector of grey literature to be filed by NRGL. The STL will collect metadata from the NRT into it's the NRGL repository using the OAI-PMH protocol. Retrieving of full documents must be based on bilateral agreements between the NRT and particular universities. This is why we have prepared standards of licence terms. The compatibility is guaranteed by the participation of University of Economics, Prague.
The STL intends to collaborate with all universities and colleges in the Czech Republic (including private colleges) that are not participating in the NRT project, in order to obtain data on their qualification theses (and also - based on licence agreements - full texts as well).
In addition, the NRGL takes into consideration normative documents of "Electronic Theses and Dissertations Working Group, Association of Libraries of Czech Universities". At present, the metadata format of the NRT is being incorporated into the format of the NRGL.

The Register of Publication Activity implemented in the Academy of Sciences of the Czech Republic

The Academy of Sciences of the Czech Republic[4] (henceforth the AS CR) is a public non-university institution formed by a system of research institutes focusing on basic research. The AS CR defines the policy of scientific research; it is involved in both national and international research programmes, it supports collaboration with application sphere and it promotes the development of education. The AS CR consists of 54 research institutes.
The AS CR does not monitor its production of grey literature separately. Grey literature is included in a general system of monitoring of output and results of research (primarily for the purposes of research funding). These results are gathered by the Register of Publication Activity implemented by the Library AS CR[5].
The STL collaborates with the Library AS CR. Metadata from the Register of Publication Activity of the AS CR[6] shall be collected into the NRGL repository using the OAI-PMH protocol. Full-texts retrieval from the AS CR institutes has to be treated in the same way like with universities, i.e. by licence agreements.

The Project Schedule

The Project is planned for period of four years. It is scheduled into three phases. In the First Phase currently running (April 2008 – June 2009), we shall define the requirements and documentation for the model application. This involves metadata specification, a choice of relevant persistent identifiers, analysis of typology of documents, software specification for the model application, a proposal of licence agreements for producers of grey literature and creation of project web pages. In the Second Phase, from July 2009 to September 2010, the model application shall be implemented, tested and evaluated. The Third Phase (July 2010 – December 2011) shall include formulation of standards, recommendations and methodology, verified in the Second Phase, which shall be published on the web as well as in a printed form. During all three phases, results and conclusions shall be published on the project's web pages. Seminars on an access to grey literature and electronic university qualification theses will take place annually. In addition, research results will be presented to the professional community.

The NRGL metadata format

We have defined our own metadata format for the NRGL system. There are some important requirements that the format shall fulfil: simplicity, few obligatory attributes, consistency with Dublin Core[7], accessibility for OAI-PMH and possibility of creation of elementary indices. As a basis we used metadata formats of the two principal the NRGL participants: the NRT and the Register of Publication Activity of the AS CR[8]. We also took in account the format of the Information Register of Research and Development Results[9] (henceforth the Register R&D). We do not incorporate this format to be able to download data from the Register R&D directly, rather to enable local repositories, that created a record into the Register R&D with the possibility to provide the NRGL with the same record without any further conversion. If the OpenSIGLE will be re-opened, we plan to incorporate this metadata format into the NRGL metadata format in order to be able to submit records into the OpenSIGLE. We also respect Dublin Core format. At present, we are defining hierarchic relations among records and rules for particular identifiers.

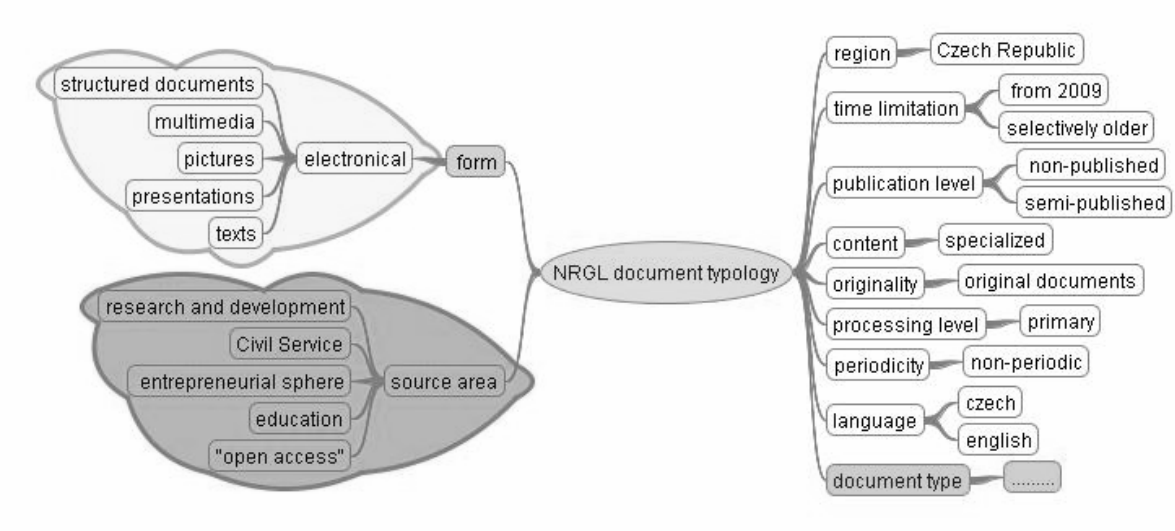
Questionnaire survey

At the beginning of the project, we addressed total of 77 producers of grey literature via questionnaire while we focused on research institutes of the AS CR and public (state-funded) czech universities. The aim was to obtain contact information on persons responsible in individual institutions together with their intention to cooperate with the NRGL. In addition, we tried to determine the method of registration, collection and access to grey literature. Total of 39 institutions (from 47 responding to the questionnaire) declared their willingness to cooperate with the NRGL. However, their approval depended on the licensing conditions that would be negotiated for this cooperation. The STL will commence the cooperation with these institutions in the second half of 2009 in accordance with the project schedule.

The NRGL Document Typology

When determining types of documents for the NRGL we started with typology of the GreyNet[10], the Register of Publication Activity of the AS CR and the Register R&D[11]. During the analysis of the document types we found that typologies concerned reflect different aspects such as events (manifestation, arrangement, organization,), form (presentation – notification, processing – translation, output), content (policy documents), places (domestic, foreign), type of format (e-text), etc. Therefore, we decided to divide our typology according to various aspects. Therefore, we decided to create typology, which will be structured in several levels so that we could express the various aspects better via a mind map, see picture below. We also anticipated that the NRGL would contain only "digital-born" documents.

Figure 2: Mind map of the NRGL Document Typology



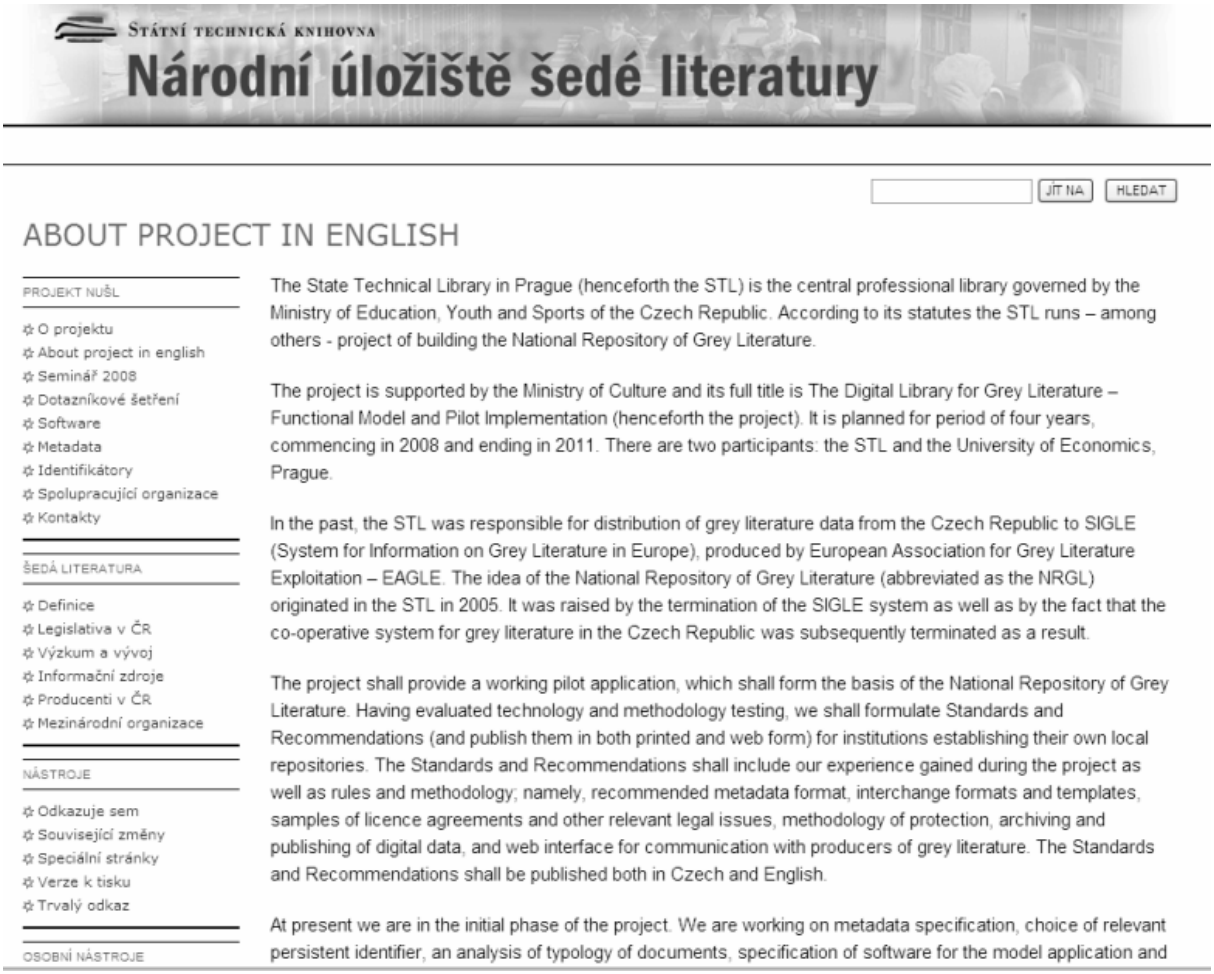
Software for the NRGL

During 2008, we were able to specify requirements for the SW functionality for the NRGL and follow-up services. These requirements were specified on two levels. In the first level requirements for pilot implementations of the system are specified, while on the second level requirements for further development of the basic system functionality are specified. During the process of SW selection, taking place in 2009, we will take into account the requirements of both the first and second level.
Requirements of the second level are important for selection of such modern technology that will ensure compatibility of digital libraries with modern trends in information technology. Since we're considering the use of open source solutions, we compare SW requirements with available open source software for digital libraries. In this analysis we included the following open source SW: DSpace, Fedora, CDS Invenio, Eprint, and Greenstone. We plan to take a close look at LOCKSS Program and some Web 2.0 based tools, too.
During this year we also deal with the problems of persistent identifiers. We carry on analysis of available persistent digital object identifiers in the Czech Republic. Our decision will also depend on the phase of implementation of NBN (National Bibliographic Numbers) nationwide variant by the National Library of the Czech Republic.

The Project Web Pages

Within the project we have created a website which can be found at http://nusl.stk.cz/. Web pages are created in the Media Wiki[12] tool and WordPress[13] graphic style modified in accordance with the STL graphic manual. Media Wiki application supports interactive interface, which in our case represents the shared content creations within the working group and discussion forums opened to the public. The website contains information on the project, which are continuously updated, while we continue to open discussions on individual articles. In addition, there is information on the grey literature from the Czech Republic as well as the rest of the world, info on the legislation related to the grey literature in the Czech Republic, links to sites supporting R&D in the Czech Republic, on-line proceedings of the "Seminar on the access to grey literature" and grey literature information resources. In 2009 we plan to translate project web pages into English. For the moment, we have translated at least basic information about the project in the article "About project in English" available on Czech web pages[14].

Figure 3: Sample of The Project Web Pages



Seminar

In relation to the project, this year we organized for the first time "Seminar on the access to grey literature 2008", held on October 8th, 2008. It followed the seminar "Systems of access to electronic university qualification theses 2008", held on October 7th, 2008. At the seminar we presented the project itself, the present state of grey literature systems around the world, PDF format and ISO standards for long-term archiving, standardization of open archives focused on description and exchange of aggregated web sources via OAI-ORE, Creative Commons license copyright issues and the state of implementation of persistent identifiers in the NL CR. All presentations and full texts of lectures are available in the online proceedings from the seminar on the Project web pages. Further seminars on the access to grey literature will be held annually in the coming years.

References

1 http://www.stk.cz/cs/
2 http://www.nkp.cz/_en/index.php3
3 http://www.ndk.cz/
4 http://www.cas.cz/en/
5 http://www.lib.cas.cz/cs
6 http://www.library.sk/i2/i2.entry.cls?ictx=cav&language=3
7 http://dublincore.org/documents/dcmi-terms/
8 http://www.iach.cz/knav/database_en.htm
9 http://www.vyzkum.cz/FrontClanek.aspx?idsekce=1028
10 http://www.greynet.org/greysourceindex/documenttypes.html
11 http://www.vyzkum.cz/FrontClanek.aspx?idsekce=29415
12 http://www.mediawiki.org/wiki/MediaWiki
13 http://www.redaktionundalltag.de/
14 http://nusl.stk.cz/index.php/About_project_in_english

Towards an Institutional Repository of the Italian National Research Council: A Survey on Open Access Experiences

Daniela Luzi, Rosa Di Cesare, Roberta Ruggieri and Loredana Cerbara
Institute of Research on Population and Social Policies, IRPPS-CNR, Italy

Abstract
The paper presents the results of a survey aiming at identifying documentation, organization as well as technological resources that could be the basis for a future development of a CNR IR. The survey makes use of a semi-structured questionnaire submitted to all CNR research units. Results show that, despite a limited number of OAI complaint repository developed under the autonomous initiative of some CNR research units, there is a mature environment for the development of an IR.

1. Introduction

Institutional repositories (IRs) constitute one of the most important applications of Open Access (OA) principles, which combines both scholars' and research institutions' concern to set up a new scholarly communication paradigm as well as enhance the visibility and impact of scientific research. However, although an increasing number of IRs have been developed internationally since 2002, the debate started by Lynch [Lynch, 2003] and Crow [Crow, 2002] is still open [Geudon, 2002, Harnard, 2005, Ginsparg, 2007, Suber, 2008], aiming at identifying the best way to make IRs an innovative channel for the dissemination, exchange and preservation of scientific contents.
In our opinion the major elements of success of Institutional Repositories (IRs) depend upon a sound synergy among the different stakeholders participating in the process of production, sharing and diffusion of the knowledge produced within the scientific organisation. In fact, in the activities of planning, designing and supporting a new IR it is necessary that a minimum set of conditions are fulfilled:

- The scientific institution has to declare its official commitment to the OA policy, advocating scholarly open access publication;
- Scholars have to be personally motivated and supported to populate IRs;
- An information network has to be built to support the activities connected with the submission and management of scientific contents, at the same time promoting an OA culture;
- A technological infrastructure has to be developed to support the implementation of OAI-PMH compliant repositories.

Even if there is no national policy on IRs, in Italy Open Access Initiatives (OAI) have been primarily promoted by Universities and intra-university consortia, which almost unanimously adhered to the Berlin Declaration [Berlin, 2003]. At the moment the ROAR registry reports about 40 operational OA archives [ROAR], which demonstrates a constant increase in these information systems, even if the number of scientific contents contained shows that they are still not fully deployed. Among other governmental research institutions only the Italian National Institute of Health [De Castro, 2007], has officially supported OA in 2007 making the depositing of its internally produced post-prints mandatory into its newly developed IR.

The National Research Council (CNR), one of the biggest multidisciplinary research institutions composed by a network of 107 geographically distributed institutes, has not expressed an official position toward OA yet. However, thanks to its scientific and organisational autonomy, some CNR Institutes might have made their scientific production freely available either through their own local repositories or through freely accessible web resources. In order to acquire a more precise picture of the AO CNR practice, a group of CNR researchers and librarians (hereafter called "supporter group") has promoted a survey aiming at identifying documentation, organization as well as technological resources that could be the basis for a future development of a CNR comprehensive IR.

2. Research hypothesis and objectives

Many surveys have been carried out in order to analyse scholars' behaviour and attitude towards publication submission in IRs, which could become "a set of empty shelves", if not adequately populated [Gibbons, 2005]. Other surveys are focused on inventory IRs at a national [Rieh, 2007, Zuber, 2008] and international level [Lynch, 2005; van Westrienen, 2005; van der Graaf, 2008] comparing national policies toward OA, IR characteristics and management.

Considering the development of a new IR as the enforcement of the synergy among different stakeholders, our survey is based on the research hypothesis that aims to:

- Identify actors and roles currently carried out in the development of the different services supporting research activities (i.e. librarians, technical and administrative staff) that contribute to the management of a new IR in terms of system implementation and maintenance, metadata control, and collection definition. The existing cooperative network between CNR libraries organised in consortia arrangement to support common access to digital resources and document delivery already constitutes a shared and value added backbone useful to support an OA culture and practice.
- Identify information and documentation services already developed to disseminate institution's scientific production such as digital libraries, different types of web pages as well as the annual activity report that gathers research outcomes for evaluation aims. All these resources could be integrated in a future CNR comprehensive IR from a technological as well as from an organisational point of view.
- Analyse the content deposited into already developed local IRs or contained in other digital free access archives. This also implies the identification of the actors who provide scientific content (researchers, librarians). The quali-quantitative analysis of the content as well as of the different actors that make it available can provide meaningful indication on the degree of awareness towards OA culture.

Summarizing, our survey intends to carry out an inventory study to identify information and technological resources currently available at CNR research units, which can became the basis for a feasibility study aiming to the development of a future CNR comprehensive IR.

3. Methods

In the phase of survey design great importance was attached to the setting up of what we call a "supporter group", composed by CNR librarians that voluntarily collaborated in sending and collecting the questionnaire. Moreover group tasks were also the direct involvement of other participants so that the supporter group could be enlarged during the survey. This approach is motivated by different factors. Past and present co-operative experiences between CNR libraries ranging from SIGLE participation to actions supporting common access to digital resources have proved to be successful and have created an informal network of collaboration. The universe of CNR libraries is very diverse; depending on their location they can manage large bibliographic collections, serve more than one research unit, have close and embedded collaboration with technological services or, on the contrary, they can be small libraries with very limited human resources. For these reasons a direct knowledge of the reality to be investigated can ensure not only a good response rate, but also accurate data collection of questionnaires. Last but not least, we think that a future development of a CNR comprehensive IR has to directly involve librarians as one of the major stakeholders in the promotion and setting up of an IR.
At the beginning of the survey the supporter group was composed of 10 librarians, which became about 25 at the end of the study, the group is widespread throughout CNR research units.

Then the questionnaire was developed together with the identification of CNR survey units that had to be reached by the questionnaire. We decided to submit the questionnaire to all CNR Institutes and their territorial sections (hereafter called CNR research units), so that the survey could have the feature of an inventory study fulfilling the proposed aims.

The survey makes use of a semi-structured questionnaire of 14 questions within two different sections that reflect the survey objectives: questions 1-9 are related to the characteristic of the developed OA archive, questions 10-14 concern information resources on the Institute's scientific production made freely available in other forms. Additionally, respondents were asked for information on gender and occupational position. The questionnaire submission started in March 2008 and the questionnaire collection was ended by June. Completion of the questionnaire took about 20 minutes.

In particular the questionnaire seeks answers to the following interrogatives:

- Quantification of already existing and/or IRs under construction and identification of their characteristics,
- Identification of other types of freely accessible web resources (institutes' and/or scholars' websites, collections of scientific production for evaluation uses),
- Identification of the actors responsible for the content submission, management, cataloguing and dissemination,
- Identification of practices for the input of internal scientific production within the library catalogue and CNR Annual report.

Another important decision concerned the adoption of the IR definition given both by DRIVER Project and SPARC in which an IR is defined as:

- Containing research results,
- Institutional and/or thematic, and
- OAI-PMH compliant.

The adoption of this definition and in particular the prerequisite of OAI metadata harvesting makes it possible to verify technological requirements that should be fulfilled in future planning and design of a CNR comprehensive IR.

4. Results

4.1. Survey numbers

Table 1. - Distribution of questionnaire received by CNR Research units.

Survey Numbers			
	CNR Research units No.	*Received Questionnaire No.*	*Coverage %*
Institutes	107	93	87
Sections	187	67	36
Total	**294**	**160**	**54**

Table 1 shows the distribution of the questionnaire received by CNR research units. 160 out of 294 Research units - equal to 54% - provided correctly completed returns. Most of the respondents were CNR Institutes (87%) and this indicates that the survey reached its major target, as CNR Institutes should be the main institutional promoter of local repositories and could become a first level aggregator of the publications of its belonging research units.

Moreover, 15 Institutes also sent their questionnaire including the data for their sections, so that the information coverage of the survey is higher than shown in table 1, reaching almost 60% of all CNR research units.

Fig. 1. Distribution of questionnaire received by CNR Department

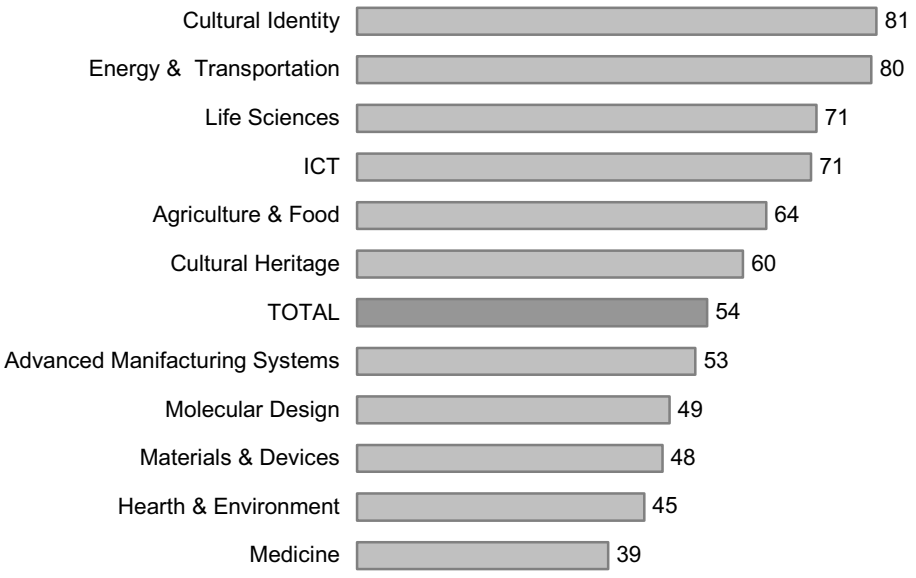
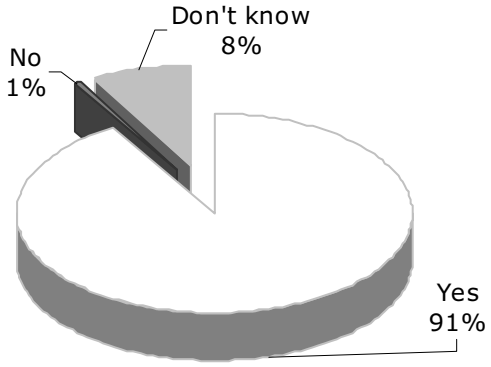


Figure 1 shows the distribution of the questionnaire received by CNR Department, which group Research units on disciplinary basis. In 7 departments out of 11 the respondent rate was higher than 50%, reaching high coverage percentages in the Department of Humanities and Social sciences (Cultural Identity 81%) as well as in Energy and Transportation (80%). The high rate of these responses could be ascribed to the active involvement of the supporter group in these specific areas.

4.2. Berlin Declaration

Fig. 2. – Distribution of respondents to the question:
"Do you think CNR should subscribe the Berlin Declaration?"

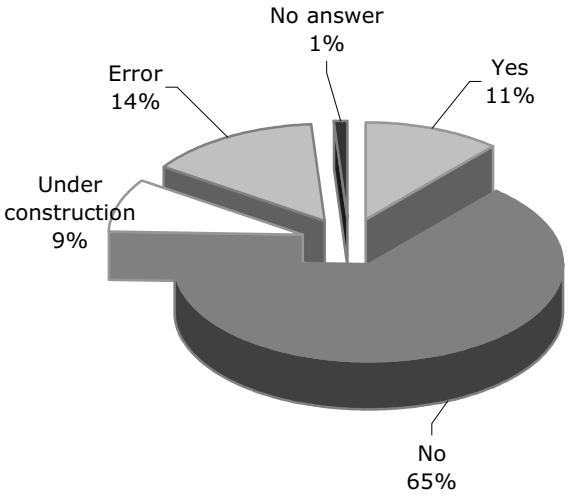


In the first question of the questionnaire respondents were asked whether CNR should subscribe the Berlin Declaration. This question gave us the possibility to summarize the principles of OA as well as to obtain the respondents' opinions on their willingness to make CNR as well as their own scientific production freely available. 91% answered that they were in favour of CNR subscribing the Berlin Declaration, while only 1% was against and 8% indecisive. This positive attitude towards OA encourages us to further promote an OA policy at an institutional level requiring CNR top management to both subscribe to the Berlin Declaration and support the setting up of a CNR comprehensive IR.

4.3. Research units' Repositories

The first section of the questionnaire aimed to ascertain whether CNR research units have developed their own repository to collect the publications produced by their researchers. Moreover, this section intended to analyse repositories' information content, access policy, as well as their organisational and technical characteristics.

Fig. 3 - Distribution of respondents to the question:
"Does your Institute have a local IR to collect the publications produced by its researchers?"



A common understanding of the definition of IR is required in the analysis of the answers to the question whether each research unit has developed its own repository or is about to develop it. As mentioned above, in our survey we decided to adopt the definition given by DRIVER (Digital Repositories Infrastructure Vision for European Research) and by SPARC (Scholarly Publishing and Academic Resource Coalition) because among other criteria ("containing research results and being institutional and/or thematic") it includes the important factor of IRs being compliant with OAI-PHM standard. This criterion is particularly crucial in the case of CNR. Given its multidisciplinary nature as well as its network organization composed of autonomous research units, a future comprehensive IR should be based on a federated system architecture made of locally managed but at the same time interoperable repositories. Moreover, according to our broad objectives, getting a precise picture of the technology used by CNR

research units can contribute to the identification of the right strategy to implement such infrastructure, taking the already developed systems into account as well as providing guidelines for new repositories to be brought into operation.

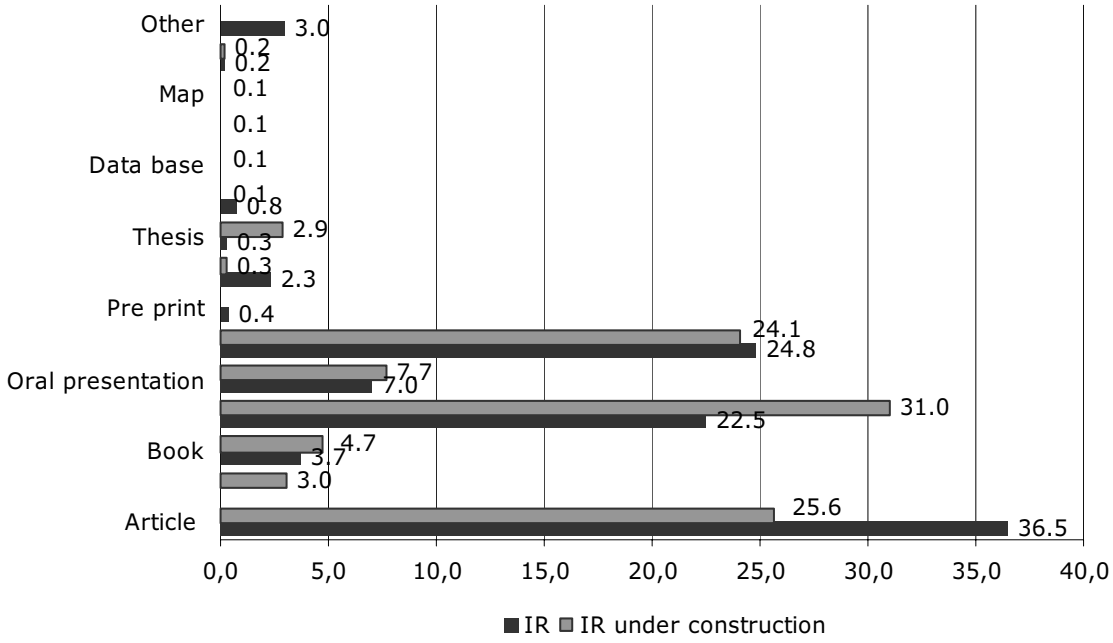
For this reason the responses affirming to have set up a local repository have been further analysed, at first verifying the information given in the questionnaire about the software used and then checking this information on the research unit's websites. This enabled us to verify the correspondence with the definition of IR adopted. In cases of doubt we contacted the respondents for more detailed information to verify whether and what type of metadata standards were adopted in order to make the repository OAI-PHM compliant.

The results of this *two-step* analysis are shown in figure 3. The majority of CNR research units do not have a local repository (65%), only 18 research units (11%) have developed a repository and 14 (9%) are planning one. A meaningful percentage of respondents (14%), despite affirming to have a local repository, has been classified as "error", because under analysis, their system turned out to be not compliant with OAI-PMH standards.

The analysis of these responses (equal to 23 Research units) showed a variety of forms of making their scientific production freely available. Some of them (8 out of 23) give a list of the publications sometimes linked to abstracts and/or full text, while 6 research units provide databases, generally searchable by type of publication, and sometimes make the abstract available. Interestingly, some research units considered they have a repository because they identify it with a CNR central collection of the scientific production sent yearly to construct the Annual report.

4.4. Content of CNR research units' repositories:
To determine the size of the repositories in terms of number of contents as well as document types considered, the respondents were asked to give an approximate number for each type of materials present in the local repository selecting them from a predefined list. This list included among the most frequently deposited textual materials, such as journal articles, books, technical reports and other GL documents also video, maps, primary datasets, software and databases. Some of these non-textual materials were not selected by the respondents and therefore are not considered in the analysis (Fig. 5). The extension to other types of research materials should be encouraged, as it would more closely reflect the research activities carried out within an institution and would therefore be worth depositing and diffusing in an IR. The analysis of the following questions pertain to 17 respondents out of 18 in the case of already developed repositories, and 11 out of 14 for the repositories under construction.

Fig 4 – Distribution of the document types present in already developed repository (n =17) and in those under construction (n = 11)



The total number of contents deposited in the local repositories is about 15.000, while in those under development there are about 3.000 documents.
The comparison between the document types deposited in the already developed repositories and those present in the repositories under construction is given in figure 4. The most frequently inserted document

types are journal articles, published proceedings and technical reports. It is interesting to note that in the repositories under construction, even if at a very low percentage, there is a tendency also to make other materials available (e-articles, courseware materials, videos and maps) or to insert some of them more frequently (theses 2,9% in repositories under construction vs. 0,3% in already developed repositories).

4.5. Period of publication of the deposited materials

Year of publication of the deposited materials provides us with important indications about the choices made by research units to populate their local repositories. This information may highlight if the repository mainly contains current research outputs or, at the opposite, tends to deposit dated scientific production. Therefore respondents were asked to estimate time range and their relative number for each deposited document type (less than 5 years, from 6 to 9 years, more than 10 years). Answering was a time consuming activity, which was however fulfilled by the majority of the respondents; 14 valid answers out of 18 for the research units that have already developed their own repository, and 8 out of 14 for those which are developing it.

Fig. 5 - Distribution of documents deposited by period of publication in IRs and in IRs under construction

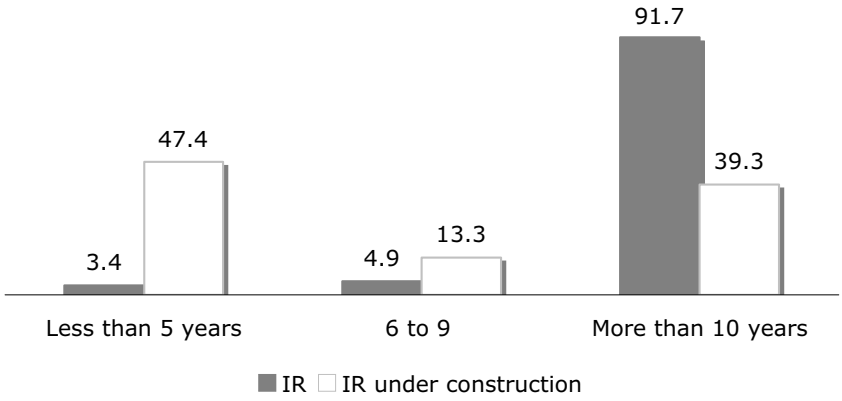
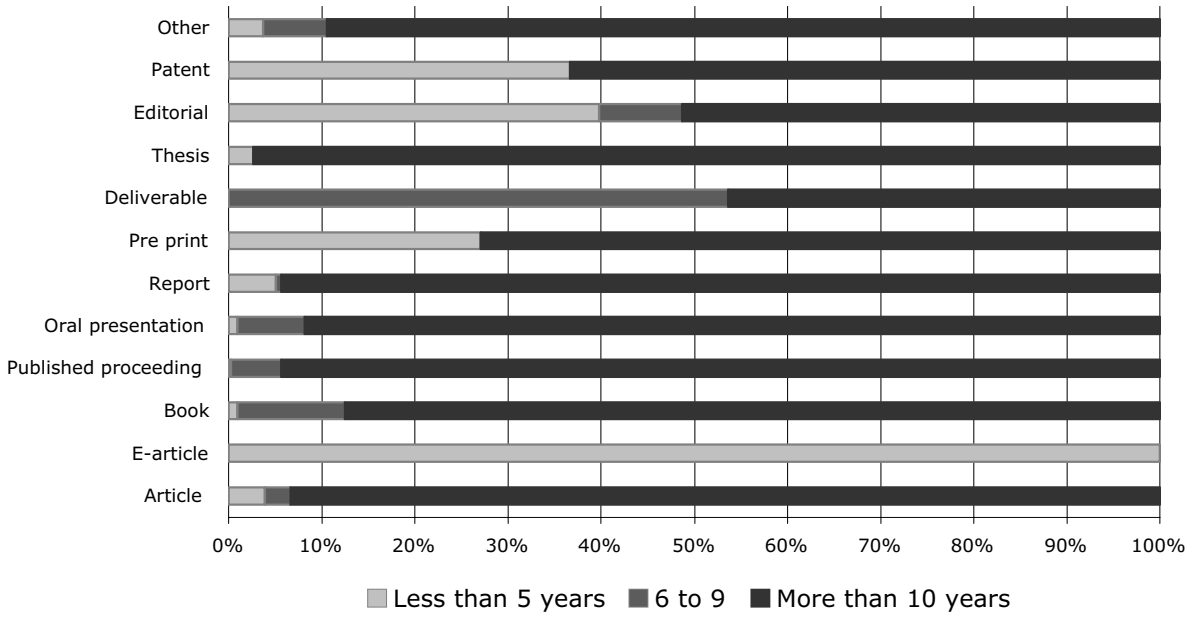


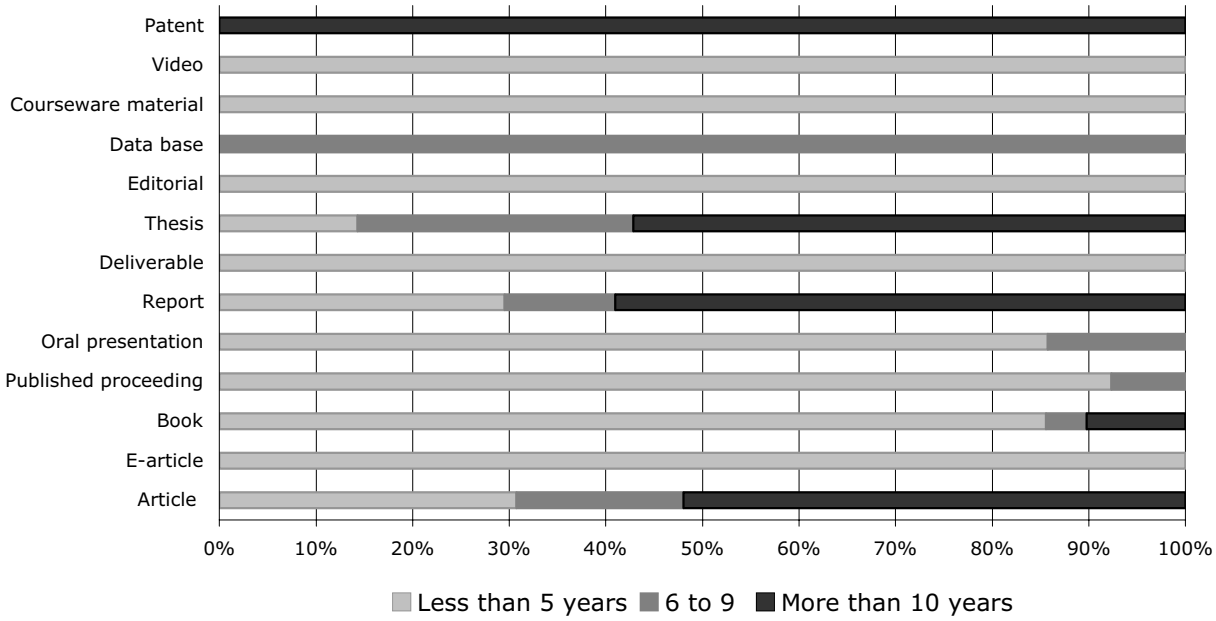
Figure 5 shows the comparison between distribution of documents deposited by period of publication in the already developed repositories and in those under construction. In the former there is a concentration of documents (91%) published more than 10 years ago, while repositories under construction contain documents with publication years more equally distributed. However in these repositories there is a higher tendency to deposit documents that have been published in a time span of 5 years (47% less than 5 years, 13,3% from 6 to 9 years, 39,3% more than 10 years).

Fig. 6 - Distribution of document types by period of publication in already developed IR (n=14)



If we analyse the distribution by document type (fig. 6) we find that in the case of already developed repositories the percentage of documents older than 10 years diminished only for deliverables (46,4%,), editorials (51,3%), patents (63,6%) and preprints (72,9%). The only exception is given by a younger type of publication, e-articles, published in a time range of 5 years.

Fig. 7. Distribution of document types by period of publication in IR under construction (IR=8)



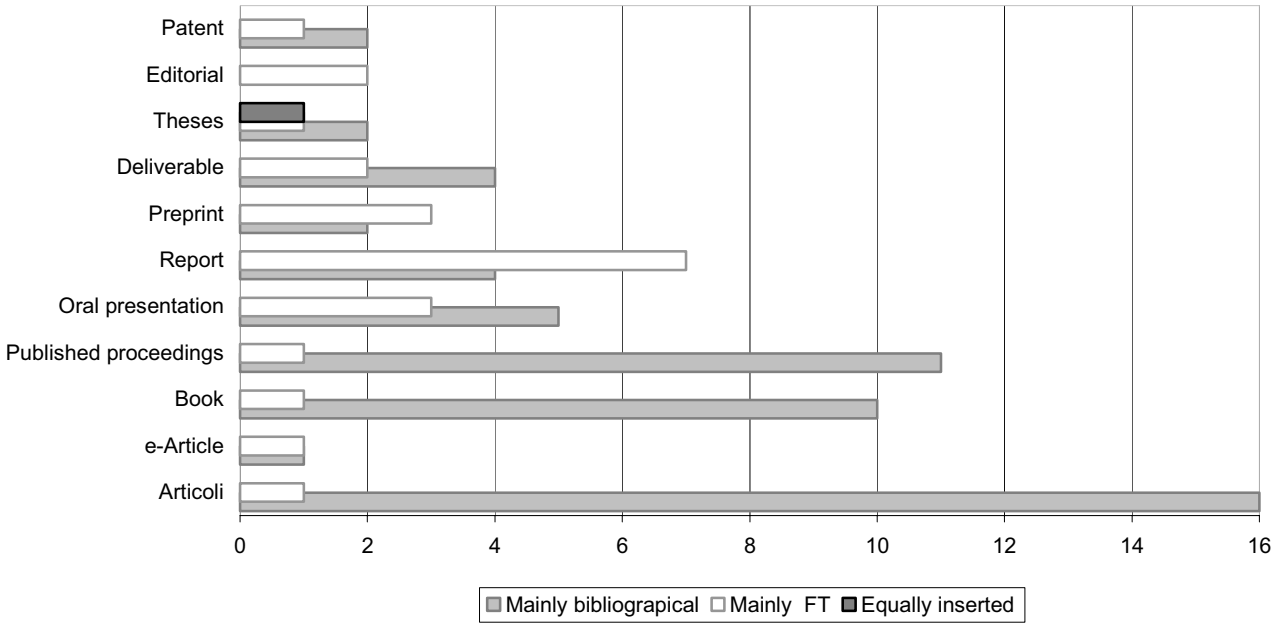
Concerning repositories under development (fig. 7) the tendency to insert documents published in the last 5 years is less evident only for technical reports (59%), theses (57%) and journal articles (52%). *New born* deposited digital objects such as e-articles, video and courseware material are on the contrary published in a time span of 5 years along with more traditional types of documents.

We might think that the difference in the period of publication of the documents deposited in already developed repositories and in those under constriction may depend not only on the *youth* of the repository but also on other factors. For instance already *functioning* repositories may have developed additional functionalities to input pre-catalogued documents from digital libraries or other external systems and/or established archival digitalisation procedures to include particular types of documents and/or documents not violating copyright in their repositories [Bell, 2005]. Other surveys too [Davis, 2007] have had similar results. This constitutes an advantage being an additional service provided by the repository and also represents a positive attitude towards the preservation issues of IRs. However, an exclusively archival use of dated textual materials might make IRs less attractive as an innovative channel of scholarly communication. Of course, this tendency may shift over time making it necessary to monitor repository information contents in order to ascertain their mission and evolution.

4.6. Forms of access

In the next question respondents were asked to estimate for each document type what percentage is deposited providing its bibliographic description and/or linking it to its full text. The availability of full texts constitutes a value added requisite that strengthens the OA vocation of IRs, maximising research access and impact. However this implies a common effort as well as a cultural change for both institutions, which should encourage authors' self-archiving in IRs, as well as authors who should recognise the benefits of making their publications freely available.

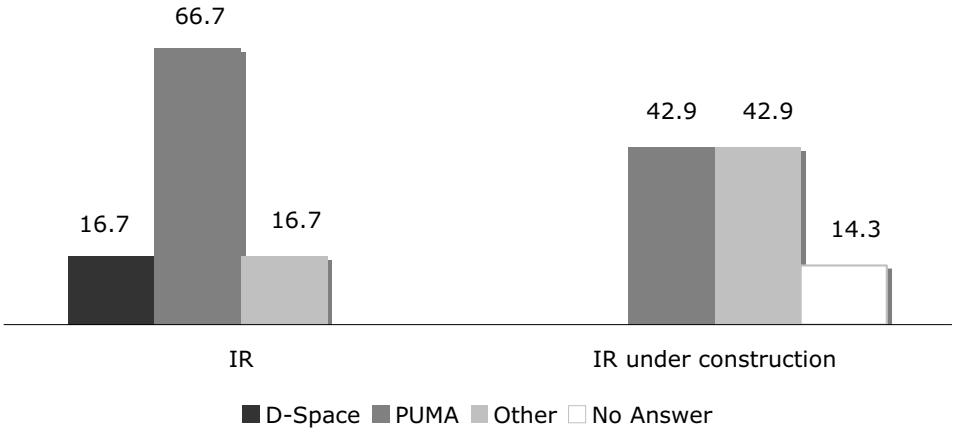
Fig. 8 – Distribution of documents types by access form in local repositories (n =17)



CNR research units' repositories contain mainly the metadata of the documents deposited (fig. 8) and this is predominant especially for journal articles, books and published proceedings. Only GL documents such as technical reports and preprints are predominantly available in full-text. The only exception is represented by the full text of editorials, which are often freely available in full text in the electronic version of published journals. This result is not surprising given the voluntary basis of the developed repositories. Moreover, this result is similar to the data reported in the DRIVER survey where 68% of the textual records contain metadata only, while 32% contain full text [van der Graaf, 2008].

4.7. Repository software packages

Fig. 9 - Software used in developed repositories and in those under constructions



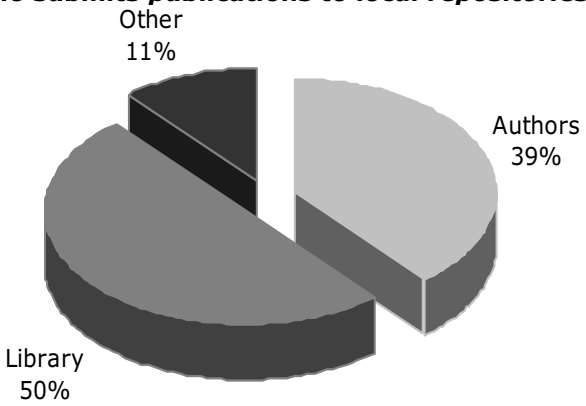
Respondents were asked to indicate from a predefined list which software package has been used to develop their own repositories. The result of this question would give important technological requirements for the development of a future CNR common infrastructure that should integrate already developed repositories and/or provide technical guidelines to make them interoperable.

The software package most commonly used is PUMA (66,7% in already developed repositories and 42,9% in repositories under construction) (fig. 9). PUMA [Biagioni, 2007] is software developed by the Institute of Information Science and Technology of CNR in Pisa that combines functionalities connected with digital libraries and IR also providing some additional services, which are very useful for the management of the scientific production. For instance it enables automatic data transfer of the institute's scientific production collected for the CNR Annual report into a central system that annually gathers the comprehensive scientific production for evaluation aims. Only few institutes use open source

software such DSpace (16,7% of established /already developed repositories), while the tendency also in repositories under construction is to have locally developed software packages.

4.8 Document depositing

Fig. 10. Distribution of respondents to the questions:
"Who submits publications to local repositories?"



To ascertain how the submission process is carried out, respondents were asked to indicate who generally deposits the scientific production (fig. 10). Documents are most frequently submitted by libraries (50%), while a meaningful 40% of authors directly deposit their publications. Considering that authors' submission is often regarded as one of the main concerns in populating IRs, this can be considered an encouraging result.

5. Availability and Management of CNR scientific production

The second part of the questionnaire intended to verify whether research units' scientific production is ever managed and made available in other forms. In our opinion, the acquisition of this information, generally not considered by other surveys, corresponds to our aims; not only providing an inventory study of CNR local repositories but also identifying resources that could be exploited and integrated in order to set up a future CNR IR.

In the first question of the second part of the questionnaire, the respondents were asked whether the research unit's scientific production is ever catalogued. Considering that the library catalogue represents another important way of diffusing the internal scientific production, especially if it is a digital catalogue available on web, this would mean under an IR perspective that documents are already available in a standard bibliographic form whose metadata could be easily transferred to an IR. Moreover this would also indicate that between authors and librarians there is long-standing collaboration for the diffusion of scientific contents.

Fig. 11. Distribution of respondents to the question:
"Is the Institute's scientific production catalogued?"

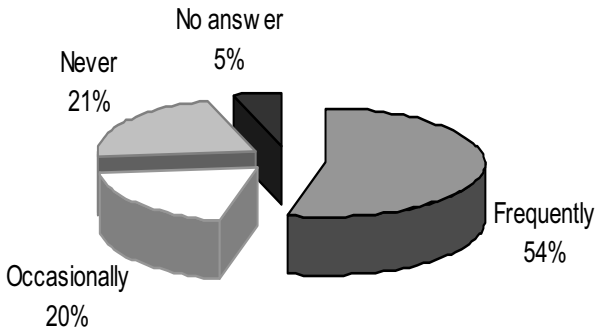
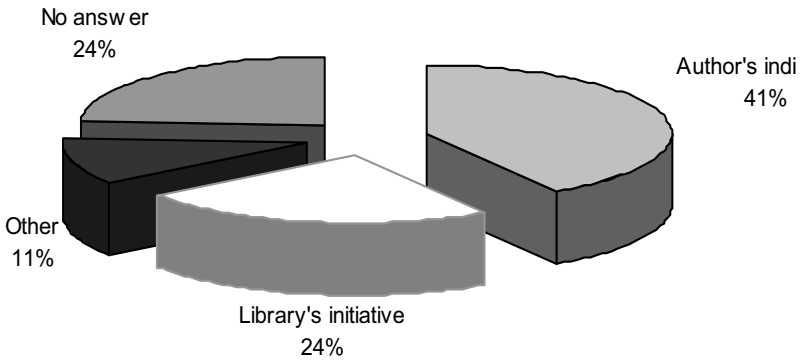
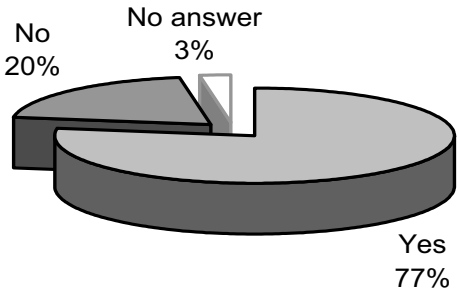


Fig. 12. Distribution of respondents to the question
"The initiative of cataloguing the scientific publications is undertaken by…



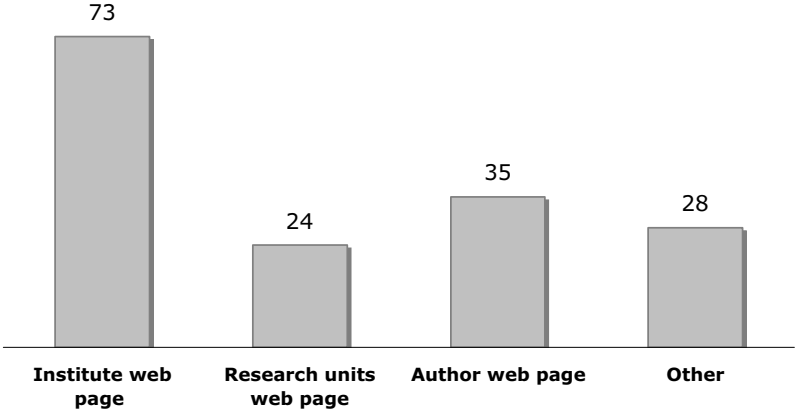
The answers were encouraging (fig. 11); of the 160 respondents 54% declares that the institute's scientific production is often catalogued, 20% affirms that the cataloguing is done sometimes, while 21% never catalogues the institute's scientific production.
In the next correlated question respondents were asked to indicate who takes the initiative of cataloguing the research unit's scientific production. The results of the answers confirmed a positive attitude towards the propensity of making the scientific production available (fig 12).
41% of respondents affirms that authors provide the library with the information about their publications, 24% declares that the library takes the initiative to catalogue the institute's scientific production.

Fig 13. Distribution of respondents to the question:
"Is the Institute's scientific production freely available in other types of digital archives?"



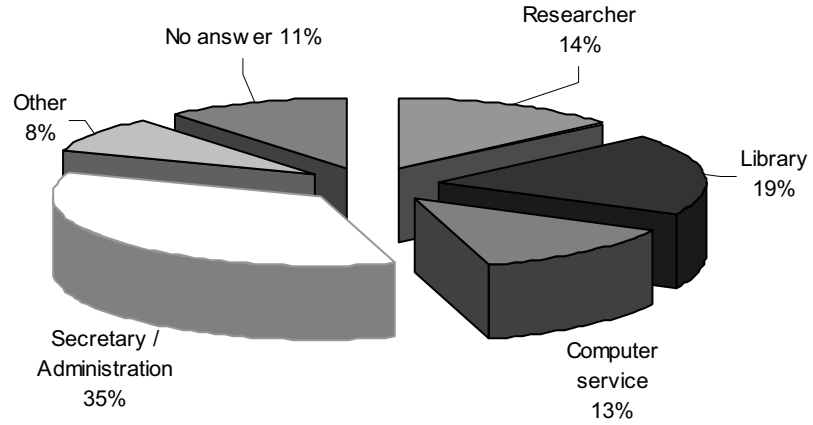
To ascertain if research units make use of other forms of web communication to distribute the internal scientific production, the respondents were asked whether the research unit's scientific production was freely available in other types of digital archive. A previous survey dating back to the 90' [Luzi, 1997], when the Internet was becoming an important and widespread scientific communication channel, showed that CNR research units made large use of the web for the diffusion of their research projects as well as of their scientific production. The results of this question confirm this attitude. 77% of respondents (fig. 13) affirm that they do make their scientific production freely available and this strengthens our hypothesis that despite a limited number of repositories, CNR research units are used to provide access to their publications.

Fig. 14. CNR scientific production available on web (%)



In the next question respondents were asked to indicate in which part of the homepage the research unit's scientific production was available. The majority (73,4%) affirmed that this information is contained in the Institute website, in the author's web page (35%) or in the websites dedicated to the description of research activities (24%). Interestingly, the respondents who gave their own free answers (28%) indicated library catalogues and CNR Annual report. (see below).

Fig. 15. Distribution of respondents to the question:
"Who submits data on the scientific production to CNR Annual report?"



Every year each research unit has to send a description of the current research activities together with a list of research results to the CNR central administration. The bibliographic description of CNR publications is available in Internet and contains CNR scientific production since 2002. Although we cannot consider it an IR, this database could be used as a starting point to develop a CNR comprehensive IR .

To analyse who takes part in this procedure (fig. 15), the respondents were asked to indicate the personnel profile and/or services involved in the submission of data about the yearly scientific production. 30% of respondents answered that this activity is carried out by the Secretary or by the local Administration, more than 18% indicates the Library, more than 13% indicates the researchers themselves and 13% the computer service. The involvement of different actors makes it necessary to foresee coordinating efforts to take advantage of the acquired competences that should be reoriented to link the Annual report to a future IR.

Conclusions

The results of the survey confirm that at CNR there is a mature environment for the development of an IR. In fact, despite a limited number of OAI compliant repositories developed under the autonomous initiative of some CNR research units, there is a general propensity to make scientific outputs freely available through various web accessible forms.

Libraries play an important role, both within already developed repositories and in the insertion of the institute's scientific production into digital catalogues and/or in the collection and data input to the CNR Annual report. Moreover, the enlargement of the librarian supporter group, which represented one of the achievements of this survey, demonstrates a growing interest in OA issues. This can envisage their active role in supporting the establishment of a future CNR IR, contributing to the formulation of guidelines for quality and metadata control of the collections to be deposited, integrating local repositories with digital libraries or other external systems, as well as promoting self-submission in their scientific community. Moreover, they can become an important stakeholder that can advocate an official CNR position on OA.

Concerning the other important stakeholder represented by authors, the survey has shown that a meaningful percentage of them self-archives the data on their publications to local repositories, prompts the library to catalogue their scientific outputs, makes them available in personal web pages, and submits them to the central administration for evaluation aims. The CNR Annual report that annually collects CNR scientific outputs can be considered a very important motivational factor not only for authors to deposit their publications, but also as first building block for a future CNR IR. Many research institutions with similar organisation models have build their IRs upon the modernization of their Annual reports [Beier, 2004; Ponsati, 2008] and/or have promoted the same workflow pattern to populate both archives and integrate their data.

However main efforts for the development of a future CNR IR should be focused on the availability of the full text as well as on the integration of already existing information resources and systems. This is not an easy task; it requires planning and design activities, which should take a CNR multidisciplinary nature as well as its complex organisation network of autonomous research units into account foreseeing the participation of the main stakeholders in the process of production, management and dissemination of CNR scientific production.

Bibliography

- Allard Suzie, Mack Thura R., Feltner-Reichert Melanie (2005). "The Librarian's Role in Institutional Repositories: A Content Analysis of the Literature," *Reference Services Review* 33, no. 3): 327.
- Beier Gerhard, Velden Theresa (2004), The eDoc-Server Project: Building an Institutional Repository for the Max Planck Society, HEP Libraries Webzine, 9.
- Bell. S., Fried Foster N., Gibbons S. (2005) Reference librarians and the success of institutional repositories. *Reference services review* 33 (3): 283-290.
- Berlin Declaration (2003), The original Berlin Declaration is located at http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html.
- Biagioni Stefania, Carlesi Carlo, Romano Giuseppe A., Giannini Silvia, Maggi Roberta (2007), PUMA & MetaPub: Open Access to Italian CNR repositories in the Perspective of the European Digital Repository Infrastructure, In: 9th International conference on Grey Literature: Grey foundations in information landscape, Atwerp 10-11 Dec. 2007.
- Crow, R. (2002). The Case for Institutional Repositories: A SPARC Position Paper, *ARL Bimonthly Report,* no. 223.
- Davis Philip M., Connolly Matthew J.L. (2007), Institutional Repositories. Evaluating the reasons for non-use of Cornell University's installation of DSpace, *D-Lib Magazine*, v. 13 (3/4).
- De Castro P, Poltronieri E. (2007), Defining a policy for the institutional repository of the Istituto Superiore di Sanità. *Rapporti ISTISAN* 07(12).
- Guédon Jean-Claude (2002), Open Access Archives: from Scientific Plutocracy to the Republic of Science, In: 68th IFLA Council and General Conference August 18-24, 2002, available at http://www.ifla.org/IV/ifla68/papers/guedon.pdf
- Harnard Stevan (2005), Fast-forward on the Green Road to Open Access: The case against mixing up green and gold, *Ariadne,* 42, available at http://www.ariadne.ac.uk/issue42/harnard/intro.html
- Hunter Philip, Day Michael (2005), "Institutional repositories, aggregator services and collection development", available at: http://www.rdn.ac.uk/projects/eprints-uk/docs/studies/coll-development/coll-development.pdf
- Luzi Daniela (1997), "The Internet as a new distribution channel of scientific Grey Literature: The case of Italian WWW servers", *Publishing Research Quarterly*, 13 (2), 1997.
- Lynch Clifford A.(2003). Institutional Repositories: essential infrastructure for scholarship in the digital age. *ARL Bimonthly Report,* no. 226.
- Lynch Clifford A. (2005). Institutional Repository deployment in the United States as of early 2005. *D-Lib Magazine*, v. 11 (9).
- Lynch Clifford A., Lippincott Joan K., (2005), "Institutional Repositories Deployment in the United States as of early 2005", *D-Lib Magazine*, 11 (9), http://www.dlib.org/dlib/september05/09lynch.html.
- Ponsati A., De Castro P. (2008). Repository increases visibility. *Research information*. Spanish National Research Council (CSIC). Available at: http://www.researchinformation.info/features/feature.php?feature_id=183.
- Rieh Soo Young, Markey Kare, Jean Beth St., Yake Elizabeth, Kim Jihyun (2007), Census of Institutional Repositories in the US. A comparison across institutions at different stages of IR development. *D-Lib Magazine*, v. 13 (11/12).
- ROAR (Registry of Open Access Repositories), http://roar.eprints.org. (last visited December 2008)
- Suber Peter (2008), Open Access News. News form the open access movement, http://www.earlham.edu/~peters/fos/fosblog.html
- van der Graaf Maurits, van Eijndhoven Kwame (2008), The European Repository Landscape, Amsterdam University Press.
- van Westrienen Gerard, Lynch A. Clifford (2005). "Academic Institutional Repositories. Deployment Status in 13 Nations as of Mid 2005", *D-Lib Magazine*, 11 (9), http://www.dlib.org/dlib/september05/westrienen/09westrienen.html.
- Zuber Peter A (2008). A study of Institutional Repository holdings by academic discipline. *D-Lib Magazine,* v. 14 (11/12).

Grey literature in French Digital Repositories: A Survey

Joachim Schöpfel, University of Lille 3, France
Christiane Stock, INIST-CNRS, France

Abstract
The impact of open archives on the availability and selection of scientific and technical information is growing. Yet, there is little empirical evidence on the deposit and processing of grey literature in digital repositories.

The purpose of this communication is to provide a survey on grey literature in French open archives, e.g. institutional and subject-based digital repositories.
The survey is based on a selection of 56 representative French digital repositories. The different archives are selected through national and international registries of OAI repositories, following a defined set of criteria. The repositories are shortly described (type of repository, scientific domain, software, size, language, institution).

Five aspects are analysed for each digital repository:
1. Typology of grey documents (in particular, theses and dissertations, reports, conference proceedings, working papers, courseware).
2. Part of grey literature in the whole archive (in %).
3. Specific metadata related to grey literature.
4. Quality control and policies (evaluation, validation).
5. Conditions of access to the full text.

These information and data are linked to the characteristics of the repositories mentioned above, and specific features of grey literature are discussed.

Furthermore, the question if the New York definition of grey literature applies to the content of digital repositories is discussed.

The communication provides an overview of the preservation and dissemination of grey literature in French digital repositories, contributes to the discovery of French grey literature and open archives, and moves forward the debate on the future of grey literature in the environment of digital repositories.

1. Introduction

"New possibilities of knowledge dissemination (…) through the open access paradigm via the Internet have to be supported. (…) A complete version of the work (…) is deposited (and thus published) in at least one online repository using suitable technical standards (such as the Open Archive definitions) that is supported and maintained by an academic institution, scholarly society, government agency, or other well-established organization." [1]

On 22 October 2003, five years ago, 19 major European scientific organizations signed this *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. In January 2006 the European Commission published the *Study on the Economic and Technical Evolution of the Scientific Publication Markets of Europe* with policy recommendations in favour of open repositories ("Research funding agencies … should promote and support the archiving of publications in open repositories", cf. Dewatripont et al. 2006).

In December 2006, the European Research Advisory Board released a report on scientific publication and policy on open access[2] and recommends, "that the Commission should consider mandating all researchers funded under FP7 to lodge their publications resulting from EC-funded research in an open access repository". A petition for guaranteed public access to publicly-funded research results launched in early 2007 was signed by more than 27,000 scientists and several hundreds organizations[3].

In France 17 scientific and academic institutions support the Berlin Declaration. French universities and research organizations signed in July 2006 an agreement on the development of a common infrastructure of open repositories. Central parts of the French "jigsaw puzzle" (André et al. 2007) are the CNRS Center for Direct Scientific Communication[4] at Lyon and, since November 2008, the institutional repository portal[5] launched by the French academic consortium COUPERIN.

Last year, the European DRIVER study evaluated France as an advanced country in the open archives landscape (see Van de Graaf & Van Eijndhoven 2007).

Grey literature represents a substantial part of the scientific production (cf. Schöpfel & Farace 2009). Since the Seventh International Conference on Grey Literature (Farace & Frantzen 2006) at Nancy, the GreyNet community intensified its research activities on the impact of the open access movement on the grey literature. Special attention was paid to institutional repositories, public policies, organisational context and e-infrastructure. Several case studies highlighted the national, cultural and domain-specific differences.[6] All the same, they also confirmed the force and dynamic of this global movement towards unrestricted access to scientific information.

The purpose of our study is to evaluate the integration of grey literature in French open archives. In particular, five aspects are analysed for each digital repository:

1. The typology of grey documents (e.g., theses and dissertations, reports, conference proceedings, working papers, courseware).
2. The relative part of grey literature in the whole archive.
3. The assignment of specific metadata related to grey literature.
4. Information about quality control and policies.
5. The conditions of access to the full text.

Whenever possible, data on development (evolution of deposit) and usage (statistics of access and downloads) are added. These information and data are linked to the characteristics of the repositories mentioned above, and specific features of grey literature are discussed.

The communication provides an overview of the preservation and dissemination of grey literature in French digital repositories, contributes to the discovery of French grey literature and open archives, and moves forward the debate on the future of grey literature in the environment of digital repositories.

2. Methodology

The survey is based on a selection of 56 representative (e.g. registered either with a dedicated platform or as data provider for harvesting) French digital repositories. The different archives were selected through eight significant international registries of OAI repositories or service providers:

BASE
Bielefeld Academic Search Engine.
http://base.ub.uni-bielefeld.de/index.html

DSpace
Repositories using Dspace – Alphabetical.
http://www.dspace.org/index.php?option=com_content&task=view&id=596&Itemid=182

Eprints
Sites Powered by Eprints.
http://www.eprints.org/software/archives/

OpenDOAR
Directory for Open Access Repositories.
http://www.opendoar.org/

ROAR
Registry of Open Access Repositories.
http://roar.eprints.org/

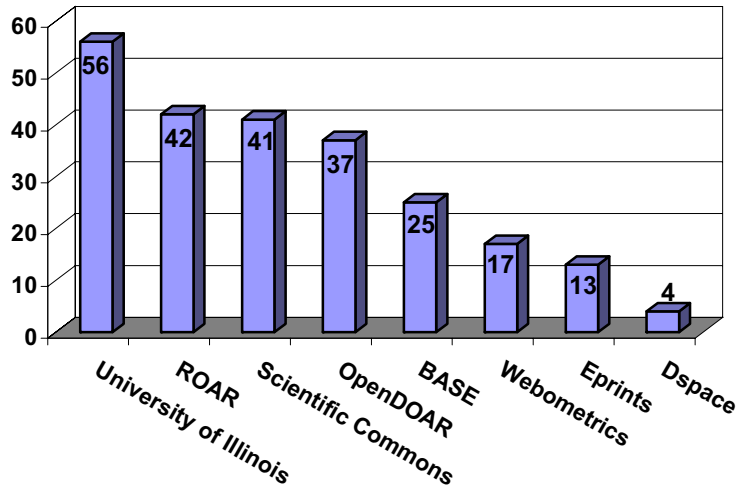
Scientific Commons
Register URL
http://en.scientificcommons.org/register-repository

University of Illinois OAI-PMH Data Provider Registry.
http://gita.grainger.uiuc.edu/registry/

Webometrics
Ranking Web of World Repositories.
http://repositories.webometrics.info/

The selection took place between March and May 2008 and followed a defined set of criteria (located/hosted in France, living archive, size>0). Figure 1 shows for each registry the number of French archives compliant with the criteria.

Figure 1: Number of French archives in international registries (March-May 2008)



Each registered archive (URL) was checked; errors (incorrect URLs etc.) and duplicates were eliminated. Information about the 56 remaining archives were incorporated into a spreadsheet with 37 data columns in 5 categories (see appendix):

1. General (background) information about the archive (10 data elements).
2. Specific information about the archive (6 data elements).
3. Content information (12 data elements).
4. Qualitative data (7 data elements).
5. Comments (2 data elements).

If the information for a specific field was unavailable or uncertain, it remained open.
The data were analyzed with basic Excel statistical functions. Qualitative information was added from the spreadsheet if necessary. Several archives had to be excluded, because the URL was no longer valid or no user interface was provided allowing us to obtain data.

3. Results

The leading questions for the data analysis are:

- How can the current situation of open archives in France be described?
- Which is their content?
- Which is the importance of grey documents in these archives?
- Which are the main aspects of grey material in French open archives?
- How are grey documents used?

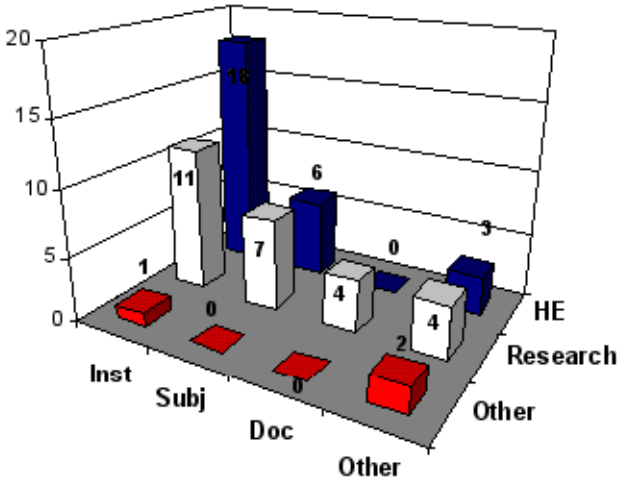
Based on empirical evidence, the following sections try to provide at least partial responses.

3.1. General characteristics of French open repositories

3.1.1. Institutions and typology of archives

One half of the French open archives are owned and/or hosted by Higher Education establishments (HE), e.g. universities and engineering schools, with Strasbourg, Lyon and Paris universities in leading positions. The other half is from public research institutes; mostly from the multidisciplinary national research centre CNRS, some other from INRA (agronomics) or INRIA (applied computer sciences). Only three archives are from other types of organizations.

Figure 2: Institutions and repository typology

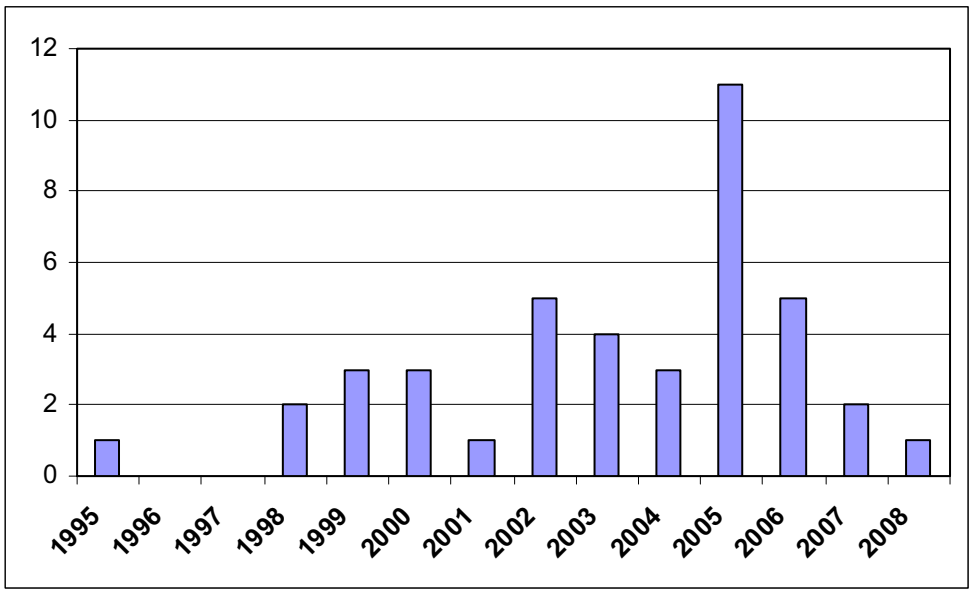


Half of the archives are institutional repositories designed for publications from the scientific authors of the specific institution. In particular, 67% from the HE archives (n=18) are in this category, confirming the academic interest to increase the visibility of scientific production (figure 2).

3.1.2. Date of creation

For 16 repositories, we could not determine the exact date of creation. Most of the others were launched in 2005 or later (figure 3). The figures for 2006 onwards would be even higher, had not HAL been agreed upon as a national repository for French research organizations.

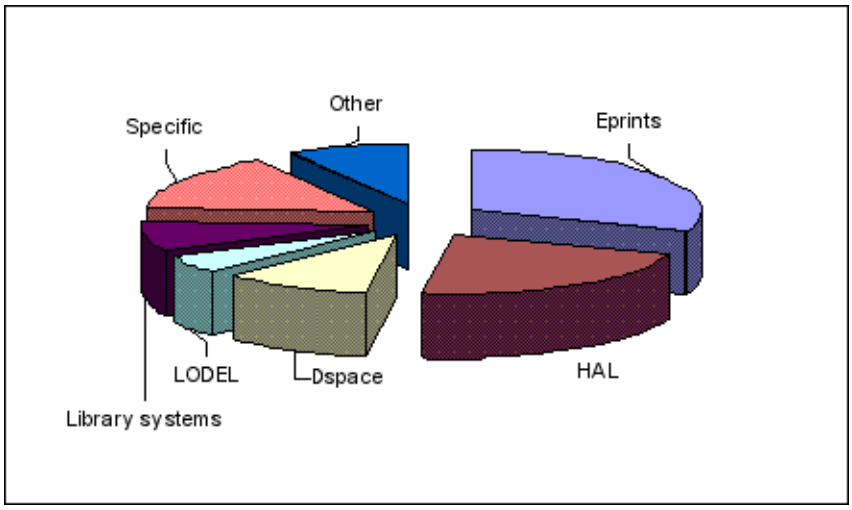
Figure 3: Date of creation of the repositories (updated)



3.1.3. Software

In spite of some early initiatives in favor of national and hegemonic software, the current situation is pluralistic with some major OA-systems and specific (local) solutions.

Figure 4: Software



Two-thirds of the repositories were developed with well-established and OAI-PMH-compliant software, namely *Eprints* (CA/UK), *HAL* (F) and *DSpace* (US). This choice offers the opportunity to collaborate with national and international user groups on problem solving and product development.
This landscape will probably change in the next months. The new French open access software *OAI-ORI*, a specific open source solution for HE institutional archives, was launched earlier this year. Nevertheless, during the period of the empirical study (March-May) *OAI-ORI* was only implemented on experimental sites.

3.1.4. Language

54 repositories provide French-speaking interfaces, 31 of them exclusively. 25 archives supply at least partial English information for users, two of them also German and Spanish information.

Figure 5: Language of interface

Language (interface)	**Number of archives**
French	31
English	2
French and English	21
French, English, German, Spanish	2

The two fully English-speaking repositories are datasets archives, created for and by international scientific communities (astronomy, crystallography).

3.2. Content: scientific domains, types of material and size

3.2.1. Scientific domains

The French open repositories and especially the multidisciplinary and often institutional archives cover most of the scientific disciplines. Nevertheless there are some characteristics of the French open access landscape.

Figure 6: Scientific disciplines

Scientific domain	Number of archives
Multidisciplinary	19
Social sciences & humanities	20
Linguistics	*4*
Library, information & communication sciences	*3*
Ethnology & cultural studies	*3*
Applied sciences	11
Engineering	*5*
Computer sciences	*3*
Agronomics	*3*
Telecommunication	*1*
Sciences, medical sciences	6
Physics	*3*
Astronomy	*2*
Chemistry	*1*
Geochemistry	*1*
Mathematics	*1*
Medicine	*1*
Sport	*1*

Compared to other countries, in particular the US and the United Kingdom, a large important archive for French medical and/or life sciences is missing so far. This probably has two explanations, the importance and force of attraction of the PubMed Central for all international scientists, and the decision of the two French public research organizations with significant research activity in medical and life sciences, the CNRS and INSERM, in favor of a national, multidisciplinary article-based repository (HAL-CCSD).

3.2.2. Types of material

The content of the repositories is widespread and of great diversity. A non-exhaustive inventory based on the repositories descriptions gives evidence for more than 20 different types of materials:

- *books*
- *manuscripts*
- *speech samples with transcriptions*
- *maps*
- *images*
- *datasets (astronomical observation, crystallography)*
- *journals (backfiles, current issues)*
- *articles*
- *proceedings*
- *courseware*
- *posters*
- *videos*
- *patents*
- *dissertations and theses*
- *bibliographical records*
- *reports*
- *preprints*
- *other unpublished materials*
- cultural heritage materials (rare books)
- websites
- *software*

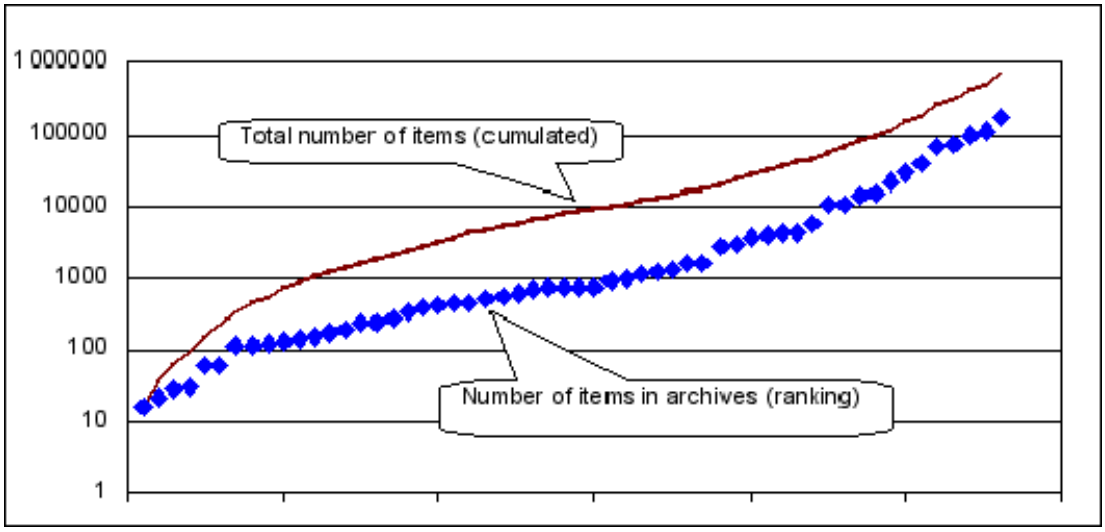
53 archives contain textual material (written documents), seven of them together with other items (datasets, images, maps etc.). Only three repositories don't contain any written document (oral documents and other datasets).

Four archives - all of them produced by the national research organization CNRS - are document-specific, e.g. designed for one specific category of documents and not limited to one institution. The CNRS created a national site for open access journals especially in social sciences and humanities (*revues.org* hosted by CLEO). The other three repositories are dedicated to grey literature: a site for French scientific and technical reports (*LARA* hosted by INIST) and two archives for French electronic theses and dissertations (*TEL* for PhD theses and *MemSIC* for Master theses, both hosted by CCSD).

3.2.3. Size of repositories

The size of the repositories varies largely, between a minimum of 16 items and a maximum of 172,215 items (average size 12,500 items, median size 713 items). Together they total 704,578 deposited items. 32 archives contain less than 1,000 items. Together, they represent 57% of the total number of archives but only 2% of the overall number of items (documents, datasets etc.).

Figure 7: Size of repositories (number of deposited items)



On the other side, 12 archives (21%) contain more than 10,000 items each or 94% of the overall number of items in French archives. These most important archives are the following:

Figure 8: The 12 most important repositories (size)

NAME	HOST	ORGANISATION
TEL	CCSD	CNRS
HAL INRIA	CCSD	CNRS
I-Revues	INIST	CNRS
HAL SHS	CCSD	CNRS
Revues.org	CLEO	CNRS
NumDam	MathDoc	CNRS
GALLICA		BNF
Horizon Pleins Textes		IRD
Crystallography Open Database		University of Maine
ProdINRA	INRA	INRA
HAL	CCSD	CNRS
PERSEE	Lyon 2	University of Lyon 2

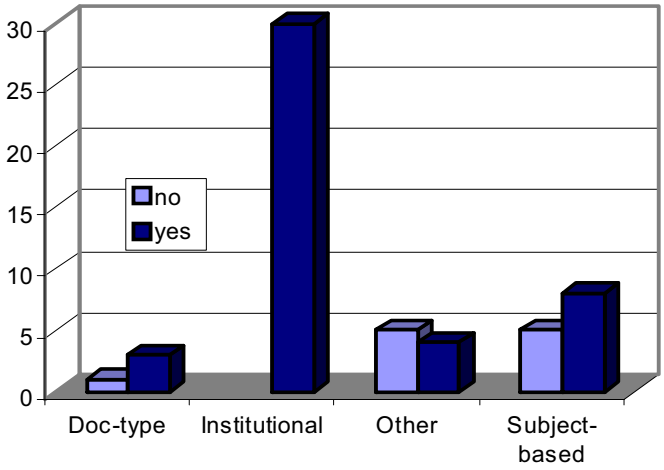
The role of the CNRS is significant; the organization hosts and/or produces more than 30% of the total number of items. However, three of the cited archives provide a mixture of bibliographic records and full text documents (HAL, Horizon Pleins Textes, ProdINRA).

3.3. Grey content

According to OpenDOAR data, 50% of the French repositories contain theses and dissertations, 35% conference or workshop papers and 32% unpublished reports or working papers (October 2008). Reports are frequently associated with journal articles and conference papers, whereas 50 % of repositories containing ETD's (10 out of 20 sites) are dedicated exclusively to this type.

Our own survey shows that a significant part of French repositories (79%) includes at least one category of "traditional" grey literature (theses or dissertations, reports, conferences, working papers, courseware etc.). Even more interesting is the fact that 100% of the institutional archives give access to grey material (figure 9).

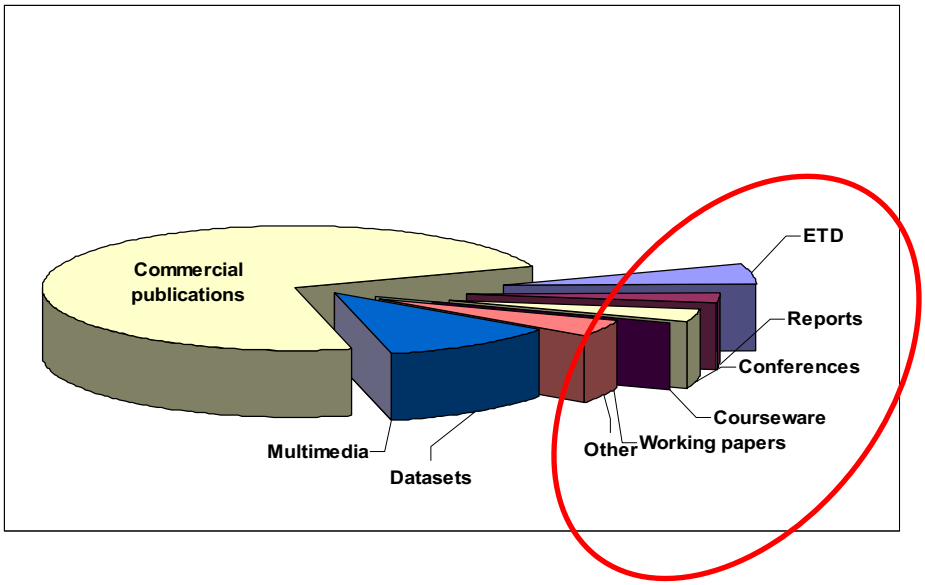
Figure 9: Type of archive and presence of grey content (nb of OA)



18 open archives are 100% grey, e.g. their content is set up by theses and dissertations (14), conference papers (2), reports (1) and courseware (1). Nevertheless, their importance is limited. Together, these "grey OA" contain but 2,5% of all publications in French OA.

The overall part of grey documents (items) in the global French OA content is 16%, e.g. one out of six deposited publications in French archives is grey literature. The other material is commercial (mainly journal articles), multimedia and datasets (figure 10).

Figure 10: Document types (nb of items in OA)



One third of the deposited grey documents are electronic theses and dissertations, followed by conference papers (22%) and reports (16%). Surprisingly low – below 1% - are the indexed deposits of courseware and working papers. On the other hand, the part of undefined grey documents is relative high – 28% (especially in three archives from IRD and CCSD).

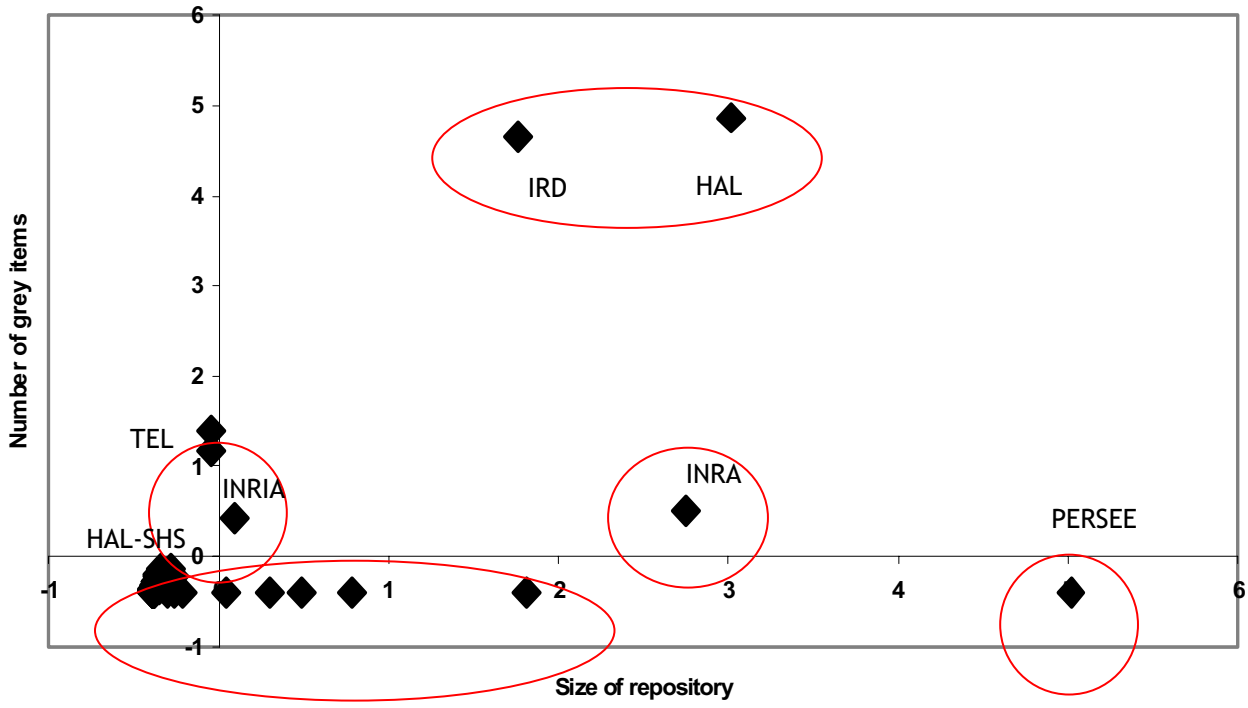
Figure 11: Typology of grey documents

Grey material	Relative part
ETD	34%
Conference papers	22%
Reports	16%
Courseware	0,3%
Working papers	0,2%
Other	28%

Related to the size of the archive and the number of grey items, we can distinguish five types of repositories (fig. 12):

(1) Important archive, no grey material: PERSEE (only journal articles).
(2) Important archive, relative high number of grey items: IRD, HAL.
(3) Important archive, average number of grey items: INRA.
(4) Medium-sized archives, average number of grey items: TEL, HAL-SHS, INRIA.
(5) Smaller archives, no grey content or low number of grey documents.

Figure 12: Size of repository and number of grey items (standard scores)



3.4. Qualitative aspects of grey content in French repositories

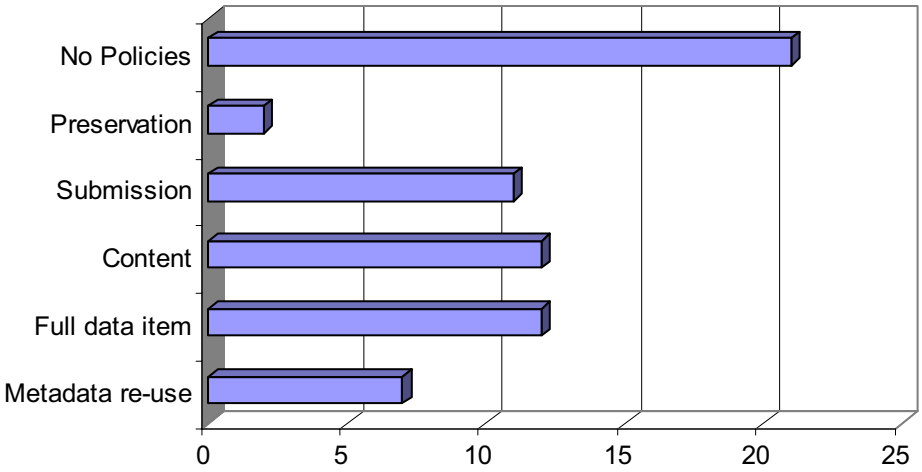
3.4.1. Policy statements

Keith Jeffery in his paper on "Greyscape" (Jeffery 2007) asked whether a repository mentions an "institutional policy to mandate deposition of material". The OpenDOAR registry provides information about policy statements of archives and distinguishes 5 aspects:

- Metadata re-use policy
- Full data item
- Content
- Submission
- Preservation

For the 38 French archives registered with OpenDOAR at the time of our survey 21 sites give no policy statement at all. For the remaining 17 sites we find the following statements (figure 13).

Figure 13: Policies defined by repositories (Source: OpenDOAR)



Policies are expressed in comparable proportions with regards to full data item reuse, content and submission. Metadata re-use is probably implicit for many in the OAI-PMH context, whereas preservation policies are mentioned only twice. Although the majority of the 17 sites make more than one statement, only one repository (OATAO, created in 2007) give information on all 5 issues.
L'Hostis (2006, 23) provides additional information about the mandatory deposit for publications within the major French research organizations. One of the most successful institutes with regards to the submission policy for its research output (effective since 1992, and attaining almost 100%) is the Cemagref institute (agricultural and environmental engineering research). Strangely enough this organization has no visibility as institutional archive whatsoever and doesn't appear in the registries used for our study. Cemagref publications may be accessed through a database (Cemadoc), whenever the full text is available.
CNRS, among the first organizations to sign the Berlin declaration on open access, and operator of HAL, still doesn't oblige it's researchers to submit their documents to HAL.

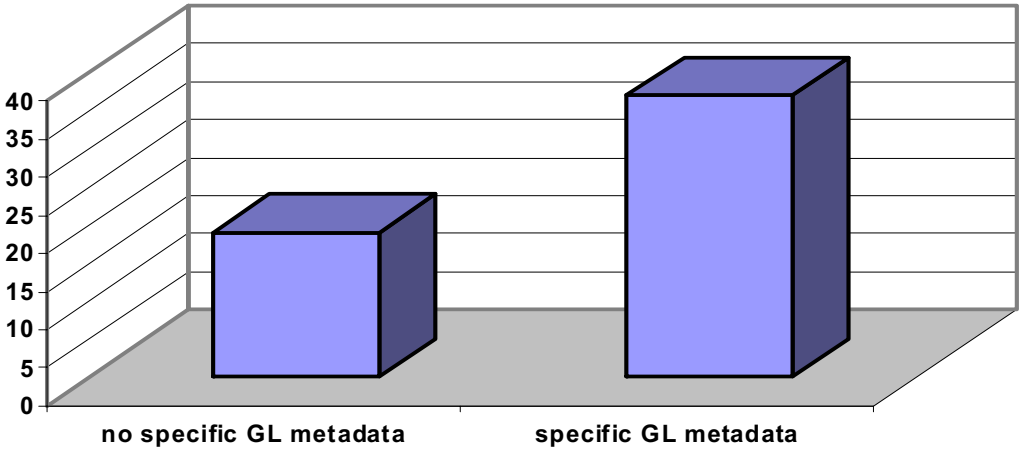
3.4.2. Metadata

Three main grey document types occur in our survey: theses, reports and conference papers. We consider that specific metadata are added when at least one of the following elements is given.

- **Report:** report number, funding organization, project name.
- **Doctoral thesis or dissertation:** defense date, university, degree, discipline, and thesis advisor.
- **Conference:** name, date, and place (town).

Although 45 repositories contain grey documents, only 37 of them add specific metadata. For the remaining 8 archives either the part of grey documents is very low, or they hold particular documents and fall into the category "other" document type.
Among those who add specific metadata, the number and quality of information vary: from adding the name of the university or defense date for a doctoral thesis to the members of the jury or very detailed information for reports on sponsors or projects.

Figure 14: Repositories with specific metadata for grey documents



3.4.3. Access to the full text

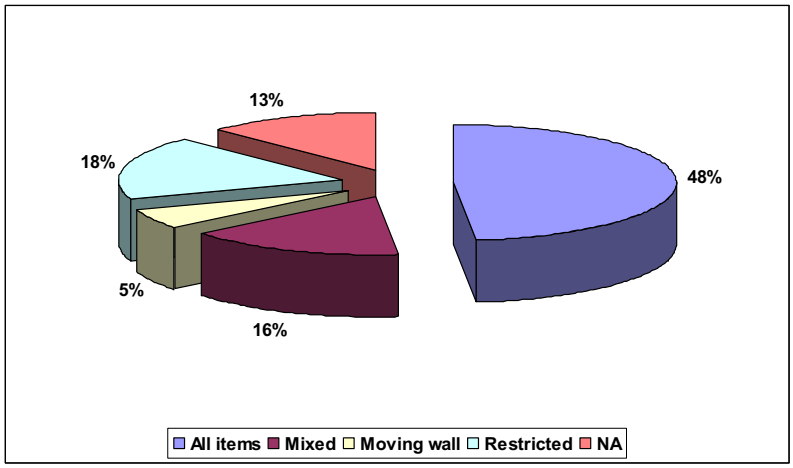
Both ROAR and OpenDOAR registries deal with the distinction between metadata records and access to full text. Data supplied by OpenDOAR refer to full text items only, whereas ROAR gives an estimate on the availability of full text.

In our survey 71% of the repositories in France provide access to full text, and for 48% of the sites the entirety of the documents is available in open access. We distinguish two categories with restrictions:

- Part of the archive is accessible through an intranet or limited to a community.
- A moving wall for commercial repositories. The goal of e-publishing platforms such as I-Revues is to provide access to the full text, but for some titles a temporary embargo is applied.

16% of the archives contain a mixture between bibliographic records and full text. A part of them (e.g. IRD - Research Institute for Development) provide access through a library catalogue, which necessarily includes bibliographic records. ProdINRA currently enhances a bibliographic database to add full text documents. HAL included a publication database of CNRS researchers in its archive. However, it's possible to search for full text records only.

Figure 15: Access to full text



3.4.4. Quality control

Along with the explicit information on the surveyed web sites, about 41% of the repositories mention some kind of quality control and/or evaluation of the archived grey documents.

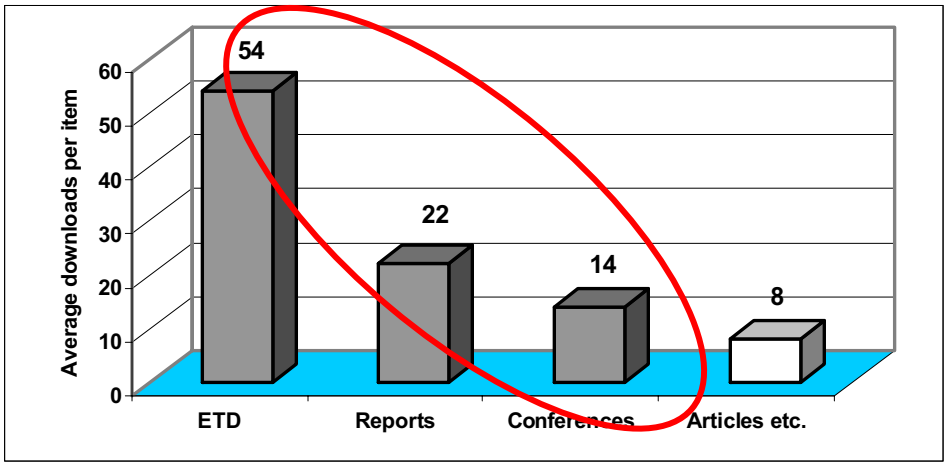
Above all, this quality control concerns electronic theses and dissertations that have been evaluated before their deposit. A few number of archives mention "archive administrators" who obviously act as a kind of scientific editor for the selection (but not for the revision, as far as one can see) of materials. Others only accept peer reviewed or published documents (mostly not grey, however), "outsourcing" by the way the quality control.

3.4.5. Usage statistics

During the period of the survey (March-May 2008), no reliable statistics or other usage related data or information were found on the repositories' web sites.

Shortly before the GL10 conference, we discovered the IFREMER report on the functioning and the usage of the IFREMER institutional archive (Merceur 2008). This report not only publishes the usage statistics (cumulated downloads of archived items) since the creation of the repository (April 2004) but also compares the usage of different types of documents. The result is rather interesting (figure 16).

Figure 16: Usage of different document types in the IFREMER archive (source Merceur 2008)



Even if the IFREMER archive contains two times more white material than grey (e.g. articles, books), the average download per item is up to seven times higher for grey documents, especially for theses and dissertations but also for reports and conferences.

4. Discussion

Difficulties we came across for this survey were numerous, making it sometimes necessary to go into details such as counting items, reading records or even the documents themselves to obtain information. In the following we shall discuss some significant problems.

4.1. Counting items

The overall number of items in a given archive is difficult to define. Data obtained may differ from one source to another depending on what is taken into account: all items, the automatic item count, only full text items, only open access items, items open to harvesting, etc.

On the national level we face the problem of double or triple entries. This situation is similar to other countries; the one and same document may be counted two to four times, because it is included in different repositories. For example a PhD thesis may be submitted to PASTEL and to TEL, then integrated into HAL.

Another confusing situation exists in Toulouse. Several technical universities maintain ETD repositories (INP and INSA and UPS). All their documents can be harvested through a specific website "Toulouse theses", with addition of the items from the veterinary school. OATAO (Open Archive Toulouse Archive Ouverte) is an institutional repository of the recently founded PRES (group of universities and engineering schools), mainly for articles, eprints, but including some ETDS as well. It is planned to have a unique repository for Toulouse in the future.

4.2. Where to find reliable information

Certain information are difficult to obtain. Policies can be expressed anywhere: on the homepage or the "about" page to an article which is deposited in the archive. Administration, validation and quality control of submitted items are often enough part of the back office and difficult to assess from the outside.

4.3. How to identify grey material

Identifying grey content and obtaining quality information, especially reliable numbers, is a real challenge. As mentioned above, it sometimes became necessary to open all items of a given category to control if they are grey. Unfortunately this was not possible for the bigger archives where we know that inconsistencies exist. The earlier archives such as ArchiveSIC and HAL in particular have constantly refined categories (document type, discipline) since their beginnings without always updating existing metadata. Therefore an unknown number of grey documents like reports or conference papers can be found in categories such as 'Miscellaneous" or "other".

4.4. The HAL case

One of the oldest archive in France, HAL became the national archive for scientific and technical organizations in France in 2006. At the same moment, other independent archives such as TEL were integrated into the global HAL, and customized views or portals were created.
As shown in the list below only 25% of the "sub-portals" are referenced in international registries. Most of them led an independent life before 2006. This "fusion" explains the low figures for repository creations from 2006 onwards.

HAL Portals
- Generic
  - *HAL*, *TEL*, CEL
- Thematic
  - *Archive-EduTice*, *HAL-SHS*, *Artxiker*, *@rchiveSIC*, HAL–CSS, HAL-SDE, hprints.org, (*MemSIC*)
- Institutional
  - *HAL-IN2P3*, *INRIA*, *Institut Nicod*, INSERM, UJM, PRUNEL, LIRMM, Académie des Sciences, EMSE, IRD, CIRAD, PASTEUR, OBSPM, Bioemco, INERIS, CEA, INSU, SSA, IRSN, METEO, UNICE, Paris Descartes, Univ-Paris1, MNHN, SUPELEC, UNIV-BREST, UNILIM

This situation may change in the next years with the evolution of independent institutional archives hosted and maintained by the universities themselves.

4.5. Usage statistics

The missing information on usage and access to open archives in France confirms the statement of the 2007 DRIVER study that 70% of the repositories do log the statistical data on access but analysis and interpretation are "in development" or "problematic".
In some cases (CCSD), the depositing authors have dynamic access to «their» statistics, e.g. the figures on hits and downloads of records and full texts. Nevertheless, no global figures are provided.
There may be different explanations: technical problems with software development, conceptual problems with standards, problems related to project planning and priorities, missing capacities for data capture and interpretation, low usage data. Even so, the main problem seems to be the silence, the general lack of any explanation why data are missing or not provided. In a competitive environment where commercial publishers and other vendors provide detailed and standard statistics on usage of journals, databases and e-books, and where the significant public investment in open access creates new business and communication models, public structures can't justify missing information about usage of their repositories.

5. Conclusion

Our survey, even if the dataset remained incomplete for reasons we indicated above, describes a landscape in movement. Pushed by the information market and fostered by new technologies, information services, communication channels and behaviors of scientific communities are undergoing rapid change. The situation of French open archives is changing, and we already mentioned the most important factors of change, e.g. the development of independent institutional archives by the French universities, supported by the academic consortium COUPERIN.
The survey shows how the grey literature takes its place in this environment. The impact of grey material – theses, reports, conferences etc. – in open archives is real and will stay. In the future, the link to new items, multimedia, datasets etc., will need attention and exploration.
On the other hand, the survey reveals three main problems of French open archives, especially in relationship with grey literature:
(1) Policy statements need improvement. Often, the strategy and positioning of repositories are not explicit or simply missing.
(2) Especially grey items in open archives need improved bibliographic control. Compared to traditional cataloguing standards, metadata for grey material are less specific or again, simply missing. This is a problem for referencing, efficient search strategies and evaluation.
(3) Mostly wanted are detailed usage statistics on access and download of documents and other items in open archives.
The survey didn't gather data on the development of the archives (evolution of deposit). This, together with a deeper investigation of usages data, will be the object of a follow-up study in 2009/2010.

6. Bibliography

- André F. et al. "Institutional repositories. The repository jigsaw". *Research Information* 2007, April/May, p. 27.
- Baruch P. "Open Access Developments in France: the HAL Open Archives System". *Learned Publishing* 2007, vol. 20, p. 267-282.
- http://hal.archives-ouvertes.fr/hal-00176428/fr/ DOI : 10.1087/095315107X239636
- Bruley C.; Huet N.; Kalfon J.; Thirionet G. "Bilan d'une enquête sur les archives ouvertes dans les établissements d'enseignement supérieur et de recherche". *AMETIST* 2007, n° 2.
- http://ametist.inist.fr/public/pdf/N2P2A1.pdf
- Correia A.M.R.; Neto M.D. "The role of eprint archives in the access to, and dissemination of, scientific grey literature: LIZA – a case study by the National Library of Portugal". *Journal of Information Science* 2002, vol. 28, n° 3, p. 231-241.
- Davis P.M.; Connolly M.J.L. "Institutional Repositories". *D-Lib Magazine* 2007, vol. 13, n° 3/4.
- http://www.dlib.org/dlib/march07/davis/03davis.html
- Dewatripont M. et al. *Study on the Economic and Technical Evolution of the Scientific Publication Markets of Europe.* Final Report. European Commission, Brussels, 2006.
- http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf
- Farace D.J.; Frantzen J. (ed.) *Seventh International Conference on Grey Literature: Open Access to Grey Resources, 5-6 December 2005*. GreyNet, Grey Literature Network Service. TextRelease, Amsterdam, 2006.
- Jeffery K.; Asserson, A. "Greyscape". In: *GL9 Conference Proceedings. Ninth International Conference on Grey Literature: Grey Foundations in Information Landscape.* Antwerp, 10-11 December 2007.
- L'Hostis D.; Aventurier P. *Archives ouvertes – Vers une obligation de dépôt ? Synthèse sur les réalisations existantes, les pratiques des chercheurs et le rôle des institutions.* Report. 2006.
- http://archivesic.ccsd.cnrs.fr/sic_00115513/fr/
- Lynch C.A. "Institutional repositories: Essential infrastructure for scholarship in the digital age". *ARL Bimonthly Report* 2003, 226, 1-7.
- http://www.arl.org/newsltr/226/ir.html
- Merceur F. *Fonctionnement et usages d'une archive institutionnelle*. IFREMER, October 2008.
- http://www.ifremer.fr/docelec/doc/2008/rapport-4632.pdf
- Schöpfel J.; Farace, D.J. "Grey Literature". In Bates, M.J. & Maack, M.N. (ed): *Encyclopedia of Library and Information Sciences*. 3rd edition. Taylor & Francis 2009 (forthcoming).
- Van de Graaf M.; Van Eijndhoven K. *The European Repository Landscape: Inventory Study into the Present Type and Level of OAI-Compliant Digital Repository Activities in the EU.* Amsterdam, AUP 2007. http://dare.uva.nl/document/93725

7. Appendix

Format of spreadsheet

General (background) information about archive (10 fields)
Name
Acronym
URL
URL alternative
Type institution
Institution
Host
Description
Source description
Creation (yr)
Specific information about archive (6 fields)
Repository Type
Content
Subjects
Software
Language
Size (items)
Content information (12 fields)
Presence GL
ETD
Reports
C-Paper
Proceedings
Working papers
Courseware
Other
Datasets
Multimedia
Total nb GL
% GL
Qualitative data (7 fields)
Policies
Specific metadata GL
Quality control
Evaluation
Validation
Limited access fulltext
Other
Comments (2 fields)
Comments
Date

References

1 http://oa.mpg.de/openaccess-berlin/berlindeclaration.html
2 http://ec.europa.eu/research/eurab/pdf/eurab_scipub_report_recomm_dec06_en.pdf
3 http://www.ec-petition.eu/
4 http://ccsd.cnrs.fr
5 http://www.couperin.org/archivesouvertes/
6 See the GL proceedings at the GreyNet website http://www.greynet.org/greytextarchive.html and the published articles in The Grey Journal, especially the two issues on "Repositories – Home2Grey" (2005, vol. 1, n° 2) and "Grey matters for OAI" (2006, vol. 2, n° 1).

Information Literacy and Librarians' Experiences with Teaching Grey Literature to Medical Students and Healthcare Practitioners

Yongtao Lin, Tom Baker Cancer Centre, University of Calgary, Canada
Marcus Vaska, Health Sciences Library, University of Calgary, Canada

Abstract
The concept of information literacy, which describes the knowledge and skills required in all contexts (i.e. educational sectors, the workplace), as well as in people's everyday lives in today's information rich society, was introduced in the United States in the early 1970s. According to the Association of College and Research Libraries Information Literacy Competency Standards for Higher Education (2000), it has been concluded that an information literate individual is able to determine the extent of information needed, access information efficiently, evaluate information and its sources critically, and use information effectively. Information literacy skills become even more central to meeting the requirements of dealing with complexity and large volumes of information from grey literature.

Our interests as health sciences librarians and thereby the focus of this paper lie in portraying the unstructured nature of grey literature and discussing methodologies and approaches towards teaching this elusive material to those in the health sciences sector, particularly medical students and healthcare practitioners, clients we serve within the Health Information Network Calgary. The Network was formed in 2005 through fee-for-service contracts between the University of Calgary and two partners, the Calgary Health Region and the Alberta Cancer Board. An integrated health knowledge service is provided for healthcare practitioners, staff, patients, and families from Knowledge Centres at major acute care sites, with the University of Calgary Health Sciences Library serving as the Network hub. In both medical school contexts and workplace settings, such as acute care facilities, information literacy is closely associated with the ability to acquire and develop competencies to enable individuals to think critically and use information appropriately.

Giving the end user knowledge related to research information, widening his/her horizons, and implementing critical thinking and carefulness in using information, is more essential than instruction on how to search various information resources. In our own teaching we employ case-based problem-based learning, described by L. Carder, P. Willingham and D. Bibb (2001). We have found this method more effective, active and more student-centered, as it falls in line with a general trend in education, which focuses on making our users independent lifelong learners, and also fits our service goals within the Health Information Network in meeting the needs of medical students and healthcare practitioners.

Keywords: information literacy, gray/grey literature, teaching, problem-based learning

Introduction: The Value of Information Literacy

> "Not having the information you need when you need it leaves you wanting. Not knowing where to look for that information leaves you powerless. In a society where information is king, none of us can afford that." (Lois Horowitz, 2007)

In today's society, there can be little doubt that acquiring the ability to retrieve and make use of information is an essential lifelong skill. Information literacy is indeed the root of information, as individuals need information "in order to achieve educational, social, occupational, and economic goals" (Lloyd & Williamson, 2008, p.3).

While the concept of information literacy has existed since the 1970s, originating in the workplace, (Lloyd & Williamson, 2008) differences characteristically arise when it comes to determining theories or best practices for implementing this term into instructional settings. In our experiences with information literacy at the University of Calgary Health Sciences Library and the Tom Baker Cancer Centre, we have shied away from a lecture-based approach, focusing instead on providing tools for the user (the medical student or healthcare practitioner) to think critically and apply what has been learned in class towards solving his/her own research problems.

Webber and Johnston (2000) define an information literate person as one who is "able to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information" (p.382). We echo this notion because our goal is to allow the user to take control of his/her own learning. Rather than merely memorizing a pattern of search techniques demonstrated in class, retention of material will be better attained if the user is able to apply what has been learned to his/her own studies

in school or in the workplace, "relating literature to life" (Dennis, 2001, p.126). What is gained in the classroom must be transferable to one's occupation, especially in the healthcare setting, as medical students and practitioners alike are faced with the daily task of "developing skills and competencies in order to access and evaluate information, think about information [critically], and demonstrate and document the process of thinking [problem solving]" (Lloyd, 2005, p.83).

The Health Information Network Calgary partnership has allowed us, as librarians of the University of Calgary and the Tom Baker Cancer Centre, to blend our skills and experiences with teaching grey literature, both from a clinical research perspective and from that of a librarian for undergraduate students in the health sciences field, in order to inform a wider audience of the importance of considering the invisible nature of grey literature in research pursuits. We strongly believe that our approach of problem-based learning, where the learner is placed front and centre, and the instructor merely facilitates, allows the main ideas of our lessons to not only come across, but be better retained in the future. Although we are but a small piece of the information literacy puzzle, we feel we are on the right track; grey literature cannot be overlooked if a research endeavor is to be all-encompassing.

In relating our experiences of teaching grey literature to individuals in the medical field, we hope that the researcher's use of a particular piece of information extends beyond a mere paper or presentation. Our aim as information professionals is to supply students and practitioners with the skills necessary to retrieve this information, even for documents that may seem "invisible" at first, skills that we believe will serve our diverse clientele. Bruce (1998) explains that, ideally, "information literacy is seen as working with knowledge and personal perspectives adopted in such a way that new insights are gained" (p.36). Our role as academic librarians and the goal of this paper is to present our views and advice on utilizing problem-based learning as a device to teach the invisible and elusive nature of grey literature to our medical student and healthcare practitioner personnel. We will additionally describe actual class sessions we have instructed, where we provide our clients with the opportunity to work on practice exercises in a group setting, encouraging the sharing of feedback among all members of the class and beyond (Dennis, 2001). Achieving success with this method depends on collaboration between the learner and the facilitator, an active pedagogic role that we see ourselves as playing. We are true believers that information literacy doesn't end with finding relevant information in the moment; it is a continuum.

The Impact of Grey Literature on Information Literacy

Researchers in the education sector propose that information literacy "is recognized as making an important contribution to decision-making, problem-solving, independent learning, continuing professional development, and research" (Bruce, 1998, p.25). Teaching and encouraging critical thinking and problem solving skills is a key ingredient that librarians follow to provide the opportunity for the learner to increase retention of what has been presented outside the classroom setting (Lloyd, 2005). Nevertheless, there is a considerable gap in the literature discussing the needs of the information literate individual alongside the teaching experiences of seasoned librarians, particularly in the field of grey literature. In adhering to Warmkessel's (1997) principle, teaching grey literature involves guiding the user to the correct sources of information, providing advice along the way, yet exercising restraint and not doing the researcher's research for him/her.

Although numerous types of grey literature are available electronically, they are often not indexed, and, without a librarian to serve as a guide, many individuals become frustrated at the lack of congruity in the invisible deep web (Alpi, 2005). Regardless of the form that a grey literature document may take, as health sciences librarians, and thereby proficient searchers in medical databases, we must be able to search beyond the obvious online resources, overcome limitations, and assist medical students and healthcare practitioners in locating grey literature on a particular topic, thereby satisfying the quest for information.

Were it not for the leaps and bounds that the world of technology has taken over the past decade, many government documents, organizational reports, and informal day-to-day communications would remain hidden. Teaching the medical student or healthcare practitioner how to use technology to locate and retrieve a desired piece of grey literature is the first crucial step on the path to becoming an enlightened information literate individual. Once the ability to use the technology has been achieved, searching for material not available through ordinary publishing channels can be a daunting task, unless information specialists instruct and guide clients with regards to the best available sources in which to find that elusive piece of grey literature.

In discussing the impact of grey literature on information literacy, we strive to remind our audience that researchers should include grey literature and published material side-by-side whenever they conduct a search on a topic, particularly when working on systematic reviews, in order to limit bias. If the theories and goals of information literacy are to remain true, in both a classroom setting and in the workplace, then our goal as information practitioners is to inform our audience that grey literature is a part of one's

daily life. In order to be accepted as a true information literate individual, one must not discount rapidly produced material, in various formats, with limited distribution, that is often not peer-reviewed. Research has shown that the notion of grey literature actually existed long before the term "information literacy" was coined, and has had a tremendous impact on critical thinking and the information search process in the real world.

Studies over the past decade have indicated that grey literature resources are expanding at a much higher rate than standard published documents (Coad et al., 2006). As with virtually any information-seeking process, a method or rationale needs to be developed to explain why searching for a particular information source is relevant. So too is the case of teaching the illusive nature of grey literature to library users. If our goal as information professionals is to supply our clientele with all of the skills necessary to experiment with various search tools and critically analyze what they have found, then the impact of grey literature on information literacy is evident. How else can we expect our medical students and healthcare practitioners to become information literate and proficient if we neglect to address tips and techniques to locate conference information, recent publications not yet cited in databases, or the personal communication of experts in a particular field? (Benzies et al., 2006)

When the authors of this paper first began teaching grey literature to their clientele, they were aware of remarks by some educators who claimed that grey literature should not be taught as a component of information literacy, since officially, it was not considered to be a scholarly form of publication. Despite these objections, the importance of grey literature cannot be overlooked. Grey literature exploded on the scene in the past decade, due in large part to the Internet. As a result, interest in grey literature has grown, and previously hidden documents can now be more easily retrieved, allowing the possibility to instantaneously access useful information in both the physical and virtual world.

Our goal in teaching grey literature to our audience of learners is to allow the client to think critically and analyze information that has been retrieved. Even when searching within an organization's website, or hand-searching through government documents, it is unlikely that all of the material located will be relevant to the search topic. Whether information is required to complete an assignment or treat a patient's rare illness, there is always a purpose to searching; grey literature is another avenue to make that search more effective.

Information Literacy and Grey Literature in a Medical Context

While medical students, and in particular, healthcare practitioners, may be real gurus of knowledge in their area of specialty, they may not all have the skills and abilities necessary to "find, select, organize, and use relevant information sources" (Macklin, 2001, p. 306). Librarians must therefore convince the user that understanding how to find information on a particular topic is just as meaningful as possessing knowledge about it.

The bond between information literacy and lifelong learning cannot be discarded, especially in a field as diverse and challenging as the medical sector. The medical student or healthcare professional who is able to think critically, actively, and independently, to solve or mediate the complex issues that he/she will inevitably encounter throughout a career, often under a great deal of pressure, is truly an information literate human being.

As previously mentioned, despite being introduced in the 1970s, information literacy was not recognized in a medical setting until nearly two decades later. First used in 1989 in the nursing curriculum at the University of Northern Colorado, it is interesting to reflect back, nearly 20 years afterwards, to what was first envisioned by the notion of information literacy in instruction, ideas that the authors of this paper advocate in their grey literature sessions today: "assist…in understanding library organization and services; promote…skills in evaluating and locating and evaluating the accuracy of information; help…understand information seeking strategies and appropriate use of those strategies" (Cheek & Doskatsch, 1998, p.246) (Fox et al., 1989, p. 423).

In the world of healthcare, finding accurate and reliable information, including grey literature, in a timely manner is essential. A preliminary survey evaluating the use of grey literature by end users in the health sciences was described in a paper published in 1990 (Alberani et. al., 1990). References in selected medical journals and databases analyzed the years 1987-1988 to determine the number of grey literature citations in selected health sciences journals, the various types of grey literature found, and the number of technical reports cited, including country of origin and intergovernmental issuing organizations. The results obtained showed that as a primary source of information, grey literature was cited in those journals that provided reliable data on research in progress. Technical reports prevailed over other types of grey literature, thus confirming the attention given to this type of material in health sciences research efforts.

Public health material takes a variety of formats (both print or electronic), in order to find the best possible solution to a medical dilemma. Alpi (2005) explores the characteristics of public health information needs and the resources available to meet those needs. Important characteristics in public health expert searching are demonstrated by the ability to identify and search for resources beyond the electronically available published literature, including older published literature, grey literature, unpublished information, and documents on the Web. Grey literature is often the only information available on a Public Health topic from a particular perspective. There have also been debates over the inclusion of grey literature in meta-analyses over the past decades. Some critics of grey literature have suggested that meta-analyses should not give unpublished literature the same weight as published studies (Sacks et al., 1996). However, excluding grey literature limits reviews to only a portion of the available evidence, which may result in overestimating effect sizes. (McAuley et. al., 2000). Hopewell (2007) also indicates that it is necessary to include grey literature for certain types of research, such as systematic reviews, since "… published trials tend to be larger and show an overall greater treatment effect than grey trials. This has important implications for reviewers who need to ensure they identify grey trials, in order to minimize the risk of introducing bias into their review." (p. 2). While many voice support for the inclusion of grey literature, most published meta-analyses do not take this information resource into account (McAuley et. al., 2000), an exclusion that may result in difficulties obtaining grey literature studies. As information professionals, this is our opportunity to address grey literature challenges in meta analyses by helping information investigators locate diverse findings and make their own research accessible to other scientists and reviewers.

Besides locating information and assisting with the indexing and cataloguing of grey literature to make this material accessible to researchers, health science librarians can take a much more proactive role in teaching end users the skills of finding and evaluating grey literature. Revere (2007) also emphasizes the need of holding grey literature instruction sessions, particularly in healthcare: "a limited amount of critical information is published through standard channels…finding a resource, let alone locating the answer to a question within a resource, is extraordinarily difficult…" (p. 411). This is where our role as information professionals comes in; we must conquer the information barrier to critical sources that are out there. Developing methods for searching grey literature to uncover the 'deep web' will present medical students and healthcare practitioners with a complete picture to solving medical mysteries. Even though a previously hidden document may become 'visible' and accessible, the source, relevance, and quality of the item must be taken into account before a citation is deemed as appropriate (Revere et al., 2007). Some of the valuable search strategies we incorporate in our teaching sessions include examination of multiple medical databases including online resources for theses and dissertations, Cochrane Library sources, and clinical trials, citation of index searches, examination of research registries, journal hand searches, contact with the organizations and associations, examination of presentation abstracts and conference proceedings, and Internet searches.

The "discerning information consumer" (Cheek & Doskatsch, 1998, p.243), one who is able to adapt new technology to gain knowledge to seek out information required in his/her profession, is prevalent in the medical industry. Medical students and healthcare practitioners are required to learn throughout their lifelong career, adapting the skills acquired to retrieve information according to various circumstances and needs. As Cheek and Doskatsch (1998) further explain, our goal as information professionals is to nurture and motivate those in the healthcare industry to want to continue learning and maintain their status as information literate individuals.

The medical industry is constantly embroiled in change, forcing all practitioners working in this field to keep abreast of new procedures, experiments, and treatments in the literature to help cure or alleviate a variety of illnesses. "Reading professional literature is a critical facet of lifelong learning" (Cheek & Doskatsch, 1998, p.245), nevertheless, time constraints often make it difficult for healthcare practitioners to be on top of all the latest developments. As educators of information literacy and the grey literature, we must remain aware of the latest information available, overcoming the challenge by reminding our learners that turning towards material not published in a scholarly fashion in large, commercial journals, is not necessarily a negative choice. Considering that it can take months for a study to become published, seeking material that may already be available on an association's website will ensure that the information is there when it is needed, keeping the chain of continuous lifelong learning intact.

Lifelong learning is a trait that will always be present in the healthcare industry. The authors of this paper believe that library instruction sessions on information literacy and the grey literature should be taught throughout one's entire medical career, instead of a one-time need-only basis. Studies in the field of nursing have indicated that information rises exponentially on a yearly basis (Barnard et al., 2005). If we apply this finding to the health sciences in general, it cannot be stressed enough that it is indeed a learned trait, one that takes a great deal of skill and practice, to be able to sort through literature, especially that of the grey variety, and locate relevant and authentic documentation on a specific field of

study. Unfortunately, even in today's day and age, access to some forms of literature can be a daunting task, despite the universally accepted idea that access to information is essential.

Using Problem-Based Learning for Teaching Grey Literature in Healthcare Settings

Healthcare reform, rapidly changing medical information, and technological advances have changed medical education from traditional lecture-centered to a more learner-driven, self-directed, problem-based approach. Problem-based learning (PBL) is a teaching technique used in many medical schools to facilitate learning basic science concepts in the context of clinical cases (Boud & Feletti, 1997). Over the past two decades, PBL has become increasingly integrated into undergraduate medical education in North America. (Donner et al., 1991). In the problem-based approach, complex, real-world issues are used to motivate students to identify and research the concepts and principles they need to know to solve these problems. Students work in small learning groups, bringing together collective skill at selecting and identifying the problem, acquiring sources of information, and communicating and integrating that information. Outcomes for students in PBL include: the capability to think critically to analyze and solve complex real-world problems, the ability to find, evaluate, and use appropriate learning resources, the ability to work collaboratively in teams and small groups, the likelihood of applying content to real-life work situations following medical schools, and developing skills for life-long learning. These learning outcomes are congruent with information competency standards, which form the basics of life-long learning. The Information Literacy Competency Standards for Higher Education (ALA, 2000) outlines that "an information literate individual is able to determine the extent of information needed, access the needed information effectively and efficiently, evaluate information and its sources critically, incorporate selected information into one's knowledge base, use information effectively to accomplish a specific purpose, understand the economic, legal, and social issues surrounding the use of information, and access and use information ethically and legally" (pp. 2-3).

PBL focuses on the learner's own ability to find relevant information. The main actor is the self-supporting learner, not the lecturer. Although information literacy is seldom expressed as a main objective, it is an intrinsic concept and learners are expected to turn out as skilled, independent library and information users. Essential to understand is the library's continuing and ever-increasing role as an important learning resource in such an educational program. A study conducted jointly at several universities analyzing the effect of PBL on library services and educational programming (Watkins, 1993), found that the PBL student, in comparison to the student in a conventional curriculum, uses library resources and reference services more frequently, and requires more library training in information-seeking skills to support self-learning.

Since information literacy is one of the primary goals of problem-based learning, it offers important new roles as well as challenges for libraries, changing library user education in the process. The PBL approach promotes critical and analytical thinking skills by applying the learner's own expertise and experience to the problem-solving and information retrieval process. The effectiveness of using PBL in library instruction sessions has been well documented in the library literature. Schilling (1995) describes a successful library program in a medical school course where librarians acted as information facilitators along with other faculty coordinators to help students find, use, manage, and evaluate appropriate information resources. Describing the difference between a PBL and traditional instruction session, Cheney (2004) states that more student-instructor-librarian interaction evokes a clear sense of "purpose on the part of the students – they knew exactly how the library session was tied to their research needs" (p. 505). Given the nature of grey literature, PBL provides the theoretical framework for a learner-centered, active instructional experience that relies on critical evaluation of the "grey" sources of information and hands-on interaction with grey literature material.

The process of applying PBL towards information literacy instruction within the constraints of a one-shot session has been effectively described by both Enger (2002) and Macklin (2001), while Kenney (2007) discusses challenges of modifying PBL instruction from its classic curriculum form to a fifty or eighty-minute session. As PBL is student-centered and focused on the whole process of learning, clear boundaries need to be established through lesson planning, since a well-structured lesson plan also provides the guidance needed when delivering instruction. Macklin (2001) provides explicit directions for creating a lesson plan with "specific learning goals and objectives that relate to information literacy" (p. 310). For example, the learning goals of our Grey Literature session are to: introduce learners to the history of grey literature, familiarize them with major grey literature health resources and search techniques, and educate them with evaluation criteria for grey literature material.

The norm of PBL teaching always starts with the facilitator introducing a problem, followed by the learner analyzing the problem for information, both individually and collaboratively. An information need is then determined and possible solutions are proposed. Afterwards, the facilitator introduces various information sources and learners investigate these sources to locate relevant material for solutions to the problem being presented (Macklin, 2001). PBL has been shown to be effective in engaging students in learning,

particularly when the problem is well developed. A well-designed, adaptable problem is an effective vehicle for learners to relate the subject matter to the real world, practicing one or two specific learning outcomes. A small group discussion question in our grey literature session was "What are the grey literature resources/channels you have used, in what way, and why?" This initial question was based on previously learned knowledge in order that all learners in the groups are drawn into a discussion of grey literature. This keeps the learners functioning as a cohesive whole and also helps them build connections to previously learned concepts and material. At the next stage of problem-setting, we had a different scenario for a diverse learner group. For instance, we asked researchers and research associates from the Department of Public Health to compile a list of resources in communication disorders from the grey literature. As another example, the Prevention Unit was given the task of critically looking at an unpublished article on indoor tanning and skin cancer, and gathering evidence for a systematic review. These problems are placed in a context familiar to the learners, requiring them to make decisions or judgments based on facts, information or logic. The process of thinking through the learning goals of a course and writing realistic problems to meet those learning priorities have changed how we, as instructors, view information literacy being integrated using problem-based learning.

An additional challenge outlined in Kenney's (2007) review is keeping a controlled schedule, particularly for teaching the "unstructured" nature of grey literature. Some of the questions we have asked ourselves when preparing our grey literature classes are: what resources will be used and which concepts will be emphasized?; what part of the research process do we want to stress?; when we have only 80 minutes, is everything important?; what can we leave out?; how comfortable are we with grey literature debriefing since there is usually no single perfect answer to a problem?

Improving one's knowledge of a topic by delving into the world of grey literature can certainly be a rewarding experience. Nevertheless, as problem-based studies have shown, when it comes to educating clients, "how someone understands or experiences information literacy is of greater importance to determining what they have learned, than how much knowledge or skill they can demonstrate" (Bruce, 1998, p.40). To place this in another light, the authors of this paper believe that the use of grey literature in information literacy results in clients learning more by taking control of their own learning. As the activity becomes user-centered, keeping the schedule well-structured is essential to drive collaboration and provide the class with clear-cut parameters and expectations. Almost all of the work of a PBL session is in the design, preparation and planning before class. Providing handouts that support the activity, such as evaluation criteria, and worksheets for group discussion, enables the learner to stay on the task and leave with the information required for future research endeavors.

One of the key principles for PBL is that learning occurs in groups, often two or three learners, a size that seems to work best in this environment. PBL learners continuously engage in discussion; they coordinate, manage, and direct themselves in order to solve problems. Nevertheless, creating teams poses yet another challenge in PBL, as participants may have different comfort levels using computers and electronic resources. Clear guidance needs to be provided for a breakdown of activities and to encourage participation from every team member. Even in a short, eighty-minute session, learners establish a bond with their team members, motivating them to seek out the best solution.

Perhaps the most challenging part of a PBL session is the information professional assuming the role of facilitator and coordinator, rather than lecturer as is the case in traditional library instruction. Macklin (2001) elaborates on this concept further by stating: "a successful PBL facilitator must be able to draw out this evolving expertise by establishing a learning environment that is conductive to exploration, creative thinking and continuous positive feedback and reinforcement" (p.309). Actively engaging a PBL session can be a rewarding experience as it provides an excellent opportunity to interact with learners in a more dynamic environment. As learning supporters and advocates, the authors of this paper set up the problem/situation at the beginning of the session, help learners focus their thinking as they explore grey literature resources, balance learner-direction with assistance, stimulate critical evaluation of ideas, and encourage users to make informed choices in their information seeking.

Conclusion: Continuing with PBL Today and into the Future

One of the satisfactions that we as instructors have from using PBL in teaching grey literature is that we get to know our clients better than we do with a traditional library lecture. By shifting ourselves from the action centre, we get to observe and listen to our users during the class. This allows us to gain insight into the structure of medical students and healthcare practitioners' understanding of grey literature, providing the experience in learning a topic on one's own while applying it to real-world situations, the core principle of information literacy. Successful PBL involves more than just repackaging the content of traditional lectures into problem-based scenarios; it requires a transformation of learners' experiences, concerns and visions. Although the experience is rewarding, the learning curve is steep; we must be prepared to deal with the changing dynamics of PBL, our users' learning skills, and the subject matter

being searched for. Nevertheless, the change in perspective of grey literature and PBL will lead us to renewed interest for providing user-education in the health sciences sector.

References

- Alberani, V., Pietrangeli, P., & Mazza, A. (1990). The use of grey literature in health sciences: A preliminary survey. Bulletin of the Medical Library Association, 78(4), 358-363.
- Alpi, K. M. (2005). Expert searching in public health. Journal of the Medical Library Association, 93(1), 97-103.
- Association of College and Research Libraries (ACRL). (2000). Information literacy competency standards for higher education. Chicago, IL: American Library Association.
- Barnard, A., Nash, R., & O'Brien, M. (2005). Information literacy: Developing lifelong skills through nursing education. Journal of Nursing Education, 44(11), 505-510.
- Benzies, K. M., Premji, S., Hayden, K. A., & Serrett, K. (2006). State-of-the-evidence reviews: Advantages and challenges of including grey literature. Worldviews on Evidence-Based Nursing / Sigma Theta Tau International, Honor Society of Nursing, 3(2), 55-61.
- Boud, D., & Feletti, G. (Eds.). (1997). The challenge of problem-based learning (2nd ed.). London: Kogan.
- Bruce, C. S. (1998). The phenomenon of information literacy. Higher Education Research and Development, 17(1), 24-43.
- Carder, L., Willingham, P., & Bibb, D. (2001). Case-based, problem-based learning. Information literacy for the real world. Research Strategies, 18(3), 181-190.
- Cheek, J., & Doskatsch, I. (1998). Information literacy: A resource for nurses as lifelong learners. Nurse Education Today, 18(3), 243-250.
- Cheney, D. (2004). Problem-based learning: Librarians as collaborators. Libraries and Academy, 4 (4), 495-508.
- Coad, J., Hardicre, J., & Devitt, P. (2006). How to search for and use 'grey literature' in research. Nursing Times, 102(50), 35-36.
- Dennis, N. (2001). Using inquiry methods to foster information literacy partnerships. Reference Services Review, 29(2), 122-131.
- Donner R.S. & Bickley, H. (1991). Problem-based learning in American medical education: an overview. Bulletin of the Medical Library Association, 81 (3), 294-298.
- Enger, K.B., Brenenson, S., Lenn, K., MacMillan, M., Meisart, M.F., Meserve, H., & Vella, S.A. (2002). Problem-based learning: evolving strategies and conversations for library instruction. Reference Services Review, 30(4), 355-358.
- Fox, L., Richer, J., & White, N. (1989). Pathways to information literacy. Journal of Nursing Education, 28(9), 422-425.
- Hopewell, S., McDonald, S., Clarke, M., & Egger, M. (2007). Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database of Systematic Reviews, 2, MR000010.
- Kenney, B. F. (2007). Revitalizing the one-shot instruction session using problem-based learning. Reference & User Services Quarterly, 47(4; 4), 386-391.
- Lloyd, A. (2005). Information literacy: Different concepts, different truths? . Journal of Librarianship and Information Science, 37(2), 82-88.
- Lloyd, A., & Williamson, K. (2008). Towards an understanding of information literacy in context: Implications for research. Journal of Librarianship and Information Science, 40(1), 3-12.
- Macklin, A. S. (2001). Integrating information literacy using problem-based learning. Reference Services Review, 29(4), 306-313.
- McAuley, L., Pham, Ba'., Tugwell, P., & Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? The Lancet, 356(9237), 1228-1231.
- Revere, D., Turner, A. M., Madhavan, A., Rambo, N., Bugni, P. F., Kimball, A., et al. (2007). Understanding the information needs of public health practitioners: A literature review to inform design of an interactive digital knowledge management system. Journal of Biomedical Informatics, 40(4), 410-421.
- Sacks, H. S., Reitman, D., Pagano, D., & Kupelnick, B. (1996). Meta-analysis: An update. Mount Sinai Journal of Medicine, 63, 216-224.
- Schilling, K., Ginn, D.S., Mickelson, P., & Hoth, L.H. (1995). Integration of information-seeking skills and activities into a problem-based curriculum. Bulletin of the Medical Library Association, 83(2), 176-183.
- Watkins, M. C. (1993). Characteristics of services and educational programs in libraries serving problem-based curricula: A group self-study. Bulletin of the Medical Library Association, 81(3), 306-309.
- Warmkessel, M. M., & McCade, J. M. (1997). Integrating information literacy into the curriculum. Research Strategies, 15(2), 80-88.
- Webber, S., & Johnston, B. (2000). Conceptions of information literacy: New perspectives and implications. Journal of Information Science, 26(6), 381-397.

Grey Literature and Development:
The Non-Governmental Organization in Action

Lynne Marie Rudasill, University of Illinois, United States

Abstract
Traditionally, the non-governmental organization working in the area of development has been viewed as a trusted source for research and information on specific topics and populations. With the advent of the World Wide Web, many of these organizations are working to make their expertise available to a large number of users. This preliminary study surveys non-governmental organizations working in several areas of health-related activity to ascertain what types of information they are making available on the Web. What types of grey literature are being made available electronically by these organizations? What resources are being used to disseminate this literature? In addition to reviewing the types of information extant on these sites, we will compare a sample of the contents of the websites to the WorldCat database to see what is, and is not, part of the traditional dissemination system.

This preliminary study is part of a larger survey that looks at the electronic production, collection, and dissemination of information by non-governmental organizations in developing and developed nations. Resources such as press/news releases, reports, books, bulletins, and journals are calculated for each organization at individual websites.

It is expected that the World Wide Web now serves the purpose of the traditional vertical file in which print copies of non-governmental organization materials used to be placed. It is also assumed that the materials on the Web serve only as the tip of the iceberg as far as the production of information by the non-governmental organization. The impact this might have on the organization's ability to serve as a point of expertise in policy decisions is explored.

[The author wishes to acknowledge the Research and Publication Committee of the University of Illinois at Urbana-Champaign Library, which provided support for this research.]

Introduction

The growth of the non-governmental organization (NGO) as an actor in the policy and information arena has been rapid over the course of the last century and a half. At the time the United Nations was chartered, forty-four NGOs were recognized as consultative organizations with the Economic and Social Council (ECOSOC). Currently over 1600 NGOs are associated with the U.N. Department of Public Information and ECOSOC.[1] The NGOs are regularly consulted about areas related to their work by various agencies of the United Nations for several reasons. These civil society organizations are often recognized for their expertise in the topical areas in which they work. They are the groups that are "on the ground" with the problems they address and they usually have a deep understanding of the causes and solutions to these problems. Although these organizations are generally recognized as representing a variety of societal problems that need to be addressed, they usually have few, if any, apparent ties to either commerce or government that might influence their fact-finding. This provides the organization with a high degree of credibility in their respective areas. These organizations also have very effective networks. Frequently the NGO has to engage in public relations, advocacy, and rationale for funding support. This type of outreach creates many channels for the sharing of ideas and information and also supports the accountability for the organization in the eyes of others.

One of the ways the NGO can achieve this networking is to support a web site. The web site serves many purposes reaching out to the target population, the press, the government, and the donor.[2] The sites are as varied as the organizations, and the richness of the website is tied to the support the organization enjoys. The sites also are somewhat representative of the digital divide because in developing nations the expense of hosting for a site, as well as the price for the expertise needed to maintain a site, is often higher than the organization can afford. In related research including the exploration of various directories for NGOs, we have found that the percentage of organizations with a web presence is much greater in the developed nation. But what does this mean in the creation and capture of grey literature by any of these organizations?

A wide variety of literature is available concerning the NGO and information dissemination. A large part of the writing revolves around how NGOs use and disseminate information and the difficulties they encounter in fulfilling the target audience needs in a timely and efficient manner.[3] One of the most useful advances in the collection of grey literature by non-profit or non-governmental organizations is the *Grey Literature Report* of the New York Academy of Medicine.[4] This particular resource is specifically targeted

at health and public health resources, and does an excellent job of covering the grey literature produced by a large number of organizations.

Scope and Method

The larger project, still under investigation, includes non-governmental organizations in several countries chosen specifically to contrast eastern and western cultures and developing and developed nations. For the purpose of this paper, the organization web sites reviewed were recognized generally as "health" organizations. Related areas of concern for these organizations included information concerning HIV/AIDS, reproductive rights, women and children, sanitation, and malaria and other tropical diseases. The majority of the organizations were drawn from those listed in the Non-Governmental Organizations Search at the Jonsson Library of Stanford University. These organizations "were chosen based on their consultative status with the United Nations Economic and Social Council (ECOSOC) and also collated from University of Minnesota Human Rights Library, Duke University Libraries' NGO Research Guide, and the World Association of Non-Governmental Organizations (WANGO)."[5] The search term was simply "health." The organizational web sites reviewed were also generated from the development directories produced by the DevDir organization.[6] The sites were closely reviewed to determine the specific material types available from these resources. In addition, each of the organizations was searched as a publisher in the WorldCat database.

The structure of disclosure is closely related to the structure of literature in the information sciences, but reflects the degree to which literature appears in the cycle of publication. The various stages of this disclosure include the following. First disclosure is the point at which the information appears as primary literature, a book, a thesis, an article, or a presentation of some sort in print or on a website. Secondary disclosure includes reviews of the literature, bibliographies, indexes, abstracts, summaries, digest, and guides to the literature. Tertiary disclosure includes the publication of the information in the form of reference works and the existence of information as a popular synthesis of knowledge. The grey literature of the health-related organization that is found on websites might fit under all of these events, but is most commonly found as primary literature. The World Wide Web and Web 2.0 provide a number of less conventional means of dissemination in the form of blogs, podcasts, and other types of social networking.

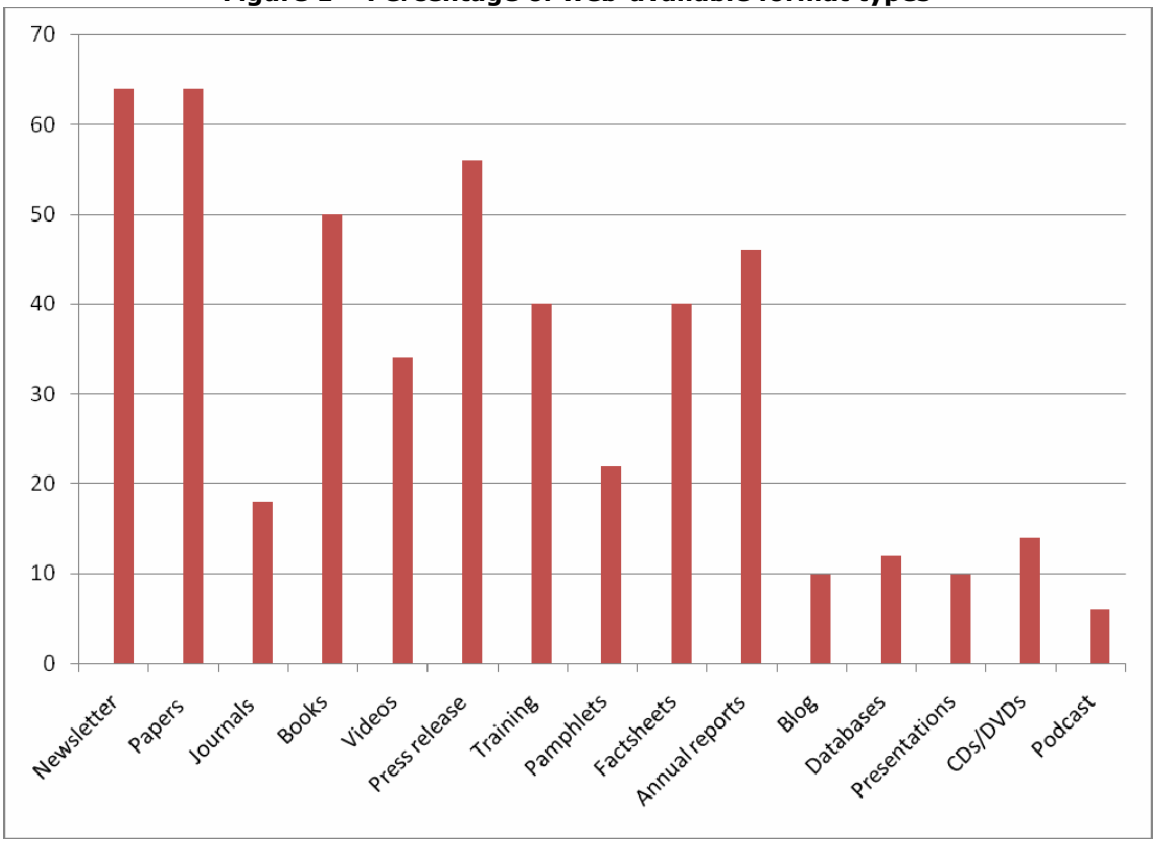
Results

In exploring the web sites of various health related organizations, a large number of information material types were revealed. The first type includes newsletters and/or bulletins, the very essence of NGO information disclosure. These are primarily directed at target audiences and possible donors. Papers, including briefing papers, policy papers and policy briefs comprise another large part of the virtual literature of NGOs. Many items are referred to as books in the publishing area of sites. These can include anything from the multiple page electronic file to the formally published monograph. In this survey, a search for a representative number of "books" was done in WorldCat to ascertain whether the item was cataloged in any way.

Many times publications referred to as journals appear on the web site. A representative number of these titles were also searched in WorldCat to ascertain whether cataloging had been done for the work. Due to the public outreach function of many NGOs, the presence of videos, press releases, training materials, pamphlets and factsheets were also present on some of the sites. Annual reports and similar materials are essential to maintaining the reputation of the NGO and so are posted to the web. Other formats for the dissemination of information are unique to the virtual world, and are frequently found on NGO websites. These include blogs, databases, PowerPoint presentations, CDs or DVDs, and podcasts. A few of the sites also included printable items such as posters and post cards related to the subject area or engaged in virtual social networking in particular via FaceBook.

The following chart indicates the frequency in percentage for the existence of each of these formats on the websites of fifty randomly selected NGOs that generally work in the area of health concerns.

Figure 1 – Percentage of web-available format types



In early results from a larger survey, it was found that a great deal of information is not published on the website. This includes any reports required by the national, regional, and state governments. Also not included were reports to the funding bodies of the organizations surveys. This frequently makes up a large percentage of the information collected by an organization. Finally, the materials that are printed for local distribution, such as handouts regarding events and scripts and recordings from radio and television appearances, are frequently not found on the web. The searches conducted for publications by the NGOs in question through WorldCat revealed that at least 80% of the organizations had publications that had been formally cataloged in the database. The publication types in WorldCat included monographs, serials, internet resources, audio files, and CDs or DVDs in addition to brief reports.

As part of the larger study, interviews with a variety of organizations of interest revealed that there are many reasons for lack of disclosure of information. Frequently, individual's remarks implied that NGOs are competing with each other to obtain funding and are therefore not very forthcoming in the interest of remaining highly desirable to their donors. The issue of confidentiality is also of major concern. It would be inappropriate and quite chilling for the organization to disclose client records in any public way. In the areas of HIV/AIDS as well as the health of women and children, there sometimes exists a fear of reprisal from the government if certain information is disclosed. This has been articulated more than once in the personal interviews done as part of the larger research project, particularly in the developing nation. Police and the courts are not always on the side of the victim in some societies, and public disclosure of the short-comings of these agencies can often put the NGO at risk.[7] Finally, the simple fact that the organization cannot afford to hire someone adept at programming prevents information from making it to the website. This is another problem that is particularly prevalent in developing nations where the cost of supporting a web site is high and the cost of hiring someone with the skills to maintain the site is out of the question.

Several characteristics were shared among the sites visited. Most of the sites clearly indicated an area in which publications might be found. Many also had news feeds of some sort on their homepages. Annual reports were often hidden on secondary pages, but were generally easy to access. Some of the items with the most direct access included areas in which donations could be made and the nearly ubiquitous press releases. The availability of RSS feeds for news releases was also quite common.

The general study, although not yet statistically significant, is revealing some interesting, if not entirely unexpected differences in information trends between developed and developing countries. As far as the focus of the organization is concerned, the NGO in a developing country is more likely to have the local

audience as a target as opposed to trying to influence policies on a regional, national, or international stage. In addition the developing nation NGO is less likely to support dissemination of information via conferences or seminars, emphasizing capacity building more than the NGO in the developed world. In addition, the developing nation NGO places less emphasis on the collection of research results but collects books, published reports, news clippings, and videos to a larger degree. The first world NGO tends to collect more journals, probably because their resources allow them to pay the subscription price.

Commonalities

In reviewing the websites of health-related organizations several aspects of information dissemination were found to be commonly held. Most pages provided a clear access point for publications although the terms varied from "library" to "publications" to "research." A large majority of the web pages also provided newsfeeds of some sort. These feeds included information about agency-related events, accomplishments, and more general press releases that might assist in making the general public and others aware of the actions of the agency. Although sometimes buried at a second or third level on the pages, many NGOs provided annual reports and related materials. The most apparent common link for the NGO web page was, not surprisingly, a "donation" link.

Suggestions and Conclusions

Two problems are paramount for the researcher working with NGOs. First is access to the print materials created by the organizations. The original impetus to this research was the story of a researcher in a developing country being unable to access a seminal report in print that had been readily available in that country when the researcher returned to the U.S. This is not uncommon and impedes the ability to provide quality information outside of the target area. Early research also indicated that the lack of formal publication of these reports results in large amounts of information piling up in the corners of NGO offices that are not available either for digitization nor cataloging. Often these organizations have private libraries that contain important information, but these collections are not available to individuals from outside of the organization.

The second area of concern revolves around those items that are born digital. Librarians in large and small institutions are struggling with the best manner in which to preserve and archive documents that have no print counterpart. OpenSIGLE is an example of the way in which this type of material might be archived.[8] The use of DSpace and open archives in Europe has been used to support subject repositories to good effect. In addition to grey literature, the NEREUS project has been compiling a repository for economic literature that reflects the cooperation of over twenty institutions of higher learning in the area as well as one university in the U.S.[9] The Grey Literature Report supported by the New York Academy of Medicine, although not a repository in the formal sense, also provides important access to digital material. In no small way, this project reflects the possibility of using institutional repositories to provide another avenue to the preservation and archiving of "born digital" information.

Two other, related tools exist that might be of use to the scholar and the librarian. These include "The Way-Back Machine"[10] and the subscription service Archive-It[11], both provided by the Internet Archive. The Way-Back Machine is basically a web crawler that will return the searcher to any site for which there is a URL that has not been blocked. The Archive-It subscription will allow the institution to build collections of digital resources according to their instructions and archive it for future use.

All four types of tools may be useful for capturing grey literature produced by NGOs whether related to health or any of the other purposes for which the organizations exist.

As mentioned in the introduction, this paper is reflective of the beginning of a larger research project involving the creation, preservation, and dissemination of information by NGOs. These organizations are still growing at an amazing rate, and hold the promise of becoming rich providers of information resources in the future. Their work in human rights, development, health, humanitarian relief, and many other areas provide unique resources for the librarian and the scholar. The capture of the grey literature they produce is a challenge that can and should be met.

References

1 United Nations, About NGO Association with the U.N. - http://www.un.org/dpi/ngosection/about-ngo-assoc.asp (Accessed November 20, 2008)

2 Ebrahim, Alnoor (2003) Accountability in Practice: Mechanisms for NGOs. *World Development* 31, 813-829.

3 A few examples include Grebremichael, Meseret D. and Jason W. Jackson. (2006) Bridging the Gap in Sub-Saharan Africa: A Holistic Look at Information Poverty and the Region's Digital Divide. *Government Information Quarterly* 23, 267-280, Fernandez, M. Isabel et al. HIV prevention programs of nongovernmental organizations in Latin America and the Caribbean: the Global AIDS Intervention Network project. *Rev Panam Salud Publica* [online]. 2005, v. 17, n. 3 [cited 2008-11-20], pp. 154-162. Available from: <http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S1020-49892005000300002&lng=en&nrm=iso>. ISSN 1020-4989. doi: 10.1590/S1020-49892005000300002, and Luzi, Daniela, et al. (2004) The Communication Flow of Research Project Results *Publishing Research Quarterly* 20, 13-24.

4 New York Academy of Medicine, *The Grey Literature Report,* <http://www.nyam.org/library/pages/grey_literature_report> (Accessed November 20, 2008)

5 . Jacobs, James R. Non-Governmental Organizations Search, <http://www.google.com/coop/cse?cx=012681683249965267634%3Aq4g16p05-ao> (Accessed November 20, 2008).

6Devdir Organization, *Directory of Developmental Organizations,* <http://www.devdir.org/> (Accessed November 20, 2008).

7 Personal interviews conducted in Siem Reap, Cambodia, May 29, 2008 - June 4, 2008 and in Phnom Penh, Cambodia, June 5, 2008 - June 10, 2008.

8OpenSIGLE, System for Information on Grey Literature in Europe, <http://opensigle.inist.fr/> (Accessed November 26, 2008)

9 Nereus Consortium, Economist Online <http://www.nereus4economics.info/econline.html> (Accessed November 20, 2008)

10 Internet Archive Way Back Machine <http://www.archive.org/index.php> (Accessed November 20, 2008)

11 Internet Archive, Archive-It <http://www.archive-it.org/> (Accessed November 20, 2008)

Green Light for Grey Literature?
Orphan Works, Web-Archiving and other Digitization Initiatives – Recent Developments in U.S. Copyright Law and Policy

Tomas A. Lipinski, School of Information Studies; University of Wisconsin, United States

Abstract
This paper reviews recent legislative and case developments in the area of copyright law affecting the collection, preservation including digitization and dissemination of grey literature. Alternative frameworks for crafting a legislative solution to impediments the copyright present to these uses are discussed. This includes review of pending legislation targeting the problem of so-called "orphan works" offering a limitation on the monetary damages or injunctive relief the copyright owner may be granted and another pending proposal aimed at relaxing the anti-circumvention prohibition of section 1201 that would allow access to compilations consisting primarily of public domain works that are protected by technical protection measures. The recent SECTION 108 STUDY GROUP REPORT also contain recommendations related to preservation (reproduction) and dissemination of both analog and borne-digital works, including a new provision for internet archiving. Finally, recent case law supporting the archiving of various online sub-literatures is reviewed, such as the disputes over caching and archiving by Google and the TurnItIn plagiarism combating service. Short of a legislative solution, the procedural elements affecting copyright enforcement are assessed to determine the legal risk in use of grey literature. These proposals and cases are analyzed and critiqued, with assessment towards solving the copyright issues related to the preservation and use of various grey literatures. Policy failures as well as successes in the United States can assist policy makers in other countries when contemplating copyright issues related to preservation and use of grey literature.

Introduction
This paper proceeds on the assumption grey literature refers to "any documentary material that is not commercially published and is typically composed of technical reports, working papers, business documents, and conference proceedings"[1] or the "quasi-printed reports, unpublished but circulated papers, unpublished proceedings of conferences, printed programs from conferences, and the other non-unique material which seems to constitute the bulk of our modern manuscript collections."[2] In the educational context it could also include recorded lectures and other course content, student papers, thesis' repositories, etc. The dominant theme of these conceptualizations is the unpublished nature of the literature, but is this true in every case? A later section of this paper explores the issue of publication status and asks whether in the eyes of the U.S. copyright law these works are indeed unpublished, with the impact of that publication status on use and legal risk discussed. Issues related to the institutional collection and dissemination of grey and other literatures protected by copyright is of increasing interest in the United States, the European Union[3] and world-wide.[4]

There are two options pursued in the United States when of crafting legislative or regulatory "solutions" to impediments that the copyright poses to the reproduction (collection, preservation, etc.) and public distribution (circulation, dissemination online, etc.) of protected content. The first is to offer an exemption (or more precisely an affirmative defense) for what would otherwise be an infringing use. Exemptions come in two forms, general (those available to all, such as fair use under section 107) and specific (limited to the particulars of the circumstance, such as the exemption granted to libraries and archives for reproduction and distribution of certain works under section 108). The second option is to offer users some sort of safe harbor or protection from the impact of such infringement. This is typically crafted as a limitation on monetary[5] and in some cases injunctive remedies[6] available to copyright owners. In rare instance immunity from any liability whatsoever may be granted.[7] This paper assess whether the existing and emerging legal climate is amenable to the use of grey literature in the ways that libraries, archives, and other institutional organizations might desire to obtain and make accessible grey literature, through archiving, digitization, etc. The paper explores the current and potential interplay of the two policy options in light of proposals for reform recent case developments and also the dynamics of copyright litigation.

Library And Archive Reproduction And Distribution Under Section 108
Other than fair use (discussed below) section 108 of the United States copyright law offers qualifying institutions specific reproduction and distribution rights that may be useful in obtaining and distributing collections of grey literature. Section 108 allows for the reproduction and public distribution (circulation for example) of copies or phonorecords[8] of the collection of a qualifying library and archive for

preservation and security of unpublished materials or of published materials in cases of damage, deterioration, loss, or theft, or if the existing format in which the work is stored has become obsolete.

Current Law

In cases of preservation and security under section 108(b), the copy or copies, phonorecord or phonorecords (up to three copies or phonorecords may be made) must be from a work in the current collections of the library or archive and if a digital copy is made, must not be made available to the public in that format outside the premises of the library or archives, i.e., remote access to the material is not allowed afterwards. A copy made under subsection (b) for deposit in another library or archive may be transferred to that library or archive in digital format but the receiving library or archive must not distribute the material in that format or likewise if this institution is the receiving library or archive, i.e., staff cannot make the reproduced material available to patrons in digital form in any manner such as an in-house intranet.[9] This would allow a qualifying library or archive with a collection of unpublished grey report or proceeding literature of the ABC Association or the XYZ Corporation to make a copy of the collection for preservation or security purposes or even to make a complete copy of the collection for another qualifying library or archive. The library or archive could digitize these collections as well in order to increase searching capabilities of users (staff or patrons) in accessing the content. However, the digital copies may not be made available outside the premises of the library or archive, but relegated to in-house use alone.

In cases of damage, deterioration, loss, or theft, or if the existing format in which the work is stored has become obsolete under section 108(c), the copy or copies made (up to three copies may be made) are subject to the same limitation on digital distribution, i.e., remote access to the material is not allowed, and the library or archive must first make a reasonable effort to obtain an unused replacement of the published work at a fair price.[10] A "reasonable effort" "will vary according to the circumstances of a particular situation. It will always requires recourse to commonly-known trade sources in the United States, and in the normal situation also to the publisher or other copyright owner (if such owner can be located at the address listed in the copyright registration), or an authorized reproducing service."[11]

The legislative history of the digital copying provision of section 108, added by the Digital Millennium Copyright Act,[12] indicates that Congress was concerned with infringement vis-à-vis the proliferation of digital libraries: "Although online interactive digital networks have since given birth to online digital 'libraries' and 'archives' that exist only in the virtual (rather than physical) sense on Web sites, bulletin boards and home pages across the Internet, it is not the Committee's intent that section 108 as revised apply to such collections of information…The extension of the application of Section 108 to all such sites is tantamount to creating an exception to the exclusive rights of copyright holders that would permit any person who has an online Web site, bulletin boards, or a home page to freely reproduce and distribute copyrighted works. Such an exemption would swallow the general rule and severely impair the copyright owner's right and ability to commercially exploit their copyrighted works."[13] Thus, an on-premises library or archive use of a section 108(b) or (c) digital copy is the rule. These obvious limitations of section 108 prompted the recent work of the Section 108 Study Group to consider recommendations for legislative reform.

Proposals for Change: The Section 108 Study Group Report

Various recommendations contained within the Section 108 Study Group Report (Report) would increase the ability of a library or archive (the Report recommends that museums also be added to the list of qualifying institutions) to collect, archive and make other use of grey literature.[14] First, the Report recommends that section 108 be amended to allow third parties through outsourcing arrangements to reproduce the work for later use by a qualifying archive, library or museum. It is often the case the large digitization projects require institutions to engage the services of low-cost often off-shore third parties. Amended section 108 would allow such services if undertaken without subsequent retention or commercial use by the outsourcer and the outsourcer agreed contractually to be subject to effective process, i.e., agree to be sued in United States court.[15]

Several recommendations implicate the preservation, digitization and dissemination aspects of collections containing or consisting grey literature and other works. In specific, the Report recommends amendment of section 108 to allow off-site lending of physical-digital content, e.g., a CD-ROM, if that was the original format of the item for preservation or security purposes under subsection (b) and for replacement copying under subsection (c), with a new category "fragile" of work added to subsection (c) "replacement" rights.[16] "[T]he Study Group defines a 'fragile copy' as one that is embodied in a physical medium that is at risk of becoming unusable because it is delicate or easily destroyed or broken and cannot be handled without risk."[17] Depending upon the format, items of grey literature may fall into this category. The Study Group could not reach consensus on access to virtual-digital, e.g., web-based works, thus the Report made no recommendation regarding this critical level of access in the Report.

Subject to numerous qualifying conditions a section 108 entity would be able to preserve "at-risk published or other publicly disseminated works in its collections."[18] The most dramatic recommendation relates to the ability of qualifying entities, subject to an opt-out mechanism by owners regarding publicly available online content, to make that content "accessible to users for purposes of private study, scholarship, or research."[19] However "publicly available online content" is content not protected by password or "requiring an affirmative act by the user to access" and would likely exclude content found on websites that are subject to terms and conditions of use, End User License Agreements (EULA)[20] or other control mechanisms that we require click of an "I agree" prompt or other affirmative click of agreement but not websites subject to mere browse wrap agreements.[21] Qualifying content could be copied or archived by the institution for individual noncommercial use by patrons. Thus sources of grey literature as well as other content located on third party websites could be subject to the archiving provisions of an amended section 108.

Solving The Problem Of Orphan Works

It may be that archiving and digitization, i.e., reproduction and public distribution of a work of grey literature in its entirety may be impeded by concerns of copyright infringement. Depending on the circumstances as such use may beyond fair use.

The Problem

It may be that the institutional collectors of grey literature like other users of copyrighted content would be willing to contact the owner and secure permission to use the work, even if compensation of the owner is involved. However, the owner cannot be identified or located. Given the nature of the provenance of grey literature such content may be particularly susceptible to the problem of orphan works. An "orphan work" is "a term used to describe the situation where the owner of a copyrighted work cannot be identified and located by someone who wishes to make use of the work in a manner that requires permission of the copyright owner."[22] Users that desire to make the use but refuse to do so under any circumstances of legal risk, i.e., that the owner could one day surface and sue for copyright infringement will forego that use. As copyright law is a law of strict liability, these good faith attempts do not impact liability though general efforts of good faith may impact damages.[23] "*Such an outcome is not in the public interest*, particularly where the copyright owner is not locatable because he no longer exists or otherwise does not care to restrain the use of his work."[24]

The Solution

Earlier this fall Senate bill S. 2913, the Shawn Bentley Orphan Works Act of 2008[25] passed in the Senate is awaiting action in the final days 110th Congress, having been engrossed in the House on September 27. The bill would create new section 514 of the copyright law (title 17 of the United States Code). Proposed section 514 is an example of the second form of policy approach to a copyright problem, i.e., addresses the problem not by creating an exemption but in limiting the so-called bottom line, i.e., damages, the user-defendant faces should litigation by the owner-plaintiff be successful. If the user meets the safe harbor requirements of the provision then the only monetary relief the plaintiff can claim is for reasonable compensation for the infringing use made of the work. Damages (actual or statutory including damage enhancement for willful violations) as well as costs and attorney fees are not available. In some circumstances no monetary relief whatsoever is available. In the instance of derivative uses injunctive relief is also limited. The question is whether or not limiting monetary liability to reasonable compensation is still too much for some would-be users to afford whereby such user would still forego use of the orphan work. Thus the impact of this solution would not be in the "public interest" to use the language of the Report.

Reasonable compensation is defined under proposed section 514(A)(3) as "the amount on which a willing buyer and willing seller in the positions of the infringer and the owner of the infringed copyright would have agreed with respect to the infringing use of the work immediately before the infringement began." The impact is obvious, users must obtain some evidence or documentation of what that amount might have been, and second keep that evidence or documentation should the orphan owner ever appear one day and the user need to prove qualification under the safe harbor. It is also a requirement of qualification that should the owner appear the user must bargain in good faith, offering to pay reasonable compensation. So again having documentation of what this amount might be is useful especially in cases where the owner appears years after the initial infringement. Considering the duration of copyright in the United States there may be a lengthy period during which this information may be relevant. Assuming the work is still protected by copyright this would be three years after infringing use of the work ceases, as the statute of limitations for commencing a civil action is three years.[26] So for a work for which the copyright does not expire until say 2045, where the infringing use commences in 2010, lasts until 2035 when the work is deaccessioned from the library or archive collection, the user would need to keep records of what reasonable compensation would have been in 2010 for 28 years: 25 years of use plus the three years to cover the tolling of the statute of limitations. For situation where the use is continuous,

i.e., the work remains a permanent part of the collection such making the work accessible to the public on a website for example, that would be for as long as the copyright lasts plus three years!

Under proposed section 514(c)(1)(B), a nonprofit educational institution, museum, library, archives, or a public broadcasting entity (or employees of such entity acting within the scope of their employment) can reduce the monetary amount to zero if three conditions are met. First, the infringement was performed without any purpose of direct or indirect commercial advantage (this is different than a situation where the use *results* in a direct or indirect commercial advantage, only the "purpose" must be so). Second, the infringement was *primarily* educational, religious, or charitable in nature (this is not the same "solely" nor does this standard look to entity, the categorizations being those employed more often to describe the nature of the entity rather than as here its conduct, i.e., here the "infringement"). Third, after receiving a notice of claim of infringement and having an opportunity to conduct an expeditious good faith investigation of the claim, i.e., some legal assessment of the merits of the claim of infringement must be undertaken) the infringer promptly ceased the infringement.

The "notice of claim of infringement" does not require that a law suit be filed rather it is more akin to the notice under section 512(c)(3) that triggers an expeditious take-down or restriction of access to content.[27] As required under proposed section 514(a)(1) the notice would be made in writing and include the name of the owner and title of the infringed copyright as well as sufficient information regarding the owner or their representative and the location of the infringing content.

Finally in the case of derivative works[28] or to be more precise under proposed section 514(c)(2)(B), where the infringer has "prepared or commenced preparation of a new work of authorship that recasts, transforms, adapts, or integrates the infringed work with a significant amount of original expression," the court may not enjoin the defendant's continued use. The concept of integration offers a somewhat broader scope of uses than contemplated by the statutory definition of derivative work. Moreover, the inability to enjoin continued preparation or use in essence creates a statutory license to use the work as long as the "infringer pays reasonable compensation in a reasonably timely manner after the amount of such compensation has been agreed upon with the owner of the infringed copyright or determined by the court." If the owner refuses to agree during good faith attempts at negotiation, the court may order the owner to accept the reasonable compensation and allow the use to continue. The user must also provide attribution "in a manner that is reasonable under the circumstances to the legal owner of the infringed copyright." However attribution is only required "if requested by such owner." It is odd to condition a court ordered attribution upon request by the owner as an initial condition of section 514 qualification is to provide attribution, as discussed below.

Qualifying for the Proposed Section 514 Safe Harbor: The Search

There are several requirements before the section 514 limitations on remedy can apply. First, the user (the proposed language repeatedly uses the word "infringer") must by a preponderance of the evidence demonstrate that before the use ("infringement") began he of she "performed and documented a qualifying search, in good faith, to locate and identify the owner of the infringed copyright" and that the search was unsuccessful. This suggests two elements to a search: substantive as to the content of the search or its protocol ("a qualifying search") and procedural as to how the search is executed ("in good faith"). A user might have access to a list of the proper steps or best practices developed by library and archive professionals but exert little effort to complete or execute those steps practices.

The requirements of a "diligent effort that is reasonable under the circumstances to locate the owner" is detailed in proposed section 514(b)(2). The diligent effort requires "at a minimum" a search of the Internet accessible Copyright Office records (assuming there is sufficient information regarding to the work to construct a search), a search of other authorship and ownership information, the "use of appropriate technology tools, printed publications, and where reasonable, internal or external expert assistance" and "appropriate databases, including databases that are available to the public through the Internet." These sources might include content made available from third party sources, for example, a web-accessible database of renewal records of published works filed between 1950 and 1992, available at http://collections.stanford.edu/copyrightsrenewals/bin/page?forward=home (renewal records). Of course not all of the content will come at little or no cost. In fact a later provision anticipates a diligent effort to include recourse to pay-per-use services ("use of resources for which a charge or subscription is imposed"). A later provision requires review "as appropriate" of Copyright Office records not available to the public through the Internet. This suggests either a trip to Washington, D.C. is in order or at least to the regional federal depository library (as not all partial depositories would have Copyright Office records in the collection). Finally, proposed section 514 anticipates that the Copyright Office ("Recommended Practices" including at least one such statement for each category of work of authorship listed in section 102[29]) as well as "authors, copyright owners and users" make available best practices to assist users in performing a qualifying search.

If through this search process the owner is located, even though never contacted or once contacted fails to respond to the inquiry the work ceases to be orphan.[30] The proposed provision reads in part: "The fact that, in any given situation . . . an owner of the infringed copyright fails to respond to any inquiry or other communication about the work shall not be deemed sufficient to meet the conditions under paragraph (1)(A)(i)(I)", i.e., "performed and documented a qualifying search." Second, attribution "based on information obtained in performing the qualifying search" must be provided about the owner "in a manner that is reasonable under the circumstances." Attribution will make it easier for owners to identify their works and the unlawful use being made of them. This in turn fulfills the purpose of section 514 (or at least the remission mechanism proposed by the U.S. Copyright Office) which makes it more likely that copyright owners will find users and vice-versa and come to agreement over the use of the work.[31] Use of the work must also indicate that it is made subject to the proposed section with "the form and manner of which shall be prescribed by the Register of Copyrights." This is likely done to dissuade others from incorrect assumption that because the qualifying user is making use of the work, such as posting of the work on the library or archive website, use of the work by all takers is welcome and free. Rather the provision promotes awareness of uses that are compliant with proposed section 514. In terms of process, the application of the section 514 safe harbor must be asserted in the initial pleading by the defendant. It is assumed this is done to encourage the parties to come to some reasonable agreement or perhaps to assist the court in pretrial motion determinations.

Additional Requirements: When (If) the Owner Later Appears

In addition to these requirements occurring prior to use of the work, other requirements exist should the owner of the orphan work later surface and give "notice of a claim of infringement." The user, after "having an opportunity to conduct an expeditious good faith investigation of the claim" must first undertake a good faith negotiation with the owner over the amount of "reasonable compensation" then "render payment of reasonable compensation in a reasonably timely manner after reaching an agreement with the owner" (or if ordered by the court to do so). Once it is determined (either by negotiation or by the court) what amount constitutes reasonable compensation that amount must be paid in a "reasonably timely manner." It is not certain what the requirement of claim investigation is meant to accomplish. It could be to assess the merits of the claim, but this would be odd as the user has likely long since concluded that the use of the work is infringing otherwise there would be no need to have undertaken measures that would qualify for the proposed section 514 safe harbor. This "good faith investigation of the claim" may address procedural aspects of the circumstances, i.e., verifying identification of the owner, the work infringed and the infringing work. Another issue is whether the good faith negotiation obligation must be successful, as the statute appears to anticipate no other option. Notice also, the obligation to negotiate is on the user not the owner. So if the owner wants nothing to with arriving at an arrangement that would compensate for past use and instead desires to sue the user for infringement the owner is free to do so. Of course if the user is able at least to document a good faith attempt to negotiate (as well as the other search and attribution requirements) then the monetary relief available to the owner will be foreclosed for a qualifying nonprofit or limited to reasonable compensation in other circumstances or for derivative uses.

Should S. 2913 or some subsequent variation become law, this discussion provides a basis upon which the user can understand its operation and fulfill the requisite legal obligations or to assess whether the cost of those obligations are not worth the benefit of the safe harbor. In essence, the bill encourages search and documentation of the search for the copyright owner. Oddly, if the search is successful the work is no longer "orphan" and the provision will not apply. Yet in this instance permission may not be forthcoming and so use cannot be made of the work without the threat of legal repercussion anyways. When a work remains orphan and the copyright owner is not located, the bill offers benefit but at potential high practical (time, record-keeping, etc.) and actual cost (outside exerts and resources) that may be no better in terms of the cost of the associated with a use of the work without application of the safe harbor, the cost of the legal risk of use under present law.

Users of orphan works should also be aware of the potential looming cost of litigation to vindicate a proper attribution and search. Those who move forward with use will be the test cases, carrying the initial cost of this "sorting out," hopefully to the benefit of subsequent users. It is critical that courts do not interpret the reasonably diligent search, attribution or other requirements too harshly otherwise the litigation-avoidance incentive will not operate properly. Once a precedent has been set to establish some reasonable norms for predicting when the provision would apply, users will have some structure as to what constitutes search and attribution, and owners might be more reluctant to litigate.

A brief comment of technical protection measures

U.S. law prohibits circumvention of technological protection measures (TPMs) that control access[32] as well as the distribution (trafficking) of technologies that control access[33] or distribution (trafficking) of protection technologies that control specific uses of a work, so called "black-box" devices.[34] Such devices

are those that are primarily designed to circumvent, have limited commercially significant purpose, or are marketed as an anti-circumvention device. It is unknown the extent to which grey literature is disseminated subject to such TPMs. However, the increasing use of TPMs by content owners of "white" literature and in particular formats such as sound recordings such as CDs and audiovisual works such as DVDs suggests that this may be a future problem for grey literature as well. In order for content to be subject to the prohibition on circumvention the content must be within subject matter and protection of the copyright law, the prohibition does not apply to content not protected by copyright such as that in the public domain.[35] Second, the control must be put in place by the copyright owner or with the permission of the copyright owner. If the control is instigated by a third party web site owner of database vendor for example without permission of the copyright owner, the prohibition will not apply.

For qualifying institutions seeking to acquire grey literature that may be subject to such prohibitions there is a statutory exception, at least for lawful circumvention during the acquisitions phase. Section 1201(d) provides a specific exception for qualifying nonprofit libraries, archives, or educational institutions to circumvent an access control in order to make a bona fide determination of whether to purchase an item for its collection or curriculum: "access to a commercially exploited copyrighted work solely in order to make a good faith determination of whether to acquire a copy of that work for the sole purpose of engaging in conduct permitted under this title shall not be in violation of subsection (a)(1)(A)."[36] Notice that this exception operates with respect to the section 1201(a)(1)(A) anti-circumvention of access control provision, it does not allow qualifying nonprofit libraries, archives, or educational institution to traffic in either an access or use control.[37] Such entities are still prohibited from engaging in conduct that remains a section 1201(a)(2) or section 1201(b) trafficking violation, i.e., sharing the means of the circumvention with another qualifying entity.

In addition there is three year cycle of rule-making, with a de novo review made of requests for regulatory exemption to the circumvention prohibition. The statutory standard for granting the regulatory exemption is whether or not "noninfringing uses by persons who are users of a copyrighted work are, or are likely to be, adversely affected." In 2006 the standard was modified somewhat as "the Register has concluded that in certain circumstances, it will also be permissible to refine the description of a class of works by reference to the type of user who may take advantage of the exemption or by reference to the type of use of the work that may be made pursuant to the exemption… must be properly tailored not only to address the harm demonstrated, but also to limit the adverse consequences that may result from the creation of an exempted class."[38] There are six exemptions granted under current law, two of possible relevance to preservation and access: "Computer programs protected by dongles that prevent access due to malfunction or damage and which are obsolete and library preservation of "computer programs and video games distributed in formats that have become obsolete."[39] Likely neither is of much relevance to collections of grey literature but nonetheless indicates that should grey literature be increasingly subject to such controls, short of legislative remedy through amendment of section 1201,[40] there is an accessible if somewhat cumbersome and limited regulatory process to achieve similar even if not permanent ends.

Web Archiving and Fair Use

Several recent cases in the past two years have suggested that initiatives to engage in systematic archiving of content can be a fair use. In *Perfect 10 v. Amazon.com, Inc.*,[41] the Ninth Circuit concluded that Google's creation of its thumbnail index of images was fair use, commenting that "the significantly transformative nature of Google's search engine, particularly in light of its public benefit, outweighs Google's superseding and commercial uses of the thumbnails in this case." However, as the index allows users of the Google search engine to be led to infringing sources of the content, Google could be found contributorily liable: "Applying our test, Google could be held contributorily liable if it had knowledge that infringing Perfect 10 images were available using its search engine, could take simple measures to prevent further damage to Perfect 10's copyrighted works, and failed to take such steps."[42] A conclusion of fair use was also found in another case involving Google, this time its practice of automatically archiving web sites unless the owner opted out. In *Field v. Google, Inc.*,[43] a district court again identified the social good that such preservation projects can achieve: "The fact that the owners of billions of Web pages choose to permit these links to remain is further evidence that they do *not* view Google's cache as a *substitute* for their own pages. Because Google serves *different and socially important purposes* in offering access to copyrighted works through 'Cached' links and does *not* merely *supersede* the objectives of the original creations, the Court concludes that Google's alleged copying and distribution of Field's Web pages containing copyrighted works was *transformative*."[44] Finally, the impact of the recent settlement by publishers and authors against Google also suggests that such archiving projects will continue to present legal challenge but through decision or settlement will be allowed to continue.[45] These developments lend support for similar efforts by institutions providing similar social good by preservation of the cultural record. It may be that the same argument could be made in the case of preservation of grey literature when that collection is unique and does not exist elsewhere and the institutions serves as the sole source of the content. A final archive decision not involving Google also stands for the proposition that such

initiatives offer a beneficial societal purpose and can likewise be a fair use. In *A.V. v. iParadigms, Ltd.*,[46] the court observed that as in the Google index, cache and archive cases the "use of Plaintiffs' written works [is] highly transformative. Plaintiffs originally created and produced their works for the purpose of education and creative expression. iParadigms, through Turnitin, uses the papers for an entirely different purpose, namely, to prevent plagiarism and protect the students' written works from plagiarism… makes no use of any work's particular expressive or creative content beyond the limited use of comparison with other works… provides a substantial public benefit through the network of educational institutions using Turnitin. Thus, in this case, the first factor favors a finding of fair use."[47] As a result the use of the student-plaintiff's papers in the TurnItIn databases was a fair use. In each of the case the use was deemed transformative and even though the entire work was taken in the instance of images in the Google cases or student papers in the *iParadigms* case the complete taking was necessary to accomplish the good purpose. This is in contrast to the recent case involving the Harry Potter Lexicon. The nature of encyclopedias and reference guides being in general transforming, though under the particular circumstances the publisher of The Lexicon: An Unauthorized Guide to Harry Potter Fiction and Related Material took more than once necessary to accomplish its good purpose.[48]

Final thoughts on the use of Grey Literature and the particulars of Copyright Enforcement

Other elements of the copyright law may make use of grey literature nonnfringing or reduce the likelihood of litigation or the fallout from that litigation should it occur. First, it may be that the content is not protected by copyright. For example, works produced by the federal government are in the public domain.[49] Other works may have fallen into the public domain due to lapse of protection. The rules can be rather complex. In general works published before 1923 are in the public domain, those works published 1923-1963 with notice and renewal are protected for 95 years from date of publication and those published 1964-1977 with notice (renewal automatic) are also protected for 95 years from the date of publication. Under the 1976 Copyright Act, works created after 1977 are protected for the duration of the author's life plus 70 years, or if corporate, anonymous, pseudonymous: lesser of 95 years from publication or 120 from creation. If the work is unpublished and created before the first of January 1978 (the effective date of the 1976 Copyright Act) then duration of copyright is 2002 or the author's life plus 70 years, whichever is longer. If created before 1978 and published before 2003, then the work is protected for the greater of author's life plus 70 years or until 2047. If the unpublished work was created after 1977, the duration is for the life of the author plus 70 years, or 120 years from creation for corporate, anonymous, pseudonymous authors. When the death date of an author is unknown: a default of 120 years from creation applies. As a result, the period of liability for infringing use of protected content may be lengthy. However, certain particulars of copyright litigation and enforcement may work against litigation and reduce the ultimate legal risk the user of grey literature may face.

Legal risk is a combination of several factors: the potential for liability, the likelihood of litigation (or threat of litigation) as well as the possibility of settlement and the impact of that litigation (or settlement), i.e., what remedies are available to the copyright owner. What is the potential for liability ("can I be sued?"), how likely is litigation ("will I be sued?") and how likely is it that the infringement will be discovered, and what remedies are available ("what's the bottom line?"). Furthermore, the scope of available damages (and award of costs and attorneys fees) is related to whether the work is published or unpublished, the publication status of the work. Much of the grey literature may be in fact unpublished. Furthermore, there may be an opportunity for damage remission, discussed below, as well.

While the work need not be registered to be protected, it is a prerequisite to litigation.[50] How many grey literature works are registered? Assuming the work is registered, and this may be a significant assumption in the case of grey literature, the timing of the registration in relation to the infringement and the status (unpublished or published) of the work determines the scope of damages available to the copyright owner. This may impact the decision to sue or not. It is unlikely an owner would undertake the cost of litigation if the monetary award were limited to actual damages alone or if costs or attorneys fees could not be recovered in addition. Registration must occur before infringement of an unpublished work and within three months of publication for published works in order to obtain statutory damages and attorney's fees.[51]

Even if there is potential for a significant award of damages, where the infringement is undertaken by an employee of nonprofit educational institution or library (or the institution itself is liable), and the employee was acting within scope of employment, believed and had reason to believe, that the use was a fair use under section 107, and infringed by reproducing the work, the court must remit the statutory damages awarded to zero. [52] The possibility of no statutory damages may dissuade a owner from ever suing.

So are works of grey literature published or unpublished? Publication is defined in Section 101 as the "distribution of copies or phonorecords of a work to the public by sale other transfer of ownership, or by

rental, lease, or lending." The distribution of copies on a busy street is publication, as is the unrestricted gift of copies constitutes. So too is leaving copies in a public place for anyone to take a publication. However, distributing text at a seminar for use only by the recipients is ordinarily not publication.[53] One district court concluded that posting content on the internet is a publication.[54] "The statutory definition of publication set forth above specifies two rules worth emphasizing: First, publication includes only acts of publication (1) by the copyright owner or (2) authorized by the owner. Unauthorized acts of publication by others do not result in publication. The copyright owner has the sole authority to authorize publication. Second, the phrase 'copies or phonorecords' refers only to plural items. What happens when a single work is distributed? For example, if a single piece of sculpture or book is distributed, rented, or lent, does publication occur? Not necessarily. There must be multiple copies available for distribution, transfer, rent, lease, or lending. Thus, publication occurs only if the single item is one of many copies available for distribution."[55] Thus Internet sources of grey literature are likely published, as are ephemeral reports of organizations that are released to the public, some conference proceedings, etc. Internal organizational documents from a corporation for example that reside in an institutional archive such as a university remain unpublished even if circulated (though this would constitute a public distribution under the copyright law. Once digitized and made accessible on the web, the publication status changes. Yet in either case such documents are likely to be unregistered! If the corporation retained copyright in those documents it could not proceed with litigation until it registered those works. In the proper circumstances the legal risk of using grey literature may be small, the works may be unprotected by copyright or if protected may offer unattractive circumstances for litigation.

Conclusion

The expanded collection and dissemination of grey literature (as well as other works protected by copyright) through archiving and digitization is bolstered by recent case law establishing the circumstances under which such initiatives can be a fair use under U.S. copyright law. In addition legislative reform is under way (section 108 and proposed section 514) to increase range of use rights available to institutions regarding protected content including grey literature. Moreover, the particulars of copyright enforcement may also work to minimize the legal risk in remaining circumstances.

References

1 Brian Matthews, Gray Literature: Resources for Locating Unpublished Research, C&CRL NEWS, March 2004.

2 PETER HIRTLE, BROADSIDES VS. GREY LITERATURE (1991), as quoted in MOYA K. MASON, GREY LITERATURE: ITS HISTORY, DEFINITION, ACQUISITION, AND CATALOGUING, available: http://www.moyak.com/researcher/resume/papers/var7mkmkw.html.

3 See, GREEN PAPER, COPYRIGHT IN THE KNOWLEDGE ECONOMY, COM(2008) 466/3, available at http://ec.europa.eu/internal_market/copyright/docs/copyright-infso/greenpaper_en.pdf, FINAL REPORT ON DIGITAL PRESERVATION, ORPHAN WORKS AND OUT-OF PRINT WORKS, SELECTED IMPLEMENTATION ISSUES (June 4, 2008), available at http://ec.europa.eu/information_society/activities/digital_libraries/doc/hleg/reports/copyright/copyright_subgroup_final_report_26508-clean171.pdf. See also, Annex: Model agreement for a licence [sic] on digitisation [sic] of out of work prints, available at http://ec.europa.eu/information_society/newsroom/cf/itemdetail.cfm?item_id=3366 (April 18, 2007).

4 See, e.g. STANDING COMMITTEE ON COPYRIGHT AND RELATED RIGHTS, STUDY ON COPYRIGHT LIMITATIONS AND EXCEPTIONS FOR LIBRARIES AND ARCHIVES (Seventeenth Session, Geneva, November 3 to 7, 2008) (Prepared by Kenneth Crews), available at http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=109192.

5 See, e.g. 17 U.S.C. § 504(c)(2).

6 See, e.g., 17 U.S.C. § 512(j).

7 See, e.g., 17 U.S.C. § 108(f).

8 17 U.S.C. § 101 defines a phonorecord as "'material objects in which sounds, other than those accompanying a motion picture or other audiovisual work, are fixed by any method now known or later developed, and from which the sounds can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device. The term "phonorecords" includes the material object in which the sounds are first fixed."

9 17 U.S.C. § 108(b).

10 17 U.S.C. § 108(c).

11 H. Rpt. No. 94-1476, 94th Cong. 2d Sess. 75-76 (1976), reprinted in 5 United States Code Congressional and Administrative News 5659, 5689 (1976).

12 Pub. L. No. 105-304, Title IV, sec. 404, 112 Stat. 2860, 2889-2890 (1998) (codified at 17 U.S.C. § 108).

13 S. Rpt. No. 105-190, 105th Cong. 2d Sess. 63 (1998).

14 U.S. COPYRIGHT OFFICE, THE SECTION 108 STUDY GROUP REPORT 31 (2008).

15 U.S. COPYRIGHT OFFICE, THE SECTION 108 STUDY GROUP REPORT 39 (2008) ("not for any other direct or indirect commercial benefit…contractually prohibited from retaining copies…preserves a meaningful ability…to obtain redress from the contractor for infringement by the contractor"). "The contractor should be contractually required to submit to U.S. jurisdiction and have assets in the United States, or be bonded and insured in this country." Id at 41.

16 U.S. COPYRIGHT OFFICE, THE SECTION 108 STUDY GROUP REPORT 52 and 61 (2008).

17 U.S. COPYRIGHT OFFICE, THE SECTION 108 STUDY GROUP REPORT 54 (2008).

18 U.S. COPYRIGHT OFFICE, THE SECTION 108 STUDY GROUP REPORT 69 (2008) Qualifying conditions relate security of process ("best practices"), storage capacity, system integrity, identification, retrieval, security of access, and the ability to migrate, audit, afford (in terms of cost), support (in terms of mission) and transfer (in cases of cessation of operations) the preserved content. Allowance should also be made smaller entities to engage in such efforts as well. Id. at 69-70.

19 U.S. COPYRIGHT OFFICE, THE SECTION 108 STUDY GROUP REPORT 80 (2008).

20 *Ticketmaster L.L.C. v. RMG Technologies, Inc.*, 507 F.Supp.2d 1096, 1108 (C.D. Cal. 2007) "Thus, by the Terms of Use, Plaintiff grants a nonexclusive license to consumers to copy pages from the website in compliance with those Terms. Inasmuch as Defendant used the website, Defendant assented to the terms." See also, *Druyan v. Jagger*, 508 F.Supp.2d 228, 237 (S.D.N.Y. 2007) (citing *Ticketmaster Corp. v. Tickets.com, Inc.*, 2003 WL 21406289 (C.D. Cal. 2003); and *Register.com v. Verio*, 126 F. Supp. 2d 238 (S.D.N.Y. 2000)): "First, courts have consistently held that the use of a website for such purposes as purchasing a ticket manifests the user's assent to the Terms of Use, and that such terms constitute a binding contract as long as the terms are sufficiently conspicuous."

21 *Specht v. Netscape, Inc*., 306 F.3d 17, 32 (2d Cir. 2002) (footnote omitted): "Internet users may have, as defendants put it, 'as much time as they need[]' to scroll through multiple screens on a webpage, but there is no reason to assume that viewers will scroll down to subsequent screens simply because screens are there. When products are 'free' and users are invited to download them in the absence of reasonably conspicuous notice that they are about to bind themselves to contract terms, the transactional circumstances cannot be fully analogized to those in the paper world of arm's-length bargaining." See also, Blaze D. Waleski , Enforceability of Online Contracts: Clickwrap vs. Browse Wrap, *e-Commerce Law & Strategy, November*, 2002, Vol 19, no. 7 (no pagination in Westlaw).

22 U.S. COPYRIGHT OFFICE, REPORT ON ORPHAN WORKS 15 (2006).

23 See, e.g., *Lowry's Reports, Inc. v. Legg Mason, Inc.*, 271 F. Supp. 2d 737, 746 (D. Md. 2003) ("The fact that Legg Mason's employees infringed Lowry's copyrights in contravention of policy or order bears not on Legg Mason's liability, but rather on the *amount of statutory and punitive damages and the award of attorneys' fees*." (emphasis added).)

24 U.S. COPYRIGHT OFFICE, REPORT ON ORPHAN WORKS 15 (2006) (emphasis added)

25 S. 2913, 110th Congress, 2d Session (April 24, 2008) (Shawn Bentley Orphan Works Act of 2008.

26 17 U.S.C. § 507.

27 17 U.S.C. § 512(c)(1)(A) ("upon notification of claimed infringement as described in paragraph (3), responds expeditiously to remove, or disable access to, the material that is claimed to be infringing or to be the subject of infringing activity").

28 17 U.S.C. § 101 defines a derivative work as "a work based upon one or more preexisting works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which a work may be recast, transformed, or adapted."

29 17 U.S.C. § 102 provides that "[w]orks of authorship include the following categories: (1) literary works; (2) musical works, including any accompanying words; (3) dramatic works, including any accompanying music; (4) pantomimes and choreographic works; (5) pictorial, graphic, and sculptural works; (6) motion pictures and other audiovisual works; (7) sound recordings; and (8) architectural works."

30 See also, U.S. COPYRIGHT OFFICE, REPORT ON ORPHAN WORKS 97 (2006) ("[O]nce an owner is located, the orphan works provision becomes inapplicable.").

31 U.S. COPYRIGHT OFFICE, REPORT ON ORPHAN WORKS 97 (2006) ("The primary goal of this study is to prompt owners and users to find each other and commence negotiation—it is *not* intended to allow use of works in disregard of the owner's wishes after that owner has been found.")

32 17 U.S.C. § 1201(a)(1).

33 17 U.S.C. § 1201(a)(2).

34 17 U.S.C. § 1201(b).

35 See, *Online Policy Group v. Deibold, Inc.*, 337 F. Supp. 2d 1195 (N.D. Calif. 2004); and *Lexmark International, Inc. v. Static ControlComponents, Inc.*, 387 F. 3d 522 (6th Cir. 2004).

36 17 U.S.C. § 1201(d)(1).

37 See, 17 U.S.C. § 1201(d)(4) ("This subsection may not be used as a defense to a claim under subsection (a)(2) or (b), nor may this subsection permit a nonprofit library, archives, or educational institution to manufacture, import, offer to the public, provide, or otherwise traffic in any technology, product, service, component, or part thereof, which circumvents a technological measure.").

38 71 Fed. Reg. 68472, 68473 (November 27, 2006).

39 37 C.F.R. §201.40 (2006).

40 See, e.g. H.R. 1201, the Freedom and Innovation Revitalizing U.S. Entrepreneurship Act of 2007 (FAIR USE Act of 2007), 110th CONGRESS, 1st Session (February 27, 2007). The FAIR USE Act would amend section 1201 to allow for

circumvention of an access control protecting compilations consisting primarily of public domain works, a work of "substantial public interest" for purposes of "criticism, comment, news reporting, scholarship, or research" and would allow circumvention in order to exercise rights under section 108(c) privileges, but excluding works in "obsolete" formats.

41 *Perfect 10 v. Amazon.com, Inc.*, 487 F.3d 701, *13 (9th Cir. 2007).

42 *Perfect 10 v. Amazon.com, Inc.*, 487 F.3d 701, *19 (9th Cir. 2007).

43 *Field v. Google, Inc.*, 412 F.Supp.2d 1106 (D. Nev. 2006).

44 *Field v. Google, Inc.*, 412 F.Supp.2d 1106, 1119 (D. Nev. 2006) (all emphasis added).

45 *The McGraw-Hill Cos. Inc. v. Google Inc.*, No. 05 CV 8881 (S.D.N.Y. filed Oct. 19, 2005); and *Authors Guild v. Google Inc.*, No. 05 CV 8136 (S.D.N.Y. filed Sept. 20, 2005). See, MOTION to Approve /Notice of Motion for Preliminary Settlement Approval (October 28, 2008); and STIPULATION AND ORDER FOR AMENDMENT OF PLEADINGS (October 30, 2008) available at http://news.justia.com/cases/featured/new-york/nysdce/1:2005cv08136/273913/.

46 *A.V. v. iParadigms, Ltd.*, 2008 WL 728389 (E.D. Va. 2008).

47 *A.V. v. iParadigms, Ltd.*, 2008 WL 728389. *6 (E.D. Va. 2008).

48 *Warner Brothers Entertainment, Inc. v. RDR Books*, 575 F.Supp.2d 513 **(**S.D.N.Y. 2008). Regarding the books in the Harry Potter series: "Other times, however, the Lexicon disturbs the balance and takes more than is reasonably necessary to create a reference guide. In these instances, the Lexicon appears to retell parts of the storyline rather than report fictional facts and where to find them. " Id. 548. Regarding the companion books to the series the use is less transformative: "The Lexicon's use of copyrighted expression from Rowling's two companion books presents an easier determination. The Lexicon takes wholesale from these short books. Depending on the purpose, using a substantial portion of a work, or even the whole thing, may be permissible… In this case, however, the Lexicon's purpose is only slightly transformative of the companion books' original purpose. As a result, the amount and substantiality of the portion copied from the companion books weighs more heavily against a finding of fair use." Id. at 548-549.

49 U.S.C. § 105 ("Copyright protection under this title is not available for any work of the United States Government.").

50 17 U.S.C. § 411: "[N]o action for infringement … shall be instituted until registration of the copyright claim has been made..."

51 17 U.S.C. § 412 ("[N]o award of statutory damages or of attorney's fees …shall be made … [for infringement of] an unpublished work commenced before the effective date of its registration; or … commenced after first publication of the work and before the effective date of its registration, unless such registration is made within three months after the first publication of the work.").

52 17 U.S.C. § 504(c)(2).

53 Compendium II, Copyright Office Practices § 905.02 (1984). See also, *Estate of Martin Luther King, Jr. v. CBS, Inc.*, 194 F.3d 1121 (11th 1999) ("I Have a Dream" speech of Dr. Martin Luther King heard by thousands and broadcast to thousands more was not a publication.)

54 *Getaped.com v. Cangemi*, 188 F.Supp.2d 398 (S.D.N.Y. 2002). Events: website revised in June, infringement in July, registration in August, litigation follows. "Consequently, when a website goes live, the creator loses the ability to control either duplication or further distribution of his or her work. A webpage in this respect is indistinguishable from photographs, music files or software posted on the web-all can be freely copied. Thus, when a webpage goes live on the Internet, it is distributed and 'published' in the same way the music files in *Napster* or the photographs in the various *Playboy* decisions were distributed and 'published.'" Id. at 402.

55 JOHN W. HAZARD, JR., COPYRIGHT LAW IN BUSINESS AND PRACTICE § 1:5 (2008).

Copyright licenses and legal deposit practices of grey multimedia materials

Debbie L. Rabina, Pratt Institute; School of Information and Library Science, United States

Abstract
The purpose of this study is to determine whether the type of copyright license under which grey multimedia materials are published makes a difference in terms of their inclusion in library catalogs. The two types of copyright licenses examined are Creative Commons and traditional copyright, and the library catalogs examined is that of the United States Library of Congress and national catalogs of countries represented in the population of the study. The population included grey multimedia materials whose main use is as educational materials, with half the items licensed under traditional copyright license and half licensed under Creative Commons license. The main finding of the study is that Creative Commons license is a barrier to access in traditional bibliographic utilities, and that traditional copyright is a barrier to access in web 2.0 environments. In addition, the study found weak enforcement of legal deposit laws regarding multimedia materials.

Introduction

Multimedia is everywhere and awareness and recognition of it growing.[1] As noted in a recent New York Times article,[2] we consume multimedia everywhere, anywhere, and all the time: in a taxicab, on a plane, on the street, in front of our computer, on our mobile devices. We watch and listen, often not by our own initiative and often reluctantly. The multimedia surrounding us is mostly push technology: publishers, vendors and others involved in the creation and distribution of content have their content visible on giant billboards, on the backs of taxi seats, and in stores. When it comes to push technology multimedia, we seem unable to shut off the current, but what about pull-technology? When we want the single commercial that we feel will best demonstrate to our students the importance of information to a global society, or that campaign ad that expressed a value-based approach to information, we are at a loss as to where to find them. An added difficulty is that much of multimedia is grey by nature: it is published outside the traditional scholarly communication powerhouses and has limited, if any, bibliographic control. Repositories for multimedia are few and usually highly specialized, limiting the scope of collections. Users of multimedia content, particularly academic faculty, such as those involved with this study, are not only users of information, but increasingly they are asked to be organizers as well. The amount and types of information used in learning and teaching now includes image files, sounds files, movies of varying lengths, web pages, multi authored documents and more. We store these materials on our personal computer's hard drive, in our space on a variety of web 2.0 applications such as YouTube or Flicker, we manage them with tags and list them on multiple bibliographies such as LibraryThing, Zotero and more, but when the moment of truth arrives, few of us are able to locate all this multimedia and we all end up doing a Google search and hoping for the best.

But beyond problems resulting from lack of organization, other barriers to access to multimedia materials also exist: Copyright restrictions and confusion regarding type of copyright license, fair use and permissible use of multimedia materials also hinders use. Multimedia licensed under traditional copyright license is more likely to benefit from existing methods of storage, organization and bibliographic control, but also more likely to have restrictions on use, while multimedia licensed under copyright alternatives, such as creative commons, may have less restrictions on use but may also be less findable. Another layer of complexity is added by the fact that while some countries have legal deposit requirement for multimedia materials, others do not. This study wishes to examine how copyright licenses and legal deposit legislation interact to effect access to grey multimedia materials.

Copyright licenses

Copyright in a personal property right that protects creators and allows them to benefit from the fruit of their intellectual and creative work. In the United States copyright, as part of intellectual property, is a constitutional right and protected by the U.S. Copyright Law. In the United States, copyright is automatically awarded to all creators, without preconditions such as registration or legal deposit[3]. The same holds true to all citizens whose countries are signatory to WIPO.[4]
Copyright is therefore inherent to all intellectual work and giving up rights in a creation is an opt-out option, an act that requires purposeful action. Copyright owners may choose to transfer their rights, with or without preconditions, to others, and allow them to use their creation in ways that are outside the scope on traditional copyright law.

The transfer of copyrights occurs most typically when creators award their exclusive rights to others who help promote their work, in exchange for profiting from it. For the most part, copyrights are transferred from authors to publishers.

The opposite of having a work under copyright is having it in the public domain. While most works reach the public domain due to legal requirements (government publications, expired term-limit of protection, failure to renew copyrights), creators may also choose deliberately to place their work in the public domain. In recent years, alternative to the traditional copyright (symbol©) have emerged, particularly as it pertains to content on the internet. Collectively, these alternative copyright licenses are known as Copyleft, although the terms itself is conceptual and has no legal or contractual implications. The most established among the copyright alternative is the Creative Commons license (symbol CC)[5]. Creative Commons is the brainchild of law professor Lawrence Lessing and was founded as a non-profit organization in 2001. It serves as an alternative to traditional copyright, allowing creators of content to set their own perimeters of use for their creative works, retaining some or no right to them. Under the Creative Commons license, holders of copyright can allow others to use their materials with some or no restrictions, specifically, any combination of attribution (allowing use of your work under the condition of crediting the original author), non-commercial (allowing use of the work for non-commercial purposes), no derivatives (allowing use of the work as long as it is left intact and no derivative works are based on it) and share alike (allowing use of the work as long as the new works based on it are shared under the same copyright license as the original).
Another alternative to copyright is GNU[6], known predominately to users of open source software. GNU General Public License (GPL) (symbol: Ram) license is specifically for computer programs and associated most often with Linux. The main provision of GNU is "share alike". What is common to all these copyright alternatives is that while traditional copyright is often viewed as censorship and inhibitor of access,[7] alternative copyright licenses are viewed as being to the benefit of users since they do not restrict access.

Legal deposit
Dating back to the days of King Francois I, the legal deposit was enacted first in France in 1537 and now exists in many countries. The main purpose of legal deposit law is to guarantee that countries will be able to preserve the cultural heritage of their nation. This is done by systematically collecting, storing and recording the print output of the country. While laws differ greatly in detail, the main provision of most legal deposit laws is that publishers are required to send a specified number of copies of each publication they produce to a specified library or libraries. Once received, the publications are recorded in either a catalog or national bibliography, some copies are stored for posterity, and some are made available to the public for research or recreational use. This mechanism ensures bibliographic control, access, and a collection of last resort to a country's print culture. Legal deposit has long been considered by librarians to be a vital link in the chain that allows bibliographic control as well as a safeguard against the loss of cultural heritage.

In the United States, legal deposit is required by the Library of Congress for all works under copyright protection, with a limited and specified list of items exempt from legal deposit requirements[8]. Works licensed under Creative Commons licenses are not addressed by the U.S. Copyright law. One could argue that as long as some right are retained to a work licensed under Creative Commons, legal deposit applies.

One of the advantages to publishers from legal deposit is the inclusion of deposited materials in the Library on Congress catalog and the assignment of a deposit number subfield in the items MARC record (field 017 of MARC 21 record format). Creators of content who do not deposit copies of their work with the Library of Congress are at a disadvantage in terms of making the record of work known and available to a wide audience.

Problem statement
With the proliferation of multimedia materials for both formal and informal learning environments, the question of access to these materials gains urgency. What are the sources that educators, researchers and interested individuals turn to for multimedia materials, and what are the barriers to access? Traditional copyright licenses and lawsuits filed against educational institutions for copyright violation have created a culture of fear among librarians and educators which is inhibits access. Images and multimedia materials are used like never before. In the past they were mostly the domain of those with an interest in visual arts, but today they are created and used by educators in all disciplines. While clearinghouses and registries for copyright holders are well established for the print works[9], such parallels do not exist for multimedia materials.

The use of multimedia in education is growing, as well as the discussion devoted to issues relating to its use, as articulated in The Journal of Educational Multimedia and Hypermedia.[10] This research wishes to examine the landscape of grey educational multimedia materials in terms of copyright licenses and

traditional tools of bibliographic control. Specifically, the research attempts to answer the following question:
Are grey multimedia works carrying creative commons licenses more or less likely to be included in traditional tools of bibliographic control such as national bibliographies?

Literature Review
In a series of articles published beginning in 2001, Michael Seadle addressed several of the issues pertinent to this study, although they were not examined in an integrated approach. Specifically, Seadle examined multimedia fair use[11], digital legal deposit[12], and grey copyright[13].

Fair use for educational multimedia in the United States is directed by a non-legislative report by the Council of Fair Use (CONFU) on "Fair use guidelines for educational multimedia" that was adopted by the Subcommittee on Courts and Intellectual Property" in 1996.[14] While not a legally binding report, it does define the scope of fair use for educational multimedia materials that carry traditional copyright licenses. Unfortunately, the CONFU guidelines are rather restricting in the use of multimedia materials and perhaps as a result, the guidelines are relatively unused,[15] and are considered to fall short in several crucial ways, mainly that in virtually all cases, fair use for educational multimedia is almost non-existent and requires permission for use.[16] While attempting to disentangle the murky waters in which legal deposit for digital materials, among them education multimedia, exists, Seadle[17] offers an interpretation that while based on existing U.S. copyright law, also introduces issues not specifically addressed by the law and encourages major internet-publishing countries, such as the U.S., U.K. and Germany, to develop policies and procedures for handling the legal deposit of digital materials.

Copyright confusion describes situations where users are uncertain about the copyright status of a work and the extent of permissible fair use. A recent report by the Center for Social Media at American University found that critical teaching is compromised as a result of copyright confusion.[18]. As a result of the complexity of copyright law, along with restrictions imposed by parent institutions, educators sometimes opt not to use certain materials, or apply copyright provisions erroneously.[19] The report's recommendation, that teachers educate themselves regarding clear and unambiguous use rights, is easier said than done.

Adding to the confusion regarding fair and permissible use is the multitude of copyright licenses that have developed over the past decade, particularly Creative Commons licenses, developed by Lawrence Lessig of Stanford University in Dec. 2002.[20] In a recent study[21] Kim examined whether Creative Commons can offer solutions to problems of copyright in digital environments and found that Creative Commons licenses offer a greater degree of flexibility for digital environment than traditional copyright, but did not address the question of bibliographic access and control.

Justification and research questions
As the literature review demonstrates, prior research examining access to grey multimedia materials based on type of copyright license is scant and piecemeal. In the absence of a legislative framework as well as prior research addressing access to grey multimedia materials carrying different copyright licenses educators who wish to access and use multimedia for classroom materials have a difficult time both learning about the existence of multimedia materials suitable for their purposes, and in understanding the conditions for their use. This research wishes to fill a gap by learning whether Creative Commons licenses that are usually seen as promoting use, also limit bibliographic access.
Specifically, the following research questions (RQ) will be addressed:
RQ1: What are the characteristics of multimedia are used by researchers participating in the study
RQ2: Are works carrying traditional copyright more, less, or just as likely to be included in national bibliographies as works carrying Creative Commons licenses?

Methodology
Syllabi of courses taught by researchers and educators in graduate departments of library and information science were harvested and lists of multimedia materials used for their teaching and research were compiled. An initial list containing two hundred items was collected. Each item was examined to identify the type of copyright license it carried and only items carrying either traditional copyright or Creative Commons licenses were included. Next, the country of origin (i.e.: publication or production) of each item was noted and the four top counties that emerged were the United States, Australia, Belgium and Israel. Other counties identified had a number of items deemed too small to provide meaningful data and were excluded from the study. The final research population consisted of a list of 117 items that met the following conditions: they carried either Creative Commons or traditional copyright and they were published in one of the four countries mentioned above. The majority of items on the list were films with music and games also being represented.

Next, list checking,[22] a methodology used primarily for collection development, was used to determine inclusion in bibliographic utilities. The bibliographic utilities chosen were national catalogs, OCLC WorldCat, and the increasingly ubiquitous YouTube, as well as NetFlix. YouTube and NetFlix were added to the more traditional national catalogs and WorldCat after faculty interviewed indicated they are just or more likely to search for movies on the non-traditional sources than in the traditional sources.

Findings

The research population used for this study was found to be relatively new, spanning in years from 2001 to 2008, with newer items receiving higher representation. Countries represented include the United States (75% of all items) followed by Australia, Belgium and Israel, each with just about 8% of items. Most items represented were movies, either on DVD or downloadable, ranging in length from 17:37 minutes to almost 3 hours.

When checked against traditional and non-traditional collection, WorldCat included the largest percentage of items (41.6), followed by Netflix (33%), YouTube (30%) and Library of Congress (8.3%).

Next, items were divided into groups according to the type of license they are licensed under: Creative Commons or traditional copyright. When examined this way, it was found that for traditional copyright, 57% of multimedia were available in Library of Congress or WorldCat and 20% were available complete on YouTube. As for items carrying Creative Commons licenses, 16% were available on Library of Congress or WorldCat and 50% were available complete on YouTube.

The representation in national catalogs was very low. Only 8.3% if items were found in the Library of Congress catalog, with none in the items represented in the national catalogs of their respective countries, regardless of whether or not these countries have legal deposit requirements for multimedia materials.

Discussion and conclusions

This study set out the answer two main research questions: First, we wanted to identify the general characteristics of multimedia material used by faculty in graduate programs of library and information science. Results indicate that the majority of materials used are less than a decade old, originate from the United States, and are mostly movies. Secondly, we wanted to know whether different copyright licenses are represented the same or differently in national catalogs. Results indicate that works licensed under Creative Commons are less likely to be included in national bibliographies than worked carrying traditional copyright. The results indicate that Creative Commons is a barrier to access in traditional tools while traditional copyright is a barrier to access in Web 2.0 environments

The analysis points to some other findings that are worthy of further examination. First, the study was structured around the assumption that legal deposit laws were systemically applied to multimedia in those countries where there are appropriate provisions in the law. This assumption proved wrong and results seem to indicate that multimedia materials are not deposited as required by law. Second, traditional bibliographic utilities such as national catalogs and WorldCat are not reliable tools for searching for multimedia materials, regardless of copyright license type, and national catalogs are particularly inept when trying to locate multimedia carrying Creative Commons licenses. At the same time, it seems that other more popular Web 2.0 tools such as YouTube and Netflix are gaining dominance and a larger percentage of materials from the study were included in them than in professional bibliographic utilities. While this study initially set out to examine only inclusion in national catalogs, Web 2.0 tools were examined since faculty indicated these are the sources they are most likely to turn to first when searching for multimedia materials. It seems that application and enforcement of legal deposit of non-print materials is weak, regardless of type of copyright license, implying disappearance of many of these materials.

While Creative Commons licenses may encourage use, they do not seem to provide access to metadata about multimedia materials implying that knowledge about the existence of Creative Commons materials may not survive in the long run.

Recommendations

Results from this study lead to several questions that should be examined in follow-up studies. First, what is the level of enforcement legal deposit requirements for multimedia materials? Second, does the young age of multimedia materials used in the study (2001 being the earliest) suggests that access to older materials is being lost? Finally, is there a way to achieve long-lasting bibliographic control of gray multimedia materials while limiting restrictions that results from traditional copyright.

References

1 Recently, UNESCO has reserved Oct. 27 at World Day for Audiovisual Heritage http://portal.unesco.org/ci/en/ev.php-URL_ID=25525&URL_DO=DO_TOPIC&URL_SECTION=201.html

2 Kelly, K. (2008, Nov. 23), Becoming screen literate. – In: The New York Times Magazine, pp. 48-57

3 The United States dropped registration and legal deposit as a prerequisite for copyright in 1978, in preparation for entering the WIPO agreement, as WIPO does not allow conditions for copyright protection.

4 WIPO Copyright Treaty http://www.wipo.int/treaties/en/ip/wct/trtdocs_wo033.html

5 More about Creative Commons at the Creative Commons website http://creativecommons.org/about/

6 GNU General Public License http://www.gnu.org/copyleft/gpl.html

7 For further discussion of copyright as censorship and inhibitor of access, see Lessing, L. (2008 Oct. 21), Copyright and politics don't mix. – In: The New York Times http://www.nytimes.com/2008/10/21/opinion/21lessig.html ; and, Mazzone, J. (2006), Copyfraud. – In: Brooklyn Law School, Legal Studies Paper no. 40; New York University Law Review, 81, p. 1026. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=787244

8 See part *202 of 37* CFR, *Chapter II* for list of exempt materials http://www.copyright.gov/title37/202/index.html

9 See Copyright Clearance Center http://www.copyright.com/

10 See Journal for Educational Multimedia and Hypermedia http://www.aace.org/pubs/jemh/.

11 Seadle, M. (2001), Copyright in the networked world: Multimedia fair use. – In: Library Hi Tech, 19 (4), pp. 422-425. ISSN 0737-8831

12 Seadle, M. (2001), Copyright in the networked world: Digital legal deposit. – In: Library Hi Tech, 19 (3), pp. 299-303. ISSN 0737-8831

13 Seadle, M. (2008), Copyright in the networked world: Gray copyright. – In: Library Hi Tech, 26 (2), 325-332. ISSN 0737-8831

14 United States Patent and Trademark Office. The Conference on Fair use. http://www.uspto.gov/web/offices/dcom/olia/confu/report.htm . See also summary from University of Texas http://www.utsystem.edu/ogc/intellectualproperty/ccmcguid.htm

15 Seadle, M. (2001), Copyright in the networked world: Multimedia fair use. – In: Library Hi Tech, 19 (4), pp. 422-425. ISSN 0737-8831

16 Sundt, C.L. (2002), The CONFU ditigal image and multimedia guidelines: The consequences for libraries and educators. University of Oregon http://darkwing.uoregon.edu/~csundt/copyweb/indy.htm

17 Seadle, M. (2001), Copyright in the networked world: Digital legal deposit. – In: Library Hi Tech, 19 (3), pp. 299-303. ISSN 0737-8831

18 Hobbs, R., P. Jaszi, P. Aufderheide (2007), The cost of copyright confusion for media literacy. Center for Social Media, School of Communication, American University. http://www.centerforsocialmedia.org/files/pdf/Final_CSM_copyright_report.pdf

19 Hobbs et al.

20 Creative Commons http://creativecommons.org/

21 Kim, M. (2007), The Creative Commons and copyright protection in the digital era: Uses of Creative Commons licenses. – In: Journal of Computer-Mediated Communication, 13 (1) ISSN 1083-6101. http://jcmc.indiana.edu/vol13/issue1/kim.html

22 For studies using list checking as a primary methodology see: Lotlikar, S.D. (1997), Collection assessment at the Gasner library. – In: Collection Buliding 16 (1), pp. 24-29. ISSN 0160-4953, and ; Moss, E. (2008), An inductive evaluation of a public library GLBT collection. – In: Collection Building 27 (4), pp. 147-156. ISBN 0160-4953

The "Grey" Intersection of Open Source information and Intelligence

June Crowe and Thomas S. Davidson,
Open Source Research Group; IIA, Inc., United States

Abstract
The U.S. Government intelligence community (IC) is relying less on "classified" information as a sole source and is moving toward an "all source" product that includes open source information. Gradually, this change will result in a more comprehensive, virtual IC that will enhance or replace the smaller, classified collections of individual government bureaucracies. This has created an emerging paradigm involving technology, outsourcing, and relationships both inside and outside of the IC that has resulted in "grey intersections" of open source information and intelligence. This paper will define "open source information" and look at specific government actions that have boosted support of open source information intelligence (OSINT), as well as the ongoing struggles within the IC to accept the new paradigm. The intelligence cycle as it applies to open source is described, and examples of the use of open source information are given in the context of their reliability and classification. The paper also discusses the future of open source and intelligence.

Introduction

Open source information has been recognized as a significant source of intelligence by some in the U.S. Government intelligence community (IC) in the mid-to-late 1990s. However, it was always considered the source of last resort by many in the intelligence community until recently. In novels like "*Out Sourced*" by R.J. Hillhouse, which describes how contractors work in Iraq, the line between what was compiled from open source and what some analysts may recognize as "classified intelligence" is blurred. Analysts may recognize some information as being classified; other readers will not know this. However, it is clear that the author appreciates the value of open source information.

Today, however, the IC is relying less on "classified" information as a sole source and is moving toward an "all source" product that includes open source information. Gradually, this change will result in a more comprehensive, virtual IC that will enhance or replace the smaller, classified collections of individual government bureaucracies. This has created an emerging paradigm involving technology, outsourcing, and relationships both inside and outside of the IC that has resulted in "grey intersections" of open source information and intelligence.

The IC is concerned about the grey intersection of open source and intelligence, including reliability, classification needs, and the quantity and quality of information to be analyzed. There are also concerns of how open source will affect the responsibilities of the various members in the intelligence community. The new paradigm has been brought about primarily by advances in information technology which allow for customized systems, federated searching, common platforms that make networking easier and allowing data to be shared and exchanged, and a general decentralization of intelligence systems.[1] The combination of technological changes and the value of open source information have begun to blur some of the distinctions between human intelligence (HUMINT) and open source intelligence (OSINT).

This paper will discuss some of the grey intersections of open source and intelligence and examine how the new paradigm will affect the intelligence community. It will also provide some specifics as to the use, evaluations, and examples of open source by the IC, as well as look at the Open Source Intelligence Cycle and the future of open source and the IC.

Definition of Open Source

The term "open source" refers to information that is derived from overt, non-clandestine sources as opposed to hidden or covert collection.[2] The IC defines open source information as "information which is publicly available and can be lawfully obtained by anyone by request, purchase, or observation."[3] Open source information can be on almost any topic, including economics, health, social and cultural, political, military, energy and the environment, and demographics. It includes media; public data included in government reports, demographics, budgets, conferences, symposia, academic papers, dissertations, theses, and experts; commercial data; and grey literature. However, once open source information is collected, even that which is obtained from outside experts, it may be classified or receive a protective marking or caveat to prevent showing information gaps in the U.S. IC databases or for other reasons. Some reports produced by government contractors may fall into this grey intersection of open source and intelligence.

At what point does open source information become "intelligence"? Open source information becomes intelligence when it is collected, exploited, and disseminated to address a specific intelligence requirement. Intelligence personnel refer to open source "collection," whereas non-intelligence analysts

prefer to use the term "acquisition," as the information has already been collected and is publicly available. The U.S. Joint Chiefs of Staff define intelligence as "the product resulting from the collection, processing, integration, evaluation, analysis, and interpretation of available information concerning foreign nations, hostile or potentially hostile forces or elements, or areas of actual or potential operations. The term is also applied to the activity which results in the product and to the organizations engaged in such activities."[4]

The U.S. Army distinguishes between open source collection—the unintrusive collection of publicly available information in the course of authorized and assigned missions with the intent to use or retain the information for foreign intelligence or counterintelligence purposes—and open source research.

> *Collection responds to reconnaissance and surveillance missions levied on intelligence and non-intelligence organizations through tasks and requests. Focused and synchronized through collection management, open source collection resources know who, what, when, where, and why to collect publicly available information and essential metadata (date, time, location, language, frequency, station identification, newspaper name, author). Essential metadata includes temporal and geospatial data that enables tracking and visualization of activities, changes in the operational environment, and coverage of reconnaissance and surveillance resources.*[5]

The U.S. Army defines open source (research) as "information gathered without the expectation of privacy—the information, the relationship, or both are not protected against public disclosure."[6] "Open source intelligence" is defined as "information of potential intelligence value that is available to the general public. Relevant information derived from the systematic collection, processing, and analysis of publicly available information in response to intelligence requirements."[7] The definitions of open source research and OSINT are similar with the exception that open source intelligence is driven by intelligence requirements. "The leveraging of OSINT to the greatest extent possible as a means to provide the needed information to the widest audience at the lowest level of classification" is part of the strategic goals, objectives, and priorities of the *Defense Intelligence Strategy 2008*.[8]

U.S. Government Actions in Support of Open Source

The United States has made some remarkable changes in policy since the July 2006 *Implementation Roadmap for National Open Source Enterprise* was released by the IC. The Roadmap includes the following initiatives, many of which were in place at the end of 2008:

- Establish and use the National Open Source Committee (NOSC) enterprise governance structure to:
  - determine and refine the IC open source capabilities and resource baseline; identify agencies and sources for open source exploitation;
  - manage IC open source resources more effectively;
  - improve input and collaboration.
- Implement an IC-wide open source training and certification program. This program will standardize open source training for all within the IC. The OSC will also provide a mobile training team for all IC open source cells regardless of agency.
- Develop and deploy the OSC Expertise Franchise Program. This program will provide mentors and subject matter experts to assist in developing and improving Open Source Cells across the IC.
- Drive rapid IC progress toward a single open source information technology (IT) architecture which would include single sign-on, federated search, one-way transfer, and an anonymous Internet research functionality.
- Publish a National Open Source Exploitation Capabilities Manual. This manual has already been published with a protective caveat.
- Establish an unclassified IC Open Source Works innovation facility. This facility is used to test ideas and capabilities. Computer technicians are brought to assist in making databases more accessible and user friendly.
- Develop open source requirements management and tasking systems. The management system will cover such items as collection management to include life cycle management of materials. The tasking system will involve receiving users' information requirements and the tasking of the appropriate open source cells. For example, the Foreign Military Studies Office (FMSO) deals in primary language open source materials. If a user's requirements included foreign language expertise, FMSO would be an appropriately tasked agency.
- Invest in IC Research Librarians. This particular problem is not the lack of research librarians but rather the lack of librarians with the appropriate clearance.
- Improve contributions of potential international and academic partners.
- Sponsor an annual Director of National Intelligence (DNI) Open Source Conference open to the IC, consumer agencies, private industry, academia, and international partners.[9]

The first DNI conference was held in July 2007 and the second in September 2008. Conference attendees included academia, contractors, foreign and domestic U.S. military personnel, and other U.S. government employees. The DNI has strengthened its partnerships with academia and private industry and has created an Open Source Collection Acquisition Requirements (OSCAR) management system that is designed to connect intelligence consumers, analysts, and collections requirements managers with providers of quality open source intelligence resources.[10] Intelligence is no longer considered just a national issue but an international one, and open source information is seen as being a major component of that effort that is embodied in DNI's new *Vision 2015* which envisions a globally networked and integrated intelligence enterprise.[11]

At the September 2008 conference in Washington, DC, the virtues of OSINT were extolled by many in the IC with great enthusiasm. Prior to 2005, the use of non-classified, open source information was seen as a source of last resort instead of a primary resource. Now officials are indicating that there can be good reasons for classifying some open source information products; however, the classification itself has become one of the grey intersections for intelligence. General Michael Hayden, Director of the Central Intelligence Agency (CIA), indicated that while the "information might be unclassified – our interest in it is not."[12]

The need for classification is usually based on the analysis of the open source information. However, the trend of integrating open source information into all-source products has led to the over-classification of open source materials in general. Nonetheless, open source information is gaining credibility for three primary reasons: ease of distribution – not everyone has top-secret clearances; easier access; and timeliness. Additionally, open source information provides insight into how others view the world and serves as a means of validating classified or sensitive source reporting.

There are some concerns among national open source policy makers that open source collectors need to be more selective in what is collected and produced. The consensus is that open source information products should be classified at the lowest level so they can be disseminated to the largest number of consumers possible. In spite of these concerns concerning selectivity, the support of open source information has become stronger because more people can be more informed in a more timely manner.[13] The IC has only recently begun to make deliberate efforts to build an infrastructure that enables them to share information easily.

U.S. Government OSINT Resources

There are a number of centers for open source information with various missions. One such center is controlled by the Office of the Director of National Intelligence (ODNI). This center provides policy documents, guidelines, reports to Congress, exploitation manuals, and other products and services in a variety of areas.

Five centers are listed on the DNI's website:

- Center for Security Evaluation (CSE) – The primary purpose is to plan and prepare for situational emergencies;
- Intelligence Advanced Research Projects Activity (IARPA) – IARPA is the innovation center for research projects; Information Sharing Environment (ISE) – The purpose is to facilitate and promote collaboration and information sharing;
- National Counterintelligence Executive (NCIX) – This center develops, coordinates, and produces annual foreign threat assessments and annual national strategies for the U.S. government, prioritizes collection, investigations, and operations, and other administrative functions;
- National Counterterrorism Center (NCTC) – The Center combats terrorism at home and abroad;
- Special Security Center (SSC) – The SSC is responsible for security policies and guidance on security practices and procedures. SSC personnel assist the DNI in sharing and protecting national intelligence information throughout the IC, the U.S. government, U.S. contractors, state and local officials, and our foreign partners.

More information concerning each of the Centers can be found on the DNI's website at http://www.dni.gov. These portals are part of the new technologies being integrated into the IC.

The Open Source Center (OSC), founded in 2005, incorporates the Foreign Broadcast Information Service (FBIS) as its foundation. FBIS was part of the Central Intelligence Agency (CIA). OSC is a center of excellence that facilitates distributed expertise and capabilities, collects intelligence data as it is archived, and fields requests for information from the Defense Department, civilian agencies, and state and local law enforcement agencies. The OSC also has open source certification training for government employees, directives issued concerning the open source information citations, and methods to be used for collecting this information. The OSC has a range of products and services, one of which is the password protected website www.opensource.gov. The head of OSC reports directly to the CIA chief.[14, 15]

Another well-known open source research and training center is controlled by the Foreign Military Studies Office at Fort Leavenworth, Kansas. The Foreign Military Studies Office (FMSO) is a research and analysis center under the U.S. Army's Training and Doctrine Command (TRADOC), Deputy Chief of Staff G-2 (Intelligence), and the TRADOC Intelligence Support Activity. Their researchers conduct analytical programs focused on emerging and asymmetric threats, regional military and security developments, and other issues that define evolving operational environments around the world.[16] The FMSO's OSINT training website is available at http://fmso.leavenworth.army.mil/OSINT-Training.pdf. Open source products are available at http://fmso.leavenworth.army.mil/products.htm. These two links are not password protected and available to the public.

New Technologies / Social Networking Tools

The IC is also changing in that it is adapting social networking technologies to collaborate and share information. Intelink was created in 1994 and has since grown to contain thousands of agency sites, along with several hundred databases, creating an information overload. Resolution of this problem has been hindered because of the system's poor indexing and dated search tools. The huge volume of open source information and the difficulty of assessing the source's credibility indicate a need for enhanced IT networking tools and technology. Assessing the credibility of information in these participatory information environments can be challenging because user-generated content can obscure or disconnect the origin of a source from established, reputable, institutional sources. This is particularly true on news sites where readers add comments about the original content.

More recently the IC has created a classified "A-Space" that includes blogs, searchable databases, libraries, and other tools to help analysts trade, update, and edit information. Another success story is Intellipedia, an experimental project started in 2005 by the ODNI (Office of the Directory for National Intelligence) that was a take-off on Wikipedia. Later that year it was deemed a success and rolled out in 2007 to the IC.

Throughout the IC, interactive Web tools, such as Facebook, YouTube, blogs and wikis, and other social networking sites, are being copied and incorporated into the new intelligence system. These tools have begun to revolutionize information sharing in the IC by providing mechanisms to incorporate open source information into the intelligence cycle. Still, not all 16 intelligence agencies have adopted inactive social networking and other tools.[17]

Impact on Intelligence Community Operations

Open source information is seen as a blessing and a curse by IC professionals; although information is more accessible, often more timely, and easier to disseminate, it has also impacted how the IC operates. No longer is there only one authoritative source of information.[18] The new paradigm is for collection to become more integrated with technology and with the other intelligence activities (HUMINT, OSINT, SIGINT (Signal Intelligence), GEOINT (Geospatial Intelligence), ELINT (electromagnetic intelligence), and MASINT (measurement and signature intelligence). These are sometimes referred to as "INTS." However, this also means that the relationship between the producer and the consumer needs to be closer. The producers of open source information products are often civilian contractors. Issues of contractor trust and reliability need to be resolved and the product requirements better defined for the contractor. Often contractors do not understand the consumers' needs for the information. Prior to 11 September 2001, the rule was "when in doubt, leave it out;" in the post 11 September 2001 environment, the rule is "when in doubt, put it all in." The only qualification is that the information needs to be user friendly and fit the consumer's daily mission.

As acceptance of open source information grows, there is a need for closer collaboration between the consumer and the producer of intelligence which may also create another "grey intersection" in the IC as uncleared contractors provide open source products to the IC. Open source means unclassified, but sometimes the government classifies the finished OSINT product using the information because it responds to specific tasking from the IC. Contractors perform a lot of research for the IC using open source research. In order to better respond to their information requests and produce a better product, closer collaboration will be required. The new products will show research results in innovative ways using geospatial images and mash-ups, a web application that combines data from more than one source into a single integrated tool.

Intelligence Cycle

In the past, the intelligence cycle was based upon the mission, such as responding to a natural disaster, war, or a terrorist attack. The on-scene commander's mission drove the intelligence cycle and, therefore, the collection of information. For every bit of information obtained and processed, new requirements were generated to provide the basis for effective planning and direction of the collection. Planning had to include personnel, equipment, beginning sources, and, finally, the development of a collection plan.

Today, the Intelligence Cycle is consumer driven as can be seen in Figure 1. The consumer is at all three levels of the government (local, state, and federal). Local government consumers include the local law enforcement personnel and medical and health elements; state government consumers include state law enforcement personnel, fusion cells, and medical and health departments; and federal government consumers include federal law enforcement agencies, Department of Defense, and the legislative and judicial branches. The consumer can be the first responder to a hurricane, the battlefield soldier, the U.S. Border Patrol, the police investigator, or the highest-level national decision maker. Open source information has become an essential part of the Intelligence Cycle. Although the basic planning procedures are still the same, today the intelligence officer uses open sources as well as classified assets as resources. The battlefield soldier or police investigator most probably does not have the same classified access as the national decision maker. Therefore, OSINT should be available to the general consumer who needs detailed information to understand his or her area of operations. Consumers at all levels benefit from accurate open source reporting and analysis.

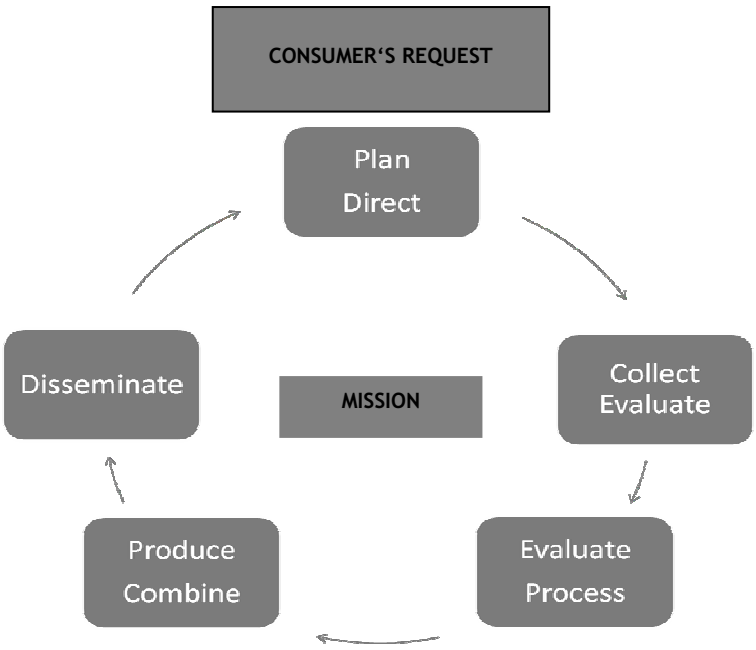


Figure 1. The Intelligence Cycle.

The intelligence cycle includes the following steps:

Step 1 – Planning and Directing is the beginning and the end of the Intelligence Cycle. Before mission start, intelligence requirements must be determined. Intertwined with the requirements are the planning of specific collection agencies, the processing, the analysis, and the dissemination requirements.

Step 2 – Collecting is the gathering and reporting of raw information which is needed to produce finished intelligence. Open source information can be obtained from a number of sources, including the Internet and government public records. In today's world, a number of countries have transparency laws, many of which are available through the Internet. These transparency laws provide the open source researcher a wider arena. During this process, the collector must evaluate both the information and the source in accordance with the evaluation criteria established earlier.

Step 3 – Processing involves the collation of the raw material into a form used by the analysts. Information management is the key to processing. Critical information is always processed first. Information processing includes the entry of the information into a database, reduction of duplications (circular reporting), collating web open source with paper files, and so on. To process information correctly, the collectors and the analysts must have a thorough understanding of the consumers' needs.

Step 4 – Production is taking the information from the previous steps to include information from classified sources and then integrating, evaluating, and analyzing all available information into a useable intelligence product in response to the original request. Once in this step, the subject matter specialists evaluate all information for contradictory information, reliability, timeliness, and relevance, and place the evaluated information into the required intelligence product in response to the commander's/consumers' requirements. These products can range from lengthy strategic evaluations for national decision makers all the way down to one-paragraph reports for the tactical level.

Step 5 – Dissemination is the end of the cycle which leads immediately to Planning and Dissemination. Commanders/consumers receive the finished reports applicable to them with appropriate classification. If the reports are of value and the analysts have done their job properly, new requirements will be generated and submitted to the appropriate intelligence agency, and the cycle begins again.

Open Source Reliability Issues

Reliability issues concerning adopting open source as a primary resource is still ongoing within the IC. Some more conservative analysts question the reliability of open source information. Open source information must be evaluated by both the collector and the analyst in a number of areas. The four categories for this evaluation are:

- Competence
- Veracity
- Objectivity
- Observational sensitivity.[19]

The U.S. government has established a reliability- and credibility-rating scheme, as shown in Tables 1 and 2. Both the source and the actual information are evaluated under this scheme. The scheme is a valuable tool for both the collector and the analyst. [20]

Table 1. Source Reliability Rating Scheme

CODE	RATING	DESCRIPTION
A	Reliable	No doubt of authenticity, trustworthiness, or competency; has a history of complete reliability; usually demonstrates adherence to known professional standards and verification processes.
B	Usually Reliable	Minor doubt concerning authenticity, trustworthiness, or competency; has a history of valid information most of the time; may not have a history of adherence to professionally accepted standards, but generally identifies what is known about sources feeding any broadcast.
C	Fairly Reliable	Doubt of authenticity, trustworthiness, or competency, but has provided valid information in the past.
D	Not Usually Reliable	Significant doubt about authenticity, trustworthiness, or competency, but has provided valid information in the past.
E	Unreliable	Lacking in authenticity, trustworthiness, and competency; history of invalid information.
F	Cannot Be Judged	No basis exists for evaluating the reliability of the source; new information source.

Table 2. Information Credibility Rating Scheme

CODE	RATING	DESCRIPTION
1	Confirmed	Confirmed by other independent sources; logical in itself; consistent with other information on the subject.
2	Probably True	Not confirmed; logical in itself; consistent with other information on the subject.
3	Possibly True	Not confirmed; reasonably logical in itself; agrees with some other information on the subject.
4	Doubtfully True	Not confirmed; possible but not logical; no other information on the subject.
5	Improbable	Not confirmed; not logical in itself; contradicted by other information on the subject.
6	Misinformation	Unintentionally false; not logical in itself; contradicted by other information on the subject; contradiction confirmed by other independent sources.
7	Deception	Deliberately false; contradicted by other information on the subject; contradiction confirmed by other independent sources.
8	Cannot Be Judged	No basis exists for evaluating the validity of the information.

These two charts are then joined together for the final evaluation of a source and the information provided. It is possible to have an "F-8" as a final evaluation. In fact, it is quite common. This evaluation (F-8) should not be considered derogatory. F-8 simply means that there is not enough information available to evaluate both the source and the information.

Source Evaluation
When a collector/analyst evaluates a source, the three items which must be evaluated are:
1. Author – What are the author's credentials, including – institutional affiliation, education background, writings, and/or experience? [21] To what extent is the author controlled – by the government, commercial superiors, financial benefit, or ideology? Has the author published previously? Does the collector/analyst have experience with the author? If associated with a business, organization, publication, government, or other institute, what is the reputation and goals of the organization?

2. Date of Publication – When was the article published? If an instantaneous report (newspaper report), then not all of the facts are in. Eyewitness reports are not always accurate. Is this the first time the article was published? If not, what edition is it? If the article is too old, then the information is most probably dated. However, the article must still be reviewed for items of interest, because as time goes by more facts come into the light.

3. Publisher – The publisher should undergo the same type of examination as the author.

When a collector/analyst evaluates content, the following must be reviewed:
1. Intended Audience – Who is the intended audience? Is the article biased? Is the information a government's propaganda message? Is the article targeted towards a particular ideology or belief system?

2. Information – Is the material presented factual, opinion, or propaganda? Are the facts consistent with what is known? Are the newly presented facts consistent and logical with previous information? Does the author present the information in an objective and impartial manner? There may be information of value in spite of the slant.

For example, The *National Enquirer* is generally considered an extremely unreliable source of information. Most analysts would rate the newspaper as a "D" or "E." Articles printed in the Enquirer would usually be rated "5," "6," or "7."

However, even these "muck raking" journals do occasionally publish articles which can be rated as "1." In September 30, 2008, an article was published about John Edwards and his mistress. In the article, he is referred to as the "vice-presidential candidate" and not the "presidential candidate." Therefore, the source and the article would be rated as D-2.

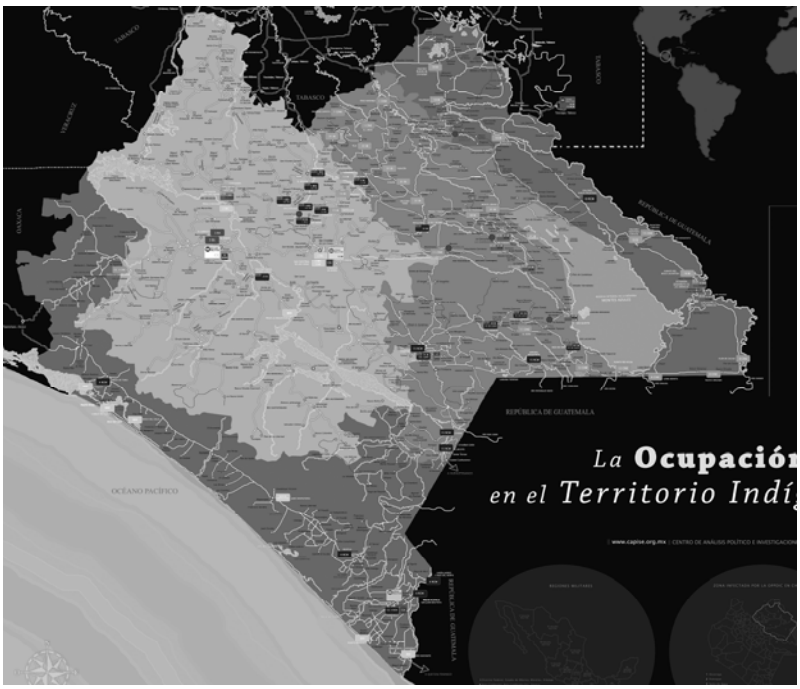


Figure 2. Map of Mexico with Identification of Military Units by Designation and Location.

A prime example of ideologically motivated but useful open source information can be found at website www.capise.org.mx.
The map depicted in Figure 2 was published by CAPISE, an organization calling for complete autonomy for all indigenous tribes in the Mexican state of Chiapas. CAPISE members travelled throughout the state of Chiapas and, with a global positioning system (GPS), identified all Mexican military installations in the state.

Of interest on this map is the identification of military units by designation and location. Through the judicious use of other Mexican government websites such as the immigration, national police, and military sites and the Mexican *Ley de Transparencia* (Transparency Law), an analyst will be able to judge the efficiency of the individual units in regards to stopping drugs and illegal immigration across the Mexican-Guatemalan border and along the Gulf coast. The efficiency of these units is especially valuable when units are transferred to the U.S.-Mexican border. This type of information is of interest to the Department of Homeland Security and the U.S. Border Patrol in particular. Most of the information contained on this map has been confirmed from other sites, but not all. In addition, all indications are that the organization has received support from some

leftist European organizations, so some of the original information obtained by CAPISE workers may have been eliminated or omitted for ideological reasons. For these reasons, the rating for this information would be C-2 or even C-3.

Classification and Security Markings

Open source information, as mentioned earlier, can sometimes be classified for various reasons. For instance, open source information about the numbers of cars and trucks crossing a bridge may not be of interest to the IC. However, if the information is coupled with a threat attack or the bridge happens to be a main connector to a large city, then the information may be of interest to the IC and some type of protective marking or classification may be applied. Currently, the U.S. government (executive, legislative, and judicial branches) has more than 100 classifications markings and protective caveats for intelligence information. Moreover, a myriad of laws and regulations govern the protection of sensitive but unclassified information, such as whether it was derived from overt sources, its public availability, and the reason and way it is analyzed. The U.S. Army has developed guidelines for the classification of open source information in its *Open Source Intelligence Field Manual*.[22]

In 2008, the White House released a new policy on "controlled unclassified information" (CUI) which establishes a new framework for designating, marking, safeguarding, and disseminating information designated as CUI.[23] Within the new framework, "sensitive but unclassified" (SBU) falls into three categories:

- Controlled with standard dissemination.
- Controlled with specified dissemination.
- Controlled enhanced with specified dissemination.[24]

The new policy is meant to facilitate information sharing within the government but will do nothing to restore public access to government records that have been improperly withheld nor will it exclude anything that is currently controlled as SBU.[25]

THE FUTURE OF OPEN SOURCE AND INTELLIGENCE

Open source information has been integrated into the intelligence cycle and is becoming considered as a source of first resort by many in the IC. The combination of open source becoming a category of intelligence (OSINT) and technological innovations have made it possible for the IC to share data and collaborate using social networking tools and common platforms. Open source products are increasingly becoming grey intersections because of the trend to assign protective caveats or low level security classifications to them. The new paradigm of integrating open source information into an all-source product, using new social networking technologies, and common computing platforms, and the outsourcing of some research, has blurred the distinctions of some intelligence categories.

Additionally, there is the grey intersection created by cleared and uncleared personnel collaborating on intelligence activities. The need to provide broad access to open source information that has been synthesized and melded with other intelligence will require close collaboration with experts in academia, the commercial sector, think tanks, and foreign intelligence networks. The new paradigm will be a virtual community with many players that will contribute to an all-source product that does not rely as much on "classified information" but more on open source information.

The Director of National Intelligence, Mike McConnell, explained the future of open source information/intelligence as follows:

> *The elusive, transitory nature of our targets, and the imbalance between the increasing demand for information and the capacity of our means to collect it, require multiple, integrated collection systems. Each of the collection disciplines — human intelligence, signals intelligence, computer network exploitation, geospatial intelligence, measurements and signatures intelligence, open source intelligence, acoustic intelligence, and foreign materiel acquisition — will continue to play key roles, although their relative importance will almost certainly change over time.*
>
> *"No aspect of collection requires greater consideration, or holds more promise, than open source information; transformation of our approach to open sources is critical to the future success of Adaptive Collection."*[26]

Open source intelligence will never replace the other "INTs" but will augment classified intelligence, provide the contextual landscape needed to understand the total situation, and ultimately support our decision and policy makers.

Conclusion

The change in open source policy is all about mission integration that will rapidly pull in information and deliver clear decision advantage to the consumer. This will be accomplished by providing broad access to open source information that will be synthesized and melded with other intelligence. The collaboration among intelligence analysts and experts in academia, commercial sectors, think tanks, and allied foreign intelligence networks will result in more targeted, accurate, and timely OSINT for consumers at all levels.

Common Abbreviations and Definitions[27]

CIA – Central Intelligence Agency

COMINT – Communications Intelligence – Technical information and intelligence derived from foreign communications by other than the intended recipients. Always voice.

CSE – Center for Security Evaluation

CUI – Controlled Unclassified Information

DNI – Director of National Intelligence

ELINT – Electronic Signals Intelligence – Technical and geolocation intelligence derived from foreign noncommunications electromagnetic radiations emanating from other than nuclear detonations or radioactive sources. Communications between machines.

FBIS – Foreign Broadcast Information Service

FMSO – Foreign Military Studies Office

GEOINT – Geographic Intelligence (also called Geospatial Intelligence) – The exploitation and analysis of imagery and geospatial information to describe, assess, and visually depict physical features and geographically referenced activities on the Earth. Geospatial intelligence consists of imagery, imagery intelligence, and geospatial information.

GPS – Global Positioning System

HUMINT – Human Intelligence – A category of intelligence derived from information collected from or provided by human sources.

IARPA – Intelligence Advanced Research Projects Activity

IC – Intelligence Community – All departments or agencies of a government that are concerned with intelligence activity, either in an oversight, managerial, support, or participatory role.

ISE – Information Sharing Environment

IT – Information Technology

MASINT – Intelligence obtained by quantitative and qualitative analysis of data (metric, angle, spatial, wavelength, time dependence, modulation, plasma, and hydromagnetic) derived from specific technical sensors for the purpose of identifying any distinctive features associated with the emitter or sender, and to facilitate subsequent identification and/or measurement of the same. The detected feature may be either reflected or emitted.

NCIX – National Counterintelligence Executive

NCTC – National Counterterrorism Center

NOSC – National Open Source Committee

ODNI – Office of the Director of National Intelligence

OSC – Open Source Center

OSCAR – Open Source Collection Acquisition Requirements

OSINT – Open Source Intelligence

SBU – Sensitive But Unclassified

SSC – Special Security Center

SIGINT – Signals Intelligence – A category of intelligence comprising either individually or in combination all communications intelligence, electronic intelligence, and foreign instrumentation signals intelligence, however transmitted. Includes COMINT and ELINT.

TRADOC – U.S. Army Training and Doctrine Command

References

1 (ANL, June 2006) Dr. Andrew N. Liaropoulos. "A (R)evolution in Intelligence Affairs? In Search of a New Paradigm." Research Paper 100. Research Institute for European and American Studies (RIEAS). http://www.rieas.gr

2 (CRS, 5 Dec 2007, pp. 5-6) Congressional Research Services. CRS Report for Congress. Open Source Intelligence (OSINT) Issues for Congress. By Richard A. Best, Jr. and Alfred Cumming. http://www.fas.org/sgp/crs/intel/RL34270.pdf

3 (ICDN, 11 July 2006) Intelligence Community Directive Number 301 and P.L. 109-163, Sec 931. http://www.dni.gov/electronic_reading_room/ICD301.pdf

4 (FAS, 22 June 2007, p. 141) Joint Publication 2.0. Joint Intelligence. http://www.fas.org/irp/doddir/dod/jp2_0.pdf

5 (FMI, 5 Dec 2006, p. 5-1) Field Manual Interim. FMI No. 2-22.9, Department of the Army. Open Source Intelligence. Expires December 2008. http://www.fas.org/irp/doddir/army/fmi2-22-9.pdf

6 (FMI, 5 Dec 2006, p. 5-1) Field Manual Interim. FMI No. 2-22.9, Department of the Army. Open Source Intelligence. Expires December 2008. http://www.fas.org/irp/doddir/army/fmi2-22-9.pdf

7 (FMI, 5 Dec 2006, p. 5-1) Field Manual Interim. FMI No. 2-22.9, Department of the Army. Open Source Intelligence. Expires December 2008. http://www.fas.org/irp/doddir/army/fmi2-22-9.pdf

8 (SDI, 2008) Secretary of Defense Intelligence. James R. Clapper. Defense Intelligence Strategy. http://www.au.af.mil/au/awc/awcgate/dod/def_intell_strat_080501.pdf

9 (DNI, July 2006) Director National Intelligence. National Open Source Enterprise. Implementation Roadmap. (only hard copy)

10 (Butler, 11 Sept 2008) Daniel Butler, Remarks and Q&A by the Acting Assistant Deputy Director of National Intelligence for Collection. DNI Open Source Conference, 2008. Washington, DC. http://www.dni.gov/speeches/20080912_speech.pdf

11 (Butler, 11 Sept 2008) Daniel Butler, Remarks and Q&A by the Acting Assistant Deputy Director of National Intelligence for Collection. DNI Open Source Conference, 2008. Washington, DC. http://www.dni.gov/speeches/20080912_speech.pdf

12 (USNWR, 16 Sept 2008) U.S. News and World Report. By Alex Kingsbury. "Spy Agencies Turn to Newspapers, NPR, and Wikipedia for Information: The Intelligence Community Is Learning to Value "Open-source" Information." http://www.usnews.com/articles/news/national/2008/09/12/spy-agenc..

13 (USNWR, 16 Sept 2008) U.S. News and World Report. By Alex Kingsbury. "Spy Agencies Turn to Newspapers, NPR, and Wikipedia for Information: The Intelligence Community Is Learning to Value "Open-source" Information." http://www.usnews.com/articles/news/national/2008/09/12/spy-agenc..

14 (WM, 17 Sept 2008) Wired Magazine. By Noah Shachtman. "Open Source Intel Rocks—Sorry, It's Classified." http://blog.wired.com/defense/2008/09/download-hayden.html

15 (FMI, 5 Dec 2006) Field Manual Interim. FMI No. 2-22.9, Department of the Army. Open Source Intelligence. Expires December 2008. http://www.fas.org/irp/doddir/army/fmi2-22-9.pdf

16 (FMSO, 2008) The Foreign Military Studies Office. Joint Reserve Intelligence Center. Open Source Research and Analysis Training. http://fmso.leavenworth.army.mil/OSINT-Training.pdf

17 (IHR, 5 Sept 2007) International Herald Tribune. "Classified Social-networking System Promises to Help U.S. Spies, Talk, Collaborate." http://www.iht.com/bin/print.php?id=7397033

18 (USNWR, 16 Sept 2008) U.S. News and World Report. By Alex Kingsbury. "Spy Agencies Turn to Newspapers, NPR, and Wikipedia for Information: The Intelligence Community Is Learning to Value "Open-source" Information. http://www.usnews.com/articles/news/national/2008/09/12/spy-agenc..

19 (EHSR, 26 Sept 2008) Evaluating HUMINT Source Reliability. David Schum and Jon Morris, George Mason University (McGill Research Blog). http://sourcesandmethods.blogspot.com/2008/09/evaluating-humint

20 (FMI, 5 Dec 2006, p. 5-1) Field Manual Interim. FMI No. 2-22.9, Department of the Army. Open Source Intelligence. Expires December 2008. http://www.fas.org/irp/doddir/army/fmi2-22-9.pdf

21 (CUL, 2008) Cornell University Libraries. Critically Analyzing Information Sources. http://www.library.cornell.edu/olinuris/ref/research/skill26.htm

22 (FMI, 5 Dec 2006, p. 5-1) Field Manual Interim. FMI No. 2-22.9, Department of the Army. Open Source Intelligence. Expires December 2008. http://www.fas.org/irp/doddir/army/fmi2-22-9.pdf

23 (WH, 9 May 2008) White House. Memorandum for the Heads of Executive Departments and Agencies. http://www.whitehouse.gov/news/releases/2008/05/print/20080509-6.html

24 (FCW, 12 May 2008) Federal Computer Week. By Ben Bain. Sensitive but Unclassified Category Simplified. http://www.fcw.com/online/news/152506-1.html#

25 (FAS, 12 May 2008) Federation of American Scientists. White House Issues Policy on "Controlled Unclassified Info." http://www.fas.org/blog/secrecy/2008/05/white_house_issues.html

26 (DNI, 18 July 2008) Director of National Intelligence. Vision 2015. http://www.dni.gov/Vision_2015.pdf

27 (DOD, 12 Apr 2001) Department of Defense Dictionary of Military and Associated Terms, as amended through 30 September 2008. http://www.dtic.mil/doctrine/jel/new_pubs/jp1_02.pdf

Grey Literature for Natural Language Processing: A Terminological and Statistical Approach

Laura Cignoni, Gabriella Pardelli, and Manuela Sassi,
Istituto di Linguistica Computazionale, CNR, Italy

Abstract
This paper presents the results of a study on grey literature (GL) in the field of Natural Language Processing (NLP). Our data has been collected in a corpus of ca 13,000 records corresponding to the titles of papers presented at International Conferences from 1950 to June 2008. A statistical representation of the most significant terms relative to GL in NLP and other interrelated disciplines associates old and new words, highlighting the terminological changes that have taken place in the course of time. Aim of our study is to contribute to the creation of language resources for the extraction of GL coming from the Web in order to help prevent the disappearance of documents containing NLP words that have undergone rapid development over the last decades. This paper is organised as follows: after a general introduction to our work, section 2 provides a historical overview of NLP; sections 3 and 4 offer an account of the most relevant terms used by specialists in different periods, and indicative of the changes that have taken place; section 5 describes the methodology we have used and also contains information on our GL database and a graphical representation of the data. Finally, the conclusions stress the need to integrate pre-existing or obsolete words and expressions, creating NLP synonym relations.

Keywords computational linguistics, grey literature, information retrieval, multilingual terminology, natural language processing

1. Introduction

Since the advent in the 90's of the www technologies, computing has had a strong impact on modern society offering new opportunities of expansion for future research. In these years the Internet has evolved to such an extent that the important changes in the field of acquisition, storage and transmission of data have provided new resources to the web society, satisfying its needs for constantly updated information. Such information can appear in journals on-line, which were started as such, or of publications originally in paper format and then made available also in electronic form. New forms of publication have emerged which provide the scientific community with free and easy access to research studies, establishing a direct relation between producers and consumers who share knowledge on the web.

As far as grey literature is concerned, in the past important resources of information like pre-prints, technical reports, etc. produced by scholars working in specific fields of research were often included in informal craft-made editions which allowed a rapid exchange of data. These originated in the laboratories in the early 1950s for limited circulation among colleagues of scientific institutions; nowadays, the Internet and the activity of those operating in the field of GL at both the national and international levels have made the same type of information stored in repositories in the form of *E-print archives* visible also to those users belonging to different types of population - for example students and citizens - navigating in the Internet and interested in research works and projects. The people who, in rapidly increasingly numbers, access the internet to obtain information about natural language processing and other associated disciplines often encounter great problems as much of what they might be looking for can be impossible to find because of inappropriate searches. Different terms used to communicate the same idea can generate linguistic ambiguity, since the same word or phrase can allow for more than one interpretation, thus affecting the information retrieval process. It follows that the queries which are made through these linguistic variations do not always obtain the response looked for, and large amounts of information, although available, do not emerge from the web because the term is not present in the document requested. Access to the semantic contents of a document can become extremely difficult in the case of polysemy (when a word has two or more similar meanings) or of synonymy (when a word means the same as another word).

Technical and scientific terminology forms a large part of the lexicon of a language and is important for the exchange of information between specialists and users of the Internet, a new market that allows those producing and those looking for information content to find a meeting-point by means of a common lexicon. Dissemination of information through the web has enhanced the role of terminology and, in particular, of the use of terms representing the main anchors for information retrieval. Therefore, language plays a major role in the formation of concepts and describes the ways in which concepts are related, representing a primary means for the transmission of data in human society. In the past, in the field of NLP, expressions such as natural language processing, computational linguistics, mathematical linguistics, algebraic linguistics were often used indifferently; for example, *automated language processing* (ALP), which anticipated the expression NLP, was used over the entire 60's[1]. The two

expressions NLP and CL are often used indifferently, as a matter of fact many works that go under the heading of CL could also fit under the definition of NLP. Natural Language Processing (NLP) was defined as "a branch of computer science that studies computer systems for processing natural languages" (Cunningham: 1999); Computational Linguistics (CL) as "a branch of linguistics in which computational techniques and concepts are applied to the elucidation of linguistic and phonetic problems" (Crystal: 1991). The expression *computational linguistics* has more often been used in an academic context, but it is more or less a synonym of NLP and of current Language Engineering (LE). Nowadays, some Departments of Linguistics[2] still use these expressions with no diversification one from the other to advertise MS degrees (The University of South Carolina offers two programs in the area of Computational Linguistics - also called *Natural Language Processing* and *Human Language Technology*").

2. Background Information

In 1964, the United States Government, which for almost ten years had financed large projects in the field of automatic translation (AT), encharged an Automatic Language Processing Advisory Committee (ALPAC) to evaluate and prepare a report on the progresses so far achieved in the sector. In the report edited by the National Science Foundation the term Computational Linguistics was used for the first time in an official source to designate a new subject - derived from automatic translation - and destined to become a new disciplinary activity. Since then, AT research has been supported by systems of syntactic and in particular of semantic analysis with particular regard to the linguistic aspects of the problem and to the introduction of a number of programming languages for data processing, which have triggered many types of research in the field of linguistics or of other related disciplines. In particular, human language started to be processed in its various expressions and at all levels of analysis, for example for the study and synthesis of the voice as well as character recognition. In fifty years, the linguistic data belonging to written or oral language have been translated in a form comprehensible to the computer (machine-readable form) and a number of notions, methods, procedures have gradually become a common patrimony for all those working in the field of CL.

In the introduction to this paper, we mentioned the relation between NLP (or CL) and its related disciplines, especially linguistics and computer sciences. In the 50's, the computer started to be used for linguistic analysis, animating the spirit of the first scholars and experimenters and contributing to the development of automatic language processing. Tools up until then unavailable to classical lexicologists[3] were made available for linguistic investigations of various types (literary, stylistic, metrical, etc.). The most successful applications were those relative to lexicology and to quantitative linguistics, probably due to the sound structural methodology achieved in those domains and to the fact that CL is based on linguistics and methodologically pertains to this discipline for its methods of analysis and criteria of usage and control.

Specialized centres were founded where numerous projects were gradually set up, the variety and size of which increased from year to year. In the years between the late 40's and the early 60's, alongside various attempts of automatic translation, the expression *electronic calculus* for linguistic data processing also referred to electronic text processing as a support for research in the various humanistic disciplines. For example in 1949 Father Roberto Busa, s.j. undertook his monumental project, the *Index Thomisticus*, at CAAL (*Centro per l'Automazione dell'Analisi Letteraria*) in Gallarate, Italy, to help scholars index classical and other ancient texts. In the 60's, the computer started to be used in all fields of human activity thanks to the possibilities offered by natural language (Giacomo Ferrari, 2003: 122). The importance of a number of projects involved in the creation of lexicons (e.g. literary and philosophical *opera omnia*) by means of electronic processing systems is confirmed by works such as "The Death of the Hand made Concordance" (Raben, 1963). In linguistic statistics the computer was exploited to describe the quantitative characteristics of a text or of a group of texts, to study the style of an author, or to solve philological problems of chronology and attribution.

3. Current NLP Research Studies

The methods used for the production of indexes, concordances and frequencies made it possible to create extensive libraries for digitized texts. Computers have always provided documentary material, which can now be accessed on-line. Starting from the 60's, language resources (LR) became available which were made up of large amounts of data and linguistic descriptions including spoken and written corpora, lexical, grammatical and terminological databases. Machine dictionaries and disambiguators for textual research were provided, standards were defined for text encoding and lexical annotation, and significant systems were devised to extract linguistic knowledge from different types of corpora. We report the call of 2008 of the Annual Meeting of the Association for Computational Linguistics to show the current topics in the area.

"*ACL-08: HLT combines the Annual Meeting of the Association for Computational Linguistics (ACL) with the Human Language Technology Conference (HLT) of the North American Chapter of the ACL.* The conference covers a broad spectrum of disciplines working towards enabling intelligent systems to interact with humans using natural language, and towards enhancing human-human communication through services such as speech recognition, automatic translation, information retrieval, text summarization, and information extraction"[4].

4. Past Terminology

The device, either *mechanical*[5] or *electronic*, called *machine*, was used by the pioneers of mechanical translation (Bar-Hillel Y. *et alii*, 1952) between the late 50's and early 60's. The term derives from *mechané* which, as reported by Agusto Guzzo (1967: 4) in an attempt to recuperate the original meaning of the word, is first of all the mental idea that invents a device or any other means to avoid the difficulties of life (in Greek tragedy, when the plot of an adverse event could not be solved naturally, Euripides used a god lowered by stage machinery to extricate the protagonist from a difficult situation: *deus ex machina*, theòs apò mechanês). The term *machine* was then used in combination with groups of words that in one way or another reflected the number of complex operations in that particular sector, for example *machine translation* (MT), and indicated the application of computers to the translation of texts from one natural language into another, a computational translation system without human intervention. The adjectives *mechanical* and *automatic* were also used in combination with the adjective *computational*.

For example, in natural language processing, terms like *meccanizzazione* (It.), *mechanical translation* (Engl.), *machine à traduire* (Fr.), used in the 50's and 60's, seem to mark the start, transition and changes in this sector which has developed thanks to the use of computers in research and linguistic studies. The consolidation of this sector gradually led to the rise of Computational Linguistics which originally used adjectives such as *electronic*, *automatic*, *mechanical* and *cybernetic*, which were specific loan words from associated disciplines that had developed for different application environments, and which could be referred to the same epistemology. In this way the role of terminology for GL retrieval in a multilingual society became increasingly dependent on knowledge and information. At that time, the first communities of computational linguists started to feel a strong need for the correct nominalisation of concepts and this desire was satisfied by the use of terms or expressions pertinent to a specific domain, conveying the term to its relative semantic and cognitive value.

In France, GL records the adjective *computationnelle*, and the need for nominalisation introduces neologisms borrowed from other languages with the addition of a linguistic element placed at the end of the word, a type of suffix, in this case *–elle*, which introduces an English-like word. This adjective is still used[6] even if the French term used to designate computational linguistics is *informatique linguistique*[7]. The expression *informatique linguistique* or *linguistique informatique* is attested starting from the early 70's in France **(**Cori and Léon, 2002: 13). However, the adjective *computational* continues to be the reference term since CL studies the use of the electronic processor for the solution of problems.

In the past the introduction and use of the adjective *computazionale* in the Italian language and its combination with the term *linguistica* have justified a sort of ambiguity derived from its etymology (*computare*), which brings it close to quantitative and mathematical linguistics (Linguistica Matematica e Calcolatori, 1973).

The Italian adjective *computazionale* referred to linguistics derives from the English *computational* and is clearly of Latin origin. The term started to be used more and more frequently as it was able to indicate clearly and concisely a complex concept that would otherwise have needed a long periphrasis to be expressed. However, there was the risk of creating some kind of ambiguity with the verb *computare*, and therefore with statistic and quantitative linguistics.

The adjective *computational* was even erroneously used as a synonym of *mathematical linguistics*. In this respect we report the words of Marcel Cori and Jacqueline Léon: "Notons également que Computational Linguistics est parfois traduit par Linguistique mathématique. Ce terme désigne soit les études statistiques, comme c'est le cas dans les pays de l'Europe de l'Est, soit les travaux sur les grammaires formelles". The term has become the access key for web query, providing the linguistic representation of the concept; language is a process *in fieri*, and any changes at the lexical level are reflected in scientific literature. At this point it is worth questioning ourselves as to whether the information content of the dated documents can be extracted and to what extent. Simple queries on-line for grey literature in NLP have obtained sufficiently satisfactory results by means of simple network queries for groups of words: *automatic language processing,* no longer in use, *traduction automatique*. As far as *mechanical translation*[8] is concerned, great help has been given by the "Electronic repository and bibliography of articles, books and papers on topics in machine translation and computer-based translation tools" from which it is possible to obtain specific information about important past conferences.

5. Methodology and data description

Our corpus is composed of 13,270 records corresponding to the titles of papers presented at International Conferences in the field of computational linguistics from 1950 to June 2008. The main sources are the ACL Anthology and other Conferences like LREC for a total of 149 events. These include the Weaver Memorial, the Alpac Report and the first Conference on Automatic Translation held at MIT in 1952, as well as various documents either unpublished or published successively in the 50's.

Each record contains the title of the paper, name of author(s), name of conference, and year.

The textual corpus has been indexed by the DBT (Data Base Testuale) system, which performs a number of standard functions of textual analysis (concordances, co-occurrences, indexes, etc.) and the results are elaborated so as to produce a statistical and chronological study of the terminology used by specialists in this field.

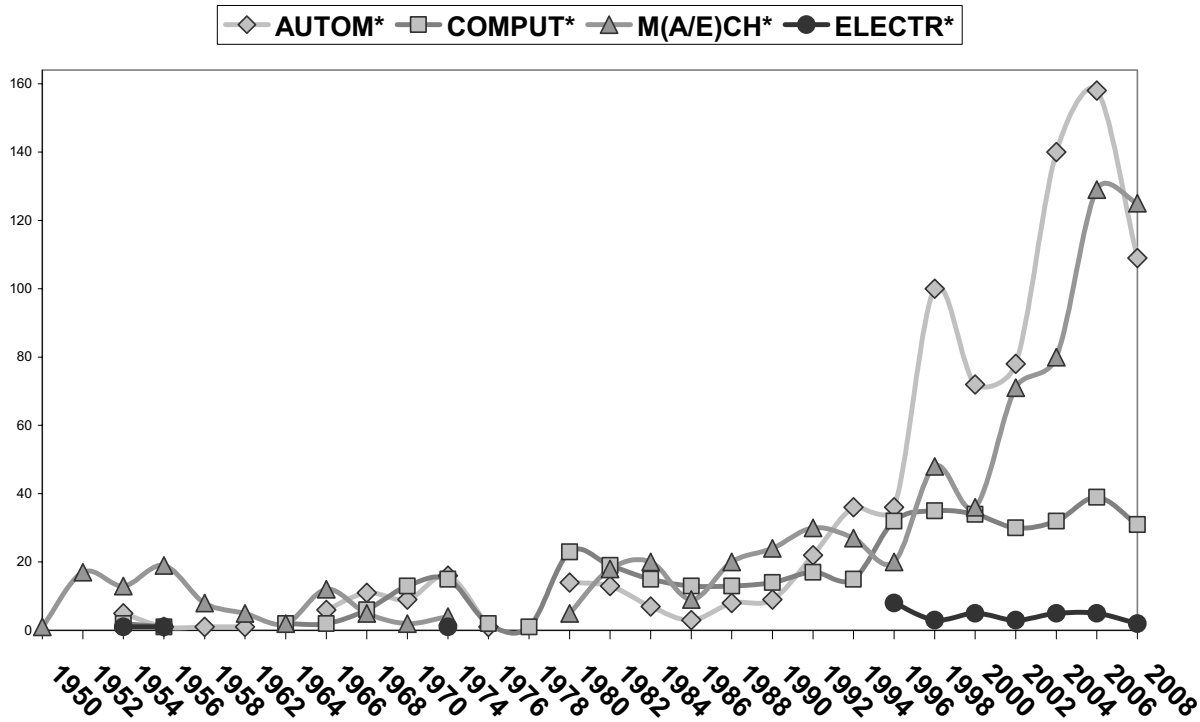
The methodology employed is the following:
- Search and saving of the most common single terms which are the object of this study;
- extraction of the contexts with year and abbreviation of the conference;
- generation of tables according to the chronological use of these terms;
- creation of charts.

The same procedure has also been applied to the abbreviations now consolidated and which represent the most important branches of research, although they may appear in the texts either as acronyms or in extended form.
Finally, the most frequent co-occurrences in the entire corpus have been calculated using the formula of DBT-integrated mutual information and the same procedure described above.

5.1 Single words

Extraction of the single contextualized words was performed by search of: AUTOM*, COMPUT*, M(A/E)CH*, ELECTRON* to compare their chronological use. The result of these cumulative queries are the following forms: *automated, automatic, automatically, automatically-extracted, automating, automation, automatique, automatisation, automatischen, automatisée, automatism, automatized, computability, computation, computationally, computational, computationally, computational-semantic, computations, compute, computed, computer, computer-aided, computer-assisted, computer-based, computerization, computerized, computer-mediated, computers, computes, computing, mechanical, mechanized, machina, machine, machine-aided, machine-guided, machine-induced, machine-learning, machine-mediated, machine-readable, machines, machine-tractable, machine-translation and electronic* (only form retrieved). Figure 1 shows the use of each group of listed words over the last sixty years.

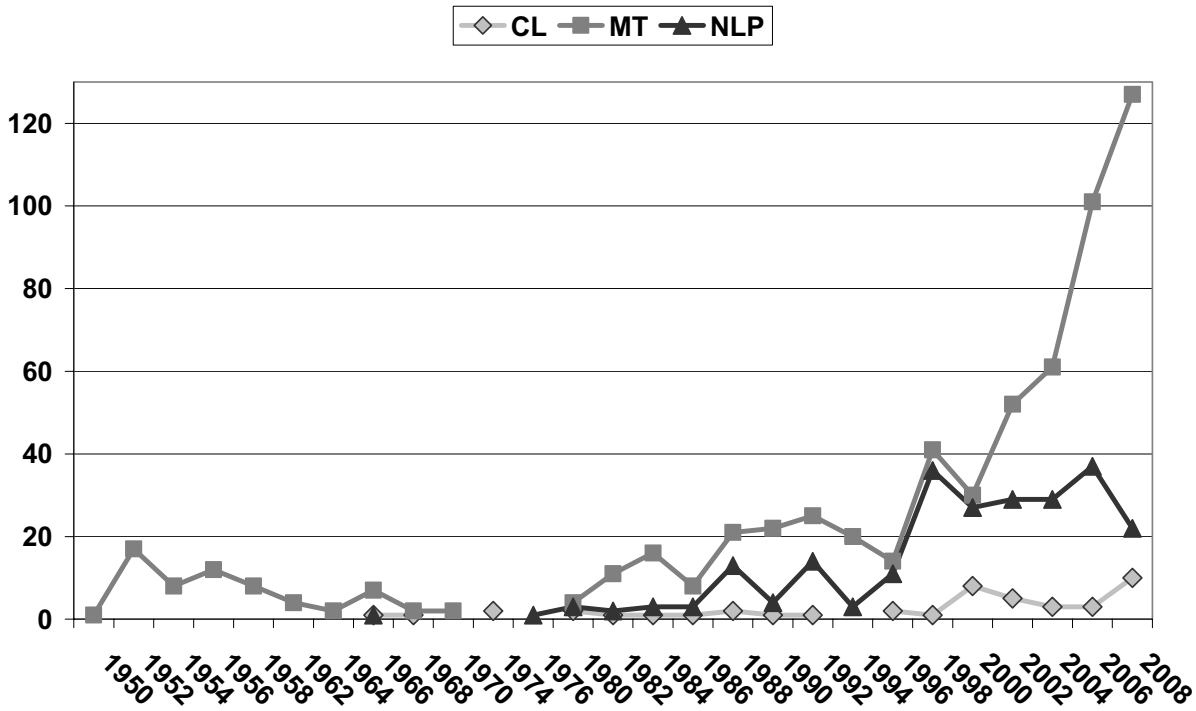
Fig. 1 - Use of listed words



5.2 Acronyms

To evaluate the use of the most important sectors of the discipline we searched *CL* and *Computational Linguistics*, or *NLP* and *Natural Language Processing*, or *MT* and *Machine/Mechanical/Mechanized Translation*.
The result of this query appears in Figure 2.

Fig. 2 - Use of acronyms



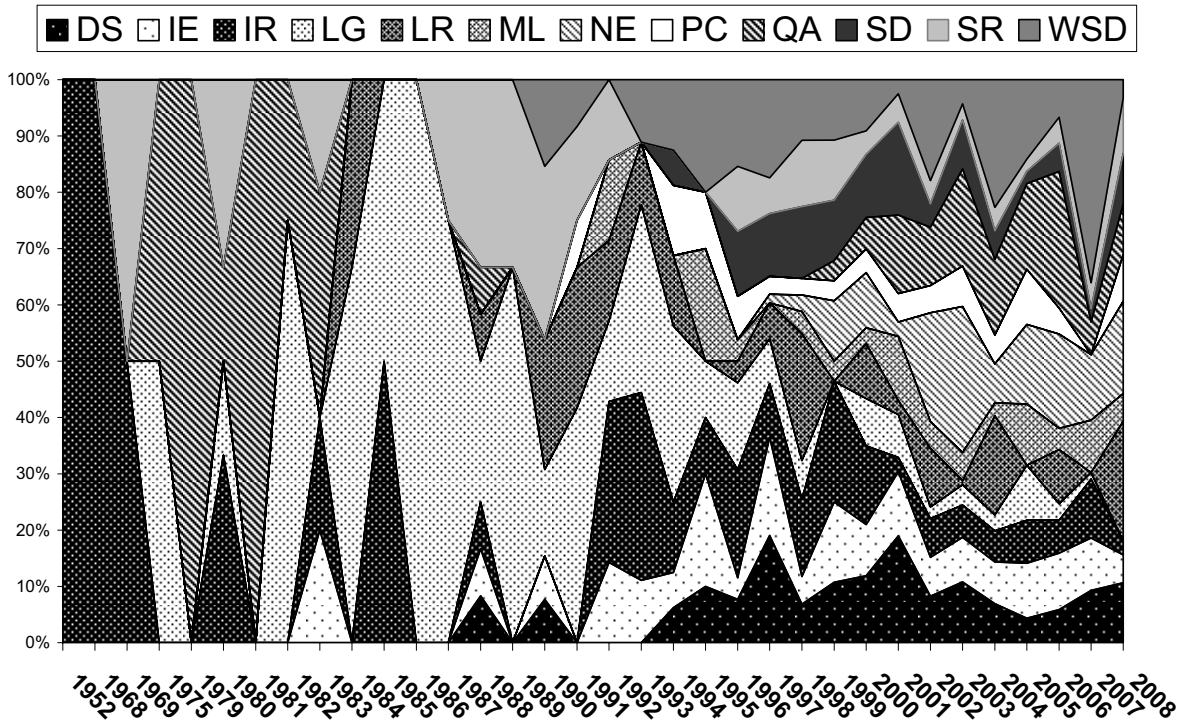
5.3 Co-occurrences

Using the mutual information formula we extracted the twelve most frequent co-occurrent words and organized their contexts in a table. The following abbreviations were used to optimize the graph legend:

DS *Dialogue System(s)*
IE *Information Extraction*
IR *Information Retrieval*
LG *Language Generation*
LR *Language Resource(s)*
ML *Machine Learning*
NE *Named Entity*
PC *Parallel Corpus(ora)*
QA *Question Answering*
SD *Spoken Dialogue(s)*
SR *Speech Recognition*
WSD *Word Sense Disambiguation*

Figure 3 shows the use of co-occurrences over the last sixty years.

Fig. 3 - Use of co-occurrences



6. Conclusions

This paper has highlighted the importance of the relationships of different terms expressing the same or similar concepts, often inherently complex and which require more than just a bit of background information. But how can similar words or expressions of the type *linguistic data processing*, *automated language processing*, *la machine dans la linguistique*, *la machine à traduire* and *literary data processing*, if unknown, possibly help retrieve documents that are distant in time? Since the queries are often incorrect, inappropriate, or simply far too general, it is necessary to integrate pre-existing or obsolete words and expressions used by specialists in the different domains to create a synonym relationship between the terms contained in the different NLP documents. In this way a term, still vital even if dated and no longer in use, becomes the key to enter the world of knowledge.

References

- Actes du Colloque International sur la Méchanisation des Recherches Lexicologiques. Besançon (1961). *Cahiers de Lexicologie*, Volume 3.
- Actes du Seminaire International sur le Dictionnaire Latin de Machine (1968). Rédigé par Roberto Busa S.J. *Calcolo*, Supplemento No. 2, Volume V.
- Université de Nancy, Faculté des Lettres et des Sciences Humaines (1966). *Actes du Premier Colloque International de Linguistique Appliquée*, Nancy, 26-31 Octobre 1964.
- Association for Computational Linguistics (1983). *First Conference of the European Chapter of the Association for Computational Linguistics*. Proceedings. Pisa, Italy.
- Bar-Hillel, Y., Perry, J.W., Reifler, E., *et al*., (1952) *Papers on Mechanical Translation*.
- Bessinger, J.B. Jr., Parrish, S.M., Arader, H.F. (eds.) (1965). *Literary Data Processing Conference*, New York, Sept. 9-11-1964. IBM, Data Processing Division.
- Bird, S., Dale, R., Dorr, B.J., Gibson, B., Joseph, M.T., Kan, M.-Y., Lee, D., Powley, B., Radev, D.R., Tan, Y.F. (2008). *A Reference Dataset for Bibliographic Research in Computational Linguistics In Proceedings of the Sixth International Language Resources and Evaluation* (LREC 2008), CD-ROM, European Language Resources Association (ELRA), Marrakech, Morocco, June, CD-ROM. http://acl-arc.comp.nus.edu.sg/ http://tangra.si.umich.edu/clair/anthology/index.cgi
- Busa, R.S.J. (1951). *Sancti Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum*: *A first example of word index automatically compiled and printed by IBM punched card machines*. Milano, Fratelli Bocca.
- Linguistica Matematica e Calcolatori: Atti del Convegno e della Prima Scuola Internazionale, Firenze, Leo Olschki, 1973.
- Calzolari, F., Sassolini, E., Sassi, M., Cucurullo, S., Picchi, E., Bertagna, F., Enea, A., Monachini, M., Soria, C., Calzolari, N. (2006). *Next Generation Language Resources using Grid*. LREC 2006: 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 24-25-26 May 2006. Proceedings, Paris, The European Language Resources Association (ELRA). CD-ROM, 1858-1861.

- Ceccato, S. (1961). *La Meccanizzazione delle Attività Umane Superiori*, parte I, in "Civiltà delle Macchine", IX, 4, Torino, 1961, 22-9.
- Ceccotti, M.L., Sassi, M., Pardelli, G. (2000). *Il soccorso informatico per lo studio di un autore difficile, C.E. Gadda*, in DIDAMATICA 2000, Informatica per la Didattica, Atti 1. Lavori Scientifici a cura di A. Andronico, G. Casadei, G. Sacerdoti: Società Editrice "Il Ponte Vecchio", 149-154.
- Ceccotti, M.L., Sassi, M., Pardelli, G. (2000). Un laboratorio multimediale dedicato a Carlo Emilio Gadda: Il modello e i primi dati implementati in formato XML, in AICA 2000, XXXVIII Congresso Annuale, Taormina 27-30 Settembre 2000: 267-272.
- Cignoni, L., Peters, C. (eds.). (1984). *Computers in Literary and Linguistic Research*, Proceedings of the VII International Symposium of the Association for Literary and Linguistic Computing (ALLC), Pisa, 1983, «Linguistica Computazionale», 3, Supplement.
- Cignoni, L., Coffey, S. (1995). *Looking for pre-selected multiword units in an untagged corpus of written Italian*: *maximizing the potential of the search program DBT*. Istituto di Linguistica Computazionale, CNR, Pisa.
- Cignoni, L., Coffey, S. (1998). *A Corpus-based study of Italian idiomatic phrases*: *from citation forms to 'real-life' occurrences*, in T. Fontenelle, P. Hiligsmann, A. Michiels, A. Moulinoulin & S. Theissen (eds.), Euralex '98 Proceedings, vol. I, University of Liège, English and Dutch Departments, 1998: 291-300.
- Cignoni, L., Coffey, S., Moon, R. (1999). *Idiom Variation in Italian and English*, *Two corpus-based studies*, «Languages in Contrast», 2 (2) 1999: 279-300.
- CITAL (1967). *2ème Conference internationale sur le traitement automatique des langues*. Grenoble.
- Cori, M., David, S., Léon, J. (2002). Pour un travail épistémologique sur le TAL. TAL Volume 43 n° 3/2002, 7-20.
- Crystal, D. (1991) *A Dictionary of Linguistics and Phonetics.* 3rd ed. Oxford. Blackwell.
- Cunningham, H. (1999), A definition and short history of Language Engineering, *Natural Language Engineering,* 5 (1), 1-16.
- Delavenay, E. (1959). *La machine à traduire*. Paris, Presses Universitaire de France.
- Garvin, P.L. (1972). *On Linguistic Method*. The Hague, Mouton.
- Garvin, P.L., Spolsky, B. (1966). *Computation in Linguistics: A Case Book.* Bloomington, Indiana University Press.
- Guzzo, A. (1967). L'uomo, la macchina, la tecnica. Relazione introduttiva, in *L'uomo e la macchina*. Atti del XXI Congresso Nazionale di Filosofia, I Relazioni. Edizioni di Filosofia, Torino.
- Hays, D.G. (ed.) (1966). *Readings in Automatic Language Processing*. New York, American Elsevier P.C.
- Hays, D.G. (1967). *Introduction to Computational Linguistics*. New York, American Elsevier P.C.
- Juilland, A., Roceric, A. (1972). Analytic Bibliography, in *The Linguistic Concept of Word.* Paris, Mouton, 11-59.
- Lanza, C., Pardelli, G. (1998). *Sviluppo delle raccolte e procedure di gestione nelle biblioteche dell'ICAS e dell'ILC*, in La Biblioteca un servizio di rete, Follonica, 1 ottobre 1998.
- Lanza, C., Pardelli, G. (2000). *Una soggettazione automatica di letteratura grigia con algoritmi di rete neurale artificiale. Due esperimenti ICAS e ILC* . In: 3° Convegno Nazionale, "La letteratura grigia: politica e pratica". Istituto Superiore di Sanità, Roma, 25-26 novembre 1999. Atti a cura di V. Alberani e P. De Castro, Roma, ISTISAN Congressi 67, 52-56.
- Luzi, D.(1995). *Internet as a new distribution channel of scientific grey literature. The case of Italian WWW servers*. In GL2, Second International Conference on Grey Literature, Washington, December, 1995.
- Minsky, M.L. (1967). *Computation*: *Finite and Infinite Machine.* Englewood, Prentice-Hall, Inc. vii-7.
- National Academy of Sciences, National Research Council (1966). *Language and Machines: Computers in Translation and Linguistics*. Washington, D.C.
- Pardelli, G., Sassi, M. (1998). *I.L.C. Library*: *Cataloghi e indici*. Pisa, S.T.A.R.
- Pardelli, G. (2000). *Subjects for the Automatic Treatment of Language within the Framework of a Virtual Humanities Library*. TR20\CNR-ILC\2000.
- Pardelli, G., Sassi, M. (2001). ILC Library. Dati bibliografici. XML.
- Pardelli, G., Cignoni, L. (2001). *Entrate Lessicali per il trattamento automatico del linguaggio* (TAL). Memorie interne. CNR/ILC/BIB-01-2001 del 31.12.2001
- Pardelli, G., Orsolini, P., Sassi, M., Enea, A., Gazzetti, S. (2002) (a cura di) TAL Bibliography (1951-2002). Parte I. Pisa, S.T.A.R, 2002.
- Pardelli, G. (2003). *BIBLOS*: *Historical, Philosophical and Philological Digital Library of the Italian National Research Council*, in A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. Linguistica Computazionale, Special Issue, XVIII-XIX, Pisa-Roma, IEPI. Tomo II, pp. 519-546.
- Pardelli, G., Sassi, M., Goggi, S. (2004). *From Weaver to the ALPAC Report*. LREC 2004: Fourth International Conference on Language Resources and Evaluation, held in Memory of Antonio Zampolli. Lisbon, Portugal, 26th, 27th & 28 May 2004. Proceedings, Volume VI, Paris, The European Language Resources Association (ELRA). 2005-2008.
- Pardelli, G., Orsolini, P. (2005) (a cura di). Special Session "In memory of Antonio Zampolli". L&T '05 - 2nd Language Technologies as a Challenge for Computer Science and Linguistics, April 21-23, 2005, Poznan, Poland.
- Pardelli, G., Orsolini, P. (2005). *Bibliography of Antonio Zampolli* (*from 1962 to 2004*), in Antonio Zampolli and Computational Linguistics, Archives of Control Sciences. Special issue on Human Language Technologies as a Challenge for Computer Science and Linguistics, Vol. 15 (LI), 4, 501-517.
- Pardelli, G., Sassi, M., Goggi, S., Orsolini, P. (2006). *Natural Language Processing*: *A Terminological and Statistical Approach.* LREC 2006: 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 24-25-26 May 2006. Proceedings, Paris, The European Language Resources Association (ELRA). CD-ROM, 2395-2398.
- Pardelli, G., Sassi, M., Goggi, S. (2007). *A survey on Human Language Technology Terminology*, in Z. Vetulani (ed.), Proceedings of 3rd Language & Technology Conference. Fundacja Uniwersytetu im A. Mickiewicza, Poznań, 364-368.
- Pritchard, D. (2008). Working Papers, Open Access, and Cyber-infrastructure in Classical Studies, *Literary and Linguistic Computing* 23 (2), 149-162.
- Proceedings of the Fourth International Congress of Applied Linguistics (1976). edited by Gerhard Nickel. Stuttgart, HochschulVerlag. Vol. I.

- Quemada, B. (1957). La technique des inventaires mécanographiques, in *Lexicologie et Lexicographie Françaises et Romanes.* Paris, CNRS, 53-63.
- Sassi, M. (2003). La consultazione dei corpora costituzionali con DBT. In *Diritto alla vita e Diritto all'ambiente nel lessico costituzionale e nella dottrina giuridica*. Strumenti e metodi per l'analisi linguistico-concettuale. Note Tecniche. ITTIG, Firenze. 149-154.
- Sassi, M., Ceccotti, M. L. 2001. L'utilizzo didattico da corpora: proposte metodologiche. In: *Didamatica 2001, Informatica per la Didattica.* Atti 1. Lavori Scientifici, a cura di A. Andronico, A.M. Fanelli, G. Piscitelli, T. Roselli: Edizioni Giuseppe Laterza. 350-358.[9]
- Sassi, M., Cinini, A. (2007). *L'informazione sanitaria. Analisi di tre quotidiani a tiratura nazionale*. ILC-CNR-Rapporto Tecnico, Pisa, 2007.
- Stindlová, J., Mater, E. (1968). (Rédaction), *Les machines dans la linguistique.* Academia, Editions de l'Académie Tchécoslovaque des Sciences. Prague.
- Stock, C., Schopfel, J. (2003). *Grey Literature in an open context : from certainty to new challenges*. In GL5, Fifth International Conference on Grey Literature, Amsterdam.
- University of Michigan (1960). Foreign language project "Mechanization": a project for research, controlled experimentation, design and testing in the mechanization of the language learning processes.
- Weaver, W. (1949). *Translation*. Repr. in W.N. Locke, A.D. Booth (eds.) *Machine translation of languages: fourteen essays,* Cambridge, Mass., Technology Press of the MIT (1955). 15-23.
- Woldering, B. (2004). The European Library: Integrated access to the national libraries of Europe, «Ariadne». http://www.ariadne.ac.uk/issue38/woldering/).
- Zampolli, A. (1969). *Due conversazioni sul panorama attuale della linguistica computazionale*. Università degli Studi di Pisa, CNR-CNUCE.
- Zampolli, A. (1973). Humanities Computing in Italy. *Computers and the Humanities*, Volume VII (6), 343-360.
- Zampolli, A., Calzolari, N. (1977) (eds*.). Computational and Mathematical Linguistics.* Leo S. Olschki Editore, Firenze. Volume 36.
- Zampolli, A., Cignoni, L. (1985). *Attività Progettuale per il Trattamento del Linguaggio Naturale*, ILC-9-3.

Endnotes

[1] We report the words of Antonio Zampolli, who helps us define the two expressions "I use this term [automatic language processing, ALP] rather than computational linguistics as it is far more general in its implications, encompassing all studies, theoretical or applied, on the use of computers or computational techniques in the processing of natural language" (Zampolli, Calzolari, 1977: XIX).

2 http://www.isi.edu/natural-language/MSCompLing/

[3] We report the words of Bernard Quemada at the International Colloquium on the mechanisation of lexicological research, held at Besançon in June 1961. "A la lumière des échanges de vues s'est trouvé confirmée la tendance qui divise, en apparence plus qu'en réalité, les utilisateur des machines. Ainsi semblent s'opposer lexicologues 'classiques' qui désirent béneficiér des moyens mécaniques pour accroître leurs possibilités de travail en suivant des normes d'exploitation ayant fait leur preuves, et lexicologues 'modernes' qui, de ce fait songent à des applications très différentes des précedentes" (Cahiers de Lexicologie (3), 1962: 3).

4 http://www.ling.ohio-state.edu/acl08/cfp.html

5 (Lat. *mechanicu*(*m*), Gr. *mekhanikós*, derived from *mechané* "machine").

6 Yahoo Italia: 7.30 pm, 15 July 2008 produced ca 851 results for the term *linguistique computationnelle*.

7 Yahoo Italia: 7.30 pm, 31 August 2008 produced ca 3,120,000 for the term *informatique linguistique*.

8 http://www.mt-archive.info/MT-1954-Reynolds.pdf

Grey Literature produced and made available by Universities – Helping future Scholars or Plagiarists?

Primož Južnic
University of Ljubljana, Slovenia

Abstract
Universities and other institutions of higher education are far the greatest producers of grey literature (GL). Most of their education process is based on various written essays or other sorts of similar tasks. Even more important, the whole process is usually finished by some sort of written theses/dissertation (graduation work, diploma) that shows a graduate is capable of research work and has a proper knowledge of the field.
The traditional paradigm was to make this material available through academic libraries. The Internet has helped to simplify this process and relieve academic librarians from trivial and routine tasks. It has also made it easier for all potential users, often students themselves, to access these materials, adding to other materials they can use that are part of GL materials. This sounds like a great leap forward if current research did not indicate that academic plagiarism is now a very serious problem worldwide.
The research presented in this paper presents how librarians are getting involved both in making materials available and at the same time in fighting plagiarism and how their expertise in dealing with different information sources, including those called "grey literature," can be used to help teaching staff in their struggle to maintain the quality of academic education. This is also one of the factors turning traditional library tasks and services toward the more professional expertise expected by information technology experts. The survey of academic libraries in Slovenia presents the availability of theses and dissertations and other services offered by academic libraries and librarians and their future plans.

Introduction

Universities and other institutions of higher education are important producers of grey literature (GL). Higher education programs are usually concluded by some sort of written paper, a thesis or dissertation, which shows that a graduate is capable of research work and has a proper knowledge of the field. A student does this task following the successful completion of all study program requirements, and the written paper is subject to approval. Although called by various names in different countries and languages and at different universities (thesis, dissertation, graduate or diploma work, final papers, etc.), the form of the paper is more or less standard. The work is written and has a title, abstract, table of contents, main part, and bibliography. The main part or body consists of

- introduction,
- review of the literature,
- methodology,
- results/findings,
- analysis and interpretation of findings, and
- summary, conclusions, applications, and recommendations for further study

These papers are not the only grey literature originating at universities—research contributes its share—but form by far the greatest part of it. Their vast numbers place universities among the greatest sources of grey literature.
The special value of theses lies in the fact that they are reviewed, checked, and evaluated. Theses are submitted to a committee consisting of a mentor/supervisor and two or more committee members who usually act as the reviewers. Normally the committee members are professors and experts in their field and have the task of reading theses, making suggestions for changes and improvements, and giving the final approval. Only after the final approval do theses become official, and the fact that a student's work goes through a review process by the university is the guarantee of its quality.
A thesis is a written text representing the independent research and authorship of a single individual. Its purpose in higher education remains the same today as it has been for centuries, across countries and disciplines. It would be beyond the scope of this paper but still worth mentioning that this remains the principle despite various critiques of both the romantic notion of authorship and the epistemological assumptions that form traditional notions of independent scientific and scholarly research: research today involves teamwork, multi-authorship is the rule in most scientific disciplines, but the thesis remains the last bastion of single authorship.
What happens afterwards? A student successfully completing a degree on the basis of a thesis receives his diploma, approval that he is ready to join the social division of work or the labour market in a certain role. The proof of this readiness, the thesis, remains at the academic institution. Traditionally, theses were regarded as library material because they were available through academic libraries. Libraries made

them part of their collections, catalogued them, shelved them, and made them available to users. They were typical GL material as it was not easy to find and access it. Libraries had also an archival role since often only one copy of a thesis existed.

Electronic Thesis and Dissertation (ETD)

The Internet has helped to solve many library and librarians' problems and relieve (academic) librarians from trivial and routine tasks. This applies to theses and dissertations as well. The solution offered is Electronic Thesis and Dissertation (ETD). The term ETD refers to a thesis or dissertation that is archived and circulated electronically rather than archived and circulated in print. Most ETDs take the form of text uploaded in a word processing format or in Adobe's portable document format (PDF) and look very much like traditional printed theses. They reside on the Internet where they are accessible to potential users. Extensive overviewof ETD can be read in a Sourcebook for educators, students, and librarians, titled Electronic theses and dissertations (Fox, 2004)

A major boost to ETD was the Networked Digital Library of Theses and Dissertations initiative. The Networked Digital Library of Theses and Dissertations (NDLTD) is a collaborative effort of universities around the world to promote the creation, archiving, distribution, and access of ETDs. Since its inception in 1996, over one hundred universities have joined the initiative, underscoring the importance institutions place on training their graduates in the emerging forms of digital publishing and information access (Suleman 2001). The NDLTD is an international organization dedicated to promoting the adoption, creation, use, dissemination, and preservation of electronic analogues to traditional paper-based theses and dissertations. Its website contains information about the initiative, how to set up ETD programs, how to create and locate ETDs, and current research in digital libraries related to the NDLTD and ETDs.

An overview (Edminster 2002) of these international efforts to develop a worldwide digital library of theses and dissertations focused on

a. the need to provide developing countries with equal access to current international scholarship;
b. the collaborative development of training materials to facilitate wider global participation in the NDLTD;
c. the work of multi-university/library and corporate collaborations to establish centralized metadata for ETDs; and
d. the development of multi-language search interfaces.

However, the objectives of the NDLTD were originally seen more broadly, including

- to improve graduate education by allowing students to produce electronic documents, use digital libraries, and understand issues in publishing;
- to increase the availability of student research for scholars and to preserve it electronically;
- to lower the cost of submitting and handling theses and dissertations;
- to empower students to convey a richer message through the use of multimedia and hypermedia technologies;
- to empower universities to unlock their information resources; and
- to advance digital library technology (Suleman 2001)

To gain an overview of activities relating to ETDs internationally, the web sites of every member of the NDLTD were examined. A study of approximately two hundred sites revealed that only a small percentage of the NDLTD institutions dealt with a large quantity of ETDs in 2002 (Copeland, 2003) The findings from the survey indicated that many universities could make better use of the guidance notes relating to all aspects of ETD production, management, and use, so it should be seen as an initiative impacting on various national ETD systems.

Why national systems? Usually theses and dissertations are seen as an important information resource because as a rule they are the result of research. Part of their content finds its way into other publications (journal articles, congress papers, and books), but not all of it. This is an important element in national use, although we tend to forget that theses and dissertations serve to disseminate research information within local communities, especially within smaller countries and language environments. A survey by Stock (2007) of theses written in English showed important differences between European repositories. In the Scandinavian countries as well as in Belgium and The Netherlands, between 50% and 90% of (doctoral) theses are in English. In German universities the percentage of English theses has grown to reach 25%. This indicates the willingness in some countries to give the widest access possible to one's work through the choice of language and through the internet.

This is perhaps positive globalization, but it also has a negative effect. While English has become the international language of research, this does not mean that all other languages have become non-scientific. If theses and dissertations are not available in national languages, this will become an issue and a problem.

There are various national initiatives and surveys presenting the current state of theses and dissertation collections, their usage, problems with access, and the academic and research community's attitude toward ETD. But do national ETD systems work? They seem to suffer from the same problems that plague the international NDLTD system, at least judging by national reports.

In India an integrated system at the national level to locate and access theses has not been fully implemented. While just a few Indian universities have actually started ETD projects at the moment, the majority have the intention of starting such projects soon (Vijayakumar 2007). In recent years, South Korean university libraries have tried to improve user services and access to ETDs in several ways. However, authors blame the absence of an adequate policy and infrastructure to handle them at the national level for the fact that little practical progress has been made at individual academic libraries (Park 2007).

As reported for France, an integrated national ETD system still does not exist, the results of the government initiative seem disappointing, and the development and implementation of national software and services is progressing more slowly than planned. At the same time, a growing number of alternative, more or less successful local initiatives, academic networks, and open archives provide access to more than four thousand ETDs. The reasons for this paradoxical situation are various. So far, neither the government nor any other institution has had enough coercive or persuasive force to impose a unique model for ETDs. Perhaps this "unique model" is simply unrealistic and not adapted to the heterogeneous needs, behaviours, and traditions of France's scientific and academic communities (Paillasard 2004).

It is clear that in recent years an increasing number of universities are building their own ETD systems or are at least considering doing so. Why are they important for every university? More and more ETD initiatives are connected with the electronic submission of theses and dissertations and other issues that help solve specific university problems, improve quality, and save time and money.

Generally speaking, five objectives for university or other higher education institution ETD systems can be named:
1. to make research reported in theses and dissertations more widely and easily available;
2. to initiate and encourage digital development;
3. to ease submission process;
4. to save space in libraries; and
5. to benefit the higher education process.

The first objective is very general and needs little explanation. An institutional repository includes a variety of materials produced by the university, not only thesis and dissertations but also research reports, congress papers, and especially teaching materials. Some university institutional repositories are also being used as resources for electronic publications and e-journals. This makes university institutional repositories different from other types of digital repositories.

Although they fail to substantiate their claims with data, many authors argue that electronic writing tools are transforming graduate education, enhancing mentoring and the shape of thesis content. A recent analysis of bibliographies from student research papers revealed what sources students used to support their research. While web sites were a definite fixture in student bibliographies, on average they were not the predominant source of information that one might expect given the current perception of student research. In the study, 55% of the bibliographies did not cite any web sites at all. This is an important finding to note, as it runs counter to the concerns of faculty (Carlson 2006). It might vary across the disciplines, but it is generally valid for the majority. One of the reasons for this might be that when students face submitting their work in the traditional printed format, they tend to work or think traditionally about the information sources they use. Another reason might be the instability of Internet resources. A study of undergraduate students' citations of web sites had astonishing results: only 18% of the URLs cited in 1996 and only 55% of the URLS cited in 1999 led to the correct documents in 2000 (Davis 2001).

Generally speaking, the paper-based thesis submission process consists of three steps: production, submission, and preservation. Availability and use are primarily shaped by the paper version. Many universities are experimenting with electronic submission, which completely surpasses traditional paper forms. Bevan (2005) describes the issues involved in the introduction of mandatory submission of electronic theses at Cranfield University in the UK. McGill University in Montreal, Canada, has undertaken a pilot project to test aspects of workflow, style sheets, metadata, and search functions (Park 2007a). In the pilot project, a new model for tracking the electronic file through the production, conversion, dissemination, and preservation processes was developed. The students first submit their theses in whichever of four authoring tools they prefer. After the completion of the examination process and thesis

revision, the students submit two paper copies of the thesis to the Thesis Office and upload the electronic version. The supervisor reads and approves either the paper form or the electronically submitted final copy. The Thesis Office performs a content check on both versions, a paper copy of the thesis is sent to the library, and the library is notified that the content check has been completed.

The advantage of single-institution ETD systems is clear and obvious. A study of ETD system implementation at individual higher education institutions discovered that library administrators who implemented ETD repositories at different universities adapted their models to the needs of their institutions and their graduate students. ETD system administrators made decisions about implementation models and software and hardware infrastructure in terms of human and technical resource allocation (Yioris 2007). These decisions are difficult to achieve at the international or even the national level, and this gives the advantage to local systems.

The next step is seen as the electronic submission of ETDs automatically building the repositories. The permanent and secure preservation of documents is often an issue; the tension between libraries' two-fold responsibility of preserving and providing access to information takes on particular significance with ETDs. As the examples have shown, many universities balk at the idea of allowing students to submit work exclusively in electronic form, and they continue to require what is perceived to be a more "permanent" print copy for archival purposes. As complementary to print, some universities will accept an archival version on CD-ROM, but there are concerns as to the long-term durability of this technology (Edminster 2002).

The preservation and availability of ETDs at all levels is not the only concern universities and other higher institutions have regarding them. There is also a concern regarding plagiarism and other forms of cheating. Plagiarism is the nightmare of higher education, often a theme not to be discussed in public. It is even hard to uncover the extent of it. Over a three-year period, McCabe (2006) surveyed more than 80,000 students and 12,000 faculty in the United States and Canada and confirmed that plagiarism is a significant issue. For example, if the four behaviours in which students engage least frequently—turning in work copied from another, copying large sections of text from written sources, turning in work done by another, and downloading or otherwise obtaining a paper from a term paper mill or website—are combined, it is clear that 16% of all undergraduate respondents and 8% of responding graduate students reported one or more of these behaviours in the past year. In contrast, a surprisingly large number of faculty (79%) report they have observed one or more instances of these behaviours in the last three years, driven in part by a perception that a large number of students (59%) have copied material almost word for word from a written source without citation. Due to their "grey literature" nature, ETDs are often seen as the main source of students' "cut and paste" work.

Survey

Higher education in Slovenia is regulated by the Higher Education Act (1993, but amended almost every two years). Higher education institutions in Slovenia comprise four universities with 53 faculties and art academies and twelve separate higher education institutions established as private institutions as of March 2008 (13 more are in preparation).[1] Higher education institutions are autonomous in managing their internal organization and operations (according to their statutes and the legal requirements). The implementation of a three-cycle higher education system according to the Bologna Declaration has been slow and reluctant, but it is progressing. There are three university libraries and almost every university faculty has its own library and often individual departments have their own libraries as well. The most important feature introduced by the new legislation was the new role of the university with the change from being an association of independent faculties into an integrated university, but this new legislation did not touch the library systems and organization.

All higher education institutions were surveyed. The creation of separate private higher education institutions has been very rapid in recent years. However, we discovered that most of the new higher education institutions have no libraries or similar services. Although the legislation demands that every higher education institution have organized library services, many new institutions wished to cut costs and proclaimed public libraries as their library services provider, which raised (among other issues) doubts about the quality of their programs.

Sixty-three higher education institutions, faculties, and departments were chosen for a further survey. All libraries use the National Union Catalogue and Co-operative Online Bibliographic System and Services (COBISS) system to input their theses and dissertations. The development and operation of the COBISS system has been the core of the library information system in Slovenia for the last three decades. Almost all libraries and information services within public institutions are part of the COBISS system, and their materials part of the database. Slovenia and its library system is also very much shaped by its national online bibliographical system for collecting and making all the information about library collections available to all interested users.

Special tags distinguish ETDs from other parts of library collections. The COBISS Union Catalogue used the following numbers in the 1998-2008 period as of November 2008:

104568	Graduation theses (Pre-Bologna system)
1458	Specialist theses
9680	Master's degree theses
6625	Ph.D. dissertations*

(*3,442 in Slovenian language)

Only ten of the higher education institutions have some form of their own ETD system, and three more intend to organize one in the near future. The great majority of these libraries allow their patrons and other users to access and use this part of their collections (theses and dissertations) only within the library premises. It can be understood from the information on their web sites that some libraries require a special author's permission before allowing access to the material.

In addition to the survey, we decided to interview twelve librarians to acquire more detailed information about this topic. They were all from the University of Ljubljana, Slovenia's largest and oldest university. Librarians from various disciplines were interviewed: one each from the natural sciences, biotechnical sciences, biomedicine, and arts and humanities faculties, and four each from the technical and social sciences faculties. We selected faculties and departments with working and planned ETD systems as well as some without such plans at the moment.

The interviewed librarians, either heads of libraries or in the larger academic libraries the person responsible for this part of the collections, rate theses and dissertations as having importance in furthering research in the respective disciplines and believed that mentors often advise their students to consult them. All agreed that theses and dissertations are an important part of their collections and are treated with special care. Seven considered theses and dissertations important for all of their users; others thought they are more important for students and less important for teaching staff and other users. They said that teaching staff rely more on journals and other information sources and refer to theses and dissertations only sporadically. One librarian mentioned that theses and dissertations are important for teaching staff as an aid for their work with students.

The next question was how users can access theses and dissertations. Four librarians were from libraries that have ETD systems (one since 2003, the second since 2006, and the other two since 2007). Interestingly enough, one institution has stated on its web site that an ETD system is in preparation, but its librarian actually did not know anything about it. Other librarians were not thinking about ETD systems, and one mentioned his doubts about a possible ETD system, since he thought that some theses and dissertations are prepared for and in cooperation with industry, which is an obstacle to making them available via the Internet. One library publishes an updated list of theses and dissertations on its web pages, and It is interesting that another library started with such a list and by adding PDF files to it created a simple but effective ETD system.

Access policies concerning theses and dissertations vary across libraries, even though they are all libraries of the University of Ljubljana. Only three libraries allow the open borrowing of theses and dissertations (for two weeks or in one case for a month). In one library only teaching staff can borrow this material, while others can use it only within the library premises. Another library allows borrowing theses done by postgraduate students but not other material. Others allow the use of theses and dissertations only in library premises, and two require written permission from the authors before any action. In only three libraries can users have direct access to theses and dissertations in the library premises; mostly due to the lack of space, all the others keep them in the part of the library where visitors have no access.

All librarians referred to these materials as an important part of their collection that needs special handling. Surprisingly, only two librarians are directly involved in helping students with their thesis and dissertations. Three of them reported that students have to submit their theses to the library for citation and other checking.

Plagiarism was an issue for most of the interviewed librarians, although two librarians denied it was a problem. The first claimed she had never heard of any cases, and the second stated that theses and dissertations are results of research work, and while the results can not be copied or faked, of course some paragraphs of the theoretical section can always be copied. A similar opinion was given by the librarian from the Art Academy, who pointed out the originality of the artistic part of theses and dissertations. Others knew about plagiarism (for some librarians, this is one of the reasons not to lend

theses and dissertations for home use) and often had anecdotal examples. Generally speaking, the librarians thought that plagiarism is the primary concern of mentors and teaching staff and not theirs.

Future developments and challenges

The wide availability of the National Union Catalogue and COBISS is likely to encourage a shift to full-text databases of electronic theses and dissertations (ETDs). However, this can be also an obstacle since many libraries and librarians might see it as a reason not to have their own ETD systems because "everything" is already available. A further obstacle might be the extreme decentralization of academic libraries in older universities and the absence of any form of library services in the new universities and higher education institutions. Decentralization in its present form could mean a zealous opposition to any form of ETD system, and the absence of library services means that an ETD system can not be built at all.

While some might be in favour of transferring ETDs to a central library and collaborating in ETD management, others will not. Those not in favour will tend to prefer a hybrid method of sharing ETDs whereby each library preserves its originals and provides full-text access to the theses and dissertations at its own repository. At the same time, each library would send a copy to the central library, which makes the full text theses and dissertations available to other libraries. The centralized input of records for other library material presenting research results (SICRIS) has been replaced by direct input from each library in Slovenia, and this should remain the principle for ETDs if the system is to succeed. Of course, the national authorities must provide the proper regulation as was done in the case of SICRIS.

The electronic submission of ETDs must be the next step. Many academic libraries might think various issues are an obstacle to creating ETD systems, including the risk of plagiarism and the lack of funding, administrative support, and regulation. However, those that have already started creating their own ETD systems should prove them wrong and demonstrate the possibility that the infrastructure support, technical expertise, and financial support to create ETD systems already exists in their own institutions. Effective awareness programs are required to increase their visibility and emphasize their usefulness. The complete electronic submission of theses and dissertations can be the decisive point toward implementing ETD systems and is therefore worth special effort and investment.

Librarians need to get more involved in helping students write theses and dissertations and create their electronic counterparts. This could be a way to improve their status, which is very low in universities in Slovenia as demonstrated by recent events regarding their salaries in the new national scheme. Active participation in the creation of theses and dissertations, the ultimate demonstration of higher education, could certainly have positive status repercussions.

It is easier for students to plagiarize from ETDs because of the increased access to electronic documents and simple copy and paste functions. The features of search functions, however, make detecting plagiarism easier as well. Every university has policies in place regarding plagiarism, and these must be enforced along with the proper application of fair-use guidelines (Yioris 2007). Why the second? There are many good technical methods of detecting plagiarism, but students can not be left alone in the fight to prevent it. Librarians can help considerably by educating students on how their work will be assessed and the potential traps of possible plagiarism. The difference between copyright violation and the threat of plagiarism is often confused in discussions about intellectual property. Plagiarism occurs when someone poses as the author of a work; copyright infringement occurs when someone uses another's work without proper authorization or citation. Students rarely understand the difference, and librarians have the expertise and authority to help them make the distinction.

In theory, librarians are seen as experts who understand user needs and perceptions. They know what works and what does not. They know how to help, inform, persuade, and teach users (Bailey 2005). They could serve as more than just "plagiarist busters," but this does require that librarians improve their own knowledge of the issues regarding academic integrity. They should be able to promote a more complex understanding of the Internet and a critical approach to research and writing. The problem is not that students today are more dishonest but that their experience—particularly with the Internet-based transfer of information—has led them to form different attitudes toward information, authorship, and plagiarism (Wood 2004).

We also need other activities to promote the concept of ETD systems. According to current data, workshops and web documents are most often used to educate students about ETDs, although faculty and administrators learn about them mainly through presentations, lectures, and seminars. The methods might be different in different environments, but the fact is that approaches must be different for different users. Even if ETD systems benefit students, professors, and the public alike by enhancing graduate education, expanding graduate research, and increasing a university's output quality, the activities must be tailored for the different audiences. Universities need to recognize the potential value

of accessible ETDs since theses and dissertations reflect an institution's ability to lead students and support original work. An interesting observation is that when ETDs are in an accessible place, students and teaching staff will make judgments regarding the quality of a university by reviewing its digital library. Universities must respond accordingly, ensuring they provide the resources and training students need to incorporate new literacy tools such as animation, graphics, sound, and streaming multimedia (Edminster 2002).

This may be seen today as a distant future, The uncertainty created by the relatively recent introduction of ETD systems and the absence of national policies and frameworks in this area hinder their rapid adoption. What we might need is an ETD submission protocol, implemented and tested for different institutions. As a result of the different ETD projects, recommendations can be made and different approaches chosen. It will be exciting to see something regarded as a grey literature in the past and treated accordingly become the core of higher education activities and a centerpiece of a university's reputation.

At the University of Ljubljana, a new portal, the Digital Library of the University of Ljubljana (DIKUL[2] – *Digitalna knjiznica Univerze v Ljubljani*) has been established employing the concept. of local ETD systems (each faculty and department should have its own). Theses and dissertations are seen as one of the digital information resources students and teaching staff use (along with international and domestic e-journals, e-books, digital teaching materials, etc.). ETDs can also be accessed through the National Union Catalogue of COBISS where a link to the digital version can be added to an original catalogue input. A series of promotion activities was launched for teaching staff and students as well as for librarians. The electronic submission of ETDs will be next step, which should be easy due to the widespread use of various e-teaching programs in which students already present their papers in electronic form for supervision and grading. However, the tradition of written theses and dissertations may be a real obstacle.

Literature

1. Bailey CW. (2005) The role of reference librarians in institutional repositories. Reference Services Review. 33 (3), 259-267
2. Bevan, S. (2005), "Electronic thesis development at Cranfield University", Program: Electronic library and information systems, 39 (2), 100-11.
3. Carlson, J. (2006) An Examination of Undergraduate Student Citation Behavior The Journal of Academic Librarianship, 32 (1), 14–22
4. Copeland, S., Penman A., Milne R. (2005) Electronic theses: the turning point. Program: electronic library and information systems. 39 (3), 185-197
5. Davis PM., Cohen SA., (2001) The Effect of the Web on Undergraduate Citation Behavior 1006-1999, JASIS 52 (4), 309-14
6. Edminster, J., Moxley, J., (2002) Graduate Education and the Evolving genre of Electronic Theses and Dissertations. Computers and Composition 19, 89–104
7. Fox. EA., (editor) (2004) Electronic theses and dissertations. Marcel Dekker, New York, Basel :
8. McCabe DL. (2006) Cheating among college and university students: A North American perspective. international journal for educational integrity 1 (1)
9. Paillassard, P., Schöpfel, J. & Stock, C.(2004) How to get a French doctoral thesis, especially when you aren't French. GL6 conference 55-63 Online. Available http://www.greynet.org/images/GL6,_Page_143.pdf
10. Park EG, Nam Y, Singe OS. (2007) Integrated Framework for Electronic Theses and Dissertations in Korean Contexts. The Journal of Academic Librarianship, 33, (3), 338–346
11. Park, EG., Zoo, G., McKnight, D., (2007) Electronic thesis initiative: pilot project of McGill University, Montreal. Program: electronic library and information systems. 41 (1),
12. Stock C. (2007) Open access to full text and ETDs in Europe: improving accessibility through the choice of language? GL9 conference, Conference CD-ROM. Online. Available: http://opensigle.inist.fr/bitstream/10068/697889/2/GL9,_Stock,_2008,_Conference_Preprint.pdf
13. Suleman H et al, (2001) 'Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access—Part1: Mission and Progress,' D-Lib Magazine, 7 Online. available: http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html
14. Vijayakumar, J.K., Murthy, T.A.V., Khan M.T.M. (2007) Electronic Theses and Dissertations and Academia: A Preliminary Study From India. The Journal of Academic Librarianship, 33, (3) 17–421
15. Wood, G., (2004) Academic Original Sin: Plagiarism, the Internet, and Librarians. The Journal of Academic Librarianship, 30 (3), 237–242
16. Yiotis, K., (2008) Electronic theses and dissertation (ETD) repositories What are they? Where do they come from? How do they work? OCLC Systems & Services: International digital library Perspectives 24 (2), 101-115

References

1 http://www.mvzt.gov.si/en/areas_of_work/science_and_higher_education/higher_education/dejavnost_visokega_solstva/register_of_higher_education_institutions_in_the_republic_of_slovenia/#c16877

2 http://dikul.uni-lj.si

INTEREST - INTERoperation for Exploitation, Science and Technology

Keith G. Jeffery, STFC; Council Rutherford Appleton Laboratory, United Kingdom
Anne Asserson, University of Bergen, Research Department, Norway

Abstract
This paper addresses the topic of interoperation of Grey resources. The title should be read as INTERoperation for Exploitation, Science and Technology. It builds on work by the authors published in previous GL conferences. The method is architectural analysis and comparison. The costs of the study are negligible, but of course the costs of implementing any solution are considerable. The result/conclusion is that CERIF (Common European Research Information Format) is the essential component to meet the requirements and is applicable – to a greater or lesser degree - in all architectural solutions.
Our GL9 (2007) paper proposed a Grey landscape architecture and identified the need for (1) excellent metadata (to improve discovery and control usage), (2) an institutional document repository of (or including) grey, (3) an institutional CRIS (Current Research Information System) for the contextual research information, (4) linkage between the document repository and the CRIS of an institution and thence (in a controlled manner with formal descriptive and restrictive metadata) to other institutions, (5) an e-research repository of research datasets and software, (6) linkage between the e-research repository and the CRIS of an institution and thence (in a controlled manner with formal descriptive and restrictive metadata) to other institutions, (7) an institutional policy to mandate deposition of the material with appropriate metadata.
These very requirements define the components for interoperation of Grey resources, and their interoperation with other resources to provide a holistic support for R&D. Indeed they can be extended (via the CRIS) to interoperation with other management systems of an organisation such as finance, human resources, project management, production control etc.
However, the capability for interoperation can be provided in several implemented architectures. This paper discusses the advantages and disadvantages of different solutions including experience of their use. This analysis and experience is then applied to the grey environment. Remote and local wrapping of resources, cataloguing techniques and a full compliant model are discussed as well as harvesting technology. It concludes that – particularly for the grey environment – the optimal architecture involves formal syntax (structure of information) and defined semantics (meaning of information) as defined by CERIF.

Background

The Basis of the Paper
(1) Grey Literature repositories are much improved for the end-user (in integrity, relevance, quality and utility) when linked with a CERIF-CRIS or a CRIS capable of interoperating with other CRIS using CERIF;
(2) Although it is possible to link Grey literature repositories independently of CRIS if they store - for example, Dublin Core (DC) metadata and use OAI-PMH protocol for interoperation and OAISTER for searching – we propose that interlinking CERIF-CRIS (or CRIS capable of interoperating using CERIF) is better because of the formal syntax and declared semantics: CERIF contains sufficient metadata (Jeffery 2000) to provide better recall and relevance in retrieval than OAI-PMH-linked DC metadata-based systems;
At GL9 (Jeffery and Asserson 2007) we presented an architecture for utilising CRIS interlinked tightly using CERIF providing access to Grey literature repositories (both publications and research datasets/software). This paper attempts to present and compare architectures utilising CRIS (with assumed linked local grey literature institutional repositories) for interoperation. Of course (and presented in section 8) the ideal architecture is the use of a native CERIF-CRIS.

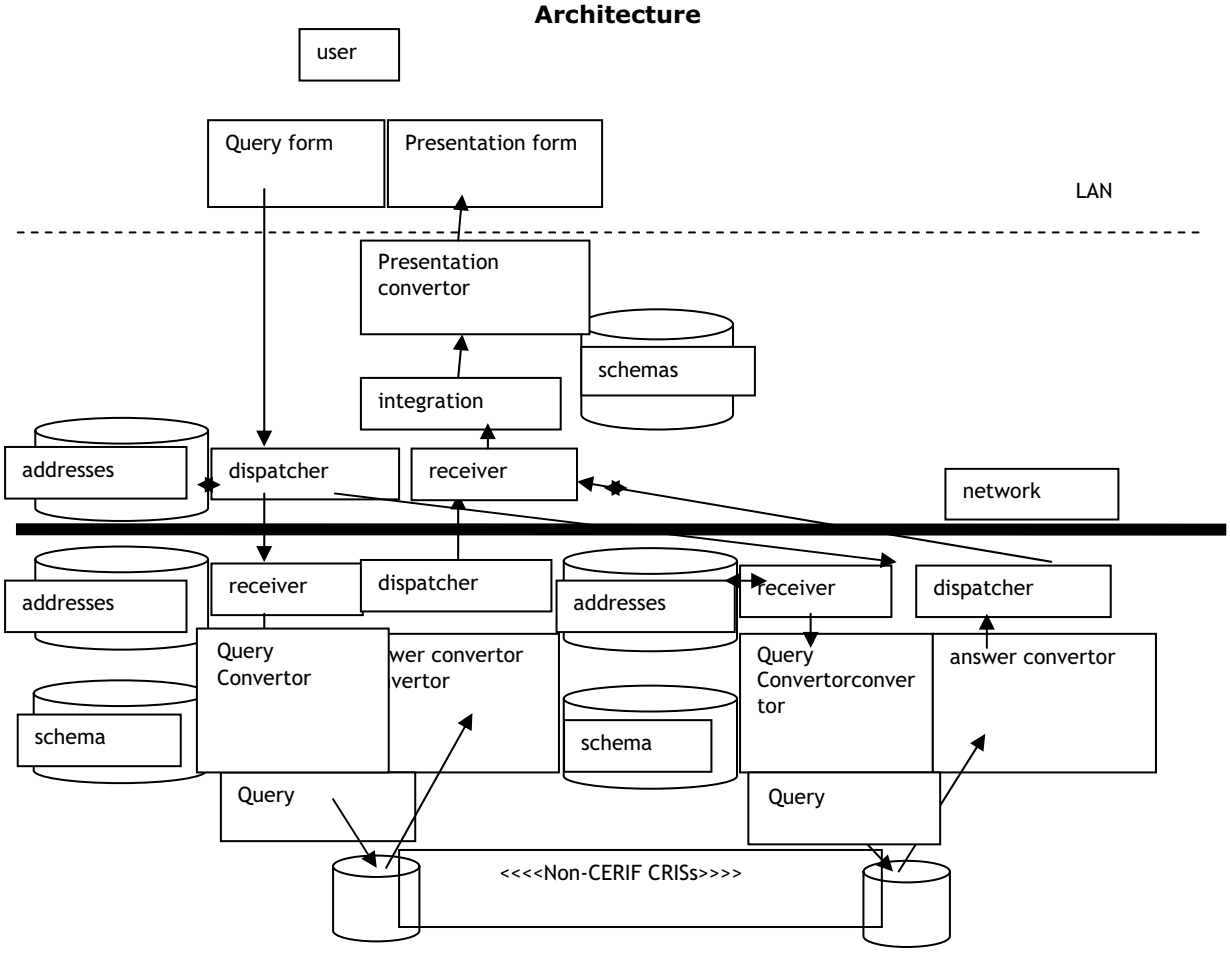
Metadata
Metadata may be classified into kinds: schema, navigational, associative with the latter partitioned into descriptive, restrictive and supportive (Jeffery 2000). Apart from harvesting, which ignores the syntax (structure) and semantics (meaning) of the data and just does text string searching, all architectures rely on a predicate query over a known schema (available or derived by schema reconciliation) thus allowing search terms or values to be related to an entity/attribute and thus to a domain. Example: the string 'green' could occur under attribute 'family name' in entity 'person' or within attribute 'abstract' or 'title' in entity 'project'. The use of query under a schema ensures that the query is meaningful and should have adequate recall (coverage) and relevance (precision).
Most techniques rely on navigational metadata to access hosts of CRISs. The catalog techniques use in addition associative descriptive metadata to perform the first pass search – rather like the harvesting

technique, but using structured and meaningful data under entity/attribute sanction. Those techniques with server-side or client-side wrappers require schema metadata to perform schema reconciliation. Although CRISs of these architectures may not have a native CERIF storage format, CERIF can be used with advantage to define the database schemas.
One technique (Full CERIF) has a uniform assumed schema and so has no need for metadata nor schema reconciliation. However, it relies on each host either being fully CERIF compliant or providing (and maintaining) by transformation a full CERIF version of the host database.
CERIF provides the optimal solution in the full implementation. All other techniques benefit from using CERIF to define schemas or export formats.

THE HYPOTHESIS
Comparison of possible architectures for interoperation of grey repositories (of publications or data and software) via structured CRISs leads inevitably to the conclusion that CERIF should be used either as the native storage format, as the storage format of a derived data warehouse (transformed copy of the CRIS) or as the export format converted from the CRIS native format using a wrapper.
The following sections describe each interoperation technique and categorise each under architecture (diagram), description of technique, metadata required, the process and the advantages and disadvantages.

REMOTE WRAPPER

Architecture



Description
This architecture provides a simple user query interface to multiple host CRISs. Each host CRIS has to provide navigational metadata to the client dispatcher database and provide software for query conversion to local host DML (data manipulation language) using the host schema.
The use of XML to encode answers (an addition to the basic architecture but indicated in the diagram) dispatched provides some syntactical uniformity but no uniformity of character set, language, semantics. Uniformity in these other aspects can only be achieved through a canonical data model (CERIF). Unfortunately XML cannot represent the full syntax (let alone semantics) of CERIF, because it represents hierarchies and CERIF represents a directed graph.

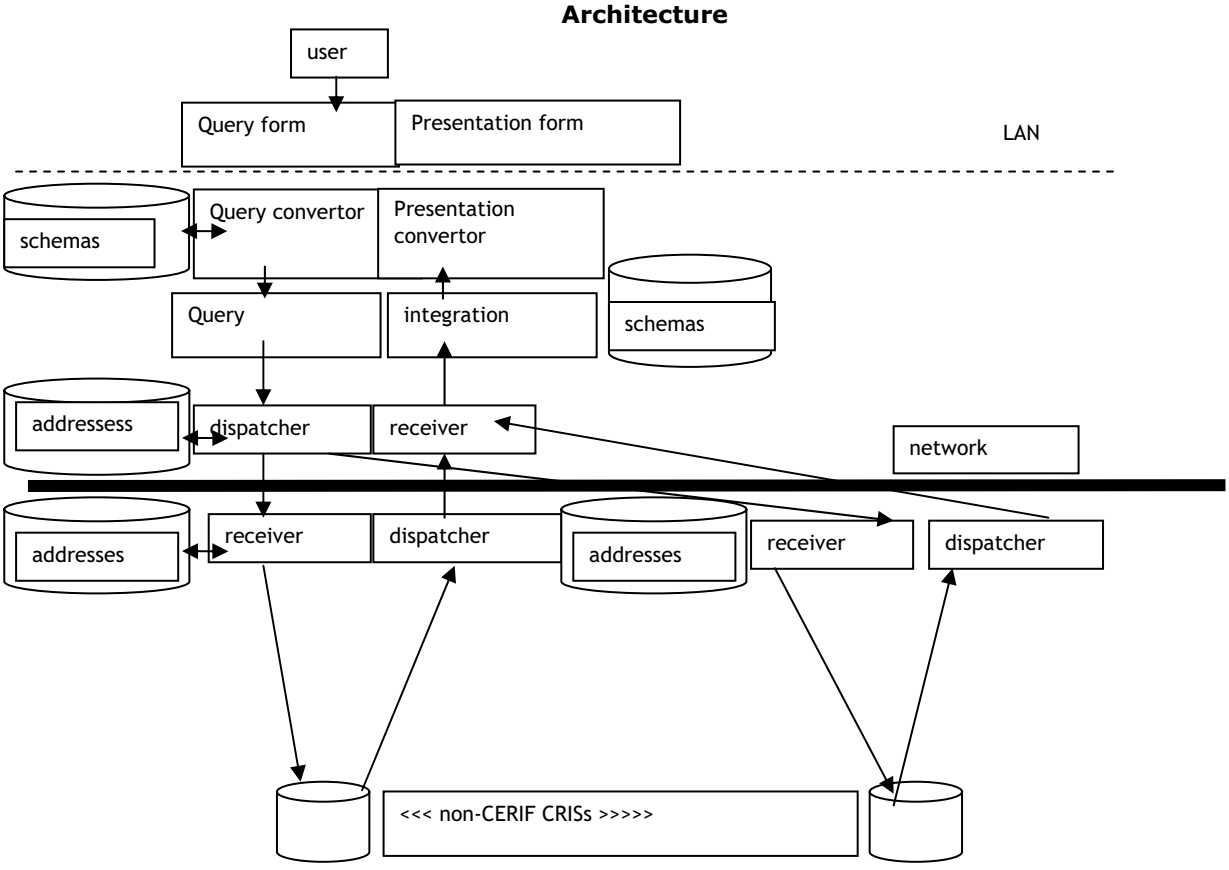
Metadata
This architecture uses schema metadata for query conversion and answer integration. It uses navigational metadata for access to hosts.

Process
The user inputs a query through a supplied web browser form of the kind 'find the string "widget" anywhere in any host database'. The dispatcher sends this query in some protocol [email | ftp | message] to all hosts with address entries in the address database. Each host converts to its own DML using the host schema and executes the query. The results at each host are converted to XML (an addition to the basic architecture but indicated in the diagram) and dispatched back to the user who receives one XML file per host (each with differing syntax and semantics). The integrator takes the result sets and using the host schemas - or preferably XML DTD (document type definition) equivalents of the schemas - reconciles them to a uniform result set which is converted for end-user viewing by the presentation converter.

Advantages and Disadvantages
- the user needs only web browser and simple query form
- the host has to write query converter
- the host has to write answer (XML?) converter (to a specific XML DTD?)
- the query expressivity is very limited
- the user client has to write an integrator for the answers

LOCAL WRAPPER

Architecture



Description
In this architecture the hosts have only to provide a receiver and dispatcher; they receive queries in their own DML and dispatch results in their own data model. All conversion responsibility is on the client. The client provides queries for each host from the user query, mediated by the host schemas and integrates the results from each host, using their schemas, to an answer for the end-user presented through a user-defined presentation converter (e.g. XML, HTML….).

Metadata
This architecture uses schema metadata for query conversion and answer integration. It uses navigational metadata for access to hosts.

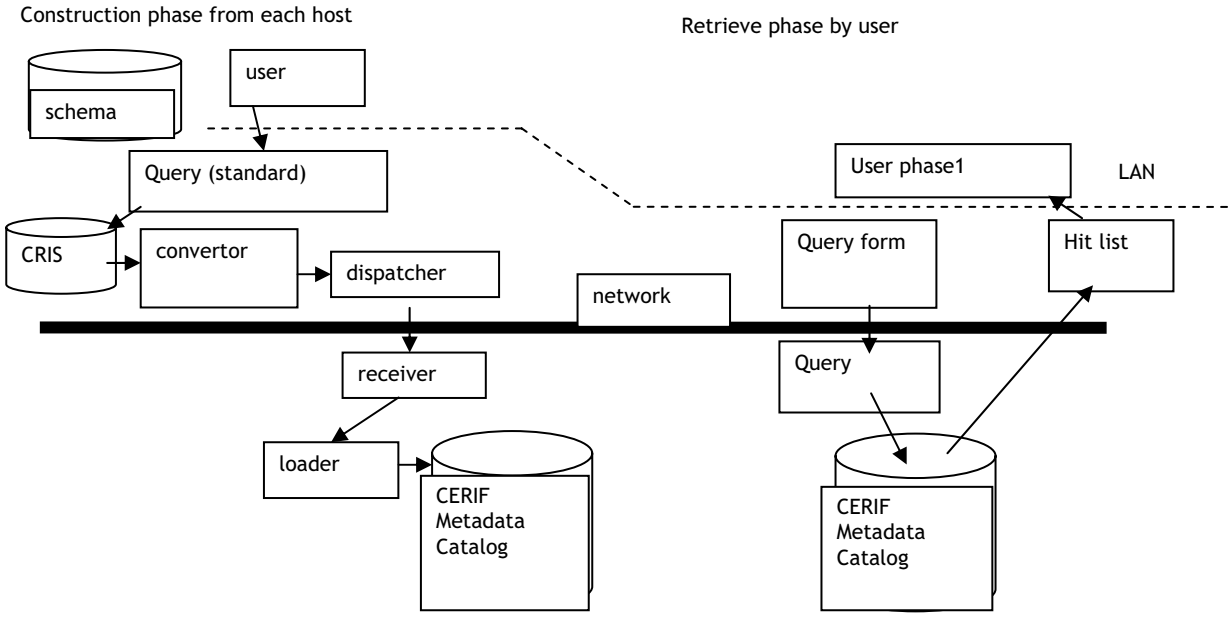
Process
The end-user generates a query in some arbitrary language, using a query refinement interface and web form. The client software converts the query to the target DML for each host using the host schemas stored (and updated by the hosts) at the client and dispatches them using the addresses database. Each host receives a query in its own DML, executes it and returns the result in its own form via the dispatcher to the client receiver. The integrator takes the result sets and using the host schemas reconciles them to a uniform result set which is converted for end-user viewing by the presentation converter. CERIF could, with advantage, be used as the uniform schema for result integration.

Advantages and Disadvantages
- each host has only to supply and update its schema to the client (all clients if there is not a central query server)
- each host has no software to provide except receiver and dispatcher
- the client (if it is a central service) has a very large workload
- if there is no central service then each client has to have all schemas supplied and updated
- the client software has to include a complex query refiner
- the client software has to include multiple complex query converters
- the client software has to include a complex answer integrator
- the client software has to include a presentation converter (complexity depends on specification of presentation required and complexity of the answer structure)

CATALOG

Catalog Only (ERGO Pilot)

Architecture



Description
This architecture provides a canonical subset data model – CERIF metadata model – with one character set, one language, one syntax (structure) and one semantics. This provides the homogeneity.

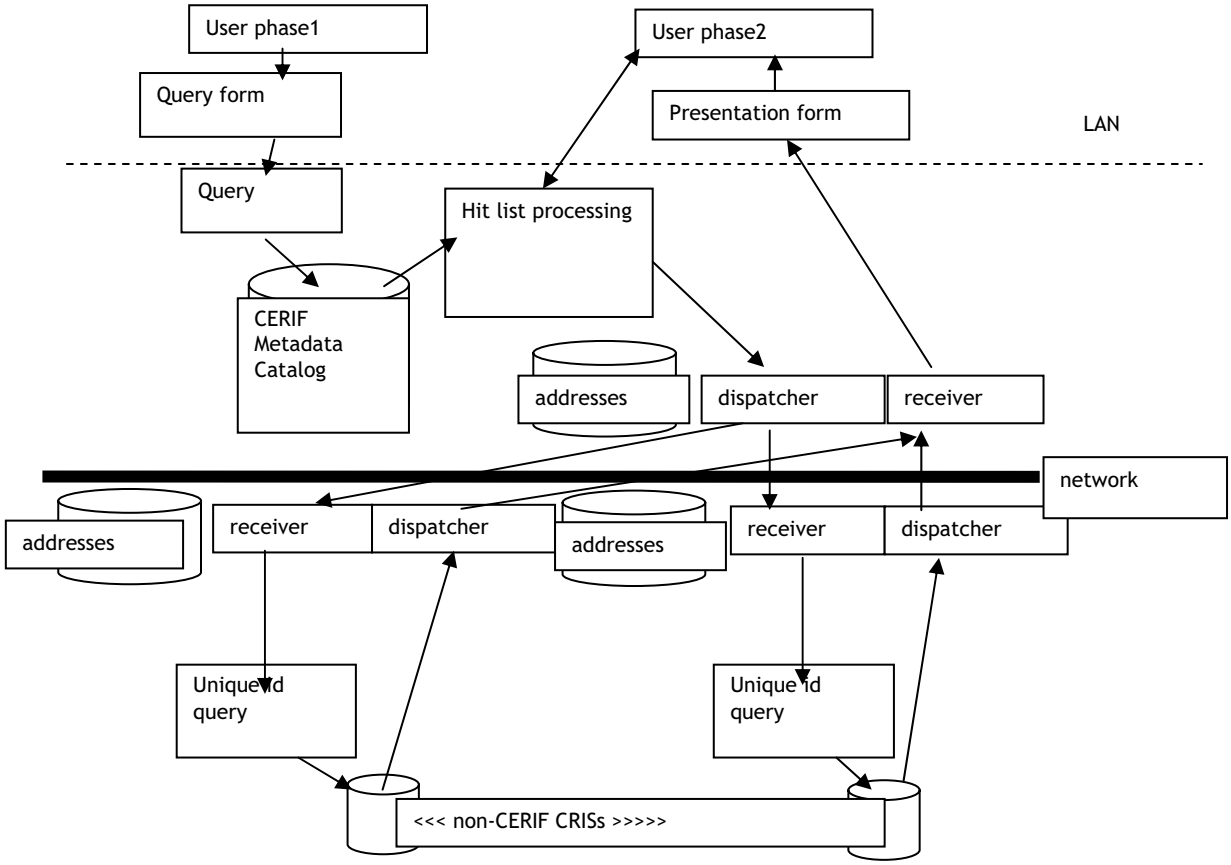
Metadata
This architecture uses associative descriptive metadata (CERIF metadata catalog)

Process
The CERIF metadata catalogue is populated from all hosts which provide a local converter from their data model to CERIF metadata (one character set, one language, one syntax (structure) and semantics (meaning)). The end-user has a query form which queries the catalog and obtains a 'hit list' of results. Experience indicates such results satisfy ~ 80% of queries; however if more detail is required the architecture provides the capability for accessing the hosts (see next section).

Advantages and Disadvantages
- simple query on union catalog (which may be centralised or replicated)
- possibly not all required entities and attributes in catalog
- effort to populate catalog; requires converter at each host to supply CERIF metadata

CATALOG PLUS PULL (ERGO 2++)

Architecture



Description
In addition to the Catalog-only model, this architecture allows a subsequent access to all hosts with hits in the CERIF metadata catalog to collect the detailed information the hosts are willing to supply. There is no further selection by attribute value nor projection of attributes, everything is 'pulled' if it relates to a hit record in the catalog. Ideally, the hosts convert to CERIF export model to provide uniformity but this is not mandatory.

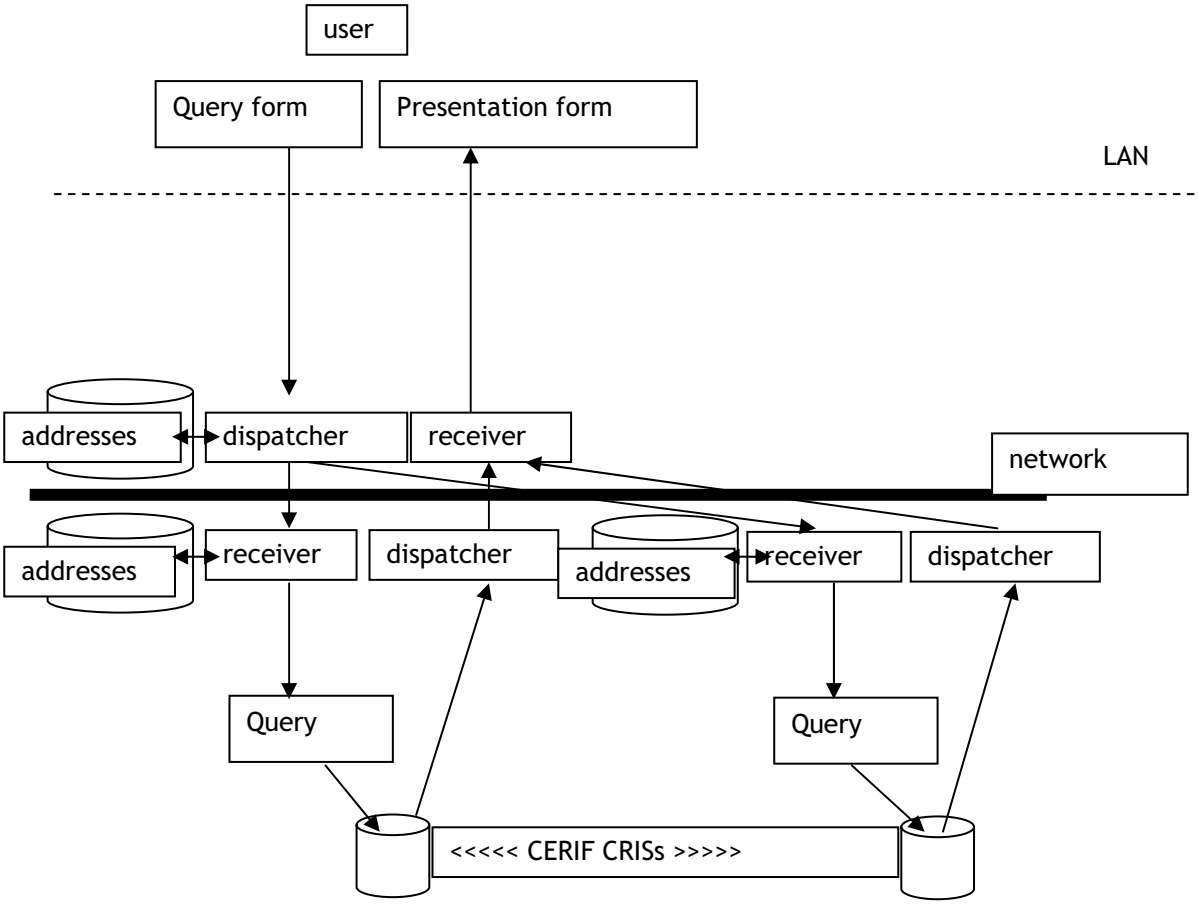
Metadata
This architecture uses associative descriptive metadata (CERIF metadata catalog) and navigational metadata for host addresses.

Process
The hit list is edited by the end-user and then sent to the dispatcher which sends to each host the Unique Ids (primary key values) of the selected records for which further information is demanded. The host sends the answers back via dispatcher to receiver and thence to the user. No attempt is made to make homogeneous this detailed heterogeneous information which may have different character sets, language, syntax and semantics. Conversion to CERIF export model could be advantageous for integration by the end-user.

Advantages and Disadvantages
- advantage of simplicity as for catalog-only architecture
- advantage of additional information provision
- disadvantage that additional information is heterogeneous (unless converted to CERIF export data model)
- disadvantage of hosts having to maintain entries representing their database content in the CERIF metadata catalog

Full CERIF

Architecture



Description
This architecture relies on the existence at each host of a full CERIF model database, either as the host database itself or a version of the host database converted to full CERIF model. This provides a completely homogeneous solution which is very simple.

Metadata
This architecture uses navigational metadata for host addresses. No other metadata is required as homogeneity is achieved through the full CERIF model.

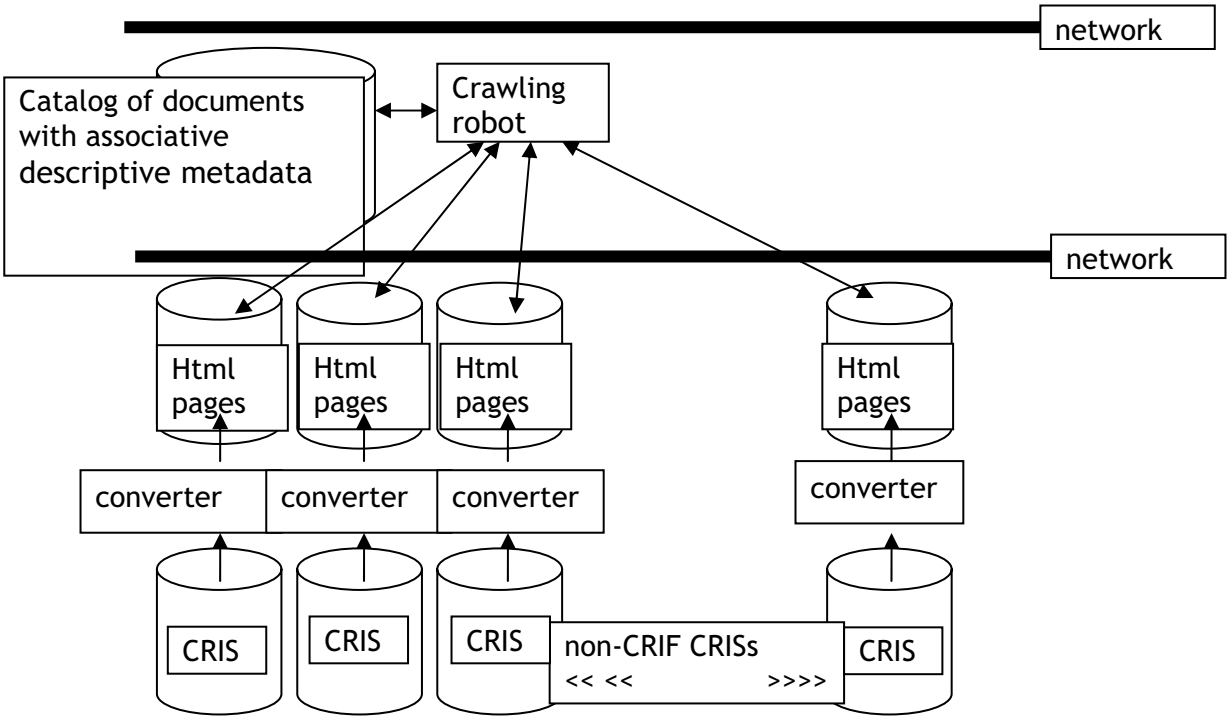
Process
The process is straightforward; through a webform the end-user queries (knowing the CERIF schema) and using normal distributed database technology the query is passed to all hosts; the answers are all in CERIF form so integration is automatic.

Advantages and Disadvantages
- very simple and easy to use for the end-user
- each host has to either run a full CERIF model database or provide a full CERIF model version of the host database

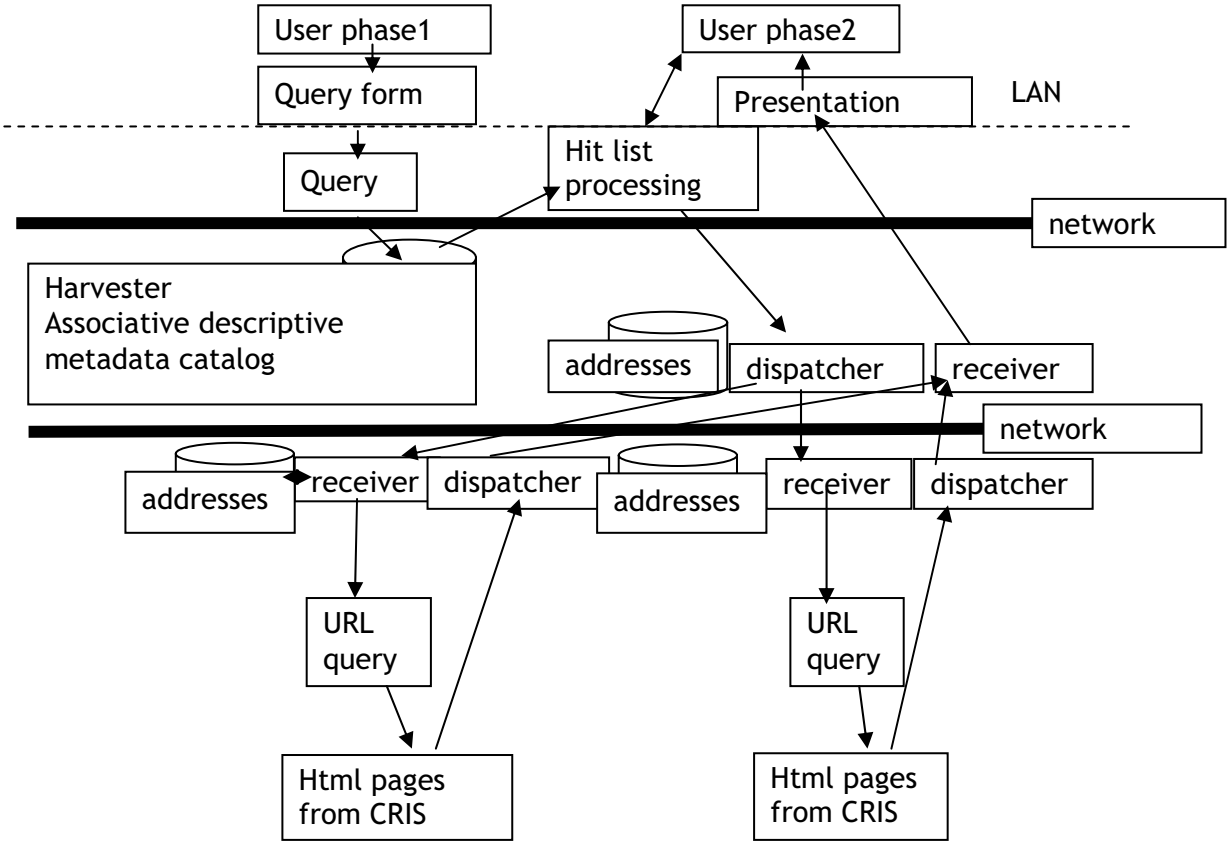
Harvesting
The concept of harvesting information from the whole WWW has been introduced. The power of modern search robots (to construct the catalog) and search engines (to search the catalog) is quite remarkable. However, much information is unavailable to harvesting being hidden in databases which may have a webform for query but which do not expose their information on webpages. Furthermore, the search robots usually take around 2 weeks to search the web and so the catalog is not up-to-date. A CRIS harvesting system should be more specific than, say, Google; this implies it searches only URLs known to be entrances to CRISs. The architecture (with catalog and reference to more detailed information) is not unlike the Catalog architecture of ERGO, but ERGO is based on structured data searchable under entity and attribute.

Architecture

Construction Phase



Search Phase



Description
This architecture relies on fast indexing of visible web pages by a search robot generating an associative descriptive metadata catalog which is then searched by the user; hits are followed up with a click on the URL to make available the detailed original web page.

Metadata
This architecture uses associative descriptive metadata in a catalog derived from the search robot and used by the search engine and navigational metadata for host addresses.

Process
First a search robot traverses the web; it may be instructed to search for certain terms but more likely is general. It constructs an associative descriptive metadata catalog as it goes, usually one entry per web page visited; the catalog also includes navigational metadata: the URL of the webpage indexed. This implies that any CRIS has to provide a set of web pages replicating the data in the CRIS to make it available to the search robot. Techniques are emerging to make structured or semi-structured databases visible to robots but there is no generally accepted technique yet.
The user then queries the catalog, and for every hit (meeting the search term(s)) receives a metadata record; clicking on the navigational metadata (URL in the metadata record) provides the original webpage.

Advantages and Disadvantages
- The host has to provide a copy of the database as webpages to be available to the search robot and subsequent accesses based on clicks from URL of metadata.
- The query is based on existence of term(s); constraining by entity or attribute is not possible (without sophisticated xml form processing).
- The results are unstructured and one page at a time (click on URL in metadata catalog to see page); this inhibits statistical processing or report generation.
- It is easy to implement and maintain (although the database may be ~2 weeks out of date) and has a familiar interface for many WWW users.

Conclusion
Clearly a full CERIF architecture provides maximum homogeneity and ease of use. However, it requires all hosts either to have their CRIS in CERIF or to provide a CERIF compatible version of their CRIS and make that version available to the federation system.
CERIF can, with advantage, be used as the canonical model for conversion from other CRIS when integrating using either remote or local wrapper techniques. It reduces the (n*(n-1)) interconversion problem to (n), where n is the number of participating CRIS.
CERIF metadata provides structured query capability in the catalog model(s), distinguishing this technique from harvesting.
Under any efficient architecture, CERIF remains the core technology for homogeneous access to heterogeneous CRIS and hence to associated Grey Literature repositories.

References & Bibliography

(Asserson and Jeffery 2004) Asserson, A; Jeffery, K.G.; 'Research Output Publications and CRIS' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 29-40 (available under www.eurocris.org)

(CERIF) www.eurocris.org/cerif

(DC) http://www.oclc.org:5046/research/dublin_core/

(DSpace) http://www.dspace.org/

(FRIDA) http://frida.uio.no

(Jeffery 1999) Jeffery, K G: 'An Architecture for Grey Literature in a R&D Context' Proceedings GL'99 (Grey Literature) Conference Washington DC October 1999 http://www.greynet.org

(Jeffery, 2000). Jeffery, K G: 'Metadata': in Brinkkemper,J; Lindencrona,E; Solvberg,A (Eds): 'Information Systems Engineering' Springer Verlag, London 2000. ISBN 1-85233-317-0.

(Jeffery 2004) Jeffery, K.G.; 'GRIDs, Databases and Information Systems Engineering Research' in Bertino,E; Christodoulakis,S; Plexousakis,D; Christophies,V; Koubarakis,M; Bohm,K; Ferrari,E (Eds) Advances in Database Technology - EDBT 2004 Springer LNCS2992 pp3-16 ISBN 3-540-21200-0 March 2004

(Jeffery 2004a) Jeffery, K.G.; 'The New Technologies: can CRISs Benefit' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 77-88 (available under www.eurocris.org)

(Jeffery and Asserson 2004) K G Jeffery, A G S Asserson; Relating Intellectual Property Products to the Corporate Context; Proceedings Grey Literature 6 Conference, New York, December 2004; TextRelease; ISBN 90-77484-03-5

(Jeffery and Asserson 2005) K G Jeffery, A G S Asserson 'Grey in the R&D Process'; Proceedings Grey Literature 7 Conference, Nancy, December 2005; TextRelease; ISBN 90-77484-06-X ISSN 1386-2316

(Jeffery and Asserson 2006) Keith G Jeffery, Anne Asserson: 'Hyperactive Grey Objects' Proceedings Grey Literature 8 Conference, New Orleans, December 2006; TextRelease; ISBN 90-77484-08-6. ISSN 1386-2316 ; No. 8-06-X

(Jeffery and Asserson 2007) Keith G Jeffery, Anne Asserson: 'Greyscape' Opening Paper in Proceedings Grey Literature 9 Conference Antwerp (GL9) 10-11 December 2007 pp9-14; Textrelease, Amsterdam; ISSN 1386-2316

(RDF) http://www.w3.org/RDF/

(XML) http://www.w3.org/XML/

VICTORINE VAN SCHAICK PRIZE 2008

ISSN 1574-1796

An International Journal on Grey Literature

Autumn 2008 – TGJ Volume 4, Number 3

'MAKING GREY MORE VISIBLE'

www.textrelease.com GreyNet Grey Literature Network Service www.greynet.org

OpenSIGLE, Home to GreyNet's Research Community and its Grey Literature Collections: Initial Results and a Project Proposal

Dominic Farace and Jerry Frantzen; Grey Literature Network Service, Netherlands
Christiane Stock and Nathalie Henrot; INIST-CNRS, France
Joachim Schöpfel; University of Lille 3, France

Introduction

For the past 16 years, GreyNet has sought to serve researchers and authors in the field of grey literature. To further this end, GreyNet has signed on to the OpenSIGLE repository and in so doing seeks to preserve and make openly available research results originating in the International Conference Series on Grey Literature. GreyNet together with colleagues at INIST-CNRS have designed the format for a metadata record, which encompasses standardized PDF attachments of the authors' conference preprints, PowerPoint presentations, abstracts and biographical notes. In April 2008, the first test batch containing records from the Eighth International Conference on Grey Literature (GL8, 2006) was uploaded. A few minor problems that were encountered and have since been successfully resolved. These metadata records and their corresponding attachments are now available for search and retrieval in OpenSIGLE. Subsequent record entries followed with GL7 (2005) down to GL6 (2004) and GL5 (2003). By December 2008, conference records over the past five years including those from GL9 (2007) will be available in the OpenSIGLE Repository. For this phase of the project, a budget of 2000 Euro was appropriated to cover the costs of formatting, conversion, and technical editing of the 100 plus metadata records and 300 accompanying PDF attachments. Records from the earlier four conferences in the GL Series (1993-1999) will require additional image scanning as well as permission from Emerald (the former MCB University Press). Should this be granted not only would the total number of GreyNet records in OpenSIGLE be nearly doubled but GreyNet's collection would then be comprehensive.

Method of Approach

If OpenSIGLE is indeed the best home for GreyNet, then some measure of empirical results should be able to confirm it. Results that would demonstrate benefits for both the GreyNet Collection as well as OpenSIGLE. For it is here, where the crossroads of more than 25 years of bibliographic information on grey literature intersects with 15 years of research on grey literature. The analysis of usage statistics and local metrics can draw on the standards and definitions of the COUNTER project for journals and databases but must take into account that little has been published so far on usage statistics of documents deposited in open archives, that standards, recommendations and empirical evidence are still missing, and that the software for the export of statistics need to be improved. Approach, methodology and preliminary usage data will be presented with special attention to comparative data when available from INIST and GreyNet websites, and to the potential and real impact of PR campaigns and referencing on usage. This may also lead to the evaluation of the role and impact of OpenSIGLE and GreyNet's Collection on the development and functioning of the international GreyNet community and the creation of community-related tools and functionality (web 2.0). If GreyNet is to factor into the design of the 'Grey Grid' for information society, then not only it's place in serving researchers and educators in the field of grey literature must be re/evaluated but also it's place in serving practitioners in the field. Such a study would help bring this home.

PART ONE

SIGLE to OpenSIGLE in Five Seconds

SIGLE (System for Information on Grey Literature in Europe) was a unique multidisciplinary database dedicated to grey literature. Up to 15 European partners participated in SIGLE, mostly national libraries or important research libraries. Created in 1980 and produced from 1984 onwards by EAGLE (European Association for Grey Literature Exploitation), the database was last available through STN International and on CD-ROM via Silverplatter/Ovid, until it became dormant in 2005. INIST then decided to make the data publicly available on an open access platform. Details of the migration from SIGLE to OpenSIGLE have been presented at the GL8 Conference held in December 2006 (Schöpfel 2007)[1]. The OpenSIGLE website went live in December 2007.

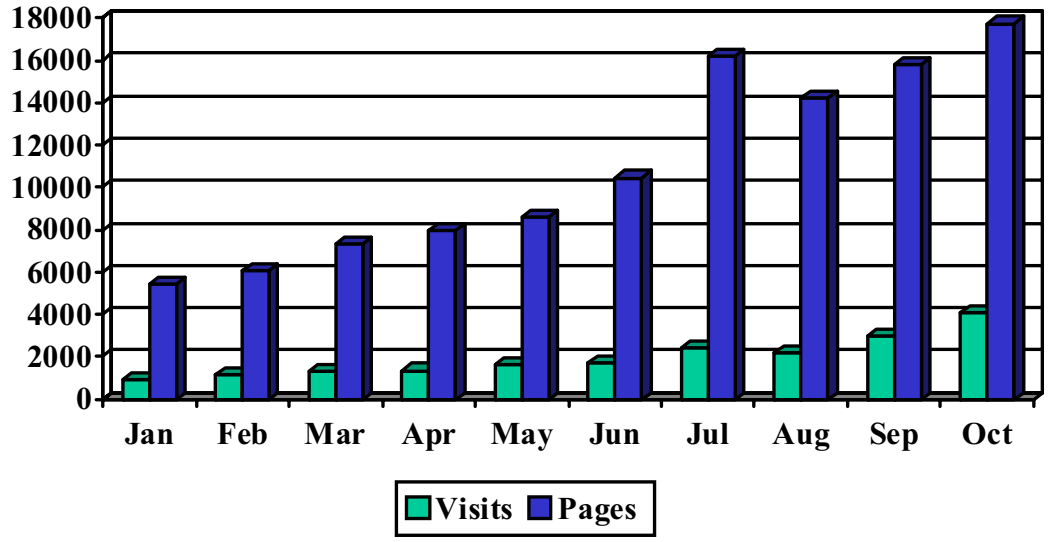
This paper further discusses two aspects of OpenSIGLE: (1) usage statistics covering one year of access to the repository and (2) a cooperative agreement with GreyNet, the Grey Literature Network Service.

OpenSIGLE Traffic Report

Our usage analysis is based on data obtained through *phpMyVisites*, an open source software for website statistics that works with a javascript image call. Only completely uploaded pages are counted and robots

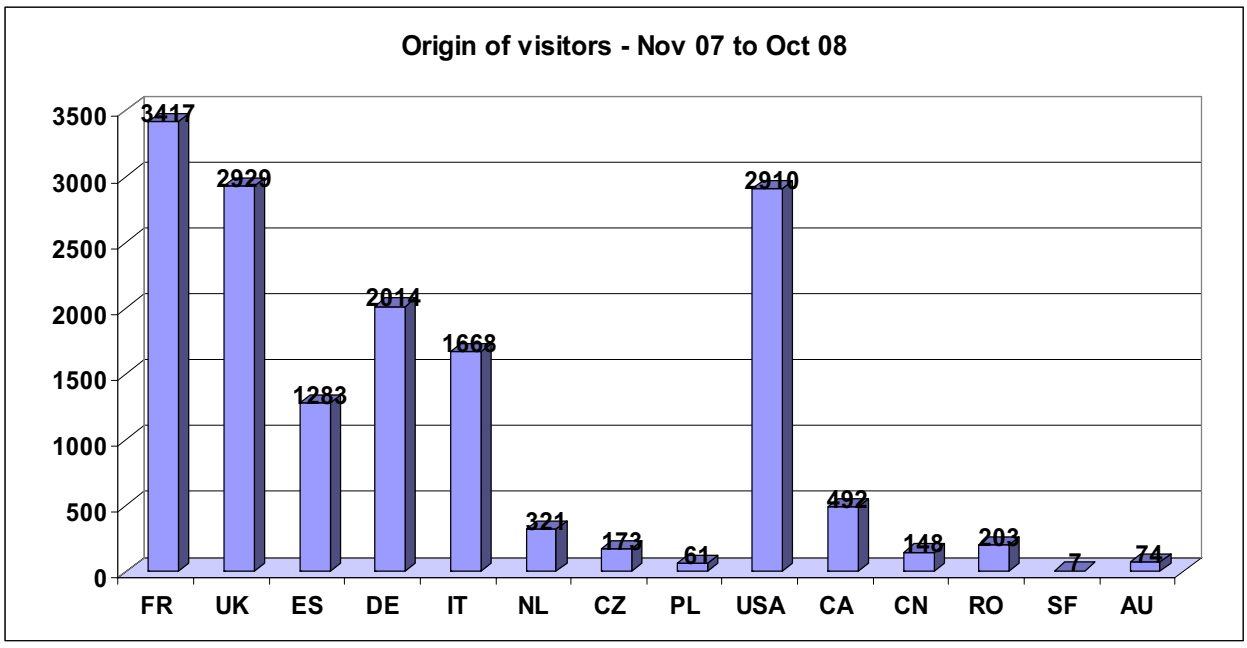
are excluded. Other statistics based on server logs might however provide higher numbers. The following data provides only a part of the information that can be obtained through *phpMyVisites*.
The first figure shows that the number of visits as well as the number of pages have increased steadily since the opening of the website in 2007. The peak in July is due to a press campaign in the middle of the French holidays. The result is both surprising and rewarding since visits usually go down during the summer months.

Figure 1: OpenSIGLE 2008 traffic report – number of visits and pages viewed



Geographic Origin of Visitors
Our software allows us to monitor the origin of the visitors for the top ten countries each month. For countries at the end of the list the global number may be higher, but they are not counted if they are in position 11 and below. The sum of 12 months worth of data shows France in the lead, closely followed by the United Kingdom and the United States. A second group is formed by Germany, Italy and Spain. We clearly see that OpenSIGLE users come not only the former EAGLE countries, but also from the United States, Canada, and since recently China. France took the lead from June 2008 onwards, before then UK users were in front.

Figure 2: Origin of visitors to OpenSIGLE



Usage and Feedback

Compared to other INIST websites and e-resources, statistics show that between 16% to 19% of our users are from North America. OpenSIGLE is in third place among users from this continent preceded by the English version of INIST's institutional website (http://international.inist.fr) and IndicaSciences - an INIST product dedicated to research evaluation and indicators (http://indicasciences.inist.fr). INIST websites geared to a French speaking audience receive about 7% of the visits from North America.

The analysis of web links as well as other feedback reveal that OpenSIGLE is fairly often used in the biomedical and public health sectors. At present, statistics don't allow us to go into further detail regarding scientific domains.

During the course of 2008, we received several requests from former users of the STN or the WebSpirs SIGLE database pertaining to complex search strategies. These requests required us to look into the limits of the Jakarta Lucene search engine implemented with DSpace, especially with regard to the length of the search query. We discovered that Lucene allows more possibilities than mentioned in the help provided by DSpace. Besides inquires involving search strategies, other users were interested in the download and export features of OpenSIGLE.

One critical view of OpenSIGLE found on a blog, slates the absence of links to the full text of documents (http://healthinformaticist.wordpress.com/2008/08/28/does-opensigle-exist-for-its-own-sake/). This of course is understandable given the fact that it was one of the very reasons why the SIGLE database was discontinued.

Promotional Activities

Before the official announcement of the launch of OpenSIGLE, we presented the project at a DSpace user group meeting and exchanged experiences with other users (Grésillaud, October 2007)[2]. Shortly afterwards, and as a result of that meeting, we received visitors from Spain and Italy. And in December 2007, attention was also focussed on OpenSIGLE during GL9, the Ninth International Conference on Grey Literature in Antwerp, Belgium.

In May 2008, a short presentation for the French public was given at I-expo (IT conference and exhibit) in Paris. And in July, INIST sent a press release to national and international lists and agencies (i.e. Information World Review and Research Information). This no doubt resulted in the abovementioned peak of visits in the middle of summer. Simultaneously a brief message was placed both on the French and international homepages of INIST. Since "news items" are less frequent during summer months, the message about OpenSIGLE remained for a longer period of time on these WebPages.

Today OpenSIGLE is indexed by Google and Google Scholar and included in the bookmarks of national libraries and research institutes. Following the creation of the WorldWideScience Alliance in June 2008, in which INIST is a partner, we proposed to integrate into the WorldWideScience portal and this was realized in September 2008. Our web statistics for October show that WWS.org[3] is the forth partner site for visitors accessing OpenSIGLE through a website and GreyNet.org[4] follows closely behind. Overall, these different promotional activities appear to have had a positive impact on the use and branding of OpenSIGLE.

Future Developments

Before discussing further the cooperation between OpenSIGLE and GreyNet, let me conclude this first part of our paper with an outlook to 2009 and beyond.

The work of designing and redesigning Websites is never finished - just as with other information products and services. OpenSIGLE can be further developed in any number of directions. One of which will be the new look planned for 2009. Other directions that we are looking into include:

1. Uploading the French data from 2005 onwards and thus closing the gap between the SIGLE and OpenSIGLE records
2. Integrating links to the full text whenever it exists
3. Inclusion of the Dutch SIGLE records
4. Inclusion of current records from other countries
5. Integrating OpenSIGLE into other networks and portals
6. Etcetera

PART TWO

GreyNet, On the Background and Forefront

In the first part of this paper, we encountered a quarter of a century history and development in warp tempo. In this second part, I would like to address the relationship between GreyNet and the former EAGLE Association including its SIGLE database and then move to a conscious positioning of GreyNet in the newfound OpenSIGLE Repository with INIST as Service Provider.

In 1992, EAGLE agreed to act as main sponsor for the launch of the First International Conference on Grey Literature, which was held in the Amsterdam RAI. GreyNet was at that time a newly established network service – driven on two fronts: (1) to promote the field of grey literature and the work of organizations involved in this branch of information the world over, and (2) to stimulate research on grey literature and make the results available both in print and digital (electronic) formats. EAGLE participated as sponsor and/or program committee member in the first five Conferences in the GL-Series.

In early 2005 GreyNet was invited as an observer to the final EAGLE Board meeting at FIZ Karlsruhe, where the Association formally voted to be dissolved. It was at that same meeting that the initial draft of an OpenSIGLE proposal was presented by Dr. Joachim Schöpfel[5], who was the last in line of EAGLE Presidents.

In the two ensuing years (2005-2007), INIST worked unilaterally on OpenSIGLE, which could then be described as a caretaker repository. In the autumn of 2007, once OpenSIGLE had become operational, GreyNet met with colleagues at INIST to hammer out an agreement that on the one hand would make GreyNet OAI-compliant and on the other hand would expand INIST's role in OpenSIGLE from solely a caretaker to an external service provider. And to this end, GreyNet's conference based collections would provide an example of OpenSIGLE's potential for other data providers in the grey literature community.

The GreyNet Community in 2008

As of December 2008, the last five years of research issuing from the GL Conference Series has been uploaded in the OpenSIGLE Repository. The bilateral contact between INIST as service provider and GreyNet as data provider was successful in customizing a metadata record for the enriched publication of conference preprints and the subsequent migration of GreyNet's collections to an open access environment. The bilateral agreement likewise holds for future conferences in the GL-Series, continuing with GL10 records onward. Retrospective input of the initial four conferences in the GL-Series (1993-1999) would however make GreyNet's collections comprehensive in OpenSIGLE. And to this end, contact has been undertaken with the Emerald Group Publishing Limited – the former MCB University Press – which still holds copyright on this material.

The initial reaction from the grey literature community to GreyNet's alliance with OpenSIGLE has been positive, however due to the brief timeframe in which GreyNet's collections are actually available in the OpenSIGLE Repository, it is yet too early to provide you with any substantial user statistics. While GreyNet has been receiving monthly reports from INIST generated via OpenSIGLE, GreyNet is looking at ways to compile use and user statistics as well as other feedback via its own channels. In this way, early down the road, there would be separate data issuing from INIST as service provider and GreyNet as data provider, which might allow us to draw comparisons and provide grounds for decisions in the future. In September of this year, an OpenSIGLE webpage was added to the GreyNet website with hyperlinks to its conference collections. From September to November visits to this webpage have been a little over hundred per month; however, now that GreyNet's current collections are all available in OpenSIGLE, what was a sub-page on the GreyNet website will in January 2009 become a main page - visible and clickable from GreyNet's homepage. This will no doubt create more traffic to the OpenSIGLE Repository; and at the same time allow for the addition of its own sub-pages for content, promotional, and instructive purposes.

GreyNet's Potential for OpenSIGLE

The Grey Literature Network Service feels that it has still more to offer OpenSIGLE than its collections of conference based research. Going back to 1992, when GreyNet was first launched, one of its primary goals mentioned earlier in this paper was to promote the field of grey literature and the work of organizations involved in this branch of information. To this end - what EAGLE was to SIGLE, GreyNet could be to OpenSIGLE and more. GreyNet operates internationally and maintains a full-time established network service specializing in grey literature with information products and resources both in print and electronic formats. GreyNet has for the past six years (2003-2008) often times together with colleagues from France and The Netherlands carried out research projects involving citation analysis, questionnaires and surveys, interviews, as well as standard review of literature.

GreyNet has incorporated its own results from these small scale projects with the results from research carried out by other of the 250 authors/researchers in the GL-Conference Series in order to develop an academic curriculum for students of Library and Information Science (LIS). This course curriculum was first offered for 3 hours of college credit via the distance education program at the University of New

Orleans (UNO) in the Fall 2007semester[6]. After an evaluation and revisions, it will once again be offered in the Spring 2009 semester. GreyNet seeks to expose other LIS Colleges and Schools to curriculum opportunities in grey literature. And, at this very conference we have ten masters students from the University of Amsterdam, who will be using results from the GL10 published research in the completion of their own assignments linked to course credit.
Over the past 16 years (1992-2008), GreyNet has developed channels for promotional outreach as well as a publishing arm. These could no doubt serve and support future developments in the OpenSIGLE Repository.

Further Considerations and Concluding Remarks
Over the past year, in the course of migrating GreyNet's collections to the OpenSIGLE Repository, a number of issues arose and give way to serious consideration, such as:

- Streamlining the SIGLE Classification Scheme for the OpenSIGLE Repository
- Customizing metadata templates for various grey literature document types
- Plus links to datasets and software underlying the results of published research
- Networking with former EAGLE members and new stakeholders in Grey Literature
- Proving a crosswalk to subject based and institutional grey literature resources and collections
- Etcetera

What began unilaterally with the vision and determination of INIST-CNRS and what has recently been expanded in bilateral cooperation with GreyNet has yet even greater potential for the global grey literature community. GreyNet together with INIST-CNRS are committed to drafting a project proposal. This proposal will explore the capacity required for the OpenSIGLE Repository to further develop in multi-lateral and international cooperation in the support of European research infrastructures committed to open access of their grey literature collections and resources.

References

1 Schöpfel, J., C. Stock, and N. Henrot (2007), From SIGLE to OpenSIGLE and beyond: An in-depth look at Resource Migration in the European Context. – In: The Grey Journal : An International journal on Grey Literature, vol. 3, no 1, Spring 2007. - ISSN 1574-1796

2 Grésillaud, S., and C. Stock (October, 2007), DSpace at INIST-CNRS: one platform, different usages and resulting specific needs/problems. Paper presented at DSpace User Group Meeting 2007, Food and Agriculture Organization of the United Nations, Rome, Italy.
Available at http://www.aepic.it/conf/viewabstract.php?id=208&cf=11

3 WorldWideScience.org, the global science gateway http://worldwidescience.org/

4 GreyNet, Grey Literature Network Service http://www.greynet.org/

5 Schöpfel, J. (2006), MetaGrey Europe, A Proposal in the Aftermath of EAGLE-SIGLE. – In: GL7 Conference Proceedings, pp. 34-39. – ISBN 90-77484-06-X

6 Farace, D.J., J. Frantzen, J. Schöpfel and C. Stock (2008), Grey Literature: A Pilot Course constructed and implemented via Distance Education. – In: The Grey Journal : An International journal on Grey Literature, vol. 4, no 1, Spring 2008. - ISSN 1574-1796

Grey Literature on Caste-based Minority Community in India

Jyoti Bhabal
SHPT School of Library Science; SNDT Women's University, India

Background
According to the ancient Hindu scriptures, there are four "varnas' (groups). Manusmriti has mentioned four varnas: the Brahmins (teachers, scholars and priests), the Kshatriyas (kings and warriors), the Vaishyas (traders), and Shudras (agriculturists, service providers, and some artisan groups). Offspring of different varnas belong to different Jātis (Castes). Another group excluded from the main society was called Parjanya or Antyaja. This group of former "untouchables" (now called Dalits) was considered either the lower section of Shudras or outside the caste system altogether. (Caste system in India, n.d.)
Despite its constitutional abolition in 1950, the practice of 'untouchability' – the imposition of social disabilities on persons by reason of birth into a particular caste – remains very much a part of rural India. (Narula, Smita, n.d.)
The communities that are socially deprived due to their caste are categorized as Scheduled Caste (SC), Scheduled Tribes (ST), Other Backward Class (OBC) , Denotified and Nomadic Tribes (DT/NT). In this study all these communities together will be discussed as CBM.
Today the educated CBM are trying to use intellectual and organizational means to fight the caste system. Some visible efforts are: using conferences and media, publication of books and journals, forming discussion groups, action groups and building websites to create awareness. They educate themselves on the constitutional and legal rights of the CBM and fight for their implementation and extension using national and international forums. They internationalize CBM issues to get world attention and support. (Melliyal Annamalai, 2002)
There were and are documentations on various issues, actions, and movements among CBM. Most of these are at local level. This paper highlights the life cycle of such documentations in the form of grey literature available in the libraries of Mumbai.

Grey Literature (GL): Grey literature is defined as 'semi-published material for example reports, internal documents, theses, etc. not formally published or available commercially and consequently difficult to trace bibliographically. (Harrold's Librarian's Glossary and Reference Book, 2000)
The Internet is now a major source for dissemination and retrieval of grey literature and often serves as the initial introduction to a topic area. Some of the examples of e-grey literature are institutional archives and repositories, search portals and databases, e-print archives and directory of institutional links. (Rajendiran, P, 2006)

Objectives of the study:
- To find the grey literature available on various issues of CBM in the seven libraries of the city of Mumbai.
- To understand the post-acquition life cycle of grey documents on CBM in these libraries.
- To know the effort/ measures taken by the libraries to enhance the use of grey literature on CBM issues.

Sample:
Libraries that were chosen for survey were the ones that were catering to the student community. These libraries were SNDT Women's University Library, New Marine Lines (SNDT Library), Jawaharlal Nehru Library of University of Mumbai Library, Santacruz (MU Library), The Aditya Birla Memorial Library of Nirmala Niketan College of Social Work, New Marine Lines (NNCSW Library), Central Library of Indian Institute of Technology, Powai (IIT Library), Library of International Intstitute of Population Studies, Deonar (IIPS Library), Sir Dorabji Memorial Library of Tata Institute of Social Science, Deonar(TISS Library), Indira Gandhi Institute of Development and Research Library Goregon (IGIDR Library).

Methodology:
A structured questionnaire was prepared to collect data on total GL collection of library on CBM, its format and its language. Some of the questions were related to technical processing of GL on CBM such as mode of acquisition, its classification, cataloguing, and maintenance. Some of the questions were asked to know the user and use of GL on CBM; further special efforts undertaken related to acquisition, analysis, storage, and dissemination of GL on CBM

Online questionnaires were sent to each library to gather basic information about their grey collection from acquisition till dissemination. In addition, the various issues related to the life cycle

of GL on CBM were discussed directly with the librarian and library staff. The data collected during individual discussion was informative and supplemented to the questionnaire data.

Individual library catalogue was searched to find total number of the Grey documents available on CBM issues in the seven libraries. To retrieve precise data several generic as well as special descriptors were used. Generic descriptors used were SC/ ST, Castes, Tribes, Scheduled Castes, Scheduled Tribes, Dalit, etc. Few other specific names of caste and tribes (i.e Mang, Matang, Koli, Paradhi, Chambhar) were also used.

SPSS software was used to analyse and generate findings.

GL Collection on CBM:

Total 341 grey documents on various CBM issues were retrieved. All documents were categorised into Report (includes monographs, research report, survey reports excluding government reports, etc), Thesis and Dissertations, Government Publications (includes reports and statistical data- census data), Working papers, Conference Proceedings, Bibliographies. There were 156 Reports (47.5%), 84 Thesis and dissertations (24.6%), 74 Government publications (21.7%), 20 Working papers (5.9%), two Conference proceedings (0.6%) and 5 Bibliographies (1.5%) available as GL on CBM.

Chart 1: Library wise holding of various types of GL covering CBM issues

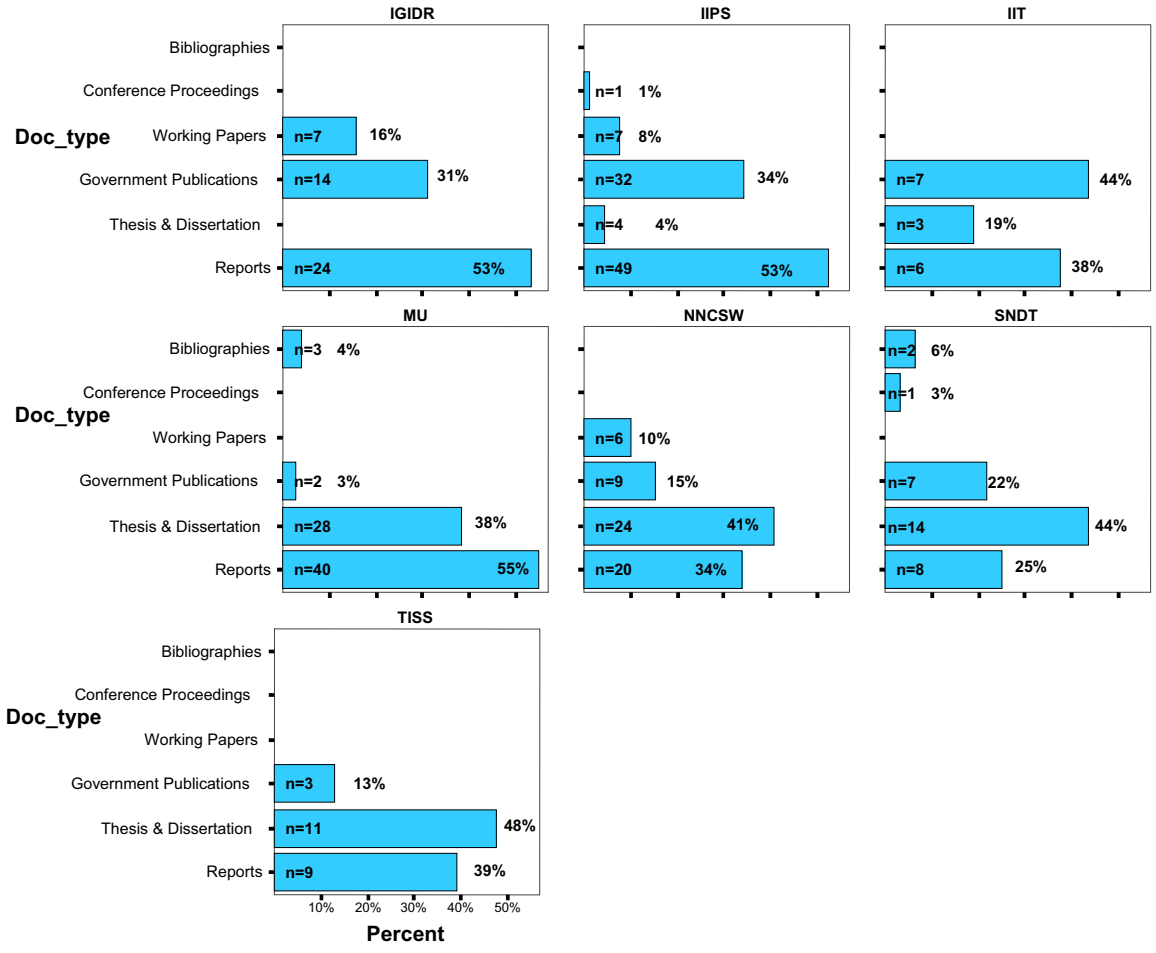


Chart 1 shows library-wise holdings of GL on CBM. It was found that IGIDR Library, IIT Library and IIPS Library had major collection of Government Publications and Reports; whereas MU Library, NNCSW Library, SNDT Library and TISS Library had major collection of Thesis and Dissertations and Reports. Bibliographies, Conference Proceedings and Working Papers were the minor collection in all seven libraries. SNDT Library and IIPS Library each had only single conference proceeding on CBM.

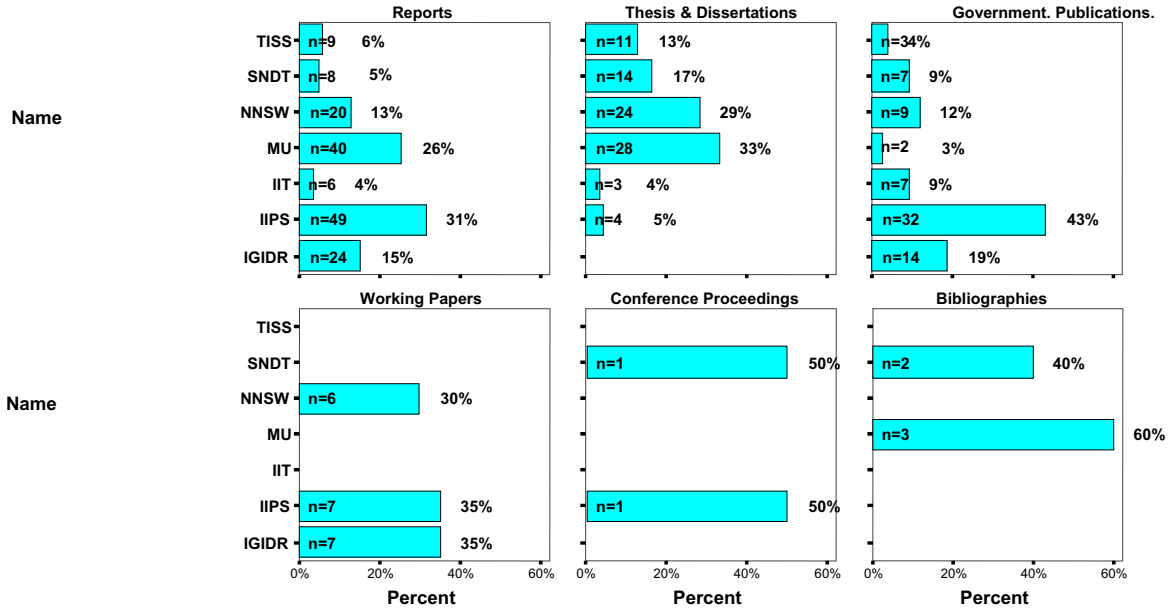
Recently a number of National and Local level conferences, seminars were conducted all over India on several issues of CBM. Some of these were:

- UGC SAP Seminar on Caste Organisations in South Indian States- Social and Economic Dimensions, Department of Politics and Public Administration, University of Madras, 8-9, January 2001,
- Global Conference against Racism and Caste based Discrimination: Occupation and Descent-based Discrimination against Dalits 1-4 March 2001.
- National Seminar on Dravidian Folk and Tribal Lore 29, 30th Nov. 2002,
- National Seminar on "Dalit Studies and Higher Education : Exploring Content Material for a New Discipline", India International Centre, New Delhi, February 28-March 1, 2004,
- National seminar on The Situation of Dalits in Bangladesh: Possible Way Forward on 6-7 September 2007,
- DEV Seminar on "The Creation of Dalit Bourgeoisies: Caste, Credit and Markets in Urban India" on Wednesday 10 October 2007,
- The Challenge of Caste System in Christianity November 2007,
- National Seminar on Emerging Trends and Issues in Reservation Policy, February 2008,
- Seminar on Caste and Conflict in Uttar Pradesh, 1901-1931: Lessons for the policy of caste-based reservations 24-09-2008,

The author is herself belonging to CBM. From her personal experience she has found that information of such events is shared only within a close network consisting of people belonging to CBM. Information of such conferences is generally not disseminated amongst scholars not belonging to CBM. Such restrictive promotion of events makes it difficult for libraries to trace the proceedings. That might be the reason for poor representation of conference proceedings in the collection.

Chart 2 shows the holdings of different types of GL on CBM available across all seven libraries. It shows that majority of the report collection was available at IIPS Library (31%), MU Library (26%), IGIDR Library (15%), and NNCSW Library (13%) as compared to other libraries which had 6% or less of such documents.

Chart 2: Different types of GL on CBM across all seven libraries



Thesis and Dissertations collection of CBM in IIPS Library and IGIDR Library was mere 4% and 5% respectively as compared to MU Library (33%), NNCSW Library (29%), SNDT Library (17%) and TISS Library (13%).

IIPS Library had a large collection of Government publication consisting 43% of total Government publication on CBM compared to other libraries. IIPS Library had acquired almost all the Government Publications on CBM whereas most of other libraries had less than 10% of total Government Publication on CBM.

In case of Working Papers, only NNCSW Library, IIPS Library, and IGIDR Library had Working Papers, which were almost 30-35 percent. Other libraries did not have any Working Papers on CBM. Working Papers and Conference Proceedings were not very commonly available in most of the libraries surveyed.

Bibliographies on CBM were very few i.e. only 5 bibliographies. Bibliographies of GL on CBM covering individual library's collection would be ideal and useful for the researchers as well as for management. Also control of GL would be possible through bibliographies. Among the 5 bibliographies on CBM, four bibliographies were actually multiple copies of the same document. That was published in 1980s by 'The Mumbai Marathi Grantha Sangralaya', which is a Public Library of Mumbai City. Further, that bibliography was not updated. The remaining one bibliography that was published in 1990s was by SNDT Library covering its own collection. That was also not updated. The efforts from individual libraries to prepare and update the bibliographies were lacking.

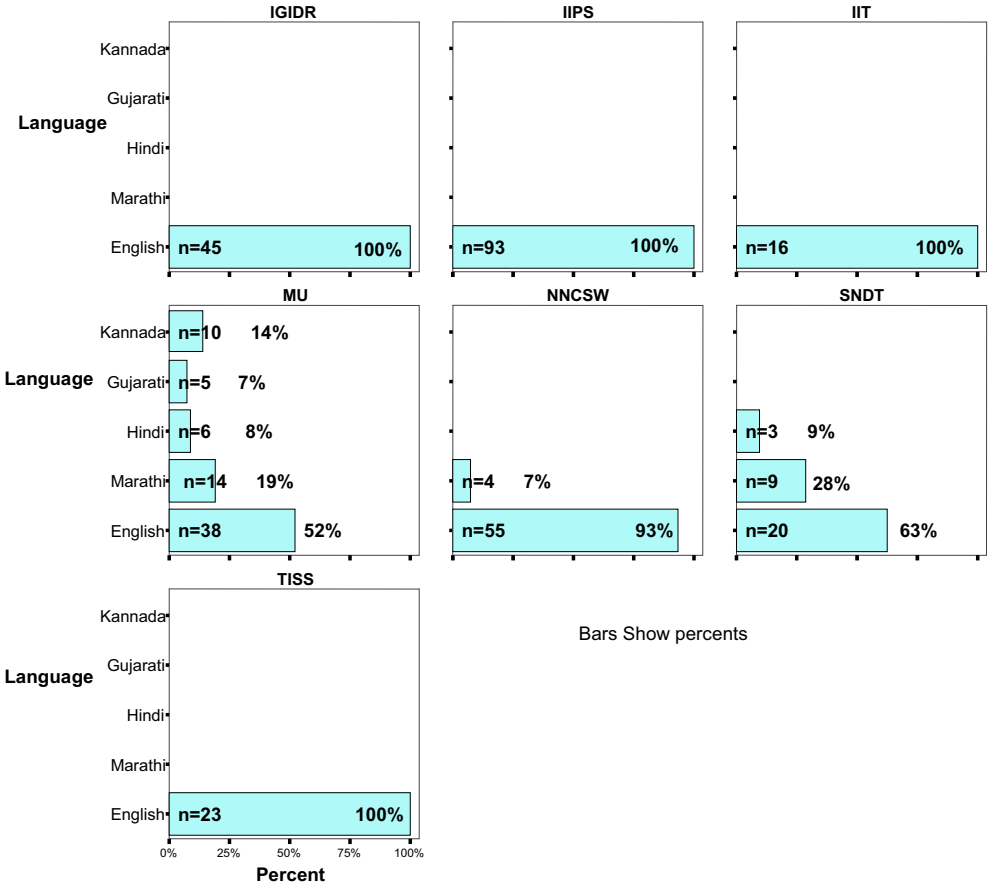
In total, there were 290 documents in English (85%), 27 documents in Marathi (7.9%), nine documents in Hindi (2.6%). There were five documents in Gujarati (1.5%) and ten documents in Kannada (2.9%) available in libraries of Mumbai.

In Thesis and Dissertations, there were documents in Hindi and Marathi in SNDT Library and MU Library. Additionally, MU Library had Thesis and Dissertations in Kannada. Chart 3 shows that all the libraries had GL on CBM mainly in English.

In GL 15% of publication was in regional languages. These GL in regional languages get unnoticed at national or international level. Majority of the publications were in English. This becomes a barrier to CBM students and researchers as many of them get educated through regional languages.

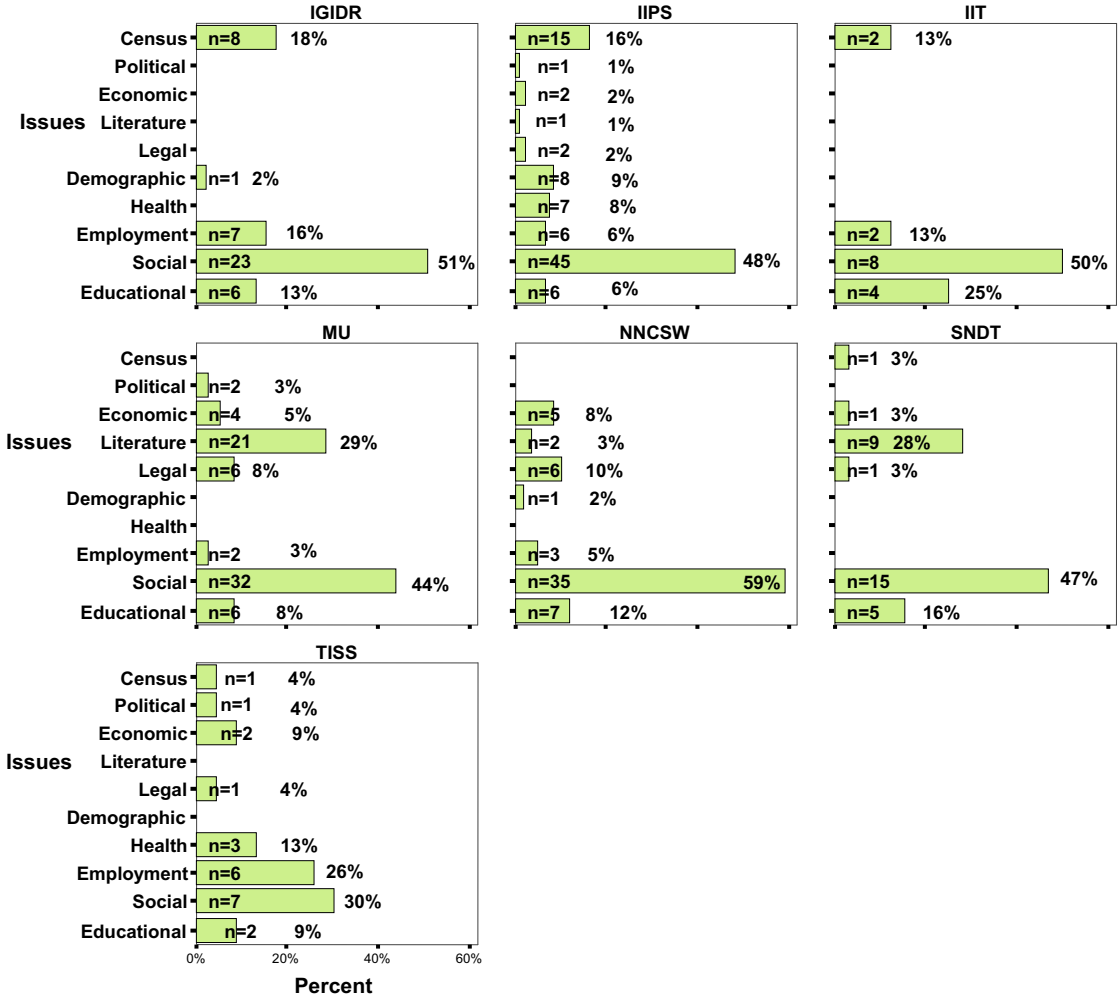
In some of the libraries visited, the software currently being used cannot handle regional languages. Therefore, separate catalogue was maintained for regional language material.

Chart 3: Library wise holdings of GL on CBM published in different languages



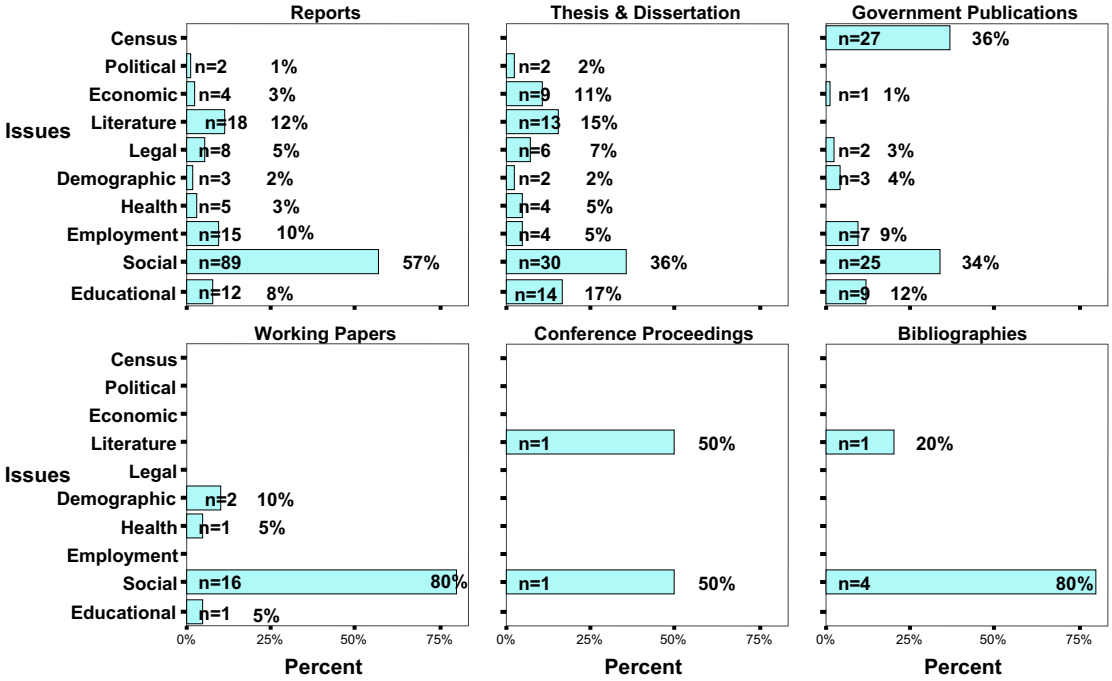
When the GL Collection was analysed as shown in Chart 4, as per subject content, it was found that social status of CBM in the society was the key issue of GL on CBM. Education, Employment and Literature issues were found to be at the second level. Publication on Political, Economic, Legal, Health issues were least represented i.e. less than 10% of the GL on CBM. Most probably studies on this issue are not being done and therefore not available to the libraries. Most of the studies revolve around status as in rural India; this is still a major concern leading to social unrest.

Chart 4: Library wise holdings of GL covering different CBM issues



The subject content was analysed in every document format. As represented in Chart 5, the status as an issue covering social discrimination, standard of living, and progress of the CBM, their interaction with the society, social policy, and government and NGOs efforts, schemes for upliftment of the CBM formed the most important area of discussion. Literature, Educational issues, and Employment issues came after social status.

Chart 5: Various CBM issues covered in different types of GL



In all types of GL on CBM, Educational issues like reservation in admissions, scholarship and freeship schemes, educational attainment of backward class students, etc. were found commonly. In Government Publication Census data was a major area of publication. Employment issues were found in Reports and Thesis and Dissertations as well as in Government Publications. In Thesis and Dissertations and Reports, Literature was the other issue discussed. Economics, Politics, Law, Demography, and Health were rarely discussed in the publication. So the findings on holdings of the library as represented in Chart 4 as well as the subject content of the different document formats as presented in Chart 5 support and reinforce each other.

All 341 documents were analysed longitudinally across 60 years publications covering the years of independence, it was found that in the first 25 years of post independence, the CBM were going through the phase in which they were searching for their identity in society, grooming themselves to acquire the basic education and employment. The retrieved literature reflected the studies done during 1947-1972 covered social status, Education and general issues of CBM.

The studies published after 1972 till 1997 continued with Social and Educational issues along with Demographic, Literature, and Employment issues. Very few studies were found on Legal, Economic, Political and Health aspects of CBM. Government of India also started publishing Reports, Statistical data, etc. to review the status of CBM after independence.

It is only in the last decade, that publications covering topics like Health, Culture, Economical status, Demography, Religious orientation, Political status, and Housing started becoming available. In this study, 54 documents i.e. 16% of GL covered these topics. This shows some shift from the basic concern of Social status, Education and Employment to more mainstream concerns.

Selection and Acquisition

Publication about CBM by CBM has become a potent tool to reflect the conditions of CBM. Most of these have been published by NGOs, Societies, trusts, etc. who publish literature by CBM about CBM. Therefore these do not get included in the national or trade bibliographies, which make it difficult for libraries to locate and acquire.

The libraries procured Government Reports directly from Government offices. Other GL were acquired from funding agencies, publishers, authors, etc.

It was commented by the librarians that copies of GL directly from author or funding agency were hardly received by the libraries. The most difficult to get were the conference proceedings. Libraries rarely received the brochures or circulars of various conferences on CBM issues. Thus libraries were facing

major difficulty in finding the information about such conferences to follow up further to acquire the proceedings.

IGIDR Library, IIPS Library and NNCSW Library have gathered the grey documents on CBM with the help of their students. Every year students from villages and remote area enroll for the courses. These students of the current year courses are told to enquire and collect the grey document on CBM available from their local areas, especially, when they visit to their villages during vacations.

Technical Processing

In all the libraries, the GL on CBM issues were classified and catalogued. Commonly all libraries have used Dewey Decimal Classification Scheme to classify the documents except IIT Library, which uses the Universal Decimal Classification Scheme to classify the GL on CBM. All GL on CBM were catalogued as per AACR II (Anglo- American Cataloguing Rules-II) standards. Bibliographic entries of GL in regional languages were added in transliterated form. The documents were assigned descriptors in English. Both generic and specific terms were used. For e.g. Scheduled Tribes- a generic term and "Koli" (Fisherman) - a specific term for a sub section of Scheduled Tribes.

The access to the bibliographic data of GL on CBM differed from library to library. SNDT Library, MU Library, NNSW Library had offline computerized catalogue. Whereas IIT Library, IIPS Library, IGIDR Library, TISS Library had Web OPAC. IIT Library and IIPS Library claimed that they have few Hindi documents on CBM, but these were not retrieved through their Web-OPAC

In some libraries separate databases of Thesis and Dissertations and Working Papers Government Publication, etc. were maintained. Thus the access to the bibliographic data was non-federated. Separate search was required under each type of documents (i.e. thesis and dissertations, working papers, seminar papers, census publications). In some libraries various research reports and other publications (i.e. books published by NGOs, institutes, and Government, bibliographies, etc.) were part of general book collection. Therefore each retrieved entry from book collection had to be checked by its bibliographic details to identify the GL on CBM.

Print copies of GL on CBM issues were arranged in classified order in all libraries except IGIDR library. This library had arranged the grey documents according to its accession number. Commonly in all the libraries, the collection of Thesis and Dissertations were kept separately which was then arranged accession number wise. MU Library did not bind the grey documents for greater longevity.

In IGIDR Library majority of the GL on CBM were stored as full text in electronic format either offline or online. The library has developed Kautilya Digital Repository, which includes Thesis & Dissertations, Working Papers and IGIDR's publications. IIPS library has developed a separate bibliographic level database (in devnagari script) for grey document in Hindi. It can be accessed through English language query from OPAC.

Users and Use

GL on CBM were generally referred by the PG students, Researchers, faculty members, management authorities, policy makers etc. The usage of the GL on CBM issues differed from libraries to libraries. Regular usage of GL on CBM was found at IIPS Library, IGIDR Library, and NNCSW Library. In SNDT Library, MU Library, and TISS Library, GL on CBM were circulated sometimes; whereas at IIT Library, these were rarely used. Except Thesis and Dissertations, other grey documents were available for home reading as well as on ILL. However in IIT Library, all grey documents were restricted to current reading only.

Libraries have taken efforts to maximise circulation of GL on CBM. When users have queries on issues pertaining to general population, the libraries do make special efforts to include pertinent GL on CBM to mainstream the GL collection. These documents were referred regularly in literature search, preparation of bibliographies, reference and information services. To further research, NNCSW's Librarian had suggested to the student to select their project/ thesis topics which would follow up the earlier work done by other researchers. This helped to judge the success of implementation of the policies and programmes of government, NGOs, etc. Few of these studies became useful as evidence in some legal and governmental matters.

In conclusion, the GL collection on CBM available in libraries of Mumbai was unique There was less than one percent duplication of the documents on CBM available in these seven libraries.

As GL on CBM was generally produced by CBM and used by CBM, special efforts should be made to mainstream their issues by being more open and inclusive for their conferences and seminar, etc.

Publications of trusts, etc. have to be better publicized and made available for library to be able to develop their collection.

There must be effort to develop bibliographic control such as union catalogue, subject specific repository, etc.

GL published in regional languages need to be given more attention as it contains discussions at grass root level.

Internet services and facilities need to be utilized by author as well as by libraries to share the e-GL on CBM to wider community. This would help to focus the attention of the world to CBM issues which in turn may help to empowerment of CBM in the coming decades

References

1. Caste system in India. In *Wikipedia.* Retrieved October 16, 2008 from http://en.wikipedia.org/wiki/Indian_caste_system
2. *Harrod's Librarians' Glossary and Reference Book* (2000). 9th ed. England: Grower Publishing Company. pp329.
3. Melliyal, Annamalai (2002). Dalits rights and issues. *India Together*. Retrieved August 23, 2008 from http://www.indiatogether.org/dalit/articles/intro.htm
4. Narula, Smita (n.d.) Caste discrimination. Retrieved September 10, 2008 from http://www.india-seminar.com/2001/508/508%20smita%20narula.htm
5. Rajendiran, P. (2006, March). Electronic grey literature in accelerator science and its allied subjects : selected web resources for scientists and engineers. *HEP Libraries Webzine* ;12. Retrieved September 10, 2008 from http://library.cern.ch/HEPLW/12/papers/4/

Polish Technologies On-line

Maciej Dominiak, Krzysztof Lipiec, Krystyna Siwek, and Maciej Ossowski,
Information Processing Centre, Poland

Abstract
In February 2008 the internet service POLSKIETECHNOLOGIE.pl was opened by Information Processing Centre in Warsaw (OPI). The strategic objective was to improve access to technologies offered to the Polish small land medium enterprises by research organizations. The below article presents the portal's principles of working and observations after the first year of its functioning.

OPI experience in scientific information and innovation activities

The Information Processing Centre (OPI) is an independent governmental public research and development organisation under auspices of the Ministry of Science and Higher Education in Poland aimed to gather and supply information on the Polish science and technology. Scientific information and innovation is the long-term statutory obligation of the Information Processing Centre. It includes such issues as:

- analysis and statistics concerning Polish science
- scientific and technical information development
- information services and databases on science and technology
- transfer of the results of research from science to industry and implementation of novel technologies
- science promotion and popularisation

OPI is a publisher of databases and printed catalogues. The most important information collections produced and maintained in the OPI cover databases and publications on science and technology.

- database on Current and Completed Research and Development Projects in Poland SYNABA
- database "Scientific and Research Units" comprising information accumulated for the Ministry of Scientific Research and Information Technology by means of the "Unit Questionnaire"
- Research Projects financed by the Ministry of Scientific Research and Higher Education relying on the database owned by the Ministry
- Research Organizations in Poland database
- Who is Who in the Polish Science database

OPI information services and databases are installed on http://www.opi.org.pl server.
The Information Processing Centre is preparing reports concerning Polish science, technologies and innovation potential for Polish government (mainly for Ministry of Scientific Research and Information Technology), but also for public and private companies.
Since the 90s OPI is becoming more and more active in participation in national and international programmes and initiatives. In this way Information Processing Centre has gain an experience participating in international and nationwide projects, especially in the field of innovation and entrepreneurship for SMEs.
The methodology implemented in creation and development of the new internet service POLSKIETECHNOLOGIE.pl benefited mainly from experience of former initiatives as:

- cooperation of OPI with the UNIDO (United Nations Industrial Development Organisation) office in Warsaw (UNIDO ITMO) in development of technology offers and technology requests database as a part of its studies concerning Polish innovation potential
- cooperation with the former governmental Agency of Technology Transfer (ATT) in development of the database of the Polish innovative products and technologies
- the Innovation Relay Centres network (IRC) 5.FP RTD UE (2000 -2004) and 6.FP RTD UE (2004-2008), as OPI used to be the co-ordinating organisation of the Innovation Relay Centre East Poland (5.FP RTD UE) and Innovation Relay Centre of Central Poland (6.FP RTD UE).
- In 1997-98 OPI participated in SCI-TECH PHARE I project on "Creation and Development of Innovation Transfer Centre" and in SCI-TECH PHARE II project "Nationwide and regional information services for SMEs – REGION EAST" in 2000
- the Polish Network of Technology Transfer and Innovation support for SMEs – STIM co-funded by the Sectoral operational Programme 'Growth of enterprise competetiveness" under Action" support for development of business advisory organizations and business Advisory networks", overlooked by the Polish Agency for Entrepreneurship development (PARP) (ERDF Objective 1, from 2005 to 2006)

Internet service POLSKIETECHNOLOGIE.pl

In February 2008 the internet service POLSKIETECHNOLOGIE.pl was opened by Information Processing Centre in Warsaw (OPI) with financial support from the public sector, namely from the Ministry of Science and Higher Education in Poland, dedicated to RTD development.

The service came into existence after one year of preparatory phase, focused mainly on developing and testing tools for specialized technology related services, gathering data (like searching for enterprises interested in implementing new technologies), gathering the technological offers (about 200 for present moment), conducting conversations with numerous experts from Poland as well as from abroad, activities connected with projecting a software and webmaster's tools. The aim of the service is to present new technological solutions made by Polish business and the scientific organizations.

Target groups of the service are:
- Small and medium enterprises
- Research and development organizations
- Industry
- Ministry of Science and Higher Education

This portal is an easily accessible source of information directed not only to researchers and businessmen who offer technological solutions and information about possibility of development of one's own business, but also to those who search for such information. The goal of the service is to facilitate communication between scientists and businessmen and to inspire the scientific environment to more commercial utilization.

The main idea was to create the portal which offers information in easy to understand way, the portal which follows the most important science events in Poland and give information about advanced technology. All information on the website are presented by an easy to follow mechanism, built in support of labour-saving tools.

Fig.1. POLSKIETECHNOLOGIE.pl portal

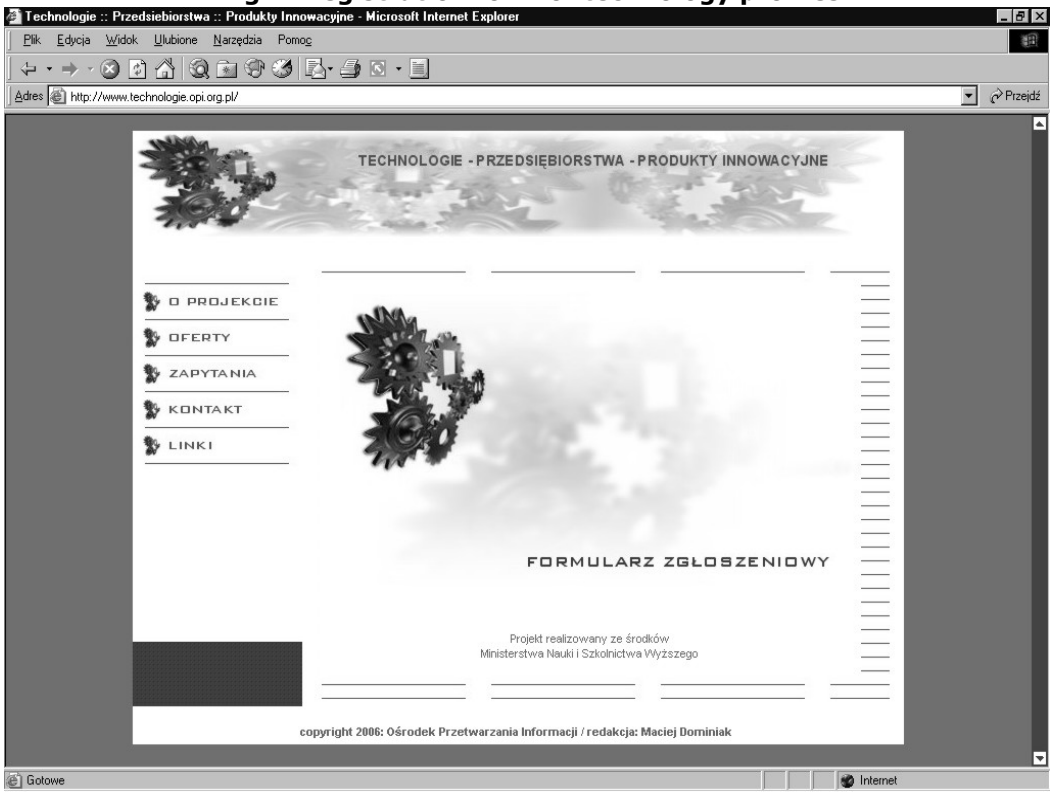
Moreover, POLSKIETECHNOLOGIE.PL contains the large number of thematic sections (e.g. the newest issues concerning research financing and intellectual property rights) as well as the information about events and valid legal regulations.

There are professional articles, invitations for events and information relating to Polish science and industry published in the Portal. All information on website are presented by an easy to follow mechanism, built in support of labour-saving tools (own search engine, expanded database, current events calendar, thematic archives).

Database of technology profiles

An innovative technology is a technology which has been developed recently and which offers significant alternative value to the already existing available technologies. One of the main parts of the service in interactive database of technological offers addressed to representatives of business and industry.

Fig 2. Registration form of technology profiles



A Technology Request description contain:

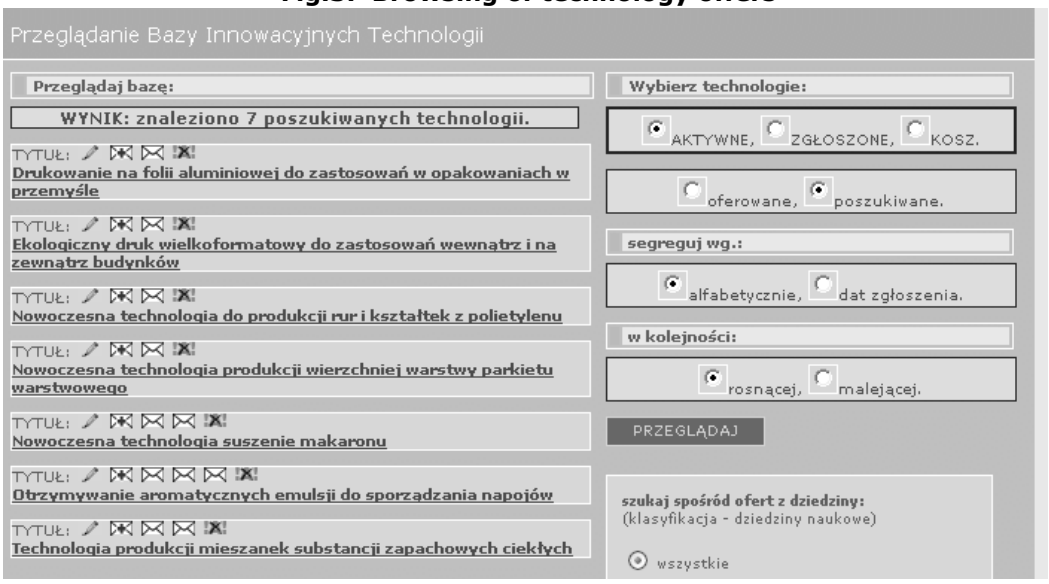
- A description of the technology process
- Details of the problems and needs
- Benefits sought
- What the company is seeking from potential partners
- Classification and keywords

Main elements of technology offer contain:

- A description of the technology or process
- Innovative characteristics
- Possible benefits
- Potential partnership
- The type of agreement
- IPR status
- Classification and keywords

The service is nationwide and it is not specialised to specific sectors, it is opened to all branches. It does not focus on any predefined technology sectors. The example of the Mazovian region comprising remote and less developed rural areas with such branches as the agriculture and food processing, ecology and environment, forestry and wood and furniture industry, tourism from one side and the Warsaw metropolitan area represented by high-tech branches, namely electronics and material science and engineering and information technology from other side illustrates broad spectrum of needs and expectations of potential clients.

Fig.3. Browsing of technology offers

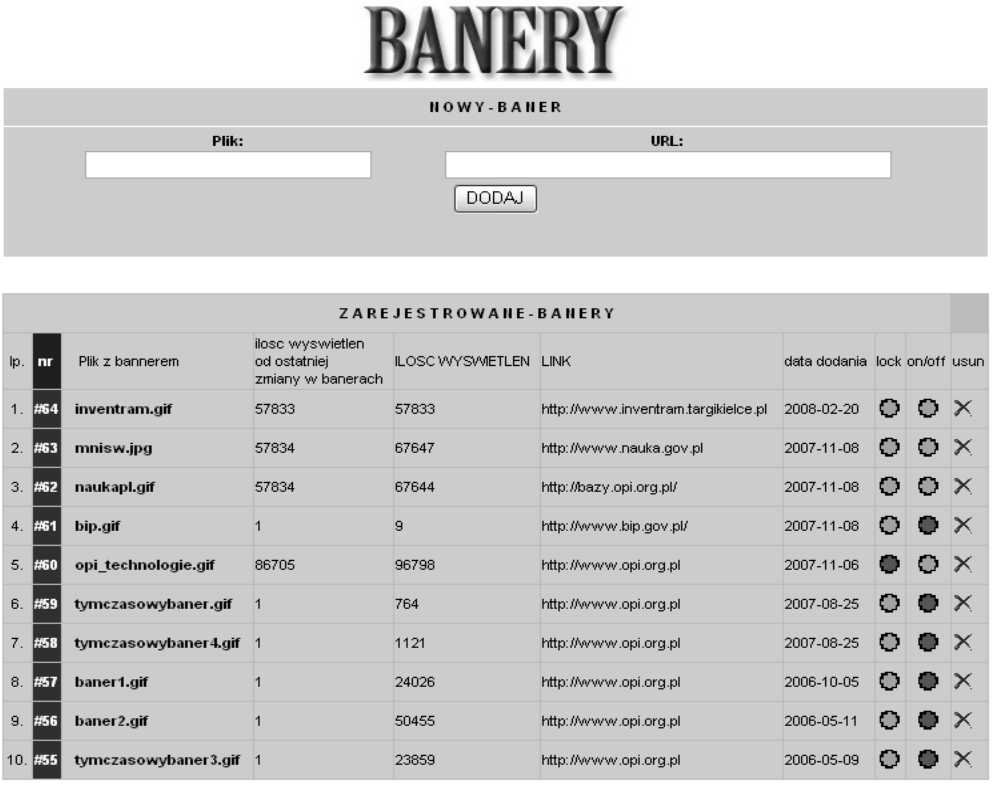


Database can be browsed in several orders: chronology, subject classification etc. Technology offer and request items can be registered and modified by their authors after logging**.** A range of options is provided by User's Menu.

Plans for future

Although most of services are standard – delivered by several technology transfer initiatives on regional or even transnational scale, but some of them are "project related" and they face common issue of continuation after project implementation. development of portal and the database. The POLSKIETECHNOLOGIE.PL is dependent on public funds as well, but the service consider commercialization of some of the services. The example is the management system of advertisement, now in development stage.

Fig. 4. Management system of advertisement



lp.	nr	Plik z bannerem	ilosc wyswietlen od ostatniej zmiany w banerach	ILOSC WYSWIETLEN	LINK	data dodania	lock	on/off	usun
1.	#64	inventram.gif	57833	57833	http://www.inventram.targikielce.pl	2008-02-20			
2.	#63	mnisw.jpg	57834	67647	http://www.nauka.gov.pl	2007-11-08			
3.	#62	naukapl.gif	57834	67644	http://bazy.opi.org.pl/	2007-11-08			
4.	#61	bip.gif	1	9	http://www.bip.gov.pl/	2007-11-08			
5.	#60	opi_technologie.gif	86705	96798	http://www.opi.org.pl	2007-11-06			
6.	#59	tymczasowybaner.gif	1	764	http://www.opi.org.pl	2007-08-25			
7.	#58	tymczasowybaner4.gif	1	1121	http://www.opi.org.pl	2007-08-25			
8.	#57	baner1.gif	1	24026	http://www.opi.org.pl	2006-10-05			
9.	#56	baner2.gif	1	50455	http://www.opi.org.pl	2006-05-11			
10.	#55	tymczasowybaner3.gif	1	23859	http://www.opi.org.pl	2006-05-09			

More focus will be put on awareness actions. Dissemination will be carried via relevant events, webpage, press releases, where information about services will be available. In 2009 an English version of the portal will be implemented.

Nancy STYLE

a tool to improve the production of Grey Literature

What is Nancy Style?

It is the informal name given to the

Guidelines for the production of scientific and technical reports: How to write and distribute Grey Literature

formally presented by the Istituto Superiore di Sanità (Rome, Italy) during the 7th International Conference on Grey Literature held in Nancy (France) in December 2005.

Who can use this tool?

Authors and GL producers in their mutual task of creating and distributing accurate, clear, easily accessible reports in different fields.

Which goal?

Permit an independent and correct production of institutional reports in the respect of the basic editorial principles.

Which language?

The original version is in English. Its translations are in:

- Italian (by Istituto Superiore di Sanità)
- French (by Institut de l'Information Scientifique et Technique)
- German (by Technische InformationsBibliothek/UniversitätsBibliothek)
- Spanish (by Universidad de Salamanca, *in preparation*)

Translations in other languages are most welcome.

Where can you get it?

All the versions are available from the official site of the GLISC: www.glisc.info.

What is it about?

- Ethical principles related to the process of evaluating, improving, and making available reports, and the relationships between GL producers and authors.
- Technical aspects of preparing and submitting reports.

What is GLISC?

The group approving these Guidelines is formally defined as Grey Literature International Steering Committee, composed of:

Istituto Superiore di Sanità (ISS), Italy

Institut de l'Information Scientifique et Technique (INIST-CNRS), France

Grey Literature Network Service (GreyNet), The Netherlands

For information:
www.glisc.info
secretariat@glisc.info

RELATED BY
Paola De Castro and Sandra Salinetti
(Istituto Superiore di Sanità, Rome – Italy) 2008

Appendices

Author Information

Asserson, Anne **108**
Anne Asserson holds a Cand. polit. with a Masters in Information Science from the University of Bergen, UiB. She has been working with Research Documentation, and has participated in substantial parts of CRIS developmental work, locally and nationally. Anne Asserson has been part of the establishing and implementing of a Research Documentation system, Fdok http://www.ub.uib.no/fdok/sok/, at the UiB. For several years she was the chairwoman of the Steering Group of the national CRIS system and project secretary of a National system for academic administration. Anne Asserson is presently representing UiB in the national group that is implementing a new national research documentation system, FRIDA. She has also participated in The CORDIS funded European-wide project on " Best Practice" 1996 . She was a member of the working group set up 1997 that produced the report CERIF2000 Guidelines (1999) www.cordis.lu/cerif, coordinated by the DGXIII-D4. euroCRIS is now the custodian of the CERIF model www.eurocris.org. Anne Asserson is a member of the Best Practice Task Group.
Email: anne.asserson@fa.uib.no

Cignoni, Laura **93**
Laura Cignoni, a former British School teacher, has been working at the Institute for Computational Linguistics in Pisa of the National Research Council since 1981. Her interests and activity range from studies in comparative lexicology and lexicography, with particular regard to multiword expressions in English and Italian, to philology and its related disciplines, to the creation of computer tools for children's dictionaries. She has participated in many national and international projects including the recent ongoing Medici Project in Florence. She has edited numerous publications, in particular the journal "Linguistica Computazionale", publication of the Institute for Computational Linguistics.
Email: laura.cignoni@ilc.cnr.it

Crowe, June **82**
June Crowe is the Senior Researcher at Information International Associates, Inc. (IIa). She received her AMLS from the University of Michigan, Ann Arbor and her M.Ed. in geographic education from the University of Georgia, Athens. She has extensive experience in the management and operations of library services across government, public, academic, and special libraries. At IIa she performs open source research on a variety of topics and manages research projects involving web-based open source document identification, collection, and processing. Her primary interests are open source information in Grey Literature, repositories, and open source intelligence tools.
Email: jcrowe@iiaweb.com

Davidson, Thomas S. **82**
Thomas S. Davidson II was a Senior Military Intelligence Analyst at the Foreign Military Studies Office (FMSO) of the U.S. Army Training and Doctrine Command at Ft. Leavenworth, Kansas. He founded and led the FMSO Mexico and Southwest Border Security Team for exploitation of primary language Latin American open source material for items of interest to U.S. National and Border Security. His training includes coursework at the Defense Language Institute, Monterey, California, for Vietnamese, German, Korean, and Czech languages, as well as other military development courses, to include the Warrant Officer Advanced course in 1993. Davidson owned and operated a language institute, Languages of El Paso, which provided language services in five primary languages to the "maquiladora" industry in Cd. Juárez, Cd. Chihuahua, and Nogales, Sonora. In 1997 he supported Bechtel Corporation on its PEMEX contract as a human resource manager and liaison to Mexican border and customs agencies. CWO4 Davidson's awards include the Legion of Merit, the Meritorious Service Medal, the Joint Service Commendation Medal, two Army Commendation Medals, the Army Achievement Medal, and the German Army Marksmanship Award.
Email: thomas.davidson@us.army.mil

Di Cesare, Rosa **27**
Rosa Di Cesare was born in Civita d'Antino (AQ) in 1952 and graduated from "La Sapienza" University in Rome in 1982. She received her diploma in Librarianship from the Vatican Library in 1996. She worked in the Central Library of National research council (CNR), where she started to become involved in research activity in the field of Grey literature (GL). Member of the Technical Committee for the SIGLE database from 1995 to 2001, she is presently responsible for the Library at the Institute of research on population and social policies (IRPPS) of the National research council. Her studies have focused on citation analysis and on the use of GL in scientific publications.
Email: biblio.irpps@irpps.cnr.it

Farace, Dominic **118**
Dominic J. Farace is Director of TextRelease, an Amsterdam based information bureau specializing in grey literature and networked information. He is a native Louisianan and holds two degrees in sociology from Creighton University (BA) and the University of New Orleans (MA). His doctoral dissertation in social sciences is from the University of Utrecht, The Netherlands, where he has lived and worked for the past twenty-seven years. After six years heading the Department of Documentary Information at the Royal Netherlands Academy of Arts and Sciences (Swidoc/KNAW), he founded GreyNet, Grey Literature Network Service, in 1993 and has since been responsible for the international GL-Conference Series. In this capacity, he serves as Program and Conference Director as well as managing editor of the conference proceedings. Since 2004, he is a Guest Lecturer on Grey Literature in the Masters Program at the University of Amsterdam; Instructor of Grey Literature via UNO Distance Education, and Editor of TGJ, The Grey Journal. Email: dominic.farace@textrelease.com

Author Information
(continued)

Frantzen, Jerry **118**
Jerry Frantzen graduated in 1999 from the College of Amsterdam in Library and Information Science. Frantzen is the technical editor of The Grey Journal (TGJ). And, since 1996, he is affiliated with GreyNet, Grey Literature Network Service, as a freelance technical consultant. Email: info@greynet.org

Henrot, Nathalie **118**
Nathalie Henrot graduated in History, then in Information Sciences from the University of Tours in 1988. She has been working for the INIST-CNRS for seventeen years, more specifically at the Monographs & Grey Literature Section from 1993, for congress proceedings acquisition. She is now the user administrator in the OpenSIGLE project.
Email: henrotn@inist.fr

Hitson, Brian **11**
Brian Hitson is Associate Director for the U.S. Department of Energy's Office of Scientific and Technical Information (OSTI) in Oak Ridge, Tennessee. In this position, Mr. Hitson is responsible for international information exchange programs, administrative and financial management, cost-reimbursable activities, limited access information programs, and the digitization and preservation of a 1.2 million scientific document repository. As part of his international responsibilities, Mr. Hitson coordinated the development of the global science gateway, WorldWideScience.org and the establishment of its multilateral governance structure, the WorldWideScience Alliance. He is Chairman and U.S. representative to the International Energy Agency's Energy Technology Data Exchange (ETDE), which manages the world's largest energy research, technology, and development database. He is also the U.S. representative to the International Atomic Energy Agency's International Nuclear Information System (INIS). In addition, he serves on the elected Bureau of the International Council for Scientific and Technical Information (ICSTI) as Chair of the Technical Activities Coordinating Committee. Mr. Hitson has a Bachelor of Arts degree in Economics and a Master's in Business Administration, both from the University of Tennessee. Email: hitsonb@osti.gov

Jeffery, Keith G. **108**
Keith Jeffery is currently Director, IT and International Strategy of STFC (Science and Technology Facilities Council), based at Rutherford Appleton Laboratory in UK. Previously he was Head of Business and Information Technology Department with a staff of 140 supporting over 360000 users, developing software for business and science and doing leading edge R&D. STFC hosts the UK and Ireland Office of W3C and develops and supports the largest OA (Open Access) institutional repository in UK. Keith is a Fellow of both the Geological Society of London and the British Computer Society. He is a Chartered Engineer. He is an Honorary Fellow of the Irish Computer Society. He is president of euroCRIS (www.eurocris.org) and of ERCIM (www.ercim.org) and holds three honorary professorships. He has extensive publications and has served on numerous programme committees and research grant review panels. He has particular interests in 'the research process' and the relationship of hypotheses, experiments, primary data and publications based on research in information systems, knowledge-based systems and metadata. Email: k.g.jeffery@rl.ac.uk

Južnic, Primož **101**
Primož Južnič is an associate professor at the Department of Library and Information Science and Book Studies at Faculty of Arts, University of Ljubljana (Slovenia). His main area of research and interest is bibliometrics, collection management and LIS education. He teaches the following courses: Bibliometrics, Special libraries, and Collection Management. Before starting his university career, he was a heading different special and academic libraries and information/computer centres. He was also working at the European Commission, for three years, as the seconded informatics expert.
Email primoz.juznic@ff.uni-lj.si

Lin, Yongtao **55**
Yongtao Lin has worked as a librarian in medical settings since 2004 and is currently a Health Information Network Librarian at Tom Baker Cancer Knowledge Centre, Calgary where she provides library services to clinicians, staff, patients and families. She is a graduate from School of Information Management from Dalhousie University, also has a Bachelor of Education and has had years of teaching experience. Her interests centre on information literacy and library instruction to support healthcare practitioners with best evidence for their healthcare decision-making.
Email: yolin@ucalgary.ca

Lipinski, Tomas A. **67**
Tomas Lipinski obtained his J.D. from Marquette University Law School, LL.M. from The John Marshall Law School, and Ph.D. from the University of Illinois at Urbana-Champaign. Professor Lipinski has worked in a variety of library and legal settings including the private, public and non-profit sectors. Professor Lipinski teaches researches and speaks frequently on various topics within the areas of information law and policy, especially copyright, free speech and privacy issues in schools and libraries. In fall of 2005, Professor Lipinski was placed on the Fulbright Senior Specialist Roster and was named a member of the Global Law Faculty, University of Leuven in Fall of 2006. Email: lipinski@sois.uwm.edu

Luzi, Daniela **27**
Daniela Luzi is researcher of the National Research Council at the Institute of research on populations and social politics. Her interest in Grey Literature started at the Italian national reference centre for SIGLE at the beginning of her career and continued carrying out research on GL databases, electronic information and open archives. She has always attended the International GL conferences and in 2000 she obtained an award for outstanding achievement in the field of grey literature by the Literati Club. Email: d.luzi@irpps.cnr.it

Author Information
(continued)

Pardelli, Gabriella **93**
Gabriella Pardelli graduated in Letters at the University of Pisa in 1980 and has been working at the Institute for Computational Linguistics of the National Research Council in Pisa since 1984. She has been active in the creation of bibliographical databases for Natural Language Processing (NLP) and Digital Library in the Humanities. Other interests regard terminology and History of Human Language Technology. She is responsible for the Library of the Institute of Computational Linguistics and the collection called "Antonio Zampolli Fund". She has participated in many research projects and has worked on the creation of bibliographical resources in the field of language technologies. She has presented many works at different national and international conferences and congresses.
Email: gabriella.pardelli@ilc.cnr.it

Pejšová, Petra **21**
Petra Pejsova studied information science and librarianship at Charles University. She works as an information specialist in the State technical Library, Czech Republic. Actually she is leading a project Digital Library for Grey Literature – Functional model and pilot.
Email: p.pejsova@stk.cz

Pfeiferová, Martina **21**
Martina Pfeiferova has a degree in information science and librarianship at Charles University. She works as an information specialist in the State technical Library, Czech Republic. Actually she is working on a project Digital Library for Grey Literature – Functional model and pilot.
Email: m.pfeiferova@stk.cz

Rabina, Debbie L. **77**
Debbie Rabina is a Assistant Professor at Pratt Institute, School of Information and Library Science. Her areas of teaching and research include scholarly communication, LIS education, government and NGO information sources, and information policy.
Email: drabina@pratt.edu

Rudasill, Lynne Marie **61**
Lynne Marie Rudasill holds a Master's Degree in Library and Information Science from the University of Illinois, with additional studies in political science from Illinois State University. She currently serves as Global Studies Librarian and subject specialist in political science and speech communication at the University of Illinois where she is an Associate Professor of Library Administration. Her current research is in the area of information use and production by non-governmental organizations and the archiving of grey literature. She continues her work in web usability and accessibility as well.
Email: rudasill@illinois.edu

Sassi, Manuela **93**
Manuela Sassi. Graduated in Foreign Languages and Literature at Pisa University, 110/110 cum laude. Since 1974 she has been working in Pisa at the Institute for Computational Linguistics of the National Research Council. Her interests and experiences range from linguistic to textual data processing and in providing linguistic resources on-line. She has been responsible for many national projects and has participated in numerous international projects.
Email: manuela.sassi@ilc.cnr.it

Schöpfel, Joachim **39, 118**
Joachim Schöpfel obtained his Ph.D. in psychology from the Hamburg University in 1992. During his studies in psychology, he participated in research on bilingual children of Turkish immigrants in Hamburg, of the German minority in Denmark, and in a French-German High School in Versailles, France. From 1991 to 2008, he worked at the French Institute for Scientific and Technical Information (INIST-CNRS) in different positions in database production and library management, at last as head of the e-publishing and document supply department. During the same time, he was lecturer at the University of Nancy. At present, he is senior lecturer in information and communication sciences at the Charles de Gaulle University of Lille 3. He published on GL, document delivery, digital libraries, scientific publishing, usage statistics and professional development.
Email: joachim.schopfel@univ-lille3.fr

Stock, Christiane **39, 118**
Christiane Stock is the Head of the Monographs and Grey Literature service at INIST, in charge of the repositories LARA (reports), mémSIC (master's theses in information sciences) and OpenSIGLE. Member of the Technical Committee for the SIGLE database from 1993 to 2005, she also set up the national agency for ISRN (International Standard Report Number). She is member of the AFNOR expert group who prepared the recommended metadata scheme for French electronic theses (TEF).
Email: christiane.stock@inist.fr

Vaska, Marcus **55**
Marcus Vaska is a librarian at the University of Calgary's Health Sciences Library, where he provides research assistance and support to clients within the Faculty of Medicine and the Calgary Health Region. In addition to exploring innovative instruction techniques for his classes, he has developed newfound appreciation for the pursuit of literature that exists beyond traditional publishing channels. Marcus' current interests focus on the teaching methods and experiences of librarians involved in delivering grey literature searching sessions to academic audiences in a medical setting. Email: mmvaska@ucalgary.ca

List of Participating Organizations

American Veterinary Medical Association, AVMA	United States
Amnesty International	Netherlands
Boekman Foundation	Netherlands
British Library, BL	United Kingdom
Centre National de Recherche Scientifique, CNRS	France
Centre of Information Technologies and Systems, CITIS	Russia
City of Amsterdam	Netherlands
Department of Economy, Science and Innovation, EWI	Belgium
Department of Energy, DOE	United States
EBSCO Information Services	United States
euroCRIS, Current Research Information Systems	Netherlands
European Organization for Nuclear Research, CERN	Switzerland
Fabchannel B.V.	Netherlands
Federal Library and Information Center Committee, FLICC	United States
Federal Library Information Network, FedLink	United States
Grey Literature Network Service, GreyNet	Netherlands
Health-evidence.ca	Canada
Information International Associates, IIa	United States
Institut de l'Information Scientifique et Technique, INIST	France
Institute of Information Science and Technologies, ISTI-CNR	Italy
Institute of Research on Population and Social Policies, IRPPS	Italy
International Atomic Energy Agency, IAEA	Austria
Istituto di Linguistica Computazionale, ILC	Italy
Istituto per lo sviluppo della formazione professionale dei lavoratori, ISFOL	Italy
Japan Science and Technology Agency, JST	Japan
Kansas State University	United States
Korea Institute of Science & Technology Information, KISTI	Korea
McMaster University	Canada
Ministry of the Interior and Kingdom Relations	Netherlands
Ministry of Education, Culture and Science, OCW	Netherlands
Mnatobi Ltd.	Georgia
National Research Council, CNR	Italy
New York Academy of Medicine, NYAM	United States
Office of Scientific and Technical Information, OSTI	United States
Oklahoma State University	United States
Open Source Center	United States
Open Source Research Group	United States
Osrodek Przetwarzania Informacji, OPI	Poland
Pratt Institute, School of Information and Library Science	United States
PricewaterhouseCoopers, PwC	Netherlands
Purdue University	United States
Science and Technology Facilities Council, STFC	United Kingdom
Scientific and Technical Information Center, VNTIC	Russia
Slovak Centre of Scientific and Technical Information, CVTI SR	Slovakia
SNDT Women's University	India
Swets	Netherlands
Université Charles de Gaulle Lille 3	France
University of Amsterdam, UvA	Netherlands
State Technical Library, STK	Czech Republic
Texas A&M University	United States
The Hague University Library	Netherlands
Universidade Federal de Santa Catarina, UFSC	Brazil
University of Bergen, UiB	Norway
University of Calgary	Canada
University of California, Irvine Libraries, UCI	United States
University of Illinois, UIUC	United States
University of Ljubljana, UNI-LJ	Slovenia
University of Missouri-Columbia	United States
University of Patras	Greece
University of Utrecht, UU	Netherlands
University of Wisconsin, UWM	United States
Washington State University	United States

GreyNet members less 20%

Designing the Grey Grid for Information Society

Publication Order Form

TENTH INTERNATIONAL CONFERENCE ON GREY LITERATURE — *Amsterdam, 8-9 December 2008*

	No. of Copies	x	*Amount in Euros*	Subtotal
Forthcoming Publications (January 2009)				
GL10 CONFERENCE PROCEEDINGS – Printed Edition ISBN 978-90-77484-11-1 ISSN 1386-2316 *Postage and Handling* ***excluded***[*)]		x	85.00 =	€
GL10 CONFERENCE PROCEEDINGS - CD-Rom Edition ISBN 978-90-77484-11-1 ISSN 1386-2316 *Postage and Handling* ***included***		x	85.00 =	€
EXCLUSIVE OFFER: **GL10 Conference Proceedings** - CD-Rom Edition With accompanying PowerPoint Presentations *Postage and Handling* ***included***		x	99.00 =	€
POSTAGE AND HANDLING PER PRINTED COPY[*)]				
Holland		x	5.00	€
Europe		x	10.00	€
Other		x	20.00	€
			TOTAL	**€**

Customer information

Name:	
Organisation:	
Postal Address:	
City/Code/Country:	
E-mail Address:	

Check one of the boxes below for your Method of Payment:

❑ Direct transfer to TextRelease, Account No. 3135.85.342, Rabobank Amsterdam
BIC: RABONL2U **IBAN:** NL70 RABO 0313 5853 42, with reference to "GL10 Publication Order"

❑ Bank check/draft made payable to TextRelease, with reference to "GL10 Publication Order"

❑ MasterCard/Eurocard ❑ Visa card ❑ American Express

Card No. _____ Expiration Date: _____

Print the name that appears on the credit card, here _____

Signature: _____ **CVC II code:** _____ *(Last 3 digits on signature side of card)*

Place: _____ Date: _____

NOTE On receipt of payment, an invoice marked paid will be sent to your postal address. If you require a *pro forma* invoice, please notify:

TextRelease
www.textrelease.com

GL10 Program and Conference Bureau
Javastraat 194-HS, 1095 CP Amsterdam, Netherlands
T/F +31-(0) 20-331.2420 Email: info@textrelease.com

Index to Authors