ISSN 1574-1796

An International Journal on Grey Literature

Summer 2012 – TGJ Volume 8, Number 2

'Data Frontiers in Grey Literature'

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

The Grey Journal
An International Journal on Grey Literature

COLOPHON

CIP
The Grey Journal (TGJ): An international journal on grey literature / D.J. Farace (Journal Editor); J. Frantzen (Technical Editor) ; GreyNet, Grey Literature Network Service. - Amsterdam: TextRelease, Volume 8, Number 2, Summer 2012. - *EBSCO Publishing, FLICC-FEDLINK, INIST-CNRS, JST, NTK, and NYAM are corporate authors and associate members of GreyNet International. This serial publication appears three times a year - in spring, summer, and autumn. Each issue is thematic and deals with one or more related topics in the field of grey literature. The Grey Journal appears both in print and electronic formats.*
ISSN 1574-1796 (Print)
ISSN 1574-180X (E-Print/PDF)
ISSN 1574-9320 (CD-Rom)

Subscription Rates:
€125 individual, including postage & handling
€240 institutional, including postage & handling

Contact Address:
Reprint Service, Back Issues, Document Delivery, Advertising, Inserts, Subscriptions:

TextRelease
Javastraat 194-HS
1095 CP Amsterdam
Netherlands
T/F +31 (0) 20 331.2420
info@textrelease.com
http://www.textrelease.com/glpublications.html

About TGJ

The Grey Journal is a flagship journal for the international grey literature community. It crosses continents, disciplines, and sectors both public and private.
The Grey Journal not only deals with the topic of grey literature but is itself a document type classified as grey literature. It is akin to other grey serial publications, such as conference proceedings, reports, working papers, etc.

The Grey Journal is geared to Colleges and Schools of Library and Information Studies, as well as, information professionals, who produce, publish, process, manage, disseminate, and use grey literature e.g. researchers, editors, librarians, documentalists, archivists, journalists, intermediaries, etc.

About GreyNet

The Grey Literature Network Services was established in order to facilitate dialog, research, and communication between persons and organizations in the field of grey literature. GreyNet further seeks to identify and distribute information on and about grey literature in networked environments. Its main activities include the International Conference Series on Grey Literature, the creation and maintenance of web-based resources, a moderated Listserv, and The Grey Journal. GreyNet is also engaged in the development of distance learning courses for graduate and post-graduate students, as well as workshops and seminars for practitioners.

Full-Text License Agreement

In 2004, TextRelease entered into an electronic licensing relationship with EBSCO Publishing, the world's most prolific aggregator of full text journals, magazines and other sources. The full text of articles in The Grey Journal (TGJ) can be found in *Library, Information Science & Technology Abstracts* (LISTA) full-text database.

Contents
'DATA FRONTIERS IN GREY LITERATURE'

I search

I find

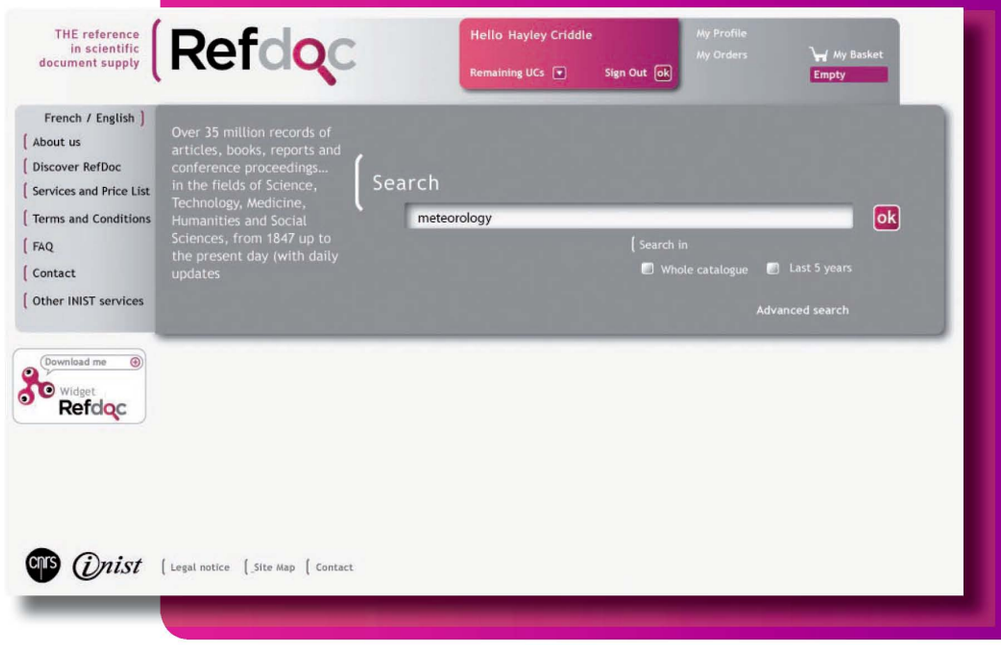
I order

January 2010 – all scientific and technical information available in just 3 CLICKS

Over 35 million records of articles, books, reports and conference proceedings from 1847 up to the present

Refdoc.fr

DISCOVER ALL THE FEATURES AND FUNCTIONS AVAILABLE AT www.refdoc.fr



bbcom ... 5677

cnrs inist diffusion

INIST DIFFUSION

2, ALLÉE DU PARC DE BRABOIS, 54514 VANDŒUVRE-LÈS-NANCY CEDEX, FRANCE → PHONE : +33 (0)3 83 50 46 64

EDITOR'S NOTE

Tracking and Backtracking Data

In 2011, GreyNet presented the first results of a two year project on Enhanced Publications (EPP). This year, the project focusses on the acquisition of research data, their cross-linking to existing full-text documents, and the establishment of a workflow for future publications. Enhanced publications combine and link research data to full-texts, other supplementary materials, as well as post-publication data.

Results in the first leg of the project indicate that sixty percent of the surveyed authors base their research on empirical and/or statistical data. And, two-thirds of those authors remarked that their data are still available for archiving purposes. These respondents also express a willingness to share their data and hold to the opinion that both the data producer as well as the prospective user would stand to benefit.

This second leg of the project rests on the approach taken in facilitating the acquisition process. The primary instrument used will be the OpenGrey Repository that houses GreyNet's collection of conference preprints. By backtracking to the existing metadata records in OpenGrey and by communicating directly with the authors of those records, another way will be opened for further cooperation between data producer and data provider. In addition, the subsequent cross-linking between OpenGrey and the DANS EASY Repository, where GreyNet's research data will be stored, stands to better serve the needs of the grey literature community in which open access to research data is a prerequisite.

Further in this summer issue, an overview of other initiatives undertaken by GreyNet over the past two decades is presented in a timeline.

Dominic Farace
journal@greynet.org

Enhancing diffusion of scientific contents: Open data in Repositories*

Daniela Luzi, Rosa Di Cesare, Marta Ricci, and Roberta Ruggieri (Italy)

Introduction

The free availability of data gathered during research activities is becoming one of the new challenges facing the Open Access Movement. New scientific instruments and technologies used in highly collaborative fields such as molecular biology, astronomy and environmental sciences, make it possible to collect a great amount of data in different formats. Moreover, data are often associated with tools that can aggregate them as well as with direct references to the publications – conventional or non-conventional – that report the results of their analysis. The benefits of the availability of these data are evident, and include assessment of research results, along with the reproduction and re-utilisation of data, potentially to draw new insight for future research.

According to the National Science Foundation: "digital data are the currency of the data collection universe, which, like currency in the financial realm, comes in many different forms". They are different in nature, generally depending on the very specific field of study; they are produced for different purposes using varying methods and/or instruments; they have their own lifecycle before they are "translated" into scientific results and diffused in scientific publications. Understanding all these aspects makes it possible to determine whether to preserve them and how, who is responsible for their curation and/or diffusion, what type of archive, or better infrastructure, should be developed. This in turn implies issues related with data ownership, as well as funding resources, types of institutions and services to be involved.

Several policy papers (NSF, 2005, OECD, 2007, US National Research Council, 1995; 1999) are advocating free access of datasets and are outlining recommendations to coordinate efforts for the development of successful data repositories and infrastructures. What is clear is that "one-size-fits-all approach to policy development is inadequate" (NSF, 2005).

That is why debate on data ranges from the analyses on issues related to data sharing (Gold, 2010, Piwowar et. al., 2010, Piwowar, 2011) to studies in specific scientific fields (NIH, 2003, 2007, Karasti, et al., 2006, Baker et al., 2009, Waaijers et al., 2011) including surveys on usage patterns (Brown, 2003, Piwowar et al., 2007,) and researchers' attitude to make them available (Savage CJ, Vickers AJ (2009).

A few studies deal with the analysis of the existing dataset archives and compare their different characteristics (Marcial, 2010). Our paper intends to follow this type of survey, but with a different approach. In fact, the decision to use data archives listed in OpenDOAR enabled us to select a random sample given by the providers that had registered their archives in OpenDOAR. This approach throws light on an emerging reality such as IRs, that theoretically at least, have started to include datasets along with other digital objects. Insight can also be gained into archives of large scale and well-established datasets. Clearly, the adoption of a random sample affected our survey. In contrast to Marcial's empirical survey mentioned above, that found a cluster of elements common to different archives, our study revealed elements of dataset archives listed in OpenDOAR, in order to bring them into line with traditional archive classification (Armbruster & Romary, 2010). This enables the tracking of possible trends in dataset archive expansion policy.

In this paper we present the result of an exploratory analysis of a dataset archive in OpenDOAR. After the dataset definition given in paragraph 2., we describe the method used to select the sample and their main variables. In the fourth paragraph the results of our survey are reported.

* First published in the GL13 Conference Proceedings, February 2012.

2. Dataset definition

Research data are complex objects (Borgman, 2010) and that explains why there is no common agreed definition. They are very generally described and definitions, especially those reported in policy documents, include a very broad variety of digital objects (see Box 1). This is evident if we consider the definition given by the U.S. National Research Council in 1995, where research data are exclusively associated with numerical quantities. Following definitions encompass a wider range of digital objects (for instance images, sounds, etc.), thus representing research outputs in all scientific fields.

A more general agreement is reached when it comes to the definitions of dataset, considered as a meaningful and systematic representation of the subject being investigated. What is importantly stressed here is the importance of its re-use for validation and future investigations.

In this paper the term dataset is used to denote the digital collections managed in data archives.

Box 1. Data definitions

Research data definitions

- *National Research Council (1995): Data are numerical quantities or other factual attributes derived from observation, experiment or calculation http://www.nap.edu/catalog.php?record_id=4871)*
- *National Research Council (1999): Data* are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors. *(http://www.nap.edu/openbook.php?record_id=9692&page=15)*
- *National Science Foundation (2005):* The term 'data' is used in this report to refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment. *(http://www.nsf.gov/pubs/2005/nsb0540/start.jsp)*
- *OECD (2007):* Research data are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. *(http://www.oecd.org/dataoecd/9/61/38500813.pdf)*
- *PARSE.INSIGHT (2009):* Digital research data is used for all output in research. In practical terms, raw data, processed data and publications are all covered by the same term. A distinction between these sorts of research data is only made when necessary (for example when policies for publications are compared with other data). *(http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf)*
- *HLWIKI Canada (2011)* Research data is often defined as the information (e.g. data sets, microarray, numerical data, clinical trial information, textual records, images, sound, etc.) generated or used as quantitative evidence in primary biomedical research. This research data is distinguished by the fact that it is accepted by the research community as a means to validate research findings, observations and hypotheses. *(http://hlwiki.slais.ubc.ca/index.php/Data_curation)*

Dataset definitions

- *ODLIS (Online dictionary for library and information science):* A logically meaningful collection or grouping of similar or related data, usually assembled as a matter of record or for research, Also spelled *dataset*. *(http://www.abc-clio.com/ODLIS/odlis_A.aspx)*
- *OECD (2007):* A research data set constitutes a systematic, partial representation of the subject being investigated *(http://www.oecd.org/dataoecd/9/61/38500813.pdf)*
- *DOE (Department of Energy):* No-text scientific and technical information *(http://www.osti.gov/data/index.shtml)*
- *University of Edinburgh:* A set of files containing both research data and documentation sufficient to make data re-use. *(http://datashare.is.ed.ac.uk/)*

3. Methods

The information source of our analysis was the directory OpenDOAR (The Directory of Open Access Repositories) that currently lists more than 2000 repositories worldwide providing a detailed description of each of them. OpenDOAR categorizes and provides access to Institutional and Subject-based repositories, but also includes open access archives developed by funding agencies, governmental institutions and digital libraries.

The inclusion of different types of archives allowed us to analyse:

- Types of archives that collect datasets;
- Types of providers;
- Relationship between dataset characteristics and types of archives.

Moreover, OpenDOAR archive description, built on the information submitted by their providers, are then categorized, allowing users to sort the listed archives according to different criteria. We used the option "dataset" reported in the OpenDOAR content type categories to identify our first sample of analysis.

The purpose of our analysis was to track current trends in the development of data archives in the general framework of open access repositories, using the random sample provided by OpenDOAR listed archives. For these reasons, the OpenDOAR archive classification needed to be supplemented with the additional categories: Directory and Digital library. This was necessary in order to group archives with features different from "traditional" IRs or Subject-based repositories. Moreover, the category Digital library was introduced, even if limited to a single case, to show trends in data archives provided by libraries that may make their collections available and re-usable in digital forms.

The second step of our analysis concerned the identification of datasets provided in each archive of the OpenDOAR sample, which was performed searching for dataset, if the archives had this search option, or analysing the archives' collections manually.

According to the NSF definitions for data origin and digital data collections the sampled archives were analysed in terms of:

- Data origin (experimental, observational, computational)
- Types of Data collection (Research data, Resource or community, Reference data collections)

Moreover, datasets were classified as follows:

- Dataset content (Numeric, Scientific image, Image of artifacts, Maps, text-image)
- Dataset format
- Contextual information associated with datasets ("traditional documents", project descriptions, etc.).

Archives with a limited number of datasets (i.e. > 4) were excluded, as they were considered not representative of a stable commitment to dataset collection. Moreover, archives containing video, audio or other multimedia, were not considered in our analysis, this can be the subject of further analysis.

The latest update of our analysis was completed in October 2011.

4. Results

4.1. The sample

In OpenDOAR there are 80 archives that claim to contain datasets in their content type. The analysis of each of the selected OpenDOAR archives showed that only 29 out of 80 actually contain datasets, while 13 archives were discarded for the limited number of datasets available. In 33 archives no datasets at all were found, whereas the remaining 7 archives were not accessible (fig. 1).

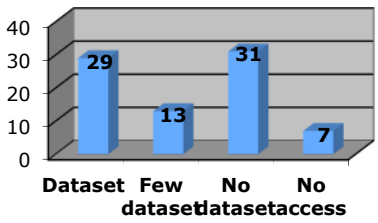


Fig. 1. - Dataset archives listed in OpenDOAR

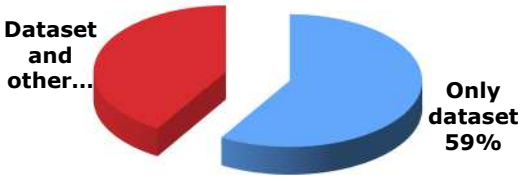


Fig. 2 Sampled archives by type of digital objects (n= 29)

Our sample of analysis consequently numbers 29 archives. Given the variety of archives listed in OpenDOAR, it should be noted that 59% of them (equal to 17 archives) exclusively contain datasets, while 41% (equal to 12 archives) contains both datasets and other digital objects, such as journal articles, reports, theses, etc. (fig. 2).

4.2. The data archives' providers

In the analysis of the type of providers that insert datasets in their archives we also wanted to verify whether they are single institutions or have built consortia. Our hypothesis is that consortia may have developed internal rules, specific metadata and or format to describe and exchange data to be shared within a specific scientific community. Results are reported in table 1.

Table 1. Dataset providers by type of organisation

Research Institution	
Single (15)	**Consortium (5)**
• *Chiba University, JP* • *Spanish National Research Council, ES* • *Cambridge University Library and Computing Service, UK* • *Data Archiving and Networked Services (DANS), NL* • *University of Southampton (Soton), UK* • *Data Library, University of Edinburgh, UK* • *Inter America Institute for Global Change Research (IAI), BR* • *International Food Policy Research Institute (IFPRI),US* • *University of Minnesota ,US* • *Monash University Library - Australia* • *Scripps Institution of Oceanography (SIO),US* • *University of Delaware Library, US* • *University of Hull ,UK* • *Centre de Données astronomiques de Strasbourg (CDS), FR* • *Marine Biological Laboratory & Woods Hole Oceanographic Institution (MBL & WHOI) Library, US*	• *Mineralogical Society of America, Mineralogical Association of Canada, University of Arizona, Schweizerbart Science Publisher, INT* • *COD Consortium, INT* • *Center for Research Libraries (CRL), US* • *Alfred Wegener Institute for Polar and Marine Research (AWI), Center for Marine Environmental Sciences (MARUM), University of Bremen, DE* • *Department of Geosciences, University of Arizona University of Arizona (UA), CALTECH (California Institute of Technology), US*
Indexing abstracting service	
Single (3)	**Consortium (1)**
• *Archaeology Data Service, UK* • *National Center for Biotechnology Information (CBI), US* • *National Library of Medicine (NLM), US*	• *Ontario Council of University Libraries, CA*
Publisher	
Single (1)	**Consortium (1)**
• *FigShare, UK*	• *Dryad, INT*
Government	
Single (3)	**Consortium (0)**
• *Coordenação de Biblioteca / CGDI / SAA / SE, Minisétrio da Saúde, BR* • *U.S. Department of Energy (DOE), US* • *Deutschen Zentrum für Luft- und Raumfahrt (EDINA), DE*	---

The majority of the providers of dataset archives are research institutions (20 out of 29), among which 8 universities and 7 research institutes.

Datasets are also available in archives developed by Indexing/abstracting services, governmental institutions and publishers. Such providers reflect the growing interest in datasets to be diffused for different purposes. At governmental level, for instance, the request for open data has been met by different countries that are progressively diffusing data collected within their institutions. In our sample we found the Brazilian Health Ministry, a German governmental agency for transport, and the U.S. Department of Energy that has a long tradition in the diffusion of technical information. Further, the presence of publishers represents the tendency to request datasets together with journal articles. In our sample a consortium of scientific journal publishers has developed Dryad that allows authors to submit their data and connect them with peer-reviewed articles. Similar features are provided by the publisher FigShare that provides citations of the datasets downloaded by authors.

4.3. Types of archives

In the analysis of type of archives we have adopted the traditional distinction between IR and subject based repositories. This classification is influenced by the information source we have chosen for our analysis, that has the advantage of exploring small dataset collections and verifying whether IRs are also beginning to consider datasets in their research results.

We introduced the category Directory to group heterogeneous types of archives, websites of governmental institutions, large databases that provide access to different data sources.

In OpenDOAR we also found an archive in the form of a Digital library, which we included in our analysis because we consider it a good example of providing a re-usable dataset from digitalised documents. In fact the South Asian Digital library not only digitalised an old text containing statistical data from the colonial period, but also provided an excel file that reported the datasets of the document. In our opinion this is a good example of making datasets re-usable, even if they are not digitally born.

Table 2. Archives by type

Subject-based Repository (15)
American Mineralogist Crystal Structure Database – http://rruff.geo.arizona.edu/AMS/amcsd.php
Archaeology Data Service - http://archaeologydataservice.ac.uk/
Crystallography Open Database (COD) - http://www.crystallography.net/
EDNA-the e-depot for Dutch archaeology - http://www.dans.knaw.nl/en/content/categorieen/projecten/edna-e-depot-dutch-archeology
eCrystals - Southampton) - http://ecrystals.chem.soton.ac.uk/
IAI Search - http://mercury.ornl.gov/iai/
Metropolitan Travel Survey Archive - http://www.surveyarchive.org/
PubChem - http://pubchem.ncbi.nlm.nih.gov/
PANGAEA® (Publishing Network for Geoscientific and Environmental Data) – http://www.pangaea.de/
RRUFF Project - http://rruff.info/
ShareGeo Open - http://www.sharegeo.ac.uk/
SIOExplorer Digital Library Project (SIOExplorer) - http://siox.sdsc.edu/
Verkehrsmodelle – http://modelle.clearingstelle-verkehr.de/
VizieR Catalogue Service - http://vizier.u-strasbg.fr/
Woods Hole Open Access Server (WHOAS) - https://darchive.mblwhoilibrary.org/
Institutional repository (7)
Chiba University's Repository for Access To Outcomes from Research (CURATOR) - http://mitizane.ll.chiba-u.jp/curator/
Digital.CSIC - http://digital.csic.es/
DSpac-e @ Cambridge - http://www.dspace.cam.ac.uk/
Edinburgh DataShare - http://datashare.is.ed.ac.uk/
Monash University ARROW Repository - http://arrow.monash.edu.au/vital/access/manager/Index
University of Delaware Library Institutional Repository - http://dspace.udel.edu:8080/dspace/
University of Hull Institutional Repository - https://hydra.hull.ac.uk/
Directory (6)
Biblioteca Virtual em Saúde - http://bvsms.saude.gov.br/php/index.php
Dryad - http://www.datadryad.org/
FigShare - http://figshare.com/
IFPRI Publications (Int. Food Policy Research Institute Publications) - http://www.ifpri.org/publications
OSTI (Office of Scientific & Technical Information) - http://www.osti.gov/
OZone (OZone provided by Ontario Scholars Portal) - https://ospace.scholarsportal.info/
Digital Library (1)
DSAL (Digital South Asia Library) - http://dsal.uchicago.edu/

4.4. Which science area?

The majority of dataset archives in our sample cover hard science (52%), but there is also a meaningful percentage of archives that provide datasets in Humanities and social sciences (fig. 3).

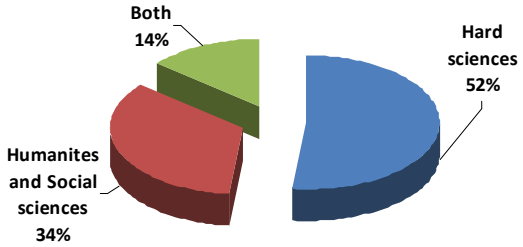


Fig. 3 Distribution of archives by science area (n=29)

If we group them in broad disciplinary fields, the most prevalent are Environment (21 %) and Demography (21%) (fig. 4).

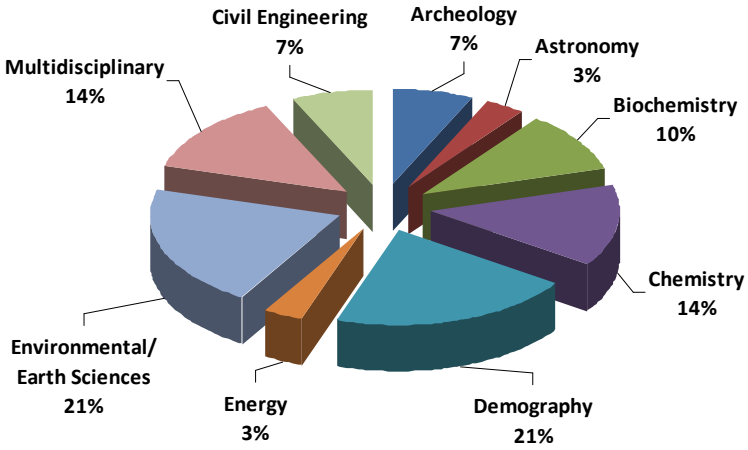


Fig. 4. - Distribution of archives by disciplinary fields (n=29)

Important criteria for the analysis of datasets depend on their origin, that is whether they are produced measuring specific phenomena at a given time, or are generated by experiments, or developing computational models or simulations to predict certain phenomena. Further, these variables are important when deciding whether it is important to preserve the data, considering that some of them cannot be so easily reproduced and/or collected. Figure 5 shows this variable linked with the scientific area.

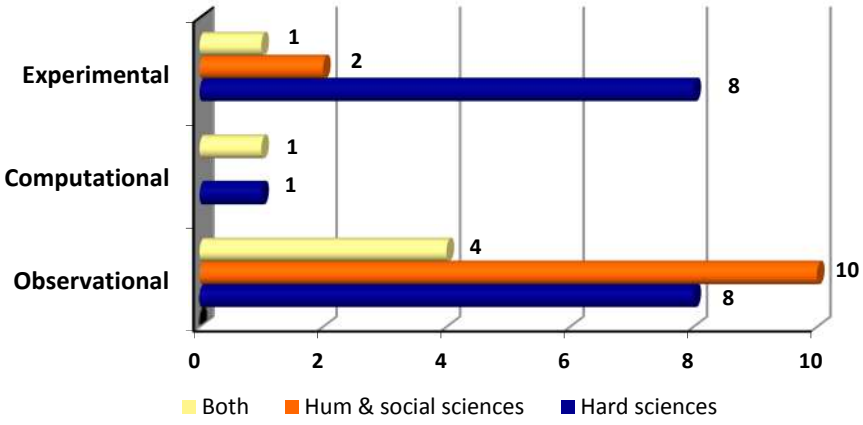


Fig. 5 Archives by data origin

The dataset archives in our sample are predominantly observational, and this is true in all science areas. Experimental data are collected mainly in hard sciences.

4.5. Functional categories of digital data collections

The National Science Foundation introduced three functional categories to analyse data collections referring to databases, infrastructures and organisations and individuals essential to managing this collection (NSF, 2005). This classification aims to distinguish between research data collected within a project of a certain size and budget as well as with different types of funds and funding sources. This distinction is also made to evaluate efforts necessary to preserve and diffuse datasets. Of course, a Research data collection can progressively become a Resource or Reference data collection, this was the case for instance of the well-known Protein data bank.

We applied these categories to the archives listed in OpenDOAR and compared them with the type of archives (fig.6).

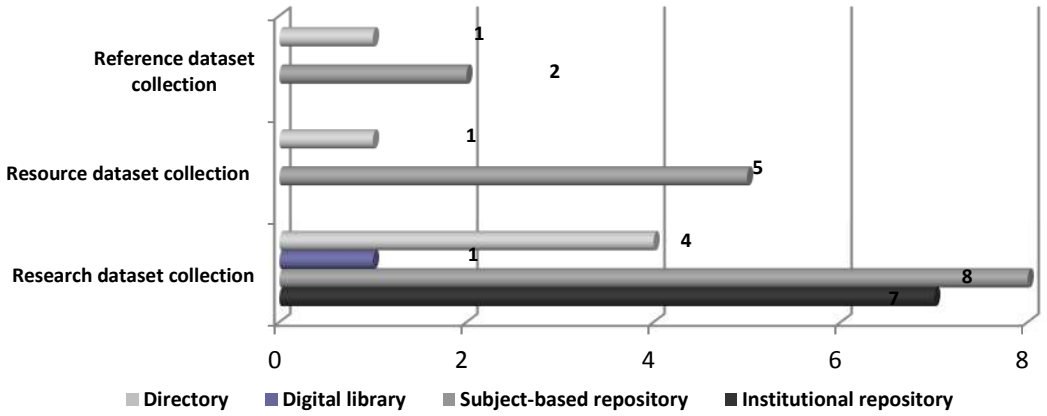


Fig. 6. – Archives by digital data collection

IRs exclusively contain datasets that fall into the category of Research data collections. Subject-based repositories contain datasets in all 3 categories, with a prevalence of Research data collections, while directories contain 1 Reference data collection.

4.6 Dataset content

For each archive in our sample we examined datasets with a view to analysing their content.

Figure 7 shows that the majority of archives contain numeric data, followed by scientific images, maps, text-images (i.e. digitized text) and images of artifacts.

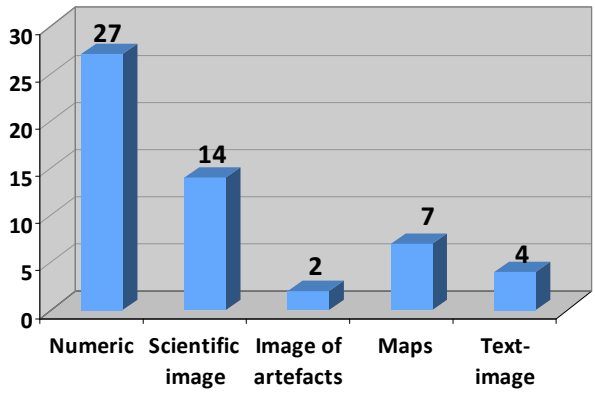


Fig. 7 Dataset content in our sample

Considering that the results of scientific observations, experiments or computational models can be expressed using different representations, not limited to numeric values, we associated the numeric content with other types of representation. Figure 8 shows the number of

archives that only contain numeric datasets and/or images and those that associate numeric data with other digital objects.

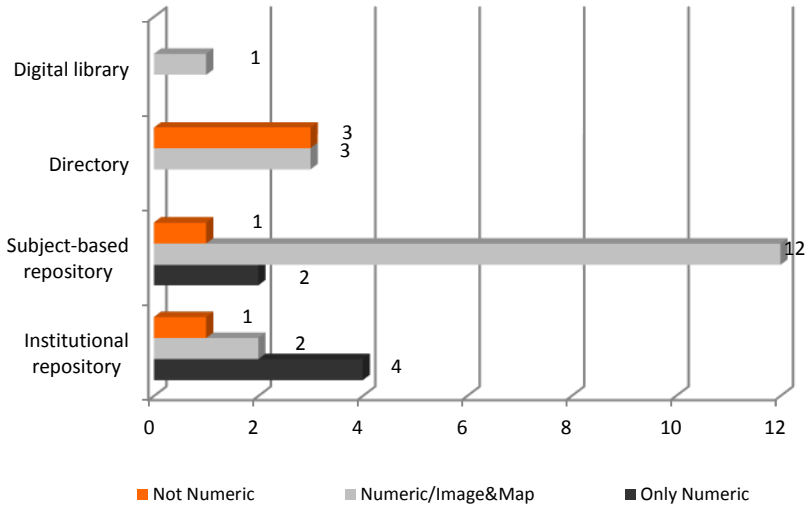


Fig. 8 Dataset content by type of archive

If we relate the content of a dataset with the type of archive, we see that there is a tendency to represent research results through numeric data associated with images. This is evident particularly in the case of Subject-based repositories (12 archives), while Institutional repositories tend to collect only numeric datasets (4 out of 7). Of course depending on the subject, some datasets are represented only by images (i.e. Not numeric) and this is present in all the kinds of archives in our sample.

The research results in the Subject-based repositories of our sample seem to provide a richer representation of datasets as a whole. For instance in the case of crystallography, the crystal structure described in the CIF format (see below) is combined with the graphical representation of its chemical structure, adding value for both crystallographers and chemists. (Cragin et al., 2010)

4.7 Dataset format

On the one hand file formats give evidence of the content of dataset (formats used to view images, texts and/or to store structured data already recognisable from their format extension). On the other, they also show how easily datasets can be exchanged. For instance the use of flat files, that is files that transform a record of a database into text, can be easily exchanged because they are not connected with proprietary systems. The disadvantage of using this format is that one needs to have additional information to interpret the data. In the archives listed in our sample we found different formats and sometimes the same archive provides the dataset in different formats so that users can easily access the data in the format he/she prefers. It follows that data format can also be considered an indicator of sharing and re-use. The formats more commonly used in our sample are reported in table 3.

Table 3. Dataset formats in our sample

File type category	File type/extension
Flat files	.txt, .ascii, .csv
Word processor	.doc, .pdf
Image	.tiff, .jpeg, .gif, .jmol
Spreadsheet	.xls
Statistical analysis	SPSS

As sharing and re-use are crucial for the dataset environment, we also looked for other file formats that facilitate their exchange. We found that some datasets were associated with the so-called readMefile, that contain important information, such as copyright, or how to install the database. We found readMefile especially in IRs (57%) and in directories related to Research data collections (67%).

It is of course the development of a specific standard format to exchange datasets that assures the highest degree of exchange and re-use. Their application indicates that a certain scientific community has a tradition in data sharing and has already agreed upon an exchange format that has a specific structure and meaning. An example of this standard is the CIF format used in crystallography to describe crystal structures or the standard used in astronomy to describe latitude, longitude and size of astronomical objects. It comes as no surprise that such exchange formats were present in Subject-based repositories (53%) and both in Resource and Reference data collections.

4.8 Datasets and "traditional documents"

Usually datasets are not self-describing, we need to know the context in which they are produced, how, and in which period, etc. Moreover their analysis can be described in other "traditional" documents, such as journal articles, reports, and theses (fig. 9).

In our sample we found that in the majority (72.4%) of archives, datasets are linked with traditional documents, and this is true for all types of archives.

Some archives also connected the dataset with the description of the project in which datasets were collected: this we found especially in large Subject-based repositories. Other archives also described the entire collection and this was the case especially in IRs.

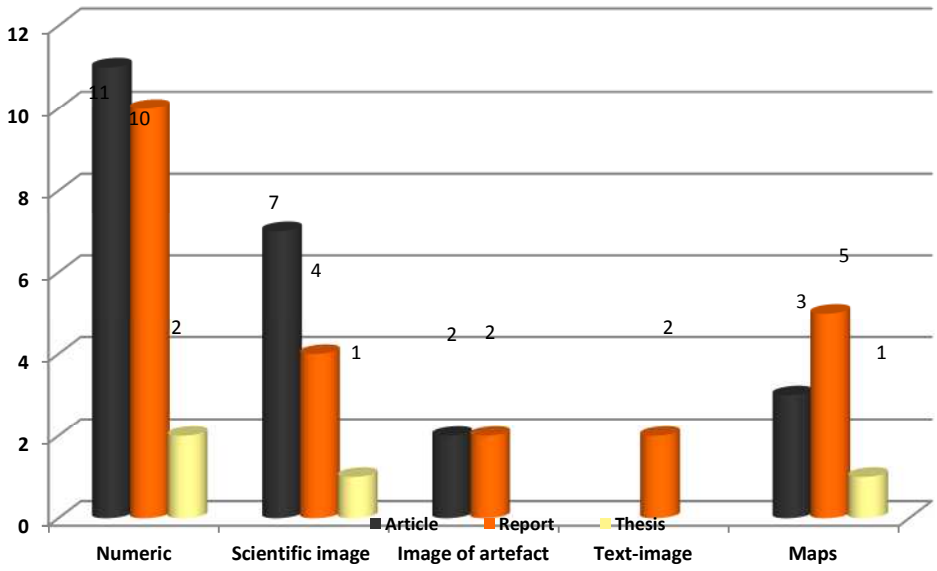


Fig. 9. – Dataset content by document type

5. Conclusions and discussion

Our sample enabled us to determine some common features, but also some characteristics that while not so widespread, may indicate possible trends in the development of dataset archives.

Considering the providers, our sample shows that along with research organisations and data services, governmental institutions and publishers too are developing archives making datasets available to the public. This is in line with the policy on open data announced in some countries as well as with the tendency of publishers to require datasets together with articles they are going to publish. Consortia are also frequently involved in building dataset archives, confirming the importance of collaboration in this field. Further analysis on the types of consortia (based on scientific collaboration, funding resources, and/or organisational models) should be carried out.

Datasets are collected in IRs along with other digital objects, while the majority of Subject-based repositories of our sample contain exclusively datasets. The introduced category Directory represents another way of organising data archives, combining different databases and linking various information sources. In our sample we also had an example of a Digital library that made datasets re-usable and an IR that contained only datasets.

Archives specifically focused on datasets and on specialized sub-disciplinary fields provide a richer environment in terms of data representations, of development of specific formats that facilitate data exchange, and of re-use and links to other digital objects and/or documents. This was evident in the Subject-based repositories and in Directories in our sample. This does not exclude that IRs cannot contribute to the collection and diffusion of datasets. Certainly, given the variety of datasets and their close relationship with the sub-disciplinary field in which they are collected, this poses different issues, such as self-archiving procedures and attitudes, ownership and copyright of data as well as their updating and maintenance. In this respect, the data collection categories proposed by the NSF provide a useful interpretative key and also suggest procedures to adequately construct and store data collections according to the type of archive and the mission of the archives' provider. In our sample, we had a prevalence of Research data collection in IRs, representing the outcomes of specific scientific projects with a limited user community and budget. Dataset availability in IRs along with other scientific results provide a more complete description of the research activities carried out in scientific institutions, while efforts concerning their visibility and usability should be further improved.

References

Anderson W.L. (2004). Some Challenges and Issues in Managing, and Preserving Access to, Long-live Collections of digital Scientific and Technical Data. *Data Science Journal,* 3.

Arms William Y., Larsen Ronald L. (2007). The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship. URL: http://www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf

Armbruster Chris, Romary Laurent (2010). Comparing Repository Types: Challenges and Barriers for Subject-based Repositories, Research Repositories, National Repository Systems and Institutional Repositories in Serving Scholarly Communication. *International Journal of Digital Library Systems*, 1 (4).
URL: http://arxiv.org/abs/1005.0839

Baker Karen S., Yarmey Lynn (2009). Data Stewardship: Environmental Data Curation and a Web-of-Repositories. *The International Journal of Data Curation*, 4 (2).

Borgman C.L., Wallis J.C., & Enyedy N. (2007). Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries. *International Journal on Digital Libraries*, 7 (1-2).

Borgman L. Christine (2010). Research Data: Who Will Share What, With Whom, When, and Why?
URL: http://works.bepress.com/cgi/viewcontent.cgi?article=1237&context=borgman>

Brown C.M. & Abbas J.M. (2010). Institutional Digital Repositories for Science & Technology Information: A View from the Laboratory. *Journal of Library Administration Special Issue: Emerging Practices in Science and Technology Librarianship*, 50:181–215.

Brown C.M. (2003). The Changing Face of Scientific Discourse: Analysis of Genomic and Proteomic Database Usage and Acceptance. *Journal of the American Society for Information Science & Technology*, 54(10): 926-938.

Cragin Melissa H., Palmer Carole L., Carlson Jacob R., & Witt Michael. (2010). Data Sharing, Small Science, and Institutional Repositories. *Philosophical Transactions of the Royal Society A*, 368(1926): 4023-4038.

Gold Anna (2010). Data Curation and Libraries: Short-term Developments, Long-term Prospects. *Data Curation and Libraries*, 4.

Graaf Maurits van der, Waaijers Leo (2011). KE Knowledge Exchange Primary Research Data Working Group. A Surfboard for Riding the Wave: Towards a Four Country Action Programme on Research Data.
URL: http://www.voced.edu.au/content/ngv48428>

Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council (2009). Harnessing the Power of Digital Data for Science and Society.
URL: http://www.nitrd.gov/About/Harnessing_Power_Web.pdf

Joint Information Systems Committee (JISC) Managing Research Data (MRD) Programme (2009).
URL: http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

Karasti Helena, Baker Karen, Halkola Eija (2006). Enriching the Notion of Data Curation in E-science: Data Managing and Information Infrastructuring in the Long- term Ecological Research (LTER) Network. *Computer Supported Cooperative Work (CSCW)* 15: 321-358.

Marcial Laura Haak, Hemminger Bradley M. (2010). Scientific Data Repositories on the Web: an Initial Survey.
URL: http://onlinelibrary.wiley.com/doi/10.1002/asi.21339/pdf

National Institutes of Health (2006). Data Sharing Policy and Implementation Guidance.
URL: http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

National Research Council (1995). Preserving Scientific Data on our Physical Universe: a New Strategy for Archiving the Nation's Scientific Information Resources. Washington D.C: National Academy Press.
URL: http://www.nap.edu/catalog.php?record_id=4871

National Research Council (1997). Bits of Power: Issues, in Global Access to Scientific Data. Washington, D.C.: National Academies Press. http://www.nap.edu/catalog.php?record_id=5504

National Research Council (1999). A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases. Washington, DC: National Academy Press.
URL: http://www.nap.edu/openbook.php?record_id=9692&page=14

National Research Council (2003). Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences. Washington, D.C.: National Academy Press.
URL: http://selab.janelia.org/publications/Cech03/Cech03-reprint.pdf

National Science Foundation (2005). Long-lived Digital Data Collections: Enabling Research and Education in the 21st century.
URL: http://www.nsf.gov/pubs/2005/nsb0540/start.jsp

National Science Foundation (2011). Dissemination and Sharing of Research Results.
URL: http://www.nsf.gov/bfa/dias/policy/dmp.jsp

OECD (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding.
URL: http://www.oecd.org/dataoecd/9/61/38500813.pdf

OpenDOAR (2011).
URL: http://opendoar.org/

PARSE.INSIGHT (2009). Insight into Digital Preservation of Research Output in Europe.
URL: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

Pirowar Heather A., Day R.S., Fridsma D.B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2(3): e308. doi:10.1371/journal.pone.000030
URL: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.00003088

Pirowar Heather A., Chapman Wendy W. (2010). Public Sharing of Research Datasets: a Pilot Study of Associations. Journal of Informetrics 4 (2): 148-156. doi:10.1016/j.joi.2009.11.010.
Savage C.J, Vickers A.J (2009). Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLoS ONE* 4(9): e7078. doi:10.1371/journal.pone.0007078

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E., Manoff M. Frame M. (2011). Data Sharing by Scientists: Practices and Perceptions; PLoS ONE (6)6
URL: http://dx.doi.org/doi:10.1371/journal.pone.0021101

UK data archive (2011). Managing and Sharing Data. [3rd ed].
URL: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Waaijers Leo, Graaf van der Maurits (2011). Quality of Research Data, an Operational Approach.
URL: http://www.dlib.org/dlib/january11/waaijers/01waaijers.html

Whitlock M.C. (2011). *Data Archiving in Ecology and Evolution: Best Practices. Trends in Ecology & Evolution* 26: 61-65. doi:10.1016/j.tree.2010.11.006.

Research product repositories: Strategies for data and metadata quality control*

Luisa De Biagi, Massimiliano Saccone, Luciana Trufelli, and Roberto Puccinelli (Italy)

Abstract

In recent years a significant effort has been spent by R&D institutions and scientific information stakeholders in general to enhance and improve the quality of Open Access initiatives and the performance of the associated services. Nevertheless much work is still needed to tackle pending data quality issues.

This paper proposes some functional and organizational solutions, based on the cooperation of all the main actors of the R&D system, which in our view should help improving quality control of data and descriptive metadata stored in research product Open Access (OA) repositories. We think that this strategy could favor a substantial innovation of the document management services offered to the scientific community and to policy makers, ensuring the interoperability between institutional repositories and Current Research Information Systems (CRIS).

Particular emphasis is given to the problem of data and metadata indexing and organization with respect to unconventional research products, which represent an important asset in the field of scientific communication.

Introduction

In Europe, despite the efforts of the scientific community and of many expert groups, effective methods and tools for R&D performance evaluation are still not available. This, in our opinion, is a top-priority issue, since reliable measurements are a pre-condition for credible process and product quality assessment.[1]

In this paper we propose a cooperative organizational approach for tackling some crucial challenges, such as research product metadata quality certification, with particular focus on metadata stored in Open Access repositories (OA).

Currently some national evaluation systems[2] [3] leverage data coming from institutional repositories, which are integrated within R&D Information Systems[4]. Disciplinary and institutional repositories can be used as data sources for R&D performance measurement, also because they keep products which are highly representative of the different scientific communities.[5]

Another interesting (but sometimes neglected) aspect of institutional repositories is that they can collect, index, keep and disseminate grey literature products. The availability of certified data about those products could provide new perspectives to science and technology phenomena investigation.[6] Actually, grey literature products could be used as a significant evaluation set both for bibliometric analysis and for investigations aimed at understanding science and innovation dynamics, change driving ideas, knowledge basis used in particular scientific developments, connections and communication patterns in particular disciplinary contexts.

In general, we think that cooperative systems facilitate the traceability of the different research product life-cycle phases and of the related metadata (versioning, persistent identification, etc.). The cooperative approach should be further extended within the scientific community to quality certification by adopting open and transparent peer-review processes (open peer review, open peer commentary, etc.).

Open Access repositories in R&D information system: strategic role of cooperation

Open Access repositories, whose number has been steadily rising in recent years, are an important component of the global e-Research infrastructure.[7] The real value of

* First published in the GL13 Conference Proceedings, February 2012.

repositories lies in the possibility of interconnecting them to create a network that can provide unified access to research outputs and be (re-) used by OA service providers, researchers' communities, management information systems (CRIS)[8], statistical information systems, bibliographic databases, etc.[9] However, in order to achieve this goal, a *multilevel* interoperability is needed. The purpose of this paper is to provide a broad overview of multilevel interoperability between Open Access repositories and other R&D information systems, identify the major issues and challenges that need to be addressed, stimulate the engagement of the repository community and trigger a process that will lead to the establishment of a cooperative network of R&D information management systems.[10]
Today, Open Access repositories are increasingly being used to collect, archive, and disseminate all types of research outputs such as research articles, conference proceedings, dissertations, data sets, working papers and reports.
Currently, research product data and metadata managed by OA and commercial repositories and databases are not used for official statistics due to several problems, such as the influence of the different national policies and strategies on the scientific production; the lack of a coherent framework of commonly agreed strategies; the different methods, tools and criteria used to collect data within the different public and private organizations; the lack of common classification criteria for product types, semantics and fields of reference; the insufficient reliability of data provided by the main bibliographic data bases (data base structure issues, lack of bibliographic & authority control tools, etc.); and more.[11]
The research process is an international and distributed endeavor, involving a variety of stakeholders such as scientists as authors and grant recipients, policy makers, research institutions, universities, publishers, and research funding agencies – each with their own set of interests. An international collaboration is needed between these stakeholders (actors) in order to develop cooperative and dynamic methodologies and processes for data and metadata quality control.
Interoperability is a pre-condition for a cooperative and widespread infrastructure of R&D information systems and for the value-added services and tools that can be built on top of the repositories.[12] The quality of these services depends on the data provided by repositories/CRIS/other information systems and on the standardization of "quality control processes" (quality of data and metadata collection and management processes).

Given the quantity and complexity of the problems affecting what in a broad sense could be called the R&D international information system, it seems evident to us that the interoperability should be implemented not only at the technical level but also at the political and organizational ones by all the institutions involved in the creation, management and use of the information resources.
Data and metadata model standardization is necessary in order to enable efficient data exchange and to allow researchers to find the desired information in the different research management systems.
From a strategic view point, the development of common logical and organizational data and metadata models i*n the Scientific and Research System* is important for:

- giving a simplified view to describe the specific area of interest;
- allowing for a better communication and multilevel interoperability between different information systems (Current Research Information Systems[13], Institutional Repositories, OA Service Providers, public and commercial Bibliographic databases, statistical databases, etc.);
- supporting *information workflow management;*
- supporting management and evaluation activities.

The aim of such cooperation should be the development of a common multilevel interoperability network and the first step should be a survey about policies and guidelines for organization and workflows, available data and metadata standards, cooperative

bibliographic, authority control and subject access systems, formats and access conditions, data use and re-use patterns, in order to gain sufficient insight into the scale of interoperability problems. Only on such basis, that is actual options, effective solutions can be developed and deployed.

The multilevel cooperation is necessary at the following levels[14]:

- Political: effective initiatives are needed at the national and international levels to favor open access to research results achieved through public funding; those initiatives should address and harmonize the different R&D stakeholders' interests;
- Institutional: academic and research institutions should define institutional and operational policies and carry out effective and widespread advocacy actions in their reference communities.
- *"For institutional record-keeping, research asset management, and performance-evaluation purposes, and in order to maximize the visibility, accessibility, usage and impact of our institution's research output"* [15];
- Economic and legal: Open Access is not zero-cost. Economic strategies are needed to sustain open access to public research products, based on the "author/institution pay" model; on the legal side, the adoption of Creative Commons (CC) licenses should protect intellectual property rights while granting open access;
- Technical-organizational: standards and commonly-agreed guidelines (based on a cooperative approach) are needed to certify data and metadata quality;
- Technological: OA greatly benefits from the development and widespread adoption of open standards and protocols and from the development of modular, interoperable and open source-based platforms for the management and diffusion of digital contents.

Green road: institutional and disciplinary archives

> *"…Two roads diverged in a wood, and I --*
> *I took the one less traveled by,*
> *And that has made all the difference."*
> (Robert Frost, The road not taken, 1920).

As a matter of fact, we could poetically say "*two roads to OA diverged in the wood of 'online scientific publishing*":

- the "golden road" of OA journal-publishing , where journals provide OA to their articles (either by charging the author-institution for refereeing/publishing outgoing articles instead of charging the user-institution for accessing incoming articles);
- the "green road" of OA self-archiving, where authors provide OA to their own published articles, by making their own eprints free for all.

In our opinion, the Green Road is the one that could bring more benefits to the scientific community.

One of the main research access/impact problem is that journal articles are not accessible to all potential users, causing a lack of potential research impact. The solution is making all articles really Open Access, granting a free, immediate and permanent online access to the full text of research articles for anyone, anywhere, webwide.

On the other hand we should consider the two roads to OA complementary, as well: the green road, representing the fastest and safest way to reach immediate 100% OA, might eventually lead to gold too.

In fact OA self-archiving is not self-publishing without quality control; nor it is meant to be scientific documentation for which the author could request payment and royalties (e.g. books or magazine/newspaper articles). OA self-archiving is bounded to peer-reviewed research, written only for research impact rather than royalty revenue[16].

The main consequence of a wider OA diffusion is that the whole society could benefit from a faster information spreading and from an accelerated research cycle through channels in which researchers can immediately satisfy their needs. It has been proved that OA articles

have a significantly higher citation impact than non-OA articles. Only 5% of journals are gold, but over 90% are already green (the green light to self-archiving is possible and authorized to authors); yet only about 10-20% of articles have been self-archived. To reach easily the '100% OA' goal, self-archiving needs to be mandated by researchers' employers and funders, as U.K. and U.S.A have recently recommended, and universities play a significant role in that. It is crucial that both funders and universities/research-boards mandate Green OA self-archiving, as not all research is funded and repositories are successful in attaining a considerable percentage of self-archiving only where a mandatory policy has been issued and enforced.

The main benefit supplied by OA, in general, and Green Road, in particular, is that researchers can increase visibility, usage and impact of their own findings, as well as their chance to find, access and use results from other researchers. On the other hand, Universities co-benefit from the increased impact of their researchers, because it also gives an excellent return on the investment to research funders, such as governments, charitable foundations etc. Finally, publishers likewise benefit from the wider dissemination, visibility and higher journal citation impact factor of their articles, and Open Access can generate new metrics to be used for assessing and improving research impact.

OA and grey literature valorization

Grey literature plays a significant role in the context of scientific documentation managed and diffused through Open Access archives, indexed and aggregated by the main service providers. Since the Seventh International Conference on Grey Literature at Nancy in 2006, GreyNet community started increasing its research activities relating to the OA effect on grey literature.

The adoption of open standards and OAI protocols by the International OpenGrey network facilitates the interoperability between OA repositories and OpenGrey (System for Information on Grey Literature in Europe). That's a first important step in developing cooperative networks for data and metadata certification.

The diffusion of the International Open Access initiative might certainly facilitate the development and coordination of cooperative networks, implementing sustainable processes and guidelines for:

- a better quality certification of grey literature products (open peer review, open peer commentary, etc.) and related metadata (adoption of common metadata standards and mappings, cooperative bibliographic and authority control, versioning, persistent identification systems, etc.);
- a better intellectual property protection especially for multimedia materials, containing a significant percent on Education, Learning and Professional Training (Creative Commons License is still weak). Moreover, a significant number of 'grey' production - as pre-prints, fact sheets, standards and working papers, committee reports, dissertation and Phd thesis - , still gets a discontinuous or null visibility due to intellectual property rights[17];
- a better information to users about copyright constraints (when and in which terms could I use it?);
- a wider access to research products, which can improve their visibility and impact.

Integrating Grey and Peer-reviewed literature often hosted in IR would enable a global view of the total available sources in a given scientific field, as well as an enhancement of research output measurements and metrics. Finally, it would also give increased researcher and affiliation visibility and (most importantly) better research outcomes.

Quality control: strategy, methods, processes and tools

Bibliographic standards and authority control tools are not sufficient to assure data and metadata accuracy, completeness and consistency.

Quality management systems are needed to define processes for the production and management of data and metadata (Trusted Digital Repositories) [18], which imply commonly agreed organizational models[19].
Only a shared effort can guarantee:

- Quality certification of the main data and metadata production and management processes;
- Commonly agreed bibliographic and authority control tools for metadata certification[20];
- Highly customizable software solutions, based on open standards and platforms.

In our opinion, after defining policies, strategies, services[21], methodologies and processes, the cooperative effort should be focused on the design and implementation of technical and organizational solutions able to support interoperability between the different R&D information Systems[22]. To achieve this goal it is important to:

- adopt a web service-based architecture (as in the JISC Information Environment Architecture);
- use open source software for information & content management systems (CRIS) and digital repositories (DSpace, E-prints, Fedora, JDIAM, Alfresco, etc.);
- use standard protocols and solutions for harvesting, aggregation, deposit, retrieval, cross-linking and context-sensitive linking (e.g. OAI-PMH and OAI-ORE[23], SRW - Search & Retrieve Web Service, SRU – Search & Retrieve URL Service, SWORD - *Simple Web-service Offering Repository Deposit,* Open URL[24], *etc.*);
- define an optimal set of context metadata, make sure these metadata are stored in CRISs and create automatic procedures for transferring these metadata to the repositories (CRIS-driven repositories – see also CERIF Metadata Model[25]);
- define common intermediary XML schemas for complex applications, in interoperable semantic and syntax context, for metadata interoperability, which allows flexible granularity[26];
- use interoperable record formats and syntaxes (e.g. SGML, XML, XML-RDF, XML-MARC, XML-MODS, XML-METS, etc.);
- use common standard models for web based interchange (e.g. RDF[27])
- participate to and leverage experiences from the cooperative development and use of Knowledge Organization Systems in the context of the semantic web (thesauri, classification schemes, subject heading lists and taxonomies, etc.) [28];
- enable citation metadata automatic detection within publications; work out/implement various multilingual controlled vocabularies (content international classifications) for the information objects in the Scholarly and R&D Information Domain (work out - or fill - the CERIF semantic layer)[29];
- define and use common research product categories and types (for example, CERIF – result–publication classification); [30]
- develop a cooperative bibliographic and authority control[31] system for Institutional Repositories and CRISs;
- develop cooperative multi-version control systems[32];
- extensively use specific unique and persistent identification codes:
    - for the different research product types (Handle, URN, DOI, Open DOI for dataset[33], SICI, ISBN, ISRN, ISTC, etc.);
    - for the researchers (international author ID, ORCID[34], etc.);
    - for research information space, CERIF entities being the core;
    - for institutions and projects (international Digital Institution Id – DII - and international Digital Project Id - DPI);
- develop a cooperative Persistent Identifiers (PI) resolution system (meta-resolver for PI)[35];
- develop cooperative semantic and meta search and discovery systems and tools[36].

CNR IA: a viable solution

In this section we will describe the situation of CNR research product archives, the current initiative aimed at implementing an Institutional Archive of research products and viable solutions to accomplish this task. A brief description of the CNR library system is given below, in order to allow a better understanding of the IA discussion.

CNR's library infrastructure reflects CNR's organization, featuring a Central Administration in Rome and a Scientific network made up of thematic institutes distributed all over the national territory. A significant percentage of CNR's institutes are hosted inside territorial Research Areas, which provide common services thus increasing efficiency.

CNR's library system features a hierarchical and distributed organization, which includes a Central Library (Biblioteca Centrale), Research Area Libraries (Biblioteche delle Aree di Ricerca), Institute Libraries (about 80). It provides a wide range of services to the entire scientific community and has recently adopted new organizational measures in order to increase the coordination of its different branches and improve the quality of the services provided to the internal scientific community. This effort has already produced some results in terms of process rationalization and digital resource sharing. The medium term objective is to complete the integration between CNR's libraries and to provide new added value services both to the internal and external scientific community.

At present, within our institution there are some research product archives but an Institutional Archive is not available. The existing repositories are based on open source platforms and are all OAI-PMH enabled.

An ad hoc working group has been established in order to define the architecture, standards, workflows and rules of a unified Institutional Archive. This group includes the personnel which has been involved in the development and management of the existing archives. The new architecture will be based on open standards and open source platforms. Web service interfaces will be provided for the communication with other systems.

From the researchers' perspective, auto-archiving will be implemented and favored. Obviously several levels of control will be enforced, in order to assure content and metadata quality. To this end, we think that the whole CNR library system should be involved, in order to have a first formal control at the local level (institutes and research areas) and a second one at the central level (Central Library). On the other hand, quality control will be automated where possible, leveraging the quality control strategies, methods, processes and tools described in the previous sections.

One of the main benefits for researchers will be the possibility to produce certified lists of their own publications (e.g. for internal career advancement procedures). We think that this could be a good incentive for self-archiving.

IA integration with CNR IS

Thanks to the web service based interfaces, the new system will be integrated with CNR Information system. Figure 1 shows the high level architecture of CNR IS. The new Institutional Archive is positioned in the right bottom corner.

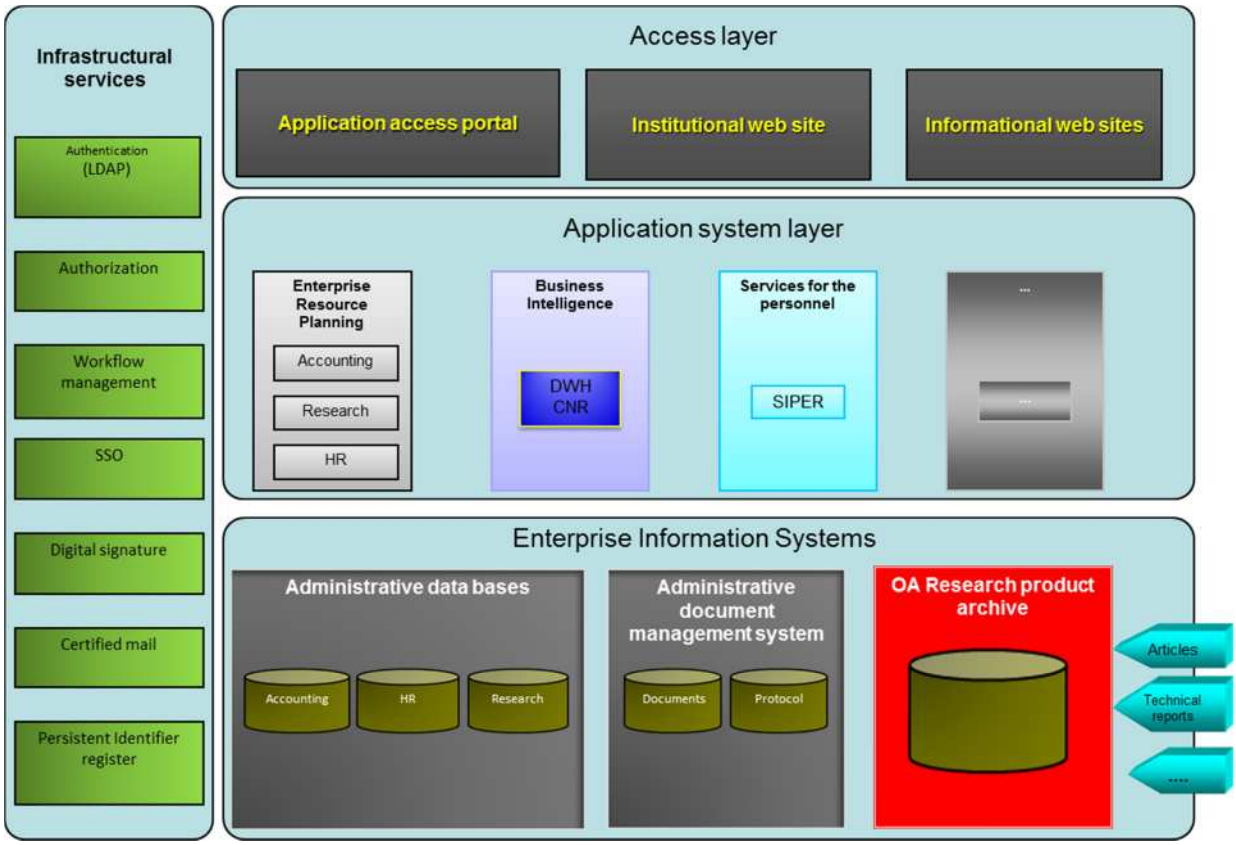


Figure 1: CNR IS high level architecture

At the bottom of this architecture there is the Enterprise Information System layer, which includes the administrative data bases and document management systems. The new IA will be positioned at this level. The Application System layer includes all the systems and applications that manage or analyze the data kept at the underlying level. The Access layer includes all the portals and websites that provide access to services and information residing in the Application layer. Orthogonal to the described layers there is the Infrastructural Services one, which provides cross-application services to the entire IS, such as authentication, authorization, single sign on, etc..

Particular care will be put in implementing an actual interoperability of the new IA with other internal and external systems. The reference schema for interoperability will be the EuroCRIS one, described in Figure 2 (single institution) and Figure 3 (inter-institution interoperability).

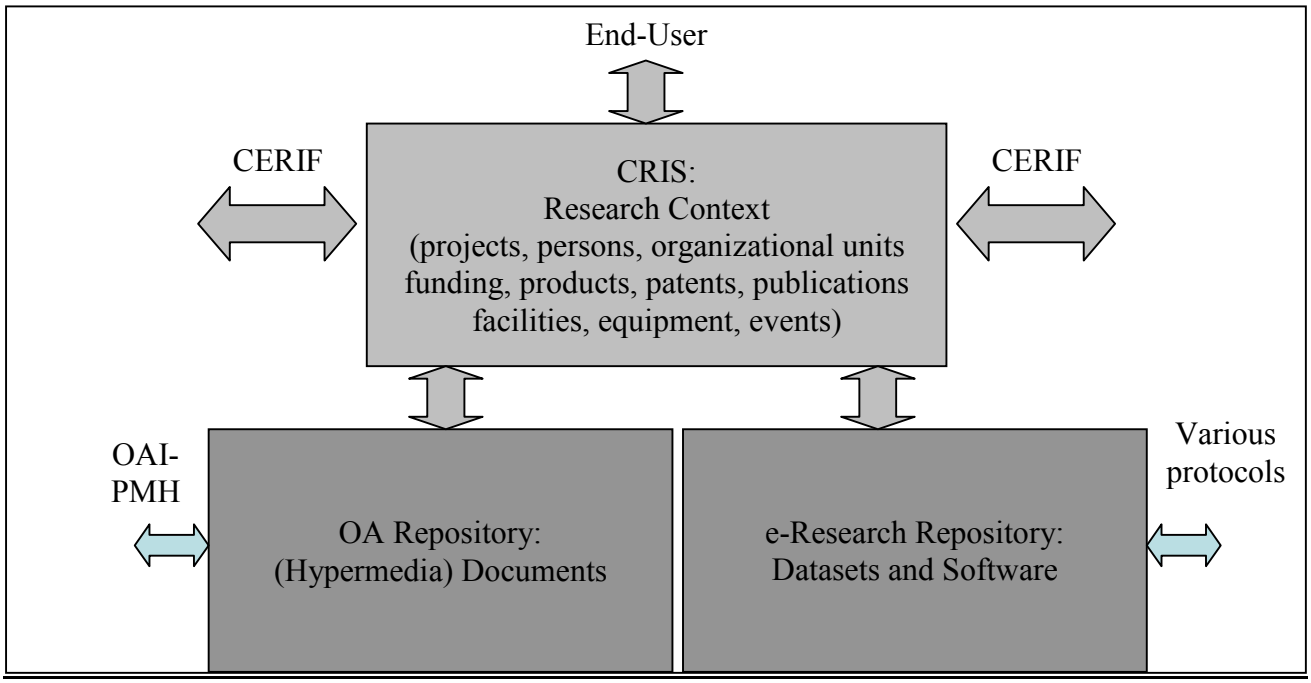


Figure 2: EusroCRIS schema - Architecture for a single institution[37]

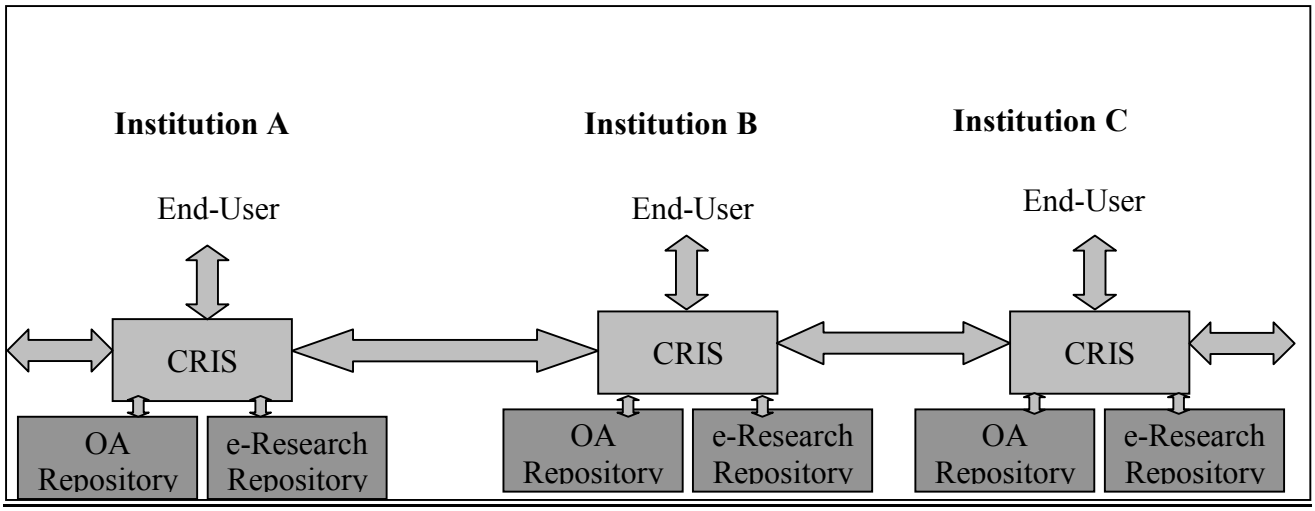


Figure 3: EuroCris schema for CRIS and OA/e-research repositories interoperability[38]

As regards the communication between systems, the figures clearly show that OAI-PMH will play a significant role at the repository level whereas CERIF will be the standard of choice for inter-CRIS communications.
Last but not least, persistent identification of digital resources, authors, institutions, projects, etc. will be taken in due account as well as standards for product classification.

Conclusions and future work

We think that it is important to be aware that the organizational and technical problems regarding multilevel interoperability are currently being discussed and addressed (or have been discussed and addressed in the past) in several other contexts[39], which are partly overlapping with the (digital) library community[40]:

- World Wide Web Consortium (W3C) (communities and working groups for interoperability);
- EuroCRIS – the European Organization for International Research Information (community for Current Research Information System interoperability)[41];

- the OAI (Open Archive Initiative) community (open archives and service providers based on harvested metadata according to the OAI-PMH, OAI-ORE protocols);
- institutional repositories/OA disciplinary repository networks (OpenAire, COAR[42], etc.);
- the Grey Literature Network Service and the OpenGrey - **multidisciplinary European database;**
- scholarly networks for Open Access publishing initiatives (SPARC - *Scholarly Publishing and Academic Resources Coalition, DOAJ* - Directory of Open Access Journals, OAPEN - *Open Access Publishing in European Networks, etc.);*
- Knowledge Exchange[43].

We think that we should learn from these communities and start with them discussions and common developments. The reason is not only the high similarity of data, services and ambitions, but also the fact that scientific products and data will be shared in all of these international contexts, thus requiring basic metadata to be produced only once, close to the source, and be re-used and augmented in other service contexts.
In our opinion, initiatives should be launched at the international level in order to:

- analyze new service scenarios/use cases for records and services or adapt existing ones;
- establish permanent cooperation for on multilevel interoperability involving R&D information system communities[44];
- establish international agencies or cooperative networks[45] for the definition and maintenance of commonly agreed workflow systems, principles, rules and vocabularies.

Within the Italian R&D system we are currently addressing the interoperability issue between the various information systems, also following the stimulus provided by recent laws and rules in the field of research evaluation. Within this context, OA archives are acquiring a great relevance thanks to their role of research product management systems and institutional data sources. In order to assure content reliability, a common effort is required for the development of cooperative certification systems.

References

- Organisation for Economic Co-operation and Development (OECD), ***Frascati manual 2002 : proposed standard practice for surveys on research and experimental development : the measurement of scientific and technological activities*****,** Paris, OECD - Organisation for economic co-operation and development, 2002, ISBN 92-64-19903-9;
- Carr, Leslie; MacColl, John, *IRRA (Institutional Repositories and Research Assessment): RAE Software for Institutional Repositories*, IRRA, 2006, http://irra.eprints.org/white/;
- *Open Access to research outputs: final report to RCUK*, LISU and SQW consulting, 2008, http://www.rcuk.ac.uk/documents/news/oareport.pdf ;
- *Open Access to research outputs: annexes: final report to RCUK,* LISU and SQW consulting, 2008, http://www.rcuk.ac.uk/documents/news/oaannex.pdf;
- Dijk, Elly, *NARCIS: linking CRISs and OARs in the Netherlands: A matter of standards and identifiers,* position paper presented at the *EuroCris Workshop on CRIS, CERIF and Institutional Repositories, CNR, Rome, 10-11 May 2010*, http://depot.knaw.nl/6365/;
- Confederation of Open Access Repositories (COAR). Working Group 2. Repository Interoperability, *The Case for Interoperability for Open Access Repositories. Version 1.0*, COAR, 2011, http://www.coar-repositories.org/files/COAR_Interoperability_Briefing.pdf ;
- Van der Graaf, Maurits; Vernooy-Gerritsen, Marjan (editor), *The European Repository Landscape 2008: Inventory of Digital Repositories for Research Output*, Amsterdam, Amsterdam University press, 2009, DOI: 10.5117/9789089641908 - E-ISBN: 9789089641908;
- ERA Expert Group 7 - EG 7: Rationales for ERA, *Developing World-class Research Infrastructures for the European Research Area (ERA)*, Luxembourg, Office for Official Publications of the European Communities, 2008, DOI 10.2777/96979, ISBN 978-92-79-08312-9;
- European Commission, *Work Programme 2012 - FP7 - Capacities: Part 1: Research infrastructures*, European Commission, 2011, European Commission, 2011, http://ec.europa.eu/research/infrastructures/pdf/wp2012_research_infrastructures.pdf#view=fit&pagemode=none
- Jeffery, Keith; Asserson, Anne, *Institutional Repositories and Current Research Information Systems*, New Review of Information Networking, 14, n. 2 (2009), p. 71-83, doi:10.1080/13614570903359357 – OAI Item Identifier: oai:epubs.cclrc.ac.uk:work/ 51773;
- White, Wendy, *Institutional repositories: contributing to institutional knowledge management and the global research commons*, In *4th International Open Repositories Conference, Atlanta, Georgia* 18th - *21st May, 2009*,

http://www.mendeley.com/research/institutional-repositories-contributing-to-institutional-knowledge-management-and-the-global-research-commons/;

- Vernooy-Gerritsen, Marjan; Pronk, Gera. Van der Graaf, Maurits, *Three Perspectives on the Evolving Infrastructure of Institutional Research Repositories in Europe*, Ariadne, n. 59 (April 2009), http://www.ariadne.ac.uk/issue59/vernooy-gerritsen-et-al/;
- Okubo, Yoshiko, *Bibliometric Indicators and Analysis of Research Systems: Methods and Examples*, in *OECD Science, Technology and Industry Working Papers*, 1997/1, Paris, OECD Publishing, 1997, http://dx.doi.org/10.1787/208277770603 ;
- Swan, Alma, *Sharing knowledge: open access and preservation in Europe: conclusions of a strategic workshop - Brussels, 25-26 November 2010 - Report*, Luxembourg, Publications Office of the European Union, 2011, doi: 10.2777/63410 - ISBN 978-92-79-20449-4;
- International organization for standardization (ISO), *Space data and information transfer systems. Open archival information system: Reference model. Standard ISO 14721:2003,*Geneva, ISO, 2003;
- *Giarretta, David; Harmsen, Henk; Keitel, Christian, Memorandum of Understanding to create* a *European Framework for Audit and Certification of Digital Repositories*, http://trusteddigitalrepository.eu/Site/Memorandum%20of%20Understanding.html.
- Mauro, Guerrini; Capaccioni, Andrea (a cura di), *Gli archivi istituzionali: Open access, valutazione della ricerca e diritto d'autore*, Milano, Editrice Bibliografica, 2010, p. 33-60, ISBN 9788870756920, http://hdl.handle.net/10760/15609;
- Park, Jung-Ran, *Metadata Quality in digital repositories: a survey of the current state of the art*, Cataloging & Classification Quarterly, 47, n. 3-4 (April 2009), p. 213 – 228, DOI: 10.1080/01639370902737240;
- **Guy, Marieke**; **Powell, Andy; Day, Michael,** *Improving the Quality of Metadata in Eprint Archives*, Ariadne, n. 38 (2004), http://www.ariadne.ac.uk/issue38/guy/;
- DINI - Deutsche Initiative für Netzwerkinformation. Working Group Electronic Publishing, *DINI Certificate Document and Publication Services - 2010: version 3.0*, march 2011, http://nbn-resolving.de/urn:nbn:de:kobv:11-100182800;
- Bijsterbosch, Magchiel; Brétel, Foudil; Natasa, Bulatovic Dale Peters; Vanderfeesten, Maurice, Wallace, Julia, *PEER. D3.1 Guidelines for publishers and repository managers on deposit, assisted deposit and self-archiving*, 2009; http://www.peerproject.eu/fileadmin/media/reports/D3_1_Guidelines_v8.3_20090528.Final.pdf
- Knowledge Exchange, *Guidelines for the aggregation and exchange of usage data*, http://wiki.surffoundation.nl/display/standards/KE+Usage+Statistics+Guidelines#KEUsageStatisticsGuidelines-GuidelinesfortheaggregationandexchangeofUsageData;
- Jeffery, Keith; Lopatenko, Andrei; Asserson, Anne, *Comparative Study of Metadata for Scientific Information: the place of CERIF in CRISs and Scientific Repositories*, 2002, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.5689;
- Gartner, Richard, *Intermediary schemas for complex XML applications: an example from research information management*, Journal of Digital Information, 12, n. 3 (2011), http://journals.tdl.org/jodi/article/view/2069/2086;
- EuroCRIS – The European Organization for International Research Information, *CERIF 2008 – 1.2 Semantics*, EuroCRIS, November 2010; http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_Semantics.pdf;
- Rumsey, Sally; Shipsey, Frances; Fraser, Michael; Noble, Howard; Bide, Mark; Look, Hugh; Kahn, Deborah, *Scoping Study on Repository Version Identification (RIVER) - Final Report*, 2006, http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf;
- Menzies, Kathleen; Birrell, Duncan; Dunsire, Gordon, *New Evidence on the Interoperability of Information Systems within UK Universities*, in Lecture Notes in Computer Science, 6273 (2010), p. 104-115, DOI: 10.1007/978-3-642-15464-5_12;
- Digital Archiving Consultancy, *Towards a European e-Infrastructure for e-Science Digital Repositories: a report for European Commission*, e-SciDR, 2008, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf;
- EuroHORCs and the European Science Foundation, *Vision on a globally competitive European Research Area and road map for actions to help build it*, EUROHORCs, 2008;
- Sutton, Caroline*, Sharing knowledge: EC-funded projects on scientific information in the digital age: Conclusions of a strategic workshop - Brussels, 14-15 February 2011 - Report*, Luxembourg, Publications Office of the European Union, 2011, doi:10.2777/63780 - ISBN 978-92-79-20451-7.

Endnotes

[1] Organisation for Economic Co-operation and Development (OECD), ***Frascati manual 2002: proposed standard practice for surveys on research and experimental development: the measurement of scientific and technological activities*****,** Paris, OECD - Organisation for economic co-operation and development, 2002.

[2] Research Excellence Framework 2014 (REF 2014), http://www.hefce.ac.uk/research/ref/; IRRA - Institutional Repositories and Research Assessment, http://irra.eprints.org/about.html; Leslie Carr, John MacColl, *IRRA (Institutional Repositories and Research Assessment): RAE Software for Institutional Repositories*, IRRA, 2006, http://irra.eprints.org/white/; *Open Access to research outputs: final report to RCUK*, LISU and SQW consulting, 2008, http://www.rcuk.ac.uk/documents/news/oareport.pdf; *Open Access to research outputs: annexes: final report to RCUK*, LISU and SQW consulting, 2008, http://www.rcuk.ac.uk/documents/news/oaannex.pdf.

[3] NARCIS - National Academic Research and Collaborations Information System, http://www.narcis.nl/; Elly Dijk, *NARCIS: linking CRISs and OARs in the Netherlands: A matter of standards and identifiers,* in *EuroCris Workshop on CRIS, CERIF and Institutional Repositories, CNR, Rome, 10-11 May 2010*, http://depot.knaw.nl/6365/.

[4] Confederation of Open Access Repositories (COAR). Working Group 2: Repository Interoperability, *The Case for Interoperability for Open Access Repositories. Version 1.0*, COAR, 2011, http://www.coar-repositories.org/files/COAR_Interoperability_Briefing.pdf.

[5] Maurits van der Graaf; Marjan Vernooy-Gerritsen (editor), *The European Repository Landscape 2008: Inventory of Digital Repositories for Research Output*, Amsterdam, Amsterdam University press, 2009, DOI 10.5117/9789089641908.

[6] *Ivi*, p. 19-21.

[7] ERA Expert Group 7 - EG 7: Rationales for ERA, *Developing World-class Research Infrastructures for the European Research Area (ERA)*, Luxembourg, Office for Official Publications of the European Communities, 2008, DOI 10.2777/96979; European Commission, *Work Programme 2012 - FP7 - Capacities: Part 1: Research infrastructures*, European Commission, 2011, European Commission, 2011, http://ec.europa.eu/research/infrastructures/pdf/wp2012_research_infrastructures.pdf#view=fit&pagemode=none [].

[8] Keith G. Jeffery, Anne Asserson, *Institutional Repositories and Current Research Information Systems*, New Review of Information Networking, 14, n. 2 (2009), p. 71-83, doi:10.1080/13614570903359357 [].

[9] Confederation of Open Access Repositories (COAR). Working Group 2: Repository Interoperability, *The Case for Interoperability op. cit*.

[10] Wendy White, *Institutional repositories: contributing to institutional knowledge management and the global research commons*, In *4th International Open Repositories Conference, Atlanta, Georgia, 18th – 21st May, 2009* [], http://www.mendeley.com/research/institutional-repositories-contributing-to-institutional-knowledge-management-and-the-global-research-commons/; M. Vernooy-Gerritsen, G. Pronk, M. van der Graaf, *Three Perspectives on the Evolving Infrastructure of Institutional Research Repositories in Europe*, Ariadne, n. 59 (April 2009), http://www.ariadne.ac.uk/issue59/vernooy-gerritsen-et-al/.

[11] Organisation for Economic Co-operation and Development (OECD), ***Frascati manual 2002, op. cit.*;** Yoshiko Okubo, *Bibliometric Indicators and Analysis of Research Systems: Methods and Examples*, in *OECD Science, Technology and Industry Working Papers*, Paris, OECD Publishing, 1997, doi: 10.1787/208277770603; Maurits van der Graaf; Marjan Vernooy-Gerritsen (editor), *The European Repository Landscape 2008: Inventory of Digital Repositories for Research Output*, *Op. cit.*, p. 100-110.

[12] ERA Expert Group 7 - EG 7: Rationales for ERA, *Developing World-class Research Infrastructures for the European Research Area (ERA)*, *OP. cit.*; European Commission, *Work Programme 2012 - FP7 - Capacities: Part 1: Research infrastructures*, *Op. cit.*

[13] A Current Research Information System (CRIS) records the R&D (Research and Development) activity either funded by or carried out by an organization, or within a thematic or subject area. Typically it covers projects, people (expertise), organizational structure, R&D outputs (products, patents, publications), R&D events and R&D facilities and equipment.

[14] Alma Swan, *Sharing knowledge: open access and preservation in Europe: Conclusions of a strategic workshop - Brussels, 25-26 November 2010 - Report*, Luxembourg, Publications Office of the European Union, 2011, doi: 10.2777/63410.

[15] Institutional Self-Archiving Mandate – Definition - ROARMAP (Registry of Open Access Repository Material Archiving Policies), http://roar.eprints.org/.

[16] S. Harnad, Open Access research, JeDEM 3 (1): 33-41, 2011

[17] Most of the Italian Phd Thesis indexed in Opengrey are not published, yet. Moreover, BNI (National Italian Bibliography) currently reports and describes all Italian Phd Thesis, also not published: in fact this document type is subjected to legal deposit at the National Library of Florence (in accordance with DPR 30.10.1997, n. 387, art. 4)

[18] International organization for standardization (ISO), *Space data and information transfer systems. Open archival information system: Reference model. Standard ISO 14721:2003,*Geneva, ISO, 2003.

[19]*David Giarretta, Henk Harmsen, Christian Keitel*, *Memorandum of Understanding to create* a *European Framework for Audit and Certification of Digital Repositories*, http://trusteddigitalrepository.eu/Site/Memorandum%20of%20Understanding.html.

[20] Mauro Guerrini, *Gli archivi istituzionali: Open access, valutazione della ricerca e diritto d'autore*, Milano, Editrice Bibliografica, 2010, p. 33-60; Jung-Ran Park, *Metadata Quality in Digital Repositories: A Survey of the Current State of the Art*, Cataloging & Classification Quarterly, 47, n. 3-4 (April 2009), p. 213 – 228; **Marieke Guy, Andy Powell, Michael Day,** *Improving the Quality of Metadata in Eprint Archives*, Ariadne, n. 38 (2004), http://www.ariadne.ac.uk/issue38/guy/.

[21] DINI Working Group Electronic Publishing, *DINI Certificate Document and Publication Services - 2010: version 3.0*, march 2011, http://nbn-resolving.de/urn:nbn:de:kobv:11-100182800.
A certificate that describes the technical, organizational, and legal aspects (including interoperability) that should be considered in setting up a scholarly repository service.

[22] Magchiel Bijsterbosch, Foudil Brétel, Natasa Bulatovic Dale Peters, Maurice Vanderfeesten, Julia Wallace, *PEER. D3.1 Guidelines for publishers and repository managers on deposit, assisted deposit and self-archiving*, 2009, http://www.peerproject.eu/fileadmin/media/reports/D3_1_Guidelines_v8.3_20090528.Final.pdf.

[23] OAI-PMH protocol limits interoperability to the unqualified Dublin Core schema, thus "flattening" research evaluation or increasing noise with an oversimplified metadata management process. Keith G. Jeffery, Anne Asserson, *Institutional Repositories and Current Research Information Systems*, *Op. cit.*; Open Archives Initiative – Object Reuse and Exchange (OAI-ORE) – Defines standards for aggregation of compound digital objects, http://www.openarchives.org/ore/.

[24] Knowledge Exchange, *Guidelines for the aggregation and exchange of usage data*, http://wiki.surffoundation.nl/display/standards/KE+Usage+Statistics+Guidelines#KEUsageStatisticsGuidelines-GuidelinesfortheaggregationandexchangeofUsageData

[25] Keith G. Jeffery, Andrei Lopatenko , Anne Asserson, *Comparative Study of Metadata for Scientific Information: the place of CERIF in CRISs and Scientific Repositories*, 2002, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.5689.

[26] In many metadata environments, particularly in that of the digital library, the problems of complex and highly flexible generic schemas are as acute as they are in that of CERIF - Common European Research Information Format. A tension arises particularly between flexibility and interoperability: the more potential approaches to encoding are offered by a standard, the more problematic is the transfer of metadata to different information systems and its interpretation and processing by them. Despite its great power as an encoding mechanism for the complex metadata needs of research environments, the CERIF model remains relatively underused in the area of research information management. Its flexibility and fragmented architecture in particular can produce significant problems for implementers and reduce its interoperability unless such key components as its semantic infrastructure are standardized between institutions. These problems were experienced by developer communities of such standards and were solved by some by using the architectural mapping features of SGML/XML. Without this facility in XML, the solution advocated here can replicate its best features but also add more powerful, non-syntactic features, such as semantic control.
The strategy has been tested thoroughly in several live research information management environments and found to be generally workable: the only problems experienced have proved to be those inherent in the metadata scheme on which the mapping to CERIF was based. The results have proved it to form a good compromise which allows the use of a key standard (with the consequent benefits of wider interoperability) in conjunction with a constrained, project-specific and more easily implemented element set. The successful application of this methodology suggests that it may be beneficial in the wider area of digital library metadata in general, where several key metadata schemas are more easily implemented when constrained it this way.
Richard Gartner, *Intermediary schemas for complex XML applications: an example from research information management*, Journal of Digital Information, 12, n. 3 (2011), http://journals.tdl.org/jodi/article/view/2069/2086.

[27] Resource Description Framework (RDF) – A standard model for web-based data interchange, http://www.w3.org/RDF/.

[28] SKOS - Simple Knowledge Organization System is an area of work developing specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web, http://www.w3.org/2004/02/skos/.

[29] The CERIF Semantics is one component of the CERIF 2008 – 1.2 Full Data Model (FDM). It aims at recommending a standardized formal semantics to be applied in the wider context of Current Research Information Systems (CRISs) with CERIF as the underlying data model to supply the relevant entities and their relationships. The semantic component in this version presents the current core semantics; that is, the types and roles considered relevant in a research context between the involved core entities. Compared to its preceding version, this release provides a major upgrade with respect to the quantity of relevant terms. EuroCRIS – The European Organization for International Research Information, *CERIF 2008 – 1.2 Semantics*, EuroCRIS, 2010.

[30] EuroCRIS – The European Organization for International Research Information, *CERIF 2008 – 1.2 Semantics*, *Op.cit.*

[31] VIAF – Virtual International Authority File, http://viaf.org/.

[32] Version Identification Framework Project, http://www2.lse.ac.uk/library/vif/index.html; VERSIONS (Versions of eprints. A user requirements study and investigation of the need for standards), http://www2.lse.ac.uk/library/versions/; The RIVER Scoping Study on Repository Version Identification - Sally Rumsey, Frances Shipsey, Michael Fraser, Howard Noble, Mark Bide, Hugh Look, Deborah Kahn, *Scoping Study on Repository Version Identification (RIVER) - Final Report*, 2006, http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf.

[33] DataCite, http://www.datacite.org/.

[34] ORCID – Open Researcher and Contributor ID, http://orcid.org/.

[35] PersID – Project aimed at building a persistent identifier metaresolver infrastructure for digital publications and electronic resources, http://www.persid.org/.

[36] Kathleen Menzies, Duncan Birrell and Gordon Dunsire, *New Evidence on the Interoperability of Information Systems within UK Universities*, Lecture Notes in Computer Science, 6273 (2010), p. 104-115, DOI: 10.1007/978-3-642-15464-5_12.

[37] Keith G. Jeffery, Anne Asserson, *Institutional Repositories and Current Research Information Systems*, *Op. cit.*

[38] *Ibidem*.

An architecture for providing a complete research information environment at an institution is presented. The linking together, at an institution, of a "OA repository of articles (that is a repository of publications deposited institutionally for toll-free open access in parallel with a peer-reviewed publication), a CRIS (to provide contextual information), and an OA repository of research datasets and software provides that institution with an information resource suitable for all the end-users and roles. Furthermore, the formalized structure of the CRIS allows a reliable workflow to be engineered which, in turn, encourages deposit of research outputs by reducing the effort threshold by using intelligent prompts or suggestions based on the information already stored and any constraints on permissible values of attributes. However the requirements of the end-user extend beyond the individual research institution or funding organization. The institutional CERIF-CRIS system can be linked to others because they have a formal structure and, hence, can be interoperated reliably and in a scalable way. This, in turn, provides a network of access to institutional OA repositories or e-research repositories linked to each institutional CRIS via the CERIF-CRIS gateways, enhancing and controlling the access using the CERIF-CRIS information as formalized, structured, and contextual metadata which is more detailed than DC and suitable for intelligent (machine-understandable) interoperation.

[39] Digital Archiving Consultancy, *Towards a European e-Infrastructure for e-Science Digital Repositories: a report for European Commission*, 2008, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf.

[40] Digital Library Federation (DLF), http://www.diglib.org/; DL.org Community – Digital Library Interoperability, Best Practices and Modelling Foundations, http://www.dlorg.eu/.

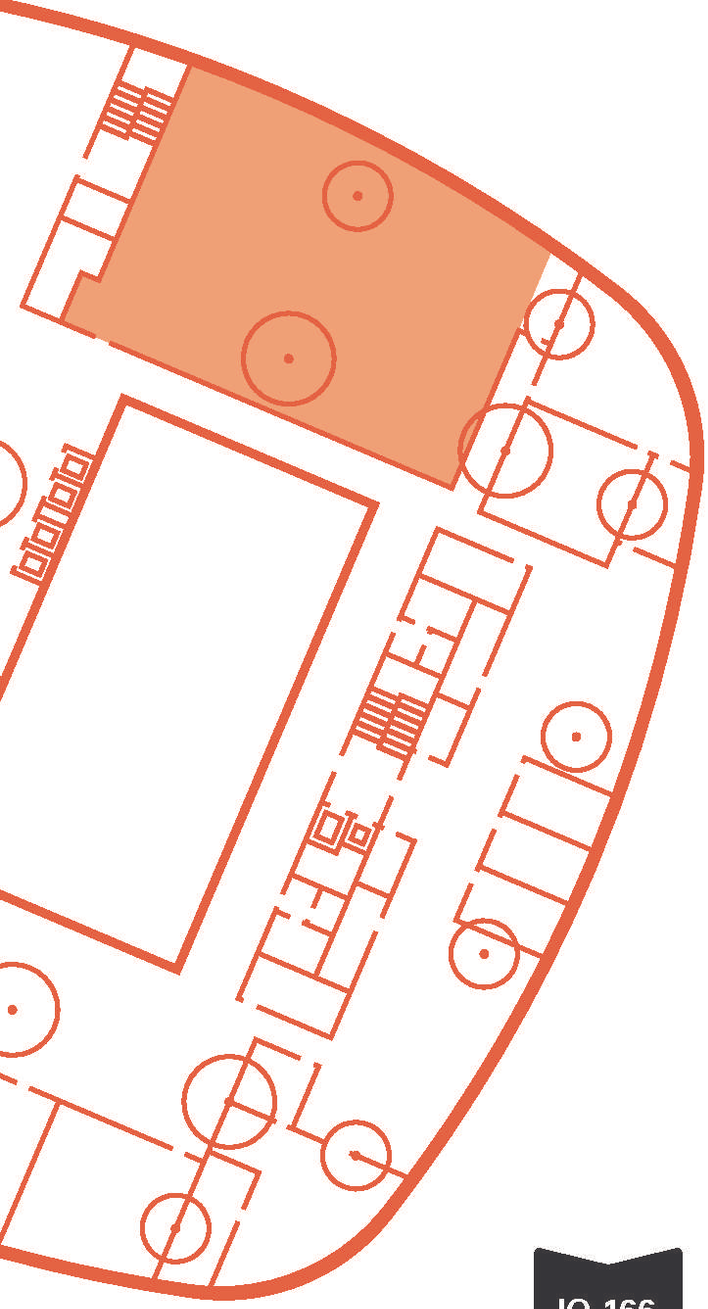
[41] EuroCRIS – The European Organization for International Research Information, http://www.eurocris.org/.

[42] COAR – Confederation of Open Access Repositories, http://coar-repositories.org.

[43] Knowledge Exchange is a co-operative effort that supports the use and development of Information and Communications Technologies (ICT) infrastructure for higher education and research.

[44] EUROHORCs, (European Heads of Research Councils), http://www.eurohorcs.org/E/Pages/home.aspx; EuroHORCs and the European Science Foundation, *Vision on a globally competitive European Research Area and road map for actions to help build it*, EUROHORCs, 2008;

[45] Caroline Sutton*, Sharing knowledge: EC-funded projects on scientific information in the digital age: Conclusions of a strategic workshop - Brussels, 14-15 February 2011 - Report*, Luxembourg, Publications Office of the European Union, 2011, doi:10.2777/63780.

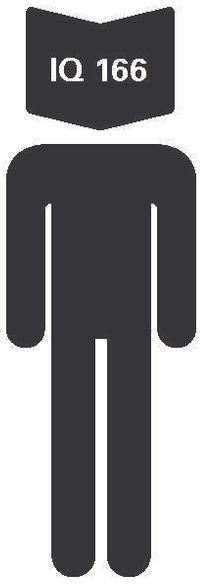


National Technical Library

National Technical Library (hereafter referred to as "**NTK**") is a central professional library open to public, which offers a unique collection of 250 thousand publications freely accessible in open circulation. Its holdings form the largest collection of Czech and foreign documents from technology and applied natural sciences as well as associated social sciences. It contains a total of 1,2 Mil. volumes of books, journals and newspapers, theses, reports, standards, and trade literature in both printed and electronic forms. Besides its own collection, parts of Central Library of the CTU in Prague and Central Library of the ICT holdings are accessible in NTK.

For detailed information on the National Technical Library visit **http://www.techlib.cz/en/**

94 mm



As corresponds to its statutes, NTK manages – among others – the project of building the **National Repository of Grey Literature.**
The project aims at gathering metadata and possibly full texts of grey documents in the fields of education, science and research.
The NTK supports an education in the field of grey literature through annual seminars in the Czech Republic.

For more information on the National Repository of Grey Literature visit our project Web site **http://nrgl.techlib.cz/** and for a search **http://www.nusl.cz/**

NTK Univers

Národní technická knihovna
National Technical Library

NUŠL

Audit DRAMBORA for Trustworthy Repositories: A Study Dealing with the Digital Repository of Grey Literature*

Petra Pejšová (Czech Republic) and Marcus Vaska (Canada)

Abstract

The credibility of a grey literature digital repository can be supported by a specialized audit. An audit of credibility declares that the digital repository is not only a safe place for storage, providing access and migrating to new versions of document formats, it also asserts the care components required of a digital repository environment, including the mandate, typology, policy, team, etc. This audit is very important in showcasing to participants and users the quality and safety of the data process.

This paper will present DRAMBORA (Digital Repository Audit Method Based on Risk Assessment), a methodology and tool for auditing a trustworthy digital repository of grey literature. DRAMBORA is an online instrument which helps organizations develop documentation and identify the risks of a digital repository. DRAMBORA is accessible from http://www.repositoryaudit.eu. The paper will also summarize prevailing advantages and disadvantages of DRAMBORA.

The second part of this paper will describe the audit of the National Repository of Grey Literature (NRGL) as a trustworthy digital repository using DRAMBORA as part of creating a digital repository of grey literature in the National Technical Library (NTK). The most important outcome of the audit was represented by the identified risks connected to the repository and potentially endangering its operation, quality, image, and other features. The main principle of the DRAMBORA audit and, at the same time, its main contribution, is its iteration (i.e. its repetition after a certain time period in new conditions when the original risks are reassessed; the measurements adopted for solution are assessed and new risks are identified).

Introduction: Audit for Trustworthy Repositories

"One of the central challenges to long-term preservation in a digital repository is the ability to guarantee the authenticity and interpretability (understandability) of digital objects for users across time" (Susanne Dobratz and Astried Schoger, 2007)

In our technologically-enhanced environment, managing, preserving, and storing material for posterity is essential, regardless of whether the material in question is a paper file or a digital object (Ambacher, 2007). In fact, efforts at maintaining a stronghold over digital records has been attempted since the 1960s, however, awareness surrounding the true digital repository has only existed for the past decade. This has lead to a number of organizations, most notably the Research Libraries Group (RLG)/U.S. National Archives and Records Administration (NARA) to establish an audit for certifying and enhancing the credibility of grey literature digital repositories. As with any marketing campaign, creating awareness of an initiative and gaining the public's trust is fundamental to ensure success. The Audit Checklist developed by RGL and NARA in 2005 supports this notion with its goal to "develop criteria to identify digital repositories capable of reliably storing, migrating, and providing access to digital collections…a method by which...customers could gain confidence in the authenticity, quality, and usefulness of digitally archived materials" (Ambacher, 2007, p. 2).

Long-term preservation of the material contained within digital repositories functions similarly to the storing of paper documents in a traditional index file within an archive. Ever since institutional repositories arose and began gaining acceptance in the 1990s, efforts at sustaining the material within these storage banks for generations to come have

* First published in the GL13 Conference Proceedings, February 2012.

been explored. The first such effort occurred in 1996 when the Task Force on Archiving of Digital Information drew attention to the need for a certification program for the long-term preservation of digital repositories, proclaiming that repositories "must be able to prove that they are who they say they are by meeting or exceeding the standards and criteria of an independently-administered program for archival certification." (Dobratz and Schoger, 2007, p. 210).

While traditional publishing ventures often result in a considerable time lag between an author's manuscript submission, peer-review by a panel of experts, and subsequent publication in a leading journal within a particular discipline, digital libraries, and in particular digital repositories, allow an author to submit a presentation, thesis, report, etc., as soon as it is written. Further, the author is able to choose from a number of creative commons licenses, maintaining control over his/her data, and deciding how and by whom the data can be accessed (Ambacher, 2007).

Credibility of Grey Literature Digital Repositories

As with any research pursuit, guidelines must be followed and adhered to in order to gain credibility and reputation that a chosen research path is indeed the right one. The same holds true when evaluating the trustworthiness of institutional repositories. Although researchers caution that the approaches used in a national repository could well transcend boundaries and apply to international pursuits, it does not necessarily lead to only one universal tool for preserving digital material over the long-term (Dobratz and Scholze, 2006). Rather, the major task of any repository should be "evaluating and disseminating examples of good or best practice and by initiating and intensifying regional, national, and international collaboration" (Dobratz and Scholze, 2006, p. 583).

In order for a repository to be deemed trustworthy, it must operate "according to its objectives and specifications (it does exactly what it claims to do)" (Dobratz and Schoger, 2007, p. 212). Further, a repository must contain information that is complete and control for any unplanned changes, whether these changes are accidental technological glitches or deliberate sabotage. It therefore becomes essential that any edits to any part of a record, once it has already been placed in the depository, is meticulously noted.

Dobratz and Schoger (2007) also make mention of groups of users whose particular interests lie in ensuring that the trustworthiness of repositories is maintained. These include users who wish to access the information, data producers and content providers, and funding agencies. In addition, repositories that wish to remain functional, trustworthy, and in business for many years down the road must "fulfill legal requirements…to survive in the market" (p. 212). A trustworthy digital repository puts the author's mind at ease, knowing that their information is secure, and will be preserved with the utmost integrity (Dobratz and Scholze, 2006). As previously mentioned, the RLG/NARA audit checklist and the Nestor certificate may be the most well-known means to prove the validity and trustworthiness of a repository, but they are by no means the only methods in existence.

What an Audit Represents

A question that should weigh heavily on the minds of any institution containing a digital repository is to assess what an audit represents to establishing criteria and trustworthiness, and what decisions must be made in order to either carry along the same work, or guide the repository in a different direction. Further, in order for a repository to be deemed trustworthy, it must meet its objectives, and contain information and material according to its mandate. There is certainly a strong tie between a trustworthy repository and its information technology infrastructure, dependent upon a number of competing factors. These include integrity, authenticity, confidentiality, and availability (Dobratz and Schoger, 2007). Authenticity precludes that the repository meets its objectives, containing information and material that, by its mandate, it is supposed to contain.

As Dobratz and Schoger (2007) explain, "availability is a guarantee of access to the repository…and that the objects within the repository are interpretable" (p. 212). This is essential for ensuring a repository's survival: repeated difficulties encountered with retrieving a specific item within a repository, or continuous maintenance resulting in repository downtime will result in clients choosing to deposit and/or access their material elsewhere. Allowing the owners of the repository to determine who should be granted permission to access the repository's contents instills a higher level of confidence for the depositing author, as he/she is able to upload and tag his/her own publications. Nevertheless, this level of access can be difficult to maintain. (Dobratz and Scholze, 2006).

Hou, Wojcik, and Marciano (2011) provide a voice that many institutions housing digital repositories can relate to: "integrity is an essential component of a trusted digital repository…all of the functional areas will have an audit trial" (p.182). Thus, establishing an audit for trustworthy repositories represents evidence gathered (usually by means of a checklist) measuring whether or not the repository adhered to pre-determined established evaluation criteria. Further, as digital repositories are primarily web-based programs relying on a server housed in the home institution, these repositories must have "a succession plan or escrow arrangements in place in case the repository ceases to operate." (Ambacher, 2007, p. 6). Ambacher also posits that data loss, whether accidental or intentional, will inevitably occur, a potential weakness that can be exploited. Therefore, maintaining a sustainable repository with a firm foundation, along with establishing a back-up alternate route in the event of a digital disaster, is essential.

While gathering appropriate hardware, establishing a reliable and secure network connection, and ensuring that a digital repository is utilized to its full potential are all essential components of certification; having the appropriate software to run the repository cannot be overlooked. The *Audit Checklist for the Certification of a Trusted Digital Repository*, jointly created in 2005 by the RLG and NARA, comments on the framework used to evaluate such common repository software packages such as DSpace, Eprints, and Greenstone (Kaczmarek et al., 2006). Regardless of the software package that is chosen, it must be applicable and adaptable, in order to "facilitate data transfer…easily…to take advantage of future, unforeseen developments in computer software and technology" (p.2).

The goal of the RLG/NARA Audit Checklist is "to develop criteria to identify digital repositories capable of reliably storing, migrating, and providing access to digital collections" (Kaczmarek et al, 2006, p. 4). Adhering to the three key areas of digital preservation (namely, technology, resources, and management), the Audit Checklist consists of four key sections: organization; repository functions, processes, and procedures; designated community and the use of information; technologies and technical infrastructure (pp. 4-5).

Reasons Why an Audit is Done

If a digital repository is mapped out appropriately, it can have tremendous benefit to both the author depositing research material, and the institution responsible for its upkeep and maintenance. Therefore, an audit need not necessarily be seen as a negative or patronizing activity, but rather as a means of establishing credibility, and educating the repository owner as to any changes that may be required in order to help the repository gain trustworthiness among its users. Of the numerous reasons for why an audit is undertaken, the following are considered to be the core criteria that is often adhered to: an audit should maintain a sustainable, secure repository, with a user-friendly interface; it should establish and maintain a policy that will result in a long-term repository for data producers; it will benefit from a solid management foundation, ensuring that high-quality information is continuously deposited; finally, an audit must identify weaknesses and risks, and establish a process to overcome these challenges (Prieto, 2009).

As the recent copyright issues in Canada indicate, particularly the current Access Copyright befuddlement that exists at some academic institutions, there are a number of legal ramifications that must be taken account when depositing material into any repository. The repository ownership must allow material to be uploaded, stored in an archive, and modified, as required, for posterity (Dobratz and Scholze, 2006, p. 587). Additional challenges faced by these institutions result from the speed in which some repositories have been established. As Downs and Chen (2010) explain, methods for storing and preserving digital content have not yet reached the level of organization used to house non-print material. This can raise doubts about the content of a digital repository, as "trust encompasses not only the integrity of the digital data, but also the authenticity of the links between the data and the data sources and documentation" (Downs and Chen, 2010).

Security of the contents within a repository will always play a prominent role. Repositories should be accessible around-the-clock, and include digital signatures as well as digital object identifiers (DOI) to be able to easily retrieve a requested file. In addition, the establishment of a consistent archiving format will ensure that documents are preserved for many years into the future. In fact, "the minimum availability of a document [should] be no less than five years" (Dobratz and Scholze, 2006, p. 590).

While supporters of the Open Access Movement would declare that the full contents of a repository should be freely and publically available to all (and indeed, this is the case with a number of institutional repositories, including DSpace at the University of Calgary), there are nevertheless a number of interest groups for whom trustworthiness holds particular merit. These include users who wish to access reliable information immediately and well into the future, content providers who rely on the audit of a repository to support their effort at ensuring high-quality information in a repository is maintained (i.e. a warranty for data producers), and corporations that determine whether or not a repository will receive adequate funding and for how long. Finally, as previously mentioned, entering the digital repository environment is indeed a competitive venture, and all repositories are therefore required to "fulfill legal requirements" (Dobratz and Schoger, 2007, p.212) in order to survive.

One methodology posited by Kaczmarek and colleagues (2006) is the creation of a matrix to function as a tool which will aid in the decision-making process of certifying a repository as a trustworthy source of information. Kaczmarek et al (2006) explain that settling on which software package best suits a particular repository will lead to a rubric "to determine how critical each particular point of functionality is and if that point is absolutely required" (p.2). Such steps were taken by the Exploring Collaborations to Harness Objects in a Digital Environment for Preservation (EXCHO DEPository) project, a joint effort between the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress, and the University of Illinois at Urbana-Champaign.

While the above examples of digital repositories comment on the importance of establishing policies that are firmly adhered to in order to establish trustworthiness and acuity, repositories must also be established in such a way that they can be easily customized if necessary. Such is the case with DCAPE, the Distributed Custodial Archival Preservation Environments project, originating out of the University of North Carolina Chapel Hill (Hou, Wojcik, and Marciano, 2011). Adhering to the three key preservation policies, namely "management of archival storage, validation, and trustworthiness" (p. 181), DCAPE supports one of the fundamental reasons why an audit of a repository is undertaken. Ensuring that high quality material is continuously deposited is certainly one way of ensuring a repository's livelihood, however without a user-friendly interface, authors and researcher's alike may become frustrated and choose to deposit their publications elsewhere, which, in turn, reflects negatively on the purpose of sustaining the repository for generations to come.

Existing Audit Methodologies and Tools

DINI, the Deutsche Initiative für Netzwekinformation, is aimed at supporting the Open Access movement in Germany. The aim of this guideline is to enhance the cooperative partnership between German educational institutions with a goal to "provide a tool for repository operators that could be used to raise the visibility, recognition, and importance of the digital repository within the university." (Dobratz and Scholze, 2006, p. 584).

As exemplified in many repositories, DINI criteria are based on two categories, the first of which explains the minimum requirements that must be captured in order for the repository to be deemed credible. These requirements include visibility and server policy, support for authors, legal issues, authenticity and integrity, indexing, impact and access to statistics, as well as long-term availability (Dobratz and Scholze, 2006, p. 585). Nevertheless, despite these rather strict requirements, Dobratz and Scholze comment on the challenges involved in deeming a repository to be both trustworthy and credible, hence the need for an audit. These include the establishment of a server policy, creating a visible service for authors, and implementing persistent identifiers (p. 586).

In addition to the aforementioned repository requirements, DINI also supports the need for creating open access to archived materials, and posits that a policy needs to be established to allow for each repository to be registered and recognized by large-scale collectives, namely the Directory of Open Access Repositories, OpenDOAR. (Dobratz and Scholze, 2006). As DINI proclaims, creating an open access policy showcases "a clear commitment to support the 'green way' to open access" (p. 587).

Originally created with cultural heritage organizations in mind, the Nestor Catalogue of Criteria for Trusted Digital Repositories serves as a guide for planning and maintaining digital repositories well into the future (Dobratz and Schoger, 2007). The criteria raised by Nestor include the following key concepts which can be applied to virtual any repository framework: compliance with terminology created by the Open Archival Information System (OAIS), abstraction, adequate documentation, transparency (essential to gain trust), adequacy, and measurability. As Dobratz and Schoger (2007) proclaim, these criteria will function as "indicators showing the degree of trustworthiness" (p. 214). The organizational structure for Nestor is divided into three top-level categories, each with a number of subdivisions. These are depicted as follows: organizational framework (defined goals, adequate usage, legal and contractual rules, organizational form, quality management), object management (integrity, authenticity, strategic plan for technical preservation, acceptance from producers adhering to established criteria, archival storage, usage, data management system), and infrastructure and security (adequate IT infrastructure, protecting the repository and the objects contained within it) (pp. 215-216).

(DRAMBORA): A Methodology and Tool for Auditing a Trustworthy Digital Repository

DRAMBORA description: tool and methodology

Launched in 2008, as the result of a joint effort between the Digital Curation Centre and Digital Preservation Europe, the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) functions as a toolkit "to make the self-auditing process easier and more efficient for repository managers" (Donnelly et al., 2009). Although digital repositories had already been in place for some time prior to the establishment of DRAMBORA, there was no standard guideline for determining the key components required for successfully implementing, initiating, and sustaining a digital archive. This issue led to the Centre for Research Libraries (CRL), widely credited as the developers of DRAMBORA to produce a list of 10 core requirements that all digital repository owners must be made aware of and should follow to preserve their archival storehouses for generations to come. As can be seen from this list, technological infrastructure plays only one part in ensuring that the data within a repository is adequately stored and maintained

over time. Creating a manageable process and action plan, along with accounting for any legal ramifications that may manifest themselves along the way, is equally important.

While DRAMBORA is a relatively recent phenomenon, it nevertheless underwent a series of pilot tests in the two years prior to its official unveiling. More than merely serving as another toolkit, it is primarily responsible for presenting "a methodology for self-assessment, encouraging organizations to establish a comprehensive self-awareness of their objectives, activities, and assets before identifying, assessing, and managing the risks implicit within their organizations" (Donnelly et al., 2009). Undoubtedly, attempting to maintain any form of electronic storage method implies a certain amount of risk, perhaps even more so than a traditional print collection. DRAMBORA has attempted to ease this risk process by positing a series of stages: authors are required to develop an organizational profile, describe and document a mandate, and list objectives/goals, activities, and assets (Donnelly et al., 2009). These stages are, however, only meant to serve as guidelines; the DRAMBORA team cautions that the entire purpose of this audit method is to serve as a living document, with revisions being made along the way as the need arises.

How to Use DRAMBORA

There are presently 18 institutions that rely on the DRAMBORA toolkit to conduct self-assessment audits of their digital repositories. A number of these organizations also hold strong ties to the grey literature community. Before describing how to use this methodology it is necessary not only to discuss the purpose of DRAMBORA, but also its three primary applications: as a web-based tool, DRAMBORA can assess the effectiveness of a repository infrastructure, and offer suggestions for its improvement; it acts as a preparatory resource for external auditors who may wish to serve as aggregators of the DRAMBORA movement; finally, it anticipates any potential weaknesses or challenges, and subsequently adjusts its plans to overcome these boundaries (Donnelly et al., 2009).

Existing in both an online and offline format, DRAMBORA is a user-friendly program that guides the user through four phases. First, the user is encouraged to register for a personal account, as well as provide details regarding the repository at his/her institution. This allows DRAMBORA to present a customized self-assessment profile for each user. Further, additional staff members from the institution in question will be identified, where they will be able to contribute to the self-assessment process. The subsequent phase of using DRAMBORA refers to the actual self-assessment audit. The goal of this stage is to ensure that the repository undergoing an audit establishes clear objectives with documented sources. The organization's mandate and/or vision/mission statement, along with any potential legal or technical issues should be listed here.

Following the actual self-assessment, any potential risks must be identified and assessed. For evaluative purposes, all of the risks identified are categorized and assessed according to their potential impact, accounting for the frequency or probability that any potential negative effects and subsequent risks could appear. (Donnelly et al., 2009). Finally, once all risks have been assessed and identified, a plan should be established to develop counter-measures, anticipated outcomes, and a timeline to reassessment the repository, to ensure that any issues have been resolved. The careful mapping of a repository using the DRAMBORA tool may already give an overview of what is finished and what is not, which documents, procedures, tools, and measurements are missing and where most critical risks reside.

DRAMBORA: Advantages and Disadvantages

Undoubtedly, as a web-based self-audit tool, DRAMBORA far surpasses a number of competitors in this field, and thus it would not do the tool justice without mentioning some of its key benefits. First and foremost, the online version of this program allows the user to view the internal activity of a repository, identifying any potential problems, and

rectifying them as quickly as possible. In addition, the user is able to interact with the content on the screen, navigating to sections of interest without having to flip through an entire text. Second, the methodology and tools are well implemented, clearly identifying the organizational role and structure of each institution taking part in the audit. Third, the descriptions used and examples presented are pertinent, intuitive, and applicable to the task at hand. This includes a clearly-defined mission statement, complete with key aims and objectives. Finally, the scale of evaluation compared to the risks is adequately assessed: "an internal understanding of the successes and shortcomings of the organization [enable it]…to effectively allocate or redirect resources to meet the most pressing issues of concern" (Donnelly et al., 2009). "For inspiration and possible direct help, the DRAMBORA tool contains a number of links to supporting documents and a range of practical examples of completed entries, whether in the preparation phase or the audit phase. In the area of risk identification, predefined risks can be directly used and modified or unique risks may be formulated." (Karlach, 2010, p. 127)

Despite the obvious benefits of DRAMBORA, there are nonetheless a few disadvantages that must also be considered. Namely, at present, the implementation and methodology is only available in English. Although English is seen as the universal language of communication, it is important to note that the majority of DRAMBORA users hail from European countries. Thus, offering the DRAMBORA interface is a variety of languages is a task that the developers of this toolkit would be wise to consider. An additional disadvantage relates to the technical, albeit programming aspect of this product. DRAMBORA functions perfectly fine on a standard Windows or Mac interface, but is not compatible with the Czech Windows operating system (i.e. it does not support the Czech character set – iso-8859-2/windows-1250). While it is understandable that this program cannot comply with every possible computer operating system, perhaps a text-based (DOS) version, in addition to the current HTML version, should be brought forward to the development team. The final two arguments surrounding the negative aspects of DRAMBORA center on access issues. Presently, read-only access is not permitted (a user must be fully registered and log in to a created account, in order to make use of al of the program's features). In addition, exporting records (either via e-mail or to a bibliographic management program) is currently not possible.

Audit of the National Repository of Grey Literature (NRGL)

Introduction to NRGL

The NRGL project, *The Digital Library for Grey Literature – Functional Model and Pilot Implementation*, started thanks to the support from the *Ministry of Culture of the Czech Republic* as part of both research and development programs. The project is divided into three phases, lasting from 2008 to 2011. Its main goals are the systematic collection, long-term archiving and provision of access to specialized grey literature, pertaining specifically to research and development, civil service and education, as well as with the business sphere and "open access" at the national level. To support this goal, the NTK created a functional network of partner organizations, a working model, and a pilot application. In addition, on the basis of verified technology and methods defined under the project, recommendations and standards are created for other institutions electing to build their own digital grey literature repositories. Recommendations and standards consist mainly of a preferred metadata format, exchangeable formats and templates, examples of licensing models and of legal issues resolved, preservation methodology, archives, and the provision of access to digital data.

NRGL Audit

The first audit of the NRGL as a trustworthy digital repository using the toolkit and methodology of DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) was performed at the end of 2009 as a part of creating a digital repository of grey literature in the National Technical Library (NTK). The audit results and experience from its course were summarized in a final report, and published in the book *Grey Literature Repositories* in 2010. The most important outcome of the audit was presented by identifying risks connected to the NRGL and potentially endangering its operation, quality, images, and other features; these risks were eliminated or moderated by the NRGL team during 2010. The main principle of the DRAMBORA audit and, at the same time, its main contribution and iteration, resulted in its repetition after a certain time period in new conditions when original risks were reassessed, the measurements adopted for solution assessed, and new risks identified.

The second audit of the NRGL digital repository was performed after one year. In this audit, the actual state of the repository was assessed, with progress achieved during 2010. New potential risks were identified, as well as possible ways to eliminate them or to reduce their impact. The NRGL documentation, the description of the whole project, its processes, procedures, and related documents developed significantly during 2010, which, in turn, made a solid basis for the audit.

Work with the DRAMBORA Tool went on without any of the issues and problems experienced with the previous audit. On the basis of lessons learned from 2009, the NTK communicated with the authors of the tool and methodology, and proposed some improvements and modifications. As a result, the web tool DRAMBORA appeared more stable after a one-year pause; however, the most unexpected modifications have not been introduced yet, especially regarding the elimination of the rather unpleasant fact that the online version of DRAMBORA does not support languages other than English at this time. Nevertheless, as the results of the audit are intended to be presented in the international field, namely in the area of grey literature projects, we will continue to use English.

NRGL Audit: Preparation and Definition

DRAMBORA Interactive was used during the preparation phase in addition to the audit phase (Karlach, 2010, p. 126-127). The preparation phase consisted of acquiring all relevant information and documents on the status of the repository, its description, standards, procedures, staff, material, budget, etc. (Donnelly et al., 2009). This information served as input data for the preparation phase of the audit and was entered into the DRAMBORA Interactive in the section *Before the Assessment*. Here, the repository was described, the scope of relevant areas (Functional Classes) of the audit were defined, and the repository staff, including a detailed description of individual team members and their roles, was listed. The definitions of staff roles were especially important since, at the subsequent risk identification stage, it was necessary to relate risks to respective roles. Even during the preparation phase, a substantial contribution might be made to an audited repository. This helps the staff see the repository from a global vantage point, to map and accumulate the most important descriptive data about the repository, and to point to possible deficiencies and defects, offering the opportunity for problems to be remedied and missing materials to be completed. The audit was run using the portion of the DRAMBORA tool called the Assessment Centre (Donnelly et al., 2009). Here, the repository mandate, including its mission, purpose, founders, etc., was defined. Other repositories were also identified that influenced its activity, both external (e.g., legislative) and internal (e.g., organization, content type restrictions, etc.). The audit continued by defining repository goals, activities, and the means used to achieve these goals.

NRGL Audit: Identified Risks

In addition to the mapped repository and the relevant environment, the producers of the methodology and the tool consider the most important output to be the analysis of identifiable risks endangering the repository, its quality, readiness, reputation and position in the eyes of both specialists and ordinary users. In the 2010 audit, 16 risks from the previous audit were assessed, primarily regarding the progress in their elimination, and an additional 8 new risks were identified. The NRGL repository is still in the pilot project stage; however, it is run on final software versions and real data are being stored. Identified and reassessed risks mainly refer to the description of activities and procedures of the repository, the state and development of the staff, project funding, hardware and software sources, including their backup and relationship to the NRGL environment.

After all necessary information is entered, the Reporting Centre function helps to create output reports on the identified risks for the repository, with respect to their relationship and plausible solutions. Two types of output report formats are available, either PDF or HTML. Other saved descriptive information cannot be exported easily, however, it is possible to copy saved snapshots of the audit page. Besides the mapped repository and its relevant environment, the producers of the methodology and tool consider the most important output to be the analysis of identifiable risks endangering the repository, its quality, readiness, reputation, and position in the eyes of both specialists and ordinary users. Since the DRAMBORA tool does not provide read-only access, it is regrettably not feasible to allow free access to the audit at this time.

Generally, risk elimination is much easier in a case where the respective area is fully under control and in charge of the NRGL management and team. If the risk relates to the cooperation within or even outside the NTK, the situation is considerably more complicated. The creation of a knowledge database NRGL Wiki indicates great progress; this database should be further developed and strictly adhered to, as the continuous documentation of procedures, activities and results of the NRGL team is of crucial importance for the elimination or minimization of the impact of most risks. Such progress may be seen in the development of the NRGL repository since the last audit along with the new activities and goals that have been added. Therefore, the documentation of activities and analysis of risks are most important. A large portion of risks reflect the topic of building the NRGL partner network, i.e. the partner network of providers of the repository content. Consequently, this area should be of priority especially for promotion and education. In relation to NRGL partners, sufficient attention should be paid to legal issues connected to the Author Act.

Conclusion

A yearly repetition of the audit under new conditions, identification of new or modified risks, and creation of another action plan make the audit an iterative process that contributes to the trustworthiness of the NRGL. Despite the valiant efforts of libraries, information technology specialists, and researchers, who devote considerable amounts of time and effort to maintain credible digital repositories, it can seem like a tall barrier to overcome. While Downs and Chen (2010) caution that “no organization can absolutely guarantee long-term preservation and access”, efforts to establish methods of audit and recognize trustworthy digital repositories must continue. As DRAMBORA and the subsequent audit of the National Repository of Grey Literature have shown, the task at hand may not yet been complete, but it is certainly moving in the right direction. It is thus perhaps fitting to conclude with the mission statement of Columbia University, which reflects not only on the goals of this particular institution, but which speaks to the efforts of raising awareness of grey literature in all topic fields and venues. Namely, one must “advance knowledge and learning at the highest level and…convey products of its efforts to the world” (Columbia Mission Statement, 2011). We therefore recommend that an

audit be undertaken on an annual basis, identifying any associated risks, and creating an action plan to make the audit an iterative process that contributes to the trustworthiness of the digital repository.

References

Ambacher, B. (2007). Government archives and the digital repository checklist. *Journal of Digital Information, 8*(2), 1-10.

Columbia University. (2011). *Mission Statement*. Retrieved November 20, 2011 from http://www.columbia.edu/content/mission-statement.html

Dobratz, S., & Schoger, A. (2007), Trustworthy digital long-term repositories: The Nestor approach in the context of international developments. *Research and Advanced Technology for Digital Libraries, Proceedings, 4675,* 210-222.

Dobratz, S., & Scholze, F. (2006). DINI institutional repository certification and beyond. *Library Hi Tech, 24*(4), 583-594.

Donnelly, M., Innocenti, P., McHugh, A., & Ruusalepp, R. (2009). *DRAMBORA Interactive User Guide*. Glasgow. Retrieved November 22, 2011 from http://www.repositoryaudit.eu/help/

Downs, R. R., & Chen, R.S. (2010). *Self-assessment of a long-term archive for interdisciplinary scientific data as a trustworthy digital repository.* Retrieved October 22, 2011 from journals.tdl.org/jodi/article/download/753/642

Hou, C.Y., Wojcik, C., and Marciano, R. (2011). Trusted digital repository design: A policy-driven approach. *Archiving, 7*, 181-186.

Kaczmarek, J., Hswe, P., Eke, J., & Habing, T.G. (2006). Using the Audit Checklist for the Certification of a Trusted Digital Repository as a framework for evaluating repository software applications. *D-Lib Magazine, 12*(2), 1-10. Retrieved August 3, 2011 from http://www.dlib.org/dlib/december06/kaczmarek/12kaczmarek.html.

Karlach, P. (2010). An audit of the National Repository of Grey Literature using the DRAMBORA tool. In Pejšová, P [ed.]. *Grey Literature Repositories*. Zlin: VeRBuM, p. 126-127. Available as an E-book at: http://nrgl.techlib.cz/images/Book.pdf.

National Technical Library. (2008). *Audit of the National Repository of Grey Literature (NRGL) in the NTK using the DRAMBORA tool: Second audit, 2010.* Retrieved August 25, 2011 from http://nrgl.techlib.cz/images/DRAMBORA_2010_EN.pdf

National Technical Library (2008). *National Repository of Grey Literature: An audit of the NRGL as a trustworthy digital repository.* Retrieved August 25, 2011 from http://nrgl.techlib.cz/index.php/Audit

NUŠL. (2011). *National Repository of Grey Literature [NRGL].* Retrieved November 20, 2011 from http://nrgl.techlib.cz/index.php/Main_Page

Pejšová, P. (2010). The development of grey literature in the Czech Republic. In Pejšová, P [ed.]. *Grey Literature Repositories*. Zlin: VeRBuM, p. 34. Available as an E-book at: http://nrgl.techlib.cz/images/Book.pdf.

Prieto, A.G. (2009). From conceptual to perceptual reality: Trust in digital repositories. *Library Review, 58*(8), 593-606.

Ross, S. (2006). The role of evidence in establishing trust in repositories. *D-Lib Magazine, 12*(7-8). Retrieved November 18, 2011 from http://www.dlib.org/dlib/july06/ross/07ross.html

Federal Information System on Grey Literature in Russia: A new stage of development in digital and network environment*

Aleksandr V. Starovoitov, Aleksandr M. Bastrykin, Anton I. Borzykh, and Leonid P. Pavlov (Russia)

Introduction

Since the late nineties when the Russian grey literature (GL) system in the sphere of scientific and technical information was first presented to the international GL community in Luxembourg we have had several opportunities to describe one or another aspect of the Federal Information System on GL in Russia [1,2,3,4,5]. This time we would like to dwell upon the system as a whole following its development from the past through present times to the prospective view all the more as this year a new ambitious project is started with the aim of renovating the system in accordance with the up-to-date requirements. The project has received a sufficient government funding for the coming three years.

The Russian Federation has inherited the federal-level information system on grey literature from the Soviet Union. The system covers the most informative kinds of grey literature - scientific research and development reports and post-graduate theses as the sources of scientific and technical information being centrally collected at the Centre of Information Technologies and Systems of Executive State Authorities (abbreviated in Russian as CITIS) in accordance with the Federal Law "On the obligatory copy of documents". The law obliges all the organizations – the collective authors of reports and persons – the individual authors of dissertations to give a free full-text copy of the documents to CITIS. In turn, the Centre is obliged not only to complete and permanently store the collection but also to disseminate the information on its content.

In the course of the past decades the system experienced several modifications in order to get adapted to the changing organizational and technological reality. In its present state the federal system combines the following three functionally separate systems run by CITIS: the traditional system for collecting, processing, storing and providing access to R&D reports and theses called the computerized information system on science and technology (abbreviated in Russian as ASINIT) that has recently been improved to store the full-text reports and dissertations in a digital form and provide full-text search and retrieval; the system for self-funded research projects registration and monitoring that was put into operation in mid-2000 to reflect a growing trend in funding R&D projects from research organizations' own financial resources; the federal register for the results of scientific and technical activities also created in mid-2000 with the idea of monitoring the life-cycle of patentogenic findings documented in scientific reports.

All the three systems are operative under the name "United Federal Database on Research and Development" (UFD R&D) and fulfil their functions however rapidly changing digital and network technologies create new environment to increase the systems' efficiency and improve its services. A new project in the process of development at CITIS is under the auspices of the newly-started State Programme of the Russian Federation "Information society (2011 – 2020)". The project is aimed at the creation of the Integral state information system on scientific research and development that is supposed to unite the three systems using unified forms of input documents so that users were to fill in the similar information only once and in interactive network conditions. The integral system will use the instruments of full-text digital documents analysis and web-technologies so that to improve data-mining and to avoid plagiarism.

* First published in the GL13 Conference Proceedings, February 2012.

The past

The computerized information system on science and technology (ASINIT) has been operating since 1975. It was then created as the grey literature part of the national library-information fund of the USSR and the part of the State System for Scientific and Technical Information (abbreviated in Russian as GSNTI). ASINIT consisted of two divisions: the full-text R&D reports and dissertations (the so-called primary documents) stored on microfiches and the bibliographic cards with abstracts (the so-called secondary documents) stored in the mainframe computer in a database format. There are two kinds of the secondary documents: the registration cards that are filled in when a new R&D project is started and the information cards that accompany full-text reports and dissertations.

Later on, in the early eighties ASINIT became the host core of the computer network called AIST in Russian for "computerized teleprocessing information network". AIST connected distant smart terminals, a prototype of personal computers situated all over the country, to the ASINIT host-computer with the grey literature databases situated in Moscow. The network operated in a dial-up mode through the public telephone lines. The distant users could conduct online searches in the centralized databases on reports and theses and order copies of documents from the System GL collection. The network throughput provided for more than 500 search, retrieval and copy-ordering transactions per day. That was the first information computer network of the pre-Internet epoch commercially working in the country.

No matter how obsolete the soft- and hardware of the System may seem now from the very beginning ASINIT met the main complex of requirements for completing the obligatory copy grey literature collection (R&D reports, candidate and doctoral dissertations – theses, descriptions of algorithms and computer programs), federal registration of the documents, the database support, online search and retrieval, abstract journals publishing, permanent storing and archiving the documents [1].

This system's configuration existed for several decades with the technological changes from mainframe computers to PCs, database and network servers and the information migrating from magnetic tapes through diskettes and CDs to the modern electronic data stores.

The present

At present ASINIT is still the heart of the United Federal Database on Research and Development (UFD R&D) along with other two systems appeared several years ago. All the systems are functioning on the technological platform of the Centre of Information Technologies and Systems of Executive State Authorities (CITIS). In 2004 by the Decree of President of the Russian Federation ASINIT was included in the list of strategically important systems. Since 2010 the System has been listed in the Federal Register of the State Information Systems.

By the end of 2010 the system supported the following information resources:

- the retrospective bibliographic database with abstracts for R&D projects registration cards and R&D reports information cards containing nearly 2,5 million documents with the depth of retrospective 30 years (each card consists of more than 30 information fields) including
  - registration cards – nearly 1,2 million;
  - information cards – nearly 1,3 million;
- the retrospective bibliographic database with abstracts for dissertations containing more than 640 000 documents with the depth of retrospective 30 years (each card consists of 35 information fields) including
  - candidate dissertations information cards – nearly 560 000;
  - doctoral dissertations information cards – more than 80 000;
- the abstract journals database – nearly 3,0 million documents;
- the database for information cards translated into English – more than 80 000 documents;

- the algorithms and computer programs database – more than 15 000 documents;
- full-text R&D reports (since 1984) – nearly 800 000;
- full-text dissertations (since 1984) – nearly 600 000 including
  - doctoral dissertations – nearly 80 000,
  - candidate dissertations – more than 500 000;
- the database for scientific organizations submitting R&D reports – more than 6 000 organizations.

The report and dissertation information cards databases are placed on a CITIS server with online network availability for the users. The databases serve as an electronic catalogue for the full-text collection providing a fast means of registration and search. The arriving full-text paper reports and dissertations are being scanned and PDF stored. At the same time the earlier documents are retroconverted (now backwards to the year of 2000) from the microfiches to PDF format. About 11 000 full-text R&D reports are entered into a full-text database. For the beginning of 2011 the total size of the electronic document store is 5 TByte. The total size of the information fund – more than 7 million documents.

The desk-top publishing system allows for issuing both electronic and paper abstract journals but now only the electronic versions are commercially disseminated by subscription. 51 titles of the journals by 25 subject series are published, totally 236 issues per year.

There are two government level documents which form a legal ground for the operation of the system: the Federal Law "On the obligatory copy of documents" of December 29, 1994 № 77-FZ (in the wording of March 26, 2008 № 28-FZ) and the Government Decision of March 31, 2009 № 279 that delegated all the functions of running ASINIT to CITIS.

The system collects and controls scientific and technical reports and dissertations concerned basically all scientific subjects ranging from mathematics, physics, electronics and engineering through to social sciences and the humanities and supports monitoring and controlling the situation (both in financial and subject respect) in *the state funded* scientific research and development activities covering extensively all the territory of the Russian Federation [2,3]. The system's collection is an indispensable source for government agencies with an interest in the latest Russian contributions to science and technology.

At the same time it is evident that no matter how much money is given to science from the state budget it can never be the only and sufficient financial source for research and development and the diversification of funding is inevitable. So, there is a growing trend in scientific research that more and more R&D projects are being funded from *research organizations' own financial resources.* Those organizations are commercial ones functioning in the forms of federal state unitary enterprises and open joint-stock companies with the state share-holding. Their self-funded research projects were out of centralized monitoring and hence were not taken into account when updating the lists of priority development directions in science and critical technologies of the Russian Federation.

To eliminate the defects in research monitoring a special Government Decision was issued on November 4, 2006 No. 645 with the idea of creating a system for self-funded research projects registration [4]. The system was designed in the years of 2007 – 2008 and now is in operation as the second part of the United Federal Database on Research and Development (UFD R&D). Based on the output information from the system the Annual Summary Report for the Government is prepared. In accordance with the Decision the information on self-funded research should be submitted in an approved unified form as an annex enclosed in the organization's annual financial report. The approved blank form is added to the Decision's text. The form's fields of data are important because their filling determines the information value of the document.

Now there is a four-year retrospective database (with the report documents of 2007- 2010 years – totally about 1,5 thousand documents), next year (2012) the documents of 2011 will be entered and so on. Thus, the system ensures the registration of report documents on self-funded research, their permanent storage in the database format and both quantitative and qualitative analysis of self-funded research in Russia prepared in the form of the Annual Summary Report. The self-funded research monitoring system is evidently grey because its input form and output Annual Summary Report are typically grey documents. Since 2010 the System has been listed in the Federal Register of the State Information Systems.

The grey literature sources contain a bulk of findings to be commercialized and/or claimed as intellectual property objects. The registration of reports and dissertations that is carried out in ASINIT now is rather document- than result-oriented. It would be useful to follow all the lifecycle of a scientific result beginning with the idea and basic research outcome through feasibility study findings to industrial implementation of the result in the form of innovative products and services [5].

In 2005 a Government Decision was issued (No. 284, now it functions in the wording of August 18, 2008 No. 622) on the development of the United Register for the Results of Scientific and Technical Activities (UR RSTA). In 2006 the Register was put into operation with its separate input forms designed to register the objects of intellectual property (patents, databases, computer programs, etc.) obtained in the course of state-funded research. Now the Register database contains the information on 50 ministries – the state R&D projects customers, 15 000 state contracts concluded by them to carry out the projects and 6 000 intellectual property objects. The Register is the third component of the United Federal Database on Research and Development (UFD R&D) operating in CITIS.

Though functionally satisfying the main requirements of the Law and scientific community the existing UFD R&D suffers from several shortcomings that are supposed to be eliminated in the course of the further system's development:

a. all the databases (DB) on R&D — state contracts DB, reports and dissertations DB, the Register DB – are formed independently one from another, so the user has to fill in several similar forms wherein the information is redundant and duplicated;
b. the lack of effective customers and executors control mechanisms, so to say, a feedback from the System to the customers in order to provide the completeness of R&D reports registration and presence in the System;
c. the total computer power of the System is insufficient to implement the modern web-technologies of forming the information resource and providing a comfortable access to it;
d. the limited analytical means of textual information processing;
e. there is no online interaction with other state information systems such like the Computerized information system for scientific research of the Russian Academy of Sciences.

The future

The newly-started State Programme of the Russian Federation "Information society (2011 – 2020)" has opened a real chance for the state financial support of the System's development in the context of rapidly changing digital and network technologies. Under the auspices of the Programme a competition was announced for a state contract to realize the System's development project. CITIS won the competition and the contract was concluded for three years to fund the project named "The development of the United R&D projects registration system (UPRS R&D) for the projects carried out in the civil sphere with the state budget funding".

In general, the project is aimed at the creation of the integral state information system on scientific research and development that is supposed to unite the three systems functioning on the platform of CITIS.

There are some main problems to be solved within the project:

- The development and implementation of effective mechanisms for securing the information completeness in the System that is all the R&D reports must be registered and present in the System. This is very important because the experience of the latest decades exposed a low executive discipline of scientific and scholar institutions that perform R&D projects.
- The elimination of redundancy and duplication in the inputted and stored information due to the existing database conducting independence. Different databases have different forms of records with many coinciding information fields.
- The registered data must include not only the subject of research information but also the data on the state contracts, size and structure of the state funding, patentogenic results of the research project.
- The development of analytical instruments to expose the innovating projects and estimate the results of conducted research.
- The development of the legal basis for the UPRS R&D. A new Government Decision regulating the procedures of the System's operation must be prepared and approved.

From a technological point of view the system's modernization must develop in the direction of network computing and digital documents processing. The essential points of novel approaches are the following.

1. The unified forms of the secondary documents – information cards – are developed so that on the one hand to eliminate the duplication of the same fields in different cards and on the other hand to include more detailed financial data on the size and sources of funding and data on the life cycle of the intellectual property objects (patents, computer programs, databases, etc.).
2. The online mode of filling the new forms of information cards is provided for the authors of R&D reports and dissertations who are able to address the CITIS site on the Internet (www.rntd.citis.ru), click "the online form filling-in" and have the form on the screen of their computer. There are many conveniences supporting the online filling-in such as the formal verification of numerical fields, the enclosed lists of priority directions and critical technologies and the list of correct names of the organizations that were previously registered in the system. The user just has to click the name instead of keying it in.
3. The formation of digital full-text databases for all the arriving documents (reports and dissertations) with the effective means of full-text search and analysis. The digital documents are entered into the single electronic repository that allows four modes of documents entering: scanning and recognizing the paper documents; inputting the documents arriving on CDs; online arrivals entering; retroconversion of the documents stored on microfiches. Now, in accordance with the existing legal acts, the full-text documents arrive on paper and must be scanned and digitized before being PDF stored. The evident tendency is to pass on to digital input documents.
4. In order to introduce exclusively digital input forms of both the full-text and metadata documents it is necessary to implement an electronic signature technology and to make alterations in the legal acts (laws etc.) currently in force.

There are two kinds of subsystems envisaged in the technical assignment for the new system: those existing and being modernized and those newly designed and implemented.
The modernized ones are:

- the subsystem for reports and dissertations collecting, processing and registration;
- the digital documents repository and archiving subsystem;
- the search and retrieval subsystem;
- the abstract journals publishing subsystem.

Among the newly designed ones are:

- the system's common Internet portal subsystem;

- the R&D projects in progress monitoring and content analysis subsystem;
- the subsystem for interaction with international scientific and technical information systems;
- the subsystem for integration with other Russian state information systems on science and technology.

In the framework of the integral system a new complex of analytical and search instruments is to be designed using artificial intelligence technologies for linguistic text processing and semantic analysis, context and fuzzy search algorithms, subject area structuring, new knowledge and data-mining, antiplagiarism and experts activity support. This will allow to create analytical information not only about a separate scientific work but also about scientific trends, scientific groups and schools, the information for updating and systematizing scientific classification schemes. A linguistic support of these possibilities suggests that computer glossaries and dictionaries, thesauri, ontologies and classifiers should be developed and maintained.

Concluding remarks

During the next three-year stage of development it is supposed to implement the advantages of digital and network technologies and significantly improve the system's characteristics. The system is designed to provide a complete R&D documents collection, a fast access to full-text documents and relevant information. It will allow to monitor the situation in the sphere of R&D works and projects all over Russia, to support the federal level administrative decisions in the sphere of science and technology, to prognosticate its development, to improve the distribution of financial means for scientific R&D, to reduce the unjustified duplication and overlapping of R&D projects and dissertations.

References

1. Pavlov, L.P. The State and Development of the Russian Grey Literature Collection and Dissemination Centre. – "Interlending and Document Supply". - MCB Univ. Press. 1998, vol. 26, No. 4. Pp. 168-170.

2. Pavlov, L.P. Literatura gris rusa en un mundo digitalizado e informatizado. - "Ciencias de la Informacion." – La Habana: IDICT, 2002, agosto, v.33, N 2, pp. 39-42.

3. Pavlov, L.P. Legal Foundations of the Scientific and Technical Grey Literature Development in Russia. - "The Grey Journal". Internat.Journal on Grey Literature, Amsterdam. Spring 2007,vol. 3, N 1, pp.37-43.

4.. Starovoitov A.V., Bogdanov Yu. M., Bastrykin A. M., Pavlov L.P. The Grey System for Monitoring Self-Funded Research. - GL11 Conf. Proc.– GL-conf. series, N 11. 11th Internat. Conf on Grey Literature, 14-15 Dec.2009, Washington D.C.- Amsterdam:TextRelease, Dec. 2009.- 142 p. P.13.

5. Pavlov, L.P. The Commercialization of Research Findings Documented in Grey Literature. - Proc. 5th Internat. Conf. on Grey Literature: Grey Matters in the World of Networked Information. – 4-5 December 2003 Amsterdam/ GreyNet.- Amsterdam: TextRelease, January 2004.VI. – Pp.64 -68.

Open Is Not Enough: A case study on grey literature in an OAI environment *

Joachim Schöpfel, Isabelle Le Bescond, and Hélène Prost (France)

Abstract

The paper contributes to the discussion on the place of grey literature in institutional repositories and, vice versa, on the relevance of open archives for grey literature. Even in an open environment, grey literature needs specific attention and curation. Institutional repositories don't automatically provide a solution to all problems of grey literature. Our paper shows some scenarios of what could or should be done. The focus is on academic libraries. The paper is based on a review of international studies on grey literature in open archives. Empirical evidence is drawn from an audit of the French repository IRIS from the University of Lille 1 and from ongoing work on the development of this site.

The study includes a strategic analysis in a SWOT format with four scenarios. Based on this analysis, the paper provides a set of minimum requirements for grey items in institutional repositories concerning metadata, selection procedure, quality, collection management and deposit policy. The communication is meant to be helpful for the further development of institutional repositories and for special acquisition and deposit policies of academic libraries.

1. Introduction

Along with other documents and items, grey literature contributes to the success of institutional repositories. Its non-commercial and alternative nature puts grey literature in close proximity to the community-driven culture of open archives.

But does this mean that "grey literature is at home in open archives" (Luzi, 2010) and that it should be re-defined through this new vector of scientific communication?

After years of debate on open access and grey literature, the international conference GL12 at Prague offered two different perspectives. Marzi et al. (2010) stated that "open access is the key to knowledge" and that "web-base sharing facilities and distributed access to openly available information" are key features of grey literature. For Marzi and her colleagues, institutional repositories became the future of grey literature, and grey literature hardly exists without or beyond open access.

On the other hand, our own communication defined additional attributes for grey literature that are not necessarily linked to open access, such as intellectual property, quality and interest for collections. Institutional repositories are an interesting and important vector for dissemination of grey literature but they are not enough. Based on literature review and survey data, we made a proposal for a new definition of grey literature ("Prague definition") with four new essential attributes: "Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers i.e., where publishing is not the primary activity of the producing body" (Schöpfel, 2010).

Concerning open archives, we added that "institutional repositories have started to take over some of the traditional roles of library holdings. In terms of function, they bear some equivalency with grey literature itself, as their main role consists in dissemination and, to a lesser extent, preservation" (ibid). Institutional repositories are important for grey literature but they are not the only option, and they have to satisfy some minimum requirements in order to offer an adequate home for grey literature.

Institutional repositories and grey literature can become a fertile and profitable encounter for scientific communities. But open is not enough. Here are the reasons.

* First published in the GL13 Conference Proceedings, February 2012.

2. Background: A review of grey literature and institutional repositories

Institutional repositories (IR) became a significant channel of digital scientific communication.[1] Part of the open access movement and alongside with subject-based repositories, research repositories or national repository systems (Armbruster & Romary, 2009), they focus on "serving the interests of faculty – researchers and teachers - by collecting their intellectual outputs for long-term access, preservation and management" (Carr et al., 2008).

They can be seen as "tools (…) for collecting, storing and disseminating scholarly outputs within and without the institution" (Jain, 2011), as "a set of services (…) for the management and dissemination of digital materials created by the institution and its community members" (Lynch, 2003) or as an "organisational commitment to the stewardship of these digital materials" (ibid.).

One of their main characteristics is their great diversity. There is not *one* model but multiple possibilities, not *one* path but a multiplicity of options. Yet it is crucial for their success that the institution clearly defines the objective of its repository, in line with its own strategy and environment. "Each of the reasons for setting up a repository carries implications for the content, design and funding of a repository, and the institution needs to be clear about the implications of different roles for a repository, while being prepared to change or add roles as the scholarly communication environment develops" (Friend, 2011).

Institutional repositories have different policies, procedures, functionalities, services and metadata, they have different business models and funding strategies (Swan & Awre, 2006), and their content may include more than current output from faculty. Smith (2008) details a "wide variety of materials in digital form, such as research journal articles, preprints and postprints, digital versions of theses and dissertations, and administrative documents, course notes, or learning objects." Other repositories include datasets, multimedia or cultural and scientific heritage.

Of course, grey literature as unpublished, special or not-for-profit documents is part of the repositories' content. But what is its place in institutional repositories, and what is the relevance of institutional repositories for grey literature?

2.1. The place of grey literature in institutional repositories

Some empirical studies contribute to a realistic vision on grey literature in institutional repositories. Luzi et al. (2008) estimate the part of grey materials eligible for the institutional repository of the Italian National Research Council at about 1/3 of all items, even if not all of these documents are freely available.

In our survey on French repositories, grey literature represents 18% of all documents (Schöpfel & Prost, 2010). Another survey on Spanish repositories reveals that at least 23% of the deposited items with full-text are grey (Melero et al., 2009). Both studies confirm, too, that the number of grey documents in repositories is rapidly growing.

Vernooy-Gerritsen et al. (2009) report results from the EU-sponsored DRIVER project on institutional research repositories. They separate full-text records (33%) from metadata only records and records of non-textual and other materials; 62% of the full-text records are grey literature. This percentage corresponds to 20% of the whole content.

Most of all these grey items are theses, dissertations, proceedings, unpublished papers (working papers) or reports. Up to now, course material is less important.

The part of 20-30% of repository content is somewhat higher than the average percentage of grey literature in citation analyses (see Schöpfel & Farace, 2010).

So far, there is but little evidence on usage of grey items in institutional repositories. Yet, recent studies on access statistics suggest that downloads per item are often higher for unpublished theses or reports than for published articles (Schöpfel et al. 2009, see also Kroth et al. 2010).[2] One reason may be that these items can't be viewed elsewhere.

2.2. The relevance of institutional repositories for grey literature

To which extent are institutional repositories the place for grey literature? According to the information of the OpenDOAR directory of open archives, 82% of all institutional repositories contain grey literature.

Type of documents	Nb IR with these items	% of all IR (n=1,978)
Theses, dissertations	958	48%
Unpublished	616	31%
Proceedings	572	29%
Learning objects	245	12%
Special items	235	12%
Total	1,628	82%

Table 1: Grey items in institutional repositories (source: Open-DOAR, June 2011)

The OpenDOAR figures are comparable to results from France and Spain. In France, 94% institutional repositories hold grey documents while their part is significantly lower in subject-based repositories (37%) or national or research repositories (23%) (Schöpfel & Prost, 2010). In Spain, more than 80% repositories contain theses, and at least 60% have unpublished working papers and/or proceedings (Melero et al., 2009).

For some of this material, especially for specific types of unpublished items like slides, posters or other, supplementary material, it is surely true that "this is academic output that would not likely be otherwise captured and made freely available were it not for publication in an IR" (Kroth et al., 2010).

Some papers praise the impact of institutional repositories for grey literature. On the word of Luzi (2010), they provide "a natural home for GL" because they amplify its dissemination. Open access makes grey literature "less grey and more white" (Gelfand, 2004); the "distinction between GL and conventional literature is becoming increasingly blurred" (Luzi, 2010; see also Swan 2008 and 2011).

Yet, this "blurring" only applies to potential usage, not to value or quality. Banks (2005) believes that even if the hierarchy between grey and white may shift into a continuum of scholarship, this hierarchy will not completely disappear insofar institution and faculty generally prefer published and peer reviewed documents. A recent study on content recruitment and usage in an institutional repository confirms this belief (Connell, 2011).

2.3. Grey issues

Studies on grey literature in institutional repositories recurrently point out six critical aspects for the success and development of such initiatives.

Community: Describing a conference proceedings repository at Cornell, Rupp & LaFleur (2004) plead for "a specific workflow (…) for the identification and gathering of proceedings" that includes public relations, "one-to-one marketing" and communication with faculty to create awareness and get the documents from the author's desktop into the repository. Without community, no repository.[3]

Quality control: A repository that is "'all things to all people' lacks focus" (Westell, 2006). Specific action from the very beginning of the workflow is required to guarantee a minimum quality of content, data and services. Control procedures and workflow technology should ensure quality of item selection and overall project management (Luzi et al., 2004).

Metadata: Grey literature in institutional repositories has need of specific metadata for identification and bibliographic description. For instance, Ruggieri et al. (2009) propose a table with mandatory and optional metadata fields, including a note field, for conference papers, oral presentations, reports and in-house publications. Jeffery (2007) adds that "the syntax must be formal and precise; the semantics must be present, formal and precise (…); the relationships form a fully-connected graph; (…); the relationships require an annotation richer than the triples of RDF (…)." Yet, unfortunately the reality is that "current metadata

elements (of electronic theses and dissertations in IR) have a significant level of inconsistency and variation" (Park & Richard, 2011), and often "individual institutions (decide) locally how metadata elements should be defined (ibid).

Interoperability: Institutional repositories are hardly ever stand-alone systems. They should be interoperable or at least three reasons: maybe because their institution is part of a network (Dijk, 2007), maybe because they are connected and exchange data and items, maybe simply because the OAI initiative stipulates interoperability. Pejsova (2011) describes a national system for grey literature that is interoperable with local repositories for documents, metadata and workflow.

Integration: Some authors insist on the integration of institutional repositories and grey literature into current research information system (CRIS) infrastructure. "An institutional repository, being a central point within the organisation for literature and data, is a component of the integration of processes, which promises benefits both to the organisation itself and to the researchers within it" (Lambert et al., 2005).

New item formats: Jeffery (2007) calls for a linkage between CRIS and e-repositories for grey literature on the institutional level, and he suggests that they should be associated to repositories for research datasets and software, via the CRIS. More recently, Doorenbosch & Sierman (2011) focus on the changing nature of scholarly publications, e.g. enhanced publications with both documents and datasets, outline the challenge of these new items for long term preservation in institutional repositories, and suggest the creation of "collaborative virtual research environments are considered to be the new workspaces for researchers".

3. Case study: The IRIS audit – grey literature at home at Lille

The IRIS repository, hosted by the Lille 1 university, successor to Grisemine, the first French open archive for grey literature. Its development and usage have been presented at the GL5 and GL12 conferences (Claerebout, 2003; Prost et al., 2010). The following case study provides a short overview on the Grisemine/IRIS history and illustrates some conditions that are favourable or not for the deposit and dissemination of grey literature in institutional repositories.

3.1. General remarks

When Grisemine was launched in 2001, it was one of the first open archives in France, a pioneer especially in the academic sector. Its notoriety and popularity among academic librarians was immediate and without doubt superior to its real impact on scientific communication.

Since 2001, Grisemine underwent deep changes. This "Grisemine/IRIS decade" demonstrates the coming out of the hybrid digital library with service marketing rather than collection building. Nearly all has changed – the name, software, architecture and workflow, content, strategy, policy and institutional positioning.

The story of Grisemine/IRIS is not over. In fact, it just began, again. But which may seem, ex post, logical and necessary often was trial and error, searching for opportunities, benchmarking, exploration and adaptation to a moving context.

3.2. Rise and decline of Grisemine (2001-2005)

Grisemine's purpose was to collect, preserve and disseminate French[4] grey literature, such as theses and dissertations, communications, notes, working papers, preprints, exam topics or educational programs. Grisemine was developed with the CinDoc electronic content management software (Cincom). Its workflow was compliant with the Dublin Core metadata standard and the MARC format.

Even as a prototype, the Grisemine project was technically viable, except for the technical maintenance and development of the CinDoc software. But it had no real institutional recognition, was a "librarians' toy" rather than a labelled, validated and accepted repository for the scientific community. Yet, its content (1,300 documents in late 2005) was widely consulted, in particular from French-speaking countries.

It became obvious, too, that the initial goal – a deposit for all French grey literature – was too ambitious and disproportionate to the allocated resources.

3.3. From Grisemine to IRIS (2006-2010)

In 2006, the French government published a decree on the processing, preservation and dissemination of electronic PhD theses and launched a national network for ETDs called STAR. Grisemine was not able to support the new workflow. For this and other reasons mentioned above (maintenance), the Lille library team considered Grisemine as a technical and documentary dead-end. The next four years were a period of transition.

The most important decision was to migrate from CinDoc to DSpace, and then make the system dialogue with STAR. The migration was operational in 2007. With the migration, Grisemine became IRIS.

Why DSpace? At the time the Lille team took the decision to migrate (2004-2006), DSpace was the most common software for open archives, and it was easy to install. Yet, DSpace is designed for self-deposits, not for an encyclopaedic-like collection (scientific heritage) or an institutional and/or national workflow (theses). Without a dedicated information technology (IT) staff, the Lille library decided to maintain DSpace at best until the new ORI-OAI software became available[5]. "At best" meant keeping the archive alive, continued uploads but no development. For instance, an early project to separate PhD theses and scientific heritage was put on ice.

The deposit of e-theses became mandatory on the Lille 1 campus in 2008, because of STAR. IRIS was able to provide an operating OAI platform for their dissemination but didn't offer a solution for their management or preservation. The open dissemination of Lille ETDs became the main function of the IRIS repository. In December 2010, IRIS had 625 theses and 711 other documents. Their long-term preservation is supported by the academic data centre CINES at Montpellier[6].

With the move from Grisemine to IRIS, the site abandoned its initial strategy as an open repository for French grey literature. The self-deposit of grey items ceased completely. Instead, the library team made another use of the IRIS platform and developed, together with a historical research centre and the academic digitization centre at Lille, a digital library with a collection of copyright cleared documents (articles, papers, books) on the history of sciences. Alongside with the PhD theses, this heritage collection was made freely available on the IRIS platform and is very appreciated by the scientists.

When the university decided the mandatory deposit of e-theses in 2008, it also acknowledged IRIS as the official Lille 1 institutional repository. Yet, this decision was not accompanied or followed by a mandatory policy for the whole scientific production of the faculty. Except some professorial habilitation theses and learning objects, IRIS never received any self-deposits from Lille researchers.

3.4. Rebirth (2010-2011)

At the end of the first decade, the strategic positioning of IRIS was atypical and confusing. The university administration considered IRIS as the official institutional repository. Yet, there was no promotion, communication, incentives or mandate, and the only open archive with a significant number of self-deposits from Lille 1 faculty was (and always is) the French national research repository HAL with 16,143 items.[7]

The library team regarded IRIS as a digital library, more like GALLICA or PERSEE than ArXiv or HAL, yet used the IRIS server for the dissemination of PhD theses, a service usually considered to be a key element of academic institutional repositories, and made some tests with other scientific output from Lille faculty, especially in the context of an emerging learning centre project.

In 2010, with the installation of the ORI-OAI system the Lille 1 repository took a new start. Why ORI-OAI? At least for four reasons: compliance with French metadata standards for theses (TEF) and learning objects (SupLOMFR), interoperability with the nationwide

infrastructure for ETDs (STAR) and the national research repository HAL, a French community of software developers and end-users, quality of development and product.
Today Lille 1 hosts a composite repository with two systems accessible through two different interfaces:

- ORI-thèses with theses, habilitations and learning objects.[8]
- IRIS with the collection of history of sciences.[9]

In fact, IRIS became a digital library without input from current scientific production.
A third platform for the self-deposit of scientific production (pre- and post-prints, communications, reports…) is under construction, on the model of the Toulouse OATAO[10] repository or the Luttich ORBi[11] site, and will be launched in 2012 probably with a new name.

3.5. Concluding remarks

As we said above, the story of Grisemine/IRIS is not over and it may be premature to debrief. Yet, we tried to highlight some main characteristics of this project and then to identify the factors in favour of grey literature and success.
The development of the Lille 1 repository was non-linear, dependent on the evolving local and national context, on technology (software) and standards. The library team's quest for legitimacy was complicated by the pluridisciplinarity of their academic community and by the fact that in France, the open archives for scientific information were initially hosted and managed by the public research organisations (CNRS, INRA, IFREMER…).
On the other hand, the national infrastructure for electronic theses (STAR system with TEF metadata standard) and the library's experience with preservation and dissemination of cultural and scientific heritage items – a traditional library function - facilitated the legitimacy and positioning of the project.
So which were the critical key elements for success or failure? Briefly:[12]

- Institutional support and recognition of the project team and the repository.
- Institutional strategy and policy in the domain of open archives and deposit mandate.
- Human resources with sufficient IT and LIS capacities.
- Metadata standard(s) for a careful and precise bibliographic description of the deposited content.
- Software fitting with local needs and IT environment as well as with national infrastructure and standards.
- A solution for perennial preservation of deposits (at least for the theses).
- Added value services for legal aspects and usage statistics.
- Knowledge of the scientific community's information needs and behaviours, and integration into the larger academy.

The Grisemine/IRIS case shows also a close link between grey literature typology, IT solutions (software) and workflow features. The repository must cope with specific conditions, such as (for the Lille repository) the national STAR system for theses or the digital university environment (UNT) for the learning objects. The need to align deposit with existing workflows was highlighted by Westell (2006) and Troll Covey (2011). This, together with the different software solutions, argues for a differential approach to grey literature in institutional repositories. Some grey documents may be at home in some open archives, while others in different ones.

4. SWOT analysis: Grey literature in institutional repositories

Based on the review of literature and standards and including the IRIS experience, our evaluative synthesis will take the form of a strategic SWOT diagnostic, keeping apart internal and external factors that are favourable or unfavourable for grey literature in institutional repositories. However, our analysis does not take into account more general aspects that are not directly related to grey literature (for instance, such as Pinto & Fernandes, 2011).

4.1. Strengths

The internal factors in favour of grey literature in institutional repository models are:

1. Grey literature amplifies the content of institutional repositories.
2. Free availability, dissemination, visibility and referencing act as incentives for grey deposits.
3. What's more, relatively high usage of unpublished items may also act as an incentive for grey deposits.
4. Institutional repositories guarantee more security and long-term accessibility of unpublished material than a personal web site.
5. Compared to published articles, there are fewer problems with copyright for grey literature.

4.2. Weaknesses

The internal factors unfavourable for grey literature in institutional repositories are:

1. The bibliographic control of grey literature, especially of conferences and reports, remains often mediocre or poor because of flawed or incomplete metadata format (non qualified Dublin Core).
2. Most often, the hosting institution doesn't provide any solution for the digital curation of metadata.
3. Deposit is time consuming.
4. Deposit of grey literature needs, more than published documents, incentives and support from institution. This support may be missing.
5. Without institutional support or incentives, self-deposits will not have the same quality as a library collection.

4.3. Opportunities

The external factors in favour of grey literature in institutional repository models are:

1. Universities need a solution for the processing, disseminating and archiving of electronic theses and dissertations (ETD). Institutional repositories offer an interesting solution and may at least be an element in the global academic information system for ETD.
2. Institutions want control on research output and content, and this includes unpublished documents.
3. Institutions want to improve presence and impact on the web. Grey literature in repositories adds to both, due to broader dissemination and increased use of grey items, increasing prestige and visibility for the institution.
4. The open access initiative is not limited to published documents.
5. The evolution from "collection development" to "content recruitment" in academic libraries may act in favour of deposit of grey literature in institutional repositories.

4.4. Threats

The external factors unfavourable for grey literature in institutional repositories are:

1. Funding and evaluation agencies put priority on published documents (articles, books) and at least partially neglect grey items. Grey literature is not indexed in the scientometric databases Web of Science and SCOPUS.
2. If institutions introduce self-archiving mandates in order to generate content, researchers may react negatively to any suggestion of compulsion. Most faculties do not respond to the invitation to "add stuff to the IR" (Jain 2011). Another side-effect is the creation of metadata only records, without full-text. This should be limited to published documents with copyright problems but it isn't.[13]
3. Alternative models, e.g. generating content through deposit by publishers (PEER project) will not impact grey items.
4. Open access through institutional repositories requires funding from particular institutions to set up and maintain a repository (Friend 2011). Poor knowledge on grey

literature will make it more difficult to sustain continuous support and commitment from the management and academic staff.

5. A significant part of the scientific community lacks awareness of open access and grey literature.

5. Findings based on four scenarios

Are institutional repositories the future of grey literature? Maybe. But because of the great variety of institutional repositories, we can distinguish at least four different scenarios.

Jain (2011) makes some recommendations for the development of institutional repositories, in particular, promotion and publicity to the faculty, provision of clear policies on ownership, contents, quality and copyright, and an adequate provision of resources. This is in line with the IRIS audit and applies to all scenarios. Therefore, our description is limited to specific criteria for grey literature within this environment.

The differences are with mandatory deposit, strategic vision, services, selection procedure, quality issues, collection management and metadata. Our description is partly based on studies on objectives and business models of institutional repositories (Friend, 2011; Swan & Awre, 2006). We don't describe real cases but potential homes – a kind of ideal archetypes of institutional repositories. The reality will be more complex and composite.

5.1. Scenario 1 – Publishing grey literature

In the first scenario, the institutional repository serves essentially the initial function of open archives, e.g. communication and publishing of scientific papers. Focus is laid on rapid and direct access to full-text, for the scientific community. For grey literature, the strategy is to become less greyish and more white, through institutional digital publishing outside of usual sales channels.

The strategic objective by the institution may be twofold (cf. Friend, 2011):

- "To increase the impact of particular research or teaching programmes through exposure of publications and other outputs on open access.
- To reduce the cost and increase the benefits from the dissemination of the institution's research and teaching outputs."

The most appropriate business model for repository provision and preservation will be institutionally-supported, perhaps with a contribution by community (learned societies).

Selection procedures for a minimum content and formal quality level (through validation or "labelling") probably will be more important than mandatory issues. Self-deposit of full-text (preprints, postprints but also conference proceedings, unpublished reports and papers…) and institutional workflows for electronic theses, perhaps also for master and habilitation theses, in-house collections of working papers or reports are essential for content recruitment while mandatory deposit policy or incentives are not.

Also, metadata are critical (only) insofar they facilitate content retrieval and access. This means that they are probably of mediocre quality and not very specific for different types of documents, except for ETD.

The primary function of this repository is communication and access to the full-text, via search engines and/or the repository's search and browse interface. The key elements are a high rate of full-text, worthy scientific content, and unrestricted access, followed by a high and representative number of deposits. Other services may be less crucial but would add value to the site:

- usage statistics services,
- preservation services,
- publishing services.

5.2. Scenario 2 – Special items container

In the second scenario, the institutional repository is a container or storehouse for all kind of material produced by faculty, staff and students. In this container, ETD, reports and

conference proceedings stand next to images, learning objects, articles, datasets, presentations, posters etc.
The focus is laid on availability and visibility of all kind of materials, "institutional stuff", rather than on selection of scientific relevant results. Quality control through validation or labelling is not an issue.
The strategic objective may be "to collect together all the publications and other research and teaching outputs as a permanent record of the institution's achievements but without any specific use in mind" (Friend, 2011).
Again, the appropriate business model for repository provision and preservation is institutional support. The institution may also decide to establish a mandatory deposit, and/or incentives for self-deposit.
The underlying idea is to "dig out" hidden material, find a solution for digital dissemination and preservation, together with other published or unpublished documents.
As for quality control or editorship, metadata probably are not an important issue. Most likely, services will be limited to preservation, publishing, resource discovery and perhaps research assessment and monitoring. It is also possible to add social indexing and data mining. There is no clear vision on collection and acquisition. But the most promising perspective may be the linking of the deposits to research data.

5.3. Scenario 3 – Scientific heritage

The third scenario the institutional repository is a showcase for the past and present scientific production, with grey literature alongside with published documents and other material.
Again, the strategic vision will be "to collect together all the publications and other research and teaching outputs as a permanent record of the institution's achievements but without any specific use in mind" (Friend, 2011). The difference with scenario 2 is the heritage character of the collection, the inclusion of older material in the public domain.
But there may (also) be other motivations:

- "To increase the impact of particular research or teaching programmes through exposure of publications and other outputs on open access.
- To make a contribution to the world-wide movement for open access to publicly-funded research" (ibid.).

The definition of an acquisition or content recruitment policy is crucial, together with an institutional strategy for the digitization of older, copyright cleared material (theses, journals, books, papers, images, maps…). This may imply a more thoroughly prepared and pondered indexing and metadata policy. The outcome may be 100% access to full text, as for the IRIS repository.
The appropriate business model is institutional support. But there may be other resources, public funding for scientific heritage or thematic or special collections. For this specific case, it may be possible to experience a subscription-supported model, appropriate for access and authentication, preservation and resource discovery services.
Also, the local presence of a digitisation centre may allow those repositories to populate content more rapidly, especially grey literature, and to attract usage (Westell, 2006).
The underlying idea is digital preservation of heritage collection, together with making these collections available to scientists, students and all interested people. This may be complementary to publishers' backfiles.
This scenario is probably the closest scenario to traditional library collection building, with issues such as quality, indexing, classification etc. Evaluation, scientometrics etc. may be less important, at least not in the heart of the project.

5.4. Scenario 4 – Institutional deposit

The last scenario for grey literature in institutional repositories is mandatory institutional or self-deposit in the way it is promoted by Stevan Harnad: green road (self-deposit) to free

online full-text access to peer-reviewed literature, through an explicit and institutional mandatory policy in order to obtain commitment by close to 100% of the authors.

This scenario is meant to demonstrate the value of the institution itself through a kind of quasi-legal deposit showcase, to facilitate control over scientific production and evaluation procedures, and corresponds to one or more institutional strategies, e.g.

- "To report the publications and other research and teaching outputs to funding agencies in support of new grant applications.
- To report the publications and other research and teaching outputs to funding agencies as part of an audit of expenditure.
- To demonstrate to governments or taxpayers the impact of the institution outside its walls (a purpose which will require the compilation of metrics).
- To increase the impact of individual members of the institution's staff through the exposure to potential academic and commercial users of the individual's publications and other outputs on open access" (Friend, 2011).

The business model will surely be institutionally-supported and may include services such as usage statistics, research assessment and monitoring, bridging and mapping, and technology transfer/business advice. Also, a connection to a current research information system (CRIS) should be possible.

The impact on grey literature in this environment is triple:

Peer-reviewed publications will play a major role in this environment, and in comparison, grey literature will be less valued or appreciated. This may have a negative impact on metadata.

The institutional policy of mandatory deposit generate a relatively high rate of metadata only records without access to full-text because of embargo, sensitive content, missing authorization by co-authors etc. Paradoxically, this "collateral damage" also impacts grey literature (see above, footnotes 7 and 12). Only 40% of the HAL grey literature records are with full text.

The number of grey documents will be significant but more or less limited to specific categories evaluated by agencies, such as theses and dissertations, conference proceedings and project reports.

The main interest of these repositories is not collection building but evaluation. Insofar grey literature enters evaluation procedures it will be valued and welcome in this environment.

6. Results and concluding remarks

Our paper started with Luzi's (2010) statement that "grey literature is at home in open archives". This may be right but as we tried to demonstrate, open archives not only offer one but at least four different homes that may be complementary, at least to some extent.

Mapped on two dimensions, policy (evaluation vs. communication) and quality (library vs. container), the four options clearly occupy different positions (figure 1).

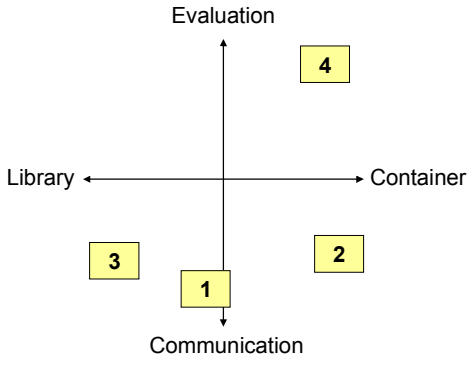


Figure 1: A map of four scenarios for institutional repositories with grey literature

In scenario 1, the political priority is laid on communication of research results, full-text, community, scientific value. Grey literature is part of the content insofar the depositing authors consider it worthy enough for direct communication and preservation. But there is no real control or selection.

In scenario 2, the main objective is the container function, the deposit of all materials produced by faculty, students and staff. Again, the institutional policy is communication-centred but without selection or validation criteria. Grey literature has its home here – in a (too) large sense and together with a lot of other stuff.

Selection or validation criteria are introduced by the 3rd scenario. Here the institution applies a policy of showcase and scientific heritage, most likely accompanied by digitization programs. The place of grey literature depends on the institution's acquisition policy and digitization program.

The 4th scenario reflects the institutional policy in favour of evaluation and ranking. Full-text and communication are secondary goals while metadata and a minimum quality control are necessary. Deposit of grey literature will be welcome insofar it enters evaluation.

Now, which is the most adequate option for grey literature? The response depends on institutional policy, library goals and professional viewpoint. For the scientific community, end-user and consumer of scientific information, perennial open access to validated items in full-text format is priority. This priority implies at least five minimum requirements:

Access to full-text. Open archives with metadata only records are like libraries with empty shelves.

Quality through selection, validation and/or labelling. Even without peer-review or other, web-based reviewing procedures, grey deposits should meet with some basic quality criteria. Incite deposit of all kind of uninteresting stuff is like keeping waste paper on the desktop. Self-deposit is not collection building.

Openness without restriction and/or embargo. Confidential, classified or non-copyright cleared material should not be part of open archives but should be managed via catalogues, databases or other systems.

Metadata quality. Repositories should guarantee a minimum level of metadata quality, e.g. compliance with standards and curation. This requirement is necessary for information retrieval, interoperability and the semantic web.

Long-term conservation. Institutional repositories should offer a solution for the ephemeral nature of grey literature, via a clear statement on and investment in perennial content preservation, if necessary also via outsourcing or "in the clouds".

For the scientific community, the best option for grey literature may be a mix of scenarios 1 "publishing grey literature" and 3 "scientific heritage". Other elements will add value (standard format and metadata, usage statistics, discovery functions, scientometrics) or increase sustainability (institutional support, integration in research community, promotion and communication, interoperability). But they are not specific to grey literature.

We didn't speak about format and legal matters; yet, they may be critical matters for the future of repositories. With the words of Swan (2011), "we (can't) relax (and) watch repositories fill with articles and datasets". Or as Anderson (2011) put it, "accessibility is not access."

The IRIS case should raise awareness that the same solution may not be appropriate to all kind of grey literature and disciplines and that the system should be evolutionary and flexible enough to easily adapt to and keep up with new conditions and opportunities.

A last and rather paradox remark: the success of institutional repositories may become a problem for grey literature, especially when the institution implements a mandatory deposit policy that gives priority to evaluation and control and not to publishing and communication. Anna Clements, a data manager from St Andrews University, described the problem some time ago on the JISC-Repositories listserv: libraries create institutional repositories with full-text or full objects as the main content, and they are then asked by the institution to look at hosting citations without full-text as well.

A library with empty book shelves may be interesting to research managers but not for scientists. In this case, grey literature would definitively not be at home in institutional repositories. Open is not enough.

6. References

K. Anderson (2011). `Does Access Create New Types of Scarcity?'. *The Scholarly Kitchen* Aug 31.

M. A. Banks (2005). `Towards a Continuum of Scholarship: The Eventual Collapse of the Distinction Between Grey and non-Grey Literature?'. In *Seventh International Conference on Grey Literature: Open Access to Grey Resources, Nancy, 5-6 December 2005*.

L. Carr, et al. (2008). `Institutional Repository Checklist for Serving Institutional Management'. In *Third International Conference on Open Repositories 2008, 1-4 April 2008, Southampton, United Kingdom*.

M.-F. Claerebout (2003). `Grisemine, a digital library of grey university literature'. In *Fifth International Conference on Grey Literature: Grey Matters in the World of Networked Information, 4-5 December 2003*.

T. H. Connell (2011). `The Use of Institutional Repositories: The Ohio State University Experience'. *College & Research Libraries* **72**(3):253-275.

E. Dijk (2007). `Accessing grey literature in an integrated environment of scientific research information'. In *Ninth International Conference on Grey Literature: Grey Foundations in Information Landscape, Antwerp, 10-11 December 2007*.

P. Doorenbosch & B. Sierman (2011). `Institutional Repositories, Long Term Preservation and the changing nature of Scholarly Publications'. *Journal of Digital Information* **12**(2).

F. Friend (2011). `Open Access Business Models for Research Funders and Universities'. Tech. rep., Knowledge Exchange, Copenhagen.

J. Gelfand (2004). `"Knock, Knock:" Are Institutional Repositories a Home for Grey Literature?'. In *Sixth International Conference on Grey Literature: Work on Grey in Progress, New York, 6-7 December 2004*.

P. Jain (2011). `New trends and future applications/directions of institutional repositories in academic institutions'. *Library Review* **60**(2):125-141.

K. G. Jeffery (2007). `Greyscape'. In *Ninth International Conference on Grey Literature: Grey Foundations in Information Landscape, Antwerp, 10-11 December 2007*.

P. J. Kroth, et al. (2010). `Institutional Repository Access Patterns of Nontraditionally Published Academic Content: What Types of Content Are Accessed the Most?'. *Journal of Electronic Resources in Medical Libraries* **7**(3):189-195.

S. Lambert, et al. (2005). `Grey literature, institutional repositories and the organisational context'. In *Seventh International Conference on Grey Literature: Open Access to Grey Resources, Nancy, 5-6 December 2005*.

D. Luzi, et al. (2004). `The integration of GL documents with a research information system on occupation safety and health'. In *Sixth International Conference on Grey Literature: Work on Grey in Progress, New York, 6-7 December 2004*.

D. Luzi, et al. (2008). `Towards an Institutional Repository of the Italian National Research Council: A survey on Open Access experiences'. In *Tenth International Conference on Grey Literature: Designing the Grey Grid for Information Society, 8-9 December 2008*.

D. Luzi (2010). `Grey Documents in Open Archives'. *The Grey Journal* **6**(3):137-144.

C. A. Lynch (2003). `Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age'. Tech. Rep. 226, ARL Association of Research Libraries.

C. Marzi, et al. (2010). `A terminology-based re-definition of grey literature'. In *Twelfth International Conference on Grey Literature: Transparency in Grey Literature. Grey Tech Approaches to High Tech Issues. Prague, 6-7 December 2010*.

R. Melero, et al. (2009). `The situation of open access Institutional repositories in Spain: 2009 report'. *Information Research* **14**(4).

B. Mukherjee & M. Nazim (2011). `Open access institutional archives: a quantitative study (2006-2010)'. *DESIDOC Journal of Library & Information Technology* **31**(4):317-324.

E. G. Park & M. Richard (2011). `Metadata assessment in e-theses and dissertations of Canadian institutional repositories'. *The Electronic Library* **29**(3):394-407.

P. Pejsova (2011). `Czech National Repository of Grey Literature'. *The Grey Journal* **6**(3):117-122.

M. J. Pinto & S. Fernandes (2011). `Gaining a sustainable IR: thinking SWOT'. In *QQML2011: 3rd International Conference on Qualitative and Quantitative Methods in Libraries (DLMC 2011: Digital Library Multi-Conference), Athens, Greece, 24-27 May, 2011*.

H. Prost, et al. (2010). `Usage assessment of an institutional repository: a case study'. *Twelfth International Conference on Grey Literature: Transparency in Grey Literature. Grey Tech Approaches to High Tech Issues. Prague, 6-7 December 2010*.

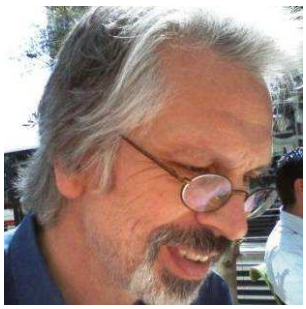
R. Ruggieri, et al. (2009). `From CNR Annual report to an Institutional repository: Results from a survey'. In *Eleventh International Conference on Grey Literature: The Grey Mosaic, Piecing It All Together, Washington DC, 14-15 December 2009*.

N. Rupp & L. J. LaFleur (2004). `Making Grey Literature Available through Institutional Repositories'. In *Sixth International Conference on Grey Literature: Work on Grey in Progress, New York, 6-7 December 2004*.
J. Schöpfel, et al. (2009). `Usage of grey literature in open archives'. In *Eleventh International Conference on Grey Literature: The Grey Mosaic: Piecing It All Together. Washington D.C., 14-15 December 2009*.
J. Schöpfel (2010). `Towards a Prague Definition of Grey Literature'. In *Twelfth International Conference on Grey Literature: Transparency in Grey Literature. Grey Tech Approaches to High Tech Issues. Prague, 6-7 December 2010.*
J. Schöpfel & D. J. Farace (2010). `Grey Literature'. In M. J. Bates & M. N. Maack (eds.), *Encyclopedia of Library and Information Sciences, Third Edition*, pp. 2029-2039. CRC Press, London.
J. Schöpfel & H. Prost (2010). `Développement et Usage des Archives Ouvertes en France. Rapport. 1e partie: Développement'. Tech. rep., Université Charles-de-Gaulle Lille 3.
J. Schöpfel & H. Prost (2011). `IRIS. Etat des lieux et perspectives. Rapport d'audit'. Tech. rep., University of Lille, Villeneuve d'Ascq.
K. Smith (2008). `Institutional Repositories and E-Journal Archiving: What Are We Learning?'. *Journal of Electronic Publishing* **11**(1).
A. Swan & C. Awre (2006). `Linking UK Repositories: Technical & Organisational Models to Support User-Oriented Services Across Institutional & Other Digital Repositories'. Tech. rep., JISC, London.
A. Swan (2008). `Study on the availability of UK academic "grey literature" to UK SMEs: Report to the JISC Scholarly Communications Group'. Tech. rep., JISC, London.
A. Swan (2011). `Institutional repositories - now and next'. In P. Dale, J. Beard, & M. Holland (eds.), *University Libraries and Digital Learning Environments*, pp. 119-134. Ashgate Publishing, Farnham.
D. Troll Covey (2011). `Recruiting Content for the Institutional Repository: The Barriers Exceed the Benefits'. *Journal of Digital Information* **12**(3).
M. Vernooy-Gerritsen, et al. (2009). `Three Perspectives on the Evolving Infrastructure of Institutional Research Repositories in Europe'. *Ariadne* (59).
M. Westell (2006). `Institutional repositories: proposed indicators of success'. *Library Hi Tech* **24**(2):211-226.
Websites visited between September-November 2011.

[1] See the quantitative study from Mukherjee & Nazim (2011).
[2] For instance, the 2010 annual report of the French Research Institute for the Exploitation of the Sea shows that the average usage for theses in their IR is 4x higher than for published articles, see http://wwz.ifremer.fr/institut/L-institut/Documents-de-reference/Rapports-Annuels
[3] See also the disillusioning survey from Seaman (2011).
[4] French means: edited in France and/or in French language.
[5] A document management system compliant with OAI-PMH, designed for the publishing, sharing and dissemination of academic digital resources and supported by the French Ministry of Higher Education http://wiki.ori-oai.org
[6] Preservation of ETDs via STAR, preservation of other deposits on a contractual basis.
[7] But only 13% of these items have full-text, the rest are metadata only records (12 October 2011).
[8] http://ori.univ-lille1.fr
[9] https://iris.univ-lille1.fr
[10] http://oatao.univ-toulouse.fr/
[11] http://orbi.ulg.ac.be/
[12] See also Westell (2006).
[13] For instance, only 45% of the deposited working papers, conferences and ETD in the Belgian ORBi repository provide access to the full-text.

On the News Front

New Director of the National Digitization Centre for PhD Theses



Dr. Joachim Schöpfel was appointed director of the National Digitization Centre for PhD Theses (ANRT) in Lille, France. The Centre founded in 1971 by the Ministry of Education reproduces French PhD theses in various formats (print, microfiche and digital), thus ensuring their dissemination via institutional and commercial channels in France and abroad.

The ANRT catalogue contains some 200,000 titles that can be ordered by academic libraries and 7,000 PhD theses in the social sciences and humanities that can be ordered in print format for private use ("Thèses à la Carte"). ANRT also digitizes scientific collections for repositories and e-libraries, journal back files, as well as other types of academic publications. ANRT is located on the campus of the University of Lille 3, where Schöpfel continues as senior lecturer and head of the library and information science department.

As new ANRT Director, Schöpfel's vision is to develop a 'competence centre', where technology, education, and research in digitization and e-publishing converge and are further exchanged with other national and international partners.

http://www.diffusiontheses.fr/

Atelier National
de Reproduction des Thèses

Fourteenth International Conference on Grey Literature

Tracking Innovation through Grey Literature

National Research Council, Rome, Italy 29-30 November 2012

Hosted by the Central Library "Guglielmo Marconi", IRPPS Rome, and ISTI Pisa

Conference Program

DAY ONE

9:00-10:30

OPENING SESSION

Aula CNR, National Research Council

Chair, *Stefania Biagioni, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", ISTI-CNR, Italy*

Welcome Address *Prof. Luigi Nicolais, President, Consiglio Nazionale delle Ricerche, CNR, Italy*

Opening Address *Dr. Carlos Morais-Pires, European Commission, Brussels, Belgium*

Keynote Address *Dr. Jan Brase, Head DOI Registration Agency; Technische Informationsbibliothek, Germany*

Rejoinder *Dr.ssa Luisa De Biagi and Flavia Cancedda, CNR Central Library 'G. Marconi', Italy*

11:00-13:00

SESSION ONE – TRACING THE RESEARCH LIFE CYCLE

Chair, *Donatella Castelli, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", ISTI-CNR, Italy*

Customized OAI-ORE and OAI-PMH Exports of Compound Objects for the Fedora Repository
Alessia Bardi, Sandro La Bruzzo, and Paolo Manghi; Istituto di Scienza e Tecnologie dell'Informazione, CNR, Italy

Grey literature in the digital culture and practices of the new global scholar: the case of molecular biology
Chérifa Boukacem-Zeghmouri; Département d'Informatique de l'Université Claude Bernard Lyon 1, France

Grey in the Innovation Process *Keith G Jeffery, STFC-Rutherford Appleton Laboratory, United Kingdom and Anne Asserson, University of Bergen, Norway*

Characteristics and use of grey literature in scientific journal articles of Algerian University of Science and Technology teachers and researchers in STM fields
Lydia Chalabi, Research Center of Scientific and Technical Information, CERIST, Algeria

Usage of grey literature in the research life cycle in Korea
Seon-Hee Lee and Hye-Sun Kim; Korea Institute of Science and Technology Information, KISTI, Korea

14:00-15:30

SESSION TWO – TRACKING METHODS FOR GREY LITERATURE

Chair, *June Crowe, Information International Associates, IIA, USA*

What goes up must come down: Publications from developing countries in the Aquatic Commons
Luigi Baldassari and Armand Gribling; Fisheries & Aquaculture Branch Library, Food and Agriculture Organization of the United Nations (FAO), Italy

Data sharing in environmental sciences: A survey of CNR researchers *Daniela Luzi; Institute for Research on Population and Social Policies, IRPPS-CNR, Roberta Ruggieri, Senato della Repubblica, and Stefania Biagioni, Institute of Information Science and Technologies, ISTI-CNR, Italy*

Tracking the Influence of Grey Literature in Public Policy Contexts: The Necessity and Benefit of Interdisciplinary Research *Bertrum H. MacDonald, Elizabeth M. De Santo, Kevin Quigley, Suzuette S. Soomai, and Peter G. Wells; Dalhousie University, Canada*

Grey communities: An empirical study on databases and repositories *Hélène Prost; Institute for Scientific and Technical Information, INIST-CNRS and Joachim Schöpfel, Charles de Gaulle University Lille 3, France*

16:00-17:30

INTRODUCTION TO CONFERENCE POSTERS

Chair, *Dominic Farace, Grey Literature Network Service, GreyNet, Netherlands*

CONFERENCE RECEPTION **17:30-19:00**

Special Presentation *by Prof. Emeritus Paul Sturges, United Kingdom*

GL14 Program and Conference Bureau
TextRelease

Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands
www.textrelease.com · conference@textrelease.com
Tel/Fax +31-20-331.2420

Fourteenth International Conference on Grey Literature

Tracking Innovation through Grey Literature

National Research Council, Rome, Italy 29-30 November 2012

Hosted by the Central Library "Guglielmo Marconi", IRPPS Rome, and ISTI Pisa

Conference Program

DAY TWO

9:00-11:00

POSTER SESSION AND SPONSOR SHOWCASE **(LIST OF POSTERS FORTHCOMING)**

The entrance hallway and rooms adjoining the Main Aula will accommodate up to 25 conference posters. For information on the availability of space and for further guidelines on posters, please contact TextRelease.

11:00-12:30

SESSION THREE – ADAPTING NEW TECHNOLOGIES

Chair, *Daniela Luzi, Institute for Research on Population and Social Policies, IRPPS-CNR, Italy*

An Environment Supporting the Production of Live Research Objects *Massimiliano Assante, Leonardo Candela, Donatella Castelli, and Pasquale Pagano; Istituto di Scienza e Tecnologie dell'Informazione – CNR, Italy*

Creating and Assessing a Subject-based Blog for Current Awareness within a Cancer Care Environment *Yongtao Lin and Marcus Vaska; Health Information Network Calgary, University of Calgary, Canada*

UFMG TUBE: Plural knowledge in connection
Maria Aparecida Moura; Federal University of Minas Gerais, Brazil

A funder repository of heterogeneous grey literature material with advanced user interface and presentation features *Ioanna-Ourania Stathopoulou, Nikos Houssos, Panagiotis Stathopoulos, Despina Hardouveli, Alexandra Roubani, Ioanna Sarantopoulou, Alexandros Soumplis, Chrysostomos Nanakos; National Documentation Centre; National Hellenic Research Foundation, Greece*

13:30-15:30

SESSION FOUR – REPURPOSING GREY LITERATURE

Chair, *Bertrum H. MacDonald, Dalhousie University, Canada*

Working for an open e-publishing service to improve grey literature editorial quality *Rosa Di Cesare, Marianna Nobile; Institute for Research on Population and Social Policies, IRPPS and Silvia Giannini; Institute of Information Science and Technology, ISTI, Italy*

Innovation and the Challenge of Knowledge Transfer: Establishing a Grey Lit Science Repository in Iraq *Donald Hagen, National Technical Information Service, NTIS; Gail Hodge, Information International Associates, USA*

Centralised National Corpus of Electronic Theses and Dissertations *Julius Kravjar; Slovak Centre of Scientific and Technical information, CVTISR, Slovakia*

Grey literature in Australian education *Gerald White, Julian Thomas, Paul Weldon, Amanda Lawrence and Helen Galatis; Australian Council for Educational Research, Australia*

Research Life Cycle: Exploring Credibility of Metrics and Value in a New Era of eScholarship that Supports Grey Literature *Julia Gelfand, University of California, Irvine (UCI) and Anthony Lin, Irvine Valley College, USA*

15:45-16:30

CLOSING SESSION – REPORT CHAIRPERSONS, CONFERENCE HANDOFF, FAREWELL

Chair, *Stefania Biagioni, ISTI-CNR, Italy and Dominic Farace, GreyNet International, Netherlands*

POST CONFERENCE TOUR - CNR CENTRAL LIBRARY **16:30-17:30**

Guided *by Flavia Cancedda and Luisa De Biagi, CNR Biblioteca Centrale "G. Marconi", Italy*

GL14 Program and Conference Bureau
TextRelease

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

Fourteenth International Conference on Grey Literature

Tracking Innovation through Grey Literature

National Research Council, Rome, Italy 29-30 November 2012

Hosted by the Central Library "Guglielmo Marconi", IRPPS Rome, and ISTI Pisa

Registration Form

The Conference Registration Fee includes attendance during the Sessions, a copy of the GL14 Program and Abstract Book, the full-text of the Conference Papers and PowerPoint Slides, as well as the conference pouch and participant badge. Also included are lunches, refreshments during the morning and afternoon breaks, and the Inaugural Reception.

Payment after October 1st 2012 add a €75 surcharge

☐ **€ 495** Participant Fee ☐ **€ 395** Author Fee ☐ **€ 275** Day Pass ☐ **€ 150** Student Fee

I am employed with a CNR institution and am entitled to 20% reduction on the fee: ☐

I am a *GreyNet* Member and am entitled to a 20% reduction on the conference fee: ☐

Registrant's Name: _____

Organization: _____

Address/P.O. Box: _____

Postal Code/City/Country: _____

Tel/Fax/Email:

Please Check One of the Boxes below for the Method of Payment:

❑ Direct bank transfer to TextRelease, Account 3135.85.342 - Rabobank, Amsterdam, Netherlands
BIC: RABONL2U **IBAN:** NL70 RABO 0313 5853 42, with reference to GL14 and Registrant's Name

❑ MasterCard ❑ Visa Card ❑ American Express

Card No. _____ Expiration Date: _____

If the name on the credit card is not that of the registrant, print the name that appears on the card, below

_____

Signature _____ **CVC II code** _____ (*Last 3 digits on signature side of card)*

Place _____ Date _____

Note: Credit Card transactions can be authorized by Phone, Fax, or Postal services. Email is not authorized.

GL14 Program and Conference Bureau
TextRelease

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

GreyNet Timeline 1992-2012

"Twenty Years serving the International Grey Literature Community"

2012-2011	2010-2008	2007-2004	2003-2000	1999-1992
2012 **GL14** Fourteenth International Conference on Grey Literature in Rome, Italy with CNR as Host GreyNet's 20th Anniversary 1992-2012 GreyNet's **International Directory** of Organizations in Grey Literature	**2010** **GL12** Twelfth International Conference on Grey Literature in Prague with NTK as Host **Monograph** on Grey Literature published by De Gruyter|Saur **GreyWorks**'10 Summer Workshop Series; GreyNet's Retrospective Collections 1993-1999 in OpenSIGLE	**2007** **GL9** Ninth International Conference on Grey Literature in Antwerp with the Flemish Ministry as Host; First Grey Literature accredited **course** via the University of New Orleans; OpenSIGLE –GreyNet Bilateral agreement with INIST (Service provider) and GreyNet (Data Provider)	**2003** **GL5** Fifth International Conference on Grey Literature in Amsterdam; GreyNet's **relaunch** by TextRelease	**1999** **GL'99** Fourth International Conference on Grey Literature in Washington D.C. USA
2011 **GL13** Thirteenth International Conference on Grey Literature in Washington D.C. with FLICC-FEDLINK as Host **GreyWorks**'11 Summer Workshop Series; **OAI7** Open Access Workshop in Geneva **OpenGrey** platform succeeds OpenSIGLE Online Password Protected Access to Serial Publications GreyNet launched **LinkedIn** Group	**2009** **GL11** - Eleventh International Conference on Grey Literature in Washington with FLICC-FEDLINK as Host **GreyWorks**'09 relaunch of the Workshop Series Grey Literature; First edition of GreyNet's **Newsletter**	**2006** **GL8** Eighth International Conference on Grey Literature in New Orleans	**2000** GreyNet discontinued	**1998** GreyNet merged with MCB University Press
	2008 **GL10** Tenth International Conference on Grey Literature in Amsterdam Science Park; GreyNet 's Collections 2003-2007 accessible via OpenSIGLE; The Grey Journal, TGJ received a Dutch National **Award**	**2005** **GL7** Seventh International Conference on Grey Literature in Nancy, France with INIST-CNRS as Host The Grey Journal, **TGJ launched**		**1997** **GL'97** Third International Conference on Grey Literature in Luxembourg with the EC as Host
		2004 **GL6** Sixth International Conference on Grey Literature in New York with NYAM as Host; First GreyNet **Award Dinner** held in New York		**1996** GreyWorks'96 1st Workshop on Grey Literature in College Park Maryland, USA
				1995 **GL'95** Second International Conference on Grey Literature in Washington D.C. at Catholic University of America
				1994 First volume of the Conference Proceedings published
				1993 **GL'93** First International Conference on Grey Literature in Amsterdam with EAGLE as Main Sponsor
				1992 GreyNet established under TransAtlantic

De Biagi, Luisa **83**

Luisa De Biagi received her degree in Literature and Philosophy at 'La Sapienza' Univ of Rome with a Specialization in 'Archivist-Palaeographer' (Vatican School of Palaeography, Diplomatics and Archivistics at the Vatican Secret Archive) as well as a Specialization Degree in Archivistics, Palaeography and Diplomatics (Archivio di Stato, Rome). Degree of the Vatican School of Library Sciences. De Biagi further holds a Master in 'Business Publishing' (LUISS Management – Rome). She was a member of the SIGLE Working Group and Italian Grey Literature Data-Base Working Group from 2002 and a Member of the CNR Working Group for Cedefop-Refernet Project (Consortium for Professional Education and Training coordinated by ISFOL). Member of the Commettee for Legal Deposit Acquisition at CNR Central Library, and European Association of Health Information and Libraries (EAHIL). Since 2010 is Responsible for the Italian National Referring Centre of Grey Literature at CNR Central Library 'G. Marconi' (Opensigle/Opengrey Network Project) and for the Library Functional Units 'Education and Training' and 'Cultural Activities Management', organizing didactics laboratories for students, professional training courses and teaching in professional trainings for librarians, students and users. Email: luisa.debiagi@cnr.it

Di Cesare, Rosa **71**

Rosa Di Cesare is responsible for the library at the Institute for research on populations and social policies of the National Research Council (CNR). She worked previously at the Central library of CNR where she became involved in research activities in the field of Grey literature (GL) as member of the Technical Committee for the SIGLE database. Her studies have focused on the use of GL in scientific publications and recently on the emerging models of scholarly communication (OA and IR). Email: r.dicesare@irpps.cnr.it

Le Bescond, Isabelle **112**

Isabelle Le Bescond is librarian at the Lille 1 University Central Library, Science and Technology, since 2005. She is currently responsible for developing the digital library IRIS. She has served in different University Libraries (Paris, Strasbourg) since 1994. She studied German language and literature. Email: isabelle.le-bescond@univ-lille1.fr

Luzi, Daniela **71**

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

Pavlov, Leonid P. **106**

Leonid P. Pavlov graduated from Moscow Physical-Engineering Institute, Dipl. Eng. in computer systems. He is a Candidate of Sciences in informatics; and since 1976 is employed with the Scientific and Technical Information Centre of Russia (VNTIC) as Deputy Director. Main works in information systems, scientific and technical information, and grey literature. Email: pavlov@vntic.org.ru

Pejšová, Petra **96**

Petra Pejšová 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. petra.pejsova@techlib.cz

Prost, Hélène **112**

Hélène Prost is responsible for studies at the Institute of Scientific and Technical Information (INIST-CNRS). The different studies concern the evaluation of collections, document delivery, usage analysis, grey literature and open access to information. Expertise in statistical tools and knowledge in library information science allowed her to participate in various research projects and writing of several publications.
Email: helene.prost@inist.fr

Puccinelli, Roberto **83**

Roberto Puccinelli is currently head of Section I at CNR's "Information System Office" and he's been working for CNR since 2001. He has previously worked in the private sector as system and network engineer. As adjunct professor, he has held courses for the First University of Rome "La Sapienza" ("Operating Systems II") and for the Third University of Rome ("Programming and Computing Laboratory"). He graduated in Electronic Engineering at the University of Rome "La Sapienza" and holds a master cum laude in Enterprise Engineering from the University of Rome "Tor Vergata". In the past he has worked in several research projects in the field of Grid technologies both at the national and international level (executive manager of Work Package 11 within the DataGrid project – V Framework Programme, et al.). He's currently involved in the design and development of CNR's information system. In particular, he coordinates projects for the development of application systems and is responsible for the design and implementation of CNR's data warehouse. He is also responsible for CNR's Local Registration Authority management. He's currently involved in projects regarding the design and development of research product open archives and persistent identifier registers/resolvers. He is author of several articles in the fields of Grid technologies, Autonomic Computing, Software Engineering, Open Archives and Persistent Identifiers. Email: roberto.puccinelli@cnr.it

Ricci, Marta **71**

Marta Ricci has an undergraduate degree in Humanities and a Master degree in Library Science from the University of Rome "Tor Vergata" with a thesis on bibliometric tools and citation analysis. She had an internship experience in the library of the Italian Prime Minister's office (Chigi's Library), where she was responsible for the Inventory of part of the library collections. Currently she is collaborating with the library of the Institute for Research on Population and Social Policies of the Italian National Research Council (CNR), in the field of GL. Email: biblio.irpps@irpps.cnr.it

Ruggieri, Roberta **71**

Roberta Ruggieri is librarian at the Senate of the Republic where she is responsible for the supervision of a digitalization project on Senate parliamentary print documents for the I to X Legislature. Her activity in managing digitalization project also includes document addition and classification in the electronic Senate catalogue. From 2004 she has been collaborating with the Institute for research on populations and social policies of CNR in research activities related to the field of grey literature and Institutional repositories.

Saccone, Massimiliano **83**

Massimiliano Saccone graduated in Letters and specialized in Library sciences at the "Sapienza" University of Rome. He works as librarian at the Central Library of CNR. He is the responsible for Legal deposit and Open Access activities. He has participated actively in several national and international projects on Information and Knowledge Management (Digital Preservation Europe – DPE, Italian Network of National Bibliography Number – NBN, etc.). His main interests are in scholarly communications, open access, digital preservation, metadata quality control, information system interoperability. massimiliano.saccone@cnr.it

Schöpfel, Joachim **112**

Joachim Schöpfel is Head of the Department of Information and Library Sciences at the Charles de Gaulle University of Lille 3 and Researcher at the GERiiCO laboratory. He is interested in scientific information, academic publishing, open repositories, grey literature and usage statistics. He is a member of GreyNet and euroCRIS. He is also the newly appointed Director of the National Digitization Centre for PhD Theses (ANRT) in Lille, France. Email: joachim.schopfel@univ-lille3.fr

Trufelli, Luciana **83**

Luciana Trufelli works as technical at the Central Library of CNR. She is responsible for Studies Office. She promotes and collaborates on many national and international projects on Information and Knowledge Management. She currently works in activities under open access, metadata quality control, and information system interoperability. She also has extensive experience in the areas of statistics and performance indicators of the R&D. Her main interests are in quality systems and institutional communication. Email: luciana.trufelli@cnr.it

Vaska, Marcus **96**

Marcus Vaska is librarian at the Health Information Network Calgary (HINC), Holy Cross Site. He was a former librarian for the Physician Learning Program, a collaborative initiative between the Universities of Alberta and Calgary, funded via an Alberta Medical Association trilateral agreement, where he was responsible for assisting physicians in their research, and addressing their perceived and unperceived learning needs. Prior to that position, he was a librarian at the University of Calgary's Health Sciences Library. Marcus' current interests focus on educational techniques aimed at creating greater awareness and thereby bringing GL to the forefront in the medical community. Email: mmvaska@ucalgary.ca

Notes for Contributors

Non-Exclusive Rights Agreement

- I/We (the Author/s) hereby provide TextRelease (the Publisher) non-exclusive rights in print, digital, and electronic formats of the manuscript. In so doing,
- I/We allow TextRelease to act on my/our behalf to publish and distribute said work in whole or part provided all republications bear notice of its initial publication.
- I/We hereby state that this manuscript, including any tables, diagrams, or photographs does not infringe existing copyright agreements; and, thus indemnifies TextRelease against any such breach.
- I/We confer these rights without monetary compensation and with the understanding that TextRelease acts on behalf of the author/s.

Submission Requirements

Manuscripts should not exceed 15 double-spaced typed pages. The size of the page can be either A-4 or 8½x11 inches. Allow 4cm or 1½ inch from the top of each page. Provide the title, author(s) and affiliation(s) followed by your abstract, suggested keywords, and a brief biographical note.
A printout or PDF of the full text of your manuscript should be forwarded to the office of TextRelease. A corresponding MS Word file should either accompany the printed copy or be sent as an attachment by email. Both text and graphics are required in black and white.

REFERENCE GUIDELINES

General

i. All manuscripts should contain references
ii. Standardization should be maintained among the references provided
iii. The more complete and accurate a reference, the more guarantee of an article's content and subsequent review.

Specific

iv. Endnotes are preferred and should be numbered
v. Hyperlinks need the accompanying name of resource and date; a simple URL is not acceptable
vi. If the citation is to a corporate author, the acronym takes precedence
vii. If the document type is known, it should be stated at the close of a citation.
viii. If a citation is revised and refers to an edited and/or abridged work, the original source should also be mentioned.

Examples

Youngen, G.W. (1998), Citation patterns to traditional and electronic preprints in the published literature. - In: College & Research Libraries, 59 (5) Sep 1998, pp. 448-456. - ISSN 0010-0870

Crowe, J., G. Hodge, and D. Redmond (2010), Grey Literature Repositories: Tools for NGOs involved in public health activities in developing countries. – In: Grey Literature in Library and Information Studies, Chapter 13, pp. 199-214. – ISBN 978-3-598-11793-0

DCMI, Dublin Core Metadata Initiative Home Page http://purl.oclc.org/metadata/dublin_core/

Review Process

The Journal Editor first reviews each manuscript submitted. If the content is suited for publication and the submission requirements and guidelines complete, then the manuscript is sent to one or more Associate Editors for further review and comment. If the manuscript was previously published and there is no copyright infringement, then the Journal Editor could direct the manuscript straight away to the Technical Editor.

Journal Publication and Article Deposit

Once the journal article has completed the review process, it is scheduled for publication in The Grey Journal. If the Author indicated on the signed Rights Agreement that a preprint of the article be made available in GreyNet's Archive, then browsing and document delivery are immediately provided. Otherwise, this functionality is only available after the article's formal publication in the journal.

An International Journal on Grey Literature

'DATA FRONTIERS IN GREY LITERATURE'

SUMMER 2012 – TGJ VOLUME 8, NUMBER 2