https://doi.org/10.26069/greynet-2025-000.540-tgj
ISSN 1574-1796

An International Journal on Grey Literature

Autumn 2016, TGJ Volume 12, Number 3

'Data Fuels and Validates Grey Literature'

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

https://doi.org/10.26069/greynet-2025-000.540-tgj

The Grey Journal

An International Journal on Grey Literature

COLOPHON

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.

CIP
The Grey Journal (TGJ) : An international journal on grey literature / Dominic Farace (Journal Editor); Jerry Frantzen (Technical Editor) ; GreyNet International, Grey Literature Network Service. - Amsterdam: TextRelease, Volume 12, Number 3, 2016. – *DANS-KNAW (NL), FEDLINK - Library of Congress (USA), CVTISR (SK), EBSCO (USA), Inist-CNRS (FR), ISTI-CNR (IT), KISTI (KR), NIS-IAEA (AT), NTK (CZ), and The New York Academy of Medicine (USA) are Corporate Authors and Associate Members of GreyNet International.*
ISSN 1574-1796 (Print)
ISSN 1574-180X (PDF)
ISSN 1574-9320 (CD-Rom)

Subscription Rates:
€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 GreyNet

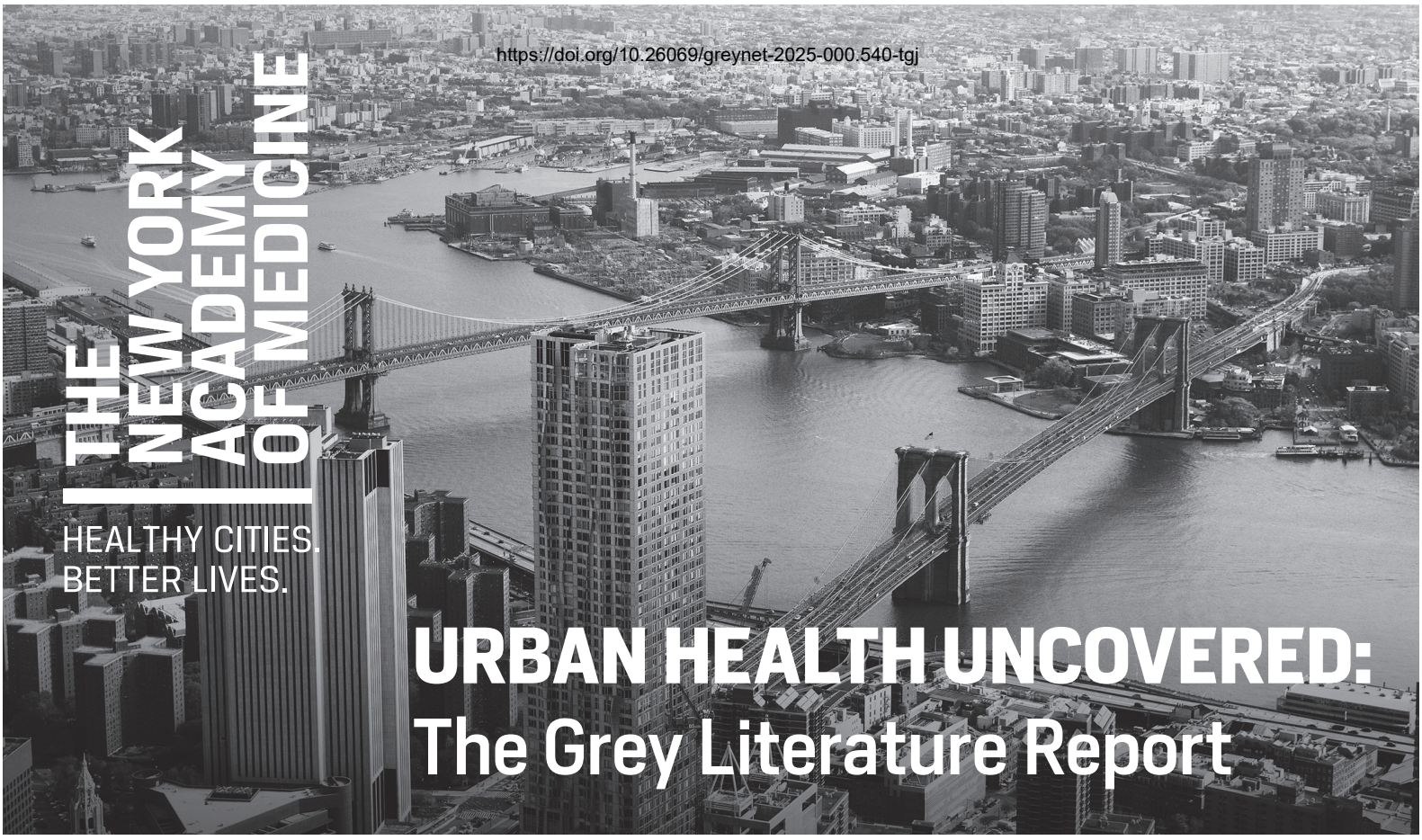
The Grey Literature Network Services was established in order to facilitate dialog, research, and communication between persons and organizations in the field of grey literature. GreyNet further seeks to identify and distribute information on and about grey literature in networked environments. Its main 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 Fuels and Validates Grey Literature'

https://doi.org/10.26069/greynet-2025-000.540-tgj



THE NEW YORK ACADEMY OF MEDICINE

HEALTHY CITIES.
BETTER LIVES.

URBAN HEALTH UNCOVERED:
The Grey Literature Report

SUBSCRIBE, SEARCH, DISCOVER AT:
Greylit.org or greylithelp@nyam.org

Enhance the impact of your research with contemporary, comprehensive urban health resources, selected from more than 750 non-commercial publishers from around the world.

- Researched and curated to give you a unique, inside track on urban health issues.
- Indexed and described to make it easy to discover resources across multiple disciplines.

Uncovering hidden resources for researchers, health professionals, librarians and the public.

NYAM.org **The New York Academy of Medicine** 1216 Fifth Avenue | New York, NY 10029

EDITOR'S NOTE

When asked to discuss 'grey literature' with a group of librarians in the social sciences, I chose not to make the assumption that we were all on the same page as to what was understood by the term. My hesitance to presume this was based on recent published articles about grey literature full of outdated content backed by references cited a quarter century ago. Thus, in preparing for the upcoming meeting, I chose instead to preface my presentation on grey literature in the light of ongoing research followed by examples specific to the social sciences. My annotated outline reads as follows:

Flip the switch, identify grey in a new light – It may not be alarming that the general public is unaware of what constitutes grey literature. What is alarming is when information professionals well into the 21st Century remain in the dark confusing grey literature with ephemera, echoing half-truths that it's hard to find, that it's unpublished, that it's not peer reviewed and of inferior quality to commercial publications. Even inserting personal bias e.g. that they don't like the term 'grey literature'. Let's take a few minutes to shed light on significant developments in this field of library and information science over the last couple of decades.

Grey Literature on the changing information landscape – One should not be surprised by changes in grey literature given the rapid pace of change on the information landscape brought on by advancements in technology and its impact on the production, processing, access, and preservation of grey as well as commercial publications. It could be said that the driving force behind these technological advancements were to a great degree focused on resolving the fundamental problems of information overload and its concomitant loss of information. Two problems seen as interconnected with the supply and demand sides of grey literature.

Examples of Grey Research Communities – Grey Literature is highly diverse bridging sectors of government, academics, business and industry. Even within one of these sectors diversity becomes further accentuated. Take for example academics, where research leading to published grey literature extends to disciplines and subject areas in the natural sciences, social sciences and humanities. As a result, numerous grey literature communities have developed over the years in a concerted effort to facilitate both the supply and demand sides for their grey literature. Let's look at some examples of the work done by these communities with particular emphasis on the social sciences.

Welcome to challenges on the Grey Horizon – In the spring of 2014, a workshop held in Pisa, Italy addressed challenges to policy development of grey literature resources. A 15 point document that has come to be known as the Pisa Declaration was drawn-up and has since been signed by over 140 signatories from some 70 organizations in 30 countries worldwide. The Pisa Declaration has since been translated and published in 20 languages. GreyNet being one of the signatories to the Pisa Declaration carried out a study to assess and further leverage its own information resources. This presentation concludes with a brief review of the outcome of that study.

Dominic Farace,
Journal Editor

Dissertations and Data*

Joachim Schöpfel, Hélène Prost, and Cécile Malleret (France);
Primož Južnič, Ana Češarek, and Teja Koler-Povh (Slovenia)

Abstract
The keynote provides an overview on the field of research data produced by PhD students, in the context of open science, open access to research results, e-Science and the handling of electronic theses and dissertations. The keynote includes recent empirical results and recommendations for good practice and further research. In particular, the paper is based on an assessment of 864 print and electronic dissertations in sciences, social sciences and humanities from the Universities of Lille (France) and Ljubljana (Slovenia), submitted between 1987 and 2015, and on a survey on data management with 270 scientists in social sciences and humanities of the University of Lille 3.
The keynote starts with an introduction into data-driven science, data life cycle and data publishing. It then moves on to research data management by PhD students, their practice, their needs and their willingness to disseminate and share their data. After this qualitative analysis of information behaviour, we present the results of a quantitative assessment of research data produced and submitted with dissertations Special attention is paid to the size of the research data in appendices, to their presentation and link to the text, to their sources and typology, and to their potential for further research. The discussion puts the focus on legal aspects (database protection, intellectual property, privacy, third-party rights) and other barriers to data sharing, reuse and dissemination through open access.
Another part adds insight into the potential handling of these data, in the framework of the French and Slovenian dissertation infrastructures. What could be done to valorise these data in a centralized system for electronic theses and dissertations (ETDs)? The topics are formats, metadata (including attribution of unique identifiers), submission/deposit, long-term preservation and dissemination. This part will also draw on experiences from other campuses and make use of results from surveys on data management at the Universities of Berlin and Lille.
The conclusion provides some recommendations for the assistance and advice to PhD students in managing and depositing their research data, and also for further research.
Our study will be helpful for academic libraries to develop assistance and advice for PhD students in managing their research data, in collaboration with the research structures and the graduate schools. Moreover, it should be helpful to prepare and select research data for long-term preservation, curate research data in open repositories and design data repositories.
The French part of paper is part of an ongoing research project at the University of Lille 3 (France) in the field of digital humanities and research data, conducted with scientists and academic librarians. Its preliminary results have been presented at a conference on research data in February 2015 at Lille, at the 8th Conference on Grey Literature and Repositories at Prague in October 2015 and published in the Journal of Librarianship and Scholarly Communication. The Slovenian research results have not been published before.
Keywords *Open science, open data, open access, institutional repository, data repository, research data, research data management, electronic theses and dissertations*

1. Data-driven science

Scientific results are increasingly disseminated as digital datasets. “Data are becoming an important end product of scholarship, complementing the traditional role of publications” (Borgman et al. 2007). Data are not only the fuel of the digital economy[1] but also of science and engineering.
Five years ago, the European Commission met the challenge and defined the main strategy for the development of an ambitious scientific data policy in the European Research Area. This strategy includes a framework for a collaborative data infrastructure, additional funding, measuring and rewarding data value and training of experts[2]. The need to manage the “data deluge”, is among the main drivers of computationally intensive science or e-Science, “the powerful paradigm in which distributed computer and knowledge systems, and information and communication technologies are integrated to provide services to enable large-scale and collaborative sciences and engineering” (Wang & Liu 2009). Mathematical modelling, numerical analysis and visualization techniques are part of this new way of doing science (figure 1).

* First published in the GL17 Conference Proceedings, February 2016.

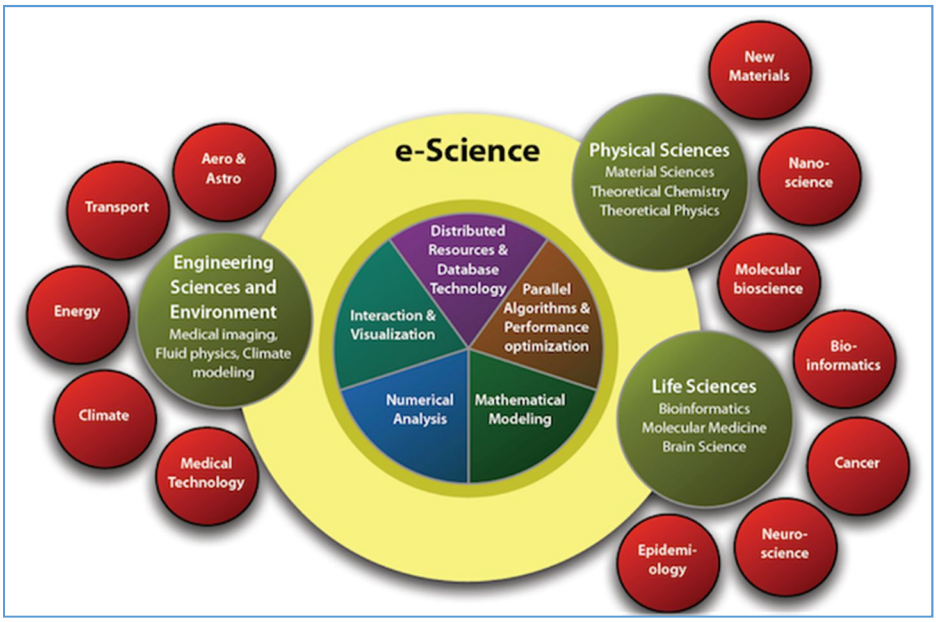


Figure 1 : Illustration of e-Science (by KTH Royal Institute of Technology in Stockholm[3])

E-Science affects all disciplines and research domains, even if some of them, such as life sciences or engineering, are more data-driven than, for instance, arts and humanities. But differences between different disciplines are diminishing and data is important for all of them, as research is based on it.

More than twenty years ago, academic publishing left the Gutenberg era and went digital. The digital revolution was the condition to enter the world of the "4th paradigm" of science where e-infrastructures enable "data-intensive scientific discovery" through data mining and integration of theories, simulations and experiments (Hey et al. 2009). Scientific information has become a continuum between publication and data. Linking data to documents is crucial for the interconnection of scientific knowledge. One can imagine this inclusion of datasets and other materials as the "perfecting of the traditional scientific paper genre (...) where the paper becomes a window for the scientist to not only actively understand a scientific result, but also reproduce it or extend it" (Lynch 2009). The possibility to reproduce research outcomes, i.e. the ability of an entire experiment or study to be duplicated, has always been the basis of science and scientific research. Today, the availability of the research data contribute to this necessary reproducibility. At the same time, it may prevent plagiarism, fraud and falsification of data.

But what exactly is research data? What does it mean? There is no clear or unique definition of the term. Following the US OMB Circular 110[4], research data can be considered as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings." The international directory for research data repositories re3data[5] distinguishes between fourteen different types of data (archived data, audio-visual data, configuration data, databases, images, network-based data, plain text, raw data, scientific and statistical data formats, software applications, source code, standard office documents, structured graphics, and structured text) but admits that there are other categories in the nearly 1,400 indexed repositories.

Large research projects, laboratories and technical apparatus produce what is commonly called "big data", i.e. research data characterized by their important volumes, their availability in real time (velocity) and their large variety (Laney 2001). Yet e-Science is not only "big data". The concept applies also to data output from individual scientists, smaller projects and research teams especially but not exclusively in social sciences and humanities. These data, defined as reusable research results, collected, observed, or created for purposes of analysis to produce original research results (University of Edinburg, cited by Burnham 2013), are produced in a large variety of formats, sources and types.

2. Data life cycle

Research data are part of the dynamic process of scientific research and discovery. In a very schematic and simplistic manner, two different functions of data can be distinguished in the research process:

- Data as material (input): at an early stage of the research process, data are collected and analysed from different sources and in different ways and formats, as material for exploration and hypothesis testing.
- Data as results (output): other data are produced during the whole process and at the end, together with publications, as research results.

The original raw or "input" data often are transformed and processed into derived products (metrics, graphics etc.). At the same time, research data, especially as research results (output) follow their own dynamic that can be described as a data life-cycle (figure 2).

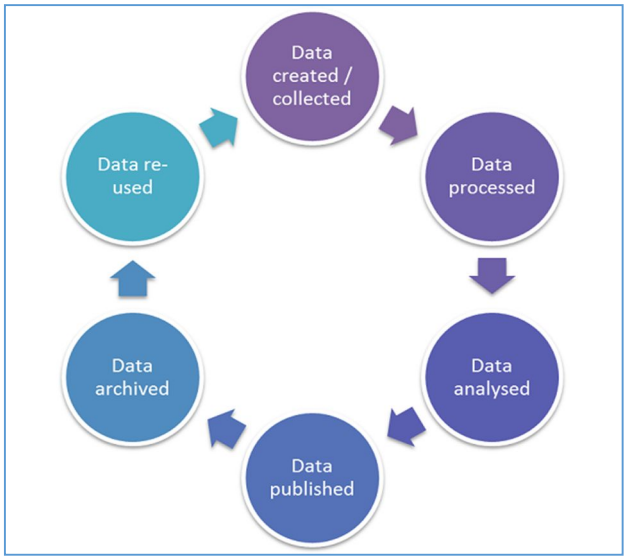


Figure 2: Data life-cycle (by Lancaster University Library[6])

We will not describe the whole life-cycle in detail, because of its complexity[7] and its great variability. "Although there may be significant differences in the individual stages, the life cycle is assumed to encompass the experimental design and capture, cleaning/integration, analysis, publication, and preservation processes, which occur in an iterative fashion" (Kowalczyk & Shankar 2011, p.251). We will just highlight two specific aspects:

- Data integration: Data integration means the process by which "disparate types of data (…) are identified and stored in a manner that facilitates novel associations among the data" (Bult 2002). This is more than preservation and sharing and goes beyond the capacities of most data repositories.
- Evaluation: Research data, as a part of the scientific "output" and together with publications, become increasingly involved in research assessment procedures, as for instance in the research portal of the King's College of London which integrates a current research information system and an institutional repository with published and unpublished results, including datasets[8].

Partly, research data management reflects the selection criteria of funding agencies and other scientific structures, with mandatory data management plans, long-term preservation guarantees and data sharing in open access.

Description and preservation of digital objects are part of the work of traditional academic libraries. For this reason, they generally consider research data curation and management as a new challenge, a kind of new frontier for the development of their campus services (Neuroth et al. 2013). It is generally seen as a culture challenge to the whole profession, due to the lack of the skills and attitudes acquired and needed to cope with it (Cox et al, 2014). They contribute to the assessment of data management practices and needs (Simukovic et al. 2014) and to the development and evaluation of data repositories (Pampel et al. 2013, Lynch 2014).

3. Data and/or publications

As said above, e-Science consists mainly of data and not of literature or documents (Hey & Trefethen 2005). The digital revolution deconstructed the unity of the text, eroding the notion of a monolithic 'document' in the hypertext paradigm and disintegrating the article in several distinct elements. At the same time and partly due to fragmentation, authors and publishers growingly enrich the article with new contents and features, such as multimedia, collaborative tools and data (Cassella & Calvi 2010). Some even predict that publications, as traditional vectors for scientific communication will disappear in favour of a direct communication between machines, at least in some specific research fields: "In the age of genomic-sized datasets, the biomedical literature is increasingly archaic as a form of transmission of scientific knowledge for computers" (Blake & Bult 2006). This may seem a little bit too futuristic, especially in the context of social sciences, arts and humanities where publications so far preserve their central role for the transmission of knowledge.

Until now, data were part of publications as support for argumentation and hypothesis testing or for illustrative purpose. In digital scholarship, new publication formats integrate data that can be updated, enriched, extracted, shared, aggregated and manipulated (McMahon 2010). Publications become live documents. The "next generation journal" will be enhanced and interactive, with video reference material, multimedia interactive graphs, links to datasets etc., making the article "more desirable for readers and more effective with regard to acquisition of the information" (Siegel et al. 2010, p.28). Publications become windows on research results.

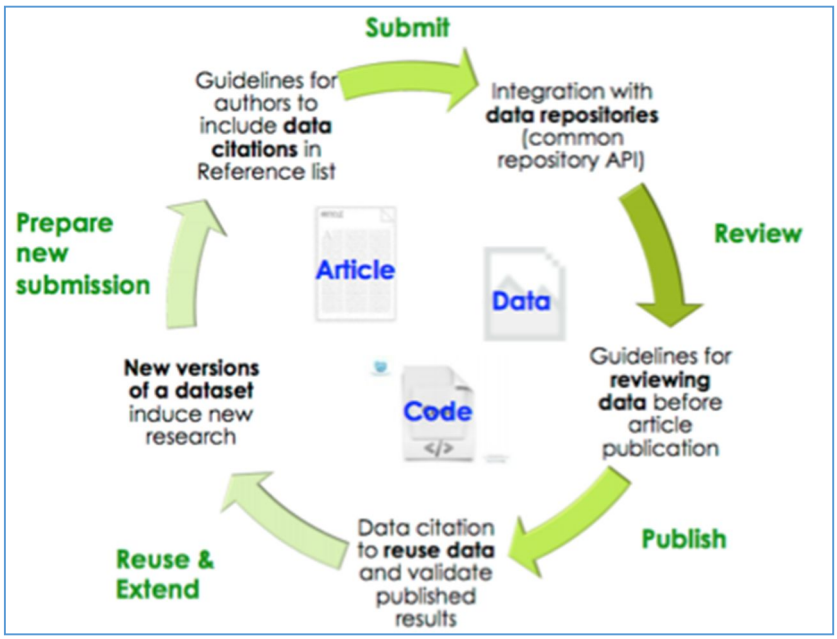


Figure 3: Diagram of an Automated Integrated Article and Data Publishing Workflow (Dataverse Project, Harvard University[9])

However, as figure 3 illustrates, other links are possible between data and publications. Perhaps the "data deluge" will not substitute academic publishing. Although we can be sure that it will change the way how academic publishing is today. At least three different ways can be distinguished in which documents contribute to data production and e-Science:

- Document as data: Documents, such as conventional articles, ETD, reports and conference abstracts or proceedings are exploited as primary data source for text mining, automatic extraction of meaningful information, intelligence etc. "Scientific journals will increasingly use standardized language and document structures in research publications" (Morris et al. 2005). The same remark applies to grey literature, including theses and dissertations[10] (see Murray-Rust 2007).
- Data vehicle: Enhanced publications or companion versions of published articles can serve as data carrier or database for content-dependent cross-querying of literature. For example, enriched articles can contain 'lively' and interactive content such as "interactive figures, semantic lenses revealing numerical data beneath graphs, pop-ups providing excerpts from cited papers relevant to the textual citation contexts (or) re-orderable reference lists" (Shotton 2012). Supplementary information files are available from the journal Web site, and/or the figures and tables containing research data within the article are available for download. Yet, their format does not necessarily allow or facilitate liberal reuse and exploitation.
- Gateway to data: Increasingly publications contain links to research data, either in the text or as part of the metadata. The reader (user) of the document can access the underlying research results but the data are not integrated into the document and both – data and document – can be used and reused separately. The full research datasets are published in a permanent archive or repository, with a unique identifier, with an open access data license or public domain dedication, and with sufficient descriptive metadata to enable their re-interpretation and reuse. This link can also be established when connecting a bibliographic database with a data repository, as INIS does with the Fukushima data archive (Savic 2015).

These different ways to link data and publications are represented in the STM data publication pyramid (figure 4). At the base line, raw data are just exploited for the publication of research results but neither

cited nor connected. They remain hidden, even if they may be made available to other scientists and/or peer reviewers, on demand.
At the top level of the pyramid, data are published together with the article (or report, dissertation etc.). The third level, well-structured and documented databases and data collections foster a new vector of data publishing, i.e. so-called data journals, peer-reviewed journals for the publication of articles (data papers) on original datasets and collections[11].

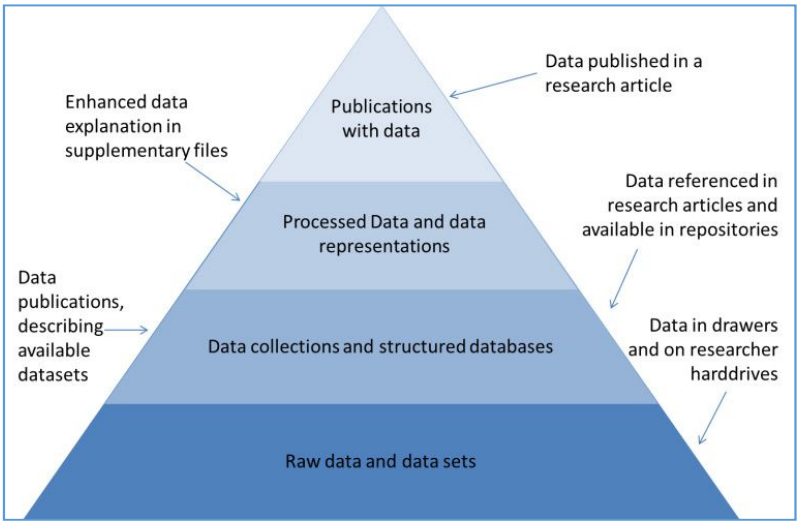


Figure 4: Data publication pyramid (from Reilly et al. 2011[12])

One part of these data is freely available, especially on servers that meet the criteria of open access. However, the access to many other research data is restricted or impossible. Savage & Vickers (2009) complain that the accessibility and the potential of datasets for reuse are often neither optimal nor effective, because of failing standards, metadata, identifiers or services. As for documents, availability and openness of data is not a simple on/off concept but a continuum between more or less open and restricted solutions ranging from the simple availability on the web of data on the lower end of the scale to data dissemination in non-proprietary formats and open standards as optimal openness.

4. The challenge of ETDs

While academic publishers make usage of new technologies to enrich the content and functionalities of their online products, universities can seize the opportunity of the supplementary files submitted together with electronic theses and dissertations (ETDs). Data sharing, long-term data storage and enrichment of dissertations by summary videos or data files were major topics of the 2015 USETDA conference at Austin, Texas, and the 2016 International Conference on Electronic Theses and Dissertation will be mainly about data and dissertations.[13]
Significant part of academic grey literature (Schöpfel & Farace 2010), produced and published by universities, dissertations are documents submitted in support of candidature for a PhD or doctorate degree presenting an author's research and findings (Juznic 2010). Theses and dissertations are “the most useful kinds of invisible scholarship and the most invisible kinds of useful scholarship” (Suber 2012). However, more and more dissertations are available in open access through institutional repositories (Sengupta 2014), i.e. open archives “serving the interests of faculty – researchers and teachers - by collecting their intellectual outputs for long-term access, preservation and management” (Carr et al., 2008). The typical life-cycle of ETDs today includes institutional repositories as main vector of dissemination (figure 5). In November 2015, the international directory OpenDOAR listed more than 1,500 institutional repositories with electronic theses and dissertations (60%). The academic search engine BASE provides more than 3.9 million ETD via the OAI-PMH protocol. “At many institutions ETD are simply the lowest hanging fruit and new submission batches can generally be counted on each semester” (McDowell 2007). The European DART-Europe portal[14] gives access to more than 600,000 open access research ETD from 586 Universities in 28 European countries while the new international search engine “Global ETD Search” by NDLTD[15] references 4.3 million theses and dissertations.

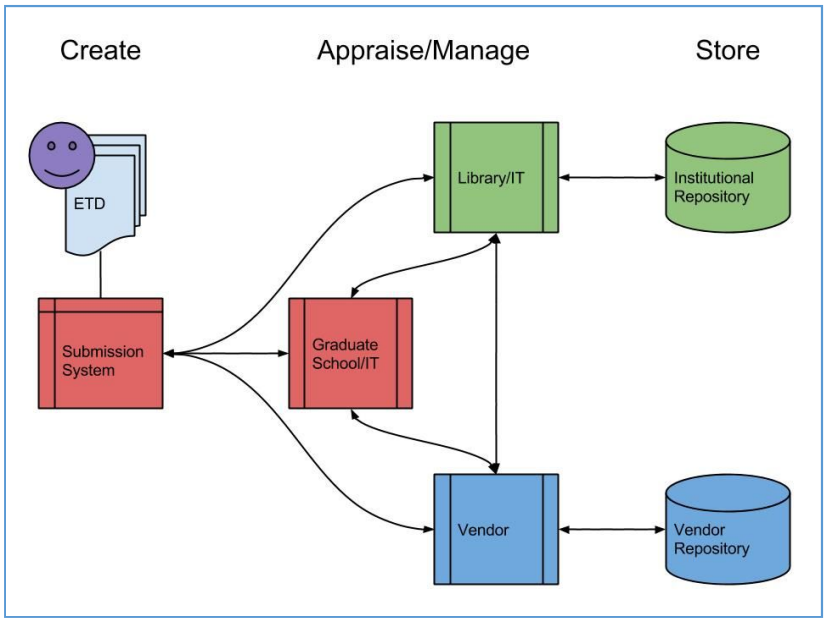


Figure 5: Life-cycle management of ETDs (Educopia project[16])

PhD dissertations contain the results of at least three years of scientific work. It is the result of cooperative work, between the PhD student and his tutor in the first place, but more generally, accomplished within a laboratory, a research team or an institute, school or company. These results may be presented as tables, graphs etc. in the paper or as additional material (annex). In the past, print dissertations have regularly been submitted together with supplementary material, in various formats and on different supports (print annex, punched card, floppy disk, audiotape, slide, CD-ROM…). Today, such material is submitted and processed as "complex content objects" (Schultz et al. 2014) together with the text files or as supplementary files in various formats, depending on disciplines, research fields and methods. If disseminated via open repositories, these research results could become a rich source of research results and datasets, for reuse and other exploitation. Thus, research results produced by PhD students could contribute to e-Science. However, there are three barriers:

- First, dissertations must be freely available in open access, deposited in institutional or other repositories and disseminated with sufficient user rights to allow re-use. However, up to now a significant portion of the digital dissertations are not online, not open, not freely available but embargoed or under restricted access (Schöpfel et al. 2015a).
- The second barrier is the fact that research data related to PhD dissertations are largely "dark data", i.e. "data that is not easily found by potential users (…) unpublished data (and) research findings and raw data that lie behind published works which are also difficult or impossible to access as time progresses" (Heidorn 2008, pp.281 and 285).
- Dissertations are a most important part of the scientific community, despite various critiques of both the romantic notion of authorship and the epistemological assumptions that form traditional notions of independent scientific and scholarly research. This makes it hard to define "authorship" regarding data produced with dissertations.

When this material is submitted as a kind of data appendix, the dissertation becomes a "data vehicle", where data are published together with the dissertation or as a part of it. Sometimes the data are available on a distant server and without the text of the dissertation, transforming the dissertation in a "gateway to data". Yet, too often the data are simply not available; or data, methodology, tools, primary sources are mingled, not indexed, badly described, and unrelated with the text, unconnected with other files.

Often, dissertations will be somewhere "in-between" with some data integrated in the text (tables, graphs, illustrations) and others published as annex. One example among thousands: a dissertation in archaeology from Slovenia contains plenty of photographs, maps and tables and has an annex with an excavation map, the results of chemical equations and a comprehensive description of analysed bodies. Beside this annex there is another volume with almost 300 pages of photographs from the excavations, drawings of the objects and their descriptions. This could be a perfect example for reuse of this comprehensive data if the data would exist in digital form, especially with the suitable searchable

platform. Even so, text and data mining would be necessary to identify and exploit all this information, and the dissertation is at the same time "data vehicle" and primary data.

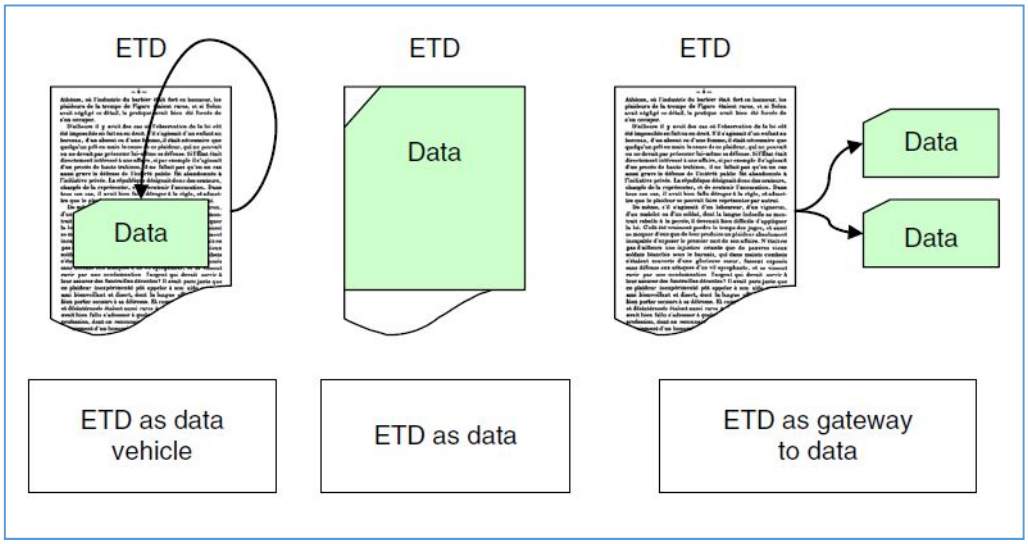


Figure 6: ETD as data, data vehicles and gateway to data

Obviously, ETDs can contribute to e-Science as primary data source, they contain (potentially) interactive content, and they are linked to datasets and research results (figure 6). But linking does not necessarily mean providing access. Supplementary material from ETD does not often even meet the criteria of simple availability on the web with an open licence, for different reasons, including dissemination under copyright or with more restrictive licences. This material continues to challenge the academic library. Research results, methodology, tools, primary sources are mingled, often not indexed, badly described, unrelated with the text, non-connected with other files, and virtually unavailable. In an academic environment that claims open access not only to scientific publications but also to research results, this situation is not satisfying.

5. Empirical evidence on data and dissertations

Up to now, there is very few empirical evidence and insight in the field of research data and other content related to dissertations, and only a small number of research projects and innovative workflows have been documented so far. Two exceptions are the Dataverse pilote project at the Emory University (Doty et al. 2015) and the ongoing research project ETDplus on preservation and curation of ETD research data and complex digital objects, funded by the Educopia Institute.[17] In the following we present some empirical results from surveys conducted by the Universities of Lille 3 and Ljubljana.

5.1. Research data management: practice and needs

In a survey on research data management and sharing in social sciences and humanities at the University of Lille 3 (Prost & Schöpfel 2015), PhD students represented 33% of the whole sample of 270 scientists. Compared to professors, senior lecturers etc., they have less experience with data management. They all store their data on the hard disks of their personal computers, sometimes also on a computer at the research laboratory or department, with back-ups on an external device like hard drive, USB flash drive or DVD, and sometimes even in the cloud (Dropbox). This is more or less personal knowledge management, good enough for personal research work and small projects but not compatible with larger research projects, such as the European H2020 programme. Also, they do not delegate this management. The Lille PhD students are not really different compared to other universities, as other survey results show - many PhD students are interested in data management and to some extent in support of sharing at least some data but have little or no experience at all. These results are compliant with a German survey with 117 PhD students of different disciplines (STM and SS&H) at the University of Humboldt, Berlin (Kindling 2013).

Our survey on research data at the University of Lille 3 confirms that PhD students have less experience with data sharing, which is not surprising as they are at the very beginning of their scientific career. More than other scientists, they often simply do not know options and opportunities for the deposit and sharing of their research results. Yet, 30% of them declare that other persons of their research team have access to their own data. This is a basic way of data sharing, not on the Internet but on their computers or via flash drives, Dropbox, the University Intranet etc. Also, they are more interested in reuse of data from other researchers than other categories.

Nearly one third (28%) of the students do not want to make their data available in the future or at least hesitate, which is the same part as for other scholars and researchers. Yet, they show a significantly higher motivation to deposit their research results in a data repository (63% compared to 43%), even in a local repository (laboratory, department) while the other scientists clearly prefer international and domain-specific sites. When asked which kind of service they would need, they ask for technical advice and help for data management plans for the publishing of their results. More than the elder staff, they also ask for assistance in ethical and legal issues. As a matter of fact, privacy issues and third party copyright are two serious legal problems that need awareness.

5.2. Research data in dissertations: a French-Slovenian survey

Following a first survey on 283 dissertations from the University of Lille 3 between November 2014 and January 2015 (Prost et al. 2015), we analysed two other, complementary samples:

- ETDs in the fields of social sciences and humanities from the Universities of Lille 1, 2 and 3.
- Dissertations in the fields of social sciences and humanities from the University of Ljubljana.

The survey is based on a German research project at the Humboldt University at Berlin (Simukovic et al. 2014). The methodological approach has been described by Prost et al. (2015). Altogether, the sample contains 780 digital and print dissertations from more than 15 different disciplines (figure 7).

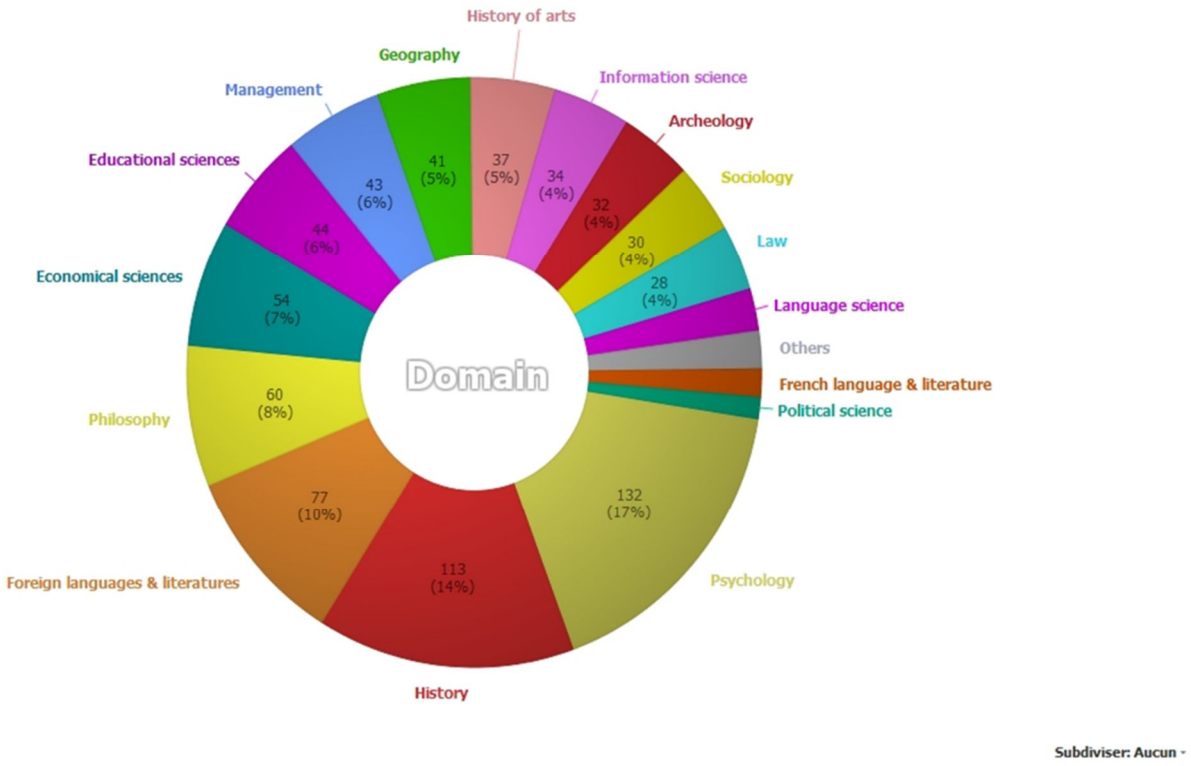


Figure 7: Scientific disciplines of the survey sample (N=780 dissertations)

The sample consisted of 353 digital dissertations (45%) and 427 print dissertations (55%), from 1987 to 2015. In our sample, Psychology, History, Foreign Languages and Literature (English and American, Spanish, Slavonic, Hebrew…), Philosophy and Economics were the most represented disciplines, followed by Education, Management, Geography and History of Arts.

All dissertations have been analysed either in digital or print format or on microform. Each dissertation has been checked by at least one of the authors, either in the library holdings (print or microform) or on the institutional repository server. We tried in particular to identify research data added to the end of the dissertation. In our sample of 780 dissertations, 522 contain one or more appendices with some kind of research data (67%). The length of these appendices varies widely, from one to 829 pages, with a median of 37 pages, and totalling more than 45,000 pages (Figure 8).

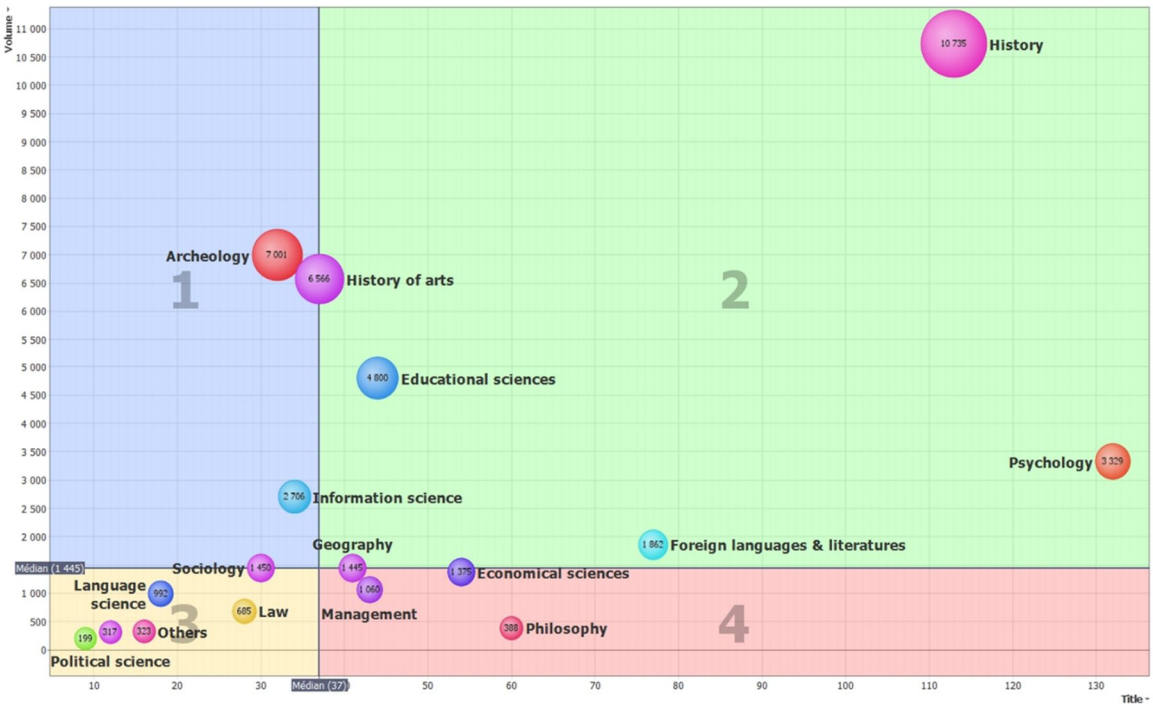


Figure 8: Size of data appendices (in pages, N=522 dissertations)

Even if each appendix holds some kind of research data, this does not mean that one can find research results (data) stricto sensu on all pages. Some pages contain empty questionnaires or survey forms, experimental procedures, bibliographies etc. which cannot be considered as data.

5.3. Disciplines

In the first Lille survey, the distribution of disciplines per dissertation with appendices was more or less the same than for the overall sample, with a significant linear determination coefficient R2 between both variables of .91 (Prost et al. 2015). The differences were not significant – in some domains such as Psychology, Philosophy and Linguistics, we found fewer dissertations with data appendices than the average (67%); in others there were relatively more (Information Sciences, History of Art). In Education and in Archaeology and Egyptology, all dissertations of the sample contained some form of data appendices.

Yet, in the larger sample the differences between disciplines are elsewhere (Figure 8). Some disciplines "produce" rather large appendices, with an average number of pages above the mean of the whole sample, such as History, Education and Foreign Languages, while others most often contain shorter appendices (Sociology, Law, Political Sciences…).

5.4. Support, presentation and format

In France, all files of digital PhD dissertations should be deposited with the text, and the French national computer centre for Higher Education (CINES) maintains a list of accepted file formats for long term preservation. However, there is no control of this deposit, if really all files with data and other material have been deposited or not. Also, nearly all files are in PDF (image or text), and other formats are very rare. In the French sample, only one dissertation has been submitted with audio-visual files in audio and video file formats on CD-ROM.

The French and Slovenian official guidelines for PhD dissertations do not specify how to structure or present an appendix. Some dissertations have poor or no table of contents for their appendices, like a dissertation in History with a table of content for the text volume but not for the two volumes that contain rich material, including 1,581 figures and images.

As mentioned above, 45% of our samples are electronic dissertations. Compared to the print dissertations, they contain slightly more appendices (Figure 9).

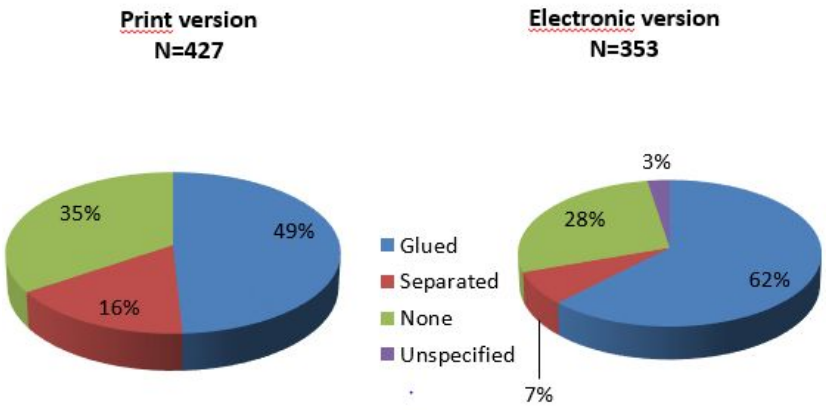


Figure 9: Link between text and appendices (in %, N=780)

Also, electronic dissertations often do not separate text and appendices but glue them together into the same file (62%), worse than the presentation of appendices in print dissertation (49%). Here the dissertations are not gateways to data but play the role of data vehicles, yet with data that are more or less useless or rather, not really reusable because of the format and missing structure.

Nevertheless, some dissertations demonstrate a real effort of data management and curation by the PhD student. For instance, a French dissertation in Egyptology on late Egyptian steles contains an exhaustive inventory of those steles on CD-ROM with indexing of geographical origin, general characteristics, specific particularities and dating, together with the transcription of the inscription and a justificatory supporting the provenance of the stele. The PhD student also delivers a user manual for the navigation in the database. Two examples from Slovenia: a dissertation in archaeology (Early Iron Age) provides short guidelines for the use of the annex which encompasses more than 50% of the dissertation; in a dissertation in history of arts, the annex is on a CD-ROM together with an installation file for the software needed to open the content files.

Another French dissertation in Linguistics presents a diachronic analysis of the vocabulary from 49 political speeches and 10 manifestos, pamphlets and articles, with a lexical analyser software (Wmatrix corpus analysis and comparison tool). The appendix contains the complete list of all words with their frequency of usage ranking.

Dissertations in History, especially for studies on historical social groups, sometimes contain detailed and well-structured biographical information, presented like a database. One example for this “prosopographical” approach at the University of Lille 3: a dissertation on the Renaissance elite of the old Flemish town of Douai with biographical records of 423 aldermen, with structured information about, among others, place and date of birth, date of death, mandate period, noble titles and occupation.

5.5. Research data sources

The PhD students used a great variety of sources for their scientific work, with four major data sources, i.e. surveys, text samples (corpora), experiments (observations) and archives (Figure 10).

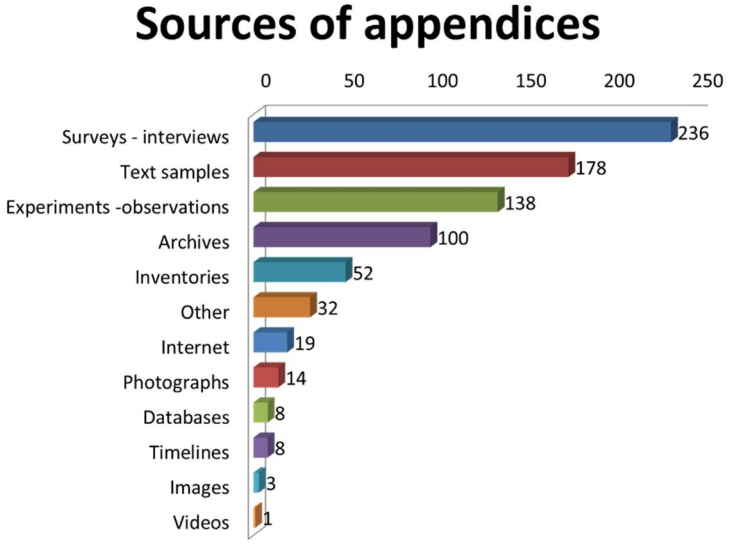


Figure 10: Data sources per dissertation (N=780)

Other less exploited sources are inventories, Internet sources, photographs and images, databases, timelines and videos. The distribution of data sources is to some extent specific for each discipline.

Y: Domain	Archives	Databases	Experiments - observati...	Images	Internet	Inventories	other	Photographs	Surveys - interviews	Text samples	Timelines	Vidéos	Tous
Archeology	1		14			25		5		3			30
Economical sciences		1	27		3	1			14	18			43
Educational sciences			6		3			1	33	13	2		38
Foreign languages & literatures	4		5		1		11		14	35	1		46
French language & literature									2	3	1		6
Geography	3	2	27		1		5		15	8		1	33
History	76		4	1		7		2	9	21	2		88
History of arts	10		4			17	7	2	3	4	2		28
Information science	1	1	1		6		6		17	10			28
Language science			1						3	5			7
Law					2				2	5			7
Management	1	3	7		1	1			21	14			30
Others		1	4						1				6
Philosophy	1		2					1	1	9			11
Political science									6	3			6
Psychology			30				3	2	71	12			91
Sociology	3		6	2	2	1		1	24	15			28
Tous	100	8	138	3	19	52	32	14	236	178	8	1	526

X: Sources

Figure 11: Data sources and disciplines (N=780)

Here are some examples of heavily used sources:

- History: archives, text samples
- Psychology: surveys, experiments
- Philosophy: text samples
- Foreign Languages and Literature: text samples
- Information and Communication Sciences: surveys, text samples, Internet
- History of Art: inventories
- Linguistics: text samples, surveys
- Archaeology and Egyptology: inventories, photographs

However, the situation is more complex, and Figure 11 reveals as well specific data-profiles for each discipline as disciplinary profiles for each data source: inventories for instance are typical for archaeology and history of arts, experiments and observations are specific for psychology, economics and geography, and so on. These are typical research data sources for the social sciences and humanities. Compared to the Berlin survey, other data sources like simulations, statistics, reference data or log files (usage data) are unusual or missing. For instance, many PhD students from the Humboldt University reported that they made use of measurement series, statistical analyses and spectra (Kindling 2013) – except for the statistics, we did not find such data sources in our SS&H sample.

5.6. Typology of research data

Which are the research data produced by the PhD students and present in the appendices? Our evaluation reveals several different and heterogeneous data types (Figure 12).

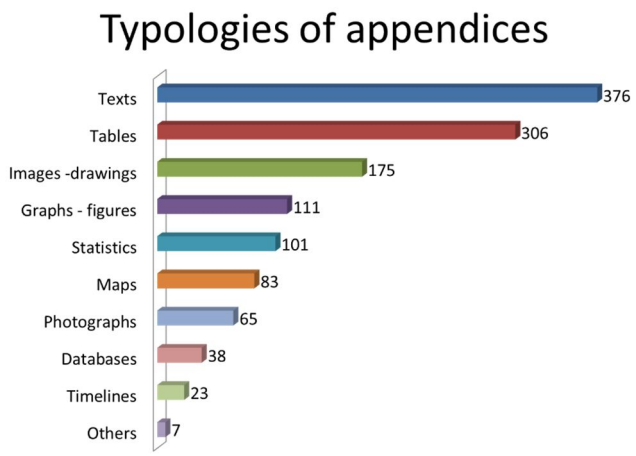


Figure 12: Data types, per dissertation (N=780)

However, two types of data are significantly more produced in these dissertations, i.e. text samples and tables (spreadsheets). Other data are produced in form of images, drawings, graphs and figures, while statistics, maps, photographs, databases and timelines (chronologies) are also part of the data appendices but less often. We found only one dissertation with audio-visual media (interviews) and no dissertation at all with geolocation data.

Again, as for the data sources, there are some discipline-specific data type profiles (Figure 13). In some disciplines, one or two data types are predominant. This is the case in Philosophy and Language Science where text samples represent more than half of the data. Other disciplines are characterized by a wider number of different research data. Some examples (in brackets, the percentage of this data type for all data appendices in this discipline) :

- History: ten different data types, including text (74%), tables (50%) and images (44%).
- Information and Communication Sciences: ten different data types, including text (71%) and tables (43%).
- Archaeology and Egyptology: nine different data types, including images (73%), maps (60%), text (50%) and photographs (37%).
- Psychology: eight different data types, including tables (71%), statistics (60%) and text (53%).
- Economics: six different data types, including text (84%), tables (72%) and graphs (37%).

Y: Domain	Databases	Graphs - figures	Images - drawings	Maps	Others	Photographs	Statistics	Tables	Texts	Timelines	Tous
Archeology	4	2	22	18		11	1	16	15	1	30
Economical sciences		16	1	5			2	31	36		43
Educational sciences		8	14	1			5	25	29	1	38
Foreign languages & literatures	1	1	20		1	1	6	21	36	1	46
French language & literature		1					1		5	1	6
Geography		13	7	13		5	3	27	23		33
History	16	22	39	27		26	14	44	65	12	88
History of arts	6		17	8	1	8		4	20	1	28
Information science	2	7	7	3	4	2	5	12	20	1	28
Language science	1	1	1				1	1	7		7
Law		1	3	2				4	5		7
Management	2	12	10	1		1	7	26	22	2	30
Others	1	2						2	4	1	6
Philosophy		2	2		1	1		1	11		11
Political science	1	1	4				1	6	2		6
Psychology	2	15	20	1		4	55	65	48		91
Sociology	2	7	8	4		6		21	28	2	28
Tous	38	111	175	83	7	65	101	306	376	23	526

X: Typology

Figure 13: Data types and disciplines (N=780)

The research data are very different. Some examples: a great number of images and photographs on the religious life in the French town of Etaples from the beginnings to 2000, statistics on prisons and prisoners in Northern France during the French Third Republic, the mapped tours and comments of children in a dissertation on two exhibitions, or a large corpus of old documents and archaeologic findings for the reconstruction of the organisation of banquets in Anglo-Saxon England. In another Slovenian example from history the annex contains on 400 pages a comprehensive list of short bibliographies from priests living in Carniola during the second half of the 18th century.

Some data types are present in all or nearly all disciplines, like text samples, tables, images or graphs – an observation which confirms the Berlin results where texts, tables (spreadsheets) and databases were dominant. Others, in particular inventories or audio-visual material, are at least in our sample specific for one or two disciplines. We compared print dissertations and e-dissertations and performed a chi-squared test but found no significant differences neither for research data sources nor for data types (on .05 level). Obviously, these differences are more related to disciplinary methodologies than to support.

5.7. The special case of ETDs in engineering sciences[18]

In order to learn more about research data in dissertations, in particular in the field of applied sciences, another study was conducted on 86 ETDs deposited in the institutional repository DRUGG[19] of the Faculty of Civil and Geodetic Engineering of the University of Ljubljana, Slovenia (UL FGG). The

repository was established in 2011, it is listed in the OpenDOAR and ROAR directories and compliant with the OpenAIRE infrastructure criteria (Koler-Povh et al. 2014). In September 2015, it contained 2,625 items, mostly Diploma theses (2,180) and Master theses (125) (Koler-Povh & Lisec 2015). 86 dissertations of 100 doctoral dissertations in civil engineering published between 2008–2014 were included in the survey on research data.

At UL FGG all dissertations since 2008 have been archived in print and digital version. All print and electronic versions are identical both in terms of content and layout. In the field of civil engineering, all dissertations contain some kind of research data. Together, we found 18,981 datasets in 86 analysed dissertations. Usually these data are integrated in the content, i.e. they are part of the text (the dissertation as the raw data source). Further, in 28 dissertations (33%), some data are published at the end of the dissertation, as a separate and distinct part of the dissertation. When submitted in digital format, the data are published in a single data file, in PDF or JPEG or, especially for large datasets, compressed in a ZIP file, as requested by the IR DRUGG's editorial board. We found 247 different data files for these 28 dissertations. All files have their own metadata.

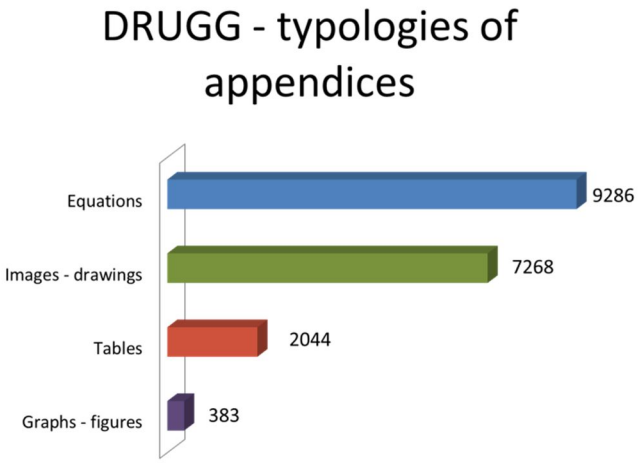


Figure 14: Data types in the DRUGG repository (civil engineering, N=86)

Most datasets published in the dissertations are equations (n=9286), which appear in 76% of all dissertations (66), with a high number in 14 dissertations. There are two dissertations with 424 and 400 equations each. In the former, the number of figures is also high (n=120), in the latter the number of other types of appendices is low, i.e. below 50 for all types.

There are two more dissertations with more than 300 equations and five with more than 200. Equations are always the most frequent type of datasets. We conclude that in civil engineering equations are the most used type of appendices.

Compared to equations, we found less images, even though they are present in all dissertations. 29 dissertations (34%) contain a high number of this type of data. Following UL FGG's guidelines (Koler-Povh and Turk, 2011), “figure” is a broader term, which includes graphs, pictures, figures, schemes, notes, windows, plans, and charts. Nevertheless, some students use the term “figure” only for schemes, pictures, and photos. In one dissertation the photos are separated from figures. Nine students distinguish between figures and graphs. In the doctoral dissertations from UL FGG, figures are the most frequently used type of appendices; one dissertation contains 252 figures, 11 others published 100 or more figures each. Sometimes figures from other author(s) and sources are used. The counting of such figures is not the same for all dissertations. Two dissertations count imported, i.e. adopted figures, separately from (author's) figures. More dissertations present the adopted figures by stating the source in the legend of the figure, but they do not list them separately in an index of imported figures.

There is also the problem of classifying maps, sometimes they are photos (e.g. satellite photos), sometimes they are maps, many times they are an appendix to the dissertation due to the high paper format.

In 35 dissertations two types of appendices are presented in a similar frequency, for 15 of them the similarity is high (the difference in the frequency is lower than 25%). In the whole sample of 86 dissertations, these 15 present 17%. We can conclude that the number of single type of appendices is regulated by authors on purpose.

6. Potential re-use of research data in dissertations

All these datasets are important to understand a dissertation’s framework, methodology and results. They are helpful for the evaluation of the author’s interpretations and conclusions, and they provide

raw material useful to verify and validate the overall research. Moreover, many, if not all data could also be of real value for further research. These data could be used to create image databases, digital maps collections or digital libraries with manuscripts, archival material and other text samples open for text mining tools. Results from experiments and surveys could be published in a way that allows for reuse, data mining and automatic meta-analysis on different datasets. Research results could thus become new data sources and generate further research. However, this potential reuse requires data management and curation to remain accessible and interpretable over time, including metadata and long-term preservation (Neuroth et al. 2013). For young scientists and PhD students, learning how to design and implement a data management plan (DMP) is even more important in so far as more and more funding bodies evaluate the existence and quality of DMPs in research project proposals. Our empirical data do not tell us if the PhD students conducted a data management plan. But only few dissertations demonstrate a real effort of data management and curation.

Our study reveals at least three barriers to open data:

- Incomplete, inadequate or missing description of the whole datasets and/or individual data. In some dissertations, especially in History, History of Art and Archaeology, inventories, photographs, maps etc. are well described and indexed. But these are exceptions and often descriptions are simply missing. This problem includes, too, the lack of citability when datasets are not correctly identified.
- Missing organisation. Research data are presented without any structuration or organisation, often together with other, not reusable material in a kind of information mash-up not suitable for further research.
- Inadequate format. In print copies, this means that data are not clearly separated from the dissertation text. In electronic dissertations, this means that data and text are glued together in a PDF file instead of being separated and published in adequate file formats (spreadsheets, image files, text files, database formats, XML…).

Other problems are related to the choice of media, e.g. compact disc, DVD, online server, USB flash drive... For instance, the dissertation on Egyptian steles inform about an online database with restricted access but does not provide login and password. For some retro-digitized dissertations, the online version does not include the data appendix submitted together with the print version.

All these problems make it difficult to find a dissertation's underlying data. The dissertation itself functions more like a kind of barrier instead as a gateway. Without metadata, without identifiers and referencing, it is virtually impossible to find these data otherwise than reading the dissertation. Lack of searchability is a direct consequence of missing data management and curation.

And they make it difficult if not impossible, too, to exploit these data with tools of text and data mining. Text and data mining tools need great volumes of open data, and hidden data in text or protected data files are not really useful for this purpose.

7. Legal aspects

Applying copyright to dissertations is already rather complicated (Schöpfel & Lipinski 2012). And regarding the sharing of research data, "the law makes all of this far more complicated than it need be. For those seeking to pick and choose which reuses of another's data may be permitted by law, regrettably, the answers (…) are more context dependent than many would like" (Carroll 2015). The legal environment of data and other supplementary materials is not the same as for ETD (see Murray-Rust 2008). Both must be considered independently, even if related and interconnected. Non-adapted licensing or (over) protection by copyright can be legal barriers to their deposit, dissemination and reuse. Linking datasets to the copyright protection of ETD creates a potential conflict with open data policy.

The European Commission and several national governments promote the dissemination of datasets under the minimalist open licence, limited to the attribution of authorship (CC-BY). On the other side, authors and service providers of ETD often adopt a more restrictive sharing policies that prohibit modifications and for-profit use, apply the full protection of the intellectual property law or limit the dissemination to campus-wide access (Schöpfel & Prost 2013). This is too restrictive to realize the potential for reuse of data and to be in conformity with the wish of the European Commission to make it "a general rule that all documents made accessible by public sector bodies can be re-used for any purpose, commercial or non-commercial, unless protected by third party copyright."

Some content may be protected by privacy or confidentiality concerns, for instance personal (human) data and sensitive or strategic information, including professional secrecy. Other research results may be subject to specific sui generis database property rights, and sometimes open access policy may be in conflict with legitimate interests (publishing for scientific career) and fear of plagiarism. In any case,

author and institution must reconsider the legal condition of the deposit and dissemination of datasets and other material, but they should do so applying a policy of open data allowing for a maximum of reuse and exploitation. Unlike ETD, datasets should not only be free (in terms of the open access movement) but also "libre", i.e. reusable.In our survey, at least two legal problems are related to the deposit and dissemination of research results in dissertations:

- Privacy issues. Some appendices contain personal data, about living or dead people, historical persons or unknown (anonymous) people. These may be survey data, experiments, interviews, biographies etc. In so far as the information allows identifying individual persons, at least with regards to the French law they need special processing and careful handling.
- Third party copyright. Some dissertations contain material that is protected by copyright and cannot be reproduced or disseminated without authorization, even by fair use or copyright exceptions (short citation, research…). These may be text samples, maps, photographs, copies from books etc. – material not created by the PhD student him/herself.

Sometimes, the authorship of the research data remains uncertain. These problems should be addressed as a part of doctoral education on data management, well ahead of decisions on preservation and dissemination. Because "restrictions in the use of research data directly affect research data curation (they) must (…) be taken into account right from the beginning (in matters such as policies, technology, etc.)" (Neuroth et al. 2013). Legal requirements, metadata, back and front office of research data handling have to be considered as a whole, interconnected, and interdependent. Important are two aspects: document and data must be distinguished and separated, intellectually, logically and physically; and the whole approach must be designed in a framework of open data, open access and open science.

Last but not least, open access to digital dissertations and data can be helpful to prevent plagiarism, as a specific form of academic integrity breaches. There was always some concern that plagiarism might occur easier and more often, if dissertations are in open access and freely accessible. On the other hand, open access makes it also easier to detect plagiarism, even if some forms of plagiarism are not easy to detect by anti-plagiarism software. Here, open access to data can be helpful when both, content (text) and supplementary data files are considered as a compound dissertation. The originality of PhD should be also proven by open access datasets especially in the cases when these data were collected by the author and are not part of some collective research project, which is usually the case in social science and humanities. This can be a good way to prevent or deter possible falsification of data.

8. ETD processing and workflows

The data publication workflows should be incorporated to dissertation submission process (Vompras & Schirrwagen 2015). But should dissertations and related datasets be processed together or separately? Should they be disseminated on the same or on different repositories? Should they be preserved on the same or on different servers? How should they be linked?

The Educopia Institute's *Guidance Documents for Lifecycle Management of ETDs* (Schultz et al. 2014) suggests the separation of the dissertation text files and the related "complex content objects" whenever possible. "Embedding multimedia components within the full text might seem advantageous in that they would then be inseparable. However, when the time comes that it is necessary to migrate either the full text itself or one of the multimedia components, having separate files would greatly simplify matters" (p.5-9). Metadata and persistent identifiers like handle, PURL, ARK or DOI are supposed to provide the "glue" that binds together the text, multimedia and data files. Preservation-worthy research data (survey data, measurements, laboratory notebooks, measured spectra etc.) might be stored as part of an "ETD package" or, after transformation into archival files, in a separate data repository, and "links to the data repository from the ETD metadata would then enable researchers to access these research data in the future" (p5-11).

The Dataverse ETD pilot program at Emory, Atlanta, is similar to this approach (Doty et al. 2015). In the Emory workflow, dissertation and research data are deposited separately, and while the dissertation is stored and disseminated on a publication (or document) server, the PhD student is invited to submit the research data to an appropriate disciplinary data repository or, if not available, to the Dataverse pilot. The Emory program supports tabular file formats (.xls, .csv, .xlsx, .dta, .R, SAS Files), software code (.hpp, .py, .rst, .cpp) and geospatial data (.mxd, .kmz, .kml, .gpx, etc…). In a very schematic way, the ETD program runs as follows (Figure 15).

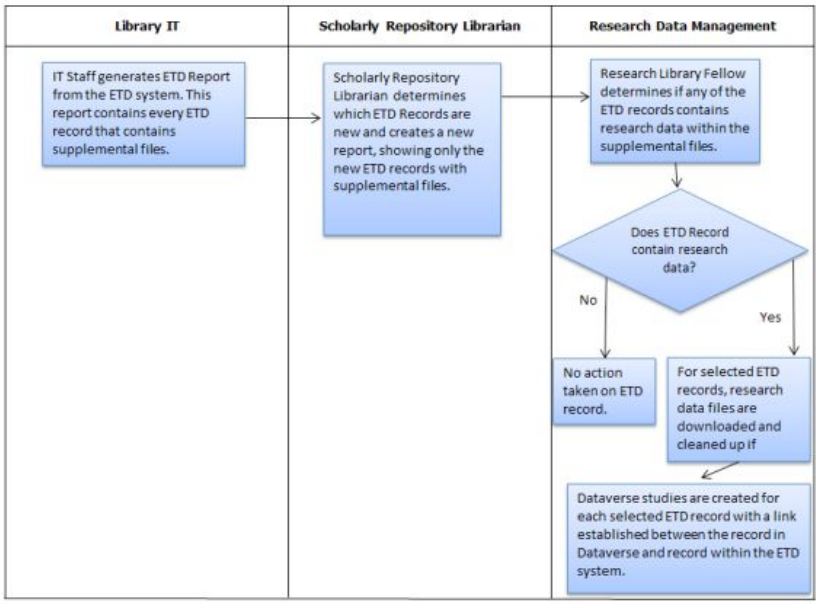


Figure 15: Workflow for Dataverse ETD Pilot Program at Emory (Doty et al. 2015)

The workflow includes also a phase of "cleaning up" research data, i.e. data curation just before the deposit and the connection to the dissertation via identifiers and direct linking between the Datavers pilot and the ETD system.

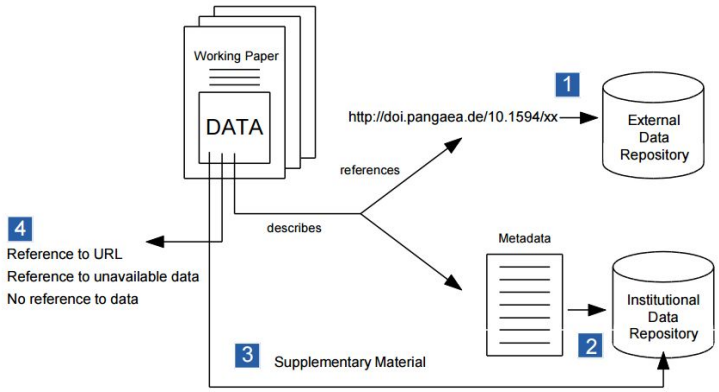


Figure 16: Linking grey literature with research data as discrete resources (Vompras & Schirrwagen 2015)

This process can be done in smooth and simple way, as a growing number of repositories show, such as the Bielefeld PUB[20] or the TU Delft institutional repository.[21] Vompras & Schirrwagen (2015) describe the linking options as follows (here for working papers): Both workflows have in common that dissertation and data are separated, that this separation is operated before or during the deposit of the dissertation, that the deposit is preceded by data curation, and that dissertation and data are not stored and dissemination on the same repository.

For universities in Slovenia, especially University of Ljubljana, there was a long way how the legal backgrounds have been prepared or revised to support a mandatory process of ETD (Ojsteršek et al, 2014). The process is still not finished and although the problem of research data was identified, due to the other, more basic, legal and even competence and authority problems it was not really tackled yet. This may change as they will have to adapt and embrace the policy of open access that includes also research data. In the Slovenian National strategy 2015-2020 on Open access, research data has become one of the priorities: "The research data financed by public funds should be as far as possible open, accessible with minimal restrictions. Open information must be given to locate or access them evaluate and understand to be useful for others and, if possible, interoperable, coherent with certain quality standards. Open access to research data is relating to the right to online access and re-use of digital research data under the conditions specified in the grant agreements. Accessing, mining, exploitation, reproduction and dissemination are free of charge. Justified exceptions must be explained, for example, in the interests of national security, protection of personal data and intellectual property rights of private co-financiers. Customer Information Control Systems (CICS) must be compliance with legal and ethical requirements to ensure open access. If the access to research data for justified exceptions is

limited, at least a freely accessible metadata must be available, from which it is clear where and under what conditions, research data are available."[22]

A particular challenge the existence of two distinct systems, one maintained by the University with its institutional repository and the other by the National Library. The actual laws on university libraries and the National Library do not mention digital dissertations, just that university libraries must obtain and process the compulsory copies of material that is created and published within the framework of the university, including graduate and masters theses and doctoral dissertations, and two copies of (print) doctoral dissertation are to be sent to the National Library. Electronic versions can be uploaded in Digital library of Slovenia, maintained by National and University Library of Slovenia, only with the written permission by authors.

A workflow comparison of the French STAR and the UK EThOS infrastructures with ProQuest's global schema (Walker 2011) and the TARDis project at the University of Southampton (Simpson & Hey 2006, Hey & Hey 2006) suggests that there may be no unique ideal solution but different options, depending on legal and technical conditions. For instance, research data can be handled and disseminated via centralized data management systems or decentralized collaborative systems (social networks) with reduced costs and customizable interfaces (Wang & Liu 2009). Data repositories can be institution-based (such as most ETD repositories) but also run by third-party service providers, such as Dryad, Zenodo or Figshare. One size does not fit all.

As a matter of fact, such heterogeneous datasets cannot be compared to the kind of big data produced by CERN and other large facilities but are more similar to personal data, even if the main challenges are roughly the same, covering issues of up-dates, enrichment and reuse, submission policy, handling of copyrighted material, standards, technical infrastructure or long-term preservation. The point is that the ideal system architecture should combine features of personal data stores (small data) with characteristics of institutional information systems (big data). For instance, how to decide on the inclusion and deposit of supplementary files? In big data repositories with automatic input, these decisions are taken ad hoc and upstream. Because of the link to copyright protected documents, and because of the personal nature of these research results, the same strategy cannot be applied here, and decisions have to be taken on a different level, probably case by case. But on which criteria?

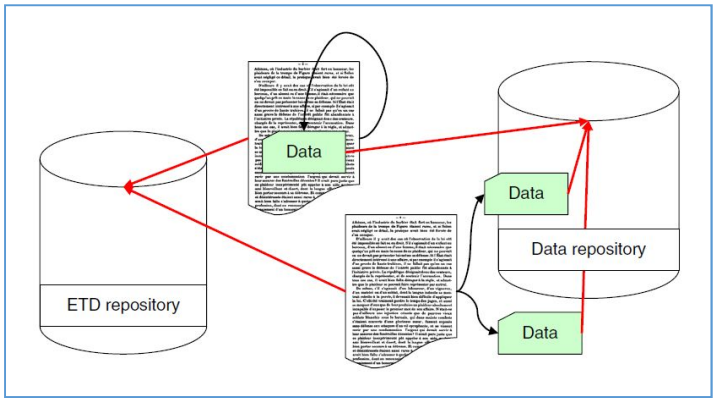


Figure 17: Storage of ETD as related datasets

Because of the specific nature of data and supplementary files (see above), it appears appropriate not to store text and data files in the same repository but to distinguish between document server and data repository and to deposit text and data files on different platforms, or at least to separate them on an early stage of the workflow and to handle them in different information system environments (see Figure 17). For instance, Sun et al. (2011) developed a database and associated computational infrastructure for datasets with different metadata submission forms for different topics. Supplementary material should not only be available as appendix or illustration to the related dissertations but also extractable and reusable without link to the thesis, as an independent dataset and interconnected to other data. In the Berlin survey cited above, scientists seem to prefer a local data repository (department, laboratory) to other solutions which means that they are realistic enough to require a combined institutional and disciplinary environment for their data (Simukovic et al. 2014, Prost & Schöpfel 2015).

9.Helping PhD students to manage their data

Considering these empirical results and the scientific interest, the University of Lille 3 decided to foster the data education of PhD students in social sciences and humanities, as a central part of its global approach to research data and open access[23]. Following the work of Reznik-Zellen et al. (2012) at the

University of Massachusetts Amherst, the Lille project team develops three tiers of research data support services for PhD students on our campus, including education, consultation and infrastructure (Figure 18).

Figure 18: Research data support services

Education: The University of Lille 3 organized three conferences on research data between February and April 2015, especially designed for PhD students in social sciences and humanities[24]. A first transdisciplinary doctoral seminar on research data management will be launched in January 2016, with seven units on data management plans, data life cycle, data description, storage, sharing etc. Another seminar will put the focus on data exploitation. At the same time, the project team will edit guidelines for good data practice and make them available for the PhD students, together with frequently asked questions and updates on data management, open data etc. The University of Lille 3 is not the first one to teach data management and exploitation, and it can build on experiences from other campuses, like the "University of Minnesota Data Management Course" with seven modules.[25] Another example is the "Data Management Bootcamp for Graduate Students" workshop series, a joint program of Virginia Tech and four other Virginia universities[26]. This Bootcamp provides data curation training with seven modules, including formats and transformation, data protection, and preservation, sharing and licensing. Also, the University of Virginia hosts a "Graduate Student Data Management Portal" that offers help to understand the research and data lifecycle (cf. Figure 20), practical guidance and links to useful tools.[27]

Advice and assistance: Probably as a part of its future Learning Centre, the University of Lille 3 will develop personalized help and assistance for PhD students, able to provide answers and advice to their specific questions and problems. This might include, too, advice and guidance regarding methods for de-identification of protected, sensitive personal data (health information, surveys etc.), such as the Safe harbour methods and following the Privacy rule[28]. Also, the role of supervisors should be valorised. Yet, a recent study reveals that most of them know a small amount or nothing about research data management or digital curation, and most of them have no or only limited skills or expertise in this area (Abbott 2015). Moreover, advice and assistance should build on external resources whenever possible, even if the same study points that the "use of specific external resources is low at under 10% and awareness for all specialised external resources was under 20%. This represents a missed opportunity in terms of outsourcing as much training as possible to dedicated experts" (loc.cit, p.15).

Infrastructures: The Lille 3 approach is based on intermediation, not on research and development. Probably, some basic tools for temporary storage and metadata will be developed and implemented on the campus. For instance, this might be a "data vault" for temporary storage of data files and/or a data asset register, synchronized with the institutional repository or the research management system[29]. Yet, the main idea is to partnership with existing data networks and repositories, including agreements if necessary and delegation of the deposit. Sometimes, especially for small "orphan" datasets, the solution may also simply be Zenodo or FigShare.

These three tiers of research data support services will be launched progressively between 2015 and 2018. Their development will follow five guiding principles (Figure 19).

Figure 19: Five leading principles for the implementation of research data support services

1. A discipline-specific approach: One size does not fit all. Research data support services must be flexible and adjusted to the scientific disciplines and domains of the PhD research. This means a very good knowledge of research methodologies, data types, formats etc. but also a good cooperation with the research teams, large projects and laboratories.
2. An integration into the doctoral education: Data management and sharing must become part of the mandatory doctoral education syllabus, such as project management, scientific writing or data analysis. The Lille 3 data literacy program will contribute to the creation of a culture of data management.
3. A proposal of data management plans: The University of Lille 3 will develop its own templates for data management plans, compliant with social sciences and humanities and the criteria of the European program H2020. The DMPonline tool developed by the JISC Digital Curation Centre to help write data management plans may be a helpful model.[30]
4. Incentives for the digital deposit of research data: Deposit of research data along with PhD dissertations should become near to mandatory. At least, there should be strong incentives to submit those data for temporary storage and long term preservation.
5. A contribution to the preservation and dissemination of data: Finally, as mentioned above, the University of Lille 3 will contribute to the preservation and dissemination of these research data – not necessarily with campus-based infrastructures (they are not excluded, though) but rather through partnerships and networking with local or national providers. We are already doing so in the field of open access, with good success, as our institutional repository is hosted by the Lyon-based CCSD[31] and part of the national open repository HAL[32].

The academic library, already present and engaged in ETD management and open access, will be a leading partner for these new research data support services, in cooperation with the graduate school and the research laboratories. Nevertheless, this leading position must become legitimate and accepted by the scientific community and the PhD students. So far, scientists and students obviously have not identified the academic library as a potentially useful structure for their data (Prost & Schöpfel 2015). In other words, the implementation of the new services must be accompanied by communication about the role and usefulness of each partner, and by the acquisition of new skills and knowledge by the information professionals for "data librarianship" (LIBER 2012).
One part of the new library function could be the promotion of research data citation by applying persistent identifiers to research data, such as DOIs. For instance, the Purdue University Research Repository provides DOI for research data. French libraries already assign a persistent identifier (code) to each dissertation. Yet, the ability to connect dissertations with the underlying data needs a consistent way to assign an appropriate level of granularity to sub-sections, appendices and related content. Moreover, due to this new function the academic library may also take responsibility in the field of persistent identifiers for authors (such as ORCID), for instance through assistance and advice for PhD students and young scientists to create and manage their identifiers.

10. Changing the way of doing PhDs

The empirical evidence of this study suggests that assistance and advice for PhD students to help them manage their research data must go beyond general rules and recommendations. Not all doctoral projects produce research data. Not all data are submitted with the dissertation to back up the research in the dissertation or to further explain and clarify the matter. Not all data can be reused especially, but not only, for legal reasons. And finally, even if our sample is not representative, it seems obvious that many characteristics of data sources and types have strong relationships with disciplinary methods, topics and approaches.

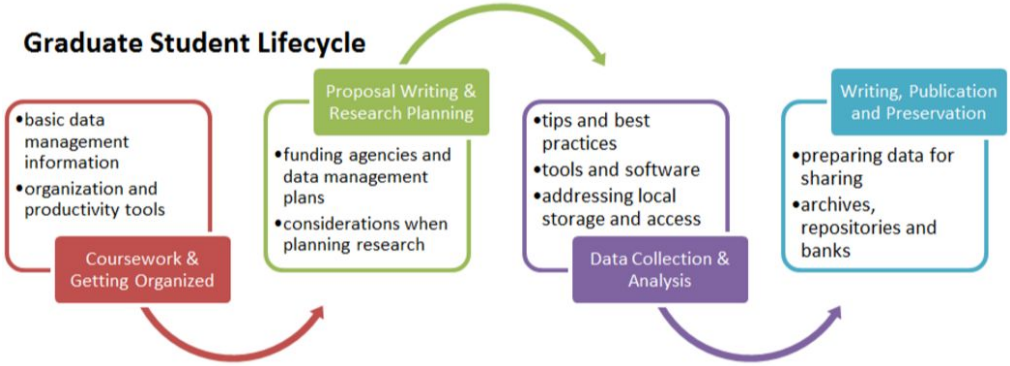


Figure 20: Research and data lifecycle (source: University of Virginia Data Management Portal)

Research data management and data curation cannot wait until the final stage of the PhD project. The student researchers must be aware of the data lifecycle from the beginning on, they must anticipate the legal and technical challenges of their collected and produced data, and they must know how to describe, store and share their data.

In an ongoing survey launched by NDLTD, 52 out of 104 US and Canadian universities that require electronic submission of dissertations allow deposit of supplemental material for doctoral work. In order to do this properly, in a way described above, text and data must be separated, with different metadata and identifiers. Also, the research data files or databases must be well structured and documented, with a detailed and organized tagging (markup) of the datasets. The data must be described in a standard language and format, with sufficient detail for retrieval and data mining.

The PhD students, with assistance from supervisors, colleagues and professionals, must make a thorough choice of formats and supports appropriate for the temporary storage, sharing and future deposit of their data in a data repository. Whenever possible, open formats should be preferred, to facilitate long-term preservation and re-use. Often, data repositories suggest data deposit in the original file format.

We already mentioned the need of clearing of privacy and copyright issues. Just like ethical aspects, these issues cannot wait and must be anticipated from the very beginning on, to be compliant with legal rules and to be able, at the end of the PhD work, to store and share the research results whenever possible.

Data management and curation change the way of doing PhDs, in two ways. The overall planning must include the different stages of the research data lifecycle, from the collection and creation to the preservation and enabling of re-use. On the other side, as the dissertation becomes a gateway to data, the structure and the format of the dissertation must allow the link to related, underlying data. The way to do this will be different between disciplines, domains and research communities. But it seems probable that the text writing and editing of a PhD dissertation will be facilitated because the data and other material moves out of the text. Some dissertations, at least some parts of them, might even become similar to data papers.

11. Perspectives

Open, digital science is work in progress. Along with documents and publications, research data become an essential part of scientific information. Electronic theses and dissertations have the potential to contribute to the emerging landscape of e-Science, as “data vehicles” as well as “gateways to data”. Higher Education and research organizations invested into infrastructures, repositories and library systems in order to facilitate the transition from print to digital dissertations. Today, new investment is needed for the curation of research data produced and deposited with ETD. The development of ETD infrastructures, open repositories and e-Science makes it possible to find an appropriate solution for the management and reuse of small data produced along with dissertations.

Dissertations often are “data vehicles” where research results are published together with the text of the dissertation. This makes sense in the print world but appears inappropriate in the digital environment of the 4th paradigm. Curation, retrieval and reuse would be largely facilitated if this material would be separated from the PhD text files and handled in a different way. This means, dissertations should be valorised as “gateways to data” which implies incentives for the deposit of related datasets and other supplementary files, minting DOIs (or other persistent identifiers) for research data and innovative procedures and workflows in graduate schools and academic libraries.

Furthermore, service functionalities from institutional repositories and data stores should be adapted to these specific items, with a flexible, user-centred approach. To increase accessibility and reuse and to avoid isolated data silos with multiple metadata entries, all developments should be as standardized as possible and with maximal interconnectivity, based on the OAI protocol. This means also that small data repositories should be, whenever possible, integrated in CRIS environments.

The JISC identified five key areas for actions in favour of research data management UK universities, i.e. policy development and implementation, skills and capabilities, infrastructure and interoperability, incentives for researchers and support stakeholders, and business case and sustainability (Brown et al. 2015). This framework describes the challenges research data projects in the field of dissertations have to face. Even if the basic idea of open access is simple, it is easy to underestimate the cultural barriers and the time required to work through them. The first step is always the hardest. Costello (2009) points out the fact that lack of support is one of the reasons why scientists don’t deposit their data in open repositories. Scientists remain committed to the values, norms and services of their institution and discipline which means that developing an infrastructure for electronic theses and dissertations and supplementary files will be successful if and only if supported by an explicit policy in favour of open

access and open data. This policy can be implemented locally and serve as a good example or show case, or nationally as the part of the accreditation systems. Either way, the awareness of the importance of open research data in dissertations should be a good basis for universities to change their policy to PhD dissertations accordingly.

References

Abbott, D., 2015. Digital curation and doctoral research. International Journal of Digital Curation 10 (1), 1–17. http://dx.doi.org/10.2218/ijdc.v10i1.328

Blake, J. A., Bult, C. J., 2006. Beyond the data deluge: Data integration and bio-ontologies. Journal of Biomedical Informatics 39, 314–320. http://dx.doi.org/10.1016/j.jbi.2006.01.003

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), 17–30. http://escholarship.org/uc/item/6fs4559s

Brown, S., Bruce, R., Kernohan, D., 2015. Directions for research data management in UK universities. JISC, Bristol. http://repository.jisc.ac.uk/5951/4/JR0034 RDM report 200315 v5.pdf

Bult, C. J., 2002. Data integration standards in model organisms: from genotype to phenotype in the laboratory mouse. TARGETS 1 (5), 163–168. http://www.sciencedirect.com/science/article/pii/S1477362702022158

Burnham, A., 2013. An introduction to managing research data for researchers and students. University of Leicester. http://www2.le.ac.uk/services/research-data/documents/an-introduction-to-managing-research-data

Carr, L., White, W., Miles, S., Mortimer, B., 2008. Institutional repository checklist for serving institutional management. In: Third International Conference on Open Repositories 2008, 1-4 April 2008, Southampton, United Kingdom. http://pubs.or08.ecs.soton.ac.uk/138/

Carroll, M. W., 2015. Sharing research data and intellectual property law: A primer. PLoS Biol 13 (8), e1002235+. http://dx.doi.org/10.1371/journal.pbio.1002235

Cassella, M., Calvi, L., 2010. New journal models and publishing perspectives in the evolving digital environment. IFLA Journal 36 (1), 7–15. http://www.ifla.org/files/assets/hq/publications/ifla-journal/ifla-journal-36-1 2010.pdf

Chaudiron, S., Maignant, C., Schöpfel, J., Westeel, I., 2015. Livre blanc sur les données de la recherche dans les thèses de doctorat. Université de Lille 3, Villeneuve d'Ascq. http://hal.univ-lille3.fr/hal-01192930

Costello, M. J., 2009. Motivating online publication of data. BioScience 59 (5), 418–427. https://researchspace.auckland.ac.nz/bitstream/handle/2292/7173/bio.2009.59.5.9.pdf

Cox, A., Verbaan, E., Sen, B., 2014. A spider, an octopus, or an animal just coming into existence? Designing a curriculum for librarians to support research data management. Journal of eScience Librarianship 3 (1). http://dx.doi.org/10.7191/jeslib.2014.1055

Doty, J., Kowalski, M. T., Nash, B. C., O'Riordan, S., 2015. Making student research data discoverable: A pilot program using dataverse. Journal of Librarianship and Scholarly Communication 3 (2). http://dx.doi.org/10.7710/2162-3309.1234

EU High Level Expert Group on Scientific Data, 2010. Riding the wave. how europe can gain from the rising tide of scientific data. European Union, Brussels. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

Halipré, A., Malleret, C., Prost, H., 2015. Les données de la recherche dans les thèses en SHS de l'Université de Lille 3 (poster). In: Journées ABES, 27-28 mai 2015, Montpellier. http://www.abes.fr/Media/Fichiers/Footer/Journees-ABES/JABES 2015 Poster SCD Lille 3

Heidorn, P. B., 2008. Shedding light on the dark data in the long tail of science. Library Trends 57 (2), 280–299. https://www.ideals.illinois.edu/bitstream/handle/2142/10672/heidorn.pdf?sequence=2

Hey, T., Trefethen, A. E., 2005. Cyberinfrastructure for e-Science. Science 308 (5723), 817–821. http://dx.doi.org/10.1126/science.1110410

Hey, T., Hey, J., 2006. e-Science and its implications for the library community. Library Hi Tech 24 (4), 515-528. http://dx.doi.org/10.1108/07378830610715383

Hey, T., Tansley, S., Tolle, K. (Eds.), 2009. The fourth paradigm. Data-intensive scientific discovery. Microsoft Corporation, Redmond, WA.

Higgins, S., 2008. Draft DCC curation lifecycle model. International Journal of Digital Curation 2 (2), 82–87. http://dx.doi.org/10.2218/ijdc.v2i2.30

Juznic, P., 2010. Grey literature produced and published by universities: A case for ETDs. In: Farace, D., Schöpfel, J. (Eds.), Grey Literature in Library and Information Studies. De Gruyter Saur, pp. 39–51.

Kindling, M., 2013. Doctoral theses' research data and metadata documentation. In: ETD 2013 Hong Kong 16th International Symposium on Electronic Theses and Dissertations 25 September 2013. http://lib.hku.hk/etd2013/presentation/Maxi-ETD-20130925.pdf

Koler-Povh, T., Lisec, A., 2015. Geodetski vestnik and its path to better international recognition. Geodetski vestnik 59 (02), 289–319. http://dx.doi.org/10.15292/geodetski-vestnik.2015.02.289-319

Koler-Povh, T., Mikoš, M., Turk, G., 2014. Institutional repository as an important part of scholarly communication. Library Hi Tech 32 (3), 423–434. http://www.emeraldinsight.com/doi/full/10.1108/LHT-10-2013-0146

Koler-Povh, T., Turk, G., 2011. Instructions for theses designing and citing on UL FGG = Navodila za oblikovanje zaključnih izdelkov študijev na FGG in navajanje virov. Ljubljana: Fakulteta za gradbeništvo in geodezijo, 63 p. ISBN 978-961-6167-97-0.

Kowalczyk, S., Shankar, K., 2011. Data sharing in the sciences. Annual Review of Information Science and Technology 45 (1), 247-294. http://courses.washington.edu/geog482/resource/9_Kowalczyk_DataSharingSciences.pdf

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. Tech. rep., Gartner META Group, Stamford CT. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

LIBER working group on E-Science / Research Data Management, 2012. Ten recommendations for libraries to get started with research data management. LIBER, The Hague. http://libereurope.eu/wp-content/uploads/The%20research%20data%20group%202012%20v7%20final.pdf

Lynch, C., 2009. Jim Gray's fourth paradigm and the construction of the scientific record. In: Hey, T., Tansley, S., Tolle, K. (Eds.), The fourth paradigm. Data-intensive scientific discovery. Microsoft Corporation, Redmond, WA, pp. 177–183.

Lynch, C., 2014. The need for research data inventories and the vision for SHARE. Information Standards Quarterly 26 (2), 29+. http://dx.doi.org/10.3789/isqv26no2.2014.05

Malleret, C., Prost, H., 2015. Les données de la recherche dans les thèses en SHS de l'Université de Lille 3. In: Séminaire DRTD-SHS "Les données de la recherche dans les humanités numériques", 2 février 2015, Lille.

McDowell, C. S., 2007. Evaluating institutional repository deployment in American academe since early 2005. D-Lib Magazine 13 (9/10). http://dx.doi.org/10.1045/september2007-mcdowell

McMahon, B., 2010. Interactive publications and the record of science. Information Services and Use 30 (1), 1–16. http://iospress.metapress.com/content/f4th457822023783/fulltext.pdf

Morris, R. W., Bean, C. A., Farber, G. K., Gallahan, D., Jakobsson, E., Liu, Y., Lyster, P. M., Peng, G. C. Y., Roberts, F. S., Twery, M., Whitmarsh, J., Skinner, K., Mar. 2005. Digital biology: an emerging and promising discipline. TRENDS in Biotechnology 23 (3), 113–117. http://cmbi.bjmu.edu.cn/news/report/2004/biotech/24.pdf

Murray-Rust, P., 2007. The power of the electronic scientific thesis. In: ETD 2007 10th International Symposium on Electronic Theses and Dissertations, June 13-16, 2007, Uppsala, Sweden. http://epc.ub.uu.se/ETD2007/sessions/keynote-2.html

Murray-Rust, P., 2008. Open data in science. Serials Review 34 (1), 52–64. http://www.dspace.cam.ac.uk/bitstream/1810/194892/1/opendata.html

Neuroth, H., Strathmann, S., Oßwald, A., Ludwig, J. (Eds.), 2013. Digital curation of research data. Experiences of a baseline study in Germany. vwh, Glückstadt. http://www.nestor.sub.uni-goettingen.de/bestandsaufnahme/Digital Curation.pdf

Ojsteršek, M., Brezovnik, J., Kotar, M., Ferme, M., Hrovat, G., Bregant, A., Borovič, M., 2014. Establishing of a Slovenian open access infrastructure: a technical point of view. *Program* 48 (4), 394–412. http://dx.doi.org/10.1002/asi.20663

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P., Dierolf, U., 2013. Making research data repositories visible: The re3data.org registry. PLoS ONE 8 (11), e78080+. http://dx.doi.org/10.1371/journal.pone.0078080

Prost, H., Malleret, C., Schöpfel, J., 2015. Hidden treasures. opening data in PhD dissertations in social sciences and humanities. Journal of Librarianship and Scholarly Communication 3 (2), eP1230+. http://dx.doi.org/10.7710/2162-3309.1230

Prost, H., Schöpfel, J., 2015. Les données de la recherche en SHS. Une enquête à l'Université de Lille 3. rapport final. Université de Lille 3, Villeneuve d'Ascq. http://hal.univ-lille3.fr/hal-01198379v1

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., Wilkinson, M., 2011. Report on integration of data and publications. ODE Opportunities for Data Exchange, The Hague. http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf

Savage, C. J., Vickers, A. J., 2009. Empirical study of data sharing by authors publishing in PLoS journals. PLoS ONE 4 (9), e7078+. http://dx.doi.org/10.1371/journal.pone.0007078

Savic, D., 2015. INIS: Nuclear grey literature repository. In: 8th Conference on Grey Literature and Repositories, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic. https://www.techlib.cz/en/83294-conference-on-grey-literature

Schöpfel, J., Chaudiron, S., Jacquemin, B., Prost, H., Severo, M., Thiault, F., 2014. Open access to research data in electronic theses and dissertations: An overview. Library Hi Tech 32 (4), 612–627. http://www.emeraldinsight.com/doi/abs/10.1108/LHT-06-2014-0058

Schöpfel, J., Farace, D. J., 2010. Grey literature. In: Bates, M. J., Maack, M. N. (Eds.), Encyclopedia of Library and Information Sciences, Third Edition. CRC Press, London, pp. 2029–2039. http://dx.doi.org/10.1081/e-elis3-120043732

Schöpfel, J., Lipinski, T. A., 2012. Legal aspects of grey literature. The Grey Journal 8 (3), 137–153. http://archivesic.ccsd.cnrs.fr/sic_00905090/fr/

Schöpfel, J., Prost, H., 2013. Degrees of secrecy in an open Environment. The case of electronic theses and dissertations. ESSACHESS - Journal for Communication Studies 6 (2 (12)). http://www.essachess.com/index.php/jcs/article/view/214

Schöpfel, J., Prost, H., Malleret, C., 2015a. Making data in PhD dissertations reusable for research. In: 8th Conference on Grey Literature and Repositories, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic. https://www.techlib.cz/en/83294-conference-on-grey-literature

Schöpfel, J., Prost, H., Piotrowski, M., Hilf, E. R., Severiens, T., Grabbe, P., 2015b. A French-German survey of electronic theses and dissertations: Access and restrictions. D-Lib Magazine 21 (3/4). http://www.dlib.org/dlib/march15/schopfel/03schopfel.html

Schultz, M., Krabbenhoeft, N., Skinner, K. (Eds.), 2014. Guidance Documents for Lifecycle Management of ETDs. Atlanta, GA. http://metaarchive.org/public/publishing/Guidance Documents for Lifecycle Management of ETDs.pdf

Sengupta, S. S., 2014. E-thesis repositories in the world: A critical analysis. Ph.D. thesis, Savitribai Phule Pune University. http://pqdtopen.proquest.com/doc/1696933497.html?FMT=ABS

Shotton, D., 2012. The five stars of online journal articles - a framework for article evaluation. D-Lib Magazine 18 (1/2). http://dx.doi.org/10.1045/january2012-shotton

Siegel, E. R., Lindberg, D. A. B., Campbell, G. P., Harless, W. G., Goodwin, C. R., 2010. Defining the next generation journal: The NLM–Elsevier interactive publications experiment. Information Services and Use 30 (1), 17-30. http://dx.doi.org/10.3233/isu-2010-0608

Simpson, P., Hey, J., 2006. Repositories for research: Southampton's evolving role in the knowledge cycle. Program: electronic library and information systems 40 (3), 224-231. http://eprints.soton.ac.uk/41240/1/ProgramAug2006simpsonrev1final2.pdf

Simukovic, E., Kindling, M., Schirmbacher, P., 2014. Unveiling research data stocks: A case of Humboldt-Universität zu Berlin. In: iConference, 4-7 March 2014, Berlin. pp. 742–748. https://www.ideals.illinois.edu/handle/2142/47259

Suber, P., 2012. Open access. MIT Press, Cambridge Mass. http://mitpress.mit.edu/books/open-access

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., Stocks, K., Allen, E. E., Ellisman, M., Grethe, J., Wooley, J., 2011. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. Nucleic Acids Research 39 (suppl 1), D546-D551. http://dx.doi.org/10.1093/nar/gkq1102

Vompras, J., Schirrwagen, J., 2015. Repository workflow for interlinking research data with grey literature. In: 8th Conference on Grey Literature and Repositories, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic. https://www.techlib.cz/en/83294-conference-on-grey-literature

Walker, E. P., 2011. What we can learn from ETDs: Using ProQuest dissertations & theses as a dataset. In: USETDA 2011: The Magic of ETDs…Where Creative Minds Meet. May 18-20, Orlando, Florida. https://conferences.tdl.org/USETDA/USETDA2011/paper/view/368

Wang, S., Liu, Y., 2009. TeraGrid GIScience gateway: Bridging cyberinfrastructure and GIScience. International Journal of Geographical Information Science 23 (5), 631–656. http://www.cigi.illinois.edu/publications/2009/GIScienceGateway-IJGIS-Wang-etal.pdf

All references are here: http://www.citeulike.org/user/Schopfel/tag/gl17

---

[1] Cf. the French Secretary of State for Digital Affairs Axelle Lemaire opening Big Data Paris 2015 http://www.digitalforallnow.com/en/big-data-paris-2015-fuel-of-the-digital-economy/

[2] EU High Level Expert Group on Scientific Data, 2010. *Riding the wave. How Europe can gain from the rising tide of scientific data*. European Union, Brussels. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[3] https://www.kth.se/en/forskning/forskningsplattformar/ict/forskning/e-vetenskap-1.323973

[4] https://www.whitehouse.gov/omb/circulars_a110/

[5] http://www.re3data.org/

[6] http://www.lancaster.ac.uk/library/rdm/plan/data-lifecycle/

[7] See for instance the much more detailed model of the JISC Digital Curation Centre, presented by Higgins (2008).

[8] https://kclpure.kcl.ac.uk/portal/en/

[9] http://datascience.iq.harvard.edu/blog/bridge-publishing-words-publishing-data

[10] In the following, we will use the term dissertation to designate the written contribution to obtain a PhD degree.

[11] See for instance the list on the FOSTER website https://www.fosteropenscience.eu/foster-taxonomy/open-data-journals

[12] Figure from https://www.elsevier.com/connect/can-data-be-peer-reviewed

[13] http://etd2016.sciencesconf.org/

[14] http://www.dart-europe.eu/

[15] http://search.ndltd.org/

[16] http://educopia.org/research/electronic-theses-and-dissertations

[17] https://educopia.org/research/grants/etdplus

[18] Study conducted by Teja Koler Povh

[19] http://drugg.fgg.uni-lj.si

[20] https://pub.uni-bielefeld.de/

[21] http://repository.tudelft.nl/

[22] https://www.arrs.gov.si/sl/obvestila/15/odprti-pristop-20152020.asp (in Slovenian)

[23] See the Lille 3 White Paper on research data in PhD dissertations http://hal.univ-lille3.fr/hal-01192930v1 (Chaudiron et al. 2015)

[24] http://drtdshs2015.sciencesconf.org/

[25] https://sites.google.com/a/umn.edu/data-management-workshop-series/home-1?pli=1

[26] https://www.research.vt.edu/announcements/data-management-bootcamp-offered-graduate-students

[27] https://pages.shanti.virginia.edu/SciDaC_Grad_Training/

[28] http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#standard

[29] See Stuart Lewis' blog posts on the Ediburgh Research Data Blog http://datablog.is.ed.ac.uk/

[30] https://dmponline.dcc.ac.uk/

[31] https://www.ccsd.cnrs.fr/

[32] http://hal.univ-lille3.fr/

https://doi.org/10.26069/greynet-2025-000.540-tgj

Can uptake of public health interventions be improved by including grey literature in the evidence-base?

Jess Tyndall (Australia)

Abstract

Interventions to prevent obesity are not working. Despite the substantial efforts of a multiplicity of interventions and strategies from the public health sector, by 2025 Australian adult obesity rates are predicted to rise by 65%. There is little doubt that increased physical activity/decreased sedentary behaviour combined with dietary intervention produces an outcome of reduced BMI: this is the science of weight loss and not in dispute, though a profusion of RCTs continue to be conducted along similar lines and academic journals continue to publish them. Policy in this area abounds but practical long-term successes remain elusive.

What is often missing from the debate is acknowledgement and understanding of the public's resistance to uptake of obesity prevention interventions. This crucial information can be gleaned from the grey literature. To be genuinely evidence-informed, policy needs to access the broadest literature base and include both published and unpublished sources. Grey literature includes surveys, public opinion, crowdsourcing, social media, theses and conference papers, issues and working papers, technical reports, unpublished studies and projects, newspapers and blogs. It is vital to consider these sources in the evidence-base to balance the science of obesity.

Grey literature is where societal attitudes, beliefs, values and opinions might be reasonably expected to be found, and also where pilot programs and case studies assessing the feasibility of public health interventions will have been reported. The degree to which this literature has been accessed and incorporated into the evidence that informs policy arguably has an impact on the ultimate success or sustained "take-up" of an intervention.

Introduction

The Australian Institute of Health and Welfare (2010) defines public health interventions as those which "focus on maintaining the wellbeing of populations … and are aimed at protecting or promoting health or preventing illness" and to be successful, these interventions need to impact people's behaviour.

Changing people's health behaviour is notoriously difficult because human beings are complex and unique. Theories which explain behaviour and predict probability can be useful, but are too often derived from optimum research environments which do not translate successfully into the uncertainty and variability of real life. And there is a need for people to 'buy in' to change. "People change if they believe there is both value in change and that it is achievable" (Kok, 1997).

One public health intervention which has achieved a clear measure of success is tobacco control. In 1971 39% of adult Australians smoked and in 2015 that number was reduced to 14.5% with a government aim of a new low of 10% by 2018. Forty five years of continuing policy implementation and legislation has succeeded in significantly changing specific health behaviour with a large degree of 'buy in' which has been sustained. While the example of tobacco control shows what can be achieved, 45 years is too long to wait for the successful uptake of obesity prevention interventions.

Background

There is an urgency accompanying the public health crisis of obesity which has been called a pandemic and a 'time bomb'. No country has yet succeeded in reversing obesity trends: in fact by 2025 obesity rates are predicted to rise in Australia by 65% (Walls, 2012). This is despite a great diversity of interventions and strategies from the public health sector, supported by legislation targeting advertising and labelling of food, national consumer education programs, limiting availability of unhealthy foods sold in schools and hospitals and much, much more. As a public health challenge obesity is proving frustratingly difficult to tackle and there is a pressing need for solutions. For those in the developed world, eating is a pleasurable everyday activity and immediate gratification often wins out over long term, delayed, negative health outcomes. Obesity is a very complex issue with causes and contributors ranging between "personal choices" and "external forces" (Kibbe, 2011) and influenced by biological, behavioural, social, psychological, emotional, environmental and economic factors. There is a plethora of strong published evidence in the form of studies, trials and systematic reviews which stress the dangers of obesity and the success of higher energy output over intake. There is little doubt that increased physical activity/decreased sedentary behaviour combined with dietary intervention produces an outcome of reduced BMI if behavioural change is maintained over the long term. This weight loss

science is not in dispute, but by any measure the published research detailing outcomes of interventions and programs is clear: these interventions are largely unsuccessful and rarely sustained over time. People are not consistently compliant.

Clearly 'knowledge transfer', defined by the Canadian Social Sciences and Humanities Research Council (2006) as "Moving research knowledge actively into the realms of policy, practice, products, and public understanding", is still a significant challenge. Somehow there has been a disconnect in the general population between this knowledge and the necessary behavioural change for obesity prevention. With a lack of consensus and a need to gain greater understanding of where the community stands on this issue it is suggested that increased engagement with the grey literature may provide some answers. Access to this information which is largely missing from the current evidence-base may lead the way to improved uptake.

Unpublished studies and trials

Grey literature in the form of unpublished trials must be retrieved and considered as part of the research literature, to avoid publication bias. This approach is mandated by both the Cochrane Collaboration and the Campbell Collaboration. It is known that larger studies with positive results are more likely to be published than smaller studies or those with negative results (Hopewell, 2007) and in the obesity prevention literature the ensuing imbalance has been less than helpful. It has placed undue emphasis on intervention studies which continue to expound the scientific and biomedical aspects of weight loss and prioritised quantitative research (Perez, 2015). This "science to service" model often sees a simplification or generalisation of outcome interpretation for groups or populations which were not indicated in the research and, for example, reports on how many participants achieved a decrease in BMI but not why those who dropped out of the programs did so. Unpublished studies with negative outcomes need to be accessible as do unpublished smaller studies which provide a close look at a very specific population. Close examination of both the published and the unpublished research may provide greater insight into the difference between success and failure of interventions. After all, to a large extent the participants themselves hold the key.

Traditional grey literature sources

Grey literature from the obesity prevention community of practice is also crucial to understanding the challenges to successful uptake of interventions. It was pleasing to note that the most recent "Australian Dietary Guidelines: providing the scientific evidence for healthier Australian diets" (NHMRC, 2013) consulted over 200 references from the grey literature, making up 20% of their evidence base. Included were multiple reports, evidence reviews, guidelines, working and background papers, conference papers, standards, opinion papers and policy briefs. Moreover their Terms of Reference stipulated they actively seek "comments provided by the broader community through public consultation".

This consultation process for a national government publication endorses at high-level, the view that societal attitudes, beliefs, values and opinions are expected to be found in the grey literature, along with practice-based evidence reports assessing the feasibility of public health interventions. The degree to which this literature has been accessed and incorporated into the evidence that informs policy may have an impact on the ultimate success or sustained "take-up" of an intervention.

Recognition that the strength and value of grey literature lays in the way it can reflect the lay voice: what people actually think, is vital. How can policy succeed when individuals' beliefs and behaviours are at odds with the underlying principles and truths of achieving healthy weight? Grey literature is where we find this voice: in public opinion polls and surveys, in focus groups and in policy reports and working papers. Listening to people is part of the community engagement so vital to any public health change: it is a challenging but well-supported aspect of behavioural change in public health.

New types of grey literature

Innovative grey literature sources like blogs and social media, talkback radio, crowdsourcing, online competitions and interactive websites can breathe fresh life into public issues, broadening and enlivening the debate and further informing the evidence base. In 2012 the Childhood Obesity Challenge was launched. It was an online competition open to everyone and designed to solicit ideas and proposals on how to address the childhood obesity epidemic. This crowdsourcing platform was designed to have wide appeal and was open to the use of anything that could relay a possible solution, including the latest technologies and use of social media. Organisers said popularity voting was to encourage participants to create solutions for childhood obesity that would "rank well with the general public" and expected them to come from "outside the standard scientific framework".

In 2013 the University of Vermont also used crowdsourcing and set up an interactive website to gather information about childhood predictors of adult BMI from the general public. People discussing the issues with each other zeroed-in on predictors of obesity, some which were previously undocumented. Researchers were surprised that "the website visitors discovered some intriguing connections that experts hadn't considered".

As Australian obesity prevalence continues to rise it is reasonable to suggest that we must accept that traditional approaches to obesity prevention have been largely unsuccessful and that the evidence supporting them are at best under-informed. Looking further afield opens up exciting possibilities. Innovative examples from the grey literature illustrate how looking outside academia and the published research provides additional rich information content drawn directly from the people whose behaviour we seek to change. Tapping into the widest variety of sources allows us to reflect the complexities of the issue and the perceptions of those we target (Thomas, 2010).

A key to uptake: Community receptiveness

All the additional grey evidence gathered from this multiplicity of sources provide a much broader understanding of the way the public and the practice community think about obesity prevention. One way to achieve increased uptake of prevention interventions and bring about a change in health-related behaviour is to tie into community receptiveness.

With a community, a traditional needs analysis approach can be creatively substantiated with events like town hall meetings and crowdsourcing, which involve the general public contributing to the debate in innovative ways. Simkhada's view (2004) is that grey literature is needed to fill the information gap as "an essential part of the evidence base for practice in complex interventions, which may have multi-stakeholders, have multi-variables, have a lack of predictability and robust data and require a broad literature approach". .Armed with a far richer knowledge of public views, and actively involving stakeholders in programme design, implementation and evaluation, it is possible to ensure that the community's needs are met, and that the public is supportive and invested. There is a demonstrable need to clearly identify the issues, values, beliefs and attitudes of those the intervention aims to engage when designing and planning a public health intervention which aims at changing their behaviour. There are numerous tools for assessing community readiness for change, and many involve the identification of 'change champions'. These are usually respected and integral members of the community who can assist the change process. Tsui (2006) suggests the message must be "one that is valuable to an audience based on their needs, delivered by a messenger they can trust, in language they are comfortable with" and Flynn (2006), discussing the community prevention of obesity, includes a "Sensitivity check". Essentially this sensitivity check can usefully be applied to all ideas, strategies and interventions to assess if "they empower … [people] to take positive action, are culturally aligned, as well as validate community/population readiness".

The innovative nature of such approaches, often developed in partnership with communities, reflect and incorporate local need and can inform the evidence on supporting people and communities to make positive behavioural changes.

Conclusion

Grey literature provides the voice of the public and the practice community and brings a bottom-up perspective to the obesity prevention debate, to temper the top-down approach. This additional evidence may help address the question: why is obesity prevention not working? With greater understanding of public values, beliefs and attitudes there is every possibility of achieving increased engagement and uptake in the middle ground of community consensus.

References

Australian Institute of Health & Welfare. Australia's Health 2010. http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=6442452957

Flynn MAT, McNeil DA, Maloff B, Mutasingwa D, Wu M, Ford C, et al. Reducing obesity and related chronic disease risk in children and youth: a synthesis of evidence with 'best practice' recommendations. Obesity Reviews. 2006; 7 (Suppl. 1): 7-66.

Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database of Systematic Reviews, 2007. http://onlinelibrary.wiley.com/doi/10.1002/14651858.MR000010.pub3/pdf

Kibbe D. (Andrew Young School of Policy Studies, Georgia State University, GA). Health by stealth, finding balance between policy and individual responsibility: experiences from the United States. Paper presented at: 2nd ILSI Southeast Asia-Center for Health Innovation and Partnership Conference; 2011 Dec. 8; Sydney.

Kok G, Van der Borne B, Mullen, PD. Effectiveness of health education and health promotion: meta-analyses of effect sizes and determinants of effectiveness. Patient Education and Counseling. 1997; 30:19–27.

National Health and Medical Research Council. Australian Dietary Guidelines Canberra: NHMRC; 2013. http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/n55_australian_dietary_guidelines_130530.pdf

Perez A, Ball, GDC. Are we overlooking the qualitative 'look' of obesity? Nutrition & Diabetes, 2015; 5(7):e174

Simkhada P, Knight J, Reid G, Wyness L, Mandava L. Chasing the grey evidence: a standardised systematic critical literature review approach: Paper presented at: Work on Grey in Progress. 6th International Conference on Grey Literature; 2004 December 6-7; New York. http://www.opengrey.eu/data/69/78/51/GL6_Simkhada_et_al_2005_Conference_Preprint.pdf

Social Sciences and Humanities Research Council of Canada. Moving forward as a Knowledge Council: Canada's Place in a Competitive World: Statement to the House of Commons Standing Committee on Finance. Ottawa: The Council; 2006

http://www.sshrc-crsh.gc.ca/about-au_sujet/publications/finance_committee_e.pdf

Thomas SL, Lewis S, Hyde J, Castle D, Komesaroff P. "The solution needs to be complex." Obese adults' attitudes about the effectiveness of individual and population based interventions for obesity. BMC Public Health 2010; 10:420 [http://www.biomedcentral.com/1471-2458/10/420

Tsui L. A handbook on knowledge sharing: strategies and recommendations for researchers, policymakers and service providers. Alberta, CA. Community-University Partnership for the Study of Children Youth and Families, University of Alberta: 2006. http://www.uws.edu.au/__data/assets/pdf_file/0018/405252/Knowledge_Sharing_Handbook.pdf

Walls HL, Magliano DJ, Stevenson CE, Backholer K, Mannan HR, Shaw JE, et al. Projected Progression of the Prevalence of Obesity in Australia. Obesity. 2012; 20: 872–878

https://doi.org/10.26069/greynet-2025-000.540-tgj

Undiscovered Value of Grey Astronomy Education Research Results *

Timothy F. Slater (United States)

For the teaching scholar focused on improving their astronomy class, a perfectly reasonable question to pose is, "how do I quickly learn about the best that discipline-based astronomy education research has to offer about improving teaching and learning?" Exhaustive reviews of astronomy education research (Adams & Slater, 2000; Bailey & Slater, 2003, 2005; Slater, 2008; Slater, Ratcliff, & Tatge, 2016), clearly position astronomy education research, AER, as a its own scholarly discipline existing within a rich field of robust student misconceptions and varied instructional strategies designed to intellectually engage students. Indeed, the recent publication of National Research Council publication, *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering* (NRC, 2012) highlights astronomy as one of the principle disciplines of education research, elevating it to the same level as physics education research, geoscience education research, chemistry education research, engineering education research, and biology education research (listed in no particular order). Yet, if it is so important, where is all the astronomy education research?

Unfortunately, because AER is still a fledgling field in comparison to its older brother of physics education research, PER, much of the AER created scholarly knowledge is just beginning to show up in refereed journals. Beyond the some 150 studies summarized in reviews from refereed journals (reviews cited above), most of astronomy education research seems just incredibly difficult to find. Some efforts like SABER (http://astronomy.uwp.edu/saber/) have worked to gather disparate astronomy education research articles in journals beyond the Astronomy Education Review (http://aer.aas.org), but it just seems plausible that more research reports should be out there, doesn't it?

Most of astronomy education research that currently exists is hiding in deep in what is known colloquially as the "grey literature." Grey literature is the scholarly work that has been done and presented at conferences or exhaustively written up in dissertations, but never formally published in refereed journals. For example, at every American Astronomical Society and American Association of Physics Teachers meetings, there are special sessions, both oral and poster, overflowing with astronomy education research scholarship reports. Bailey (2010) reports that in a recent search of the SAO/NASA Astrophysics Data System (ADS), a query of "education" in abstracts for oral or poster presentations at American Astronomical Society (AAS) meetings yielded more than 1300 abstracts since 1992; nearly 600 of those were from the year 2006 or later. And there are similarly impressive numbers of astronomy education abstracts, both invited and contributed from meetings of the American Association of Physics Teachers sub-section now called space science and astronomy. This is the domain of the grey literature.

The best example of the AER grey literature is probably the most well-known study of understanding in astronomy ever done. It was publicized by Philip Sadler and is presented not in a printed publication, but instead in a video called *A Private Universe* (Schneps 1989). The video begins with clips of interviews with several alumni, faculty, and graduating seniors from Harvard University. Of the 23 individuals interviewed, 21 could not give a scientifically acceptable explanation for the cause of the seasons or the phases of the Moon. Although the misconceptions so cleverly illustrated by *A Private Universe* were garnered through methods that probably do not conform to the rules of reliability and validity that define today's DBER, the video's widespread influence cannot be understated. The *Private Universe* video served as wake-up call for the astronomy education community, much like Hestenes and colleagues (1992) *Force Concept Inventory* served as a rallying point for PER. For many, *Private Universe* marks a major milestone in the evolution of AER. Yet, nowhere does *Private Universe* show up as a refereed journal article.

As another example, consider the *Test Of Astronomy STandards*, the TOAST. Conceived of by Stephanie Slater and colleagues, this pretest-posttest, multiple-choice survey of astronomy knowledge is widely used across the international astronomy education research community. Yet, its only formal citation to date is from an appendix of the book, "Discipline-Based Science Education Research: A Scientist's Guide, 2nd Edition" published by Pono Publishing. Books are only sometimes considered by university tenure and promotion review committees as refereed scholarship, yet often hold a treasure trove of wonderful work. It is worth remembering that sometimes AER scholars have so many great projects going on that they just don't get around to formally publishing everything that is meritous. In fact, this isn't true of just AER, but is a rarely whispered truth to all of scientific scholarship that we rarely tell students about.

---

* First published in CAPER Astronomy Faculty Lounge https://astronomyfacultylounge.wordpress.com/2015/09/24/is-the-best-astronomy-education-research-grey/ under the title: Is the best astronomy education research 'grey'? (September 24, 2015)

If you really want to see the magic of the grey literature, consider graduate theses and dissertations. For one, dissertations are a secret place where exhaustive reviews of the literature are found. It is often in the second chapter of discipline-based education research dissertations where work is widely reviewed within the sometimes dreaded "literature review" chapter. Found as PDF's at istardatabase.org, nearly 300 astronomy education research dissertations review tens, if not hundreds of studies, conference presentations, poster contributions, and other theses and dissertations. Many authors publishing reviews of research rarely include graduate theses and dissertations, despite the fact that most have been more thoroughly reviewed by a larger number of scholars—those four, five, or six scholars who sit on graduate review committees—than those two or three who review manuscripts for journals. And this doesn't even consider the numerous Internet blogs out there, such as astronomy's AstroLrner.org blog. Together, this represents a tremendous amount of work within dissertations that review, synthesize, and recast earlier work that is well worth the time to consider, even though it is considered tertiary "grey literature."

Overall, when encountering difficulty in trying to find the very best and most influential astronomy education research work, nascent scholars know that you are not alone. Most Ph.D. dissertations are never published, and an even smaller percentage of conference presentations end up as refereed journal articles. Much of what was retrieved for the CAPER Center for Astronomy & Physics Education Research's *international STudy on Astronomical Reasoning* istardatabase.org is decades old and have never been cited in any recent research literature reviews. Just because something never makes it into a formal refereed journal, doesn't mean that it is bad work; in fact, there are plenty of published refereed journal articles that represent bad scholarship. Rather, it just means that the scholarship cycle was never completed. In reviewing astronomy research, scholars need to be a critical consumer of the research being considering and realize that the formal citation rate is but one of many aspects you should use to judge a research report's relative importance.

References:

Adams, J. P., & Slater, T. F. (2000). Astronomy in the National Science Education Standards. Journal of Geoscience Education, 48(1), 39-45.

Bailey, J.M. (2010). Astronomy Education Research: Developmental History of the Field and Summary of the Literature. Paper commissioned by the National Research Council, http://www7.national-academies.org/bose/DBER_Janelle_Bailey.pdf

Bailey, J. M., & Slater, T. F. (2003). A review of astronomy education research. Astronomy Education Review, 2(2), 20-45. doi: 10.3847/AER2003015. http://aer.aas.org/resource/1/aerscz/v2/i2/p20_s1

Bailey, J. M., & Slater, T. F. (2005). Resource letter AER-1: Astronomy education research. American Journal of Physics, 73(8), 677-685

Hestenes, D., Wells, M. and Swackhamer, G. (1992). Force concept inventory. The Physics Teacher, 30: 141-158.

National Research Council (2012). Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering, 2012, http://www.nap.edu/catalog.php?record_id=13362

Test Of Astronomy STandards, TOAST. In "Discipline-Based Science Education Research: A Scientist's Guide" by Slater, Slater, Heyer, & Bailey, Hawai'i: Pono Publishing at: https://www.amazon.com/Discipline-Based-Education-Research-Guide-Scientists/dp/1515024563

Slater, T. F. (2008). The first big wave of astronomy education research dissertations and some directions for future research efforts. Astronomy Education Review, 7(1), 1-12. doi: 10.3847/AER2008001, http://aer.aas.org/resource/1/aerscz/v7/i1/p1_s1

Slater, T. F., Ratcliff, M., & Tatge, C. B. (2016). Research on Learning in the Planetarium. Netherlands: Springer Publishing.

Schneps, M. P. 1989, *A Private Universe* [Video], Found at http://www.learner.org/resources/series28.html

WikiPedia entry for Grey Literature, http://en.wikipedia.org/wiki/Gray_literature

https://doi.org/10.26069/greynet-2025-000.540-tgj

A semantic engine for grey literature retrieval in the oceanography domain*

Sara Goggi, Gabriella Pardelli, Roberto Bartolini, Francesca Frontini, Monica Monachini, Giuseppe Manzella, Maurizio De Mattei and Franco Bustaffa (Italy)

Abstract

Here we present the final results of the MAPS (Marine Planning and Service Platform) project, an environment designed for gathering, classifying, managing and accessing marine scientific literature and data, making it available for search to Operative Oceanography researchers of various institutions by means of standard protocols. The system takes as input non-textual data (measurements) and text - both published papers and documentation - and it provides an advanced search facility thanks to the rich set of metadata and, above all, to the possibility of a refined and domain targeted key-word indexing of texts using Natural Language Processing (NLP) techniques. The paper describes the system in its details providing also evidence of evaluation.

1. Introduction and background

Efficient textual search and information retrieval over large quantities of data is one of the most important algorithmic challenges of our times, and we are all aware of the importance of generalist search engines in everyday life. Web search is but one of the main applications of what is generally known as *text mining*, namely the study of algorithms that are capable of extracting structured data from unstructured text (Ramjan et al., 1998). More specifically, Information Retrieval (IR) is a sub-field of text mining that focuses on retrieving a given piece of information from a document base (Manning, et al., 2009). It is easy to see why these technologies are relevant for grey literature studies: they are capable of enhancing efficient use and querying of internal document bases in any institution and it is clear that making documents easily retrievable and accessible is an important step for their correct storing, organization and preservation.

This is particularly true for any scientific domain and in cases when document bases describing and accompanying scientific data exist. The present paper focuses on a case-study on Operational Oceanography, the branch of marine research which deals with the development of integrated systems for monitoring, analyzing, modeling and forecasting oceans and seas. These integrated systems need access to real-time as well as delayed mode data. The operational oceanography activities includes also the analysis of data and models to assess marine variability or long term trends. These activities require the access to quality controlled data and to information that is provided in reports and/or in relevant scientific literature. This finds application in many areas, ranging from environmental studies, security and safety to protection of off-shore and coastal infrastructures.

Hence, creation of new technology is needed by integrating several disciplines such as data management, information systems, knowledge management and others. Full-text technologies are often unsuccessful when applied to this type of queries since they assume the presence of specific keywords in the text; in order to fix this problem, the MAPS project suggests to use different NLP technologies for retrieving text and data and thus getting much more complying results.

Up to recently, basic textual search systems were built on a so-called “bag of words” approach. Documents were retrieved and segmented into words; content words were identified based on function words lists and a very basic indexing was performed. Measures based on term frequency and representativeness were then used to map the words in the users’ query with the most appropriate documents.

Nowadays the development of NLP technologies has made it possible to improve on such basic techniques, allowing for a more refined analysis of both the texts of the document base and of the query and for a more linguistically aware match between the two. We refer to such innovative systems as semantic engines, meaning that they take into account the meaning of words and not only their superficial form. This can be achieved in many ways, but generally a more refined morpho-syntactic analysis of the text and a domain specific extraction of complex terms is involved, often with the use of ad-hoc constructed terminological and ontological resources. In a previous publication on the general architecture of MAPS (Goggi et al., 2015) we provided an overview of the state of the art on semantic engines.

* First published in the GL17 Conference Proceedings, February 2016.

In the present paper, we shall provide a full description of the final version of the MAPS search engine, its general architecture, its NLP modules, its terminology extraction system and query analysis system. Moreover, we shall report on experiments carried out to evaluate MAPS as in information retrieval system using a corpus of texts and a series of pre-selected queries. We shall conclude with some reflections on this experiment, its applicability for other disciplines and its relevance for grey literature studies.

2. The MAPS library

First of all, a brief introduction to the MAPS library is necessary. The MAPS library contains Oceanographic data-sets from various agencies and various instruments measuring different aspects of the sea environment. The data are defined by types of measure, measurement tools, geographic areas and also linked to specific textual documentation which can be composed of both published scientific papers and of actual documentation (working papers, manuals, dataset descriptions, reports), all fitting in the definition of grey literature. Such documents are defined by their link to the corresponding data that they contribute to describe but also by a set of metadata, described using the current international standards: Title, Authors, Publisher, Language, Date of publication, Body/Institution, Abstract, etc.; serial publications are described in terms of ISSN, while books are assigned ISBN; content of various types on electronic networks is described by means of DOI and URL. Each description is linked to the document.

Thanks to this, the MAPS library already enables researchers to go from structured oceanographic data to documents describing it. But this was not enough: documents may contain important information that has not been encoded in the metadata. Thus an advanced Search Engine was put in place that uses semantic-conceptual technologies in order to extract key concepts from unstructured text such as technical documents (grey literature material) and scientific papers and to make them indexable and searchable by the end-user in the same way as the structured data (such as oceanographic observations and metadata). The general architecture of this system will be described in the next paragraph.

2.1 The general architecture

More specifically, once a document is uploaded in the MAPS library key domain concepts in documents are extracted via a NLP pipeline and used as additional information for its indexing. The key term identification algorithm is based on marine concepts that were pre-defined in a domain ontology, but crucially it also allows for the discovery of new related concepts. So, for instance, starting from the domain term salinity related terms such as sea salinity and average sea salinity will also be identified as key terms and used for indexing and searching documents. A hybrid search system is then put in place, where users can search the library by metadata or by free text queries. In the latter case, the NLP pipeline performs an analysis of the text of the query, and when key concepts are matched, the relevant documents are presented. The results may be later refined by using other structured information (e.g. date of publication, geographic area).

A running system has been currently put in place with data from satellites, buoys and sea stations; such data is documented and searchable by its relevant metadata and documentation. A quantitative evaluation in terms of information retrieval measures has then been performed: more specifically, given an evaluation set defined by domain experts and composed of pre-defined queries together with documents that answer such queries, it was shown how the system is highly accurate in retrieving the correct documents from the library.

Though this work focuses on oceanography, its results may be easily extended to other domains; more generally, the possibility of enhancing the visibility and accessibility of grey literature via its connection to the data it describes and to an advanced full text indexing are of great relevance for the topic of this conference.

3. Architecture of the Semantic Search Engine

The Search Engine developed for the project MAPS has allowed cataloguing and selection of interesting information for a more or less expert community users operating in the marine domain. Indexing operations and successive recovery were carried out not only on the basis of the typology of text, their geographical location or other similar information contained in the associated metadata, but also on the basis of textual content, given the significant source of technical and scientific documentation in the base document.

The semantic search engine developed for the project MAPS has required the development of a battery of linguistic annotation modules operating both on Italian and English. In particular, the Search Engine uses semantic-conceptual technologies in order to extract key concepts from unstructured text such as technical documents and scientific papers. The documents are indexed and made searchable by the

semantic engine and then queried from the web interface by users together with other types of objects (data).

3.1 Indexing Pipeline

In Figure 1, we give a detailed overview of the Semantic search architecture. It contains a NLP pipeline for performing textual analysis; the result of the NLP pipeline is an annotated text that allows for terminological extraction of relevant concepts: such terms are then used for the indexing process. The MAPS System includes the following components both for the Italian and the English pipeline (as you can see from Figure 1): these two parallel pipelines should logically perform exactly the same tasks and this is true if we consider the pipeline as a unit. However, the pipelines are not composed of the same modules: the slight difference is hence due only to optimization problems of modules that previously operate individually and are then re-used, re-adjusted and/or optimized to operate within an operating unit of NLP:

- sentence splitting: division of text into sentences;
- word tokenization: splitting of sentences into words;
- lemmatization and morphological analysis (part of speech tagging);
- toponym detection: identification of geographic names;
- basic syntactic analysis (chunking): division of sentences into non recursive constituents;
- deep syntactic analysis (Ideal functional analysis) with adapted rules functional to terminology extraction;
- term filtering.

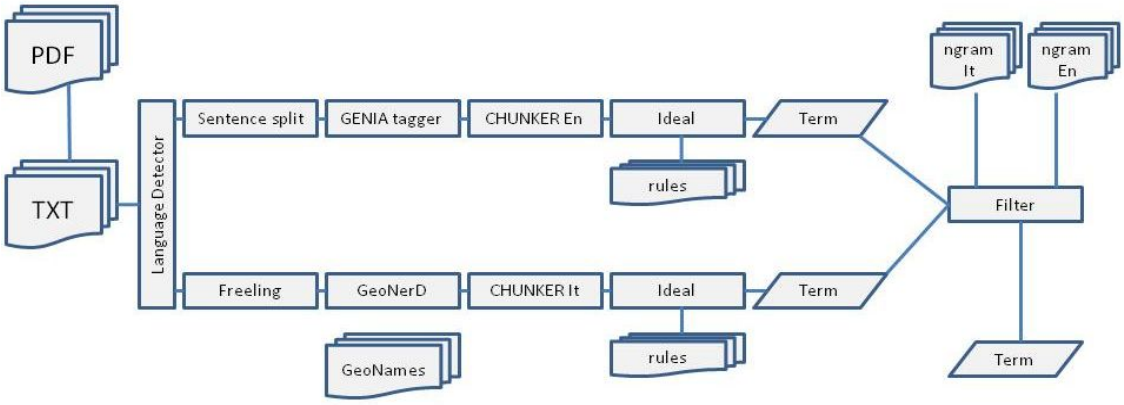


Figure 1: Steps of Natural Language Processing and term extraction.

The first step is the transformation of the original document (often in pdf format) in plain utf-8 format text. This is done with an open source command-line utility (pdftotext) converting PDF files to plain text. Then, since documents may be in English or Italian, a language detector is used in order to call for the correct language-specific NLP pipeline. In the upper pipeline we can see the component used for document in English and in the lower pipeline the component used for document in Italian:

- SentenSentenceDetector of the Apache OpenNLP suite was used for sentence splitting (https://opennlp.apache.org/);
- Genia Tagger was adapted (for tokenization, lemmatization and morphological analysis (http://www.nactem.ac.uk/tsujii/GENIA/tagger/);
- GeoNerD: an ad-hoc developed method using a knowledge-based (Geonames) and a rule-based algorithm for toponym detection and identification of geographic names;
- Chunker Eng|Ita: an ad-hoc developed module using a rule based algorithm for syntactic analysis (Lenci et al., 2003);
- Freeling: an open source language analysis tool suite available for several languages among which Italian (http://nlp.lsi.upc.edu/freeling/);
- Ideal Eng|Ita: an ad-hoc developed module using a rule-based algorithm for pattern term extraction.
- "Filter": an ad-hoc developed module used for filtering domain term.

3.2 Query Pipeline

Whereas NLP and indexing processes are asynchronous and are activated whenever a new set of texts enters the document base, the query analysis pipeline is instead a synchronous process, that is called whenever the Search Engine is interrogated by a user with a query in natural language. For this reason the query analysis system is built in a simplified way that guarantees to obtain results similar to those of

more complex NLP and indexing pipelines (as to ensure the extraction of relevant texts), but with better performances in terms of the response time - at least for such smaller snippets of texts that are likely to be entered by humans in a query.
There have been many interventions in order not to lower the performance and at the same time gain in speed (on average queyPipe runs 10 times faster than IndexingPipe). Interventions have been manifold: for instance, parsing algorithm was wired to program without using "Chunker" and "Ideal" that are actually compilers grammars which can run on generic rules: we were able to do this because the rules of terminology extraction are fixed and not very numerous. We also reduced the number of lexical resources to use, in fact the query language is certainly less rich than natural language. Another intervention involved the language detector: language detector in IndexingPipe is done by means of the Compact Language Detector (CLD2)[1] an application capable of recognizing a large number of languages. Since it was necessary to recognize only two languages, Simplified ad-hoc versions for the two query pipelines have been implemented using the most frequent terms in both languages. This tool was developed specifically for MAPS, being the most crucial component of the NLP pipeline.

3.3 Terminology extraction

The NLP pipeline produces an intermediate annotated document, which is preliminary to terminology extraction; the latter in turn is necessary in order to be able to correctly index the document in the document base for then being semantically searched. In particular each complex term is analyzed into a lemmatized head term ("salinity") and all the possible specifiers, with the frequency with which they occur in the text.

The part of the NLP pipeline relating to terminology extraction is made of two modules: "Ideal" and "Filter". "Ideal" is a finite-state grammar compiler capable of interpreting a set of instructions written in a language ad-hoc (Bartolini et al, 2004). In this language you can write rules like:

Pattern with { Constraints } ➔ Actions

where *Pattern* is a regular expression on macro components called "chunks" (the chunker performs a shallow parsing dividing the morphologically analyzed text into non-recursive syntagms; the chunk is identified by potgov that is his lexical head: for example "the red buoy" forms a single chunk with "buoy" as potgov). The *Actions* relate to the possibility of establishing a relationship between chunks. The rule is executed if the text recognizes a pattern that meets the linguistic *Constraints*. The rules are designed to extract all simple and complex noun phrases in the text and the extractor works on the chunked text searching for patterns such as:

- nominal phrase;
- nominal phrase followed by prepositional phrases;
- nominal phrase followed by prepositional phrases followed by prepositional phrases.

We are therefore interested in extraction of all monograms, bigrams and trigrams in the text (meeting certain linguistic constraints) without worrying about the reliability of the extraction. In this phase even trigrams as MEMBER DISTRICT TUSCANY can be extracted and they could match the text in the form "the members in the districts of the rich Tuscany"; these terms will form a set of candidate terms.

The "Filter" module examines the candidate terms and select those it deems to be in the domain. To this purpose a list of concepts is stored in the terminological base of the semantic engine, drawn from SeaDataNet vocabularies[2]: these lists (Ngram-it, Ngram-eng, as you can see in Figure 1) constitute the basic terminological domain (domain ontology) which has been made available by the marine experts of the MAPS project.

The filtering algorithm works this way: a term candidate is indexed if and only if its subpart belongs to the ontology domain; this allows to extract complex terms that were not previously present in the ontology of domain thus enriching it: for example, having "datum" and "height wave" as domain terms, the extracted terms - not being part of the domain - "the satellite data", "marine oceanographic data ", "average wave height", " changes in wave height" will be indexed.

The NLP and terminology extraction pipeline produces a set of domain terms in a standardised JavaScript Object Notation JSON format as output (see Figure 2), which is then used as input for the indexing function. Similarly, the query analysis pipeline produces the same output that can be used to search the existing index for matching terms.

```
                "form" : "220 210 Surface Solar Radiation Downward W/m2",
                "freq" : 1
              }
            ]
},
{
  "term" : "SEA SURFACE TEMPERATURE",
  "type" : "trigramma",
  "freq" : 39,
  "forms" : [
              {
                "form" : "the sea surface temperature",
                "freq" : 10
              },
              {
                "form" : "derived sea surface temperature",
                "freq" : 5
              },
              {
                "form" : "derived Sea Surface Temperatures",
                "freq" : 5
              },
              {
                "form" : "the sea surface temperature science",
                "freq" : 4
              },
              {
                "form" : "simulated sea surface temperatures",
                "freq" : 3
              },
              {
                "form" : "sea surface temperature",
                "freq" : 3
              },
              {
                "form" : "nder-avhrr sea surface temperature",
                "freq" : 3
              },
              {
                "form" : "the Sea Surface Temperature signals",
                "freq" : 2
              },
              {
                "form" : "high resolution sea surface temperature pilot project workshop",
                "freq" : 2
              },
              {
                "form" : "The godae high resolution sea surface temperature pilot project development",
                "freq" : 2
              }
            ]
},
{
  "term" : "AIR-SEA PHYSICS PARAMETRIZATION",
  "type" : "trigramma",
  "freq" : 38,
  "forms" : [
              {
                "form" : "an alternative air-sea physics parametrization",
                "freq" : 7
```


Figure 2: Steps of Natural Language Processing and term extraction.

4. Evaluation results

In the last phase of the MAPS project an evaluation activity was carried out in order to assess the capabilities of the developed search engine. This activity was divided in two parts: the first part was responsible of assessing whether stakeholders' requirements were correctly implemented in the system; the second part had the objective to evaluate whether the Search Engine is capable of retrieving relevant documents. The assessment of requirements is carried out by planning and executing a number of tests in order to verify the behavior of the system in given operational conditions finalized more on evaluation of the quality attributes of the implementation than on verifying that every function was properly executed. There are several quality models useful for this purpose, such as the quality models developed by (McCall, 1977) and (Bohem, 1978). More recently, the International Organization for Standardization (ISO) has developed a number of standards defining software quality models, the most known being the ISO/IEC 9126 standard. This model has been used for assessing the search engine: a number of characteristics and associated sub-characteristics of interest for the project was identified. Subsequently requirements were grouped in four classes and then evaluated by stating whether each requirement implementation had or not each sub-characteristic. Finally, a rating, in the range 0 to 10 has been computed for each requirement group and sub-characteristic by evaluating the ratio between the number of requirements in the group satisfying the sub-characteristic and the total number of requirements. Although several areas have to be improved, in particular the presentation and usability extents, the total rating reached was good.

In the second part of the evaluation, the Search Engine was experimented to assess how much it was capable of retrieving relevant documents against the submitted queries. Information Retrieval (IR) has

elaborated several techniques to evaluate the performance of a retrieval system such as Precision and Recall, Accuracy, f-measure, ranked results and so on. It was decided to adopt the Precision/Recall method for its simplicity and because the MAPS Search Engine has been specifically customized for the marine domain and this does not allow re-using existing benchmarks such as, for instance, the ones adopted in the various TREC evaluation campaigns.
Precision and recall are defined as follows:

$$p = \frac{n_{Rr}}{n_r} \qquad r = \frac{n_{Rr}}{n_R}$$

where $\mathbf{n_r}$ is the number of documents retrieved by the search, $\mathrm{n_R}$ is the number of relevant documents in the document collection and $\mathbf{n_{Rr}}$ is the number of relevant documents retrieved by the search.
For computing precision and recall, a collection of 15 documents and an assortment of 9 queries (Q1, …, Q9 in Table 2) were defined; for each document-query pair, a number of experts determined the corresponding relevance n_R. Then the 15 documents have been loaded in the MAPS system and the 9 queries have been executed, annotating the number of retrieved documents $\mathrm{n_r}$ and the number of relevant documents $\mathrm{n_{Rr}}$ by each query. Finally, overall precision and recall values have been computed by averaging the values resulting from each query. The results obtained are reported in Table 1.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
# of Relevant docs (n_R)	2	2	4	2	2	2	2	4	1
# of Retrieved & Relevant docs (n_{Rr})	2	2	2	0	0	2	2	3	1
# of Retrieved docs (n_r)	2	4	2	0	0	2	2	3	1
Precision ($p = \frac{n_{Rr}}{n_r}$)	1.00	1.00	0.50	0.00	0.00	1.00	1.00	0.75	1.00
Recall ($r = \frac{n_{Rr}}{n_R}$)	1.00	1.00	0.50	0.00	0.00	1.00	1.00	0.75	1.00
Precision (Average)	0.75								
Recall (Average)	0.81								

Table 1. Precision and Recall values

By averaging the precision and the recall computed for each query, we obtained that about the 80% of relevant documents were retrieved while the 75% of the retrieved documents were relevant: a promising result.

5. Conclusion

The MAPS project, supported by regional funding POR-FESR for industrial development of enterprises associated to the Liguria Cluster of Marine Technologies, is part of a long term activity aiming at building a computer platform for supporting a Marine Information and Knowledge System.
The paper deals with the application of semantic technologies to grey literature documents in the oceanography domain. The semantic engine developed for the MAPS project has allowed annotation and extraction of information and cataloguing of scientific documentation in the marine domain. Terminology indeed is inextricably linked with specialist knowledge and is the means by which specialists transfer it. This knowledge is distributed on scientific documentation and (unpublished) document repositories represent the intellectual output of academic institutions in the various areas. The semantic engine uses semantic-conceptual technologies in order to extract key concepts which are used, on turn, to index and make searchable and retrievable the document base itself by the community of expert users. Indexing operations, hence, were carried out not only on the basis of the metadata associated to the texts, but also on the basis of the content of the documents. Terminologies/ontologies and knowledge resources in a rule-based information extraction technique enable rich semantic indexing of grey literature documents.
The case-study represented by MAPS appears to stand at the crossroads between three quite different domains, 1) *Semantic engines design*, dealing with technologies for the extraction of knowledge from texts, 2) *Grey Literature*, the study of scientific and technical documentation produced by the various domains of knowledge and 3) *Operative Oceanography*, the field of oceanography devoted to gathering and analyzing oceanographic data.

Acknowledgment

This paper was supported by the "Programma Operativo Regionale POR-FESR (2007-2013), Asse 1 Innovazione e Competitività, Bando DLTM Azione 1.2.2 *Ricerca industriale e sviluppo sperimentale a favore delle imprese del Distretto Ligure per le Tecnologie Marine (DLTM) anno 2012*. Concessione di agevolazione POS. N° 19. CUP G45C13000940007 MAPS."

References

BELOV, S., MIKHAILOV, N. (2012). *Technical Workshop on the IODE Ocean Data Portal,*. Retrieved Oostende, Belgium, 27-29 February2012 Status of development of the ODP V2
<http://www.jcomm.info/index.php?option=com_oe&task=viewDocumentRecord&docID=8545>

BOEHM, B., BROWN, J.R., KASPAR, H., LIPOW, M., G. MCLEOD, G.M., M. MERRITT, M. (1978). *Characteristics of Software Quality*, North Holland, New York.

BOYER, T.P., ANTONOV, J.I., BARANOVA, O.K., COLEMAN, C., GARCIA, H.E., GRODSKY, A., JOHNSON, D.R., LOCARNINI, R.A., MISHONOV, A.V., O'BRIEN, T.D., PAVER, C.R., REAGAN, J.R., SEIDOV, D., SMOLYAR, I.V., ZWENG, M.M. (2013). *World Ocean Database 2013*. Sydney Levitus, Ed.; Alexey Mishonov, Technical Ed. NOAA Atlas NESDIS 72, 209 pages.

BOSMA, W. E., VOSSEN, P., SOROA, A., RIGAU, G., TESCONI, M., MARCHETTI, A., MONACHINI, M. , ALIPRANDI C. (2009). *Kaf: a generic semantic annotation format*. In Proceedings of the Generative Lexicon *GL2009*, Workshop on Semantic Annotation, Pisa, Italy.

BRIN S. (1998). Extracting Patterns and Relations from the World Wide Web. In *Proceedings of the WebDB Workshop at 6th International Conference on Extending Database Technology*, EDBT'98. Pages 172-183.

BUSTAFFA F., DE MATTEI M. (2013). MAPS. Raccolta delle esigenze. Programma Operativo Regionale POR-FESR (2007-2013), Asse 1 Innovazione e Competitività, Bando DLTM Azione 1.2.2 "Ricerca industriale e sviluppo sperimentale a favore delle imprese del Distretto Ligure per le Tecnologie Marine (DLTM) anno 2012. Concessione Di Agevolazione POS. N° 19. CUP G45C13000940007 MAPS. Deliverable D1.2,Versione 1.0.

CREATIVE COMMONS (2001) .*Some Rights Reserved*: *Building a Layer of Reasonable Copyright*.
<http://creativecommons.org>

CLARKE, R. (2005). *A proposal for an open content licence for research paper* (Pr)ePrints. *First Monday 10* (8), 1-11. <http://www.firstmonday.org/issues/issue10_8/clarke/>

EUROPEAN COMMISSION (2010). EUROPEAN MARINE OBSERVATION AND DATA NETWORK: IMPACT ASSESSMENT. SEC(2010) 999 FINAL.

EUROPEAN COMMISSION (2012) *Green Paper Marine Knowledge 2020*. COM(2012)473 Final
<HTTP://EC.EUROPA.EU/MARITIMEAFFAIRS/DOCUMENTATION/PUBLICATIONS/DOCUMENTS/MARINE-KNOWLEDGE-2020-GREEN-PAPER_EN.PDF>

FELLBAUM, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

FRONTINI F., BARTOLINI R., MONACHINI M., PARDELLI G. (2014). MAPS. Stato dell'Arte dei motori semantici. PROGRAMMA OPERATIVO REGIONALE POR-FESR (2007-2013), Asse 1 Innovazione e Competitività, Bando DLTM Azione 1.2.2 "Ricerca industriale e sviluppo sperimentale a favore delle imprese del Distretto Ligure per le Tecnologie Marine (DLTM) anno 2012. CONCESSIONE DI AGEVOLAZIONE POS. N° 19. CUP G45C13000940007 MAPS. Deliverable D1.1,Versione 1.0.

GOGGI S., MONACHINI M., FRONTINI F., BARTOLINI R., PARDELLI G., DE MATTEI M., BUSTAFFA F., MANZELLA G. (2015). Marine Planning and Service Platform (MAPS): An Advanced Research Engine for Grey Literature in Marine Science. In Dominic Farace and Jerry Frantzen (eds.), *Proceedings of the Sixteenth International Conference on Grey Literature Grey Literature* GL16 - Lobby: Engines and Requesters for Change - (Library of Congress Washington D.C., USA, December 8-9, 2014). Pages 108 - 115. TextRelease, Amsterdam. {GL-conference series, ISSN 1386-2316), vol. 16).

BARTOLINI R., LENCI A., MONTEMAGNI S., PIRRELLI V., CLAUDIA SORIA C. (2004). *Semantic mark-up of Italian legal texts through NLP-based techniques.* In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, Raquel Silvia (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation,* LREC 2004. ELRA, Paris. Volume III, Pages 795-798. ISBN 2-9517408-1-6.

KOGUT, P. & HOLMES, W. (2001). AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages. *First International Conference on Knowledge Capture (K-CAP 2001).* Workshop on Knowledge Markup and Semantic Annotation. <HTTP://CITESEERX.IST.PSU.EDU/VIEWDOC/DOWNLOAD?DOI=10.1.1.21.8180&REP=REP1&TYPE=PDF>

LENCI, A., MONTEMAGNI, S. & PIRRELLI, V. (2003) *CHUNK-IT.* An Italian Shallow Parser for Robust Syntactic Annotation, In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), *Computational Linguistics in Pisa - Linguistica Computazionale a Pisa*. *Linguistica Computazionale*, Special Issue, XVI-XVII, (2003). Pisa-Roma, IEPI. Tomo I, 353-386.

MAILLARD, C., LOWRY R.K., MAUDIRE, G., SCHAAP, D. (2007). SeaDataNet: Development of a Pan-European Infrastructure for Ocean and Marine Data Management, *Oceans2007- Europe* Conference, Aberdeen, 18-21 June. IEEE, Pages 1-6. E-ISBN:978-1-4244-0635-7 <DOI:10.1109/OCEANSE.2007.4302435>

MANNING CHRISTOPHER D., RAGHAVAN P., SCHÜTZE H.(2009). *An Introduction to Information Retrieval. Online edition*, C.U.P. <HTTP://NLP.STANFORD.EDU/IR-BOOK/PDF/00FRONT.PDF>

MCCALL, J.A., RICHARDS, P.K. & WALTERS, G.F. (1977). *Factors in software quality.* National Technical Information Services (NTIS), RADC-TR-77-369.

PROCTOR, R., ROBERTS, K. , WARD, B.J. (2010). A data delivery system for IMOS, the Australian Integrated Marine Observing System, *Advances in Geosciences*, An open access journal for refereed proceedings and special publications, vol. 28, pages 11-16.<DOI:10.5194/ADGEO-28-11-2010>

RAJMAN M., BESANÇON R. (1998). Text mining: natural language techniques and text mining applications. In *Data Mining and Reverse Engineering*. Springer US. Pages 50-64.

VOSSEN, P., AGIRRE, E., CALZOLARI, N., FELLBAUM, C., HSIEH, S., HUANG, C., ISAHARA, H., KANZAKI, K., MARCHETTI, A., MONACHINI, M., NERI, F., RAFFAELLI, R., RIGAU, G., TESCON, M. & VAN GENT, J. (2008). *KYOTO:* A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures, in Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, Piek Vossen (eds.), *Proceedings of the Fourth International Global Word Net Conference - GWC 2008*. University of Szeged, Department of Informatics, Pages 474 – 484. ISBN 978-963-482-854-9.

WANG, Y. (2008) *Software Engineering Foundations, a software science perspective*, Auerbach Publications.

[1] <https://github.com/CLD2Owners/cld2/> last accessed on January 15, 2016.
[2] <http://www.seadatanet.org> last accessed on January 15, 2016.

https://doi.org/10.26069/greynet-2025-000.540-tgj

The National Portal for Recording Theses (PNST): Its Role, Importance and Constraints for Algerian Researchers*

Azzedine Bouderbane, Nadjia Gamouh, and Hadda Saouchi (Algeria)

Abstract
If researchers no longer face difficulties to have access to published resources which are exploding in all disciplines, they, however, make tremendous unsuccessful efforts when looking for unpublished documents (grey literature) to develop their research. For the case of theses, several researchers do not know about topics treated by their peers. Moreover, they often find constraints to get access to a specific title's or author's thesis. They also make a long distance, waste time and money in the attempt to reach a desired thesis. For years, this has constituted a real problem for Algerian researchers. The Research Center for Technical and Scientific Information (CERIST) in Algeria has innovated by establishing the National Portal for Recording Theses (P.N.S.T.) as a device that aims above all at caring about national scientific production in terms of theses. It is a central catalogue for registered research topics and maintained theses that provides a certain visibility of research in the country. This portal was also set up with the intention to put an end to the researchers 'difficulties when seeking for theses.
In our paper, we have attempted to evaluate whether the researchers' situation has known any improvement. To reach this goal, a questionnaire in Arabic was administered by e.mail to a sample of researchers at the University of Constantine 2. 50 university teachers and 50 Ph.D students were concerned by this qualitative study. An informal interview helped us to communicate with researchers. The analysis of the survey led to significant results that showed the positive points of this portal and the various constraints that still face researchers when attempting to explore this information gate.

Keywords: *Grey literature, thesis, researcher, National Portal for Recording Theses, constraint, qualitative study, Algeria.*

Introduction

The Algerian researchers in the different scientific disciplines face difficulties in accessing to scientific information, above all the information known as grey literature. If researchers make tremendous efforts to get published information, they are pushed to make more efforts for acquiring unpublished information. As a consequence, the CERIST (Research Center for Scientific and Technical Information) launched a project for setting up the PNST (National Portal for Recording Theses) for the purpose of reducing the researchers' difficulties in terms of scientific information needs, and offering opportunities of access to information which is in subjects undertaken as academic research works or projects. Through this paper, we intend to highlight the role and importance of the CERIST portal in promoting scientific research on the national level, and to underline the researchers' needs for such initiative that not only enables these researchers to develop scientific communication, but also to exploit the rich unpublished scientific literature production available in theses.

1. Methodology

a. Statement of the problem

- The development of publishing process and the technical means in this matter led to an explosion of literature production and to an extraordinary progress of scientific knowledge in the various disciplines. Researchers have found themselves in front of a challenge: getting access to specialized scientific literature and being able to select from this enormous published and unpublished literature tide the most useful, appropriate and accurate information needed by the researchers who have not the right to miss the optimal updated information. Researchers are bewildered by the variety of research tools available, and by the diverse sources of information. Academic institutions produce a great amount of published and unpublished literature that pushes researchers to make tremendous efforts to accede to this literature production. On the level of space, Algeria is a real continent with a surface of 2 381 741 Km2.The number of academic institutions spread all over this large space is very high: (48 Universities, 10 University Centers and 20 National Higher Schools). This constitutes a handicap for researchers who feel exhausted when looking for academic scientific information that is disseminated all over the country. The difficulty is greater when speaking about the exploitation of unique copies of theses stored in university libraries. This harsh problem led the CERIST to look for a solution that might contribute to the reduction of the researchers' difficulties in the country. The CERIST initiative consisted in launching a project for setting up the PNST (National Portal for Recording Theses) with a double purpose: informing researchers about the various subjects treated

* First published in the GL17 Conference Proceedings, February 2016.

or on the way to be treated by Magister and Ph.D. theses in the different disciplines in Algeria; and exploiting this collection of theses by researchers for the sake of contributing to the development of scientific research in the country. After about five years of existence, a key question should be asked: Do researchers feel satisfied by this portal? This principal question has led us to state several interrogations:

- Does this portal help researchers in acceding to theses?
- Do researchers face difficulties when exploiting this portal? If yes, what are they?
- Can the CERIST improve this portal? How and why?

b. Hypotheses

To lead this study, we have stated some hypotheses:

- The PNST contributes efficiently to the promotion and development of scientific research on the national level by exhibiting useful information about Magister and PhD theses with a completed or on the way to be completed research.
- The variety of disciplines covered by this portal provides to the Algerian researchers a good opportunity to accede to a fruitful collection of theses.

c. The objectives of the study

Our study attempts to achieve the following objectives:

- To identify the difficulties faced by the University Constantine 2 researchers when attempting to collect scientific information and to accede to academic research results.
- To highlight the local initiatives that try to promote scientific research in the country.
- To evaluate the efficiency of the PNST through its use and exploitation by the Algerian researchers.

2. The Research Center for Scientific and Technical Information (CERIST)

This research center was founded in 1985. It belonged at first to the government presidency before joining the Ministry of Higher Education and Scientific Research in 2003. This center is considered as a national institution with a scientific and a technological mission. Its main tasks consist in valorizing scientific works, in establishing the information national system, in exploiting scientific and technical information for the benefits of scientific research and sustainable development, in building and reinforcing information society via the establishment of information networks in a variety of areas. The center promotes the use of information and communication technology in higher education, in addition to the development of the national university documentary system and the foundation of virtual libraries. It has also to develop on the one hand the national data-bases for the purpose of disseminating them, then facilitating the access to them, and on the other hand, to encourage research that secures information and networks.

Image 1: The Research Center for Scientific and Technical Information (CERIST)

3. The National Portal for Recording Theses (PNST)

This portal which is the result of the CERIST initiative includes three essential elements: the researcher who needs to accede to Magister and PhD theses, the university library that conserves the original copy and feeds the portal with the required information, and the information specialist who processes the collection adequately checking the duplication of titles or subjects, and accepts to collaborate with the CERIST. The portal gives the opportunity to the researchers to register the subject that he/she is going to treat and that is approved by the scientific council of his/her institution. This gives him/her a copyright protection of the research. The portal can be reached via the following link: http://www.pnst.cerist.dz

Image 2: The front page of the portal (PNST)

Statistics about the number of subjects' theses stored in the data-base are exhibited on the portal. The number of theses in full text is also given. The following link portrays the above statistics: http://www.pnst.cerist.dz/stat.php

4. Recording Magister and PhD theses

The services of post-graduate studies belonging to all academic institutions in Algeria inform the CERIST about the availability of a thesis. Then, these services get a personal account for each thesis. The services register with great precision all the bibliographic data related to the thesis such as the title, the abstract, the keywords, the language of the thesis, the researcher's name and his/her electronic mail, the supervisor's name, in addition to the discipline and the academic institution to which the researcher belongs. The services of post-graduate studies and researchers should attest that they keep confidentially the account provided by the CERIST. The latter should be informed by the various academic scientific councils about any modification introduced in the researchers' theses.

5. Filing theses in the PNST

To record electronic copies of these submitted in front of a jury, the PNST opens the portal related to the theses by providing accounts to university library managers who will enter the portal's site and insert the complete electronic copy with all the required bibliographic data such as the name of the academic institution, the discipline, the researcher's name, the chronological order number, the supervisor's name and academic status, the co-supervisor's name when available, the rank of the thesis (Magister – Doctorate), the language of the thesis, the keywords, the abstract, the date of submission of the thesis and other supplementary bibliographic data such as: clarifications, observations, thesis' dimensions … . After filling in all the data, the university library manager submits the whole file with the thesis. The university library informs the PNST whether it permits the access just to the bibliographic file of the thesis or to the full text. The university library may impose the confidentiality of the thesis in full text for five years. After the end of this period of time, the CERIST gets the permission to free the users to accede to the theses.

6. The research procedure through the PNST

The research procedure in the PNST permits the retrieval of information needed in various scientific disciplines taught at the Algerian universities. In fact, the portal allows researchers to use a search engine to carry out an advanced search through the introduction of data about the academic institution or the author's name or the title of the thesis or the keywords to get research results at the bottom of

the page. These results consist in a list of theses that deal with the research subject inserted in the data-base of the portal. The research may get the title of the thesis, the author's name, the supervisor, in addition to the situation of this thesis: submitted, allowed for full text access, allowed just for acceding to the bibliographic file. The researcher can also get the information about the date of insertion of the thesis in the PNST. If the researcher wants to down-load one of the theses from the portal, he/she should use his/her personal account via the site of the National System for Online Documentation (SNLD). After registration in the SNLD, the researcher selects the portal "PNST" and clicks on it. He/she gets the data already stated above. At the bottom of the page, he/she will notice "full text" on the bibliographic file of the thesis.

7. The Algerian researchers 'use of the PNST: the results of a survey

To identify the various benefits of Algerian researchers from this portal and to measure its impact on the Algerian scientific scope, we relied on a qualitative study. 100 researchers from the university Constantine 2 were selected. We adopted the descriptive approach and used the questionnaire as an instrument for collecting information. We also relied on an informal interview with ten (10) researchers from our institute to confirm certain results collected via our questionnaire.

8. The analysis of the collected data

When analyzing the answers to the first question, one can notice that 65% of the respondents were females. Furthermore, 80% of PhD researchers were females. Through these results, one can predict a majority of female university teachers in future. The great majority of researchers showed their interest for the underground literature when involved in research projects. 95% of the respondents selected 'theses' as "the most essential and suitable documents" they would like to use in research. They explained that their access to published documents is surrounded by several difficulties. However, when talking about 'theses', they affirmed that the situation would become alarming because these collections were protected rigidly by academic institutions. They stated that they could not speak about 'reports and official publications' which are practically inaccessible. For 'theses', they explained that university libraries did not accept to lend these documents though library managers knew that researchers might come from very far parts of the country. These libraries did not allow researchers to make printed or electronic copies of 'theses'. About the use of this portal, 25% of the respondents mentioned that they exploited it once a week, whereas 65% of them used it at least once a month. 5% mentioned that they used it once a year while some of them did not answer. How were the respondents informed about the PNST? That was a question asked and to which 90% of the respondents stated that their academic institution informed them about this portal. 10% did not remember exactly how they got informed about that innovation. That may show the relative good collaboration that should exist between the PNST and the university libraries. Concerning the statistics provided by the portal, 50% of the respondents mentioned that they did not give importance to the statistics available on the portal. 30% exploited these statistics. 20% knew about the statistics, but did not exploit them. The question related to the linguistic aspect of the portal pushed all the respondents (100%) to mention that surfing via the portal could be made through three languages: Arabic, French and English. They felt satisfied by this because that responded to the linguistic need of every researcher. 55% of them criticized the inclusion of a 'guide' that could be exploited just in French. One may mention in this context that several researchers no longer handle foreign languages as before in the country. They just use Arabic as a unique instrument for communication or for information search instead of making efforts to master other foreign languages that are fundamental in the research field. 45% of the respondents underlined the usefulness of the guide without treating its linguistic aspect. Concerning the various links provided on the portal, 75% of the respondents appreciated their introduction because they might orientate them rapidly to other interesting sites. 40% of the sample showed that some links were not activated. For their satisfaction in terms of response to their information needs, 60% of the respondents stated that the PNST was suitable and fairly appropriate, whereas 20% showed some pessimism. 10% of the respondents declared that it responded to some of their information needs. Some of them did not answer because of their lack of conviction concerning this matter. For the design of the portal, 85% of the respondents appreciated it, whereas 10% mentioned that it could be improved, and 5% did not reply: they felt unable to make an evaluation. They might not know the criteria for designing an electronic site. 90% of the respondents did not understand why the portal did not give them the opportunity to accede directly to the theses in full text. Several researchers mentioned that they often forget their personal account provided by the library. Concerning the technical constraints that may face the researchers, 95% talked about the weak Internet connection which disturbed them in their research while surfing through the portal. For their global appreciation of the PNST, 90% of the respondents were

quite satisfied, whereas 10% showed their dissatisfaction asking for more seriousness and engagement. 95% affirmed that the portal contributed to the development of scientific research in Algeria. 80% of the researchers who participated in this survey recommended the necessity of enhancing collaboration between the different partners..

9. Comments and suggestions

Nobody can deny that the PNST has been tremendously helpful for a lot of researchers. Six years before, researchers suffered a lot when travelling along the vast country of Algeria looking for a copy of a thesis conserved intimately by a university library. A researcher in the Wilaya of Tamanrasset, in the south of the country, is obliged to cross more than 2000 Klms to reach a university library in Algiers. He/she may return to his/her city frustrated without the expected document to cross again 2000 Klms. We agree that the portal is the solution. Of course, this is not a perfect device on the basis of the results of the questionnaire. A lot of researchers attest that the portal contributes to the promotion of scientific research. We should recognize that there is no better alternative that may help researchers more efficiently in this matter. Through these results and this analysis, we may assume that the first hypothesis is confirmed.

The results also show that the portal constitutes a source of information for the Algerian researchers who need to use 'theses' to specify the previous studies led by other researchers. No one can begin a research without checking whether his/her topic has not been treated by other researchers. It is also approved that a research begins from where the others arrived. The imaginative idea of recording theses in all disciplines and, then, providing them to researchers will lead to a constitution of a data-base that can be very fruitful for research projects. This can generate a lot of benefits for researchers and that confirms the second hypothesis. The Algerian universities should contribute in developing a tight collaboration between their libraries and the CERIST which controls the PNST. The university Scientific Councils should also be sensitized about the importance of cooperating with this useful portal. Universities may participate in improving it at all levels for the sake of helping researchers and activating the research process in the country.

The Algerian universities should be conscious about the role and the importance of this outstanding source of information. They may enhance the CERIST and university library specialists to act as professionals in order to offer a fruitful service to researchers all over this large country.

Bibliography

Baltz, C., ,1998 . « Une culture pour la société de l'information ? Position théorique, définition, enjeux », *Documentaliste – Sciences de l'information*, 35,1

Comberousse, Martine.. 1993. La littérature grise. *Bulletin des bibliothèques de France* [en ligne], n° 5, 1993 [consulté le 21 novembre 2015]. Disponible sur le Web : <http://bbf.enssib.fr/consulter/bbf-1993-05-

Duspaire, J.L. 2004. « La documentation : une fonction essentielle du système éducatif ». *ARGOS*, 36,

Farace, D. J. & J. Schöpfel (eds.) (2010). Grey Literature in Library and Information Studies. De Gruyter Saur

Fischer, Gilles ; Harlan, Cleveland. 2008. Information stratégique à valeur ajoutée – Enssib: *www.enssib.fr/.../61310-l-information-strategique-a-valeur-ajoutee-l-inf*.

http://www.adbs.fr/litterature-grise-17647.htm?RH=OUTILS_VOC

http://www.cerist.dz/index.php/fr/

http://www.pnst.cerist.dz/stat.php

Joachim Schöpfel,. 2012. « *Vers une nouvelle définition de la littérature grise* », *Cahiers de la Documentation*, vol. 66, n° 3, p. 14-24

Mebarki, 2003. M. *Sauver l'université*. Oran : Dar-El-Gharb, Serres, A. *2008* « Educations aux médias, à l'information et aux TIC : ce qui nous unit est ce qui nous sépare », *Colloque international de l'ERTé, L'éducation à la culture informationnelle, Lille, 16-17-18 octobre*.

Sulouff, P., et al.,. 2005. Learning about gray literature by interviewing subject librarians: A study at the University of Rochester. *College & Research Libraries News*, 66(7), pp. 510–515.

https://doi.org/10.26069/greynet-2025-000.540-tgj

Korea Institute of Science and Technology Information (KISTI)

English version - http://en.kisti.re.kr/

*** Vision**

World-class information research institute creating values for customers

*** Main functions**

*** Management and service of Korean R&D reports**

KISTI exclusively manages, preserves, and serves Korean R&D reports for citizens and government officials. It provides Korean R&D reports and their information with National science & Technology Information Service (NTIS) and National Discovery for Science Leaders (NDSL).

***Contact information**

KISTI email address: hcpark@kisti.re.kr

Headquarters: Tel : +82-42-869-1004, 1234 Fax: +82-42-869-0969

https://doi.org/10.26069/greynet-2025-000.540-tgj

Analysis of National R&D Project Report Output Utilization and Economic Contribution *

Kiseok Choi, Cheol-Joo Chae, Yong-hee Yae, and Yong Ju Shin (Korea)

Abstract

As the major developed countries of the world have realized the importance of open common information and its recycling, they have begun to seek a plan that will facilitate manipulation of common information. Despite development of the National R&D business and its quantitative increase in Korea, its practical use still remains at a lower level. Therefore, it is worthwhile to analyze practical use and economic contribution in order to achieve higher profits.

There are two purposes of this study. The first purpose is to draw a conclusion by analyzing/investigating the practical uses of the National R&D Reports, and the second purpose is to evaluate economic contributions by the National R&D Reports. These are possible by manipulating the improved quantitative/qualitative indicator designated to evaluate the economic contributional aspect of the National R&D Reports.

The degree of outcome measurement for the National R&D Reports was investigated through science and technology workers in both the field and the academic world. The results of this investigation show that the average time of use for the total National R&D Reports is 6.3 years and average monthly reference occurs 9.2 times. Additionally, results showed that researchers spend an average of 11.6 hours per month using the reports. Moreover, the proportion of the National R&D Reports in Collection of Science and Technology Information, the importance of the National R&D Reports in Practical Science and Technology Information, and the proportion of the National R&D Reports in Quotation Science and Technology Information are considered between 'average' and 'slightly higher than average.'

The results of the investigation/ analysis for usefulness of the National R&D Reports show that the quality of content, the quality of the system, the satisfaction, and the utility are recorded as 'slightly higher than average' and 'high.' These results demonstrate that users of the National R&D Reports consider these reports to be an important resource.

The results of the investigation/ analysis for the preservation value of the National R&D Reports indicate 83.2% positive opinion about preserving this report. Therefore, KISTI should preserve National R&D Reports at the national level.

The results of the investigation/ analysis for the demand of National R&D Reports show that the intention to use the system records falls between 'slightly higher than average' and 'high.' These results address the high number of users who intend to keep using the National R&D Reports.

The results of the investigation/ analysis for the economic contribution of the National R&D Reports show that value for practical use, value of common ownership, and contributional value—which are all qualitative indicators, respectively—fall between 'slightly higher than average' and 'high' on the scale. These results show that the users of the National R&D Reports consider it of high economic value.

The usefulness value of the National R&D Reports is 2.908 won for a piece of information, and the total usefulness value of the National R&D Reports thus far is 1.713 hundred million won; the reports are estimated to hold 15.5 times the economic value over the cost. The payment value of the National R&D Reports is 2,156 won for a piece of information, and the total payment value of the National R&D Reports thus far is 1,270 hundred million won; the reports are estimated to hold 11.5 times the economic value over the cost. Compared to other domestic and foreign information services' estimation for the economic contribution—which has only recorded estimates of 3 to 9 times the value rate—the construction of these National R&D Reports is currently economically effective for businesses and is also expected to be economically feasible for businesses in the future.

1. Introduction

Recently, major advanced countries in the world have recognized the importance of the opening and recycling of public information and searched for a way to provide and utilize public information in a smooth and efficient manner. In the Republic of Korea, despite a significant increase in the number of government-led R&D programs and outcomes, the utilization of these research outcomes is still very low. Therefore, it is urgent to figure out how to utilize research results and analyze their economic contribution.[1]

In accordance with Paragraph 13 of Article 25 (Management of R&D Information) of 'Act on the Management of National R&D Programs,' the head of the research institute or research management

* First published in the GL17 Conference Proceedings, February 2016.

agency is required to register or submit the research results to the 'research outcome management & distribution agency' according to the terms and conditions of the agreement signed under Article 9 (Signing of the Agreement). Furthermore, the agency in which the research results are submitted or registered is needed to build and operate a research outcome management and distribution system in connection with the National Science & Technology Information System (NTIS) and fulfill its duties for the maintenance, storage and management of the research results.[6][7] The R&D reports have also been registered and managed in the National Science Research & Development Reports Registry Management System (NRMS) in the Korea Institute of Science and Technology Information (KISTI) in accordance with the law mentioned above. Major advanced countries in the world have recognized the importance of the opening and recycling of public information and searched for a way to provide and utilize public information in a smooth and efficient manner. For example, the U.S. has attempted to transfer technology through the Federal Laboratory Consortium organized by region for the distribution and utilization of research results. In addition, government-funded research project reports have been managed and provided in an independent or integrated manner by various agencies such as NIH, NIST, NTIS (NTRL), OSTI and NASA.

The United Kingdom has also reviewed a plan to increase people's access to research results by launching an independent working group. It is recommended to upload government-funded research results on the Internet immediately to allow all users to get access to them free of charge.

In Japan, the contents form the electronic archive of R&D reports are provided through KAKEN database which has been prepared and provided by National Institute of Informatics (NII) in cooperation with Japan Society for the Promotion of Science (JSPS) and AIST Repository.[5]

The research outcome spread & utilization policy differs by country, but it has been promoted under the basic principle of 'open access.' However, studies on the utilization of research results and analysis of their economic contribution are still very much needed.

2. Analysis on the utilization of national R&D reports

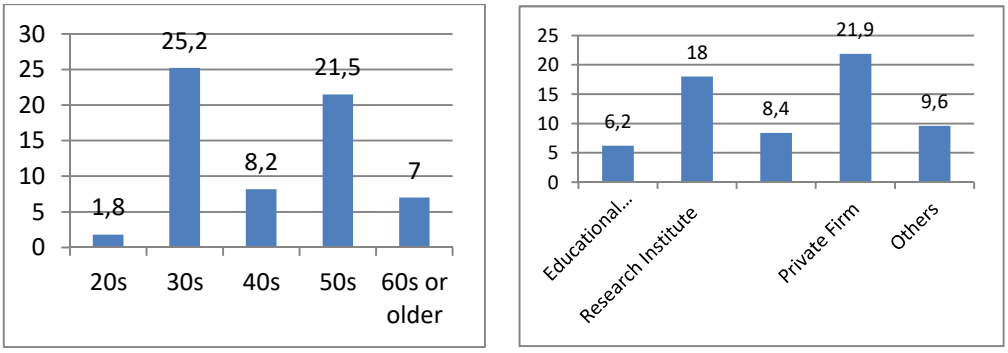
2.1 Analysis on the utilization of national R&D reports

This study investigated and analyzed how much the national R&D reports developed and provided by the KISTI are utilized against scholars and engineers in science and technology.

2.1.1 Average usage frequency of national R&D reports

The average monthly detailed-view frequency of national R&D reports which have been surveyed against all respondents was 16.6 times. According to the comparison of average monthly detailed-view frequency by gender, men (18.7) were far greater than women (5.0). According to the analysis of average monthly detailed-view frequency by age, '30s' was the highest with 25.2, followed by '50s,' '40s,' '60s and older and '20s.'

According to the analysis of average monthly detailed-view frequency by organization, 'private firm' was the highest with 21.9, followed by 'research institute,' 'other organization,' 'administrative body' and 'educational organization.'



<Figure 1> Average Monthly Detailed-view Frequency of National R&D Reports by Age and Organization

According to the comparison of average monthly detailed-view frequency by topic, 'machine engineering' was the highest with 101.4, followed by 'life science,' 'chemical engineering,' 'construction/transportation' and 'electricity/electronics.'

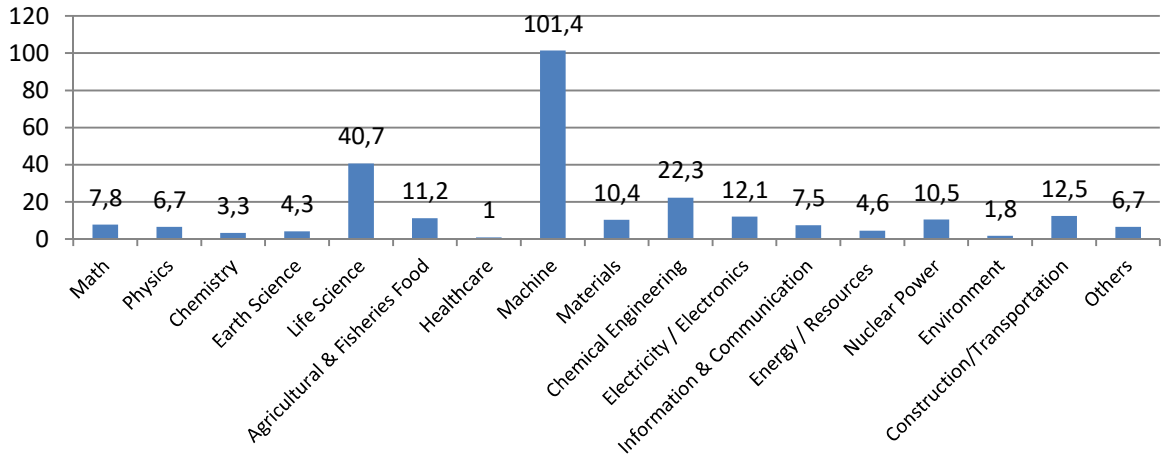


Figure 2 Average Monthly Detailed-view Frequency of National R&D Reports by Topic

2.1.2 Importance of KISTI-provided national R&D reports among the science and technology information in use

According to analysis on the importance of national R&D reports developed and provided by the KISTI among the science and technology information utilized for research, business and learning purposes against all respondents, 'high' was most responded with 26.5% (52 respondents). Based on the response 'neither low nor high,' positive responses ('very high,' 'high' and 'slightly high') accounted for 61.2% (120 respondents) while negative responses ('very low,' 'low' and 'slightly low') were 17.4% (34 respondents). The survey results are as high as 62.2 scores out of 100.

<Table 1> Results of Questionnaire on the Importance of National R&D Reports among Science and Technology Information in Use

Scale	Very Low	Low	Slightly Low	Neither Low Nor High	Slightly High	High	Very High
Frequency	12	6	16	42	48	52	20

2.2 Implications

According to a survey against scholars and engineers in science and technology, the average usage period of national R&D reports provided by the KISTI was 6.3 years with 16.6 times in average monthly detailed-view frequency, 9.2 times in average monthly original text-view frequency and 11.6 hours of average monthly usage time.

In terms of the ratio of national R&D reports developed and provided by the KISTI among science and technology information collected for research, business and learning purposes, 'neither low nor high' was the highest with 26.0% (51 respondents). Based on the response 'neither low nor high,' positive responses ('very high,' 'high' and 'slightly high') accounted for 49.5% (97 respondents) while negative responses ('very low,' 'low' and 'slightly low') were 24.5% (48 respondents). The survey results are as high as 56.8 scores out of 100.

In terms of the importance of national R&D reports developed and provided by the KISTI among science and technology information collected for research, business and learning purposes, 'high' was the highest with 26.5% (52 respondents). Based on the response 'neither low nor high,' positive responses ('very high,' 'high' and 'slightly high') accounted for 61.2% (120 respondents) while negative responses ('very low,' 'low' and 'slightly low') were 17.4% (34 respondents). The survey results are as high as 62.6 scores out of 100.

According to the said results, the ratio and importance of KISTI-provided national R&D reports among the collected and utilized science and technology information are situated between 'neither low or high' and 'slightly high.' The analysis results mean that respondents collect national R&D reports in 'neither low nor high' or higher levels and understand that they are pretty important.

3. Economic evaluation of national R&D reports

3.1 Economic analysis indicators of national R&D reports

For the economic analysis of national R&D reports, the following qualitative and quantitative indicators were developed based on previous studies and analyses. The qualitative indicator was classified into

utilization value of national R&D reports.[3][4] And the quantitative indicator were divided into utilization value of national R&D reports and B/C analysis.[12][13][14]

<Table 2> Economic Analysis Indicators of National R&D Reports

Category	Specific Indicators
Qualitative Indicator	▪ Utilization value of national R&D reports
Quantitative Indicator	▪ Utilization value of national R&D reports ▪ Benefit-Cost (B/C) analysis of national R&D reports

For analysis of qualitative indicators, a 7-point Likert scale-based questionnaire survey is used. For calculation formula, the 7-point Likert scale is converted into 100 scores, and the results are analyzed based on the following formula:

$\sum_{i=1}^{n} \overline{x_i} \times 14.28$ $\overline{x}$= Mean sore on a 7-point scale, n=Total number of respondents

For the analysis of utilization value as well, users' utilization value is estimated based on the following calculation formula:

Calculation Formula for the Utilization Value of National R&D Reports

$$UV = CS^{UV} \times N$$

UV: Total utilization value of national R&D reports
CS^{UV}: Utilization value of national R&D reports per capita
N: Total number of subscribers

According the B/C analysis, total benefits are divided by total costs spent to build and provide national R&D reports. If the result is '1 or higher,' it is deemed 'economically efficient.' Total benefits are analyzed using the utilization value of national R&D reports.

Calculation Formula for the B/C Analysis of National R&D Reports

B/C = Net Present Value (NPV) of total benefits / NPV of total costs

$$NPV = \sum_{i=1}^{n} Values_i \times (1+rate)^i$$

Valuesi: Benefits or costs on the ith year, rate: discount rate, n:No.ofyears

※ The said NPV formula is formulated after converting past value into present value.

3.2 Utilization value of national R&D reports

The utilization value of national R&D reports was 76.4 scores, standing between 'slightly high' and 'high.' According to analysis on the utilization value of national R&D reports with demographic characteristics, men (76.7 scores) were higher than women in gender. In age, '50s' was the highest with 77.5%, followed by '40s (77.3)' and '60s or older (77.1).' In terms of organization, 'private firm' was the highest with 83.0 scores, followed by 'others (79.6)' and 'research institute (75.0).'

3.3 Utilization value of national R&D reports

The utilization value of each original text view of national R&D reports was KRW 2,908. According to analysis on the utilization value of national R&D reports with demographic characteristics, men (KRW 2,933) were higher than women in gender. In age, '60s or older' was the highest with KRW 3,800. In organization, 'educational organization' was the highest with KRW 3,827. In terms of tope, 'environment (KRW 4,375)' was the highest, followed by 'chemical engineering' and 'physics.'

According to analysis on the utilization value of national R&D reports based on the formula below, total utilization value was KRW 171.3 billion. As the number of subscribers increased, utilization value also improved.

Calculation Formula for the Total Utilization Value of National R&D Reports

$$\text{Total NPV} = \sum_{i=2008}^{2014} \text{Valuesi X AC X MPY X Useri}$$

Valuesi = Utilization value of a report(2,908 KRW)
AC = Average monthly usage frequency per capita (9.2)
MPY = 12 months
Useri = No. of subscribers on the ithYear

<Table 3> Analysis of Total Utilization Value of National R&D Reports

Category	2009	2010	2011	2012	2013	2014	Total
No. of Subscribers	36,614	53,955	78,294	110,509	156,616	188,002	623,990
Utilization Value (x KRW 100 million)	118	173	251	355	503	604	1,713

3.4 B/C analysis of national R&D reports

According to B/C analysis on national R&D reports, the utilization value-based B/C was 15.5 with about 16 times of economic effects. In addition, payment value-based B/C was 11.5 with approximately 12 times of economic effects. Even though KRW 11.1 billion is spent annually to provide national R&D reports, their economic value is deemed 12-16 times greater than the costs. Therefore, the development and supply of national R&D reports appear to be economically very efficient.

<Table 4> B/C Present Value of National R&D Reports (Unit: X KRW 100 million)

Category	2009	2010	2011	2012	2013	2014	Total
Cost	5	7	22	23	21	18	95
Cost (Present Value)	7	10	27	27	23	18	111
Utilization Value (Present Value)	118	173	251	355	503	604	1,713

※ The present value of costs are estimated by converting past value into present value according to the calculation formula for the conversion of NPV after applying 7.7% of discount rate (Technology Valuation Guideline by the former Ministry of Knowledge Economy (2008), discount rate for science & technology services applied).

※ For utilization value, its present value per usage was applied. Therefore, no separate conversion was needed.

B/C of utilization value = 171.3 billion KRW / 11.1 billion KRW = 1.55 billion KRW
B/C of payment value = 127 billion KRW / 11.1 billion KRW = 1.15 billion KRW

4. Conclusion

According to a survey against scholars and engineers in science and technology, the average usage period of national R&D reports provided by the KISTI was 6.3 years with 16.6 times in average monthly detailed-view frequency, 9.2 times in average monthly original text-view frequency and 11.6 hours of average monthly usage time. The ratio and importance of national R&D reports developed and provided by the KISTI among science and technology information collected for research, business and learning purposes were 56.8 and 62.6 scores respectively.

The said results reveal that both ratio and importance of national R&D reports developed and provided by the KISTI among the collected or utilized science and technology information stand between 'neither low nor high' and 'slightly high.' The analysis results mean that respondents collect national R&D reports in 'neither low nor high' or higher levels and understand that they are pretty important. To improve the utilization of national R&D reports, therefore, this study proposes the improvement of national R&D report contents in relatively weak sectors: math, chemistry, earth science, life science, healthcare, construction/transportation and environment.

According to economic analysis on national R&D reports, 'utilization value' was the highest with 76.4 scores, followed by 'shared value (75.1)' and 'contribution value (73.0)' in qualitative indicators. These results show that the contribution value to new R&D planning is low. Because the utilization value and shared value with others are also as low as '70-80 scores,' therefore, it is needed to develop a strategy to keep enhancing values.

References

[1] Kwak, Seung-Jin, Kim, Jeong-Taek and Park, Yong-Jae. 2007. "Study on Performance Evaluation of Academic Information Distribution Project in Scientific Technology Field." Journal of Korean Library and Information Science, 38:(4): 441-462.
[2] Ministry of Security and Public Administration, "Introduction of Prism." <http://www.prism.go.kr/homepage/info/retrieveIntro.do?leftMenuLevel=310>
[3] Yang Oh-seok, 2005, "A Study of Economic Value of Utilization of Public Information," advisory report of Korea Database Agency.
[4] Yang Oh-seok, 2006, "Measurement of Economic Value of Commercial Utilization of Public Information and Analysis of its Ripple Effects," Korea Database Agency.
[5] Lee Joon, Kook Yoon-gyu, Park Min-woo, Choi Gi-seok, Kim Jae-soo, "Analysis of Current National R&D Programs in Japan," Korea Technology Innovation Society, Conference of Korean Technology Innovation Society, November 2011, pp. 258-273.
[6] Im Chang-joo, Oh Se-hong, "Successful Construction of General Management System for National R&D Programs," Korean Society for Information Management, Journal of Korean Society for Information Management, June 2002, 19(2): 93-108.
[7] Cho Hwang-hui, Jang Byeong-yeol, Chun Joo-yong, Hong Jeong-im, "A Study of Innovation of Science and Technology Information," Korea Information Society Development Institute (KISDI), 2005.
[8] Korean Intellectual Property Office, 2011, "Analysis of Patents for National R&D Programs in 2011 and Patent Performance Improvement Plan."
[9] Korea Institute of Science and Technology Information (KISTI), 2007, "Development of Performance Indicators for Academic Information Distribution Program in Science and Technology," Daejeon: KISTI.
[10] Korea Institute of Science and Technology Information (KISTI), 2012, "A Study of High Value-added Service Supply Plan Using National R&D Reports," Daejeon: KISTI.
[11] Korea Institute of Science and Technology Information (KISTI), 2013, "Utilization of High Value-added Contents in National R&D Reports and Service Strategies," Daejeon: KISTI.
[12] Kim Hui-seop, Chung Yeong-mi (2005), "Development and Application of Evaluation Model for the Economic Value of Online Information," Journal of Korean Society for Information Management, Vol. 22, No. 2, pp. 165-184.
[13] Ryu Hui-gyeong, 2006, separate volume, "A Study of Measurement of Economic Value," Ph.D. degree paper at Chung-Ang University.
[14] Ministry of Culture, Sports and Tourism, 2009, "A Study of Economic Values of Public Library."
[15] Holt, Glen E., Donald Elliott, & Amonia Moore (2004), Placing a Value on Public Library Services.<http://www.slpl.lib.mo.us/libsrc/restoc.htm>.
[16] Keating, R., O'Brien, D. & Tessler, A. (2013), Economic Evaluation of the British Library, British Library.
[17] NIH."iEdison Overview" <https://public.era.nih.gov/iedison/public/checklist.jsp>
[18] Pung, C., Clarke, A. & Patten, L. (2004), Measuring the Economic Impact of the British Library, New Review of Academic Librarianship, Vol 10, No. 1, pp. 79-102.
[19] Chung Yong-il, Moon Yeong-ho, Bae Sang-jin, Kim Yoon-jong (2005), "A Study of User Satisfaction with Information Analysis Report and Evaluation of its Economic Value," Information Management Institute, Vol. 36, No. 3. pp. 167-180

https://doi.org/10.28069/greynet-2025-000.540-tgj

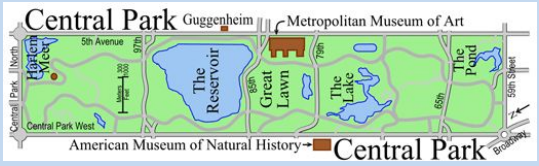
Eighteenth International Conference on Grey Literature

Leveraging Diversity in Grey Literature

THE NEW YORK ACADEMY OF MEDICINE
HEALTHY CITIES. BETTER LIVES.

1216 Fifth Avenue
New York, NY USA
November 28-29, 2016

Program Outline

DAY 1 Monday	DAY 2 Tuesday
09:00 Registration Desk opens, coffee service	**09:00 Registration Desk opens, coffee service**
09:20 Opening Session Welcome Address, Opening Address, Keynote Address	**09:00 Special Panel Session** Invited Panel Members to be announced
10:30 Morning Break	***10:00 Morning Break***
11:00 Session One Diversity in Publishing and Disseminating Grey Literature	**10:30 Poster Session and Sponsor Showcase** Multiple presentations exhibited in the President's Gallery
12:30 Lunch	***12:00 Lunch***
1:30 Session Two Open Access and Shared Grey Resources	**12:45 Session Three** Altmetrics, Aggregating Data, and Semantic Analysis
3:00 Afternoon Break	***2:15 Afternoon Break***
3:30 Introduction to Posters Briefings by Poster Presenters	**2:30 Session Four** Fueling Public Awareness to Grey Literature
4:30 Tour of the Rare Book Room	***Brief Pause***
	**4:15 Closing Session** Moderators' Report, Best Poster, Conference Handoff, Farewell **Post-Conference Tour of the Rare Book Room**

TextRelease
GL18 Program and Conference Bureau

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

https://doi.org/10.28069/greynet-2025-000.540-tgj

Eighteenth International Conference on Grey Literature

Leveraging Diversity in Grey Literature

1216 Fifth Avenue
New York, NY USA
November 28-29, 2016

Call for Posters

Title of Poster:	**Conference Topic(s):**
Author Name(s):	**Phone:**
Organization(s):	**Fax:**
Postal Address:	**Email:**
Postal/Zip Code – City – Country:	**URL:**

Participants who seek to present a poster during GL18 are invited to submit an English abstract between 300-350 words. The abstract should describe the project and/or related information product or service. The abstract should likewise include a title, name(s) of the presenter(s) and their full address information. Abstracts are an important source of information available prior to the conference that is accessible to both the delegates and conference participants.

Conference Related Topics

- ☐ **Diversity in Publishing and Disseminating Grey Literature**
- ☐ **Open Access and Shared Grey Resources**
- ☐ **Altmetrics, Aggregating Data and its Semantic Analysis**
- ☐ **Fueling Public Awareness to Grey Literature**
- ☐ **Other Poster Topic:**

Due Date and Format for Submission

Timely registration is a guarantee of your placement on the conference program. The conference venue is only able to accommodate a limited number of physical posters. Abstracts in Word should be emailed to conference@textrelease.com on or before **November 1st 2016**. The author will receive verification upon receipt and this will be accompanied by further guidelines for submission of PowerPoint slides, Conference Registration, etc.

https://doi.org/10.28069/greynet-2025-000.540-tgj

Eighteenth International Conference on Grey Literature

Leveraging Diversity in Grey Literature

THE NEW YORK ACADEMY OF MEDICINE
HEALTHY CITIES. BETTER LIVES.

1216 Fifth Avenue
New York, NY USA
November 28-29, 2016

Registration Form

The Conference Registration Fee includes attendance during the Plenary and Poster Sessions, a copy of the GL18 Program Book, the full-text of the Conference Papers and PowerPoint Slides, as well as the conference pouch and participant badge. Also included are lunches and refreshments during the morning and afternoon breaks.

Payment after October 15th 2016 add a USD 60 Surcharge

☐ **USD 595 Participant Fee** ☐ **USD 495 Author Fee** ☐ **USD 275 Day Pass** ☐ **USD 100 LIS Student**

☐ *GreyNet* Member entitled to a **20% reduction** on the conference fee.

☐ Affiliate of **The New York Academy of Medicine** entitled to a **20% reduction** on the fee.

Registrant's Name: _____

Organization: _____

Address/P.O. Box: _____

Postal/Zip Code - City - Country: _____

Tel/Fax/Email:

Check the box below for your 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 GL18 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.

TextRelease
GL18 Program and Conference Bureau

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

Author Information

Bouderbane, Azzedine **162**
Azzedine Bouderbane is professeur in library science at the University Constantine 2, Algeria. He holds A Ph.D, a Magister and a Higher Degree in the same discipline. He is fluent In three languages Arabic, French and English. He presented more than 60 papers in national and international conferences. He also published more than 30 papers In scientific journals and collective books. Expert with various international scientific institutions, he is also involved in various scientific research projects. Email: bouderbane.azzedine@yahoo.fr

Choi, Kiseok **168**
Kiseok Choi is a principal researcher of NTIS Center, Korea Institute of Science and Technology Information. He received the B.S. degree from the school of Computer Science and Statistics at Seoul National University, in 1988, and, M.S. degree from the school of Information and Communication Engineering at Korea Advanced Institute of Science and Technology, in 1997, PH.D. degree from the school of Computer Engineering at Chungnam National University 2013, respectly. His research interests include information retrieval, database and distributed data interoperability. choi@kisti.re.kr

Goggi, Sara **155**
Sara Goggi is a technologist at the Institute of Computational Linguistics "Antonio Zampolli" of the Italian National Research Council(CNR-ILC) in Pisa. She started working at ILC in 1996 working on the EC project LE-PAROLE for creating the Italian reference corpus; afterwards she began dealing with the management of several European projects and nowadays she is involved with organisational and managerial activities mainly concerning international relationships and dissemination as well as organization of events (e.g. LREC conference series). Currently one of her preminent activities is the editorial work for the international ISI Journal Language Resources and Evaluation, being its Assistant Editor. Since many years (from 2004) she also carries on research on terminology and since 2011 - her first publication at GL13 - she is working on topics related with Grey Literature. Email: sara.goggi@ilc.cnr.it

Južnič, Primož **126**
Primož Južnič is an associate professor at the Department of Library and Information Science and Book Studies at Faculty of Arts, University of Ljubljana (Slovenia). His main area of research and interest is bibliometrics, collection management and LIS education. He teaches the following courses: Bibliometrics, Special libraries, and Collection Management. Before starting his university career, he was a heading different special and academic libraries and information/computer centres. He was also working at the European Commission, for three years, as the seconded informatics expert. Email: primoz.juznic@ff.uni-lj.si

Monachini, Monica **155**
Monica Monachini is a Senior Researcher at CNR-ILC. Field of expertise: computational linguistics, computational lexicography, semantics, lexical semantics, language resources, ontologies, lexicon, terminologies, metadata, validation, methods for retrieving information in different areas (biology, environment, civil protection, oceanography, social media, humanities and social sciences, ...), infrastructural issues related to language resources. Active in many standardisation activities for harmonising lexical information. Involved and responsible of the Pisa team in many international projects for language engineering. Over the last years, she has published articles in the field of lexical resources and information extraction in different areas. Currently, she focused her activities on digital humanities. Member of various Scientific Committees; UNI delegate for ISO/TC37/SC4. Email: Monica.Monachini@ilc.cnr.it

Pardelli, Gabriella **155**
Gabriella Pardelli was born at Pisa, graduated in Arts in 1980 at the Pisa University, submitting a thesis on the History of Science. Since 1984, researcher at the National Research Council, Institute of Computational Linguistics "Antonio Zampolli" ILC, in Pisa. Head of the Library of the ILC Institute since 1990. Her interests and activity range from studies in grey literature and terminology, with particular regard to the Computational Linguistics and its related disciplines, to the creation of documentary resources for digital libraries in the Humanities. She has participated in many national and international projects including the recent projects:- BIBLOS: Historical, Philosophical and Philological Digital Library of the Italian National Research Council, (funded by CNR); - For digital edition of manuscripts of Ferdinand de Saussure (Research Programs of Relevant National Interest, PRIN - funded by the Ministry of Education, University and Research, MIUR).
Email: gabriella.pardelli@ilc.cnr.it

Author Information *CONTINUED*

Prost, Hélène **126**
Hélène Prost is information professional at the Institute of Scientific and Technical Information (CNRS) and associate member of the GERiiCO research laboratory (University of Lille 3). She is interested in empirical library and information sciences and statistical data analysis. She participates in research projects on evaluation of collections, document delivery, usage analysis, grey literature and open access, and she is author of several publications. Email: helene.prost@inist.fr

Schöpfel, Joachim **126**
Joachim Schöpfel is senior lecturer at 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 access, grey literature and eScience. He is a member of GreyNet and euroCRIS. He is also the Director of the National Digitization Centre for PhD Theses (ANRT) in Lille, France. Email: joachim.schopfel@univ-lille3.fr

Slater, Timothy F. **153**
Timothy F. Slater, Ph.D., is a professor at the University of Wyoming where he holds the Wyoming Excellence in Higher Education Endowed Chair of Science Education. At the University of Wyoming, he is working in a unique Ph.D. program in which his graduate students conduct discipline-based education research on the teaching science. His research focuses on student conceptual understanding in formal and informal learning environments, inquiry-based curriculum development, and authentic assessment strategies, with a particular emphasis on non-science majors and pre-service teachers. Prior to joining the faculty at the University of Wyoming, Dr. Slater was a tenured professor in the Astronomy Department at the University of Arizona where he built the first Ph.D. program in astronomy education research. Professor Slater earned his Ph.D. at the University of South Carolina in geophysics and his M.S. from Clemson University in astrophysics. He holds two bachelors' degrees from Kansas State University, one in science education and one in physical science.
Professor Slater is the Editor-in-Chief for the Journal of Astronomy & Earth Sciences Education and has been elected to serve on the Boards of Directors and Councils for the American Astronomical Society, Astronomical Society of the Pacific, National Science Teachers Association, and Society of College Science Teachers and has served multiple terms as chairman of the Astronomy Education Committee of the American Association of Physics Teachers and Education Liaison Committee for the American Institute of Physics. He represented the United States as the initial US National Chairman of the International Year of Astronomy. He is an author on more than 100-refereed articles, winner of numerous awards, and is frequently an invited speaker on improving teaching of science through educational research and improving teacher education.
Email: timslaterwyo@gmail.com

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; 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 FUELS AND VALIDATES GREY LITERATURE'

TGJ Volume 12, Number 3 2016