

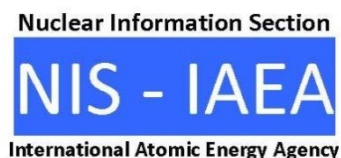
Seventeenth International Conference on Grey Literature

A New Wave of Textual and Non-Textual Grey Literature

The Royal Netherlands Academy of Arts and Sciences, Amsterdam, December 1-2, 2015



Program Sponsors:



GL17 Program and Conference Bureau

TextRelease

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



CIP

GL17 Proceedings

Seventeenth International Conference on Grey Literature : A New Wave of Textual and Non-Textual Grey literature : December 1st - 2nd 2015 at the Royal Netherlands Academy of Arts and Sciences in Amsterdam / compiled by D. Farace and J. Frantzen ; GreyNet International, Grey Literature Network Service. Amsterdam : TextRelease, February 2016. – 184 p. – Author Index. – (GL Conference Series, ISSN 1386-2316 ; No. 17).

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 NYAM (USA) are Corporate Authors and Associate Members of GreyNet International. These proceedings contain the full text of conference papers presented during the two days of plenary and poster sessions. The papers appear in the same order as in the conference program book. Included is an author index with the names of contributing authors/co-authors along with their biographical notes. A list of more than 50 participating organizations as well as sponsored advertisements are likewise published in these proceedings.



Foreword

A NEW WAVE OF TEXTUAL AND NON-TEXTUAL GREY LITERATURE

As the internet becomes increasingly grey and every cloud now has a grey lining, there arises the need to address a new and challenging wave of textual and non-textual grey literature. GL17 will examine a number of new types of textual grey literature both web-based and submerged in the sea of social networks. No less attention will be drawn to the expanding quantity of non-textual grey literature accessible in visual, audio, and diverse data formats and frequencies. Actually, in order to grasp this new wave of grey literature it may be even more advantageous to look at the convergence of these new types of textual and non-textual content rather than focus separately on each. The problems textual grey literature faced and addressed over the past quarter century are to a certain extent very similar to what non-textual grey literature faces today. The wide range of graphics, photographs, and other data-intensive grey literature is obscure, hard to find, and often short lived because it lacks proper indexing and sustained access. Such non-textual grey literature requires interpretation and documentation, which can in part be achieved by linking and crosslinking to their related textual counterparts. In this way, grey literature becomes leveraged and its value and return on investment made transparent.

While bridging textual and non-textual content is technically possible, it also requires an information policy in place that supports these new digital assets. Likewise, information professionals and practitioners must be able to (re)appropriate human resources and streamline their workflow in innovative ways. These should allow for content and feedback generated in social networks and in particular the information communities served.

These proceedings share the work of over fifty authors and researchers recording their experience and vision on how to channel this new wave of grey literature.

Dominic Farace
GREYNET INTERNATIONAL

Amsterdam,
FEBRUARY 2016

GL17 Conference Sponsors



DANS, The Netherlands
Data Archiving and Networked Services,
Royal Netherlands Academy of Arts and Sciences



KISTI, Korea
Korea Institute of Science and Technology
Information



CVTISR, Slovak Republic
Slovak Centre of Scientific and Technical Information



EBSCO, USA



NIS-IAEA, Austria
Nuclear Information Section
International Atomic Energy Agency

GL17 Conference Sponsors



NYAM, USA
The New York Academy of Medicine



NTK, Czech Republic
National Library of Technology



FEDLINK, USA
Federal Library Information Network
Library of Congress



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



ISTI, Italy
Institute of Information Science and Technologies
National Research Council, CNR

GL17 Program Committee



Marnix van Berchum **Chair**
Data Archiving and Networked Services, DANS-KNAW Netherlands



Danielle Aloia
New York Academy of Medicine, NYAM United States



Stefania Biagioni
Institute of Information Science and Technologies, ISTI-CNR Italy



Ján Turňa
Slovak Centre of Scientific and Technical Information, CVTI SR Slovak Republic



Tomas Lipinski
University of Wisconsin - Milwaukee United States



Petra Pejšová
National Library of Technology, NTK Czech Republic



Dobrica Savić
Nuclear Information Section, NIS-IAEA Austria



Joachim Schöpfel
University of Lille 3 France



Blane Dessy
FEDLINK, Library of Congress United States



Christiane Stock
Institut de l'Information Scientifique et Technique, INIST-CNRS France



Dominic Farace
Grey Literature Network Service, GreyNet International Netherlands



Table of Contents

	Foreword.....	3
	Conference Sponsors.....	4
	Program Committee.....	6
	Conference Moderators.....	8
	Conference Program.....	9
<i>Program</i>	Opening Session.....	11
	Session One - Convergence of Textual and Non-Textual Grey Literature.....	39
	Session Two - Influence of Social Media and Networks on Grey Literature.....	69
	Poster Session - Presentations Showcasing Grey Literature.....	95
	Session Three - Innovative Ways in leveraging Grey Document Types.....	113
	Session Four - Visualizing Content in and for Grey Communities.....	137
<i>Advertisements</i>	EBSCO LISTA Full-Text.....	10
	NYAM, The New York Academy of Medicine.....	74
	GreyNet LinkedIn Discussion Group.....	82
	CVTISR, Slovak Centre of Scientific and Technical Information.....	94
	FEDLINK, The Federal Library and Information Network - Library of Congress.....	100
	KISTI, Korea Institute of Science and Technology Information.....	112
	NTK, National Library of Technology, Czech Republic.....	124
	INIS, The International Nuclear Information System.....	146
	TGJ, The Grey Journal.....	174
<i>Appendices</i>	List of Participating Organizations.....	175
	GL18 Conference Announcement.....	176
	GL18 Call for Papers.....	177
	Author information.....	178
	Index to Authors.....	183
	GL17 Publication Order Form.....	184

**Moderator Day One**

**Marnix van Berchum,
Head of Data Services at DANS**

Marnix is responsible for the coordination of the services DANS offers (including EASY, the narcis.nl portal, and the Dutch Dataverse Network). After his studies in Musicology at Utrecht University, Marnix started working at Utrecht University Library. In the years 2010-2012 he also worked at SURF. At both employers he was primarily involved with projects related to Open Access and innovations in scholarly communications. Marnix combines his work at DANS with a PhD trajectory at Utrecht University, where he seeks to apply the concepts and methods of network theory on the dissemination of music in the 16th C. He is Associate Director of the CMME Project.

marnix.van.berchum@dans.knaw.nl

**Moderator Day Two**

**Jens Vigen,
Head Librarian CERN**

For over a decade, Jens has been deeply involved in designing digital library services. In parallel to developing new services for members of the particle physics community, he has a particular interest for redesigning business models in the digital era for purchasing of library materials. Recently his activities have been strongly focused on establishing models for open access journal publishing. Before joining CERN, Jens held a position at the library of the Norwegian University of Science and Technology. In addition to his library qualifications he has a master degree in civil engineering; geodesy and photogrammetry.

jens.vigen@cern.ch

Opening Session

Opening Address	Non-literary text and Non-textual literature	11
	Peter Doorn and Theo W. Mulder, Royal Netherlands Academy of Arts and Sciences, KNAW	
Keynote Address	Dissertations and Data	15
	Joachim Schöpfel, Univ. of Lille, France; Primož Južnič, Univ. of Ljubljana, Slovenia; [et al.]	

Session One - Convergence of Textual and Non-Textual Grey Literature

Move beyond text – How TIB manages the digital assets researchers generate,	39
Margret Plank and Paloma Marín Arraiza, German National Library of Science and Technology (TIB), Germany	
Situation surrounding grey literature in academic research in Japan,	44
Yoshio Isomoto, Hokkaido University, Japan	
Copyright Reform and the Library and Patron Use of Non-text or Mixed-Text Grey Literature: A Comparative Analysis of Approaches and Opportunities for Change	46
Tomas A. Lipinski and Katie Chamberlain Kritikos, School of Information Studies; University of Wisconsin, USA	
Library as Publisher: Convergence of New Forms and Roles of Textual and Non-Textual Grey Literature in Digital Scholarship	61
Julia Gelfand, University of California, Irvine; Anthony Lin, Irvine Valley College, United States	

Session Two - Influence of Social Media and Networks on Grey Literature

Academic blogging consequences for Open Science: a first insight into their potential impact	69
Carla Basili, National Research Council of Italy, CNR-IRCrES Institute; Luisa De Biagi, CNR-Biblioteca Centrale, Italy	
Share #GreyLit: Using Social Media to Communicate Grey Literature	75
Danielle Aloia and Robin Naughton, New York Academy of Medicine, NYAM Library, United States	
Public sharing of medical advice using social media: an analysis of Twitter	83
Gondy Leroy, Eller College of Management, University of Arizona; Philip Harber, Zuckerman College of Public Health, University of Arizona; Debra Revere, School of Public Health, University of Washington, United States	

Poster Session

International identification and ‘white and grey literature’ : Identities, retrieval, reuse and the certainty of knowledge while sharing and connecting information,	95
Flavia Cancedda, CNR - Biblioteca Centrale, ISSN National Reference Centre; Luisa De Biagi, CNR - National Research Council of Italy, Biblioteca Centrale, Italy	
Sustaining Scholarly Communication Support by Academic Libraries In Sub-Saharan Africa: A Case of Makerere University and University of Zimbabwe Libraries,	101
Andrew Mwesigwa, Makerere University, Uganda; Elizabeth Mlambo, College of Health Sciences Library, University of Zimbabwe	
A semantic engine for grey literature retrieval in the oceanography domain	104
Sara Goggi, Gabriella Pardelli, Roberto Bartolini, Francesca Frontini, Monica Monachini, CNR-ILC, Italy Giuseppe Manzella, ETTsolutions; Maurizio De Mattei and Franco Bustaffa, DP2000, Italy	

Session Three - Innovative Ways in leveraging Grey Document Types

Analysis of National R&D Project Report Output Utilization and Economic Contribution	113
Kiseok Choi, Cheol-Joo Chae, Yong-hee Yae, Yong Ju Shin, KISTI, Korea	
Scientific Audiovisual Materials And Linked Open Data: The TIB Perspective	119
Paloma Marín Arraiza, German National Library of Science and Technology (TIB), Germany	
The National Portal for Recording Theses (PNST): Its Role, Importance and Constraints for Algerian Researchers	125
Azzedine Bouderbane, Nadjia Gamouh, and Hadda Saouchi, University Constantine 2, Algeria	
Grey Literature Sources in Historical Perspective : Content Analysis of Handwritten Notes	131
Snježana Ćirković, Faculty of Philology, University of Belgrade, Serbia	

Session Four - Visualizing Content in and for Grey Communities

Grey Literature citations in the age of Digital Repositories and Open Access	137
Silvia Giannini and Stefania Biagioni, Institute of Information Science and Technologies ISTI-CNR; Sara Goggi and Gabriella Pardelli, Istituto di Linguistica Computazionale, ILC-CNR, Italy	
Public Interest in Accessing the INIS Collection,	147
Dobrica Savić, Nuclear Information Section, IAEA, Austria	
Extracting value from grey literature: processes and technologies for aggregating and analyzing the hidden “big data” treasure of organizations,	154
Gabriele Motta, Roberto Puccinelli, and Lisa Reggiani, Network and Information System Office, CNR; Massimiliano Saccone, Central Library, CNR, Italy	
Leveraging Grey Literature – Capitalizing on Value and the Return on Investment: A Cumulative Case Study	165
Dominic Farace and Jerry Frantzen, GreyNet International, Netherlands, Stefania Biagioni and Carlo Carlesi, ISTI-CNR, National Research Council, Italy, Christiane Stock, Inist-CNRS, National Centre of Scientific Research, France	

Library, Information Science & Technology AbstractsTM with Full Text

Available via EBSCOhost[®]

The definitive professional information resource designed for librarians and information specialists...

Library, Information Science & Technology AbstractsTM with Full Text is an indispensable tool for librarians looking to stay current in this rapidly evolving field.

Comprehensive content includes:

- Full text for more than 270 journals and nearly 20 monographs
- Indexing for more than 550 core journals, 50 priority journals and nearly 125 selective journals
- Includes books, research reports, proceedings and author profiles
- Access to 6,800 terms from reference thesauri
- Coverage extends back as far as the mid-1960s

Subject coverage includes:

- Bibliometrics
- Cataloging
- Classification
- Information Management
- Librarianship
- Online Information Retrieval
- And much more...

Contact EBSCO Publishing to learn more about *Library, Information Science & Technology AbstractsTM with Full Text*, or to request a free trial.

Phone: 800.653.2726

Email: request@ebscohost.com

www.ebscohost.com

Opening Address

Non-literary text and Non-textual literature

Peter Doorn and Theo W. Mulder,
Royal Netherlands Academy of Arts and Sciences, KNAW

Introduction

The computer and accompanying revolution in information and communication technologies have played a crucial role in widening, deepening and accelerating research. Computing has improved enormously the potential for disseminating, finding and retrieving scientific information and documentation. All aspects and areas of scientific inquiry and scholarly research have been affected by the technological developments. Also the world of libraries, archives and documentation centres has completely changed in the past decades, and that process of change has not at all reached the end. We are living, indeed, in the decades of information.

In current times we see how, for instance, academic libraries are searching for new functions and indeed even for a new identity, now that paper as the main carrier of information and knowledge is rapidly being replaced by digital media. Most University Libraries for instance have the ambition to play a role as intermediary in data services. But it is not always certain that old players are best placed to answer the new questions or cater for new demands. Information and communication technologies even tend to squeeze out the intermediaries between producers and consumers. Especially the internet makes it possible that consumers get to the products and services they want directly, without intervening parties, except perhaps a general search engine. Take for example the music industry, where it is the trend that we can stream our music directly from a few very big suppliers, bypassing the whole chain of wholesalers and retailers.

Looking at the title of the conference: *“A New Wave of Textual and Non-Textual Grey Literature”*, it seems apparent that GreyNet is also witnessing the effects of the changing nature of the core material it is used to focus on: grey literature.

In the past, when paper was the dominant information carrier, grey literature was easy to distinguish from official literature. It is clear that, indeed, there is a new, and enormous, wave of unofficially published documents. At the same time, when looking at the second part of the conference theme: *“textual and non-textual grey literature”*, we get the feeling of some bewilderment. We more or less understand the distinction between a literary text and a non-literary text. But what should be understood by non-textual literature? Perhaps this is a reflection of the fact that in the digital world the character and meaning of the term document has become fluid. If we replace the word *“document”* by *“data”*, as we suspect that is meant by *“non-textual literature”*, then the question is whether the term grey literature does not become more or less boundless and loses its meaning.

KNAW and Grey Literature in the 1970s

The Royal Academy has a history with *“grey literature”*. From 1972 until 1996, the Academy was host of the Sociaal-Wetenschappelijk Informatie- en Documentatiecentrum (SWIDOC), which included the *“Rapportencentrale”*. This department collected unofficially published reports that either contained results from social and policy research, or policy reports that could be relevant for such research. Such reports were produced by academic researchers, but also by all kinds of national, provincial and local government agencies. These reports, often lead a hidden life on the bookshelves or in the drawers of public servants. Perhaps even not officially printed, but stencilled or photocopied in limited numbers – was it the poor reproduction quality that was responsible for the very name grey literature? At best they were kept in the reference library of the department, but in general they were hard to find and hardly accessible, and many of them got lost or were thrown away after some time. Hence it was the duty of the Rapportencentrale to trace such reports, to collect them, to catalogue and store them, and to disseminate such catalogues, originally in the form of a bibliographic journal.¹

¹ http://www.politiekcompendium.nl/id/vh4vaujpk7sj/algemene_bibliografische_hulpmiddelen



Figure 1. The SWIDOC Catalogues in 1991.

Source: <https://www.knaw.nl/shared/resources/actueel/bestanden/nr11.pdf>

The digital turn and Grey Literature

In the 1990s the very base of this grey information also turned digital, and the ways in which users could access the reports also changed, because both researchers, institutes and government agencies started to create websites on which they “published” their reports.

This digital turn of both social research and government agencies had a collateral effect on the very nature of “grey literature”, which began to shift. Similar changes took place in the area of official publications. One could say that, as the medium turned digital, the border of the publication, either grey or official, began to blur, especially because of the properties of hypertext and http-links on the world wide web.

Text could now easily be connected with all kinds of related materials that in the past simply could not be printed in a paper journal. An article could now be linked with the original research data on which the argumentation was based, to models and algorithms used, to visualizations and graphical representations, to spreadsheets containing tables with which the reader/user could interact, and even to additional information about the researcher, his institute, and the projects she worked on. In the 2000s, the term “enhanced publication” was coined to refer to such articles with all kinds of extensions.²

The blurring boundaries of types of information

The boundary between grey and official literature became blurred too, as is clearly illustrated by the foundation in 1991 of the e-print archive arXiv.org, created by Paul Ginsparg at Los Alamos National Laboratory. In the current discussions about Open Access, many unofficial or “author” versions of articles published in official journals can now be found in the institutional repositories of universities (= green open access).

Against this background it would not be surprising to imagine that the grey literature community witnessed an identity crisis. What exactly does the term stand for in the digital age? And how to provide optimal access to this “digital grey literature” in all its pluriformity? Who is or rather who feel responsible for this?

Parallel to the changing challenges of ICT and research, in 1997 SWIDOC was merged with a number of other Academy institutes into the Netherlands Institute for Scientific Information Services (NIWI). The new institute was tasked with improving and innovating the information supply to Dutch researchers (and beyond), and around that time the journal and other bibliographic tools were turned into online-accessible databases. The social science data archive

² https://en.wikipedia.org/wiki/Enhanced_publication; <https://www.surf.nl/en/themes/research/research-data-management/enhanced-publications/index.html>

(Steinmetz Archive, also part of SWIDOC) and the Netherlands Historical Data Archive were also incorporated in the merger.

In 2005 the Data Archiving and Networked Services (DANS) was one of the institutes succeeding NIWI. It was created by KNAW and the Dutch Funding Organisation for Research NWO. The core mission of DANS was to create the ideal conditions in the field of research data: a single, transparent data infrastructure of outstanding international quality. It was tasked to become the foremost national facility for research data. Researchers should be able to turn to DANS for quality and timely data, regardless of where those data were produced, stored or made accessible.

Research data and grey literature are clearly not the same, although they are related. Research data is hardly ever officially published in a journal, and therefore in this sense research data is “grey” by definition. But data is not literature. However, if we accept the term “non-textual literature”, as in the title of the 17th GreyNet conference and its proceedings, the blur is almost complete.

“Non-textual literature” and other definitions of Grey

Of course textual data is an often-used term, and literary texts are used by scholars for all kinds of analytical purposes, including text mining, linguistic analysis, etc. “Non-textual literature” seems to refer to everything “grey” that is not a digital representation of the classical, textual, report, which was the core business of the Rapportencentrale in the 1970s-90s.

Looking at definitions of grey literature over time, we come across the following: the “Luxembourg definition”, discussed and approved at the Third International Conference on Grey Literature in 1997, defined grey literature as *“that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers”*.

The Fourth International Conference on Grey Literature (GL '99) in Washington, DC, in October 1999, maintained the same definition. Grey literature publications were further specified as *“non-conventional, fugitive, and sometimes ephemeral publications. They may include, but are not limited to the following types of materials: reports (pre-prints, preliminary progress and advanced reports, technical reports, statistical reports, memoranda, state-of-the art reports, market research reports, etc.), theses, conference proceedings, technical specifications and standards, non-commercial translations, bibliographies, technical and commercial documentation, and official documents not published commercially (primarily government reports and documents) (Alberani, 1990).³*

In 2004, at the Sixth Conference in New York, a postscript was added for purposes of clarification: grey literature is *“...not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body”*.⁴ This definition has since been used extensively and is widely accepted.⁵

But in 2010, Farace and Schöpfel pointed out that existing definitions of grey literature were predominantly economic, and argued that in a changing research environment, and with new channels of scientific communication, grey literature needed a new conceptual framework. They proposed a new definition (“Prague Definition”): *“Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers i.e., where publishing is not the primary activity of the producing body.”*⁶

This definition opens the door to an enormous variety of digital materials. The current list of document types of GreyNet counts about a 150 entries, including databases, datasets, software and websites. One could almost say: the whole content of the internet (a lot of which is by the way textual) can be seen as grey literature, except for what is commercially published!⁷

³ <http://www.greylit.org/about>

⁴ Schöpfel, J. & Farace, D.J. (2010). “Grey Literature”. In: Bates, M.J. & Maack, M.N., *Encyclopedia of Library and Information Sciences* (3rd ed.). Boca Raton, Fla.: CRC Press. pp. 2029–2039.

⁵ https://en.wikipedia.org/wiki/Grey_literature

⁶ Schöpfel, J., “Towards a Prague Definition of Grey Literature”, in: *The Grey Journal (TGJ): An international journal on grey literature*, volume 7:1 (2011).

⁷ It is remarkable that e-mail, Twitter feeds and Facebook pages are not yet included as categories of grey literature

GreyNet: expanding the scope or focusing on a niche?

The question we would like to raise for discussion is the following: apparently GreyNet has over the years expanded its scope, from a concentration on unofficially published reports relevant for social science research, to virtually all imaginable information on the web. Print reports are also still in scope, but analogue sources have obviously become less relevant for research purposes.

Meanwhile, the number of players in the world of digital research data and information has increased enormously. The registry of research data repositories re3data.org alone now lists almost 1400 entries all over the world⁸; probably this list is not even complete, and given its focus it may not contain the institutional publication repositories.

But given the huge amount of “grey information” on the Internet and the growing number of players such as repositories and data infrastructures, which are taking care of particular information types, would it not be better for the grey literature movement to focus on a niche much closer to the original area of grey documents, instead of trying to cover almost everything on the web? Is even the term “grey” not a reflection of the past, reminiscent of the times of wet photocopying? Would not a concentration on a well-chosen selection of document types, that seem to be forgotten by most other repositories, libraries and archives, be a more fruitful and realistic strategy?

Many reports of the grey kind as described at the beginning of this paper, are still not very accessible, documented, or archived for long-term preservation. In this country, the National (Royal) Library aims to store all official publications in its e-depot, but it tends to neglect grey reports. The DANS archives contain a rather good coverage of archaeological reports, not because it systematically focuses on reports in general, but because these contain the basic data on excavations carried out in the past. DANS also provides access to materials in academic repositories in the Netherlands, and aims to connect the various information types in a kind of “enhanced publications” (from the DANS perspective they are rather “enhanced datasets”).

About 40% of the publications in the Dutch academic repositories are openly accessible, but the universities do not yet register whether the type of open access is green or gold, and neither whether the publication is grey or official.⁹ It is even questionable how relevant this last distinction is. About 54% of the content of these repositories is classified as article, 14% as book or book part, 6% as doctoral thesis. About one quarter consists of reports, conference papers, preprints, etc.

Concluding suggestions

GreyNet could even consider a “negative definition”, as a defender of the interest of classes of materials that nobody else cares for. Indeed, the access to a lot of government reports with a value for (social) scientific research is still far from optimal. Although the Dutch government actively pursues an “Open Data” policy and provides access to public data via its Open Data Portal, a lot of information is hidden and is bound to quickly disappear.¹⁰ GreyNet could effectively focus on stimulating open access to government reports, perhaps in collaboration with organizations such as the Open State Foundation.¹¹ As “Open Science” will be the theme of the Dutch Presidency of the EU in the first half of 2016, the climate for this seems to be right.

In sum, our recommendations to the grey community are as follows:

- Focus on a more limited set of “grey documents”, closer to the original definition, instead of an overambitious attempt to cover the majority of the information on the world wide web.
- Support the open access and long-term archiving of such grey reports.

A large sign with the following message used to hang in the main hall of the *Institute for Human Learning and Cognition* in Minneapolis: “the first law of progress is to give a challenging name to a non-existing thing”. In line with this, our question is whether it would be possible to think of another term than grey literature which is reminiscent of the grey tinted wet photocopies of the 1970’s or of a digital support group of elderly citizens.

⁸ <http://www.re3data.org/> (last checked on January 27, 2016)

⁹ Based on www.narcis.nl (last checked on January 27, 2016)

¹⁰ <https://data.overheid.nl/>

¹¹ <http://www.openstate.eu/nl>

Keynote Address

Dissertations and Data

Joachim Schöpfel, GERiiCO Laboratory, University of Lille 3, France
Primož Južnič, Department of LIS, University of Ljubljana, Slovenia
Hélène Prost, CNRS, associated member of GERiiCO Laboratory, France
Cécile Malleret, Academic Library, University of Lille 3, France
Ana Češarek, Department of LIS, University of Ljubljana, Slovenia
Teja Koler-Povh, Academic Library, University of Ljubljana,
Faculty of Civil and Geodetic Engineering, 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 economy¹ but also of science and engineering.

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/>

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 experts². 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).

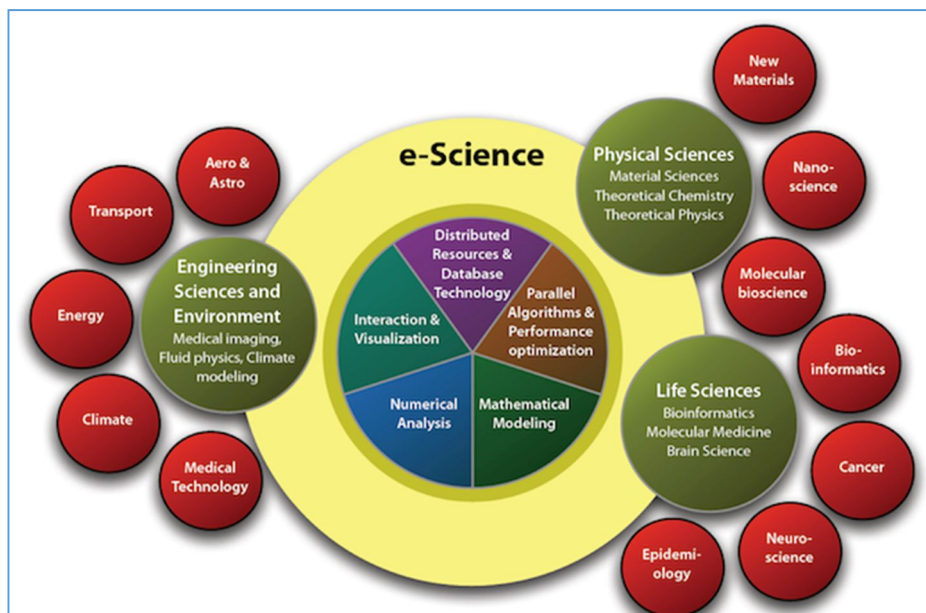


Figure 1 : Illustration of e-Science (by KTH Royal Institute of Technology in Stockholm³)

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 110⁴, 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 re3data⁵ 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.

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/>

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 Library⁶)

We will not describe the whole life-cycle in detail, because of its complexity⁷ 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 datasets⁸.

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.

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/>

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.

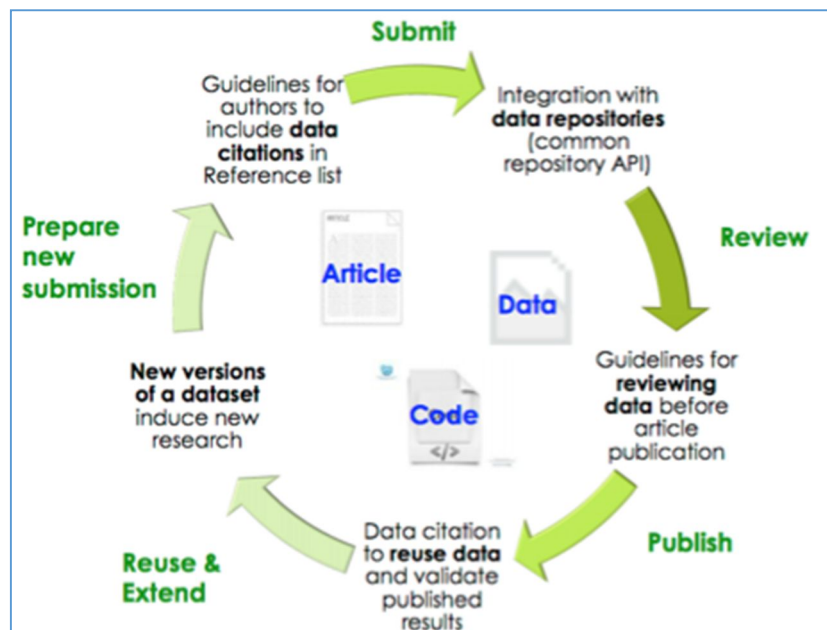


Figure 3: Diagram of an Automated Integrated Article and Data Publishing Workflow (Dataverse Project, Harvard University⁹)

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

⁹ <http://datascience.ig.harvard.edu/blog/bridge-publishing-words-publishing-data>

use standardized language and document structures in research publications” (Morris et al. 2005). The same remark applies to grey literature, including theses and dissertations¹⁰ (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 collections¹¹.

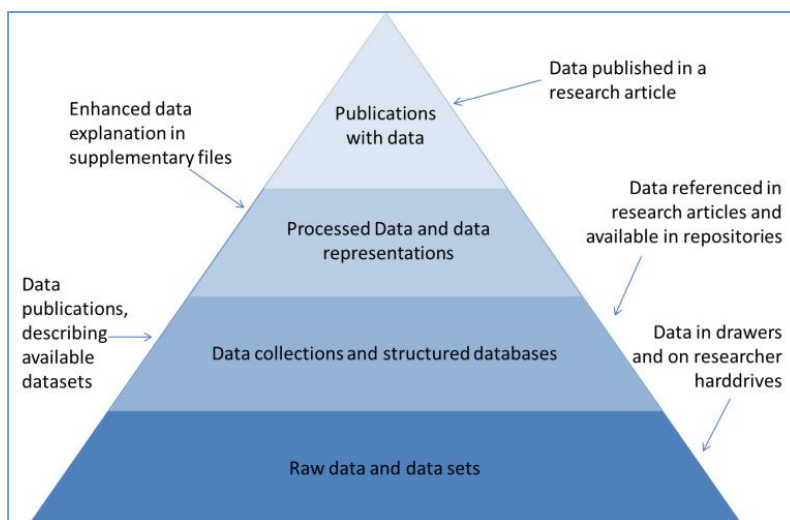


Figure 4: Data publication pyramid (from Reilly et al. 2011¹²)

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.

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>

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 portal¹⁴ 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 NDLTD¹⁵ references 4.3 million theses and dissertations.

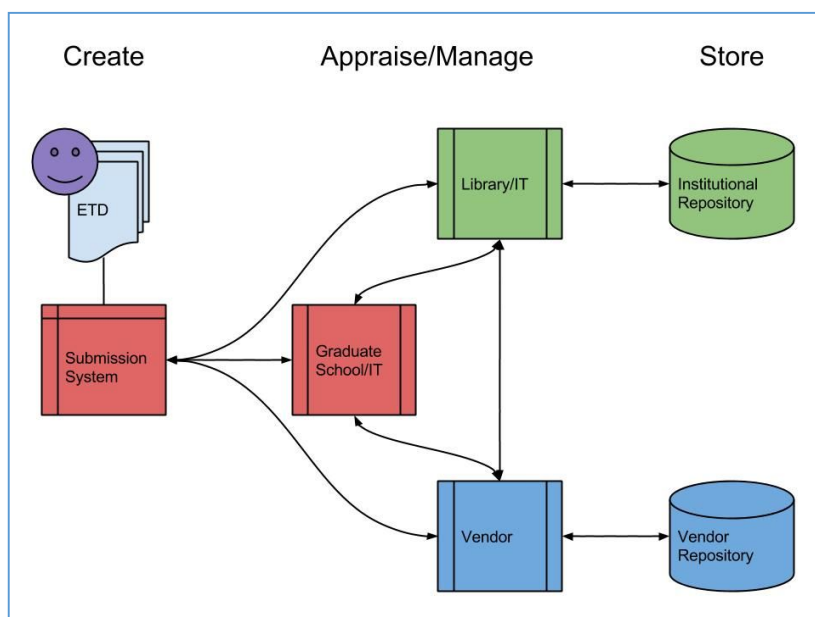


Figure 5: Life-cycle management of ETDs (Educopia project¹⁶)

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:

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>

- 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.

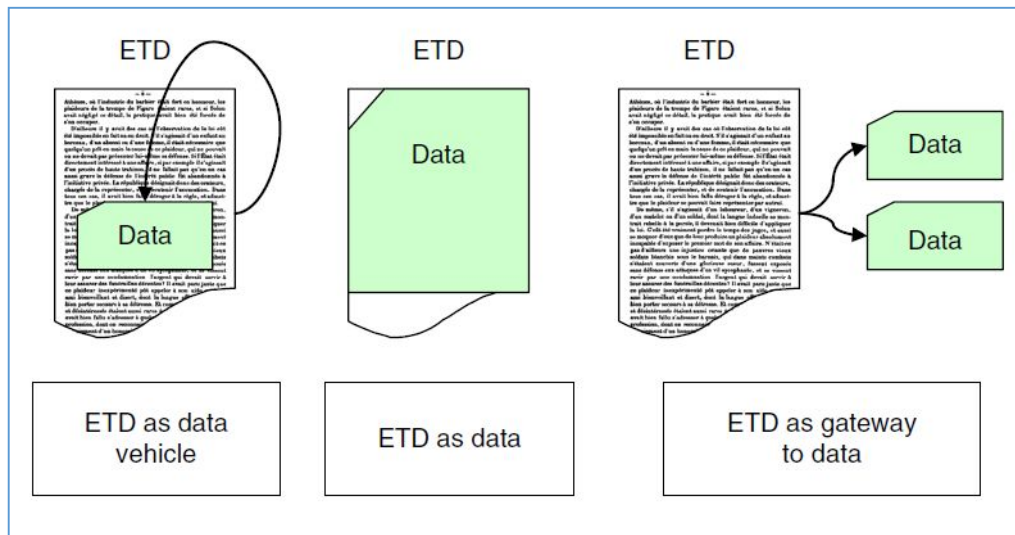


Figure 6: ETD as data, data vehicles and gateway to data

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.

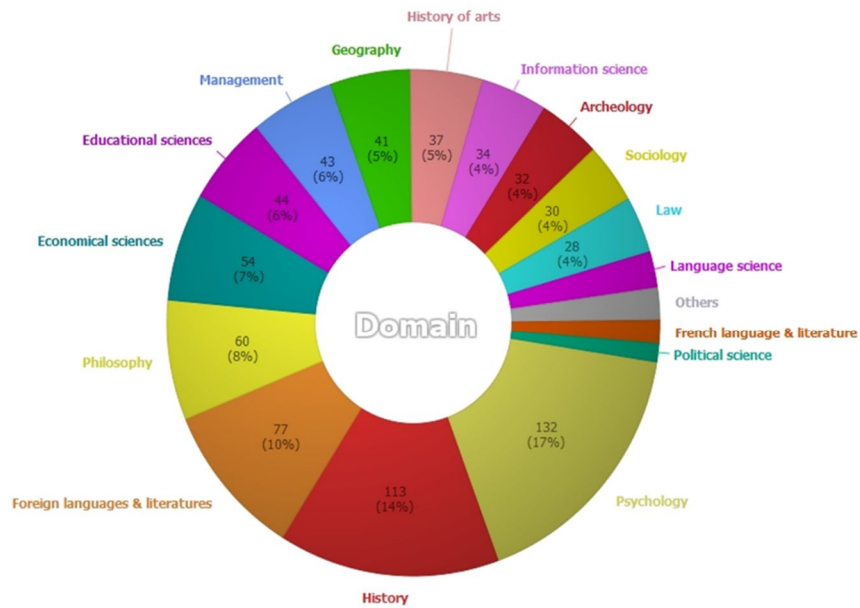
5.2. Research data in dissertations: a French-Slovenian survey

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).

¹⁷ <https://educopia.org/research/grants/etdplus>



Subdiviser: Aucun

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).

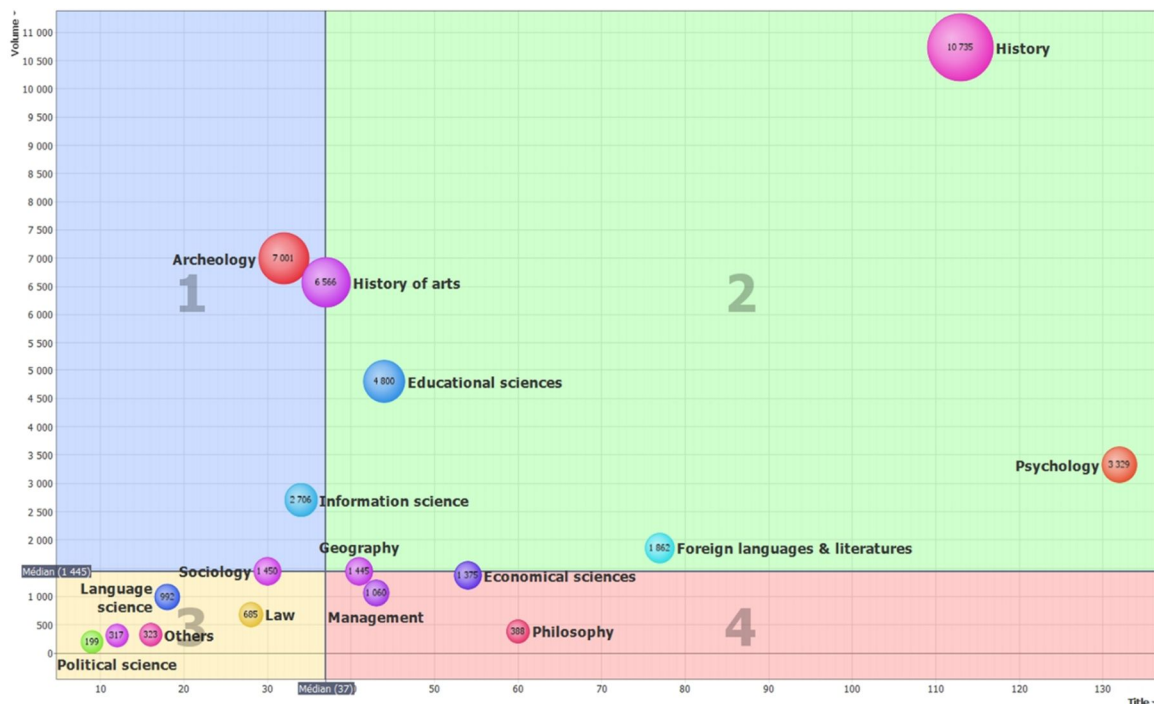


Figure 8: Size of data appendices (in pages, N=522 dissertations)

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 R^2 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).

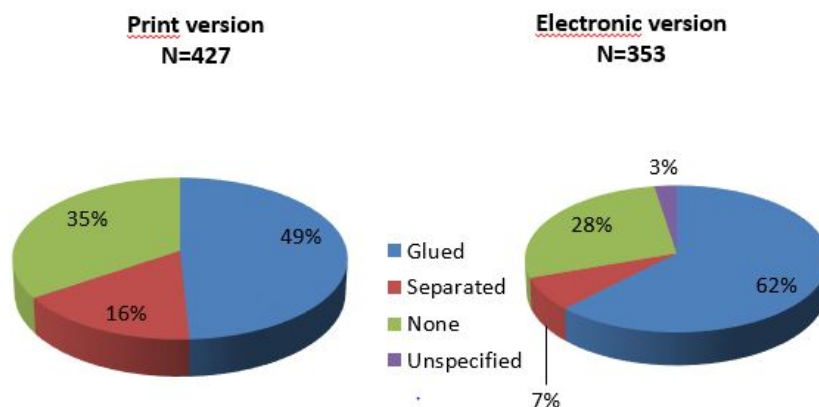


Figure 9: Link between text and appendices (in %, N=780)

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).

Sources of appendices

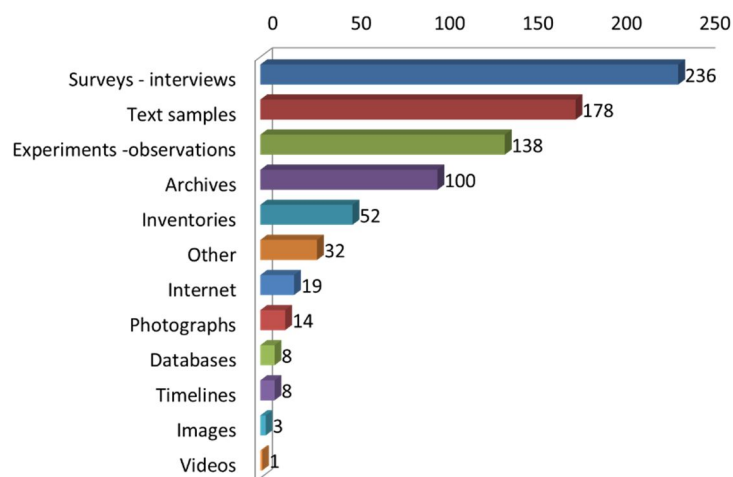


Figure 10: Data sources per dissertation (N=780)

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.

Discipline	Archives	Databases	Experiments - observations	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

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).

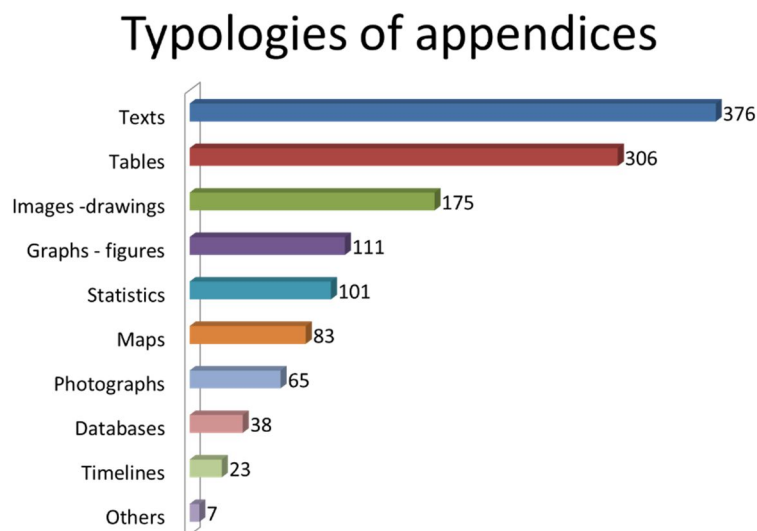


Figure 12: Data types, per dissertation (N=780)

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%).

Discipline	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

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 archaeological 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 sciences¹⁸

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 DRUGG¹⁹ 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.

¹⁸ Study conducted by Teja Koler Povh

¹⁹ <http://drugg.fgg.uni-lj.si>

DRUGG - typologies of appendices

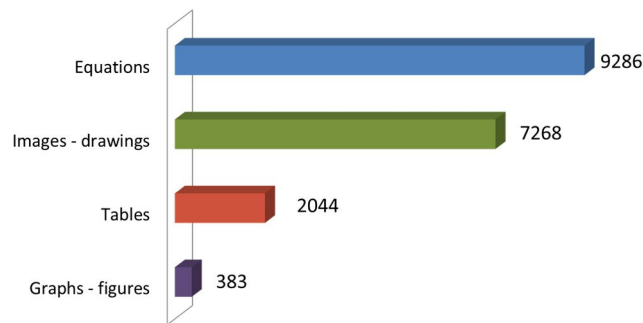


Figure 14: Data types in the DRUGG repository (civil engineering, N=86)

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).

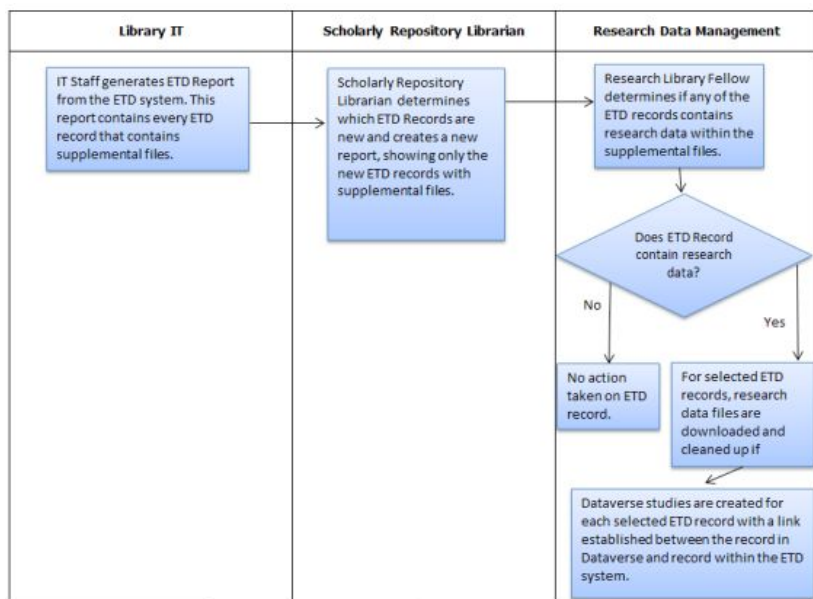


Figure 15: Workflow for Dataverse ETD Pilot Program at Emory (Doty et al. 2015)

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.

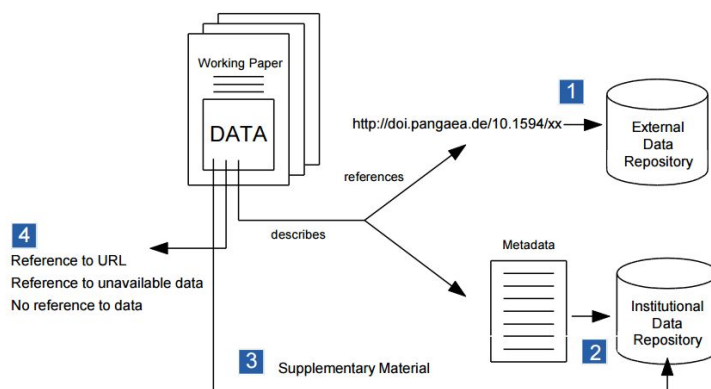


Figure 16: Linking grey literature with research data as discrete resources (Vompras & Schirrwagen 2015)

This process can be done in smooth and simple way, as a growing number of repositories show, such as the Bielefeld PUB²⁰ 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

²⁰ <https://pub.uni-bielefeld.de/>

²¹ <http://repository.tudelft.nl/>

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?

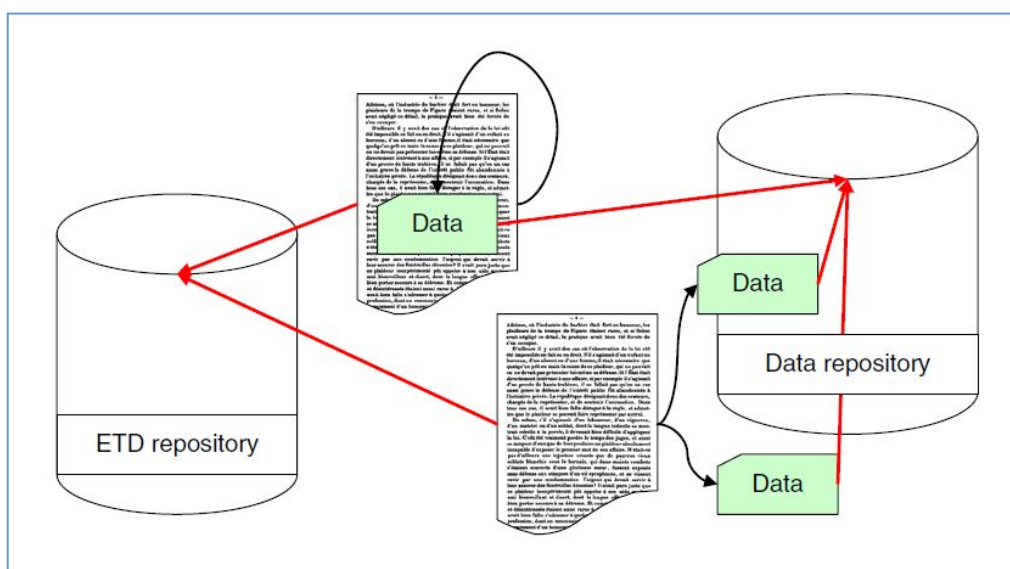


Figure 17: Storage of ETD as related datasets

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

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

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 access²³. 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).

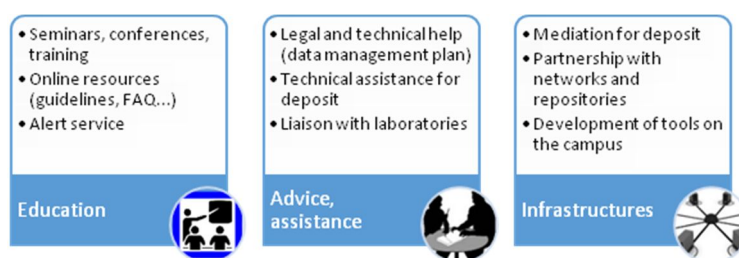


Figure 18: Research data support services

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 humanities²⁴. 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 universities²⁶. 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 rule²⁸. 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

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>

“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 system²⁹. 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).

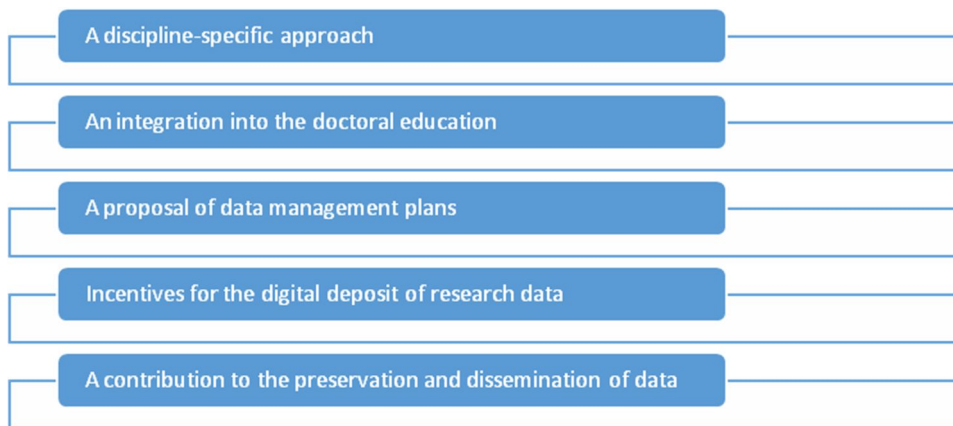


Figure 19: Five leading principles for the implementation of research data support services

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 CCSD³¹ and part of the national open repository HAL³².

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

²⁹ See Stuart Lewis’ blog posts on the Edinburgh 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/>

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.

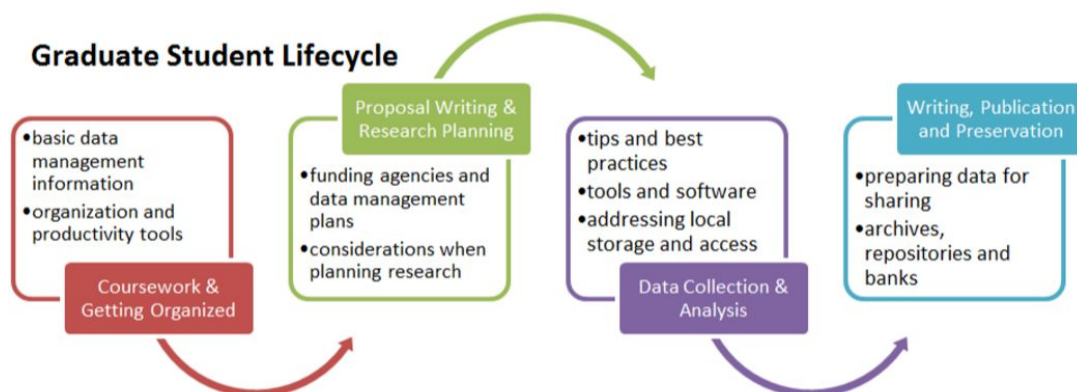


Figure 20: Research and data lifecycle (source: University of Virginia Data Management Portal)

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 ND LTD, 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.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. vvh, 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., Krabbenhoef, 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-IGIS-Wang-etal.pdf>

All references are here: <http://www.citeulike.org/user/Schopfel/tag/g17>

Move beyond text – How TIB manages the digital assets researchers generate

Margret Plank and Paloma Marín Arraiza
German National Library of Science and Technology (TIB)

Abstract

The supply, use and significance of non-textual materials is steadily increasing in the areas of research and teaching. Digital assets like scientific videos, 3D objects and re-search data are highly relevant in order to make science reproducible. Yet, they often present a lack of appropriate metadata, unique identification, and long-term preservation, remaining beyond the relevant and journal-based scientific publication system. The TIB has established a competence centre for non-textual materials in order to improve the access and the use of those digital assets. As a use case of a service for scientific videos the TIB|AV-Portal is presented. Further best practices are also discussed.

Keywords: non-textual materials, scientific videos, access, multimedia retrieval, library services, scientific communication, infrastructure.

1. Introduction

Today's research papers move beyond text as they include a variety of comprehensive digital assets such as research data, audiovisual media, 3D-objects and even software code. Those digital assets have a unique life cycle within their scientific communities. However, only a negligible proportion of those digital assets are accessible at present, whilst scientific texts are, in principle, sufficiently well-documented and available. This can lead to serious problems when it comes to preservation and potential of re-use. Consequently, valuable scientific information cannot be found, cited or re-used and remain hidden when a research project ends or the responsible scientists transfer to other institutions (Kraft et al., 2015).

In 2011 TIB established a competence centre for non-textual materials, which focuses on strategies regarding "Move beyond text". The competence centre aims at improving ease of access and use of non-textual material, such as audiovisual media, 3D objects and research data. It concerns itself with the development of tools and infrastructure that actively support users in the scientific work processes, enabling the easy publication, finding and long-term availability of non-textual media. To accomplish this, the TIB actively engages in communities, providing support and advice to fellow libraries and other preservers of scientific knowledge and its associated digital assets. This paper addresses how the TIB manages the digital assets researchers generate using the example of scientific audiovisual media. The TIB|AV-Portal¹ shows how to overcome the appearing challenges regarding scientific videos in particular.

The structure of the paper is as follows: this section gives an overview of the situation and presents the problem. Section 2 approaches the management of scientific audiovisual media from a library perspective and presents the TIB|AV-Portal as use case for scientific video service. Section 3 summarises the essential conclusions.

2. Managing scientific audiovisual material in the library context: The TIB|AV-Portal

2.1 Background

Scientific audiovisual media such as computer visualizations, simulations, research laboratory experiments, interviews and recordings of lectures and conferences have become an important part of scientific communication (Spicer 2014). These media contribute to a better understanding and discussion of research work and results because they can describe dynamic phenomena which cannot be precisely detailed in words and pictures (Whitesides for the American Chemical Society, 2011).

The publication of scientific audiovisual media goes beyond the traditional journal-based scientific publication system (Löwe et al., 2015). Audiovisual scientific media are usually shared using web portals that do not contribute to a long-term preservation. There is also a lack of metadata attached to these media (Nixon and Troncy, 2014), which makes the search and retrieval process very difficult. According to these issues, scientific audiovisual media can be considered non-textual grey literature. Their management is a challenge affecting academia

¹ <https://av.getinfo.de/>

(Löwe et al., 2015). Often, scientific videos are uploaded to general video-sharing platforms, such as YouTube, Vimeo or Dailymotion, where videos are provided with little metadata – let alone individual segments. A manual detailed description and annotation of the content is considered too time-consuming (Snoek, 2007). Therefore, there are several initiatives that introduce automated metadata extraction for scientific and non-scientific videos. Neumann and Plank (2013) enumerate two of them: (i) ScienceCinema², the video portal of the Office of Scientific and Technical Information (OSTI) of the U.S. Department of Energy and the European Organization for Nuclear Research (CERN), where audio indexing and speech recognition is used; (ii) Voxlead New³, which presents a navigation within the video based on speech recognition.

Further examples of portals outside the library world are SciveeTV⁴, JOVE⁵ and YoVisto⁶. SciveeTV identifies scientific videos with a Digital Object Identifier (DOI) and offers the possibility of uploading supplementary materials. The portal JOVE presents a subscription-format 'video journal', in which a video is created after the submission of a manuscript. Videos are properly indexed and provided with use metrics. YoVisto presents a portal for academic video search, where the videos are segmented and automatically tagged according to the content of each segment. Innovative semantic searches can be also performed.

With the project The Open Video Digital Library⁷, the library approach to video indexing begins to change. Traditionally, librarians have applied textual bibliographic metadata to videos (Marchionini and Geisler, 2002). However, there is a need of new indexing models that can describe the characteristics of a video. The framework of The Open Video Digital Library includes tasks such as visual and linguistic analysis. This has a direct effect on the way video characteristics are indexed.

Following new ways of indexing and in order to implement innovative video access and retrieving as a library service, the Competence Centre for Non-Textual Materials at TIB teamed up with the Hasso Plattner Institute to develop the TIB|AV Portal.

2.2 The TIB|AV-Portal

The TIB|AV Portal is a web-based platform for scientific videos from the fields of engineering, architecture, chemistry, information technology, mathematics, and physics. The hosted videos mainly consist of computer visualisations, learning material, simulations, experiments, interviews, and recordings of lectures and conferences. Additional materials related to the video can also be uploaded.

The TIB|AV-Portal presents diverse innovative features. Key features are the identification of each hosted video with a DOI, which provides a stable identification for the video, and the use of different kinds of automatic video analysis. These video analyses technologies provide a wide range of metadata associated with the video as well as an audio transcription. Moreover, most videos have been published under a Creative Commons⁸ license.

Scientific audiovisual media require more metadata.

As defined by Guenther and Radebaugh (2004, p-1), metadata are 'structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource'. The information resources in the TIB|AV-Portal are video materials that are described by two types of metadata: manual and automatic.

The manual metadata, which are provided by publishers and video providers, describe the whole video. They are sub-classified in three groups: formal (e.g. title, author or publisher), technical (e.g. file size or duration) and content-related (e.g. subject, abstract/description or keywords). Due to the fact that these metadata are intellectually created, they present a high reliability during the search and retrieval process.

The automatic metadata are generated after a process containing a structural, text, speech, and image analysis. The process consists of five steps. First, the video is split into segments based on the visual content. Thus, each segment becomes an individual element that can be retrieved by the search engine. Secondly, the written language on the video is indexed using Optical Character Recognition (OCR). This allows the searches within the written language. Thirdly, speech to text

² <http://www.osti.gov/sciencecinema/>

³ <http://voxlead.labs.exalead.com/>

⁴ <http://www.scivee.tv/>

⁵ <http://www.jove.com/>

⁶ <http://www.yovisto.com/>

⁷ <http://www.open-video.org/>

⁸ <http://creativecommons.org/>

converts spoken language in the video into a speech transcript. Fourthly, moving images are detected by visual concepts and are indexed according to subject-specific and multidisciplinary visual concepts (Blümel et al, 2012). Finally, named-entity recognition indexes the video segments with semantically associated terms based on the transcripts from OCR and speech recognition (Strobel and Plank, 2014).

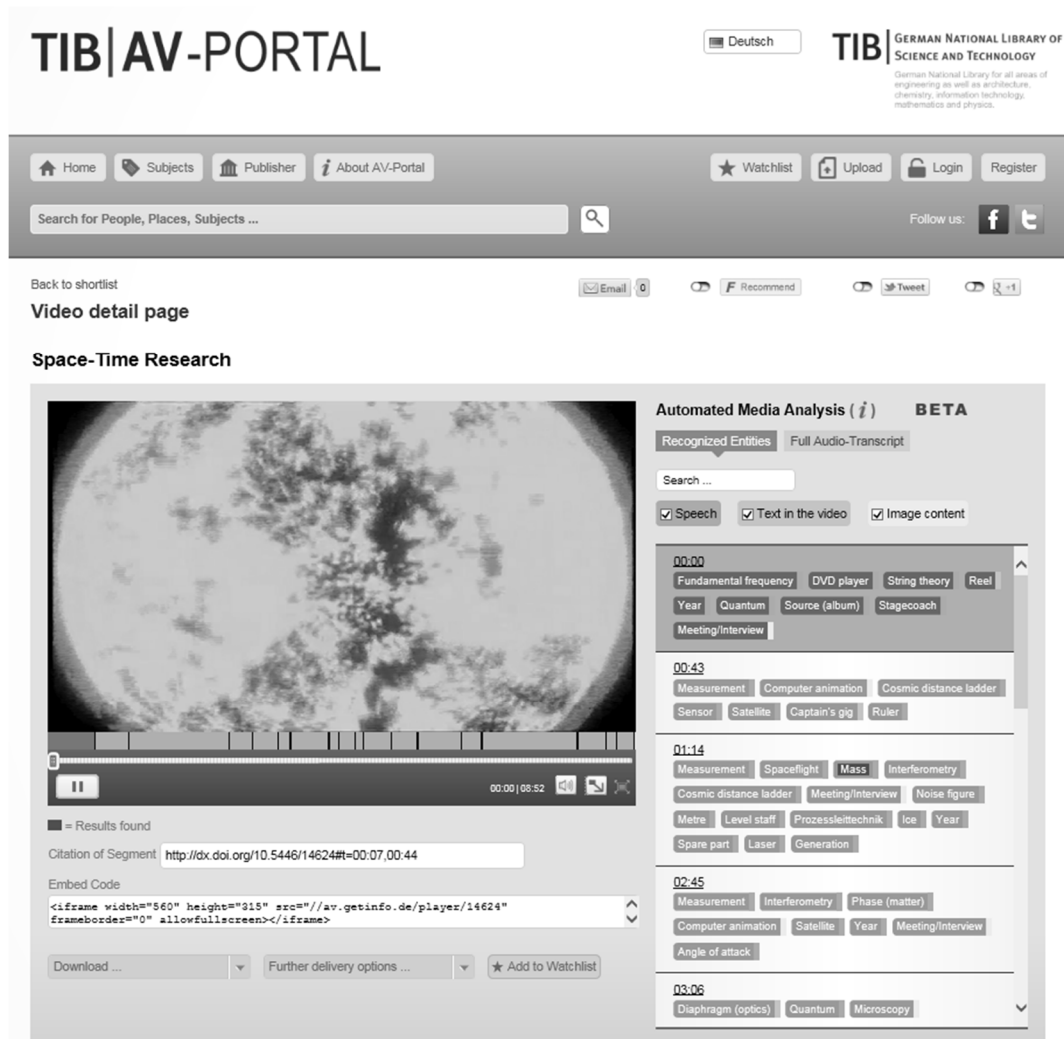


Fig. 1. Current video⁹ detail page of the TIB|AV-Portal. Automatic metadata are shown in the right side.

As a result of this process, new metadata are associated with individual video segments, allowing precise searches within the content of the video. Therefore, the searchability of the video is enhanced. Similar approaches are also carried out by publishers such as SAGE Journals or Alexander Street Press (Davies, 2015).

Searching and discovering scientific audiovisual media.

In most cases, the search is only performed within metadata that describe the whole video and not in the individual segments (Strobel and Plank, 2014). There is a need to improve these searches, hence the use of automatic generated metadata, and recognised entities to perform semantic searches in the portal as well as within the video content.

⁹ <https://av.getinfo.de/media/14624?359>

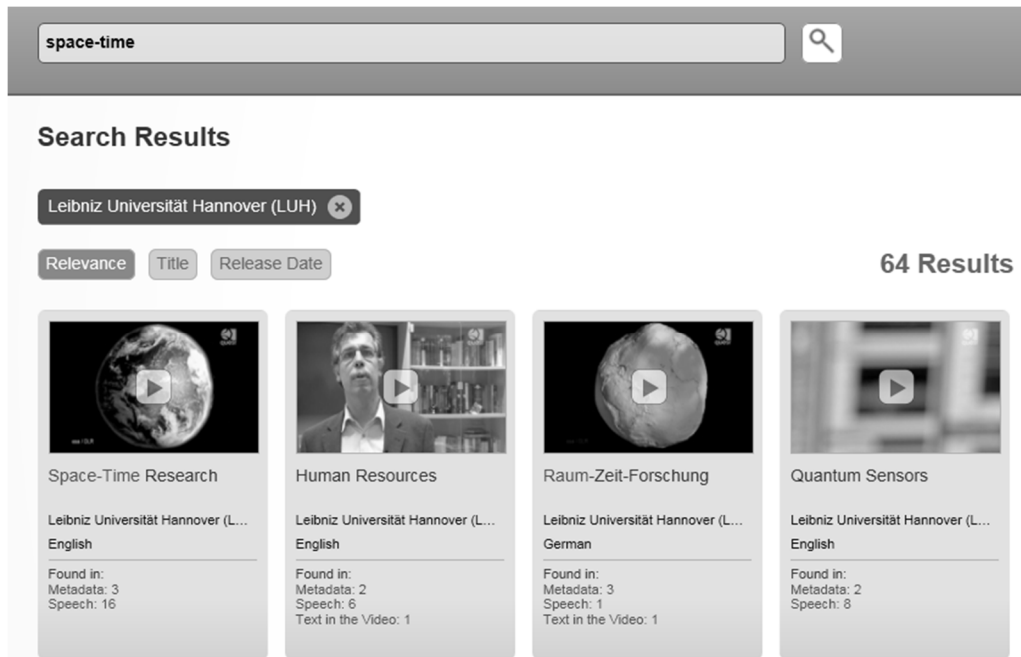


Fig. 2. Search results page of the TIB|AV-Portal. Searched term matches highlighted in red colour.

As shown in figure 3, the terms in the query can be found in the authoritative metadata like author, title and abstract as well as in the spoken text, text overlays or image information. The terms in the query are also semantically associated with other index terms (synonyms, English-German translations, etc.). This allows the users to discover new videos that otherwise would have remained hidden. For instance, retrieving a video in German when the query was in English (see figure 3).

2.3 Linking scientific content, media, and data

The upload of additional material related to the video complements the information enhancement. Causal associative metadata relationships, such as 'isSupplementBy', 'isSupplementTo', 'cites' or 'isCitedBy', allow the user to connect a video and other related materials. In order to cover the wide spectrum of these materials, different file formats (doc, .cdv, .odt, .pdf, .jpeg, etc.) are accepted in the portal.

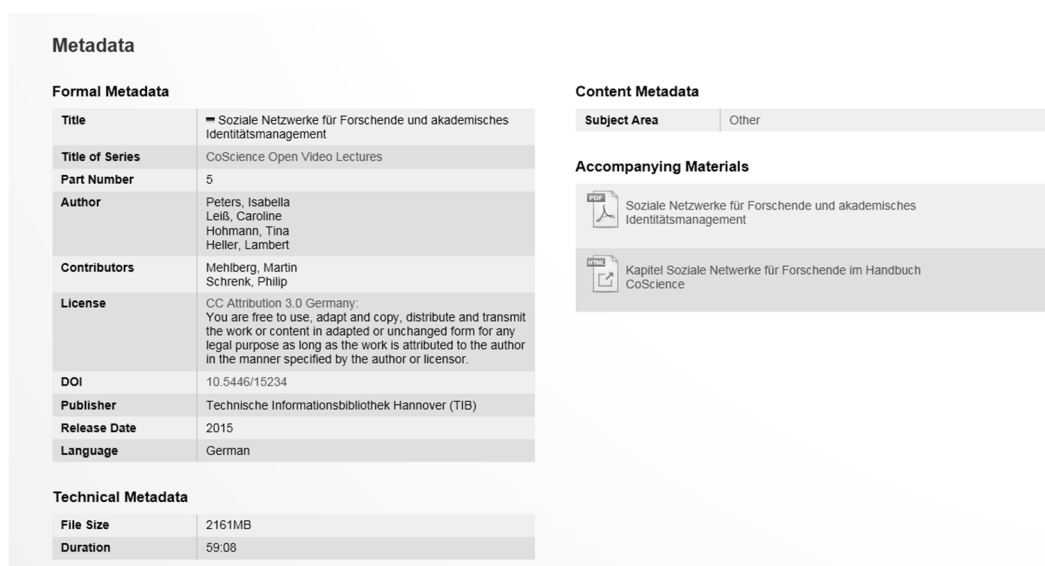


Fig. 3. Metadata section of a video¹⁰ at TIB|AV-Portal.

As accompanying materials, not only text files are available, but also links to material in other platforms. The CoScience Handbook¹¹ is shown as instance.

¹⁰ <https://av.getinfo.de/media/15234?338>

Future directions in the management of audiovisual scientific information consist of the publication of the data as Linked Open Data, following W3C Best Practices¹². This will provide a better data interoperability, and reuse.

3 Conclusions

The use of new types of scientific information, beyond the traditional journal-based system, is an ongoing process. Videos, images or research data are part of the research process and as relevant as the final paper. Academic libraries, as information managers, should provide new infrastructures that archive this information and give it an appropriate identification and term of use.

This paper focuses in particular on the treatment of scientific videos from the library perspective. The TIB|AV-Portal shows how videos can be identified and indexed with manual and automatic generated metadata. These metadata contribute to a better description of the video as a whole and at the segment level, facilitating the access and reuse.

It is also important to establish links between video, as non-textual information, and complementary sources of information, denoted in the portal as 'Accompanying materials'. This shows how the portal achieves one of the technical principles of the Pisa Declaration (2014), namely: 'encourage system for linking data and other non-textual content to their grey literature publications together with interoperability standards for sharing grey literature'.

Along this article, we have focused on the importance of new types of information in science. New ways of indexing have been shown using the TIB|AV-Portal as example, and it has been discussed why academic libraries need to provide new information services and infrastructure.

References

1. American Chemical Society. (2011, April 29). Publishing Your Research 101. Impact of video on scientific articles. [Video file]. [Retrieved: September 9, 2015] Available in: <https://www.youtube.com/watch?v=HboNzrq0MKE>
2. Blümel, I., Hentschel, Ch., H. Sack (2013). Automatic Annotation of Scientific Video Material based on Visual Concept Detection. Proceedings of i-KNOW 2013, ACM, 2013, article 16. Available in: <http://dx.doi.org/10.1145/2494188.2494213>
3. Ayuso García, M. D. and Martínez Navarro, V. (2014) La literatura gris en entornos digitales: estrategias de calidad y evaluación. In: Revista Interamericana de Bibliotecología. V. 27, num. 2 (jul.-dec. 2004), pp. 49-70.
4. Davies, N. (2015). Video Enhances Publications. Research Information. [Retrieved: September 9, 2015] Available in: http://www.researchinformation.info/features/feature.php?feature_id=499
5. Guenther, R., Radebaugh, J. (2004) Understanding Metadata. National Information Standards Organization. Bethesda, MD: NISO Press. [Retrieved: September 9, 2015]. Available in: <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>.
6. Grey Literature Network Service. Pisa Declaration on Policy Development for Grey Literature Resources. May 16, 2014. [Retrieved: September 9, 2015]. Available in: http://www.greynet.org/images/Pisa_Declaration_May_2014.pdf
7. Kraft, A., Löwe, P., Plank, M., Heller, L., Dreyer, B. (2015). Preserving the long tail in a big data world: Frameworks for e-infrastructures in research libraries. Proceedings 2015 PV Conference, Darmstadt, Germany, 3-5 November 2015. (To be published)
8. Löwe, P., Plank, M., Marín-Arraiza, P. (14th – 17th July 2015) Acquisition of audiovisual Scientific Technical Information from OSGeo by TIB Hannover: A work in progress report. In: Geomatics Workbooks n° 12, FOSS4G Europe Como (Italy).
9. Marchionini, G., Geisler, G. (2002). The Open Video Digital Library. In: D-Lib Magazine, vol. 8, n. 12. Available in: <http://www.dlib.org/dlib/december02/marchionini/12marchionini.html>
10. Neumann, J., Plank, M. (2013). TIB's Portal for audiovisual media: New ways of indexing and retrieval. In: IFLA WLIC 2013. Available in: <http://library.ifla.org/92/1/124-neumann-en.pdf>
11. Nixon, L., Troncy, R. (2014). Survey of Semantic Media Annotation Tools for the Web: Towards new Media Applications with Linked Media. In: 11th ESWC 2014, Demos Track. V.CCIS 476, pp.100-114. Available in: http://dx.doi.org/10.1007/978-3-319-11955-7_9
12. Snoek, C.G.M., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., Worring, M. (2007). Adding Semantics to Detectors for Video Retrieval. In: IEEE Trans. Multimedia, V. 9, num. 5, pp. 975-986. Available in: <http://dx.doi.org/10.1109/TMM.2007.900156>.
13. Spicer, S. (2014). Exploring Video Abstracts in Science Journals: An Overview and Case Study. In: Journal of Librarianship and Scholarly Communication V.2, num.2, eP1110. Available in: <http://dx.doi.org/10.7710/2162-3309.1110>
14. Strobel, S., Plank, M. (2014). Semantische Suche nach wissenschaftlichen Videos: Automatische Verschlagwortung durch Named Entity Recognition. In: Zeitschrift für Bibliothekswesen und Bibliographie. V.61, num. 4-5, pp. 255-259. Available in: <http://dx.doi.org/10.3196/18642950146145154>
15. Wittenberg, K. (2008). The Role of the Library in the 21st-Century Scholarly Publishing. In: No Brief Candle: Reconceiving Research Libraries for the 21st Century. Council on Library and Information Resources. [Retrieved: September 9, 2015]. Available in: <http://www.clir.org/pubs/reports/pub142/wittenberg.html>

¹¹ <http://handbuch.io/w/Hauptseite>

¹² <http://www.w3.org/TR/ld-bp/>

Situation surrounding grey literature in academic research in Japan

Yoshio Isomoto,
Hokkaido University Library, Japan

Abstract

In recent years, significant changes have been occurred to the environment surrounding the academic information distribution in Japan. In 2013, publishing of doctoral dissertations via the internet was obliged to authors. In 2015, Expert Panel on Open Science based on Global Perspectives, Cabinet Office, Government of Japan released the report named "Promoting Open Science in Japan". In this report, they recommend to promote "Open Science". Also, in 2008 a subject repository project named "Repository of Archaeological Reports" launched. To activate these movements about grey literature, it is important to support researchers in deepening their understanding about Open Science, and to cooperate with other section of university in order to build physical and technical support system for researchers.

Introduction

So far, initiatives toward publishing grey literature have not been received a lot of attention in Japan. But in recent years, significant changes have been occurred to the environment surrounding the academic information distribution in Japan. This paper introduces situation surrounding grey literature of academic research in Japan.

1. Doctoral dissertations via the internet

Before April 2013, doctoral dissertations have been published in paper, and preserved university which conferred their degree and NDL(National Diet Library). Dissertations have been located in closed stacks, and patrons have not been able to browse freely. If patrons want to get photocopies of dissertation, they required author's permission. It was too difficult to get dissertations for patrons.

Related regulations about doctoral dissertations were amended, from April 1st 2013, publishing of dissertations via the internet was obliged to authors, instead of publishing in papers as before. Authors are required to publish their dissertation via the internet within one year from the date they was awarded the degree, unless there is some particular reason. If they could not open their full-text of dissertation, they must open their abstract of dissertation. And universities are required to open abstracts and summaries of dissertation review.

Hereafter, researchers have come to publish their dissertations by institutional repositories. Hokkaido University has institutional repository named "HUSCAP"(Hokkaido University Collection of Scholarly and Academic Papers) and utilize this IR to open dissertations. And patrons (including general citizens) come to get dissertations easily.

However, not all doctoral dissertations published ever are opened all people. There is case where dissertations could not open due to copyrights, patents and other reasons. And dissertations issued before 2013 remains as paper and in closed stacks.

Sometimes, author cannot judge whether they can publish their dissertation via the internet or not, because they do not know well about copyrights, patents and other restrictions. As the efforts for dealing with these situations, Hokkaido University officials hold briefing sessions. These briefing sessions deal with important points of publishing via the internet (copyrights, patent and so on), case that their dissertations contain other academic papers, case that they submit their dissertations to academic journals, etc. Also, Hokkaido University Library set up consultation hot-line in order to advise these issues.

2. Promoting Open Science in Japan

In Japan, organized discussion about Open Science has not been almost done so far.

They're afraid Japan is left out of discussing Open Science. At March 30th 2015, report named "Promoting Open Science in Japan -Opening up a new era for the advancement of Science-" was released by Expert Panel on Open Science based on Global Perspectives, Cabinet Office, Government of Japan. In this report, they recommend to promote "Open Science". This report consists of 3 chapters "The Importance of Open Science", "The Need to Promote Open Science, based on Global", "Response to the Global Trends in Open Science".

This report emphasizes the importance and the necessity of Open Science, and mentions the importance of gazing international movements. Also, this report points out the importance of

ensuring the quality and transparency of research results as well as importance of building a system of reusing of the research result, articles and based data. And they mentions disadvantages of failing to keep up with world tendency of “Open Science”. Also, they point out effectiveness of “Open Science” to preventing research misconduct. They say the core principle of promoting “Opne Sciecce” is enhancing the utility of research results (including research data) supported by public research funds.

In response to this report, Ministry of Education, Culture, Sports, Science and Technology (MEXT), Science Council of Japan, Ministry of Foreign Affairs of Japan, and other related organizations have prompted discussions about these topics. JSPS(Japan Society for the Promotion of science) announced a policy to promote Open Access to researchers supported by their fund “Grants-in-Aid for Scientific Research”(KAKENHI). They distribute brochure explaining importance of Open Access. These movements have just begun. The concept of Open Science is not penetrated yet in researchers.

And MEXT announced “Enforcement policy of KAKENHI reformation ” on its website, and mentioned Open Science in this policy at September 2015. This policy contains the proposal to ensure advancing the visualization of research results and cooperating with other public research fund.

As part of effort to these movements, Hokkaido University’s Institutional repository “HUSCAP” has begun to include supplemental sound data of articles. Some researchers consult library staff about preserving their huge experimental, observational data. At the present time, Hokkaido University has no proper methods to respond these requests. This is a subject for future analysis.

3. Repository of Archaeological Reports

In 2008 a subject repository named “Repository of Archaeological Reports” opened as a result of “Repository of Archaeological Reports Initiative”. This project launched by 5 national university corporations centered upon Shimane University. This repository has gathered and opened archaeological reports in Japan. This project was not influenced by governmental tendency. This is a great example of original efforts by universities.

This repository gathered about 14,000 archaeological reports from all over Japan. Now, it was integrated into “Comprehensive Database of Archaeological Site Reports in Japan” operated by “Nara National Research Institute for Cultural Properties” in 2015.

4. To activate these movements (my opinion)

In order to activate these movements about grey literature, there are two important missions played by university officials.

One is supporting researchers in deepening their understanding about Open Science. The other is cooperating with other section of university in order to build physical and technical support system for researchers.

In Japan, these movements have just begun. And these are too huge and complicated, to deal with by single department of university (like Library). So, cooperation with researchers and other university officials is essential.

References

- [1] Satoshi Nakayama, Takayuki Manaka (2013) Makoto Shuto, The Possibility of Networked Electronic Theses in Japan - before & after April, 2013-.
<http://lib.hku.hk/etd2013/presentation/Nakayama%20-%20The%20Possibility.pdf> (accessed 2015-11-10)
- [2] Kazuhiro Hayashi (2015). Recent Trends of Open Access and Open Science Policy and the Role of Library, Current awareness, (324), 15-18. (in Japanese) <http://current.ndl.go.jp/ca1851> (accessed 2015-11-10)
- [3] The Expert Panel on Open Science, based on Global Perspectives Cabinet office, Government of Japan (2015). Promoting Open Science in Japan -Opening up a new era for the advancement of Science-. (in Japanese / English)
<http://www8.cao.go.jp/cstp/sonota/openscience/index.html> (accessed 2015-11-10)
- [4] Council for Science and Technology, Ministry of Education, Culture, Sports, Science and Technology (MEXT) (2015). Documents related to openness of academic information. (in Japanese)
http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu4/036/shiryo/_icsFiles/afieldfile/2015/06/30/1359347_03.pdf (accessed 2015-11-10)
- [5] Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP
<http://eprints.lib.hokudai.ac.jp> (accessed 2015-11-10)
- [6] Comprehensive Database of Archaeological Site Reports in Japan
<http://sitereports.nabunken.go.jp/> (accessed 2015-11-10)

Copyright Reform and the Library and Patron Use of Non-text or Mixed-Text Grey Literature: A Comparative Analysis of Approaches and Opportunities for Change

Tomas A. Lipinski and Katie Chamberlain Kritikos,

School of Information Studies; University of Wisconsin Milwaukee, United States

Abstract

Problem/Goal: What is the current state of use rights in the copyright law specific to libraries and related institutions regarding the use of non-text or mixed text grey literature? Are those exceptions sufficient? If not, what recommendations for change can be made?

Research Method/Procedures: This paper assesses the adequacy of existing use rights in the copyright law specific to libraries and related entities applicable to the collection, preservation, access, dissemination and use of grey literature in non-text or mixed text formats. The study is timely, the World Intellectual Property Organization Standing Committee on Copyright and Related Rights is reviewing world copyright laws and considering drafting an international protocol addressing use rights for libraries and archives. While previous studies have focused on copyright and grey literature or on copyright reform, an analysis that combines these lenses (grey literatures, non-text or mixed text formats and copyright review and reform) has not been undertaken. This analysis presents a content review of the copyright use provisions (known as “exceptions and limitations”) from the copyright law of those countries represented by presentation participants (excluding poster participants) at GL-12 through GL-16. The specific focus is on a particular country’s section of the copyright law that is dedicated, where extant, to libraries, archives, and related entities. The impact of the law on the collection, preservation, access, dissemination, and use of non-text works, such as images, sound recordings, audio visual works, etc., or on mixed-text works (*i.e.*, multimedia) in grey literature collections of libraries and archives is discussed. The following factors, among others, will be applied in the analysis: qualification, preservation, replacement, reproduction, distribution, inter-library loan (including cross-border sharing), and digitization. The review will evaluate the shortcomings of the existing copyright law. Recommendations for change are offered to grey and policy advocates at the national and international legislative and policy-making venues to raise awareness of the shortcoming of existing copyright laws and offer direction for positive change regarding use of grey collections. Such change would be consistent the Articles 8 and 9 of the *Pisa Declaration on Policy Development for Grey Literature* relating to collection, access and use of grey literature.

Results: The results will demonstrate the current copyright of many countries in the GL conference community are inadequate when applied to non-text or mixed-text sources. Recommendations indicate opportunities for change that grey literature proponents can use to influence policy makers, effect positive change, and ensure the future retention, access, and use of grey collections.

I. Introduction

This paper assesses the adequacy of existing use rights in the copyright law specific to libraries and related entities and applicable to the collection, preservation, access, dissemination, and use of grey literature in non-text or mixed text formats. The study is timely, as the World Intellectual Property Organization Standing Committee on Copyright and Related Rights is reviewing world copyright laws and considering drafting an international instrument addressing use rights for libraries and archives. Use rights such as these are referred to as “exceptions” and “limitations” on the exclusive rights of copyright holders.

Throughout this paper, the word “exception” is used to capture the concept of “exceptions and limitations,” and the word “library” includes archives, museums, and other institutions as the case may be (for example, the copyright laws of some countries include museums and galleries). The copyright laws of many countries contain an exception that secures reproduction, dissemination, and other use rights for qualifying libraries. The dissemination right allows the further distribution of protected works, but seldom includes the rights of performance and transmission (*i.e.*, display). Often, rights of preservation and replacement are also included.

The review of world-wide copyright exceptions for libraries, however, revealed that some countries do not have any exceptions or that the existing provisions are inadequate. Professor

Kenneth D. Crews observed that in 2008, twenty-one out of 149 countries surveyed had no copyright provision dedicated to library use rights.¹ The 2014 Crews Report showed that thirty-two out of 186 countries surveyed had no copyright provision for libraries.² There are many reasons for such discrepancy or variation: “Determining factors included legal and cultural traditions, social-economic factors, national IP strategy and policies.”³ This paper reviews a select number of copyright exceptions for libraries from various countries. Part II presents the problem analyzed herein and Part III describes our research methodology. Part IV provides an analysis of the data and Part V offers a variety of observations, assessments, and recommendations based thereon. Part VI concludes that the current copyright laws of many countries in the GL conference community are inadequate when applied to non-text or mixed-text works. Recommendations for change are forwarded to grey literature and policy advocates at national and international legislative to offer direction for positive change regarding use of grey collections.

II. Problem Statement

The authors inquire: What are the characteristics of the *use rights* (known as “exceptions” and “limitations”) in the copyright law specific to libraries and related institutions regarding the use of non-text or mixed text grey literature? Are those *exceptions and limitations* sufficient? If not, what recommendations for change can be made?

The three basic elements of an exception and limitation provision are *preservation*, *replacement*, and *dissemination*. Such a provision “enables memory institutions to reproduce copyrighted works as part of their institutional responsibility in collecting and preserving their collections so that they are permanently accessible . . . it allows material to be archived in any format, and seeks to maintain the integrity of the public record for online resources, the medium of today, that disappear for different reasons everyday [sic].”⁴ *Dissemination* can include the reproduction of materials for a patron or the delivery of materials to the patron of another library (e.g., interlibrary loan). While a number of countries have secured basic rights of *preservation*, including replacement and dissemination, far less have provisions for document supply or interlibrary loan (28 countries according to the 2014 Crews Report). It is not clear, however, that the loan provisions allow for *cross-border transfers*, i.e., the sharing of resources beyond the borders of the country where either the requesting patron/library or the fulfilling library is situated. While previous studies have focused specifically on copyright and grey literature⁵ or generally on how changes in the copyright law affect libraries,⁶ an analysis that combines these lenses (non-text or mixed text grey literature with copyright review and reform) has not been undertaken.

In order to focus existing copyright provisions and the need for reform, the authors examined communities of users and/or scholars of grey literature. The analysis presents a content review of the copyright use provisions from the copyright laws of those countries represented by

¹ See Kenneth D. Crews, *Study on Copyright Limitations and Exceptions for Libraries and Archives*, SCCR/17/2 (Aug. 26, 2008), available at http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=109192 (last visited Nov. 2, 2015).

² See Kenneth D. Crews, *Study on Copyright Limitations and Exceptions for Libraries and Archives*, SCCR/29/3 (Nov. 5, 2014), available at http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=290457 (last visited Nov. 2, 2015) [hereinafter “2014 Crews Report”]. The 2008 and 2014 reports were consolidated in Kenneth D. Crews, *Study on Copyright Limitations and Exceptions for Libraries and Archives: Updated and Revised*, SCCR/30/3 (Jun. 14, 2015), available at http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=306216 [hereinafter “2015 Crews Report”].

³ Electronic Information for Libraries, *Draft Law on Copyright: Including Novel Exceptions and Limitations for Libraries and Their Users* (2014, Work in Progress), at 45, available at http://www.eifl.net/system/files/resources/201411/eifl_draft_law_2014.pdf (last visited Nov. 2, 2015) [hereinafter “EIFL Draft Law on Copyright”].

⁴ EIFL Draft Law on Copyright at 31.

⁵ See, e.g., Tomas A. Lipinski, *Green Light for Grey Literature? Orphan Works, Web-Archiving and Other Digitization Initiatives – Recent Developments in U.S. Copyright Law and Policy*, GREY J.: INTERNAT’L J. GREY LIT. (Vol. 5, Iss. 1, 2009), available at <http://www.moyak.com/papers/grey-literature-digitization.pdf> (last visited Nov. 2, 2015); Joachim Schöpfel and Tomas A. Lipinski, *Legal Aspects of Grey Literature*, GREY J.: INTERNAT’L J. GREY LIT. (Vol. 8, Iss. 3, 2012), available at https://hal.inria.fr/file/index/docid/905090/filename/Schopfel_and_Lipinski_Legal_Aspects_of_Grey_Literature_.pdf (last visited Nov. 2, 2015).

⁶ See, e.g., Tomas A. Lipinski, *The Climate of Distance Education in the 21st Century: Understanding and Surviving the Changes Brought by the TEACH (Technology, Education, and Copyright Harmonization) Act of 2002*, 29 J. ACAD. LIB’SHIP 362 (2003); Katherine A. Chamberlain, “Lawfully Made under This Title”: *The Implications of Costco v. Omega and the First Sale Doctrine on Library Lending*, 37 J. ACAD. LIB’SHIP 291 (2011), available at <http://www.sciencedirect.com/science/article/pii/S009913331100070X> (last visited Nov. 2, 2015).

presentation participants (excluding poster participants) at GL-12 through GL-16. The specific focus is on a particular country's section of the copyright law that is dedicated, where extant, to libraries. The impact of the law on the collection, preservation, access, dissemination, and use of *non-text works*, such as images, sound recordings, audio visual works, etc., or on *mixed text works* (i.e., multi-media) in the grey literature collections of libraries and archives is discussed. The following factors, among others, will be applied in the analysis: qualification (of the entity), preservation, replacement, reproduction, distribution, inter-library loan (including cross-border sharing), and digitization. The review will evaluate the shortcomings of the existing copyright laws. The following section describes our research methodology.

III. Research Methodology

Sampling. For this paper, the selected sample of copyright laws represented the countries from the five previous Grey Literature conferences (GL-12 through GL-16). This sample initially consisted of twenty-three countries. If a country from the initial sample did not have a copyright law or its law was not available in English, that country was excluded from the final sample. If a country from the initial sample had a copyright law and its law was also available in English, that country was included in the final sample. Because this paper explores and analyzes mixed-text or non-text grey literature, we wanted not to compare the copyright laws of a random sample group of countries, but rather a self-selected sample of countries that are active in the grey literature community by their paper submissions at Grey Literature conferences.

In addition, two countries were added to the list: South Africa and the United Kingdom. When discussing possible copyright reform, these countries offer a good basis of comparison. South Africa's pending legislation represents the most sweeping copyright reform in terms of securing a wide range of use rights for libraries.⁷ The United Kingdom also recently amended its copyright law relating to library and archive rights, including the so-called "Teflon" protections that prohibit contractual override of library use rights. The rights of *reproduction* for inter-library loan ("ILL"), which include "whole or part of a published work" and "one article . . . or reasonable proportion," and *replacement* cannot be overridden by a contract like a database subscription agreement or license: "To the extent that a term of a contract purports to prevent or restrict the doing of any act which, by virtue of this section, would not infringe copyright, that term is unenforceable."⁸

A second reason that the United Kingdom was included was that it ranked second behind the United States in a recent study measuring the correlation between library and other limitations and exceptions for users and a variety of economic indicators relevant to the knowledge economy.⁹ The study found that a positive correlation existed between a flexible system of copyright exceptions and "higher rates of growth in value-added output throughout their economy."¹⁰ The more robust the range of limitations and exceptions that existed, the more robust the knowledge economy and the production of information.¹¹ The United States received a top index score of 8.13, with the United Kingdom close behind with a score of 7.19.¹² Hence, after the addition of South Africa and the United Kingdom, our sample resulted in analyzing copyright provisions from eighteen different countries.

Data Collection. Obtaining the text of the copyright laws from the eighteen countries in the final sample was a two-step process. First, the citation for each provision was found in *Study on Copyright Limitations and Exceptions for Libraries and Archives: Updated and Revised* ("2015 Crews Report") prepared by Kenneth D. Crews on behalf of the World Intellectual Property

⁷ See South Africa Copyright Amendment Bill, No. 646 (July 27, 2015), available at <http://blogs.sun.ac.za/iplaw/files/2015/08/gg39028.pdf> (last visited Nov. 2, 2015).

⁸ See §§ 41(5) ("Copying by librarians: supply of single copies to other libraries"), 42(7) ("Copying by librarians, etc.: replacement copies of works"), and 42A(6) ("Copying by librarians: single copies of published works"), Copyright, Designs, and Patents Act of the United Kingdom, Chapter 48 (Nov. 15, 1988), as amended through Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, Statutory Instrument 2014 No. 1372 (May 19, 2014), available at <http://www.legislation.gov.uk/ukpga/1988/48/part/I/chapter/III/crossheading/libraries-and-archives> (last visited Nov. 2, 2015).

⁹ See Benjamin Gibert, *The 2015 Intellectual Property and Economic Growth Index: Measuring the Impact of Exceptions and Limitation in Copyright Growth, Jobs and Prosperity*, Lisbon Council (rev'd May 2015), available at <http://www.innovationeconomics.net/component/attachments/attachments.html?id=263&task=view> (last visited Nov. 2, 2015) [hereinafter "2015 Growth Index"].

¹⁰ 2015 Growth Index at 3.

¹¹ See 2015 Growth Index at 2-4.

¹² 2015 Growth Index at 5.

Organization (“WIPO”).¹³ With the citation to the specific library provision in hand, the WIPO Lex database was used to find the full text of the sample countries’ copyright laws. Conversion from downloadable PDFs to searchable Word documents facilitated later analysis. If a country’s copyright law was not available via WIPO Lex, the citation referenced in the 2015 Crews Report and government websites to find the laws online in either website, PDF, or Word document form. This information was catalogued in an Excel spreadsheet containing categories for country, source of copyright law, link to law, section for exceptions, and notes. After securing copies of the copyright laws, we pinpointed the sections pertaining to copyright exceptions and limitations for libraries and archives and created a separate, searchable Word document of just the text of the exceptions and limitations. The analysis was limited by the necessity of using translated versions of certain legal texts. This resulted in often peculiar phrasing that required assumptions to be made regarding the meaning or intent of a particular word or phrase. The following section addresses our methods of document analysis.

IV. Analysis

Categories of Use Rights. After collecting and organizing the data, we analyzed the eighteen different copyright provisions and placed their use rights into one of three categories: rudimentary, basic, and elaborate. *Rudimentary* use rights do not contain all three basic elements of use rights (preservation, replacement, and dissemination) or it focuses on some other singular aspect of use rights. Several countries fall into this category: Cameroon (official archives only); Greece (replacement and ILL); Norway (preservation and for patrons); and Slovenia (reproduction rights for a broad range of entities including museums and education or scientific establishments but limited to “internal use”). Countries with *basic* use rights and reproduction for patrons including ILL include Algeria, Czech Republic, Iceland, India, Japan, Poland, Slovakia, and South Africa. Countries with *elaborate* use rights, in terms of the length or number of sections of text or the range of rights, are Australia, Canada, Russia, South Africa (which mirrors some of the U.S. copyright law), the United Kingdom, and the United States. Elaborate countries also, at a minimum, provide rights of preservation, replacement, and dissemination whereas *basic* countries supply those rights at maximum. As a summary of our analysis, see Appendix A, a chart of the exceptions and limitations with categories for country, library’s rights under the copyright law, application to text and non-text works, and notes.

Word Cloud of Common Terms. In addition, Figure 1, a *word cloud*, reveals commonalities among the texts of the sample countries’ copyright exceptions and limitations for libraries and archives. To create the word cloud, we copied and pasted the Word document of just the text of the exceptions and limitations sections of copyright laws pertaining to libraries and archives into Wordle, an open access word-cloud generator.

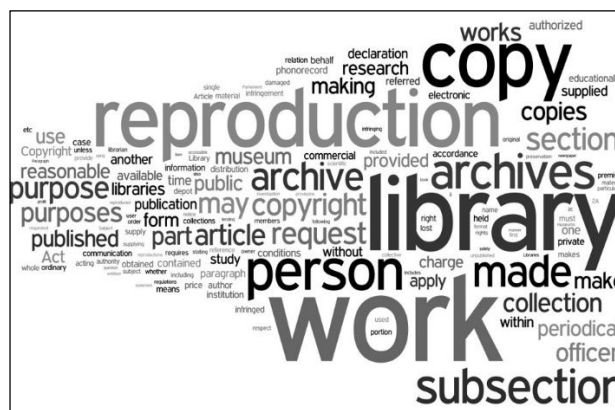


Figure 1. Word cloud of common terms in use rights for libraries and archives.

The largest words are those that appear most often; here, the most frequently recurring words are “library,” “work,” “reproduction,” “copy,” “person,” “archive” and “archives,” and “made.” Such repetition of copyright-relevant words shows the common themes among the current exceptions for libraries and archives in the sample countries’ copyright laws, such as the focus on reproduction and copying in libraries and archives. While it may be easy to forget about the “person,” this word cloud reinforces the human element that connects libraries and archives,

¹³ See 2015 Crews Report.

patrons, and laws, a connection not to be neglected when reexamining and refining copyright laws treatment of mixed-text or non-text grey literature.

Digitization and Reproduction of Works. Several other observations can be made. A number of countries allow the *digitization* of content for the purposes of preservation or replacement, but restrict *access* to the work in some fashion. The most common restriction is limiting access to in-house or on-premises use. An example is Australia's rules for copying for users:

If an article contained in a periodical publication, or a published work (other than an article contained in a periodical publication) is acquired, in electronic form, as part of a library or archives collection, the officer in charge of the library or archives may make it available online within the premises of the library or archives in such a manner that users cannot, by using any equipment supplied by the library or archives: (a) make an electronic reproduction of the article or work; or (b) communicate the article or work.¹⁴

Additionally, Australia provides for copying for preservation "to officers of the library or archives by making it available online to be accessed through the use of a computer terminal installed within the premises of the library or archives with the approval of the body administering the library or archives."¹⁵

The Czech Republic, Norway, and the United Kingdom all provide other examples of restricting access to in-house or on-premises use.¹⁶ Poland appears to make a similar restriction: "to make the collection available for research or learning purposes through information technology system terminals (endings) located at the premises of those entities."¹⁷ Russia does as well, indicating that "copies of works in electronic form may be solely provided for temporary gratuitous use at the premises of a library or archive, provided that it is impossible to create any more copies of the work in electronic form."¹⁸ Slovakia, while allowing a broad range of activities including many uses of grey literature ("educational purposes or scientific and research purposes"),¹⁹ also requires access be made "exclusively on the premises of the library or the archive."²⁰ Slovenia's law is less explicit, but could be interpreted to impose a similar restriction (*i.e.*, "works from their own copies for internal use").²¹ Even the United States contains such a restriction in both its preservation and replacement provisions: "[A]ny such copy or phonorecord that is reproduced in digital format is not otherwise distributed in that format and is not made available to the public in that format outside the premises of the library or archives."²²

The impact on the use of non-text or mixed-text grey literature is clear: While a non-text work could be found in some *analog format* such as tape (audio or video), conversion of the content into a *digital format* triggers the in-house or on-premises restriction of several countries' copyright provisions. Moreover, because more and more works are born "digital," any

¹⁴ § 49(5A), "Reproducing and communicating works by libraries and archives for users," Copyright Law of Australia, No. 63 (Jun. 27, 1968), *as amended* through No. 31 (May 27, 2014), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=336977 (last visited Nov. 2, 2015).

¹⁵ § 51A(3), "Reproducing and communicating works for preservation and other purposes," Copyright Law of Australia, *supra* note 14.

¹⁶ *See* § 37(1)(c), "Library License," Act on Copyright and Rights Related to Copyright of the Czech Republic, No. 121/2000 (Apr. 7, 2000), *as amended* through No. 216/2006 (May 22, 2006), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=137175 (last visited Nov. 2, 2015); §16, "Making copies in archives, libraries and museums, etc.," Act Relating to Copyright in Literary, Scientific, and Artistic Works etc. of Norway, No. 2 (12 May 1961), *as amended* through 22 Dec. 2006, *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=248181 (last visited Nov. 2, 2015); § 40B, "Libraries and educational establishments etc: making works available through dedicated terminals," Copyright, Designs, and Patents Act of the United Kingdom, *supra* note 8.

¹⁷ § 28(3), Copyright and Related Rights Act of Poland, No. 83 (Feb. 4, 1994), *as amended* through Alteration of the Law on Copyright and Neighboring Rights, No. 91 (Oct. 10, 2010), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=129378 (last visited Nov. 2, 2015).

¹⁸ Art. 1275(1), "Free Use of Works by Libraries, Archives and Educational Organisations," Civil Code of the Russian Federation, No 230-FZ (Dec. 18, 2006), *as amended* through Amendment No. 35-FZ (Mar. 12, 2014), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=345444 (last visited Nov. 2, 2015).

¹⁹ § 31(1)(a), "Use of work by a library or an archive," Law of Copyright and Related Rights of Slovakia, No. 618/2003 (Dec. 4, 2003), *as amended* through No. 453/2008 (2008), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=189474 (last visited Nov. 2, 2015).

²⁰ § 31(1)(a), "Use of work by a library or an archive," Law of Copyright and Related Rights of Slovakia, *supra* note 19.

²¹ Art. 50, Copyright and Related Rights Act of Slovenia, No. 21 (1995), *as amended* through No. 16 (Dec. 15, 2006), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=180840 (last visited Nov. 2, 2015).

²² 17 U.S.C. § 108(b)(2), "Limitations on exclusive rights: Reproduction by libraries and archives," Copyright Act of the United States, Pub. L. No. 94-553 (Oct. 23, 1976), *as amended* through Pub. L. No. 111-295 (Dec. 9, 2010), *available at* <http://www.wipo.int/edocs/lexdocs/laws/en/us/us352en.pdf> (last visited Nov. 2, 2015).

reproduction thereof would naturally be digital as well as a library would not likely shift formats in the opposite direction, *i.e.*, from digital back to analog.

Digitization of Non-Text Works: Reproduction and Inter-Library Loan. Other copyright provisions address the *non-text nature of works* such as musical works, sound recordings, images, and audio-visual works. Sometimes this treatment of non-text works is by default if the particular provision focuses on text-based works. One example is the Algerian law that uses words or phrases suggesting literary works alone are covered; library rights of reproduction are limited to an “article” or “written work . . . published in a collection of works, newspaper volumes or periodicals.”²³ The text of the Australian statute has a similar focus on the article or work or a part thereof.²⁴ Yet another provision uses the broad phrasing “of a work” with rights of reproduction anticipating the digital more likely with new non-text works: “[T]he work from which the reproduction is made is in electronic form.”²⁵ Canada allows the use of digital format for the delivery of inter-library loan (“ILL”), but does not allow retention in digital format.²⁶ Some countries specifically exclude non-text formats whether digital or not. For example, the Czech Republic specifically excludes audio or audio-visual works from ILL.²⁷ Likewise, the United States has similar provision in § 108(i) that applies to ILL or copying for patrons:

The rights of reproduction and distribution under this section *do not apply to a musical work, a pictorial, graphic or sculptural work, or a motion picture or other audiovisual work* other than an audiovisual work dealing with news, except that no such limitation shall apply with respect to rights granted by subsections (b), (c), and (h), or with respect to pictorial or graphic works published as illustrations, diagrams, or similar adjuncts to works of which copies are reproduced or distributed in accordance with subsections (d) and (e).²⁸

India suggests a conversion from analog to digital, but not the right to preserve born-digital works, by allowing the storage of a work in digital format “for preservation if the library already possesses a non-digital copy of the work.”²⁹ The National Diet Library of Japan can “record on a memory a work included in its library materials, to the extent deemed necessary, in the case where an electro-magnetic record . . . for the purpose of avoiding the destruction, the damage, of the stain of such original by the public”³⁰ This suggests that in Japan, preservation copies can be made in digital format and include images.

Other Memory Institutions. As discussed initially, the use of the word “library” or “libraries” includes an archive or archives, but several countries include other memory or knowledge institutions. A common addition is *museums*, as in Canada, Czech Republic, Norway, Slovenia (“Publicly accessible archives and libraries, museums and educational or scientific establishments”³¹), and the United Kingdom. The Czech Republic includes *galleries* as well as the United Kingdom by definition (“‘Museum’ includes a gallery”³²). A number of countries consider *educational institutions*: Slovenia, Canada (in some provisions), the Czech Republic (“school, university, and other non-profit school-related and educational establishment”³³), Norway, Poland, and Russia.

Reproduction Limited to Certain Purposes. A number of countries limit reproduction for users to certain purposes. Most include a theme of *research* or *private use*: Algeria (“educational, academic research, or personal purposes”)³⁴; Australia (“research or study”)³⁵; Canada (“research

²³ Art. 45, Copyrights and Related Rights Act of Algeria, No. 03-05 (July 19, 2003), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=178342 (last visited Nov. 2, 2015).

²⁴ See, *inter alia*, §§ 50(7), (7A), (7B), Copyright Law of Australia, *supra* note 14.

²⁵ § 50(7B)(b), “Reproducing and communicating works by libraries or archives for other libraries or archives,” Copyright Law of Australia, *supra* note 14.

²⁶ See § 30.2(5.02), Copyright Act of Canada, c. C-42 (1985), *as amended* through Jan. 2, 2015, *consolidated* as of Mar. 31, 2015, *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=366684 (last visited Nov. 2, 2015).

²⁷ § 37(3), “Library License,” Act on Copyright and Rights Related to Copyright of the Czech Republic, *supra* note 16.

²⁸ 17 U.S.C. § 108(i), “Limitations on exclusive rights: Reproduction by libraries and archives,” Copyright Act of the United States, *supra* note 22.

²⁹ § 31(iii), Copyright (Amendment) Act of India, No. 27 (Jun. 7, 2012), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=304385 (last visited Nov. 2, 2015).

³⁰ Art. 31(1), “Reproduction, etc. in libraries, etc.,” Copyright Act of Japan, Act No. 48 (May 6, 1970), *as amended* through Act No. 43 (Jun. 27, 2012), *available at* http://www.cric.or.jp/english/clj/doc/20130819_July,2013_Copyright_Law_of_Japan.pdf (last visited Nov. 2, 2015).

³¹ Art. 50, Copyright and Related Rights Act of Slovenia, *supra* note 21.

³² § 43A(3), “Sections 40A to 43: interpretation,” Copyright, Designs, and Patents Act of the United Kingdom, *supra* note 8.

³³ Art. 37(1), “Library License,” Act on Copyright and Rights Related to Copyright of the Czech Republic, *supra* note 16.

³⁴ Art. 45, Copyrights and Related Rights Act of Algeria, *supra* note 23.

or private study”³⁶, India (“research or private study”³⁷; Japan (“own research study”³⁸; Norway (“research or private study”³⁹; Slovakia (“educational purposes or scientific and research purposes exclusively on the premises of the library or the archive”⁴⁰; South Africa (“private study or the personal or private use”⁴¹; and the United Kingdom (“non-commercial purpose or private study”⁴²).

Commercial Availability. Finally, some reproduction rights are predicated on the absence of *commercial availability*, or require some sort of market search for a suitable replacement. Because most grey literature is not commercially available, this condition does not likely impact a library with grey literature in its collections or its patrons. For example, several provisions in the Australia statute condition reproduction rights on the inability for the library to obtain the item “within a reasonable time at an ordinary commercial price.”⁴³ Greece also has the double criteria of availability and price: “The reproduction shall be permissible only if an additional copy cannot be obtained in the market promptly, and on reasonable terms.”⁴⁴ Iceland, however, conditions the reproduction rights upon an availability standards alone (the work is unobtainable or “unavailable in the open market and from the publisher”⁴⁵). The United States conditions its replacement right on the determination “that an unused replacement cannot be obtained at a fair price,”⁴⁶ and the right to copy a substantial or entire works for patrons or ILL upon whether “a copy or phonorecord of the copyrighted work cannot be obtained at a fair price.”⁴⁷ As the South African statute is based on the U.S. statute, its replacement, copying, and ILL provisions contain identical phrasing.⁴⁸

V. Results

SCCR Criteria. The WIPO Standing Committee on Copyright and Related Rights (“SCCR”)⁴⁹ has identified eleven criteria (“SCCR Criteria”) to protect a library’s rights under a copyright law provision: (1) preservation, including replacement; (2) copying for patrons (*i.e.*, document delivery); (3) legal deposit; (4) lending or ILL; (5) parallel importation; (6) cross border; (7) orphan works; (8) limitation on liability; (9) Technological Protection Measures (“TPMs”); (10) contractual override; and (11) translation rights.⁵⁰

Applying the SCCR Criteria to grey literature, legal deposit and translation rights may be eliminated because legal deposit relates to published works and, as noted above, most grey literature (at least in its initial creation) is *unpublished*. Translation rights for grey literature may

³⁵ § 49(1)(b)(i), “Reproducing and communicating works by libraries and archives for users,” Copyright Law of Australia, *supra* note 14.

³⁶ § 30.2(2), “Libraries, Archives and Museums,” Copyright Act of Canada, *supra* note 26.

³⁷ § 52(1)(p), Copyright Act of India, No. 14 (4 Jun. 1957), *as amended* through Act No. 49 (30 Dec. 1999), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=128098 (last visited Nov. 2, 2015).

³⁸ Art. 31(1), “Reproduction, etc. in libraries, etc.,” Copyright Act of Japan, *supra* note 30.

³⁹ § 16, “Making copies in archives, libraries and museums, etc.,” Act Relating to Copyright in Literary, Scientific, and Artistic Works etc. of Norway, *supra* note 16.

⁴⁰ § 31(1)(a), “Use of work by a library or an archive,” Law of Copyright and Related Rights of Slovakia, *supra* note 19.

⁴¹ Copyright Regulations of South Africa (1978), *as amended* through GN 1375 (1985), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=130435 (last visited Nov. 2, 2015).

⁴² § 32A(3)(c), “Copying by librarians: single copies of published works,” Copyright, Designs, and Patents Act of the United Kingdom, *supra* note 8.

⁴³ § 49(5)(b), “Reproducing and communicating works by libraries and archives for users,” Copyright Law of Australia, *supra* note 14.

⁴⁴ “Reproduction of Libraries and Archives,” Law of Copyright, Related Rights, and Cultural Matters of Greece, No. 2121 (Mar. 4, 1993), *as amended* through No. 4281 (2014), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=367777 (last visited Nov. 2, 2015).

⁴⁵ Art. 12, Copyright Act of Iceland, No. 73 (29 May 1972), *as amended* through No. 93 (21 Apr. 2010), *available at* http://www.wipo.int/wipolex/en/text.jsp?file_id=332081 (last visited Nov. 2, 2015).

⁴⁶ 17 U.S.C. § 108(c)(1), “Limitations on exclusive rights: Reproduction by libraries and archives,” Copyright Act of the United States, *supra* note 22.

⁴⁷ 17 U.S.C. § 108(d), “Limitations on exclusive rights: Reproduction by libraries and archives,” Copyright Act of the United States, *supra* note 22.

⁴⁸ *See* Copyright Regulations of South Africa, *supra* note 41.

⁴⁹ World Intellectual Property Organization Standing Committee on Copyright and Related Rights, *available at* <http://www.wipo.int/policy/en/sccr/> (last visited Nov. 2, 2015).

⁵⁰ World Intellectual Property Organization Standing Committee on Copyright and Related Rights, *Consolidation of Proposed Texts Contained in Document SCCR/26/3*, prepared by African Group, Brazil, Ecuador, India, and Uruguay, Twenty-ninth Session, Geneva, Switzerland (Dec. 8-12, 2014), *available at* http://www.wipo.int/edocs/mdocs/copyright/en/sccr_29/sccr_29_4.pdf (last visited Nov. 2, 2015).

be allowed under concepts of fair use or fair dealing if performed by an individual. Other SCCR Criteria are often addressed through other copyright law provisions. For example, *parallel importation* is a first sale or exhaustion concept in many countries;⁵¹ likewise the issue of *orphan works* is often addressed through collective licensing or damage remission.⁵² Limitations on library liability are often addressed outside the library provisions, such as in a damages provisions. The United States, however, includes an immunity provision in its library provision,⁵³ and lending or ILL includes the right of dissemination without restriction of geographic boundaries (*i.e.*, cross-border). The remaining SCCR Criteria of preservation (including replacement), copying for patrons, lending or ILL (including cross border), TPMs, and contractual override are relevant to uses of grey literature. The results discussed above are assessed in light of these remaining elements.

IFLA Proposal. As part the SCCR's focus on libraries and archives, IFLA issued a *Treaty Proposal on Copyright Limitations and Exceptions for Libraries and Archives* ("IFLA Proposal")⁵⁴ in accordance with the International Council on Archives, Electronic Information for Libraries, and Innovarte Corporation, a library NGO.⁵⁵ The Treaty Proposal lays out the key issues for libraries and archives and provides a suggested framework for promoting creativity with protecting the rights of authors and other copyright holders.⁵⁶ For purposes of this paper, the authors eliminated or combined certain elements into our discussion and inquired and whether a particular country's copyright laws addressed the issues; the results are in Appendix B. While their application to grey literature is not specified, the SCCR Criteria and the suggestions in the IFLA Proposal should apply to non-text works in addition to text-based works because libraries and archives preserve the audio and images in their collections in the same way.

Preservation. Based on analysis of the SCCR Criteria and IFLA Proposal, a basic requirement of any effective library copyright provision is the right to preserve both *text* and *non-text works*. *Preservation* is most often associated with *unpublished works*. In fact, five countries (*i.e.*, Canada, India, South Africa, United Kingdom, and United States) make the preservation right dependent on the status of the work being unpublished. As most grey literature is unpublished, this right is paramount and should not depend upon the format of the work. For *published works*, a somewhat parallel right is *replacement*. A replacement provision allows a library to replace an item in its collection that is lost, damaged, or in obsolete format (*e.g.*, Slovakia). Typically, a replacement right of a published work depends on a check for market availability. Since most, if not all, grey literature is unpublished and/or not made available through the commercial marketplace, at least initially, it is unlikely that library reproduction and dissemination rights are impacted by exercise of the replacement right.

Reproduction. In addition, libraries can engage in *reproduction for individual patron use*. A common approach, discussed above, is to restrict the use to research or personal use. Considering various statutes suggests that these rights are directed toward *literary works* and not to non-text works (*e.g.*, recordings, images, or motion media). An extension of this right is *reproduction for a patron at another library*, or ILL. Given the global nature and community of scholars, such exchanges are often across international borders (where the law allows). Again, many statutes suggest that such exchanges are limited to text-based literary works or not allowed across borders.

Licensing. As Lipinski and Copeland have discussed, grey literature collections are available subject to a *licensing agreement*, and some of the terms may by contract limit the rights of libraries and users in the copyright law.⁵⁷ As more and more content becomes subject to a license agreement, whether *commercially* (such as a database subscription agreement) or *non-commercially* (through a Creative Commons or other End User License Agreement), such clauses

⁵¹ See, *e.g.*, *Kirtsaeng v. John Wiley & Sons, Inc.*, 133 S. Ct. 1351 (2013).

⁵² See, *e.g.*, U.S. Copyright Office, *Orphan Works and Mass Digitization: A Report of the Register of Copyrights* (Jun. 2015), available at <http://www.copyright.gov/orphan/> (last visited Nov. 2, 2015).

⁵³ See 17 U.S.C. § 108(f)(1), "Limitations on exclusive rights: Reproduction by libraries and archives," Copyright Act of the United States, *supra* note 22.

⁵⁴ International Federation of Library Associations, *Treaty Proposal on Copyright Limitations and Exceptions for Libraries and Archives*, Version 4.4 (Dec. 6, 2013), available at www.ifla.org/copyright-tlib [hereinafter "IFLA Proposal"] (last visited Nov. 2, 2015).

⁵⁵ IFLA Proposal at 2.

⁵⁶ See IFLA Proposal at 4-6.

⁵⁷ See Tomas A. Lipinski and Andrea Copeland, *Look Before You License: The Use of Public Sharing Websites in Building Patron Initiated Public Library Repositories*, PRESERVATION, DIGITAL TECH. & CULTURE (Vol. 42, No. 3, Nov. 2013), available at <https://scholarworks.iupui.edu/bitstream/handle/1805/4574/lipinski-2013-look.pdf?sequence=1> (last visited Nov. 2, 2015).

will become increasingly important, ensuring that library collections are not rendered unusable by restrictive contractual terms and conditions.

Contractual Override. The revised statute of the United Kingdom and a proposed South African bill include a so-called “Teflon” clause that prevents any contractual provision from overriding use rights provided for in the library provision.⁵⁸ In other words, a contract (such as a database subscription agreement, i.e., a license) cannot offer more restrictive terms than those articulated in that statute without being null. The United Kingdom statute, for example, contains the following language in its interlibrary loan, replacement, and copying for patron provisions.: “To the extent that a term of a contract purports to prevent or restrict the doing of any act which, by virtue of this section, would not infringe copyright, that term is unenforceable.”⁵⁹ This is an important provision found in only two statutes reviewed for this study and discussed above.

TPMs. The final SCCR Criteria relates to Technological Protection Measures (“TPMs”). As more and more content is available in digital format, including born digital works, the use of TPMs may increase in commercial settings to control access to or use of the work. According to the tabulated information in the 2015 Crews Report, 112 countries have some form of TPMs.⁶⁰ Over fifty of those countries offer some circumvention of TPMs, yet little more than half (twenty-nine) allow circumvention when the desired use is ultimately a lawful, non-infringing use. This result is inconsistent with the purpose of TPMs as a way to thwart the mass infringement of digital content. Moreover, the lack of uniform and universal ability by libraries to circumvent TPMs for lawful use can impede those libraries identified as having some form of library exemption. Some countries have regulatory, administrative, or other mechanisms (such as mediation) to authorize exceptions, yet these result in inconsistencies and additional costs and delays when libraries seek exemptions for what are ultimately lawful uses. A “TPM override” provision would allow a library to circumvent a TPM for lawful use.

Formatting Use Rights. The authors also recommend that library use rights should be format-neutral. This would be the simplest way of providing the inclusion of non-text material, including grey content, in the cluster of rights afforded to libraries.

VI. Conclusion

These results demonstrate that the current copyright laws of many countries in the GL conference community are inadequate when applied to non-text or mixed-text works. Recommendations based upon the modified list of SCCR Criteria and the IFLA Proposal indicate many opportunities for change. While the *preservation* of grey literature is well-established, the *digitization* of grey content is either not allowed or limited to in-house or on-premises uses alone. While some mass digitization projects can fall under a concept of fair use,⁶¹ few countries have or are proposing fair use. As a result, a robust library exception is needed.

Preservation and replacement of *all* works and *all* formats, including new and mixed media and born-digital works, should be allowed, subject to the work’s non-commercial availability and restriction to particular uses (e.g., personal, educational, or scholarly activities). Remote access to the digitized content should also be allowed, or, at the very least, users should be able to search and locate collections and use interlibrary loan to obtain use of the works. In addition, lawful use of works under the copyright law should not be restricted by TPMs or contractual provisions. These changes offer an array of rights consistent with Articles 8 and 9 of the *Pisa Declaration on Policy Development for Grey Literature Resources*, relating to the collection, access, and use of grey literature.⁶²

⁵⁸ See § 37, South Africa Copyright Amendment Bill, *supra* note 7; §§ 41(5) (“Copying by librarians: supply of single copies to other libraries”), 42(7) (“Copying by librarians, etc.: replacement copies of works”), and 42A(6) (“Copying by librarians: single copies of published works”), Copyright, Designs, and Patents Act of the United Kingdom, *supra* note 8.

⁵⁹ §§ 41(5) (“Copying by librarians: supply of single copies to other libraries”), 42(7) (“Copying by librarians, etc.: replacement copies of works”), and 42A(6) (“Copying by librarians: single copies of published works”), Copyright, Designs, and Patents Act of the United Kingdom, *supra* note 8.

⁶⁰ See 2015 Crews Report.

⁶¹ See *Authors Guild, Inc. v. Google, Inc.*, Dkt. No. 13-4829-cv (2d Cir. 2015) (“If the library had created its own digital copy to enable its provision of fair use digital searches, the making of the digital copy would not have been infringement,” *id.* at 44); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) (holding that facilitating full-text search to identify specific page reference and providing access by users “with certified print disabilities” are fair uses).

⁶² See GreyNet Grey Literature Network Services, *Pisa Declaration on Policy Development for Grey Literature Resources* at Arts. 8 and 9 (May 16, 2014), available at http://www.greynet.org/images/Pisa_Declaration,_May_2014.pdf (last visited Oct, 23, 2015).



References

Scholarly Articles

Chamberlain, K. (2011). "Lawfully made under this title": The implications of *Costco v. Omega* on the first sale doctrine and library lending. *Journal of Academic Librarianship*, 37(4), 291-298. Available at <http://www.sciencedirect.com/science/article/pii/S009913331100070X>.

Lipinski, T.A. and Copeland, A. (2013). Look before you license: The use of public sharing websites in building patron initiated public library repositories. *Preservation, Digital Technology, and Culture*, 42(3), pp. 174-198. Available at <https://scholarworks.iupui.edu/bitstream/handle/1805/4574/lipinski-2013-look.pdf?sequence=1>.

Lipinski, T.A. (2009). Green light for grey literature? Orphan works, web-archiving and other digitization initiatives: Recent developments in U.S. copyright law and policy. *The Grey Journal: International Journal on Grey Literature*, 5(1), 11-21. Available at <http://www.moyak.com/papers/grey-literature-digitization.pdf>.

Lipinski, T.A. (2003a). The myth of technological neutrality in copyright and the rights of institutional users: Recent legal challenges to the information organization as mediator and the impact of the DMCA, WIPO, and TEACH. *Journal of the American Society for Information Science and Technology*, 54(9), 824-835. Available at <http://dl.acm.org/citation.cfm?id=874019>.

Lipinski, T.A. (2003b). The climate of distance education in the 21st century: Understanding and surviving the changes brought by the TEACH (Technology, Education, and Copyright Harmonization) Act of 2002. *Journal of Academic Librarianship*, 29(6), 362-374. Available at <http://www.editlib.org/j/ISSN-0099-1333/v/29/n/6/>.

Schöpfel, J. & Lipinski, T.A. (2012). Legal aspects of grey literature. *The Grey Journal: International Journal on Grey Literature*, 8(3), 137-153. Available at https://hal.inria.fr/file/index/docid/905090/filename/Schopfel_and_Lipinski_Legal_Aspects_of_Grey_Literature_.pdf.

Reports and Studies

Crews, K.D. (2015, June 14). *Study on Copyright Limitations and Exceptions for Libraries and Archives: Updated and Revised*. SCCR/30/3. Geneva, Switzerland: World Intellectual Property Organization. Available at http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=306216.

Crews, K.D. (2014, Nov. 5). *Study on Copyright Limitations and Exceptions for Libraries and Archives*. SCCR/29/3. Geneva, Switzerland: World Intellectual Property Organization. Available at http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=290457.

Electronic Information for Libraries, *Draft Law on Copyright: Including Novel Exceptions and Limitations for Libraries and Their Users* (2014, Work in Progress). Available at http://www.eifl.net/system/files/resources/201411/eifl_draft_law_2014.pdf.

Gibert, B. (2015). *The 2015 Intellectual Property and Economic Growth Index: Measuring the Impact of Exceptions and Limitation in Copyright Growth, Jobs and Prosperity* (revised May 2015). Brussels, Belgium: The Lisbon Council. Available at <http://www.innovationeconomics.net/component/attachments/attachments.html?id=263&task=view>.

GreyNet Grey Literature Network Services (2014, May 16). *Pisa Declaration on Policy Development for Grey Literature Resources*. Available at http://www.greynet.org/images/Pisa_Declaration,_May_2014.pdf.

International Federation of Library Associations (2013, Dec. 4). *Treaty Proposal on Copyright Limitations and Exceptions for Libraries and Archives*, Version 4.4. The Hague, Netherlands. Available at www.ifla.org/copyright-tlib.

U.S. Copyright Office (2015). *Orphan Works and Mass Digitization: A Report of the Register of Copyrights*. Washington, D.C.: U.S. Copyright Office. Available at <http://www.copyright.gov/orphan/> (last visited Oct. 23, 2015).

World Intellectual Property Organization Standing Committee on Copyright and Related Rights (2014). *Consolidation of Proposed Texts Contained in Document SCCR/26/3*. Prepared by African Group, Brazil, Ecuador, India, and Uruguay, Twenty-ninth Session, Geneva, Switzerland (Dec. 8-12, 2014). Available at http://www.wipo.int/edocs/mdocs/copyright/en/sccr_29/sccr_29_4.pdf (last visited Oct. 23, 2015).

United States Case Law

Kirtsaeng v. John Wiley & Sons, Inc., 133 S. Ct. 1351 (2013).



International Copyright Laws

Algeria: Copyrights and Related Rights Act of Algeria, No. 03-05 (July 19, 2003), available at http://www.wipo.int/wipolex/en/text.jsp?file_id=178342.

Australia: Copyright Law of Australia, No. 63 (June 27, 1968), as amended through No. 31 (May 27, 2014), available at http://www.wipo.int/wipolex/en/text.jsp?file_id=336977.

Czech Republic: Act on Copyright and Rights Related to Copyright of the Czech Republic, No. 121/2000 (April 7, 2000), as amended through No. 216/2006 (May 22, 2006), available at http://www.wipo.int/wipolex/en/text.jsp?file_id=137175.

Canada: Copyright Act of Canada, c. C-42 (1985), as amended through Jan. 2, 2015, consolidated as of Mar. 31, 2015, available at http://www.wipo.int/wipolex/en/text.jsp?file_id=366684.

India: Copyright Act of India, No. 14 (4 Jun. 1957), as amended through Act No. 49 (30 Dec. 1999), available at http://www.wipo.int/wipolex/en/text.jsp?file_id=128098.

India: Copyright (Amendment) Act of India, No. 27 (June 7, 2012), available at http://www.wipo.int/wipolex/en/text.jsp?file_id=304385 (last visited Oct. 23, 2015).

Japan: Copyright Act of Japan, Act No. 48 (May 6, 1970), as amended through Act No. 43 (June 27, 2012), available at http://www.cric.or.jp/english/clj/doc/20130819_July,2013_Copyright_Law_of_Japan.pdf.

Norway: Act Relating to Copyright in Literary, Scientific, and Artistic Works etc. of Norway, No. 2 (12 May 1961), as amended through 22 December 2006, available at http://www.wipo.int/wipolex/en/text.jsp?file_id=248181.

Poland: Copyright and Related Rights Act of Poland, No. 83 (Feb. 4, 1994), as amended through Alteration of the Law on Copyright and Neighboring Rights, No. 91 (Oct. 10, 2010), available at http://www.wipo.int/wipolex/en/text.jsp?file_id=129378.

Russia: Civil Code of the Russian Federation, No 230-FZ (Dec. 18, 2006), as amended through Amendment No. 35-FZ (Mar. 12, 2014), available at http://www.wipo.int/wipolex/en/text.jsp?file_id=345444.

Slovakia: Law of Copyright and Related Rights of Slovakia, No. 618/2003 (Dec. 4, 2003), as amended through No. 453/2008 (2008), available at http://www.wipo.int/wipolex/en/text.jsp?file_id=189474.

Slovenia: Copyright and Related Rights Act of Slovenia, No. 21 (1995), as amended through No. 16 (Dec. 15, 2006), available at http://www.wipo.int/wipolex/en/text.jsp?file_id=180840.

South Africa: Copyright Amendment Bill, No. 646 (July 27, 2015), available at <http://blogs.sun.ac.za/iplaw/files/2015/08/gg39028.pdf>.

Copyright Regulations of South Africa (1978), as amended through GN 1375 (1985), available at http://www.wipo.int/wipolex/en/text.jsp?file_id=130435.

United Kingdom: Copyright, Designs, and Patents Act of the United Kingdom, Chapter 48 (Nov. 15, 1988), as amended through Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, Statutory Instrument 2014 No. 1372 (May 19, 2014), available at <http://www.legislation.gov.uk/ukpga/1988/48/part/I/chapter/III/crossheading/libraries-and-archives>.

United States: 17 U.S.C. § 108, Copyright Act of the United States, Pub. L. No. 94-553 (Oct. 23, 1976), as amended through Pub. L. No. 111-295 (Dec. 9, 2010), available at <http://www.wipo.int/edocs/lexdocs/laws/en/us/us352en.pdf>.

APPENDIX A. *Analysis of library exceptions and limitations in copyright laws of sample countries.*

Country	Rights	Application to text-based works	Application to non-text or mixed-text works	Comments
Algeria	“Library” and “document keeping center” (archive?) can reproduce “article” or “excerpt” from a “collection of works, newspaper volumes or periodicals.” Art. 45.	In absence of collective license; “for educational, academic research, or personal purposes” if isolated and non-recurring. Art. 45.	“Library” and “document keeping center” (archive?) can reproduce in response to request from another library or document keeping center (ILL?), or for purposes of preservation or replacement if isolated and non-recurring.	Art. 45 applies to text-based on example provided. Art. 46 does not delineate; could argue this applies to any category of work. However, parallel subsection in U.S. law, for example, does not apply specifically to “musical work, a pictorial, graphic or sculptural work, or a motion picture or other audiovisual work.” See 17 U.S.C. § 108(i).
Australia	Right of reproduction for “article or part of an article”; replacement copying allowed. Art. 50.	User request or ILL. Reproduction for “the whole, or of more than a reasonable portion” requires no commercial copy available; likely does not apply to most grey literature.	Subsec. (3) refers to “whole or part of a work” and (4) refers to “or of any other published work”; (7B)(b) “work from which the reproduction is made is in electronic form” suggests digital media is included.	Under Subsec. (7)(C), reproduction can be in electronic form but entire context of Arts. 49 and 50 appear to relate to literary works alone.
Cameroon	No specific library provision, but “Without prejudice to the author’s right to an equitable remuneration, reproductions may be preserved in official archives.”			Limited to copy made for official archive.
Canada	Preservation / replacement reproduction rights for unpublished works; reproduction rights for published works.	“Research or private study rights” for published works.	Reproduction right for “research or private study” of published works (does not apply to musical work). ILL allows use of digital copy for delivery.	Rights apply to “library, archive or museum” but in context apply to literary works: “article... periodical... newspaper or periodical.”
Czech Republic	Archiving, conservation, replacement; work cannot be commercially available or subject to license.	Limited to “dedicated terminals located on its premises.” Theses and dissertations limited for research or private study.	Works in audio or audiovisual format may be lent but without “possibility of making reproductions of works fixed in audio or audio visual formats.”	Wide variety of entities: “library, archive, museum, gallery, school, university and other nonprofit school-related and educational establishment.”

Greece	Rights to make “one additional copy” from collection of library or archive.	Right to reproduce or “transfer” (ILL).	No limitations on type of work if not commercially available	Not limited by text. Applies to any work (“of the work”) but context suggests provision modeled on basic exception of replacement and ILL (usually limited to literary works).
Iceland	Rights of production for public (“enjoying support from public funds”) in libraries and archives	Reproduction for ILL limited to works not commercially available. Replacement copying limited to “minor proportion of a work” and “parts of works.”	Rights not limited by text; work must be in collection of library or archive.	No exclusions <i>per se</i> , but context suggests modeled after basic preservation, replacement, and ILL.
India	Replacement and preservation rights; ILL and copying for patron “for purposes of research or private study.”	Unpublished literary, dramatic, or musical work. Digitization allowed through replacement provision (“in any medium by electronic means”) if library owns non-digital copy.	Sec. 52(o) applies to “pamphlet, sheet of music, map, chart or plan,” but other visual works noted. Preservation appears to support digitization: “in any medium by electronic means...if the library already possess a non-digital copy of the work.”	Applies to library, museum, or other institution.
Japan	Basic rights of ILL and preservation.	Reproduction for patrons (“research study”), preservation, and out-of-print supply to another library.	Exception for National Diet Library “to record on a memory work” and make interactive transmission. Other libraries may make “a single copy of a part of that work” for patron research.	By definition, “library materials” limits use of non-text works: “books, documents and other materials held in the collections.”
Norway	Reproduction rights for preservation (“conservation and safety”).	Digital access allowed (“using terminals”), but must be on library premises and not commercially available.	Text does not works (“copies of works”), but subject to collective licensing.	Applies to archives, libraries, museums, and educational and research institutions.
Poland	Basic rights of reproduction and preservation.	Published (“disseminated works”) for preservation (“maintain or protect”) of works in collection and for patron copy.	Digital access “for research or learning purposes through information technology system terminals” restricted to premises.	Applies to libraries, archives, and schools. No limits in text, but context suggests basic preservation, copying for patron, and on-premises access.
Russia	Broad preservation and replacement rights.	Digital copies of published works (“put into the civil circulation”) limited to premises. ILL and copying for patrons limited to articles; small-size works lawfully published in collections, newspapers, and other periodical prints; short extracts from other lawfully published written works (with or without illustrations). Similar right for educational organizations.	Reproduction “of copies of works on machine-readable media for whose use there are no required facilities” could represent digitization or print-disabled. State archives can preserve work on web (“works inserted on the Internet”); could include mixed-media or media-based works.	Public libraries and archives. Dissertation copies may be “electronic,” but premise limitation applies. Could include images or other media. Attribution required.



Slovakia	Basic reproduction and preservation rights.	Translation of text suggests right includes anticipatory copy for “protection... against loss, destruction or damage...”		Rights limited to works in collection of library or archive.
Slovenia	Very broad rights.	Reproduction limited to non-commercial use and works from collection.	Digitization rights without restriction (“on any medium”). Would apply to born digital. Possible premise limitations: “for internal use.”	Applies to “[p]ublicly accessible archives and libraries, museums and educational or scientific establishments.”
South Africa	Preservation and security for unpublished works.	Replacement of published work does not contain premises liability. See Copyright Amendment Bill, No. 646 (27 July, 2015).	Limitation from U.S. law does not contain § 108(i), so ILL and copyright for patron could be of a mixed-media or media-based work. Private study or personal use limitation. Market test required for entire or substantial portion reproduction.	Text mirrors United States provision 17 U.S.C. § 108.
Switzerland	Basic rights of preservation.	No replacement; broad back-up right exists, but “the original or the copy must be stored in an archive not accessible to the general public and be marked as the archive copy.”	No limitations; would apply to any work in listed entities; could include all media or mixed-media works.	Applies to “[p]ublic libraries, educational institutions, museums and archives accessible to the public.”
United Kingdom	Reproduction for another library or replacement (market test required). Reproduction of published works includes ILL and copying for patrons (one article or “reasonable proportion of any other published works”).	Reproduction made “available to the public” limited to “a dedicated terminal on its premises” suggest a general premises limitation. Reproduction for ILL (“whole or part of a published work” and “one article... or reasonable proportion”) and replacement cannot be overridden by contract: “To the extent that a term of a contract purports to prevent or restrict the doing of any act which, by virtue of this section, would not infringe copyright, that term is unenforceable.”	Rights for unpublished works, but not available where copyright owner has prohibited copying.	Applies to library, archive, museum, and educational establishments. Museum includes gallery. Reproduction of published and unpublished works subject to declaration for non-commercial research or private study.
United States	Reproduction for preservation and security of unpublished works; replacement of published work if not commercially available.	Reproduction for ILL and patrons restricted to private study, scholarship, or research.	Per § 108(i), exclusions for preservation, security, or replacement of musical, pictorial, graphic, or audio visual works apply, but cannot be made available outside premises.	Libraries and archives only; museums not included. ILL and copying for patrons limited to one copy for “private study, scholarship, or research.”



APPENDIX B. Application of select SCCR Criteria and IFLA Proposal to copyright laws of sample countries.

Country	Library Lending	Reproduction	Preservation	Cross-Border	Contract Override	TPM Override
Algeria	X	X	X	?		
Australia	X	X	X	X		
Cameroon	X	X (archives)				
Canada	X	X	X	X		
Czech Republic	X	X	X			
Greece	X	X	X	X		
Iceland	X	X	X	X		
India	X	X	X	X		
Japan	X	X	X	X		
Norway	X	X	X			
Poland	X	X	X			
Russia	X	X	X	X		
Slovakia	X	X	X			
Slovenia	X	X	X			
South Africa	X	X	X	X	X	X
Switzerland	X	X	X			
U.K.	X	X	X	X	X	X
U.S.	X	X	X	X		

Library as Publisher: Convergence of New Forms and Roles of Textual and Non-Textual Grey Literature in Digital Scholarship

Julia Gelfand, University of California, Irvine
Anthony Lin, Irvine Valley College, United States

Abstract

Special Collections and University Archives typically constitute libraries', an organizations' or institutions' treasured assets. As libraries assume greater responsibilities for the curation and preservation of resources, they also become indispensable in issuing, publishing, releasing, and making original content accessible in digital formats, while the "born digital" content grows. Managing online content creation, in addition to handling traditional textual content shifts libraries into new roles based out of necessity and cost-constraints. With the dominance and reliance of electronic information now the status quo, the library's role as publisher has become increasingly important. Hosting and creating original content in partnership with other institutional networks harnesses the collaborative power of libraries by reducing operating expenses and increasing the availability of materials to any user with an Internet connection. Although it is often thought that more non-textual resources will tend to become increasingly grey, our experience indicates the contrary. Libraries strive to make such resources less grey than originally anticipated. The increased success in finding resources or content provides better identification and classification for sources that may not have been discovered initially with little to no metadata. The impact on library collections affirms the institutional mission of teaching, research and service by celebrating and honoring the memory of historical snapshots and experiences that would be otherwise forgotten or difficult to find or document. There are many challenges in this trend including scalability and sustainability, plus financial or business models that influence levels of output and innovation. This paper will highlight examples of the convergence of textual to non-textual content by showcasing different levels of library publishing and artifact models that celebrate and include a variety of library and organizational hosted published resources on different scales. The path of convergence is intertwined but blends new publishing potential, enhanced value and best practices with outputs that are defined more through textuality and new publishing models than by hues of grey.

Introduction and Background

Academic libraries are the sector with which we are most familiar and thus, this paper emphasizes that environment, but we acknowledge that such trends and development about "Library as Publisher" are taking place within all library communities and sectors. Libraries around the world are exploring how to handle and treat digital assets. There are several ways in which this is being demonstrated. By exploring and growing digital scholarship that has roots in the institution, it gives a different dimension to depending on commercial content that is often the research and creative output of its scholars, faculty, researchers and its students. The organizational culture of recent decades illustrates how Special Collections and University Archives typically constitute the treasured assets of libraries, organizations, or institutions, and is where rare books, manuscripts, and objects related to bibliographic resources and collections are found. These artifacts and collections have experienced a renaissance as many of these resources have become digitized and supplemented with a largesse of metadata to allow readers and users to access them remotely rather than be required to visit the holding library in person. This has constituted a major change in the research process.

Increasingly university presses have become aligned and affiliated with academic libraries and currently more university presses now share the organizational structure of libraries, with leadership, staffs and strategic planning synchronized between collections and new research outputs. Since most university presses carry non-profit status, and release books/monographs and journals and are an integral part of the research mission of a university, it has not gone unnoticed how great a challenge it is to remain fiscally solvent in these difficult economic times. The integration of publishing with libraries has created a very cohesive commitment to knowledge generation. Libraries are moving away from the warehousing of materials to information creation and that too is a new direction and way that libraries strive to be relevant in these changing dynamics of higher education and methods of reading and conducting research. Publishing and libraries is not a new relationship but in recent years is one that is an obvious outgrowth of both scholarly communications and the institutional repository movement that

began after the launch of the Internet, the maturation of ePublishing and the waning of traditional grey literature as it became easier to find, identify and access.

Library as Publisher

Library as Publisher – what does this mean? This question has been posed for over a decade and early work by the Association of Research Libraries (ARL) has captured it while its member libraries explored options for research library publishing services (Hahn). As libraries assume greater responsibilities for the curation and preservation of resources, they also become indispensable in issuing, publishing, releasing, and making original content accessible in digital formats. Managing online content creation, in addition to handling traditional textual content shifts libraries into new roles based out of necessity and cost-constraints. With the dominance and reliance of electronic information now the status quo, the library's role as publisher has become increasingly important. "This growth of library publishing and of collaborations between libraries and university presses signals a desire both to challenge and complement long established scholarly publishing practices." (Horava & Barker). This quote suggests goals libraries have for this new role that embrace peer review, value added services, Open Access and other related aspects of scholarly communications. Significant documentation of this trend was compiled in a final research report issued by SPARC and numerous libraries in 2012 (Mullins).

Much was learned from the Library Press Collaboration Survey conducted in late 2013 by the American Association of University Presses, Library Relations Committee in conjunction with research libraries and these findings include

(http://www.aapnet.org/images/stories/data/librarypresscollaboration_report_corrected.pdf):

1. Library publishing services are on the rise. 65% of respondents say library-publishing programs are an increasingly important service. 62% of all respondents (77% of library respondents and 34% of press respondents) to this question agree that publishing should be part of the library's mission. Presses must imagine a way of engaging with these and other emerging publishers on campus. (p.3)

2. Collaboration rather than duplication is recommended. 69% of respondents believe that library-publishing initiatives should complement press publishing programs, rather than reinventing (or duplicating) a service for formal peer-reviewed literature. (p. 3)

3. Recognize and discuss mission overlap. 95% of respondents see the need for presses and libraries to engage with each other about issues facing scholarly publishing beyond the usual topics of open access, fair use, and copyright. Common interests—such as how to best serve scholars—rather than areas of divergence, would be fruitful topics of discussion. (p.3)

4. Understand the scope of publishing activity on your campus. Respondents indicated some knowledge of many on-campus publishing operations, but a large percentage of respondents had no sense of number or scale. Recognize where these operations present opportunities for your press. (p. 3)

5. Look beyond the financial figures. Many libraries provide support to presses, but it is usually in-kind rather than money. While in only 11% of cases did the library provide cash support to a press, more than 53% of libraries provide other kinds of service ranging from digitization, metadata, and preservation services to office support and rent-free space (p. 3)

Today, "library as publisher" signals the best of scholarly publishing practices that increasingly promotes the open access or open source content that libraries and academics are moving forward as paramount to their work.

Some examples of these best practices from my own institution at the University of California include:

- eScholarship Repository (<http://escholarship.org/>)
- the UC Open Access Policy that went into effect October 23, 2015 (<http://osc.universityofcalifornia.edu/open-access-policy/>)
- the University of California Press launch of two OA initiatives 1) Luminos (books - <http://www.luminosoa.org/>) and 2) Collabra (a mega journals on the PLoS scale – each generously sponsored by initial funding from the Mellon Foundation (<http://scholarlykitchen.sspnet.org/2015/01/21/university-of-california-press-introduces-new-open-access-publishing-programs/>))
- creating digital scholarship units within libraries to promote library collections and resources, faculty scholarship, students' research (<http://www.lib.uci.edu/dss/>)

On a more global scale there are some very valuable tools that have become established resources to aid all the players in this expanding universe. The higher education enterprise with institutional members, established support and member driven organizations and the foundation/philanthropic arm that has generously supported many initiatives are all players. Some of the most recent additions to make their mark include:

- The Library Publishing Coalition – Launched in 2013-14 to promote the development of innovative, sustainable publishing services in academic and research libraries to support scholars as they create, advance, and disseminate knowledge. Now includes >60 institutional members. (<http://www.librarypublishing.org/>)
- Library Publishing Directory, 3rd ed., 2016 (http://www.librarypublishing.org/sites/librarypublishing.org/files/documents/Library_Publishing_Directory_2016.pdf)
- SPARC – under the aegis of the Association of Research Libraries it is the longest international alliance of academic and research libraries working to create a more open system of scholarly communication (<http://www.sparc.arl.org/>)
- Coalition of Networked Information (CNI) - dedicated to supporting the transformative promise of digital information technology for the advancement of scholarly communication and the enrichment of intellectual productivity (<http://www.cni.org>)
- Society of Scholarly Publishing (SSP) – increasingly connecting the dots between libraries as not only customers but as publishing partners of scholarly content (<http://www.sspnet.org/>)

Relationships to Grey Literature

There are many actions for Grey Literature in this hybrid publishing and collections arena. Libraries are increasingly learning the value of grey literature as it becomes more of a focus where libraries currently are engaged in to open up resources by digitizing content and rebirthing it with layers of metadata that offer new methods of finding, navigation and access.

This conference has hosted many presentations and discussions about definitions, levels and hues of grey. Earlier research that one or both of us has conducted reflects how Grey Literature is treated in Distance Education, eScience, ETDs, Social Media and Data Management to name a few domains and applications (Gelfand, Gelfand and Lin, Gelfand and Tsang). Current trends and experiences suggest that they are morphing into more traditional content that makes it less grey and more transparent to readers and users. This is accomplished by:

- Hosting and creating original content in partnership with other institutional networks to harnesses the collaborative power of libraries by reducing operating expenses and increasing the availability of materials to any user with an Internet connection.
- Although it is often thought that more non-textual resources will tend to become increasingly grey our experience indicates the contrary. Libraries strive to make such resources less grey than originally anticipated by opening up new opportunities and access points while more seriously promoting resources through websites, exhibitions, access in multiple places
- The increased success in finding resources or content provides better identification and classification for sources that may not have been discovered initially with little to no metadata.

Several implications suggest the importance to library collections:

- Libraries, regardless of largesse and budgets are clearly buying less – the expense of commercial content is not sustainable as library budgets remain stable if not declining. Libraries are challenging how they are “buying back” content that was produced by scholars paid to generate this new content. As committed as libraries are to resource sharing, they continue to explore ways to add content via demand driven and patron initiated actions, use document supply options and consider how to legally provide access in the digital age. The financial investment in electronic resources challenges how the acquisitions and licensing process intersects.
- The impact on library collections affirms the institutional mission of teaching, research and service by celebrating and honoring the memory of historical snapshots and experiences that would be otherwise forgotten or difficult to find or document.
- There are many challenges in this trend including scalability and sustainability, plus financial or business models that influence levels of output and innovation. Academic libraries

around the globe are focusing on several trends that address the importance of collection significance. These include:

- digital scholarship
- innovations in technology
- distance education models
- student learning outcomes & assessment
- globalization
- partnering with funding agencies & new mandates
- ability to license through Creative Commons

Examples that highlight the convergence of textual to non-textual content by showcasing different levels of library publishing and artifact models that celebrate and include new information products can be summarized by the following packaging or common authorship, all of which could be considered having hues of grey at an earlier stage of analysis:

- **National level collections** – We have explored how discovery is the conduit for success in some of these examples. The reason is that the success of online resources is dependent on how readers find, navigate and utilize resources. The increased success in finding resources or content provides better identification and classification for sources that may not have been discovered initially with little to no metadata. We are most familiar with the Library of Congress' (LC) American Memory Project of photography, memoirs, biography, and social media the online resource compiled by the LC National Digital Library Program. With the participation of other libraries and archives, the program provides a gateway to rich primary source materials relating to the history and culture of the United States. Currently containing more than five million historical items in coming years, the National Digital Library Program plans to digitize more of the Library's unique American history collections and make them freely available to teachers, students, and the general public over the Internet. Special collections to be digitized include the documents, films, manuscripts, photographs, and sound recordings that tell the American story as an excellent example but we want to explore discovery as the conduit for success in some of these examples. The reason is that the success of online resources is dependent on how readers find, navigate and utilize resources. The increased success in finding resources or content provides better identification and classification for sources that may not have been discovered initially with little to no metadata. (<http://memory.loc.gov/ammem/index.html>)
 - The National Library of Singapore (NLS) which recently created a masterplan for their collections to take into account the digital strategies to improve the experience of its patrons (Tang, 2015). The five capabilities noted in the strategies follow and we have coined them the “ables”:
 - FindABLE - “I can discover NLS/NAS Content in NLB, as well as outside NLB, across format and languages.”
 - ExpandABLE - “I can go further than my first hit and see a wider context from what I am searching for.”
 - UsABLE - “I can use the content because I can reproduce it, manipulate & repurpose.”
 - DeliverABLE - “I can ask for content I want delivered to me in a platform or channel I desire.”
 - PreserveABLE - “I can rest assured that everything in the National Collection that is precious will last beyond this generation.”

Items such as audiovisual content can prove challenging as metadata needs to be added to improve the findability. The NLS will attempt to add metadata for these items such as extracting textual information from photographs and videos (i.e. building photographs and building names), image matching for photographs and videos (identifying shapes of objects), and voice to text technology for automatically dictate speeches that are in audio. In addition to the metadata, the NLS will create a new federated search, to be known as “One Search” to make findability easier for internal searches between their platforms, offering these features:

- digital content is currently spread over a number of different format portals from (NewspaperSG (for newspapers), BookSG (for books, periodicals, and text based materials), PictureSG (for photographs), MusicSG (for music), NORA (for arts personalities), Web Archive Singapore (for archived websites), to SingaporeMemory.SG (for personal memories))
- At this time, each microsite is currently resides in isolated silos that do not allow for easily search ability across platforms. (Tang 2015)



Implications of these issues:

- 1) The impact on library collections affirms the institutional mission of teaching, research and service by celebrating and honoring the memory of historical snapshots and experiences that would be otherwise forgotten or difficult to find or document.
 - 2) There are many challenges in this trend including scalability and sustainability, plus financial or business models that influence levels of output and innovation.
 - 3) The new “One Search” connects together via federating formats on a single platform reducing complexity and confusion for the user and increasing transparency, serendipity, and ease of use.
- **Large State-level research library digital collections**
 - **Online Archive of California** (<http://www.oac.cdlib.org/>) – a collaborative ongoing interdisciplinary archive about anything relevant to the State of California created by more than 200 organizations and hosted by the University of California
 - **Community-based museum and library collections** that document important human history events such as genocide, catastrophic accounts of loss, displacement, internment, slavery, refugee and migration status with oral histories, ethnographies, genealogies. Most of these examples had their start with print content and eventually expanded with electronic resources and now include a wide variety of formats, media and information available online.
 - **Holocaust Resource Centers** in nearly every major city around the world – e.g. Yad Vashem (http://www.yadvashem.org/yv/en/holocaust/resource_center/index.asp)
 - **Armenian Genocide Collection** (<https://www.facinghistory.org/for-educators/educator-resources/resource-collections/armenian-genocide-resource-collection/library-resources>)
 - **Southeast Asian Refugee** experience documented with photos, oral histories, social media, in Southern California as immigrants from Laos, Cambodia and Vietnam have settled and formed enormous new communities (<http://seadoc.lib.uci.edu/>)
 - **Genealogy resources** – provides tools for families, hobbyists, teachers, faith-based groups, ethnic and racial populations, and increasingly important in independent and public libraries (<https://www.newberry.org/genealogy-and-local-history> ; <https://familysearch.org/>)
 - **Government documents collections** – Records of a government activity, movements, social awareness
 - **WikiLeaks** – Begun in 2006 in Iceland as an international, journalistic news organization to publish and disseminate secret and classified information and news leaks from anonymous sources many of which became front-page news items. The wide variety of video and news coverage released is significant and credited to its team headed by Julian Assange. Some of the content included news that had never previously been released including prisoners’ files of those detained at Guantanamo Bay. As WikiLeaks expanded it collaborated with other news and media channels to release redacted content that occasionally was accused of errors and concerns arose about what impact some of this information could have on innocent persons. (<https://www.wikileaks.org/index.en.html>)
 - **Federation of American Scientists** – Congressional Research Service Reports – Archived US Congressional Research Service reports that cover a variety of subjects of interests from the US Congress (<http://fas.org/sgp/crs/index.html>)
 - **Multimedia news archives** - Holdings of news reels, documentaries, life experiences and anniversaries from world events.
 - **EUScreen** – European portal that allows access for audiovisual heritage resources from the 20th and 21st Centuries. (<http://www.euscreen.eu/>)
 - **BBC Archives** – A news archive from the British Broadcasting Corporation that contains both video and audio content from past programming. (<http://www.bbc.co.uk/informationandarchives/archivenews/>)
 - **NPR Archives** – US based news radio archive from the National Public Radio broadcasts from years past. (<http://www.npr.org/series/4564213/historical-archives>)
 - **Global born-digital collections** - there are many thoughtful, exciting products utilizing many combinations of software tools to describe, showcase and allow for personalization as one uses them today. Examples are plentiful in the digital

humanities and applied sciences. Libraries are at the forefront of creating such resources as noted with the following example:

- **International Digital Ephemera Project** - an initiative to digitize preserve and provide broad public access to print items, images, multimedia and social networking resources produced in the Middle East – (<http://digital.library.ucla.edu/dep/>)
- **ACI Scholarly Blog Index** – a relatively new commercial product that indexes and accesses content found in blogs by distinguishing those that are trustworthy from those that can spread misinformation, reinforcing value of grey content. It provides academic researchers who are inclined to discuss their work, discoveries opinions in their own blogs or get covered in those of their colleagues or competitors. With inclusion of nearly 15,000 blogs this database is a growing source for information in non-traditional publications that incorporates social media and can be widely shared. (<http://aci.info/aci-blog-index-content/>)
- **African Commemorative Fabrics** – a portal created at the University of Wisconsin Digital Library to host a collection of machine-made commemorative textiles from various African countries. This collection provides researchers access to digitized fabrics that are printed with images and text documenting events and individuals of historical, political, religious, economic, educational, and sociological significance to African societies. Throughout the continent, fabric serves multiple functions in people’s daily lives. It is used for clothing, shelter, storage, and packing material. The type of African fabric found in this collection also serves as a communication device. When used as a textual and visual document, the fabric becomes a vehicle to commemorate an event or to celebrate a person’s life or achievement. One does not need to know how to read in order to understand the messages found in these textiles. In societies in this way, the fabric also serves to preserve historical narratives that are important to the community, and documents critical history regarding life, public health issues, lifecycle events, leadership successions and special celebrations. (<https://uwdc.library.wisc.edu/collections/AfricanStudies/Fabrics/>)
- **Internet archive** – Resource that provides a snapshot of a historical period on the Internet
 - **Internet Archive - Way-back Machine** – Founded in 1996 and located in San Francisco, the Archive has been receiving data donations from Alexa Internet and others. In late 1999, the organization started to grow to include more well-rounded collections. Now the Internet Archive includes: texts, audio, moving images, and software as well as archived web pages in our collections, and provides specialized services for adaptive reading and information access for the blind and other persons with disabilities. (<https://archive.org/>) (2015)
- **Data & Macroscopic Sciences** – With a growing emphasis on Big Data and data in general, how to manage data is a challenge most libraries are facing. Within the publishing sphere, authors and scholars are expected to make available their research data for repurposing and reuse especially if federal research funding supported the initial research. Data Management plans are increasingly important on the front end as a central component of a research proposal. The goal of librarians and publishers is to foster cleaner data for use by enhancing data with valuable information. There are numerous processes such as visualization, statistical analysis, graphical depictions, and different preferences that subject disciplines promote that focus on how data is described, stored, used and repurposed. Parking data in Data Sources, data questions, data credibility and the future of data are all concerns of managing and using data, large and small, empirical and real. What we try and avert is data malaise.
 - Tools that support and add value to data – an example of the program may include **DeepDive** (<http://deepdive.stanford.edu/>) which uses machine learning but is not algorithmic and “explores how next generation search and extraction systems can help with real-world use cases.” Increasingly of value it already has applications in the following areas:
 - Human trafficking
 - TAC-KBP Challenge (Text Analysis Conference, Knowledge Base Population)
 - Wisci(-pedia)
 - Geology & Paleontology
 - Medical Genetics
 - Pharmacogenomics



- **Software and Data Carpentry tools** – gaining lab skills for understanding and working more effectively with scientific and other forms of data
 - **LabArchives** – one of the fastest growing tools of the electronic lab archive movement allowing wet and dry labs to manage their research data by more easily creating, storing, and sharing their data both in instructional and professional settings as it stores Rich text data, tables, images, sketches, as well as annotations of images (<http://www.labarchives.com/>)

Conclusions

The path of convergence is intertwined but blends new publishing potential and best practices with outputs that are defined more through textuality and new publishing models than by hues of grey. The skepticism that was expressed in an opinion piece in *The Nation* in Spring 2014 about the harsh realities of university presses and scholarly publishing confirm that much of the system may be broken due to the shrinking buying power of academic libraries but we see more positive outcomes in these past 18 months (Sherman). Sometimes the problems in academe are more related to a “university in turmoil more than a library in distress” (Babb). Even though the experiences shared by Murzyn-Kupisz and Dzialek in their work apply to Malopolska, Poland, we concur that “a new outlook on museums and libraries and a new understanding of their social roles goes beyond their basic statutory functions or their traditional roles as preservers, researchers and disseminators of knowledge on collected artifacts or books.”

Okerson and Holtzman summarize the changing world that allows for more publishing success by libraries as due to:

- Digital technologies and ubiquitous access to them
- Cost reductions that lower publishing barriers
- The squeeze on library collections budgets
- A desire to reduce prices to libraries and “liberate” academic publishing
- A new vision of open access
- Increasingly complex challenges of balancing institutional priorities. (Okerson & Holtzman, pp. 4-6)

Publishing and creating new intellectual capital is increasingly the role and domain of libraries and this is clearly seen by the flurry of conferences taking place around the world on “Library as Publisher” by a variety of organizations (annual 2016 Library Publishing Forum (www.librarypublishing.org/events/lpforum16), the IFLA 2016 Satellite in Ann Arbor, MI, USA (<http://2016.ifla.org/programme/satellite-meetings>)) and the proliferation of supporting resources and guides issued by the Library Publishing Coalition (<http://libguides.uky.edu/libpub>) and the work of the American Association of University Presses in partnerships with other library and publishing organizations. Becoming a content strategist that bridges the world of libraries and publishing is the entrepreneurial spirit of the next generation of grey literature and its publishing partners.

Pundits, both optimists and skeptics have speculated that the library of 2100 will be compressed into a vision of libraries as collections or perhaps there will be 3 million libraries. These extreme answers to the question about the library of the future is dependent on calculations based on census predictions, current physical library buildings in the United State and doing the numbers by extrapolating. As Jim O’Donnell states, “we will see the consolidation of collections and a consolidation of the technical infrastructure of presenting those collections...and we will see the emergence of business models for paying for what we now think of as “publishing” that allow complete free and open access to the contents of this global library” (O’Donnell). We expect that the role of library as publisher will intensify as the OA movement grows and libraries expand their reach with increased capabilities and scope. As Okerson states, their new challenge will be how to market their outputs, so that the “tree does not fall unheard” (Okerson, ATG).

References

- American Association of University Presses, Library Relations Committee (2014), Library Press Collaboration Project Report http://www.aaupnet.org/images/stories/data/librarypresscollaboration_report_corrected.pdf
- Anderson, Rick. (2015). A quiet culture war in research libraries - and what it means for librarians, researchers and publishers. *Insights: The UKSG Journal*, 28(2), 21-27. doi:10.1629/uksg.230
- Babb, Meredith (2015). Challenges Facing University Presses. Interview with Meredith Babb by Baker & Taylor executive, posted on the Yankee Book Peddler/Gobi website. http://www.gobi3.com/StaticContent/GOBICContent/YBP/Public/misc/Challenges_Facing_UPs.pdf
- Bonn, Maria and Furlough, Mike, eds, (2015). *Getting the Word Out: Academic Libraries as Scholarly Publishers*. Chicago, IL: ACRL.

- Gelfand, Julia and Tsang, Daniel (2014). Data: Is it Grey, Maligned or Malignant? Paper presented at the 16th International Conference on Grey Literature, Washington, DC, December 9, 2014. Published in *The Grey Journal* 11, # 1, 2015: 30-40. <http://escholarship.org/uc/item/80w006rz?query=gelfand>
- Gelfand, Julia and Lin, Anthony (2012). Research Life Cycle: Exploring Credibility of Metrics and Value in a New Era of eScholarship that Supports Grey Literature, paper presented at the 14th International Conference on Grey Literature," November 30, 2012, Rome, Italy. Published in *The Grey Journal* 9 (3) Fall 2013
- Gelfand, Julia and Lin, Anthony (2011). Social Networking: Product or Process and What Shade of Grey? Paper presented at the 13th International Conference on Grey Literature, December 5, 2011, Washington, DC. GL13 Conference Proceedings, February 2012. Published in *The Grey Journal* 8 (1), Spring 2012: 14-27.
- Gelfand, Julia, (2009). New Shades of Grey: The Emergence of E-Science, Scientific Data and Challenges for Research Libraries. Presentation at the 11th International Conference on Grey Literature, Washington, DC, December 12, 2009. Published in the Conference Proceedings, 2010.
- Gelfand, Julia M., (2007). Updating Grey Literature as Distance Education Matures. Paper presented at the Ninth International Conference on Grey Literature, Antwerp, Belgium, December 11, 2007.
- Gelfand, Julia M., "Grey Literature: Taxonomies and Structures for Collection Development," Paper presented at the Eighth International Grey Literature Conference, New Orleans, December 4, 2006. Published in *GL8, Conference Proceedings: Eighth International Conference on Grey Literature: Harnessing the Power of Grey*, 4-5 December 2006, compiled by D. J. Farace and J. Frantzen. Amsterdam: TextRelease, February 2007: CD-ROM. Also published in *The GreyJournal*, vol. 3 #1:7-16 Spring 2007 <http://www.greynet.org/thegreyjournal.html>)
- Gelfand, Julia M., "Challenges for Collections in New Collaborative Teaching and Learning Environments: Does Grey Literature Fill a Void?" Paper presented at the Seventh International Grey Literature Conference, INIST, Nancy, France, December 5, 2005. Published in *GL7, Conference Proceedings: Seventh International Conference on Grey Literature, Open Access to Grey Resources* 5-6 December 2005, compiled by D.J. Farace and J. Frantzen. Amsterdam: TextRelease, January 2006:71-76.
- Gelfand, Julia M., "What is New in Grey Literature: Everything from the Textbook Market to the Blogosphere," *The Grey Journal*, vol. 1 #3, Autumn 2005.
- Gelfand, Julia M., "Knock, Knock, Has Grey Literature Found a Home in Institutional Repositories?" Paper presented at the Sixth International Grey Literature Conference, December 6, 2004, New York. Published in *GL6 Conference Proceedings: Sixth International Conference on Grey Literature: Work on Grey in Progress* 6-7, December 2004, compiled by D. J. Farace and J. Frantzen. Amsterdam: TextRelease, January 2005: 10-16. Also in *The Grey Journal*, vol. 1 #2 Summer 2005:61-66.
- Hahn, Karla (2008). Research Library Publishing Services: New Options for University Publishing. Washington, DC. *Research on Institutional Repositories: Articles and Presentations*. Paper 37. <http://digitalcommons.bepress.com/repository-research/37>
- Horava, Tony and Barker, Andrew (2015). Library/Press Collaboration: A Magical Mystery Tour. Presentation made at the Charleston Conference, November 5.
- Internet Archive. (2015). Retrieved November 21, 2015, from <https://archive.org/about/>
- McCormick, Monica (2015). Toward New Model Scholarly Publishing: Uniting the Skills of Publishers and Libraries. In Bonn and Furlough, eds., *Getting the Word Out: Academic Libraries as Scholarly Publishers*. Chicago, IL: ACRL, 57-82.
- Mullins, J. L., Murray-Rust, C., Ogburn, J. L., Crow, R., Ivins, O., Mower, A., Nesdill, D., Newton, M. P., Speer, J., & Watkinson, C. 2012. *Library Publishing Services: Strategies for Success: Final Research Report*. Washington, DC: SPARC. http://docs.lib.purdue.edu/purduepress_ebooks/24/
- Murzyn-Kupisz, Monika and Dizialek, Jaroslaw (2015). Libraries in and Museums as Breeding Ground of Social Capital and Creativity: Potential and Challenges in the Post-socialist Context, in *Creative Economies, Creative Communities: Rethinking Place, Policy and Practice*, edited by Saskia Warren and Phil Jones. Farnham, Surrey, UK: Ashgate: 160.
- O'Donnell, Jim (2015). What Will Libraries be Like in 2100?: It's not so far away. *Slate*, November. http://www.slate.com/articles/technology/future_tense/2015/11/what_will_libraries_be_like_in_2100.html based on forum at the New America Foundation. Full video available (about 2 hours) at <http://www.ustream.tv/recorded/77548793>
- Okerson, Ann (2015) Back Talk – I'm a Publisher Too! *Against the Grain (ATG)*, November: 94.
- Okerson, Ann and Holtzman, Alex (2015). *The Once and Future Publishing Library*. CLIR Report 166, July. <http://www.clir.org/pubs/reports/pub166>.
- Sherman, Scott (2014). University Presses Under Fire: How the Internet and slashed budgets have endangered one of higher education's most important institutions. *The Nation*, May 6. <http://www.thenation.com/article/university-presses-under-fire/>
- Songolo, Emilie Ngo-Nguidjol (2015). Preserving African Commemorative Fabrics: Implications for Social Understanding, Public Health and Archival Stewardship. Presentation made as part of the Program in Public Health Fall 2015 Lecture Series, University of California, Irvine, November 23

Academic blogging consequences for Open Science: a first insight into their potential impact

Carla Basili, National Research Council of Italy, CNR-IRCrES Institute;
Luisa De Biagi, CNR-Biblioteca Centrale, Italy

Abstract

Social media have been analysed in different studies and from different perspectives, and the findings show that they can constitute new forms of impact indicators of scientific activities, intended at large and not limited to the publication. Therefore, new forms of research evaluation are emerging as alternative to the traditional citation-based metrics. Initially based on web links or download numbers (webometrics) the new evaluation methods have evolved into alternative metrics, or “altmetrics”, based on the set of activities covered by social media services.

In view of the above, the paper concentrates on academic blogging as a form of scholarly communication in the Open Science context, and on the disciplinary areas of the Humanities and Social Science as privileged domain to be investigated.

The OpenEdition initiative is analysed as a complete infrastructure for the digital resources for the Humanities and Social Sciences and as a concrete exemplification of quality filters applied to these (relatively) new media.

1. Introduction and background

1.1 On-going trends in the Scientific System

Pisa Declaration raises a number of policy issues about non-formal channels of scholarly communication, including “openness”, a principle that permeates the Scientific System in three main areas:

- Knowledge dissemination – (Open Access);
- Knowledge creation - (Open Science), largely unexplored issue in the literature in the current EU vision.
- Research impact evaluation – (AltMetrics).

The first Open Access area, related to the free availability of research literature and data, is a mature subject matter, deeply studied in the literature and even transposed as obligation into the European research policies.

Therefore, in this paper the focus is on the two areas of Open Science and Altmetrics, rather than on (the indeed relevant) Open Access theme.

Open Science

Open Science is among the six pillars of the European Commission policy approach to Responsible Research and Innovation, an approach defined by the Observatory of Responsible Research and Innovation (RRI) as follows:

Responsible Research and Innovation means that societal actors work together during the whole research and innovation process in order to better align both the process and its outcomes, with the values, needs and expectations of European society. RRI is an ambitious challenge for the creation of a Research and Innovation policy driven by the needs of society and engaging all societal actors via inclusive participatory approaches.¹

Open Science is one of the thematic elements of RRI², and is conceived by the European Commission as follows:

Open science is about the way research is carried out, disseminated, deployed and transformed by digital tools, networks and media. Open science relies on the combined effects of technological development and cultural change towards collaboration and openness in research.³

Altmetrics

Galligan and Dyas-Correia (2013) provides a number of definitions for Altmetrics in the literature and suggest the following overall view on Altmetrics:

¹ The Observatory of Responsible Research and Innovation (RRI) <http://observatory-rri.info>

² <https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation>

³ The Digital Agenda for Europe: Open Science <https://ec.europa.eu/digital-agenda/en/open-science>

[...] altmetrics examine the content of the social Web in order to provide either an alternative or enhancement to the use of journal impact factors and click-through rate analysis to measure the impact and value of scholarly work⁴.

These trends in the scientific system converge towards a model of science called Science 2.0, now known as Open Science, for which the European Commission launched the consultation “*Science in transition*” between July and September 2014. According to the background paper of the consultation “*Science in transition*”:

‘Science 2.0’ describes an on-going evolution in ways of doing and organising research. These changes are enabled by digital technologies, and they are driven by globalisation and growth of the scientific community as well as the need to address the grand challenges of our time. The changes impact the modus operandi of the entire research cycle, from the inception of research to its publication, as well as the way this cycle is organised⁵.

This definition is very effective for the purposes of this paper, as it corresponds to the perspective of analysis which will be conducted in the next paragraphs.

1.2 Academic Social Media

Academic Social Media constitute a quickly growing area of (relatively) new channels of Scholarly Communication, including different platforms:

- Social networking
- Blogging
- Micro-blogging
- Collaborative authoring tools for sharing and editing documents
- Social tagging and bookmarking
- Scheduling and meeting tools
- Conferencing
- Image or video sharing

This classification forms the basis of a survey⁶ conducted in 2010 by CIBER - and published in 2011 - for an exploratory study on the use of social media tools by a set of 1,923 researchers. Among the results of the study is the distribution of the percentage use of the diverse categories of social media tools.

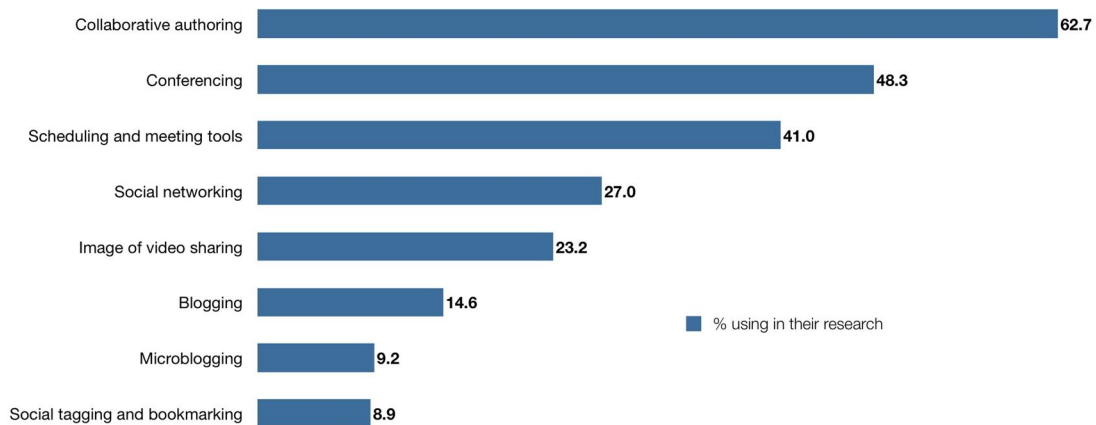


Figure 1 - Academic Social Media - active users (Source Ciber, 2010)

Figure 1 shows the Popularity of various types of social media in research, providing the percentages of active social media users for each category of tool.

⁴ Galligan, F., Dyas-Correia, S. (2013). Altmetrics: Rethinking the way we measure. *Serials Review*, 39(1), 56-61.

⁵ European Commission Directorate General for Research and Innovation (RTD) and DG Communications network, content and technology (CNECT) (2014) *Science in transition* background document, p. 1, <http://ec.europa.eu/research/consultations/science-2.0/background.pdf>

⁶ *Social media and research workflow*, CIBER, University College London, Emerald Group Publishing Ltd, 14 December 2010 available at <http://ciber-research.eu/download/20101111-social-media-report.pdf> (accessed November 2015)

Open Edition: Social Media in the Humanities and Social Sciences

The paper concentrates on academic contexts usually considered as resistant to technological innovation, namely the disciplinary areas of the Humanities and Social Science.

The OpenEdition system <<http://www.openedition.org>> will provide an exemplification in the case of the Humanities, in order to provide a first picture of the potential impact of these new “publication” channels, both for the Open Science context, and for research evaluation and scholarly reputation in academic contexts.

The system is supported by the CNRS, EHESS, Université d’Aix-Marseille, Université d’Avignon, the French Research Ministry and the Bibliothèque Scientifique Numérique and is a European initiative to promote the online publication and digital distribution of open access research in the humanities and social sciences, encompassing all disciplines.

OpenEdition offers the academic community four international-scale publication and information platforms in the Humanities and Social Sciences⁷ (numeric data update as of January 2016 – in square brackets the data collected on October 2015).

Reeves.org – 420 - [413] **Journals**: *Reeves.org* <<http://www.reeves.org>>: lauched in 1999, *Reeves.org* distributes journals in all disciplines of the humanities and social sciences in fourteen languages, which represents over 100,000 open access articles (95% in full text).

OpenEdition Books – 2644 - [2538] **books** enables the distribution of book collections on open access. Opened in February 2013, the platform hosts 50 publishers in the humanities and university presses based in several European countries.

Hypotheses – 1426 - [1365] **academic blogs**: *Hypotheses* <<http://hypotheses.org>> is a directory of research blogs for the humanities and social sciences, in different European languages. French, English, German, Spanish and Portuguese researchers make intensive use of the platform to exchange and distribute information about their research results or the latest developments in their field.

Calenda -30628 - [29846] **events**: *Calenda* <<http://calenda.org>> is an announcement platform for the humanities and social sciences posting announcements for seminars, conferences, calls for contributions and employment opportunities. *Calenda* was launched in 2000.

The interest of this paper concentrates on *Hypotheses*, the platform of academic blogging in Open Edition, in order to provide some evidence about the use of blogs in the Humanities and their potential impact on research conduct and dissemination.

Academic blogs in Open Edition by subject categories

In Open Editions, inside the vast disciplinary domain of Arts & Humanities are distinguished the following research areas:

- Arts & Humanities (535) [517]
- Education (104) [98]
- History & Archaeology (578) [552]
- Library, Information & Communication sciences (154) [145]
- Multidisciplinary (438) [436]
- Political Science, Public Admin. & Development (127) [124]
- Psychology (33) [32]
- Public Health & Health Care Science (29) [29]
- Social Work & Social Policy (57) [55]
- Sociology & Anthropology (376) [363]
- Language & Linguistics (85) [80]
- Economics (42) [39]
- Law (36) [34]
- Literature (158) [149]
- Management (21) [21]
- Psychiatry (8) [7]

For each research area, the number in round brackets indicates its cardinality (i.e. the number of blogs belonging to the area) as of January 2016, while numbers in square brackets indicate the cardinality of the single disciplinary subdomain as of October 2015.

⁷ The description of the platforms given below is a revised version of Open Edition online documents.

Academic blogs in Open Edition by types of blogs

Open Edition provides a categorisation corresponding - to a large extent - to the multiple activities that contribute to the production of new knowledge, particularly within an Open Science process life-cycle.

Research program blogs (317) [301]	Non-specialist blogs (44) [42]
Research blogs (234) [222]	Debate blogs (39) [38]
Laboratory blogs (182) [178]	Master's blogs (40) [36]
Seminar blogs (123) [118]	Methodology blogs (33) [31]
Thesis blogs (107) [99]	Library blogs (25) [25]
Monitoring blogs (64) [63]	Field work blogs (27) [24]
Publication blogs (41) [44]	Media blogs (8) [9]
Event blogs (44) [44]	Bibliography (1) [1]

It is merely a coarse grained description, nevertheless useful for the purposes of understanding how academic blogging can map into scholarly activities in a networked and open environment.

Mapping between blog types and Open Science activities

As already mentioned, in fact, Open Science has become a strategic priority in the European Commission research policy, where it is conceived as follows:

Open Science describes the ongoing transitions in the way research is performed, researchers collaborate, knowledge is shared, and science is organised. It is enabled by digital technologies, and driven by:

- *the enormous growth of data,*
- *the globalisation and enlargement of the scientific community to new actors (e.g. citizen science), and*
- *the need to address societal challenges⁸*

A recent (2015) report⁹ by IPTS (Institute for Prospective Technological Studies of the European Commission) provides a conceptual framework for scholarly activities, identified and listed as follows:

Managing the research process:

- Identifying a researchable topic
- Planning a research project
- Producing research output collaboratively
- Releasing laboratory notebooks to the scholarly community
- Keeping up with new developments
- Getting help for solving topical problems
- Participating in open peer reviewing
- Monitoring one's impact

Disseminating research findings

- Disseminating research results, ideas and opinions informally via blogs

Lack of regulation and, mainly, standardisation in academic blogging does not facilitate and, above all, undermines a reliable determination of a grid of unique matches between the different types of blogs and the different scholarly activities. However, the lists of blog types and scholarly activities provided above could facilitate the understanding of the kind of support that these new media can offer and a first insight into their potential impact.

Academic blogging and scholarly impact

The traditional filters for scientific literature - the well-known processes of Peer review, Citation counting, and Impact factor of journals - are increasingly subject to serious critiques, enabling a set of new filters to enter the arena of Research Assessment. These filters can be distinguished into two classes:

⁸ EUROPEAN COMMISSION. Research and Innovation. Science With And For Society. Policy. Open Science <http://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=science> accessed November 2015

⁹ HERMAN, E - JAMALI, H.R. - NICHOLAS, D. - OSIMO, D. - PORCU, F - PUJOL, L.(2015). Analysis of Emerging Reputation and Funding Mechanisms in the Context of Open Science 2.0, IPTS, Luxembourg: Publications Office of the European Union, 2015

Webometrics – quantitative techniques and tools for collecting data and calculating “indicators” like usage metadata (page views and downloads, Twitter counts, Facebook comments, science blog postings, bookmarkings and reference sharing numbers).

Altmetrics - evaluation methods of scholarly activities (not only publications) based on social media data that serve as alternatives to citation-based metrics.

The development of both these classes of tools, coupled with innovative forms of research results (beyond scientific literature), have given rise to two different dimensions of Scholarly impact:

- Research *impact*: evaluating research performance through webometrics is still in its infancy and still lacks of an established framework of evaluation.
- Scholar *reputation*: sort of “*de facto*” assessment (even self-assessment) through statistical evidence regarding the impact, usage, or influence of one’s own work.

Based on this distinction, detractors are used to claim that social media “serve as “technologies of narcissism”, more than “technologies of control”.

Open problems and concluding remarks

The whole set of Academic Social Media shows to have a great potential impact on most academic activities, nevertheless, a number of open problems still remains:

Here the most common barriers are listed:

- Authority and trust - lack of quality filtering mechanisms
- Unclear benefits
- Technology barriers (e.g. bandwidth)
- Uncertain moral rights – copyright protection
- Difficulties in citing non-traditional content
- Lack of time
- Lack of familiarity with social networking tools .

Despite these limits, a positive perception of academic social media is widely spread, and the fast growing of the Open Edition directory provides evidence of this recognition.

Social media are forms of free exchange of ideas, opinions, information, allowing communication and remote collaboration, and self-legitimizing and filter through a natural mechanism for the recognition of academic reputation.

Enabling technologies and financial constraints constitute major drivers for the ongoing move towards Open Science, where the “Openness” principle is moving from knowledge dissemination (Open Access) to the whole research cycle (Open Science).

Collaboration, transparency, globalisation, scientific reputation are the main keywords in this paradigm shift, and scholarly social media, and academic blogs in particular, constitute “*de facto*” means to achieve these goals. Nevertheless, academic social media can support a new approach in the assessment of the scholar reputation and visibility, but not yet in the evaluation of the impact of the research output.

Bibliography

- BASILIC C. (2015). *Open Science*. In: D. Archibugi et al. “The Contribution of the European Commission to Responsible Research and Innovation. A review of the Science and Society (FP6) and Science in Society (FP7) Programmes”, CNR Edizioni, 125-153
- CENTRE FOR INFORMATION BEHAVIOUR AND THE EVALUATION OF RESEARCH. (2010). *Social media and research workflow*. London, England: CIBER. Retrieved from <http://www.ucl.ac.uk/infostudies/research/ciber/social-media-report.pdf>
- CONOLE, G. (2007, 10 20). The nature of academic discourse. Accessed 17 03, 2015, on <http://e4innovation.com/?p=45>
- EUROPEAN COMMISSION (2015). Research and Innovation. Science With And For Society. Policy. Open Science <http://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=science> accessed November 2015
- GALLIGAN, F., DYAS-CORREIA, S. (2013). Altmetrics: Rethinking the way we measure. *Serials Review*, 39(1), 56-61
- HERMAN, E - JAMALI, H.R. - NICHOLAS, D. - OSIMO, D. - PORCU, F - PUJOL, L.(2015). Analysis of Emerging Reputation and Funding Mechanisms in the Context of Open Science 2.0, IPTS, Luxembourg: Publications Office of the European Union, 2015
- PISA DECLARATION On Policy Development for Grey Literature Resources (May 2014), <http://greyguide.isti.cnr.it/>
- SHEMA H, BAR-ILAN J, THELWALL M (2012) Research Blogs and the Discussion of Scholarly Information. PLoS ONE 7(5): e35869. doi:10.1371/journal.pone.0035869
- The Observatory of Responsible Research and Innovation (RRI) <http://observatory-rri.info>
- VON SCHOMBERG, Rene (2013). “A vision of responsible innovation”. In: R. Owen, M. Heintz and J Bessant (eds.) Responsible Innovation. London: John Wiley Available at: <https://renevonschomberg.wordpress.com/implementing-responsible-research-and-innovation/>
- WELLER, K. (2015). Social media and altmetrics: an overview of current alternative approaches to measuring scholarly impact. In: Welpel, I. M., WOLLERSHEIM, J., RINGELHAN, S., & OSTERLOH, M. (Eds). *Incentives and Performance*. Springer International Publishing. pp. 261-276

FIND THE PIECE THAT FITS YOUR PUZZLE



THE GREY LITERATURE REPORT FROM THE NEW YORK ACADEMY OF MEDICINE

Focused on health services research and selected public health topics, the Report delivers content from over 750 non-commercial publishers on a bi-monthly basis.

Report resources are selected and indexed by information professionals, and are searchable through the Academy Library's online catalog.

Let us help you put it all together; subscribe to the Grey Literature Report today!

For more information visit our website: www.greyliterature.org
or contact us at: greylithelp@nyam.org



**The New York
Academy of Medicine**

At the heart of urban health since 1847

Share #GreyLit: Using Social Media to Communicate Grey Literature

Danielle Aloia and Robin Naughton

New York Academy of Medicine, NYAM Library, United States

Abstract

Grey literature, publications that are not produced commercially, is becoming a common form of literature included in the systematic review process and called upon to influence policy- and decision-making. A major obstacle to finding this literature is the lack of a systematic way in which to search for current grey literature. Searching organization websites, bibliographies, and Google are just some of the methods used to find grey literature resources. In this research study, we looked at how social media can be an effective resource disseminating and finding information on grey literature. We found that many researchers disseminate their findings through social media, but rarely do they look for resources using social media.

Introduction

There is discussion in the scholarly community regarding researchers getting their work noticed beyond journal publishing. Grey literature, publications that are not produced commercially, and social media can be key resources for researchers seeking to disseminate their work. Digital scholarship is also being tapped to increase the recognition that social media is a valid scholarly tool. According to Rinaldi (2014), Social media is a way that researchers network with other researchers and gain public visibility. In addition, non-profits use social media for information sharing, community building, and action (Thackery 2013). In this way, researchers need strategic methods to get their work the attention it deserves and to find effective ways to use social media.

Grey literature has historically been used in decision- and policy-making, especially within governmental agencies. Often grey literature is research published by think tanks or non-profits in order to provide evidence and support for a specific policy agenda. Government technical reports of program evaluations on government-funded projects help to shape future policies. Sometimes these reports can be hundreds of pages long and contain political jargon. Often times, these reports are published along with fact sheets or executive summaries that can be easier to understand and less time consuming to read. Think tanks or research centers will publish fact sheets or summaries of these larger reports to distribute among their networks. When publishing as grey literature it is essential to include a dissemination process whereby the right report gets into the right hands (MacDonald, 2015).

To understand the dissemination of grey literature in social media, last year we researched the use of Twitter among health and health policy think tanks. The resulting poster, *Think Tanks, Twitter, and Grey Literature* (Aloia, 2014), was presented at the 16th International GreyNet Conference in Washington, DC. It was found that there is a lot of skepticism among researchers in using Twitter or social media effectively, conversely, we also found that there was a significant amount of evidence on using Twitter or social media effectively. Twitter is used to communicate scientific evidence, keep up with research in a specific field, share ideas, and gain visibility. Building on our Twitter research and adding to the growing importance of grey literature to the health sciences, this study examines the use of social media to communicate grey literature.

Research Questions

1. How is social media used to communicate grey literature?
2. To what extent are subscribers of the Grey Literature Report sharing resources found on greylit.org?

Literature Review

The literature review is based on research collected from combing the grey literature and PubMed with articles and grey literature reports focused on social media and the dissemination of research. Criteria for inclusion included evidence of social media use as a communication tool for research or statistical information on general use.

Call for Grey Literature in Systematic Reviews

The health sciences are calling for grey literature to be included in the systematic review process and as part of evidence-based practice. The value of the reported findings that might not have the opportunity to be published in peer-reviewed journals or in a timely manner can be published in other ways. More and more researchers are finding it easier to publish their research online through an organization or through a repository system.

It is also being called upon as an area of research that needs to be disseminated and communicated effectively to maximize its reach. Recently, at the conference *Advancing the Science to Improve Population Health* held at the National Academies of Sciences (Sep 2015), a presentation on Population Health Research Agenda Survey Results stated that raising the profile of “grey literature” was something respondents wanted. The AcademyHealth’s Translation and Dissemination Institute says that they are there to help “the field of health services research move its findings more effectively into policy and practice.” In 2014, the Institute offered two webinars focusing on the use of Twitter in disseminating and translating evidence as a way for researchers to promote and share their work.

State of Social Media

According to Pew Research, social media use has consistently grown in popularity and is up 66% over the past 10 years (Perrin, 2015). The Pew Research Center began systematically tracking social media usage in 2005 of 100,000 adults in the United States. Over the past ten years, Facebook (71%) has consistently been the leader with the most users, Twitter had 23% and LinkedIn had 28%. The proportion of users using Twitter on a daily basis decreased from 46% in 2013, to 36% in 2014, but weekly usage went up 3%.

In contrast, the worldwide survey *Digital, Social and Mobile in 2015* by We Are Social from 2015, found that there were 3 billion Internet users (a 22% increase from Jan 2014) in 2015 with over 2 billion on social media (a 12% increase from Jan 2014). Facebook is by far the most used with over 1.3 billion users. Google +, Skype and Instagram rank 7, 8, and 9 respectively. Twitter ranked 10 with 264M users and Tumblr ranked 11 on the list of 17 most used social media sites for January 2015. These numbers refer to the people who use social media, but how do organizations use it? We looked at two reports that follow corporate or nonprofit entities on social media and found that 78% of the Fortune 500 companies are using Twitter and 74% are using Facebook. (Barnes, 2015) This is in stark contrast to the general public’s use of both Twitter and Facebook. Interestingly, blogging is down 10% from last year, only 21% have a corporate blog (Barnes, 2015). Go-to-Think Tank ranks the top 40 best think tanks who use media (print or electronic) and the top 60 for best use of Social Networks (McGann, 2015). Not surprisingly, Pew Research Center ranked #1 with the best use of media. The Pew reports are cited multiple times in the press and shared through social media. “Social media and news reports are also used by more than 50% of (GreyLit) producing organizations to find an audience for their work” (Lawrence, 2014).

Researchers’ use of Social Media

Research on the use of social media in the health sciences fell under three producers: journal publishers, educators, and researchers. It was important to look at all three to make concrete recommendations since one producer may have more evidence than another. Journal publishers’ research on using social media analyze its impact factor on paper downloads and sometimes “the paper with the best performance in social media is also the top downloaded paper” (Mula, 2015). Educators are looking for criteria on the best ways to use social media in scholarship (Sherbino, 2015). Researchers use social media to communicate with colleagues, follow experts and trending topics, and disseminate information or study results. Some researchers use social media to communicate directly with journalists to discuss their work (Chapman, 2015; Wilkinson, 2013). The AcademyHealth report, *Health research and policy making in the social media sphere*, states that social media is a great way to disseminate information to the public and stakeholders whereby a dialog can be created between policy-makers, researchers and the public. Four factors for social media engagement are absorption (positive affectivity--personalized, how does it feel), self-expression and representation (how are you expressing yourself), empowerment (self-efficacy--are you getting likes and participating in online interactions), and interactivity (bridging ties vs. bonding ties). The authors point out that while there is an extreme amount of research on information-seeking behavior, there is a lack of research on health policy- and research-seeking behaviors. This type of research can help provide tools and evidence for researchers and organizations to effectively use social media to disseminate their works.

Amanda Lawrence’s seminal work on the value on grey literature in public policy (Lawrence, 2014), which was presented at 16th International Grey Literature Conference, adds to the research on health policy- and research-seeking behaviors. It was found that those surveyed (Australian researchers and policymakers) used some form of grey literature in their work, but one of the major issues was access to relevant sources. Access to relevant sources is a common problem with grey literature. It is interesting to note that though 50% of respondents used social media to disseminate their work, only 22% thought it was an important source for finding new information and a whopping 75% used websites, colleagues, newsletters and alert services to find new information (Lawrence, 2014).

Grey Literature Report

The Grey Literature Report produced by the New York Academy of Medicine has been a leader in curating, preserving and providing access to high quality, cutting edge research in the health sciences for researchers, health practitioners, and policy makers in urban and public health. The GreyLit Report has 2000 active subscribers. Thirty-five percent of subscribers are librarians and 15% are researchers, working in Academic or Library settings.

Methods

This research study is guided by a user-centered information behavior methodology with a focus on website usability. An online survey of active users of the Grey Literature Report was used to capture data on use of social media to disseminate grey literature.

Data Collection

All subscribers (n=2000) were sent an email regarding participation in the research study with a link to the online survey. When users clicked the link, they were taken to a Google form to provide consent and complete the survey. The survey was available for submission for five weeks beginning in mid-October 2015. Two reminder emails were sent to subscribers on November 15 and December 15. The survey consisted of 15 closed and open-ended questions. In December 2015, the survey was shared through various social media channels and listservs.

Data Analysis

Data collected from the online survey was analyzed using descriptive statistics and qualitative data analysis. The main goal of the analysis was to understand the use of social media to disseminate grey literature among GreyLit users.

We had a 4% response rate or eighty-five respondents. The first half of the survey contained demography questions about our audience and their use of social media.

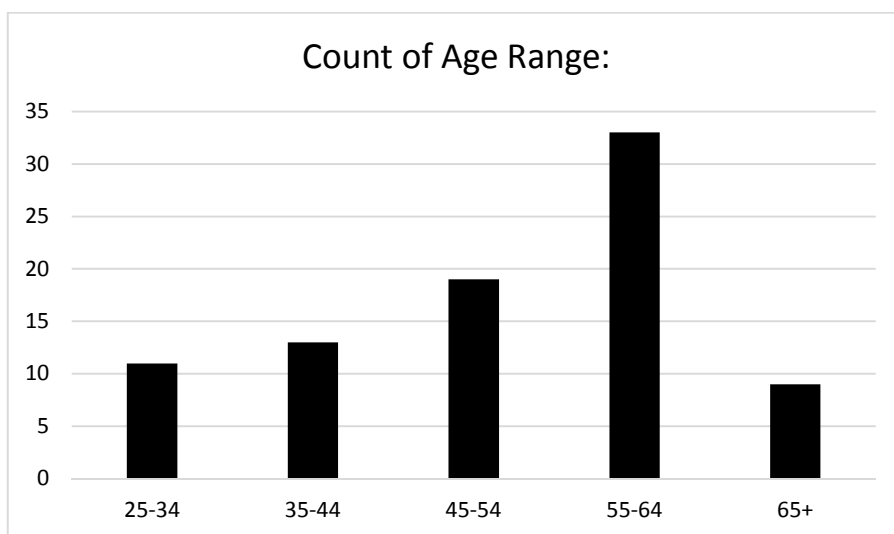


Figure 1: Age Range of Respondents

Forty-nine percent of our respondents identified as 55 or older and 23% as 45-54. Nationally, social media use has risen among all age groups, but particularly so for the 65+ from 2% in 2005 to 35% in 2015, a total of 86% of those 50 and older are using social media today. There were no respondents in the 18-24 age group, which is not consistent with the U.S. population in social media use. Ninety percent of social media users in the U.S. were 18-29 in the Pew study.

Count of Where do you work?

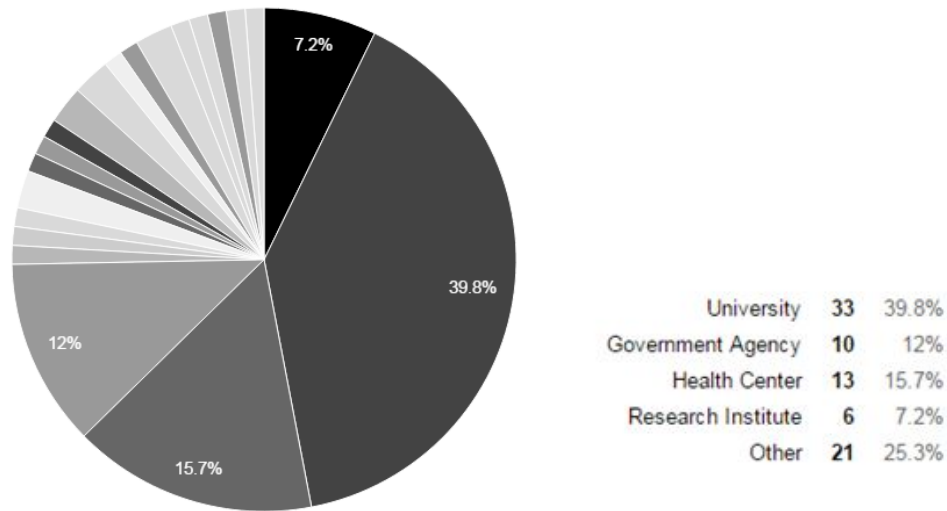


Figure 2: Employment

The majority of GreyLit respondents were librarians, working in a university setting. This coincides with the Pew study in that 76% of social media users were college graduates. (Perrin, 2015) Those with lower levels of education are less likely to be users of social media (Perrin, 2015).

How do you use social media?

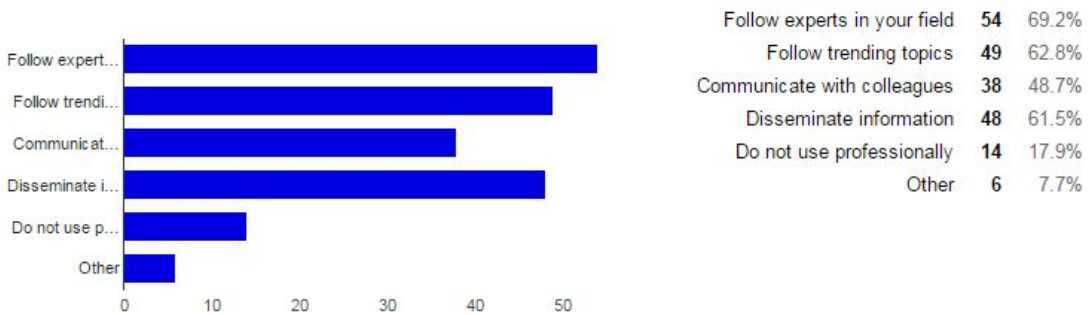


Figure 3: Use of Social Media

The three main reasons respondents used social media were to follow experts in the field, follow trending topics, and disseminate information.

What social media do you use to follow grey literature?

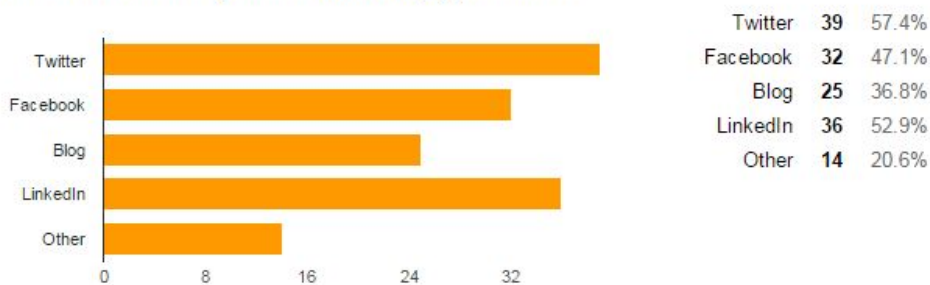


Figure 4: Social Media to find Grey Literature

A little more than half of respondents said they use Twitter or LinkedIn to follow grey literature, while over 61% said they disseminate information through social media.

The second half of the survey asked specific questions about the GreyLit Report and how users share the report.

How often do you visit greylit.org website?

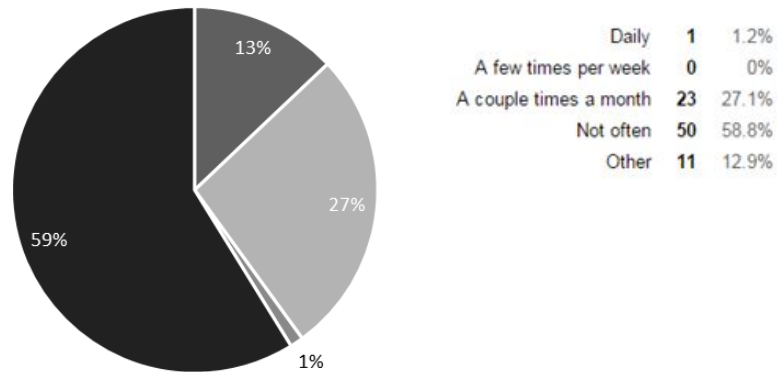


Figure 5: Visiting www.greylit.org

The www.greylit.org website is not visited often. Almost 60% of respondents reported that they do not visit the site often, while 27% said they visit a couple of times a month.

Count of Have you recommended the GreyLit Report to others?

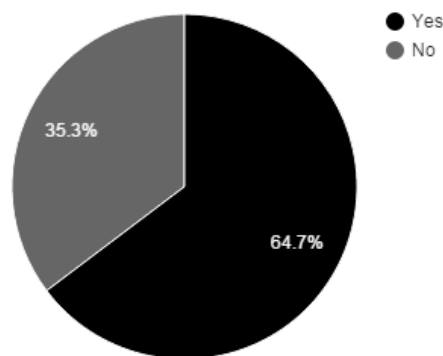


Figure 6: Recommend GreyLit Report

The majority of respondents (65%) indicated that they recommend our report to others, which can be evidence that they trust the Report.

Count of Have you liked any of our tweets or posts?

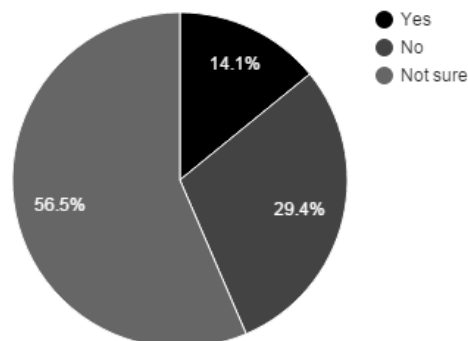


Figure 7: Like GreyLit Report Tweets

We also found that 57% were unsure whether they liked any of our tweets. This may seem odd but we do not have a direct social media channel. Our tweets are sent through @NYAMHistory or @NYAMNYC with #GreyLit. It is not unexpected that respondents weren't sure if they liked our tweets because having a single-issue focus social media presence generally have a higher following (Chapman, 2015) and trust.

Discussion

How is social media used to communicate grey literature?

While the majority of GreyLit users are older than the national average, they are highly educated and work in the research field which could mean that they are more savvy users of technology than their counterparts in other fields. The qualitative responses provided by the respondents were coded to represent the themes and topics of the research. The majority of respondents used Google or listservs and emails to find out about other grey literature. Responses to how users found grey literature included 30 email/newsletter/listserv, 34 website/search engine/RSS feed, 10 colleagues and 15 social media responses. Even though respondents indicated they followed grey literature on social media they tended to follow experts and trending topics more so. This does coincide with Lawrence's report that few of those surveyed used social media to find information (Lawrence 2014). This may be because there is no standard or evidence-base to using social media as an information source.

In order to engage users on social media, it is necessary to provide a "hook" to content that can inspire action. This can be done by framing a report or issue in the context of its social relevance to inspire users, be they journalists, researchers or policy makers, to take action. Having one topic of focus through social media channels is a good way to maintain a high following. Using social media with links to larger reports or studies will lead to increase of downloads of the report or study.

To what extent are subscribers of the Grey Literature Report sharing resources found on greylit.org?

Through this study we learned that few of our GreyLit subscribers are going to our website, but the majority are recommending the GreyLit Report to others. This suggests that the GreyLit Report is valued by our users even though the website may not be the first place our users go to access new grey literature. In addition, the research highlighted how our social media presence, Twitter in particular, may be improved to make it easy for our users to find grey literature.

Conclusion

A review of the literature found that the use of social media is becoming more and more a part of everyday life and work. There are many ways in which researchers are using social media in their daily work, from sharing their work with colleagues, directly contacting journalists, and communicating with their audience.

The literature review and the survey results both reflected the use of social media as a dissemination tool. Even though the literature review found that it was mostly used as a forum to exchange ideas, the importance of social media as a dissemination tool can still be warranted. What is not well understood is why users are not using social media to find information. Future research in this area will focus on learning why users are not using social media to find new grey literature, and exploring a social impact factor of downloaded reports and effects on public policy.

Bibliography

Allen HG, Stanton TR, Di Pietro, F, Moseley, GL. Social media release increases dissemination of original articles in the clinical pain sciences. *PLOS One*. 2013;8(7):e68914.

Barnes, NG, Lescault AM, Holmes, G. The 2015 Fortune 500 and Social Media: Instagram Gains, Blogs Lose. Dartmouth, MA: University of Massachusetts Dartmouth. 2015. Available at: <http://www.umassd.edu/cmr/socialmediaresearch/2015fortune500andsocialmedia/>

Chapman S, Freeman B. Who has Australia's most-followed Twitter accounts in health and medicine? *Public Health Res Pract*. 2015;25(3). doi:10.17061/phrp2531534.

Cosco TD. Medical journals, impact and social media: an ecological study of the Twittersphere. *CMAJ*. 2015 Dec 8;187(18):1353-7. doi: 10.1503/cmaj.150976.

Djuricich AM. Social media, evidence-based tweeting, and JCEHP. *J Contin Educ Health Prof*. 2014;34(4):202-204. doi:10.1002/chp.21250.

Duggan, M, Ellison NB, Lampe C, Lenhart, A, and Madden, M. *Social media update 2014*. 2015. Washington, DC; Pew Research Center.



- Grande, D, et. al. translating research for health policy: researchers' perceptions and use of social media. *Health Affairs*. 2014;33(7):1278-1285.
- Hays CA, Spiers JA, Paterson B. Opportunities and constraints in disseminating qualitative research in web 2.0 virtual environments. *Qual Health Res*. 2015;25(11):1576-1588. doi:10.1177/1049732315580556.
- Kapp JM1, Hensel B2, Schnoring KT2. Is Twitter a forum for disseminating research to health policy makers? *Ann Epidemiol*. 2015 Dec;25(12):883-7. doi: 10.1016/j.annepidem.2015.09.002. Epub 2015 Sep 14.
- Lawrence A, Houghton J, Thomas J, Weldon P. *Where is the evidence? Realising the value of grey literature for public policy and practice*. 2014. Melbourne, Australia: Swinburne Institute for Social Research. Available at: <http://apo.org.au/files/Resource/where-is-the-evidence-grey-literature-strategies-nov-2014.pdf>
- McDonald BH. et al. How information in grey literature informs policy and decision-making: a perspective on the need to understand the processes. *The Grey Journal*. 2015;11(1): 7-16.
- McGann JG. 2014 Global Go To Think Tank Index Report. Philadelphia, PA: 2015. Available at: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1008&context=think_tanks
- Mula M. The impact and dissemination of scientific research: from impact factor to social media. The Top 10 articles in *Epilepsy & Behavior* published in 2014. *Epilepsy Behav*. 2015;50:113-115. doi:10.1016/j.yebeh.2015.07.012.
- Nagendran M, Dimick JB. Disseminating research findings: preparing for Generation Y. *JAMA Surg*. 2014;149(7):629-630. doi:10.1001/jamasurg.2013.5019.
- Narayanaswami P, Gronseth G, Dubinsky R, et al. the impact of social media on dissemination and implementation of clinical practice guidelines: a longitudinal observational study. *J Med Internet Res*. 2015;17(8):e193. doi:10.2196/jmir.4414
- Nisbet MC. *Rethinking the translation and dissemination paradigm: recommendations from science communication research for health services policy debates*. 2015. Washington, DC: AcademyHealth. Available at: <http://www.academyhealth.org/files/FileDownloads/LessonsProjectScienceCommunication.pdf>
- Pei S, Muchnik L, Andrade JSJ, Zheng Z, Makse HA. Searching for superspreaders of information in real-world social media. *Sci Rep*. 2014;4:5547. doi:10.1038/srep05547.
- Pereira I, Cunningham AM, Moreau K, Sherbino J, Jalali A. Thou shalt not tweet unprofessionally: an appreciative inquiry into the professional use of social media. *Postgrad Med J*. 2015;91(1080):561-564. doi:10.1136/postgradmedj-2015-133353.
- Perrin A. *Social media usage: 2005-2015*. 2015. Washington, DC: Pew Research Center. Available at: <http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/>
- Rinaldi A. Spinning the web of open science: Social networks for scientists and data sharing, together with open access, promise to change the way research is conducted and communicated. *EMBO Rep*. 2014;15(4):342-346. doi:10.1002/embr.201438659. <http://embor.embopress.org/content/15/4/342.long>
- Sherbino J, Arora VM, Van Melle E, Rogers R, Frank JR, Holmboe ES. Criteria for social media-based scholarship in health professions education. *Postgrad Med J*. 2015;91(1080):551-555. doi:10.1136/postgradmedj-2015-133300.
- Smith BG, Smith SB. *Engaging health: health research and policymaking in the social media sphere*. 2015. Washington, DC: AcademyHealth. Available at: http://www.academyhealth.org/files/FileDownloads/AH_Translation%20Engaging%20Health%20report%20v5.pdf
- Thackeray R, Neiger BL, Burton SH, Thackeray CR. Analysis of the purpose of state health departments' tweets: information sharing, engagement, and action. *J Med Internet Res*. 2013;15(11):e255. doi:10.2196/jmir.3002.
- Tunnecliff J, et al. The acceptability among health researchers and clinicians of social media to translate research evidence to clinical practice: mixed-methods survey and interview study. *J Med Internet Res*. 2015 May 20;17(5):e119. doi: 10.2196/jmir.4347..
- Verhagen E, Bower C, Khan KM. How BJSM embraces the power of social media to disseminate research. *Br J Sports Med*. 2014;48(8):680-681. doi:10.1136/bjsports-2013-092780.
- We Are Social. *Digital, social and mobile in 2015*. London:We Are Social. Available at: <http://wearesocial.net/blog/2015/01/digital-social-mobile-worldwide-2015/>
- Wilkinson C, Weitkamp E. A case study in serendipity: environmental researchers' use of traditional and social media for dissemination. *PLoS One*. 2013;8(12):e84339. doi:10.1371/journal.pone.0084339.

Linked



GreyNet

Grey Literature Network Service

Public sharing of medical advice using social media: an analysis of Twitter

Gondy Leroy, Eller College of Management, University of Arizona;

Philip Harber, Zuckerman College of Public Health, University of Arizona;

Debra Revere, School of Public Health, University of Washington, United States

Abstract

Introduction: Social media tools, such as Facebook®, Twitter™, blogs and online communities, are increasingly utilized for networking and to distribute information in medicine and public health. Participation in these media has increased sharply over the past decade. Six years ago, Twitter did not exist yet now an estimated 15% of the world population subscribes to Twitter. This has created a large-scale, complex, and unindexed publicly available data source.

Goal: We sought to understand the richness and novelty of health-related Tweets by analyzing the characteristics of health information-focused tweets using automated and manual analysis.

Research methods: Utilizing the Twitter Search application programming interface (API) we retrieved two sets of English language tweets using keywords related to asthma (#asthma and asthma). Tweets were categorized by the assumed source (retweeted by a person, sent by organization, originated by an individual) and content (containing medication, symptoms, triggers, a combination, or none of these) using natural language processing. Regarding tweet source we assumed that tweets retweeted to a person (i.e., @username) were sent by an individual; those not retweeted that contained a URL were sent by an organization; and those tweets remaining were original content tweeted by an individual. Regarding content categorization, we used lexicons containing terms for asthma medication, symptoms, and five different types of asthma triggers (activities, air pollutants, allergens, environmental and irritants). In addition, we conducted content analysis using a combined text mining and manual approach. Applying association rule mining to the tweets, we generated an overview of the most frequency combination of terms presented as if-then rules. The manual, in-depth analysis evaluated a random sample of 200 tweets for originality, content, credibility and relevance.

Costs: The costs associated with this project were time to process tweets. While over 500 million tweets are generated daily, the cost of this information distribution is shared among millions of Twitter subscribers.

Results: The analysis showed that the majority of tweets contain URLs and many are retweeted. The proportion of tweets containing personal, new content is small. The majority of tweets are sent by organizations, both commercial and noncommercial, and the content are broad facts and statements. Both medication and environmental triggers are common topics.

Conclusion: The high diversity in topics and terminology combined with the small proportion of personal tweets should be taken into account when using Twitter as a resource for tracking and discovering health behaviors or problems in the population. The large proportion of tweets referring to external information may make this a very useful tool for accessing grey literature and using the tweets as descriptors. Further research is needed to create comprehensive vocabularies and methods to efficiently labels tweets.

Introduction

Social media tools are playing a greater and significant role in both clinical medicine and public health. Patients are using these tools to exchange information, advice and support; health care providers are distributing information, consulting with other providers, and interacting with patients through social media; numerous commercial vendors and professional organizations promote products and viewpoints using these tools; and public health authorities both disseminate and acquire information via social networks.¹ Many of these social media platforms also generate data that is made available to research and industry data users. For example, Twitter™, Facebook®, LinkedIn™ and Google+ each provide an application programming interface (API) to access this data with some restrictions. The popularity of social media can be illustrated by Twitter, in which an estimated 15% of the world population is subscribed (both active and inactive).² This popularity has created newly available large-scale data sources that are characterized by a high volume of data, high variety in data elements, high velocity of the data stream, and potential veracity issues (i.e., the objectivity and validity of the data content cannot be presumed).

Social media tools provide value and entertainment to their members and users, as well as new opportunities for consumers, industry, government and researchers to post and exchange

information and gather feedback. For example, the Twitter account owned by the United States' National Institutes of Health (NIH) has sent several thousands of tweets (about 5,200 in October 2015) to its more than 610,000 followers. Twitter is seen as the "zeitgeist" in numerous communities,³ for example, providing a platform to discuss issues surrounding vaccines, as well as tweet specific drug information and side effects.⁴ Participants can rapidly switch between the roles of content consumer/audience and content generator. Social media traffic is often ruled by the crowd and popularity may trump the truth⁵; as user-generated content increases its quality is questionable given the ease of online deception.⁶

Information distributed in social media platforms, such as wikis and social networking sites, Twitter, listserv archives, blogs, podcasts (audio or video), and other forms of electronic networked communication, are considered a new form of "grey literature". Traditionally, grey literature was an umbrella term used to describe materials such as government reports, working papers, and evaluations that were not indexed by major databases, controlled by commercial publishers, or peer reviewed. Variability in quality and limitations on availability are other characteristics associated with grey literature. Despite these features, grey literature is considered an important source of information for research as it is sometimes the only source of information for a research question.⁷

Twitter as Information Source

The motivations to participate in a social network such as Twitter may vary widely, although the primary intentions are to share information, engage socially, and communicate personal opinions and critiques. While not all followers of an individual, organization or topic on Twitter post tweets, social media is unlike other passive, communication mediums as the recipients of information can actively engage with the information. Commercial entities use Twitter to distribute news and updates on problems or policy. Non-commercial, government and NGO/non-profit organizations have adopted Twitter to similarly distribute information, alerts and updates. The variety of tweets is wide, ranging from tweeting personal comments on social issues, public health agencies issuing alerts and advisories, celebrities sharing personal information with their fans, the Red Cross tracking Twitter posts during hurricanes to gather information about where the greatest needs are, political leaders hold town hall gatherings with constituents, "citizen activists" discussing issues, and entities such as news agencies requesting comments and feedback using Twitter.¹

Market researchers are now using Twitter to identify the demographic characteristics and audience segmentation related to products and commercial services, as well as to disseminate promotions and product information. Scientific research on Twitter is growing, along with the size of the data source, with the majority of research focusing on tweet quantity or numbers in relation to specific topics and their relationships with other information sources. For example, Ram et al.⁸ leveraged social media (Twitter and Google) and environmental sensor data to accurately predict the number of asthma-related emergency department visits with approximately 70% precision. While retrospective, another example is work conducted by Odium and Yoon (2015)⁹ mining tweets to inform public health education regarding Ebola. The researchers analyzed over 42,000 Ebola-related tweets posted in the early stage of the Ebola outbreak (between July 24-August 1, 2014) and found that nearly 1,500 tweets regarding Ebola were disseminated prior to official announcements from the Nigerian Ministry of Health, the World Health Organization and the Centers for Disease Control and Prevention. Other research using Twitter has considered content of tweets. For example, Harris et al (2014) coded the content of one month of #childhoodobesity tweets (n=1110) distributed in June 2013. Each tweeter was classified by type of user (individual, organization, etc), health focus, and sector (individual, government, for profit, etc). Each tweet was coded as original or retweeted (i.e., the forwarding of tweets received by one user to their own personal social network). The study reported and overall credible content in the tweets, suggesting a limited presence of government, media, and educational information sources and a need to focus the content of tweets on scientific evidence.¹⁰ Another area of research on Twitter is consideration of hashtags in addition to counts and content. Hashtags in social media provide a folksonomy, or user-driven classification, that can be especially effective in placing boundaries around a health topic by providing a common self-organizing language that defines that topic. In addition to organizing content, hashtags can aid in building communities around a shared topic of interest.¹¹

As a newcomer under the grey literature umbrella of materials, understanding how Twitter is related to other grey literature and information sources is a new endeavor. We report our investigation which takes a comprehensive meta-approach—including characteristics of source,

content, and origin and utilizing natural language processing, text mining, and manual approaches—to evaluate the content of tweets and the website to which they refer to understand the richness and novelty of health-related Tweets.

Data Collection

We selected asthma as our health topic because it is a common chronic disease, affecting 8.4% of the U.S. population¹², affects all age groups, and varies in severity ranging from mild to fatal. We used the Twitter Search API (application programming interface), which allows programmatic access by specifying queries and a few additional parameters such as the language and geolocation. The (free) Twitter search API collects tweets approximately 7 days prior to the search date and retrieves a maximum of 100 tweets per query and 180 queries within a specific window (currently 15 minutes). Thus, one researcher can retrieve 18,000 tweets every 15 minutes. We used two queries: (a) a hash tag specific query including the term “#asthma “; (b) a broad query for tweets including the term “asthma”. Both queries were submitted every two weeks to avoid overlap and duplicate tweets.

After retrieving the tweet data set (in JavaScript Object Notation (JSON) format), we processed the tweets using a system that combines in-house developed Java code in combination with existing resources provided by General Architecture for Text Engineering¹³ (GATE) and its social media libraries.¹⁴ GATE facilitates the creation of customized components for text processing such as our two Java Annotation Pattern Engines (JAPES). The first JAPE aimed to recognize the tweet types and assign one of three labels: tweets that were retweeted at a person (started with RT @...) which is duplicated information, tweets containing URLs, and tweets without URLs that were personal comments (given no reference to additional information). The second JAPE aimed to recognize content and mapped words in the tweet’s text or hashtags to small lexicons (i.e., gazetteers) created by a medical expert and containing commonly used terms for medications, symptoms and triggers of asthma (see Table 1). When matches were found, the tweet was assigned labels indicating the presence of these terms which allowed the tweets to be categorized by this content.

Type	Subtype	Nr Terms	Example Terms
Medication		104	Metaproterenol, Xolair, ...
Symptoms		58	Chronic coughing, wheezing, ...
Triggers	Activities	25	Exerting, smoking, ...
	Air Pollutants	9	Diesel, ozone, ...
	Allergens	28	Animal dander, mold, ...
	Environmental / Occupational	31	Glues, adhesives, exhaust, ...
	Irritants	17	Deodorant, fragrance, ...

Table 1: Content of In-house Developed Gazetteers

Tweets were processed in the following steps:

- Tweet Tokenizer: to divide tweets into tokens.
- Hashtag Tokenizer: to split hashtags containing multiple words into their individual components, e.g., #healthcommunication is recognized as containing ‘health’ and ‘communication’.
- Normaliser: to normalize text in tweets, e.g., abbreviations or slang are rewritten in proper English
- Tweet POSagger: to label each term with its Part-of-Speech (POS), e.g., noun, verb.
- JAPE Type Labeling (in-house): to label tweets containing a URL, retweets with a URL, or tweets without a URL.
- JAPE Content Labeling (in-house): to label a tweet as containing medication, symptoms, triggers, a combination of the three, or none at all.

Research Methodology

We conducted three sets of analyses to evaluate the tweets:

1. Frequency analyses provided an automated overview of tweets using type and content labels.
2. Descriptive data mining (association rule mining) was used to show relationships between individual concepts in tweets. Association rule mining is a well-known data mining technique used to discover interesting associations that are not yet common knowledge. The technique^{15,16} combines conditional probabilities optimized for item sets. An item set is a set

of elements that belong together, e.g. nouns in a tweet. While it does not aim to predict labels, association rule mining will find associations between items and present them as IF-THEN rules (e.g., air pollution → asthma attack). The rules are generated by the a priori algorithm which uses support to create frequent item sets and confidence to transform frequent item sets into rules. For a rule such as $X \rightarrow Y$:

- Support is defined as the percentage of item sets (tweets) that contain the elements X and Y
- Confidence is defined as the conditional probability: support (X and Y) / support (X)

3. In-depth, manual analysis provided a detailed look at a random sample of tweets. Each tweet was scored along four axes which were created based upon manual review of a pilot sample of tweets. Our intent was to use sufficiently fine-tuned labels to support different definitions of grey literature either by collapsing categories or keeping them fine-grained. The axes are summarized in the Table 2. The first two axes, Tweet Origin and Breadth of Message Topic, focus on labeling the tweet text itself to reflect the information available with an incoming tweet as seen by a reader. The latter two axes, Referenced Item Credibility & Authority and Asthma Content Specificity, focus on labeling the content of sites referenced in the tweet.

A sample of 200 tweets was randomly selected for coding by the authors. Each was coded by two of the authors. A single choice was allowed for axes 1-3, and multiple selections were permitted for axis 4.

Axis	Level	Example
Axis 1. Tweet Origin	Person	
	organization-noncommercial	
	organization-commercial	
	Retweet	
	Unknown	
Axis 2. Breadth of Message Topic	Level 1: Personal Only	
	Level 2: Time and/or Location-limited Content	
	Level 3: General Fact Applied to a Specific Time or Location, but not population	
	Level 4: General Fact that can be Population-Specific	
Axis 3. Referenced Item Credibility & Authority	Personal	blogs, personal video, personal website, social networking sites (Facebook, PatientsLikeMe, CleanAirMoms)
	News	Newspapers, news services, magazines
	Commercial	Commercial YouTube videos, advertising, businesses
	Nongovernmental organization	WebMD, Universities
	Governmental organization	NIH, CDC, EPA, Local and State Public Health Agencies
	Professional Associations/Organizations	Medical professionals
Axis 4. Asthma Content Specificity	Treatment	1) Medication: inhalers, drugs, 2) Non- pharmacologic: keeping the house clean? mold deterrents? 3) Alternative/complementary: acupuncture, yoga, 4) Treatment/NOS: e.g., health system access
	Triggers	1) home & community: e.g., dogs, pets, 2) Environmental: e.g., air pollution, community violence, 3) occupational, 4) personal or behavioral : e.g., stress
	Other	1) Symptoms: e.g., wheezing ..., 2) Personal health consequences: e.g., hospitalization, Dr. visit, missed school,..., 3) Personal Risk factors: e.g., smoking, exposure to triggers , other conditions/co-morbidities (hay fever, rhinitis), 4) Diagnostic tests: e.g., spirometry, 5) Support & Management, 6) Not Asthma Specific

Table 2: Axis Definitions

Results

Tweets were collected from June 2015 - July 2015 and included 16,427 and 72,000 tweets retrieved by Twitter in response to the two queries: #asthma and asthma. The query including the hashtag (#asthma) retrieved approximately 4,000 tweets every two weeks. Although the search API was used, which searches up to seven days back, the maximum allowed by the query (18,000) was not reached. In contrast, searching for “asthma” retrieved 18,000 tweets for each query. While more tweets might have been collected by querying over additional time windows, to ensure higher diversity (and fewer retweets of the same information) we limited our collection to one query every two weeks. Please see the appendix for a list of example tweets.

1. Frequency Analyses

Type Analysis

Figure 1 shows an overview of the results. For the specific tweets (#asthma), most contain a link to an outside source (59%) and retweets were the second most common source (29%). Only a small percentage (12%) of the tweets does not contain links to any outside information. In contrast, the broader query shows an almost equal distribution of tweets containing URLs (30%), retweets (33%), and the remaining group (37%).

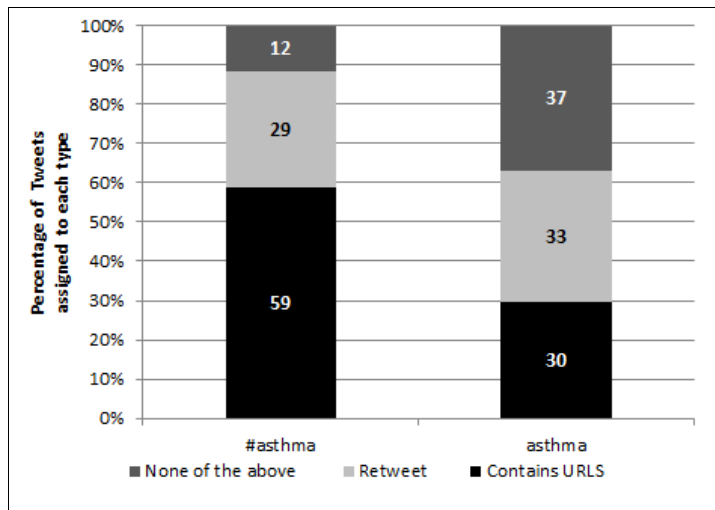


Figure 1: Percentage of Tweets by Source

Content Analysis

Figure 2 illustrates results of the content analysis, showing the proportion containing specific medication, symptoms or triggers. The number of tweets identified as containing this information was low (89% and 79% for #asthma and asthma), and most tweets contained more high level information regardless of the query: approximately 10% from the #asthma query and approximately 20% for the broader “asthma” query contains specific information. The more general asthma query contains more mentions of symptoms (12%).

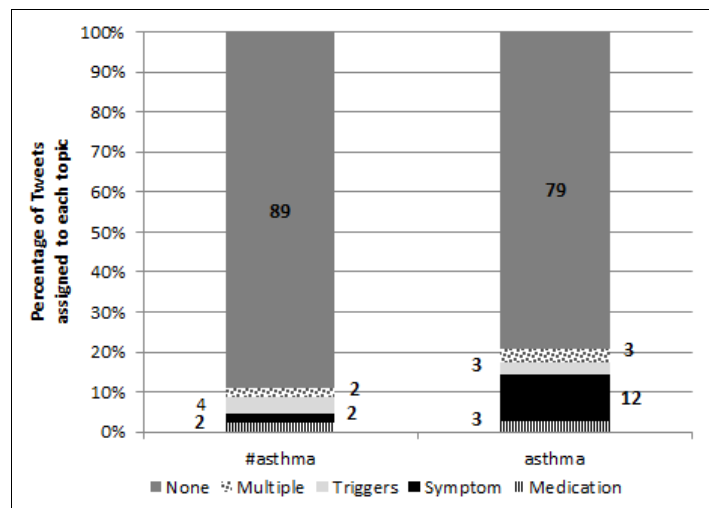


Figure 2: Percentage of Tweets by Content

2. Relationship Mining (Automated Content Analysis)

For both queries, there was very limited commonality of terms among tweets, and therefore the association mining algorithm was set to use a low support setting of 1% (i.e., generate all rules containing two or more items that appear together in $\geq 1\%$ of the tweets (≥ 164 tweets for #asthma and ≥ 720 'asthma')). We identified rules with calculated confidence $\geq 80\%$. We conducted the analysis for each query twice: using only nouns and once using all words in the text. For both, preliminary analyses showed the need for preprocessing and removal of proper nouns (e.g., Matt, Lopez, askdrnick, ...), as well as pronouns (e.g., yall), common verbs (e.g., wanna gotta, shudnt), interjections (e.g., Okay, ugh, hey), references to time (e.g., yesterday, tomorrow), and numbers (e.g., ten, thousand).

Table 3 shows a sample of the association rules that were generated. Few rules were created using only nouns. There were no rules for the #asthma query generated from 753 unique nouns in the 16,427 tweets. There was one rule generated for the asthma query from the 610 unique nouns in 72,000 tweets.

Using the unrestricted term set generated many rules using the same settings (support $\geq 1\%$, confidence $\geq .80$). For the #asthma query, 73 rules are generated based on 870 unique terms. The top rules reflect that when people tweeted about alert and symptoms in this set, they also include asthma. Tweets about air quality were also frequent.

Query & Terms	Unique, Stemmed Terms (N)	Antecedent	Consequent	Support (%)	Confidence (%)
#asthma					
Nouns Only	753	No rules found with $\geq 1\%$ support			
All terms	870	73 rules found with $\geq 1\%$ support			
		alert & symptom	asthma	5.5	100.0
		advic	asthma	5.5	100
		alert	asthma	5.8	99.5
		obes	asthma	1.5	99.4
	
		copd	asthma	2.9	81.4
		qualiti & air	asthma	1.2	80.6
Asthma					
Nouns Only	610	1 rule found with $\geq 1\%$ support			
		busi	attack	1.3	93.1
All terms	925	149 rules found with $\geq 1\%$ support			
		mind & continu	busi	1.5	100
	
		mind & busi & continu & attack	asthma	1.5	100
	
		hogti & die	asthma	1.1	99.0
	
		end & unsaf & prescrib	asthma	1.0	97.2
	
		polic	hogti	1.1	81.1

Table 3: Overview of Top Association Rules

Some rules were created based on tweets reflecting periods of high social commentary on Twitter. They are of limited relevance to the medical topic asthma. For example, rules were generated for 'hogti' reflecting an incident where "A Tennessee man with asthma died after being hogtied and placed face down on a stretcher by police following a Widespread Paniclosif Britton".

3. Manual Content Analysis

To optimize the manual coding scheme and calibrate the coders, we first completed two rounds of pilot testing of 20 tweets. Results were shared and differences discussed. Then a random sample of 200 tweets from the #asthma set was manually coded by the authors independently. (GL coded all 200 and the other co-authors each coded 100). Thus, each item was coded by two individuals.

Axis 1: Tweet Origin

As shown in Figure 3, retweets were most common (approximately 35%) and easily identified by the symbol “RT”. Some of the others allowed a straightforward classification (for example, “bbcbreakfast: tens of thousands of people with #asthma in the uk are not getting the right medicines <http://t.co/PxFJLIQyK>” is easily recognized as a commercial organization). Others were not as direct but still classified into the same groups by the raters. For example, “35 worst cities (and the worst state) for asthma and allergy sufferers <http://t.co/RtOKo62IBA> #asthma #allergys #usethermapureprecision env.” was thought to originate from a commercial company according to both raters, while “less ambitious #airpollution levels will result in more deaths and people suffering from respiratory symptoms <http://t.co/3qINNd7fnL> #asthma” was considered most likely to be a personal tweet by both raters.

However, for many, the tweet content did not in itself support easily deciphering the source. Many tweets required speculation and can be argued to belong to a different group than the one assigned. For example, “the baby food is organic; the shots are not <http://t.co/zy7ftmebjs> @assemblydems @assemblygop no #sb277 #vaccines #autism #asthmaj dub”

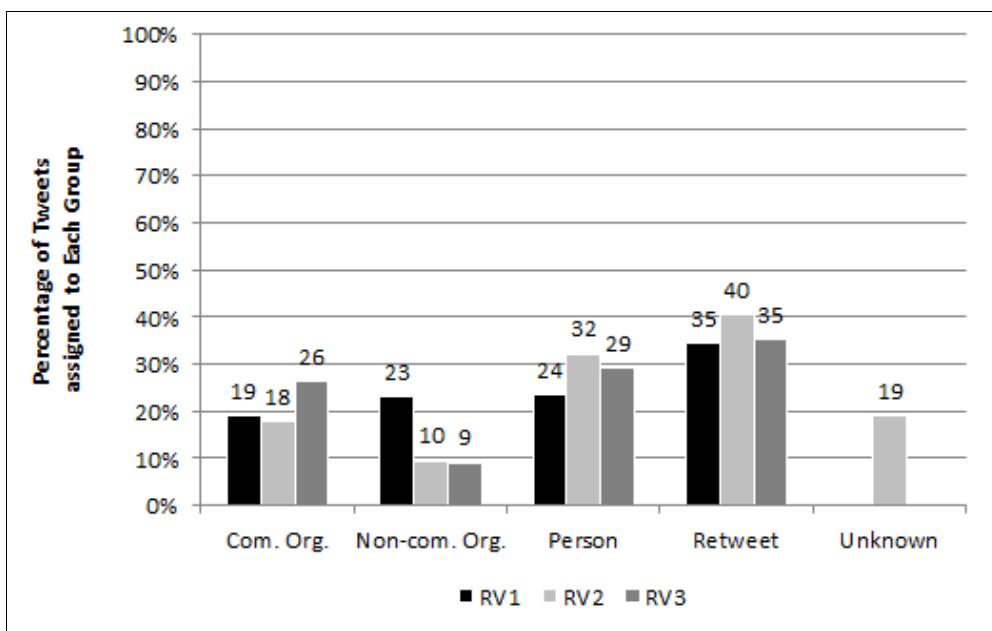


Figure 3: Original of Tweet

Axis 2: Breadth of Message Topic

Table 4 explains the coding scheme for axis 2, and Figure 4 shows results for the tweets categorized by breadth of the topic. Level 4 (broadly applicable) was most common. Level 2 and 3 were the least common. The three evaluators’ labels differed most for level 1 (personal information). For many, classification was ambiguous.

Level	Example	Explanation
4	“asthma treatments fail older patients more often: study http://t.co/oiYXgyu1GC #asthma #geriatrics”	Applies to all asthma patients. Implies underlying mechanistic finding (e.g., effect of age)
3.	uk has one of highest rates of #asthma prevalence and mortality @asthmauk raising awareness http://t.co/IUDkNURcCw .”	Of general interest to a broad group, but not universally
2.	rt @asthmaaus: very important reminder with all the colds and flu around #asthma https://t.co/xo5EGSl6vt asthma foundation sa”	Only applies to a limited time or specific place
1.	“stupid #asthma. i can't breathe. ??manilowese ?”	Describes the condition of only the tweeter at the moment

Table 4: Example Tweets for Breadth Levels

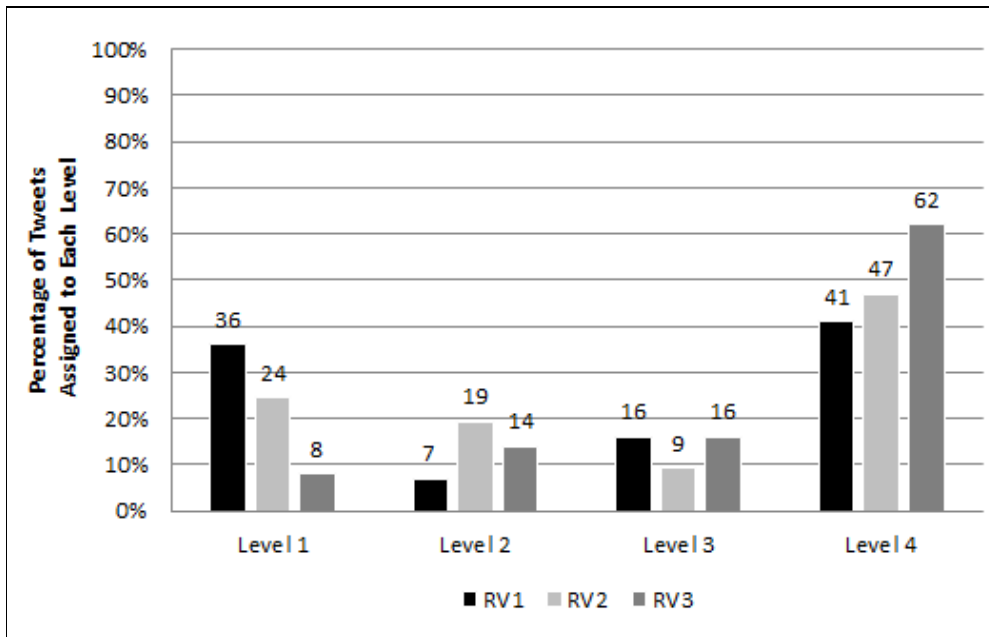


Figure 4: Breadth of Message Topic

Axis 3: Referenced Item Credibility and Authority

For this axis, we considered both the content of the tweet and the Internet page associated with the first URL in the tweet when available. Figure 5 summarizes estimated credibility. Personal and news items were predominant. Commercial organizations constituted 20-25%. Conversely, governmental and nongovernmental organizations contributed relatively few tweets.

There were several sources of ambiguity. In some cases, more than one type of authority was included (e.g., news combined with an organization). Furthermore, we limited the information to only reading the first page linked to the link, and some pages are not easily categorized. For example, “*how to prevent #asthma at your own #home <http://t.co/s3HC0oQaNg> weirdscience*” and the referenced page were seen as news and NGO. Others tweets require expert knowledge in assessing the source credibility. For example, in the tweet “*#Asthma Pittsburgh Lung Institute Grand Opening <http://t.co/UVQ6CSrrWd> #COPD*” both the tweet and the linked page appear as credible information from a respected research institute. However, the “Institute” is a commercial entity, and the proposed method of stem cell infusion is considered highly suspect as documented by editorials in a major United States medical journal. Thus, the tweet and the webpage may have been intentionally designed to masquerade as a respected scientific institution.

The largest group tweets was rated as having ‘personal’ credibility, such as for example “*i guess my lungs missed you... a lot. #asthmaproblems #asthma #seretide #inhaler #gsk <https://t.co/8YHCtq7p1K> the fluffiest mamon*”; the URL page shows an image of a metered dose inhaler. The smallest set contained tweets sent by professional individuals, such as for example “*stop #migraine with the classical buteyko method <http://t.co/QPQVAwF7s7> #asthma #copd #chronic #illnesslearnbuteykoonlinen*”. Tweets with credibility level due an NGO, government and professional organizations were about equally prevalent (about 10% of the tweets).

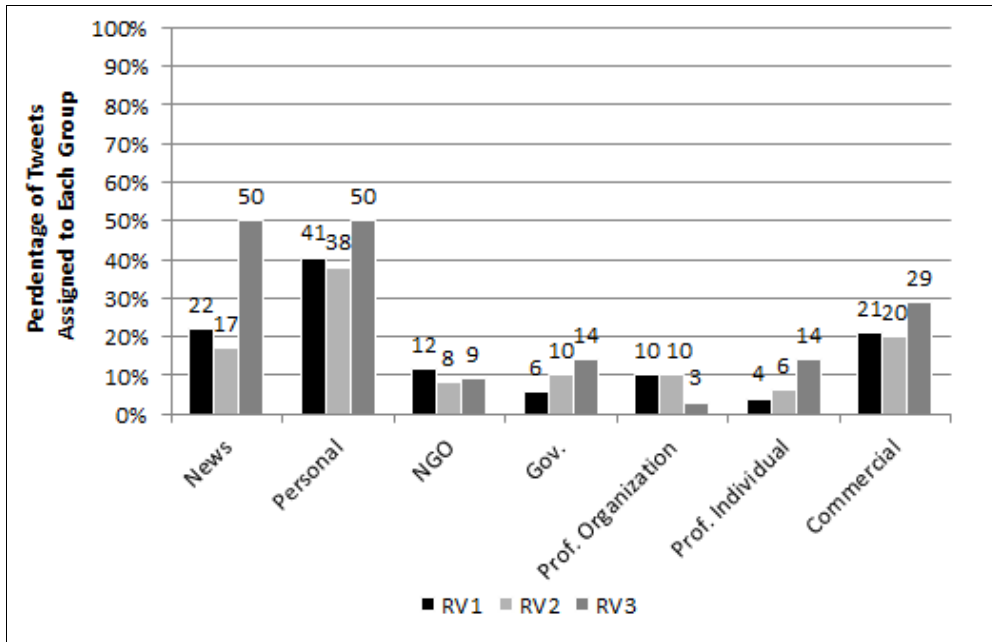


Figure 5: Referenced Item Credibility and Authority

Axis 4: Asthma Content Specificity

Assessment of the asthma related content is summarized in Figure 6. More than a single content item could be assigned to an individual tweet. The asthma related content was diverse. In addition to treatment items (medical, pharmacologic, and alternative), environmental and other triggers were frequently mentioned. For example, “rt @s_rattigan: thank you rep @matmuratore for helping to spread the word about #asthma and cleaning products! @massdph <http://t.co/lh1ei2n>” contains information about triggers (cleaning products) while “q3 - #asthma is often being treated as an acute condition. do you think one annual assessment is enough or acceptable? #arnsarns” discusses treatments.

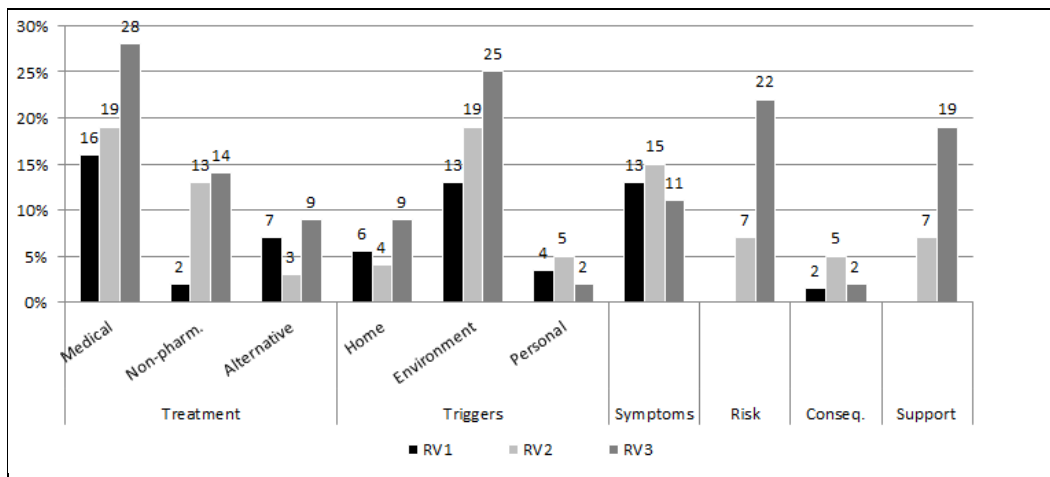


Figure 6: Content Specificity

Conclusion

Origins: The algorithm described herein appears to be a useful method to enhance the utility of public health monitoring of tweets. The automated and manual methods gave reasonably concordant results.

Use of media such as Twitter has been proposed as a public health monitor of population health status (e.g., to identify epidemics or geographic clusters). However, both the automated and manual assessments show that only a minority of tweets are generated by individuals. Therefore, to meaningfully reflect conditions of individuals, it may be advisable to restrict analyses to tweets operationally identified as not coming from organizational sources

(commercial or governmental). Further empirical study is necessary to determine whether retweets should be considered useful or only those generated by individuals.

The majority of tweets originate from organizations with the Tweet reflecting some of the contents, e.g., a summary or title containing the main point of the linked information. Further study may show how the tweets may act as an indexing mechanism for the linked information.

Content analysis: The expert assessment of a random sample of the #asthma tweets demonstrates that the public has a wide range of concerns about asthma. Notably, they are concerned about prevention of episodes as demonstrated by the frequency of references to triggers and the influence of the environment on asthma.

The results of automated and manual content analyses were discordant (see figures 1 and 6). While automated content assessment identified medication and symptoms as the predominant elements, the expert manual assessment showed a more diverse content. This likely indicates that the range of terms for treatments and triggers is much greater than for asthma related symptoms. Automated analysis using association rules further demonstrated the wide range of topics and terminology used. The terms and topics of tweets were diverse, with limited overlap unless they were retweeted. Our study therefore suggests the need to create an extensive lexicon for triggers, treatments, and consequences. Such resources will be beneficial to automated tracking and analysis. Lexical analysis and expert annotation of a large number of tweets may provide an empirical basis for improving automated content interpretation in the future.


References

1. 2014;39(7):491-520. VCPT. Social media and health care professionals: benefits, risks, and best practices. *Pharmacy & Therapeutics* 2014;39:491-520.
2. Zun LS, Sadoun T, Downey L. English-language competency of self-declared English-speaking Hispanic patients using written tests of health literacy. *Journal of the National Medical Association* 2006;98:912-7.
3. Cheong M, Lee V. Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base. In: 2nd ACM workshop on Social Web Search and Mining. Hong Kong, China: ACM; 2009.
4. Bian J, Topaloglu U, Yu F. Towards Large-scale Twitter Mining for Drug-related Adverse Events. In: International Workshop on Smart Health and Wellbeing (SHB 2012). Maui, Hawaii, USA; 2012:25-32.
5. Keen A. *The Cult of the Amateur: Doubleday Currency*; 2007.
6. Tsikerdekis M, Zeadally S. Online Deception in Social Media. *Communications of the ACM* 2014;57:72-80.
7. Lawrence A, Houghton J, Thomas J, Weldon P. Where is the evidence: realising the value of grey literature for public policy and practice. (Discussion Paper). In: Australian Policy Online. Melbourne AUS: Swinburne Inst for Social Research; 2014.
8. Ram S, Zhang W, Williams M, Pengetnze Y, 2015;19(4):1216-23. Predicting Asthma-Related Emergency Department Visits Using Big Data. *IEEE J Biomed Health Inform* 2015;19:1216-23.
9. Odium M, Yoon S. What can we learn about the Ebola outbreak from tweets? *Am J Infect Control* 2015;43:563-71.
10. Harris J, Moreland-Russell S, Tabak R, Ruhr L, Maier R. Communication about childhood obesity on Twitter. *Am J Public Health* 2014;104:e62-9.
11. Ferragina P, Piccinno F, Santoro R. On analyzing hashtags in Twitter. In: 9th Intl AAAI Conf Web Social Media (ICWSM 2015). Oxford England; 2015:110-9.
12. Moorman JE, Akinbami LJ, Bailey CM, et al. National surveillance of asthma: United States, 2001-2010. *Vital Health Stat* 3 2012:1-58.
13. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Comput Biol* 2013;9:e1002854.
14. Bontcheva K, Derczynski L, Funk A, Greenwood MA, Maynard D, the NAPo. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In: International Conference on Recent Advances in Natural Language Processing (RANLP 2013). Hissar, Bulgaria; 2013.
15. Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD International Conference on Management of Data; 1993; 1993.
16. Agrawal R, Srikant R. Mining Sequential Patterns. In: Eleventh International Conference on Data Engineering; 1995: IEEE; 1995. p. 3-15.

Appendix

Tweet numbers 1-3 and 13-15 are tweets that are retweeted at a person. They are duplicates of other earlier tweets as is shown with tweet number five. Tweets 4-6 and 10-12 contain URLs and demonstrate how the reader is enticed to click on the link for further information. Finally, the remaining tweets 7-9 and 16-18 show tweets that are general more personal in nature.

Nr	Query	Type	Tweet Text
1	#asthma	Retweet@	RT @MyID_Research: Hello @luckytilldeathx I read your post. I would like to discuss more about #asthma #medicalresearch https://t.co/ndKp03 ...Zimba
2	#asthma	Retweet@	RT @ShelleyWebbRN: Please RT! New #Asthma #ClinicalTrial available for those suffering from mild or moderate asthma: http://t.co/KnNq8XjZZtaudrey mcevoy
3	#asthma	Retweet@	RT @TheWiseAngel: I'm ready to receive miracles. #asthma #myastheniagravis #Chronicillness #ALS #CancerShaun... ? @yusalife
4	#asthma	With URL	Reasons Why Soy Wax Is Better For Your Home http://t.co/4jy2Di7xA3#asthma #candles #soycandles #environment http://t.co/k5VDS4LeIE Holly
5	#asthma	With URL	@RMEGY thanks for the post. I would like to discuss more about #asthma #medicalresearch https://t.co/ndKp03UQwX Matt Poulton
6	#asthma	With URL	Study Uncovers #Mechanism Responsible for #pollen -induced #Allergies #animalresearch #asthma http://t.co/VJep9XSBC Raemdoncke
7	#asthma	Other	I'm ready to receive miracles. #asthma #myastheniagravis #Chronicillness #ALS #CancerJulie Ann
8	#asthma	Other	Exercise can trigger bronchoconstriction in both people with and without #asthma.No More Asthma
9	#asthma	Other	I think all of this #BlackCarbon #AirPollution is making my #Asthma worse! How do you feel?WE ACT Aethalometer
10	asthma	With URL	http://t.co/ZuTfv1cmaP symptoms of Asthma & alerts,as of July 24, 2015 at 03:17PM. #AsthmaExpat Inc
11	asthma	With URL	RT @jusa_life: #LetsClearTheAir Secondhand smoke exposure causes more than 202,000 asthma episodes each year. http://t.co/oKWdLYyk4KT
12	asthma	With URL	RT @ViewFromTheHook: Coalition for Healthy Ports Blasts @EPA / @PANYNJ Absurd New Diesel Emissions Regulations http://t.co/JV9JzeYok7 #dies...CleanWaterAction NL
13	asthma	Retweet@	RT @AEA: Five Charts That Blow Apart EPA's Asthma Claims http://t.co/Qp6mBijNpf http://t.co/JOaGM88xbq Mary Jeff Dunlap
14	asthma	Retweet@	RT @YouStalkSam: Tulisa's gay best friend is meant to be her body guard lol? If I boxed him in his chest he would download asthma.
15	asthma	Retweet@	
16	asthma	Other	@spiceloft Hi there! We saw your post about asthma and would love to feature you on our Faces of Asthma page!Control A+
17	asthma	Other	pulling the muscles between your ribs and having bad asthma is not a great combination #inpainruthie
18	asthma	Other	I AM BRINGING LIKE 6 INHALERS BC I DON'T DON'T TO HAVE AN ASTHMA ATTACK AND DIEFUCKING TODAY



Slovak Centre of Scientific and Technical Information **SCSTI**

Achieve
your goals
with us



INFORMATION SUPPORT OF SLOVAK SCIENCE

SCIENTIFIC LIBRARY AND INFORMATION SERVICES

- technology and selected areas of natural and economic sciences
- electronic information sources and remote access
- depository library of OECD, EBRD and WIPO

SUPPORT IN MANAGEMENT AND EVALUATION OF SCIENCE

- Central Registry of Publication Activities
- Central Registry of Art Works and Performance
- Central Registry of Theses and Dissertations and Antiplagiarism system
- Central information portal for research, development and innovation - CIP RDI >>>
- Slovak Current Research Information System

SUPPORT OF TECHNOLOGY TRANSFER

- Technology Transfer Centre at SCSTI
- PATLIB centre

POPULARISATION OF SCIENCE AND TECHNOLOGY

- National Centre for Popularisation of Science and Technology in Society

IMPLEMENTATION OF PROJECTS

- National Information System Promoting Research and Development in Slovakia - Access to electronic information resources - NISPEZ
- Infrastructure for Research and Development - the Data Centre for Research and Development - DC VaV
- National Infrastructure for Supporting Technology Transfer in Slovakia - NITT SK
- Fostering Continuous Research and Technology Application - FORT
- Boosting innovation through capacity building and networking of science centres in the SEE region - SEE Science

www.cvtisr.sk
Lamačská cesta 8/A, Bratislava

International identification and ‘white and grey literature’: Identities, retrieval, reuse and the certainty of knowledge while sharing and connecting information

Flavia Cancedda, CNR - Biblioteca Centrale, ISSN National Reference Centre
Luisa De Biagi, CNR - National Research Council of Italy, Biblioteca Centrale, Italy

Abstract

During the 20th century the development of identifier codes - in most cases internationally spread under ISO auspices - encouraged the idea that few things were yet “unidentifiable” and “uncontrollable” in document field. Then, the expansion of new technologies for information retrieval on the Net, made thinkable the advent of a sort of social control for documents, data and metadata identifiers: this seemed to be in conflict with the concept of a unique identification coming from an authoritative origin. In fact, we’ve seen more recently a new increase of bibliographic identifiers, mainly concerning digital environment documents, and especially, some publishing sectors not directly interested to the topic of identification have been involved in. Finally, also responsibility entities (individuals, groups, corporate bodies...) have been included under identifying activity. The identification of publishing or documentary products/actors seems to be now consolidated. So, it becomes urgent to establish common guidelines - description, metadata, cross-identification - widely shared and implemented by agencies or any other component of the information chain. The inclusion of grey literature in meta-analysis is fundamental.

We consider the usefulness of a dynamic model for cross-sharing and cross-use of data, metadata and identifiers, that allow international agencies to pool or exchange their information collections, avoiding duplications when same data match in more than one archive. Moreover, this model could be easily supported by current techniques for information retrieval via linked data. Obviously the aim would not to create the nth super archive, but to encourage the disseminated allocation of multiple information, that could be found or summarized just when searched by users. Necessary condition is the cooperation among the agencies.

The goal of this study is showing which consequence could represent a general improvement of the public information quality level, thanks to the exponential circulation of authoritative data: “low-cost” for agencies, publicly available for everybody, and easy update according to rigorous certified criteria.

Flexibility of grey literature: *pleasure & pain*.

The Flexibility of grey literature (preprints, posters, theses, patents, policy documents, Research and Internal reports, etc.) could be actually both an advantage and a problem because GL lacks lots of infrastructures and best practices used by scholarly publishing. The term ‘grey literature’ covers an area between scholarly and popular literature and Grey e-resources aren’t typically controlled by academic publishers, traditional ‘gatekeepers’ of scholarly literature.

In fact, Grey literature is made by researchers and ‘fed’ by Research, but isn’t still usually viewed in the ‘Upper Class’ of the scholarly literature because, for example, it’s difficult to cite in academic journals. Grey documents are not totally and constantly considered in citation indexes like Web of Science or Scopus: some editors and publishers even discourage any formal citation of preprints and similar (e.g. American Chemical Society¹, American Association for Cancer Research - AACR - , the *Journal of Clinical Investigation* by American Society for Clinical Investigation, *The Journal of Experimental Biology* by the Company of Biologists²).

The wide panorama rewarding GL is still without a ‘sustainable ecosystem’, though good initiatives like greyLit.org are to be considered a significant effort³. So, important challenges and processes have previously to be done:

¹ Most ACS journals do not explicitly allow preprints.

² By submitting a research article to JEB, the authors undertake that it has not been published previously (that means posting the article on a preprint server) and is not submitted for publication elsewhere.

³ A bimonthly publication and a data-base platform provided by The New York Academy of Medicine, “alerting readers to new grey literature publications in health services research and selected public health topics”. It works as an archive for the cataloged reports

- Identifying relevant documents
- Extracting metadata and references
- Granting permanence and digital preservation
- Recognized evaluation

Grey literature could also represent a great opportunity for alternative metrics, providing research data and indicators not reliable and available anywhere else. Altmetrics analyze tweets, blogs, posts, presentations, news articles, comments, or any 'social voice' on the Web rewarding scholarly community. But without knowing even how much Grey material is created each year, it's difficult to realize how complete any citation index is, even using altmetrics. The constant use of a PI as, for example Handle System - in which resources are assigned a unique identifier that can be resolved to a URL by the creator – would represent a significant solution to ensure track of documents, even if they move around Internet. In Italy, a good *policy* example for academic institutional repository is given by Trieste University (IT). In accordance with *CRUI guidelines*⁴, Trieste University promotes self-archiving of PhD thesis, which are very quickly deposited in *OpenStar Ts* before their discussion, with a Persistent Identifier assigned as URI, in some cases matched with NBN - National Bibliographic Number - . So, authors can authorize an immediate open access availability or set the 'embargo' by 1 year.

In particular NBN, defined by RFC 3188 standard (<http://www.rfc-editor.org/rfc/rfc3188.txt>) is a persistent identifier based on URN (Uniform Resource Name), and identifying in an unambiguous way a publication. In Italy NBN:IT is linked to legal deposit and digital preservation. The project, coordinated by the National Libraries of Florence, Rome, and Venice, by the Foundation Rinascimento Digitale and by the Conference the Italian University Rectors, completes the PI current offer⁵.

Past and present: identifier codes in bibliographic fields

During the 20th century the development of identifier codes - in most cases internationally spread under ISO⁶ auspices - encouraged the idea that few things were yet "unidentifiable" and "uncontrollable" in documentary fields. It could be useful to remember that "bibliographic identifier code" means a sequence of characters - mostly numbers, but not only - preceded by an alphabetic acronym that indicates the type of code. The codes are used to identify uniquely publishing/intellectual works or identities (persons, bodies etc.) in the world of communication and media. Each type of code is issued by a responsible agency - guarantor of integrity and uniqueness of it -; each code matches to a set of descriptive data and metadata managed by the agency (e.g.: name, title, author, year of publication, date of code assignment, etc.). The bibliographic codes are widely used in the whole publishing and media market for commercial distribution. They are also necessary for tracking the publication/work/entity in whatever step of the circulation chain, from the producer to the end-user and to other agencies which process bibliographic data for providing advanced information services (e.g.: bibliographic or informative archives of scientific publications or scholars/scientists). The most common and known identifiers? ISBN for books and ISSN for periodicals, but also, e. g., ISAN for audiovisuals, ISRC for recorded music, ORCID for scholars.

In the last years of the 20th century the expansion and evolution of new technologies for information retrieval on the Net made thinkable the advent of a sort of *social control* for documents/data, and for the metadata identifiers too. This new horizon seemed to be in conflict with the concept of micro-tools for unique identification – as the identifier codes - coming from an authoritative origin (the agencies assigning the codes): in fact, the wide distributed technology allowed continuous bottom-up controls in a pervasive way never before experienced in the history of culture and publishing diffusion.

But: the "Net public" could check and control directly and from different sources the correct identification and essence of publishing objects, had the identifier codes a useful role yet? The answer of the media market was: yes.

⁴ Conference of Italian University Rectors

⁵ Bellini, E., Cirinnà, Lunghi, M., Luddi, C. (and others), *The National Bibliography Number Italia (NBN:IT) Project. A persistent identifier supporting national legal deposit for digital resources* in 'JLIS.it'. Vol.3, n.1 (Giugno/June 2012). DOI: 10.4403/jlis.it-4789 NBN: urn:nbn:it:unifi-3866

⁶ International organization for standardization, iso.org.

As long as the Net enormously increased the informative opportunities for the world public, professional operators and issuing bodies decided to increase and enlarge coverage and aims of bibliographic identifier codes, to allow the unique identification of all new online publishing objects, often disaggregated and miniaturized in front of the traditional products, but overall changing and updating their form and content. If the new publishing products could be continuously updated by their producers, it could be obviously difficult or impossible for the end-users to trace the intellectual content of the products (as well as to trace their reliability and trustworthiness). This *uncertain status* could undermine the confidence that readers had in publishers and authors; for that reason, publishers decided to renovate and increase the credibility value of their publications, entrusting again to the updated bibliographic identifier codes the unique and certain identification of the digital and online publications, and of intellectual contents inside.

The main consequence of these two divergent cultural trends (dilatation of online publishing products potentially unidentifiable; intensification of the identifying power of the traditional bibliographic codes) was the proliferation of bibliographic identifiers, *in* and *out* of the ISO context. Some of them were old identifiers born for the paper environment and recently updated for the electronic and digital media; others were definitely new, as ISTC, DOI, NBN, ISLI. Another interesting consequence of this trend was that some new publishing *entities* not directly interested to now in the topic of identification have been finally involved in this process: people (connected with intellectual, cultural or media activities, and now identified by at least two different codes, ISNI and ORCID) and digital links (connecting multiple digital editions, identified by the ISLI code).

The conclusion of this complex process of renewal is: identifier codes of publishing or documentary products/actors are definitely consolidated and strongly recognized in the media environment. Moreover, each bibliographic identifier can innately become a Persistent Identifier; in fact, to create an online Persistent Connection a persistent link between two information sources is sufficient: for example, between an identifying code – whose metadata registered by the relevant agency in its database act as guarantors - and the publishing object itself or its public description (both generally available in the publisher's website or in other specific digital stores).

Dynamic model for re-using and disseminating P.I. (*the public control and re-use of bibliographic warranties*)

In the world there is a sizeable quantity of bibliographic identifier codes (the ISO family alone contains ten different codes), implemented and continuously managed by several different international agencies, often spread widely by country. The overall mass of bibliographic and informative data they collect/update/maintain is impressive⁷.

It is now urgent to establish common guidelines - about description, metadata collecting systems, cross-identification - that could be widely shared and implemented by international and national agencies, or by any other component of the information chain. In fact, by now the different agencies – even in cases when they are established and promoted by bodies having similar organization profiles, and even though those agencies share similar socio-cultural backgrounds, aims and objectives – do not usually cooperate in common policies; mainly they do not collaborate in gathering the bibliographic information, which is the basis activity for building the metadata grids necessary to ensure a reliable correspondence between publishing objects and identifier codes.

How to draw an information cross-exchange now? First of all, it is crucial that international and national agencies acknowledge that a cross-cooperation is needed: the exchange of data is useful for each actor of the informative chain, and particularly for agencies themselves. The informative data common to more than one system identification (i. e., the main part of bibliographic data, as names and address of publishers, trademarks, names and address of issuing bodies, names of intellectual contributors, titles of works edited in different media, title of collections...) could be easily pooled by different registration agencies. This operation should not be implemented collecting the data in a new *nth* database – obviously uncontrollable because of its size and

⁷ Example: at present the ISBN network collects essential public data of more than 1 million of publishers (the enormous number of the books identified by ISBN code is actually not known...); the ISSN network records the bibliographic data of about 1.9 million of serial publications, etc.

because of the different policy needs of the co-owners -, but using the cross-reference system allowed by linked data technology⁸.

The agencies responsible for assignment of codes could choose this kind of data-sharing, agreeing about a minimum set of data and metadata acknowledged as “common” and shareable; so, same agencies would evidently save substantially their own human and technological resources, avoiding millions of recording operations, often identically repeated in different identifying systems. Moreover, this system could be implemented in open access regime, so the results would become immediately available for the whole Net public.

This big saving would make available the same H&T resources for other kinds of advanced services; the loss of profits depending from the public availability of bibliographic data (now mostly priced) could be balanced by the pricing of new customized services, particularly when they are addressed to publishers, providers, data aggregators, institutional bodies, etc.

The feasibility and usefulness of a dynamic model for sharing and cross-reusing of data, metadata and identifiers is evident when we add to a traditional bibliographic record the related identifier codes already assigned to each component of it, as in the following example:

cataloguing record (as in the homepage <http://www.greynet.org/>)

[GL16 Conference proceedings] Sixteenth International Conference on Grey Literature : Grey Literature Lobby: Engines and Requesters for Change, Library of Congress, Washington D.C., 8-9 December 2014 / compiled by D. Farace and J. Frantzen ; GreyNet International, Grey Literature Network Service. Amsterdam : TextRelease, March 2015. – 164 p. – Author Index. – (GL-conference series, ISSN 1386-2316 ; No. 16)

Several components of this description are already identified by codes:

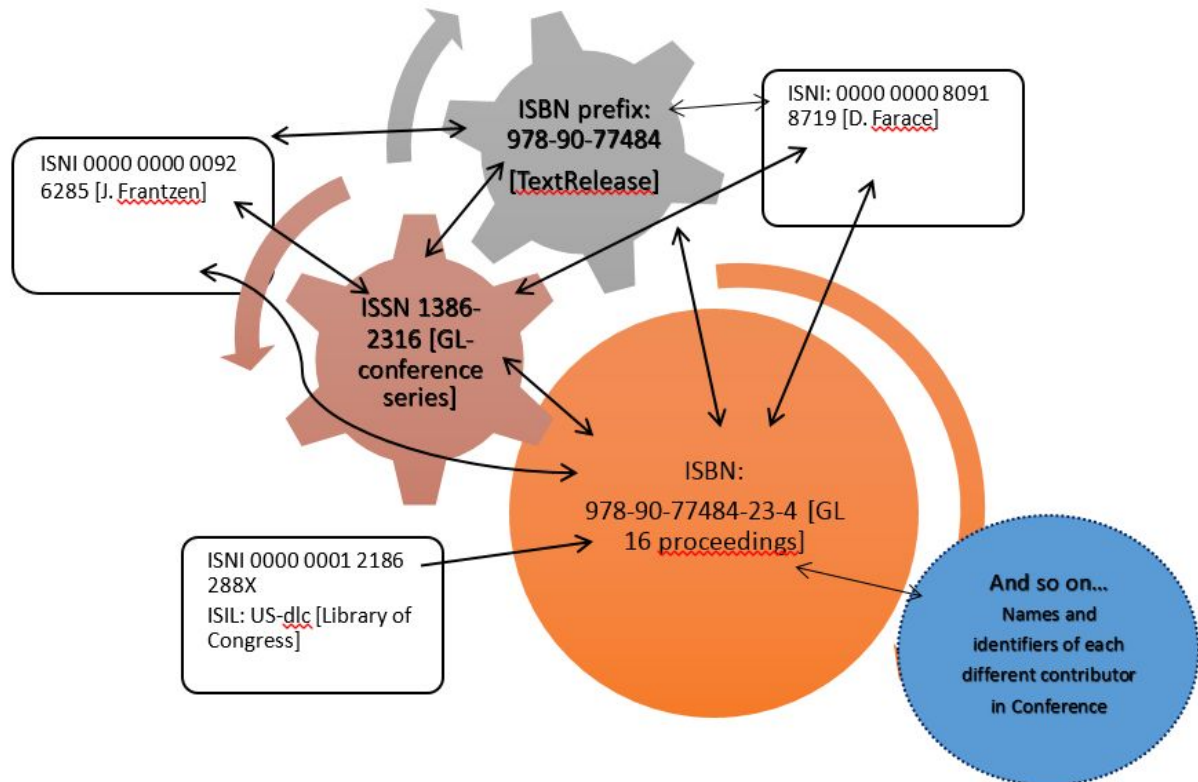
Object	Name	ontology	Identifier codes
Publishing object: book	<i>[GL16 Conference proceedings] Sixteenth International Conference on Grey Literature etc.</i> (published by TextRelease in 2015)	Published by Textrelease Edited by D. Farace Edited by J. Frantzen Expression of the participants in Sixteenth International conference on grey literature.... Contained in GL-conference series	ISBN: 978-90-77484-23-4
Entity: body	Library of Congress	Hosted and sponsored the Conference	ISNI: 0000 0001 2186 288X ISIL: US-dlc
Entity: person	D. Farace	Edited the book Sixteenth International Conference on Grey Literature etc.	ISNI: 0000 0000 8091 8719
Entity: person	J. Frantzen	Edited the book Sixteenth International Conference on Grey Literature etc.	ISNI: 0000 0000 0092 6285
Entity: issuing body	Greynet international	Issued the Sixteenth International Conference on Grey Literature etc. [conference and book]	ISNI: 0000 0001 1508 0451
Entity: body; publisher	TextRelease	Published the book Sixteenth International Conference on Grey Literature etc.	Publisher's ISBN prefix: 978-90-77484
Publishing object: series	GL-conference series	Published by TextRelease; Contains the book Sixteenth International Conference on Grey Literature etc.)	ISSN: 1386-2316 (print ed.)

Each of the above identifier codes (ISBN, ISNI, ISIL, ISSN) has been issued and granted by a recording system where personal and bibliographical data – according to the specific internal rules of each identifying system - has been previously checked and stored.

It is not difficult to imagine that this traditional list of elements, codes and related metadata could be easily make available for a linked data approach, where the ontologies (i.e.: roles and

⁸ General info about the world of linked data in <http://www.w3.org/standards/semanticweb/data>.

functions) are the digital glue linking chains of digits (the codes) which, in their turn, drag behind other metadata – id est: descriptive textual information or URI – already validated and not to be re-written or for the umpteenth-time repeated.

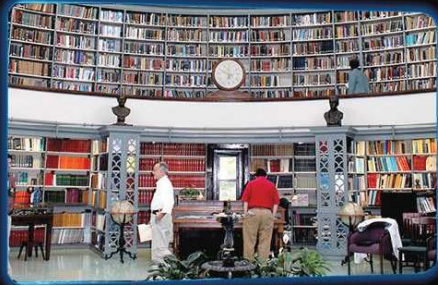


References:

- RFC 2288, Feb 1998, C. Lynch, C. Preston, R. Daniel, *Using Existing Bibliographic Identifiers as Uniform Resource Names*;
- RFC 3044:2001, S. Rozenfeld, *Using The ISSN (International Serial Standard Number) as URN (Uniform Resource Names) within an ISSN-URN Namespace*;
- RFC 3187:2001, J. Hakala, H. Walravens, *Using International Standard Book Numbers as Uniform Resource Names*;
- RFC 3188:2001, J. Hakala, *Using National Bibliography Numbers as Uniform Resource Names*;
- I-D IEFT-URNBIS-RFC3044BIS-ISSN-URN-01, 7 Nov 2012, P. Godefroy, *Using International Standard Serial Numbers as Uniform Resource Names*;
- I-D IEFT-URNBIS-RFC3187BIS-ISBN-URN, 19 Oct 2012, M. Huttunen, J. Hakala, A. Hoenes eds., *Using International Standard Book Numbers as Uniform Resource*
- ISO standards (last or current editions):
- ISO 15707:2001, *International Standard Musical Work Code (ISWC)*
- ISO 3901:2001, *International standard recording code (ISRC)*
- ISO 15706-1:2002 *International Standard Audiovisual Number (ISAN), Part 1: Audiovisual work identifier*
- ISO 2108:2005 *International standard book number (ISBN)*
- ISO 3297:2007 *International standard serial number (ISSN)*
- ISO 15706-2:2007 *International Standard Audiovisual Number (ISAN),- Part 2: Version identifier*
- ISO 10957:2009 *International standard music number (ISMN)*
- ISO 21047:2009 *International standard text code*
- ISO 26324:2012 *Digital object identifier system*
- ISO 27729:2012 *International standard name identifier (ISNI)*
- ISO 17316:2015 *International standard link identifier (ISLI)*
- Auger, C.P., Ed. (1989) *Information Sources in Grey Literature* (2nd ed.). London: Bowker-Saur. ISBN 0862918715
- Bellini, E., Cirinnà, Lunghi, M., Luddi, C. (and others), *The National Bibliography Number Italia (NBN:IT) Project. A persistent identifier supporting national legal deposit for digital resources in 'JLIS.it'*. Vol.3, n.1 (Giugno/June 2012). DOI: 10.4403/jlis.it-4789 NBN: urn:nbn:it:unifi-3866

FEDLINK

an organization of federal agencies working together to achieve optimum use of resources and facilities of federal libraries and information centers by promoting common services, coordinating and sharing available resources, and providing continuing professional education.



Strategic Sourcing



Currently, more than 20 federal agencies, both military and civilian – including FEDLINK – participate in the Federal Strategic Sourcing Initiative (FSSI). FSSI was created in 2005 by the Dept. of the Treasury, the Office of Management and Budget, and the General Services Administration to identify products and services that can be purchased more efficiently through strategic sourcing. FSSI agencies also provide centralized acquisition functions for a variety of products and services to streamline efficiency and reduce costs to the federal government.



101 Independence Ave, SE ~ Washington, DC 20540-4935
FEDLINK Main Number (202) 707-4800
FEDLINK Hotline (202) 707-4900

Sustaining Scholarly Communication Support By Academic Libraries In Sub-Saharan Africa: A Case Of Makerere University And University Of Zimbabwe Libraries

Andrew Mwesigwa, Makerere University, Uganda

Elizabeth Mlambo, College of Health Sciences Library, University of Zimbabwe

Abstract

Institutional repositories in academic institutions in the global South have made great contributions to the visibility of research emerging from developing economies. However, the process of establishing and managing them has proved so daunting that it requires strategic planning to sustain such efforts. The Libraries of Makerere University (Mak) and the University of Zimbabwe (UZ) examine, discuss and report efforts towards strategic planning for sustainability of institutional repositories and related digitisation initiatives. A preliminary study at Mak and UZ indicated inadequate awareness of the distinction between institutional repositories and other research management tools among faculty Musoke and Mwesigwa (2014). With the support of development partners like INASP, EIFL and the British Library for Development Studies (BLDS) at the Institute of Development Studies (IDS), these two public universities located in Sub-Saharan Africa have invested in infrastructure and committed resources that support scholarly communication strategies. The institutional repository depends largely on contributions by academics and researchers to reach a critical mass of content without which it could not serve its purpose as a scholarly communication forum for sharing research. The paper therefore attempts to answer some of the intricate policy issues, which are pre-requisite for academic libraries in the global south to sustain their supportive function in the scholarly communication landscape, which has implications for the management of institutional repositories. We hope the lessons learnt could be used by other developing country universities that are running or attempting to build institutional repositories.

Keywords: Capacity-building for institutional repositories, institutional repositories in the Global South, sustaining digitization initiatives.

Introduction

Makerere University (Mak) and the University of Zimbabwe (UZ) are found in sub-Saharan Africa. The two universities are the oldest and largest universities in Uganda and Zimbabwe respectively. Both universities had affiliation with the University of London in their early years whereby Mak was a university College of London and University of Zimbabwe had a special relationship with the same. Both Mak and UZ have big student populations (over 35,000 and over 12,500 respectively). Both Mak and UZ have digital institutional repositories¹, which enhance their visibility online. Both UZ and Mak as compared to other universities in Zimbabwe and Uganda respectively, have a well-documented and time-tested research culture and UZ has a publishing house with a decent output, factors among which have created a conducive environment for the implementation of a successful IR.

Capacity Building Strategy at Makerere University

Mwesigwa (2012) presented Makerere University's experience in setting up and running the Institutional repository, providing an account of the process and the challenges experienced in the first 6 years. Several efforts have been made by Makerere University Library (Maklib) to implement the recommendations given by Kakai (2009), Musoke (2010) and Mwesigwa (2012). These efforts have included capacity building of librarians and ICT staff to expose and empower them with necessary knowledge and skills to provide the supportive environment needed for scholarly communication at Makerere University. Capacity building has taken the form of training at PhD and Masters level as well as short task-specific ICT training such as DSpace server-side configuration and repository workflow management. Maklib recognises the specific financial-support provided by various development partners towards this effort (Swedish Government-Sida, Institute of Development Studies, Norwegian Government, etc). In Maklib's view, capacity built lasts longer than out-sourcing service support as reiterated by Playforth (2014) and Musoke (2010).

¹<http://ir.uz.ac.zw> and <http://makir.mak.ac.ug>, respectively

The short ICT task specific training ensures that staff trained gain the ability to manage the repository service in-house. Whereas there have been arguments by some scholars in favour of cloud-hosting of repository services (Zainab, Chong & Chaw, 2013), Maklib believes that if staff are trained, then hosting the repository on a local server is a better option for sustainability. Also the security and long-term ownership of content submitted in a repository hosted locally have highlighted the need for capacity building. Maklib plans to continue supporting staff for refresher training in order to sustain in-house digitisation and repository services.

Capacity Building to Manipulate Open Access Resources at The University of Zimbabwe

INASP, EIFL and IDS have emerged as major players in providing capacity building initiatives for the University of Zimbabwe's Institutional Repository. INASP has provided funding for, training of library staff in Open Access workshops and to raise awareness of Open Access initiatives. INASP supports the Open Access Movement and is heavily involved in signposting and raising the visibility of Open Access research output. It provides training support and guidance to University of Zimbabwe librarians and publishers about the options available to them. IDS has supported the initiative by providing a state of the art computer and scanner for uploading of UZ scholarly output on OpenDocs². Access to the UZ Social Science Research sub-community has been overwhelming with the following five countries from the international community topping the list of access: United States of America with 2771 downloads, France with 1021 downloads, China with 928 downloads, Germany with 435 and United Kingdom with 214 (as of 30th November 2014). BLDS is a well-known brand and OpenDocs content can be found through Google scholar. About 31 000 downloads are made on a monthly basis from the digital library. As highlighted by Mwesigwa and Musoke (2014), we believe that with the awareness of usage patterns of the intellectual outputs from UZ and Mak, which are archived in BLDS digital library there shall be a multiplier effect of increased access and usage of UZ's and Mak's own repositories respectively hence the two repositories will also increasingly become important scholarly communication channels not only locally but globally.

EIFL is also giving financial support to the University of Zimbabwe library to embark on a campus wide Open Access Advocacy Campaign with the aim of adopting a campus wide Open Access policy. UZ and Mak have participated in the Open Access Week every October in the last three years.

Policy Support

The paper also covers efforts at both Mak and UZ to provide supportive policies, which encourage mandatory content submission and sustainability of IRs as well as related scholarly communication services offered by the libraries of Mak and UZ respectively.

Research Data Management and Content Recruitment Strategy

Maklib has embarked on collecting research data from the research performance tool, InCitesTM (a Thomson Reuters software solution), continuous digitisation and uploading of grey literature and the rich archival collection as well as encouraging academic staff to self-archive their scholarly materials as some of the strategies to sustain content recruitment into the institutional repository.

As an incentive to actively self-archiving researchers, the UZ library and Maklib have honored and awarded researchers who contribute most to the IR.

Both UZ library and Maklib keep track of the usage and download statistics and conduct e-resources training and Information Literacy training to encourage usage of the respective repositories.

Integration of Digitisation and Repository Work: Strategy and Workplan

There is a dedicated library section at Maklib called Digitisation. The Digitisation section has librarians with experience and skills to work on repository workflows. The section was created when Maklib had integrated digitisation and repository services in its 10-year strategic plan, which directly fits within Makerere University's strategic plan. As a result, the repository is high on the agenda of the university as it contributes to the institution's profile and visibility.

Repository/Digitisation Librarians have been trained in workflow management hence handle the day-day tasks as part of their workload both at UZ and Mak. There is also decentralization of IR activities to faculty/college librarians to improve workflow management. This ensures continuity of the repository service from content submission, curation to archiving. Librarians have continued to market the repository among academic staff at Makerere University and University

²<http://opendocs.ids.ac.uk/opendocs/handle/123456789/3>.

of Zimbabwe with the result of having academic staff voluntarily submitting their scholarly papers for self-archiving. Maklib appreciates the refresher training, which was supported by IDS through the DFID grant. It enabled refresher training for repository Librarians (Mwesigwa&Musoke, 2014; Playforth, 2014).UZ's Special Collections department is responsible for the activities relating to digitization of any scholarly output. This is to fulfill the objectives of the Institutional Repository which were to comprehensively collect the research output of UZ scholars and researchers, digitizing the scholarly output of UZ scholars, disseminating these products and publications widely and preserving the products of the UZ scholars (Tevera and Mlambo 2007:41).The UZ has embarked on a digitization project of UZ Social Sciences publications with support of IDS. Faculties and Institutes are invited to submit their older research dating back to the 60s for digitization. It further supports Mbambo (2006)'s assertion that for UZ, Open Access was important as it gives researchers through the Internet an opportunity to share, increase Zimbabwe content in e-form and enhance access to learning materials as well as 'grey' literature.

Conclusion

As reported by Mwesigwa and Musoke (2014), that digitisation is a never-ending process, the changing technology environment requires the continued commitment of research institutions. The paper reports efforts, which depict the commitment that Mak and UZ have made towards sustaining the development of repository services at both institutions respectively, even beyond donor support. Through its IR, the UZ has managed to adopt the Green Route or self-archiving approach to Open Access. This effectively means UZ authors can now make their articles freely available in digital form on the Internet. The University academics' articles are now more visible, discoverable, retrievable and useful. University lecturers through Open Access now have a wide audience larger than that of any subscription journal and their increased visibility is now impacting on their work with a possible increase in citation of their work. This is derived from the fact that access to subscription journals is limited to institutions with the subscription credentials, whereas UZ and Makdigital repositories offer online and world-wide open access to materials.

Visibility of both institutions has been increased.UZ and Makhave successful IRs because of the diverse subjects of relevant digital collections and their accessibility on the web. Both UZ and Mak have made brisk steps towards improving their bandwidth hence improving user download experience. Both Mak IR and UZ IR are registered repositories on OpenDOAR. Mak and UZlibraries have managed to advance the mission of sharing knowledge. The commitment demonstrates deliberate efforts by both university libraries to innovatively support the scholarly communication landscapes at both institutions, which is a primary role of academic libraries.

References

- Kakai, M. (August, 2009). The challenges of advocating for open access through institutionalrepository building: experiences from Makerere University, Uganda. Paper presented at the World Library & Information Congress, Milan, Italy.
- Mbambo, B (2006) Open Access and Libraries in Zimbabwe: a Case Study of UZ; Paper presented at the Zimbabwe University Libraries Consortium (ZULC)International Conference on " Open Access and Creating a Knowledge Society", held at the Crowne Plaza Monomatapa Hotel, Harare, Zimbabwe,24 - 26 April, 2006
- Mlambo, E. (2013) Open Access: maximizing research impact by maximizing research access: experiences at the University of Zimbabwe. Paper presented at a Conference on Information and Knowledge Management in the Innovation era-space – technology and time Johannesburg, June 2013.
- Musoke, M. G. N. (August, 2010). Reconstruction@Maklibusing minimal resources. Paper presented at the World Library & Information Congress, Gothenburg, Sweden.
- Musoke, M. G. N. & Mwesigwa, A. (May, 2014).Exploring the Unknown: Usage of the Incites™Research Management Tool at Makerere University. Paper presented at theResearch Evaluation and Performance Measurement (REPM) International Conference, Cape Town, South Africa.
- Mwesigwa, A. & Musoke, M. G. N. (November, 2014). Opening Access to Rare Collection through collaboration: the experience of Maklib and BLDS.Paper presented at the 10th International Conference on Knowledge Management (ICKM), Antalya, Turkey.
- Mwesigwa, A. (July, 2012). Makerere University's experience in setting up and managinganinstitutional repository running on DSpace software. Poster presented at the 7th International Conference on Open Repositories, Edinburgh, Scotland.
- Playforth, R. (June, 2014). Grey literature, green open access: the BLDS Digital Library. Paper presented at the 9th International Conference on Open Repositories, Helsinki, Finland.
- Tevera, S. & Mlambo, E. (2007).Digitising local collections. In: Mbambo-Thata, B. (Ed.). Building a digital library at the University of Zimbabwe: a celebration of team work and collaboration. Oxford: INASP
- Zainab, A. N., Chong, C. Y. & Chaw, L. T. (2013). Moving a repository of scholarly content to a cloud. *Library Hi Tech*, 31(2): 201-215.

A semantic engine for grey literature retrieval in the oceanography domain

Sara Goggi, Gabriella Pardelli, Roberto Bartolini, Francesca Frontini, Monica Monachini
CNR, Istituto di Linguistica Computazionale, "Antonio Zampolli", Italy

Giuseppe Manzella, ETTsolutions;

Maurizio De Mattei and Franco Bustaffa, DP2000, 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.

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 an 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.

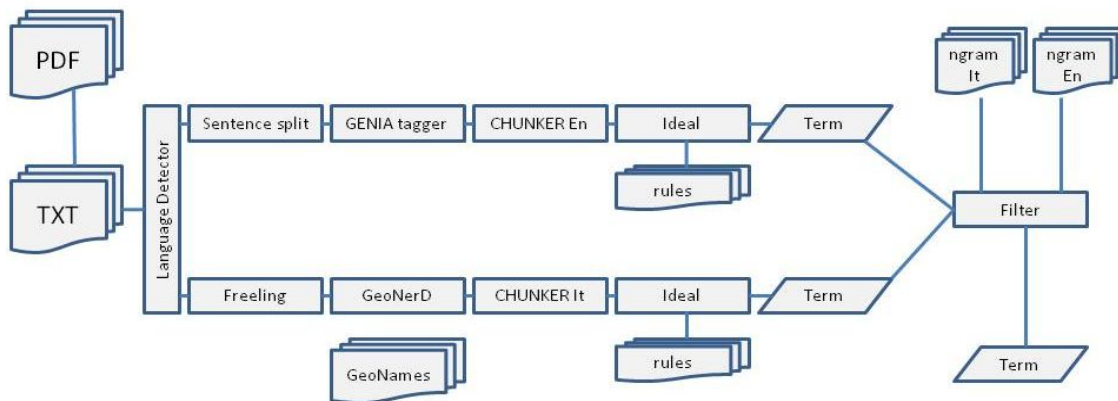


Figure 1: Steps of Natural Language Processing and term extraction.

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:

- SentenceDetector of the Apache OpenNLP suite was used for sentence splitting (<https://openlp.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 queryPipe 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,

¹ <<https://github.com/CLD2Owners/cld2/>> last accessed on January 15, 2016.

drawn from SeaDataNet vocabularies²: 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" : "Z20 Z10 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.

² <<http://www.seadatanet.org>> last accessed on January 15, 2016.

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:

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

where n_r is the number of documents retrieved by the search, n_R is the number of relevant documents in the document collection and 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 n_r and the number of relevant documents 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 February 2012 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](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://CITSEERX.IST.PSU.EDU/VIEWDOC/DOWNLOAD?DOI=10.1.1.21.8180&REP=REP1&TYPE=PDF](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](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), RAD-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.

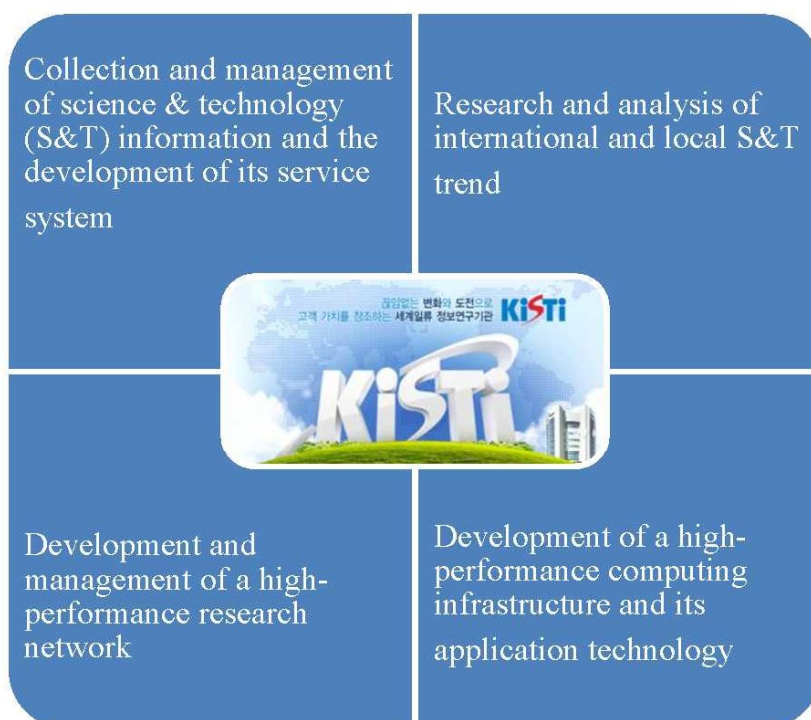
Korea Institute of Science and Technology Information (KISTI)

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

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

Kiseok Choi, Cheol-Joo Chae, Yong-hee Yae, Yong Ju Shin
Korea Institute of Science and Technology Information (KISTI), 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 contributonal 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 contributonal 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 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

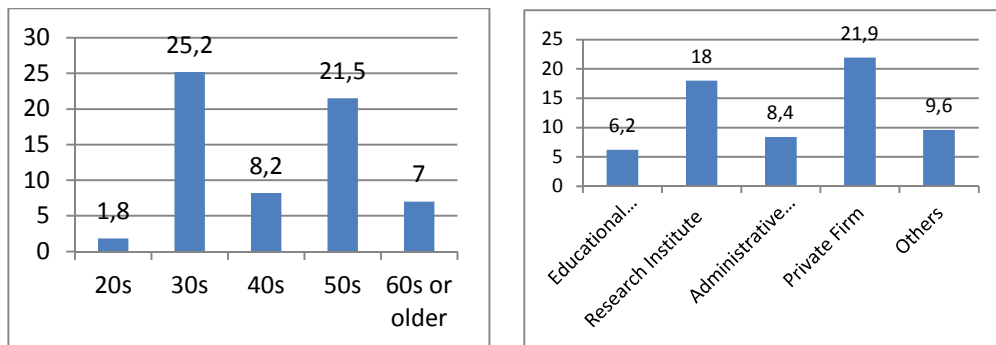
2.1 Analysis on the utilization of national R&D reports

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.’

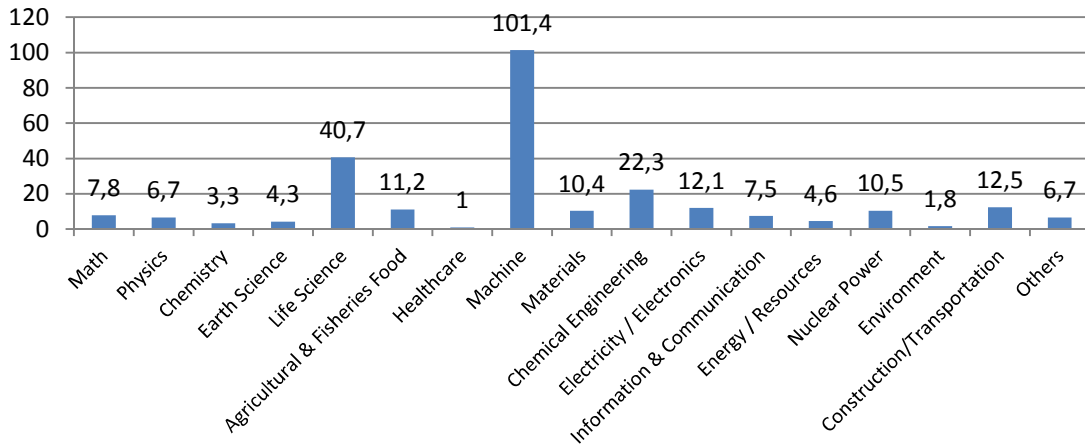


Figure 2> Average Monthly Detailed-view Frequency of National R&D Reports by Topic

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:

$$\sum_{i=1}^n \bar{x}_i \times 14.28$$

\bar{x} = Mean score 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

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

Values_i: Benefits or costs on the ith year, rate: discount rate, n: No. of years

※ 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
$\sum_{i=2009}^{2014} \text{Values}_i \times \text{AC} \times \text{MPY} \times \text{User}_i$ <p>Total NPV =</p> <p>Values_i = Utilization value of a report(2,908 KRW) AC = Average monthly usage frequency per capita (9.2) MPY = 12 months User_i = No. of subscribers on the ithYear</p>

<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/retrievalIntro.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.

Scientific Audiovisual Materials And Linked Open Data: The TIB Perspective

Paloma Marín Arraiza,

German National Library of Science and Technology (TIB), Germany

Abstract

Libraries are starting to use Linked Open Data (LOD) to provide their data (library data) for reuse and to enrich them. However, most initiatives are only available for textual resources, whereas non-textual resources stay aside. Firstly, this paper discusses the potential of library data to be published as LOD. Secondly, it focuses on the library data related to the management of audiovisual scientific materials in the TIB|AV-Portal. The use of LOD Standards to support multilingual functionalities and data reuse is outlined. Future developments lead to the building of semantic applications based on LOD Structures

Keywords: scientific videos, linked open data, data publishing, semantic applications, grey literature, library data.

1 Introduction

A large amount of data is generated by global science every day from multiple sources. Libraries' task is to organise and classify all the resources and their respective data, generating new data known as 'library data'.

Data on the web exists in a structured form in databases, and in semi-structured forms in textual and non-textual collections. To deal with both the flood of information as well as the range of heterogeneous data formats, a new approach was needed for information searching and access. The data should no longer be isolated but connected to other data, and become accessible, explorable and discoverable by both people and machines. Thus, the HTML based web changed into the Web of Data, also referred to as Semantic Web. It derives value from data and creates pathways between datasets and resources (Heath and Bizer, 2011).

The derivation of value from the data is possible due to the structure this data presents. In the web of data, information and knowledge are stored in simple structures known as 'triples', which consist of three parts: subject, predicate and object. To publish and interlink structured data on the web, RDF (Resource Description Framework) is used. Data linked to other and published under open licenses (such as CC0¹) are known as Linked Open Data (LOD).

Tim Berners-Lee (2006) outlined the principles structured data need to follow in order to be published as Linked Data:

1. Use URIs (unique resource identifier) as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
4. Include links to other URIs, so that they can discover more things.

The use of semantic technologies and LOD concerns libraries and has a lot of advantages. LOD is better retrievable and easier to interlink since URIs, that define things, are more stable than URLs (uniform resource locator), that define addresses. The interoperability and further use of the data is improved since LOD is based on open web standards and besides the used data model, RDF, is more flexible than other library standards (e.g. MARC) (Pohl and Danowski, 2011). Therefore, W3C Standards become a reliable and flexible alternative.

The aim of this article is to present the use of LOD in libraries and at the TIB, specifically in the TIB|AV-Portal, and how LOD support multilingualism, data reuse and information retrieval applications.

The structure of the paper is as follows. Section 1 introduces the topic and describes the structure of the paper. Section 2 defines library data and specifies the use of LOD in libraries for textual and non-textual materials. Section 3 presents the use of LOD in the TIB|AV-Portal. Section 4 concludes the article.

¹ Creative Commons Zero, public domain.

2 Linked Open Data in Libraries

2.1 Library Data and new types of scientific information

In the Library Data Incubator Group Final Report (2011), library data is defined as ‘any type of digital information produced or curated by libraries that describes resources or aids their discovery’. So far, scientific written texts such as books and journal articles were the main resource of academic libraries. Non-textual materials were considered to be inappropriate for academic purposes, and were usually regarded as general-interest publications rather than as ‘proper’ scientific publications (Löwgren, 2011). However, the scientific landscape changes constantly and academic librarians must be aware of the importance of this multimodal scholarship (Spicer, 2014). This means, they should provide platforms to support these materials, index them properly, and promote their accessibility and reuse.

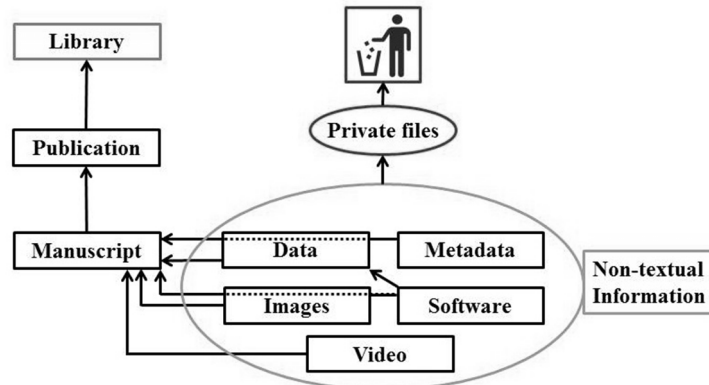


Fig. 1. Flow of scientific information from research to publication (modified after Klump et al., 2006²).

Therefore, as mentioned in the Pisa Declaration (2014), ‘grey materials’ should be persistent identified and linked, as far as possible, with other publications. This generates a big amount of library data that aids the classification and finding of relevant information. Not doing so lead to a loss of scientific information, as shown in figure 1.

2.2 The potential of library data for Linked Open Data

The W3C consider libraries an important focus for LOD because of being important content providers. Being usually structured information, library data present a high potential to be shared as LOD. Peset et al. (2011) highlight libraries projects where LOD play a crucial role: Authorities and vocabularies of the Library of Congress, Linked Data service of the Deutsche Nationalbibliothek (DNB), and Libris, the collaborative Swedish catalog. Moreover, the Web is considered a global data space where libraries should publish their data using RDF and open licenses. Doing so, they promote and support data reuse (Saorín, 2012).

However, most library content is text-based providing few non-textual materials. An example of aggregation environment with textual and non-textual materials is Europeana. Europeana is classified by the Library Data Incubator Group under ‘archives and heterogeneous data’ since it manages metadata from diverse types of materials and formats (Ríos-Hilario et al. 2012). Some tasks performed by Europeana to promote LOD in aggregation scenarios are the experimentation with semantic search based on RDF within collections, knowledge organization systems alignment, and data aggregation system with LOD export.

In this context, we present the TIB|AV-Portal as future use case of the LOD approach for non-textual materials. Some applications, such as multilingualism, are already implemented in the portal. Further implementation is focused on the weaving the data into LOD to enable metadata enrichment in scientific audiovisual materials.

3 The TIB|AV-Portal: Current practices and perspectives with Linked Open Data

3.1 Multilingualism

Textual metadata from text and speech recognition are linked to entities of the *Gemeinsame Normdatei* (GND, German for: Integrated Authority File), which is mainly managed by the German National Library. The data of the GND are available with the formats MARC 21

² Klump et al. (2006) Data Publication in the Open Access Initiative. Data Science Journal, Volume 5, 15 June, 2006, page: 80.

Authority, MARC21-XML and RDF-XML under Creative Commons Zero³ (CC0). This publishing as Linked Open Data enables the linking between entities in the portal.

From the complete dataset of the GND, the TIB|AV-Portal uses the terms related to the six TIB core subjects. The GND includes synonyms, homonyms, hierarchical relations between terms and cross-references between related terms. These properties are used in the portal to perform a semantic search.

In order to achieve a multilingual and semantic portal, further terms in English were needed, since the GND possessed few English labels. A mapping into other controlled vocabularies was required. For this purpose, labels from the DBpedia⁴, Library of Congress Subject Headings⁵ (LCSH), results of the project 'Multi Lingual Access to Subjects'⁶ (MACS) and the WTI thesaurus 'Technology and Management'⁷ were selected. These terms were also available as LOD and are now saved in the local RDF store (Strobel, 2014).

The combination of vocabularies and bilingual labels enables a cross-lingual search in the portal, as shown in figure 2.

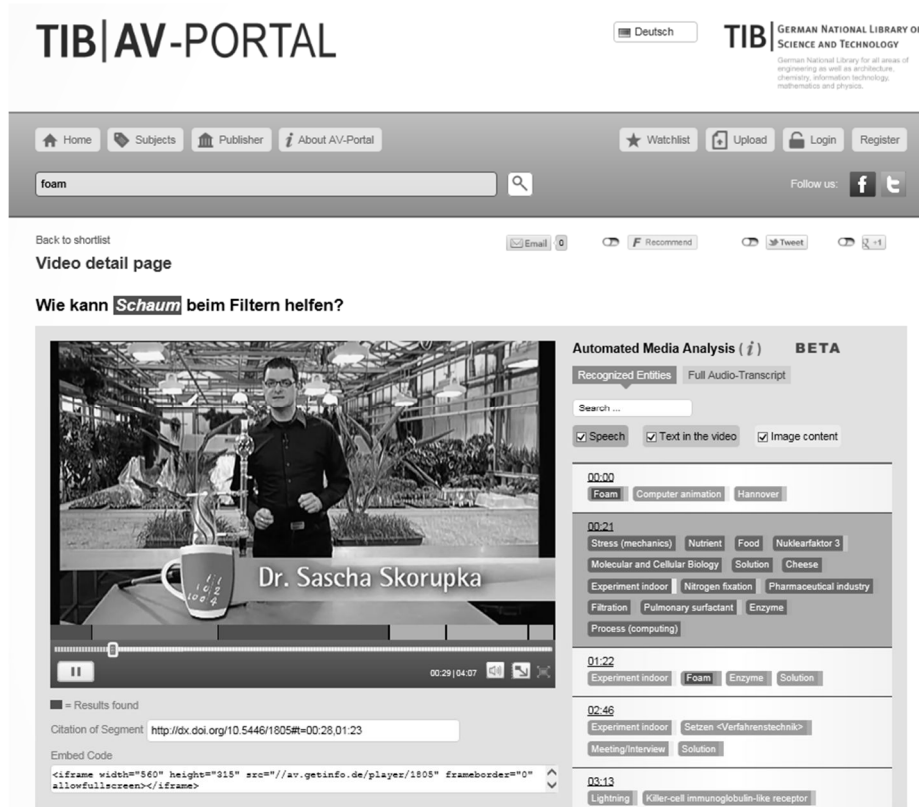


Fig. 2. Video detail page of the TIB|AV-Portal. Results of the cross-lingual retrieval (foam-Schaum) are shown.

This is likely to increase the completeness of relevant search results and improve video indexing based on LOD properties. The semantic search based on linked data is an interesting research topic in continuous development (Waitelonis and Sack, 2011; 2014).

3.2 Data publishing and semantic applications

A fundamental question is how to publish metadata as LOD. According to Zuiderwijk et al. (2012) there is no standardised way to do it and one finds different approaches. Heath and Bizer (2011) also support this idea and describe possible Linked Data publication patterns. The type of input data (queryable structured data, static structure data or text documents) determines the appropriate publishing pattern. Static structured data coming from XML files must be converted into RDF files directly into an RDF store (Health and Bizer, 2011). The data from a RDF store (also known as triple store) can be directly published as Linked Data if they fulfill the Linked Data principles.

³ <https://creativecommons.org/publicdomain/zero/1.0>

⁴ <http://de.dbpedia.org>

⁵ <http://id.loc.gov/authorities/subjects.html>

⁶ http://www.dnb.de/DE/Wir/Kooperation/MACS/macs_node.html

⁷ <http://www.wti-frankfurt.de/index.php/produkte-thesaurus>

Due to the structured nature of the data of the TIB|AV-Portal and the existence of an RDF store, the weaving into Linked Data requires the fulfillment of the principles. To do so, parts of the internal ontology and URI scheme need to become dereferenceable. Therefore, we are mapping and merging our internal ontology with existing LOD vocabularies⁸. Following best practices for library data, Dublin Core Terms (dcterms), Schema.org, Bibliographic Framework Initiative (Bibframe) and The Bibliographic Ontology (BIBO) were identified as best vocabularies for our purposes. Dealing with videos and their automatic annotation made also important the use of the Ontology for Media Resources and the Open Annotation Model as part of the mapping process.

The datasets will be exposed as dumps, as already done by other libraries such as The European Library⁹. Dumps are an efficient and fast way to share data with consumers. However, it requires a content and synchronization planning. As data provider, one has to decide which data are part of the dump, its format and actualization frequency. Following other libraries' actions, we decide to provide 3 dumps (one with the IWF¹⁰ collection, one with the TIB core subjects collection and one with the whole collection). The chosen formats are RDF application/rdf+xml and RDF in text/turtle, and we will offer a quarterly actualization frequency. Further, the metadata can be embedded into the portal in form of RDFa. The last specification of RDFa 1.1¹¹ allows the use of RDF in HTML. This provides more control on data access and may be built on top of an RDF store. RDFa defines attributes for the semantic markup such as *about*, *property* or *vocab* to identify the subject, relations between subject and object, and the vocabularies that are used (Strobel and Marín-Arraiza, 2015). We consider RDFa an appropriate form to provide more relevant information –for instance, about an author– and to retrieve information available elsewhere in our portal or on the web –for instance, articles written by the same author and retrieve through the property *schema:author*–. This should establish a net between our content and related content of other providers, supporting reuse and content promotion.

4 Conclusions

Data organization is a requirement that libraries have to cope with. Nowadays, library data results not just after having managed textual information but also non-textual information.

The TIB|AV-Portal manages scientific videos as part of the non-textual information environment and generates a big amount of data associated with each hosted video. The data is indexed according to international standards and W3C Best Practices. However, it is still possible to go a step further and make the data available to third parties.

In order to execute this action, LOD publishing patterns are being followed. This guarantees that our data are no longer isolated or left behind in the information world, which, unfortunately, happens often with non-textual information in science. Moreover, the data can be reused by others, and being part of Linked Data Cloud¹² improves the visibility of the institution and promotes its content.

The TIB|AV-Portal benefits from the use of LOD. LOD enables the multilingual annotation of the content, the semantic search, and will enable the retrieval of new related content thanks to the interlinking.

Libraries such as the German National Library, The European Library or Europeana are currently working with LOD in aggregation scenarios and providing the data to third parties. Big broadcasting companies such as the BBC also use Linked Data to annotate their content. With the TIB|AV-Portal, the TIB aims to do the same and weave into a linked ecosystem.

There is still a long way to go, particularly in the use of LOD for non-textual materials. However, there are already successful use cases¹³ generated after the Library Data Incubator.

⁸ <http://lov.okfn.org/>

⁹ <http://www.theeuropeanlibrary.org/>

¹⁰ http://www.filmarchives-online.eu/partners/iwf-knowledge-and-media-1/view?set_language=en

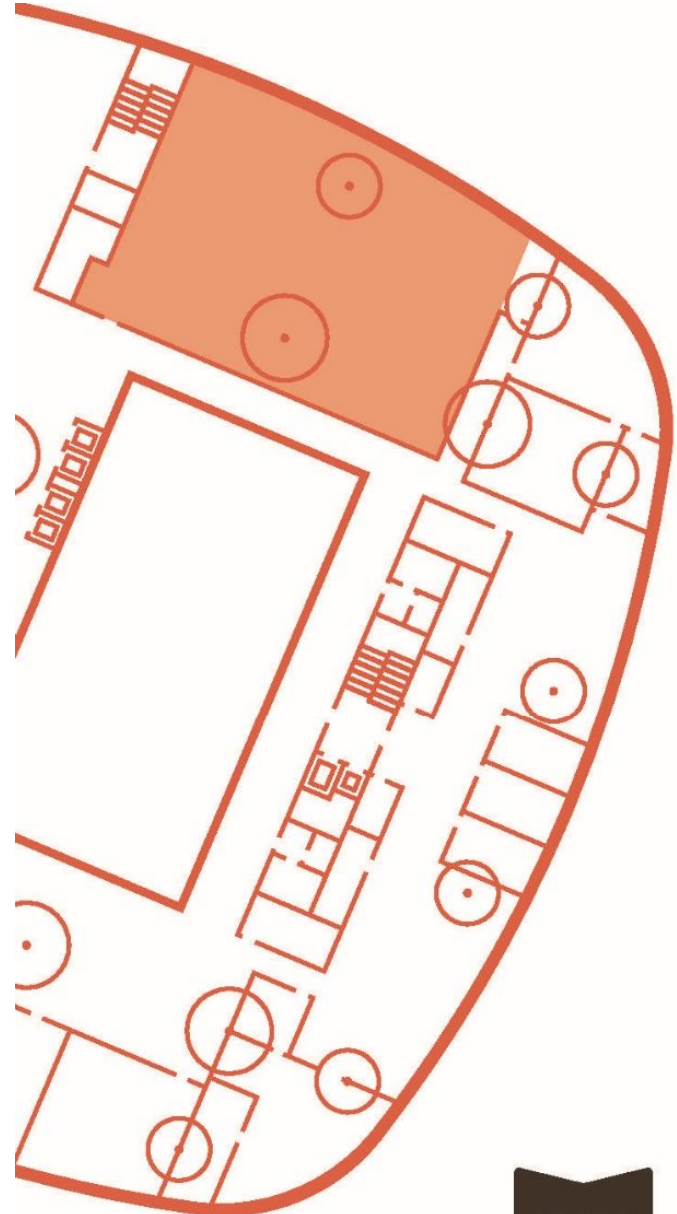
¹¹ www.w3.org/TR/html-rdfa/

¹² <http://lod-cloud.net/>

¹³ <http://www.w3.org/2005/Incubator/ld/wiki/UseCases>

References

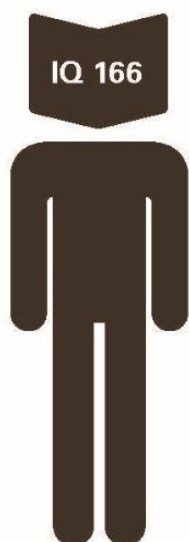
1. Berners-Lee, Tim: *Linked Data-Design Issues*, 2006.
2. Heath, Tom, Bizer, Christian: *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, 2011.
3. Löwgren, Jonas: "The ground was shaking as the vehicle walked pasted me." The need for video in scientific communication. *Interactions*, vol. XVIII n.1, pp. 22–25, 2011. DOI: 10.1145/1897239.1897246
4. Peset, Fernanda; Ferrer-Sapena, Antonia; Subirats-Coll, Imma: Open data y Linked open data: su impacto en el área de bibliotecas y documentación. *El profesional de la información*, 2011, vol. 20, n. 2, pp. 164-172. DOI: 10.3145/epi.
5. Pohl, Adrian, Danowski, Patrick: *Linked Open Data in der Bibliothekswelt: Grundlagen und Überblick*, 2013.
6. Ríos-Hilario, Ana; Martín-Campo, Diego; Ferreras-Fernández, Tránsito: Linked data y linked open data: su implantación en una biblioteca digital. *El caso de Europeana*. *El profesional de la información*, 2012, vol. 21, n. 3, pp. 292-297.
7. Sack, Harald, Waitelonis, Jörg: **Linked Data als Grundlage der semantischen Videosuche mit Yovisto**, in T. Pellegrini, H. Sack, and S. Auer (Hrsg.): [Linked Enterprise Data – Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien](#), Berlin: Springer Berlin, 2014, pp. 263-287, ISBN: 978-3-642-30273-2.
8. Saorín, Tomás. Cómo linked open data impactará en las bibliotecas a través de la innovación abierta. *Anuario ThinkEPI*, 2012, vol. 6, pp. 288-292.
9. Strobel, Sven, Marín-Arraiza, Paloma: Metadata for Scientific Audiovisual Media: Current Practices and Perspectives of the TIB|AV-Portal. E. Garoufallou et al. (Eds.): *MTSR 2015, CCIS 544*, pp. 159–170, 2015. DOI: 10.1007/978-3-319-24129-6_14
10. Strobel, Sven: Englischsprachige Erweiterung des TIB|AV-Portals. Ein GND/ DBpedia-Mapping zur Gewinnung eines englischen Begriffssystems. In: *obib*. *Das offene Bibliotheksjournal*, vol. 1, pp. 197-204, 2014. DOI: 10.5282/o-bib/2014H1S197-204.
11. Waitelonis, Jörg, Sack, Harald: Towards Exploratory Video Search Using Linked Data. In: *Multimedia Tools and Applications* 53, pp. 1-28, 2011. DOI: 10.1007/s11042-011-0733-1.
12. Zuiderwijk, Anneke, Jeffry, Keith, Janssen, Marijn: The Potential of Metadata for Linked Open Data and its Value for Users and Publishers. In: *JeDEM*, 2012, vol. 4, n. 2, pp. 222-244.



National Technical Library

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

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



As corresponds to its statutes, NTK manages – among others – the project of building the **National Repository of Grey Literature.**

The project aims at gathering metadata and possibly full texts of grey documents in the fields of education, science and research. The NTK supports an education in the field of grey literature through annual seminars in the Czech Republic.

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

NTK Univers

Národní technická knihovna
National Technical Library



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

Azzedine Bouderbane, Nadjia Gamouh, and Hadda Saouchi
University Constantine 2, 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 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 researcher 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 accessing 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 download 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 Kms to reach a university library in Algiers. He/she may return to his/her city frustrated without the expected document to cross again 2000 Kms. 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-strategie-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 « Éducatifs 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.



Appendix

The questionnaire

- I. General information
 1. **Position**
 - a. University teacher researcher
 - b. PhD researcher
 2. **Sex**
 - a. Male
 - b. Female
- II. Documents 'use by researchers
 3. **Select from the list below the most essential and suitable documents you prefer using when you do research**
 - Books
 - Articles in journals
 - Theses
 - Reports and official publications
 - Specialized electronic sites
 4. **Cite documents that are easy for access:**
.....
 5. **Cite documents that are difficult for access:**
.....
 6. **Specify the type of difficulties you face with the documents stated in question 5**
.....
- III. The use of the PNST
 7. **How often do you use the portal?**
 - Very often
 - Often
 - Rarely
 - never
 8. **How did you hear about the PNST?**
.....
 9. **Do you use the statistics provided by the portal?**
 - Less than 25 %
 - 25 – 50 %
 - 50 – 75 %
 - More than 75 %
 10. **What do you think about the linguistic aspect when surfing on the portal?**
.....
 11. **Give your appreciation of the guide provided by the portal**
 - Useless
 - A little bit useful
 - Useful
 - Very useful
 12. **How can you evaluate the provided links on the portal?**
 - Uninteresting
 - A little bit interesting
 - Interesting
 - Very interesting
 13. **Does the portal respond to your needs in terms of information**
 - Not at all
 - A little bit
 - Suitable
 - appropriate
 14. **Give your global appreciation of the portal's design**
.....
 15. **What are the constraints you face when exploiting the portal?**
.....
 16. **Are there technical problem you encounter when surfing on the portal?**
.....
 17. **Give your appreciation about the usefulness of the portal**
.....
 18. **Is there a relation between the PNST and scientific research?**
.....
 19. **Could you give suggestions about the portal?**
.....

Grey Literature Sources in Historical Perspective: Content Analysis of Handwritten Notes

Snježana Ćirković

Faculty of Philology, University of Belgrade, Serbia

Abstract

This paper describes historical perspective of grey literature in comparison to the postulates of Serbian enlightenment in the 18th century. The theoretical part of this paper describes comparative analysis between main postulates of grey literature given in *Pisa Declaration on Policy Development for Grey Literature Resources* (Pisa, Italy) on May 16, 2014 in comparison to the first manifest of Serbian enlightenment written by one of the greatest Serbian enlightener Zaharija Orfelin. Orfelin has written this manifest in the preface of the Journal *Slaveno-Serbian Magazin*, founded in Venice (Italy) in 1768.

The second part of this paper describes the handwritten notes, as one of the grey literature forms, made by Zaharija Orfelin on the books in his private library. This library presents one of the first private Serbian libraries from the 18th century, it is a rich collection of the most representative books from that period in different languages (German, Italian, French, Russian, Latin and Slaveno-Serbian) and represents invaluable cultural heritage. Using the method of content analysis, this paper describes these handwritten notes, and their role as grey literature source in reconstructing the biography of Zaharija Orfelin compensating for lack of other primary biographical and historical printed sources.

Introduction

The 18th century is the century of migrations and great exodus of the Serbian people and re-establishment of Serbian national identity - the century in which modern Serbian civic culture was born. One of individuals, who contributed to this creation, was Zaharije Orfelin.

Mr Ofelin was a unique personality in his own way, both in terms of erudition and the legacy he left in Serbian art, specifically in the fields of engraving, calligraphy, poetry and education. "His contemporaries glorified him as the most talented and learned among Serbian artists in the early 18th century. He was the first who reached the highest level of engravers' skills and became respected in Vienna, and the sole artist who established himself as a serious historian and wrote a monumental work on Peter the Great" (Todorović, 2006).

Orfelin was also a teacher, owner of the largest library of his time in the Archbishopric and the founder of the first publishing house. He was recognized and respected member of Jacob's Schmutzers Imperial engraving Academy and he received imperial rewards and grants for his calligraphic work. Orfelin's life was filled with unusual diversity and strong contrasts due to both private and professional circumstances. He frequently changed places, countries and posts.

He was an erudite and the list of his occupations is very long. Let me highlight some of them. Orfelin was first Serbian poet who published his poems; he established the first Serbian and South Slavic Journal *Slaveno-Serbian Magazin*, in Venice in 1768; compiled the first Serbian recommended bibliography; wrote the first history of Peter the Great in a Slavic language, and all this for common good, which was the main postulate of his entire work.

Zaharija Orfelin – A Short Biography

Zaharija Orfelin was born in the city of Vukovar, Austrian Monarchy, in 1726. First twenty years of his life are still unknown to the Orfelin's biographers. What we know is that he was a teacher in a Slavic School in Novi Sad, and in 1757 he moved to Sremski Karlovci, where he became the secretary to the Serbian Metropolitan Pavle Nenadovic. This small town in today's Vojvodina was Serbia's most important political and cultural centre at the time. These five years which Orfelin spent in Sremski Karlovci were one of his most productive periods. Together with the Metropolitan Pavle Nenadovic Orfelin founded the *Copper Publishing House*, where he published his first poems, translated books, and created graphics and engravings, inspired by the work of his contemporaries.

Some of Orfelin's biographers assume that his inquisitive mind took him to Venice, where he had an opportunity to publish more books and get in touch with the contemporary West-European thought and ideas of enlightenment.

Other Orfelin's biographers assume that he had to move from Sremski Karlovci, because of political pressure he was exposed to. Namely, in 1761 he published a poem *The Lament of Serbia*. He first published it anonymously in Venice in 1761. In this poem he criticized Austrian

establishment and the high ecclesiastical circles in the Archbishopric, while also emphasizing the difficult position and unjust treatment of the Serbs under the Imperial protection. Despite the fact that it was unsigned, the Orfelin's authorship was soon discovered, which gave him a status of *persona non grata* and had to leave Sremski Karlovci.

In 1764, Orfelin came to Dimitrios Theodosios's publishing house in Venice, where he spent the next six years before moving back to Karlovci. Theodosios, who began to print Serbian books in Venice in 1758, needed Orfelin to edit and proofread the abundant material being brought to him. Dimitrios Theodosios's publishing house remained the only place for publishing of Serbian literature and printed books in Russo-Slavonic until the emergence of Kurtzbek's publishing house in Vienna in 1770. The Orfelin's contributions as a redactor and editor in this publishing house were huge. These six years which Orfelin spent working together with Dimitrios Theodosios are described as a "golden age" of this publishing house. (Ćurčić, 2002)

Slaveno-Serbian Magazin

Slaveno-Serbian Magazin is the first Serbian and Slavic journal published in Venice in 1768. Zaharija Orfelin published this journal anonymously. A role model for this journal was Russian journal *Ežeměsečnaja Sočineniâ*.

Slaveno-Serbian Magazin contained short novels, poems, numerous articles from different scientific disciplines and first Serbian recommended bibliography. The most important part of this journal is the preface, written on 96 pages, and described as the *Manifest of Serbian Enlightenment*. That was the first time that ideas of rationalism, characteristic for the 18th century, were published in Serbian language.

For Orfelin, the 18th century was a century where education and knowledge cease to be any more the privilege of certain social classes, but rather should become available to everyone, to all social classes. Everything is focused on community, common good and humanization of a man as a being. For rapidly growing need for knowledge sharing, books were not any more the best and the fastest way. Journals became much better instrument and form. They offer a great deal of knowledge from different subjects, consist of short articles and offer a lot of different types of information. The magazine form was Orfelin's attempt to compensate for a lack of public schools and books. We can summarize the main postulates of this preface as following:

- Sharing knowledge among all social classes;
- Common good;
- Spreading knowledge through new printed forms (journals, almanacs, manuals) instead of books, in lack of schools and official education;
- Relief from clerical influences.

Although the *Slaveno-Serbian Magazin* was published two centuries before, the ideas standing in background, the enlightenment ideas about spreading knowledge, or better to say "knowledge for everyone", remained till today as an intellectual legacy in many different ways. One of them, appeared some 250 years later, as the *Pisa Declaration on Policy Development for Grey Literature Resources* published in Italy in May, 2014. In this paper I will make a comparative analysis of these two documents, but let us first give a look in Orfelin's library and handwritten notes founded on the books as a grey literature source.

Private Library of Zaharija Orfelin

This library presents one of the first private Serbian libraries from the 18th century, it is a rich collection of the most representative books from that period in different languages (German, Italian, French, Russian, Latin and Slaveno-Serbian), and represents invaluable cultural heritage. Orfelin's library is in every aspect an important source for the better understanding of his personality. Book collection of one man opens a door to his spirit and character, gives us a way to perceive the height, depth and direction of his education, and his interests - literary, scientific and personal one.

When we analyze the character of the Orfelin's library, we see that he was buying some books to expand his intellectual horizons, some of them to gain a practical knowledge and skills, some of them he was buying to educate others, while some he used as sources for writing *the History of Peter the Great*. (Ćirković, 2013)

In the preface to this book Orfelin says: "Many historians, when they want to write a history book, are happy because they can use state, public or academic libraries. For me who lives in faraway regions was needed to travel to Germany and buy German books or to Russia to buy Russian books, and all that as much as was allowed by my funds."

If we accept Orfelin's statement that he was an autodidact, then this rich library, which contains valuable and significant scientific works from various disciplines (Theology, History, Linguistics, Medicine, Astronomy, Botanic, Arts, Bibliography, Pedagogy, Geography, Physics, Law, Literature, Stylistics, Numismatics, etc.), could be seen as his attempt to compensate the knowledge which he missed during his youth days.

Important additional values of this library are the handwritten notes written by Orfelin himself on the margins of the books. Every book from Orfelin's library has his signature, in a form of *Ex Libris*, and confirmation that the book belongs to him.

We can find also numerous notes about book content, date and time when and where Orfelin bought the book, even to whom he borrowed it. On many books in his library we find different types of bibliographies and book lists, which Orfelin used as a reference tools for his literary and scientific work.

Handwritten notes as historical sources

Any remnants of the past can be considered as a historical source. That might be a document, a piece of art, some ephemeral or any other object. These are all sources because they provide us in different ways with information which can add to the sum of our knowledge about the past. However sources only become historical evidence when they are interpreted by the historians and put in the historical perspective and contents. There are three main types of historical sources:

- I. **Primary source** - an original document containing firsthand information about a topic (diaries, interviews, letters, handwritten notes, original works of art, photographs, speeches, works of literature);
- II. **Secondary source** - commentary/discussion about a primary source; interpretation of primary sources (biographies, dissertations, indexes, abstracts, bibliographies, journal articles, monographs);
- III. **Tertiary source** - presents summaries or condensed versions of materials, usually with references back to the primary and/or secondary sources (dictionaries, encyclopedias, handbooks, etc.)

Orfelin's handwritten notes - Content analysis

As previously mentioned, Orfelin's library serves as source for better understanding of his personality and his intellectual profile. Numerous handwritten notes give us an insight into his travels, attitudes, habits, personality, literary and scientific interests. Based on typology of most common handwritten notes on the books in Orfelin's library, we can distinct four types of handwritten notes:

1. **Form of Ex Libris** – Signature on the books - *Z.O., Zacharias Orphelin, Zaccaria Orfelini...*;
2. **Notes about the book content** – *Author lies (лжесть ауторъ), materialist (материалистъ)*, explanation to some terms in books..;
3. **Notes about date and place of book purchase** - *Das Buch habe ich gekauft um zwanzig Groschen von den Buchhaltern 24 July 1751 in Ofen*;
4. **Different types of title lists, short bibliographies** – *КНИГИ на нѣмецкомъ и славенскомъ ЯЗЫКЪ...*

All these notes are important for Orfelin's biographers and historians. Their importance is reflected in following:

- For reconstructing of Orfelin's biography; in lack of other primary biographical sources;
- For creating an intellectual Profile of Zaharija Orfelin;
- For better understanding his personality, habits, personal interests, etc.

Grey Literature – Overview

The metaphor "the chameleon of information resources" maybe describes the grey literature (GL) the best. As the author further states, "it can constitute virtually anything and be written for and by anyone in almost any format" (Rucinski, 2015: 552). The ephemeral and variable nature of grey publication types, editions, and formats makes it hard to describe and define. Despite that, there are many definitions of grey literature. The most widely accepted one is the *Luxemburg definition*. The newest one is Prague definition from *The Twelfth GL International Conference (Prague, 2010)*, which defines GL as follows:

Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers i.e., where publishing is not the primary activity of the producing body (Schöpfel, 2011: 17)

Grey literature is commonly described as a field in library and Information science that deals with the production, distribution, and access to multiple document types produced on all levels of government, academics, business, and organizations in electronic and print formats not controlled by commercial publishing, i.e. where publishing is not the primary activity of the producing body. Examples of grey literature include patents, technical reports from government agencies or scientific research groups, working papers from research groups or committees, white papers, and preprints.

According to the Wikipedia “Common grey literature publication types include reports (annual, research, technical, project, etc.), working papers, government documents, and evaluations. Organizations that produce grey literature include government departments and agencies, civil society or non-governmental organizations, academic centers and departments, and private companies and consultants.

Grey literature may be made available to the public, or distributed privately within an organization or group, and often lacks systematic means of distribution and collection. The standard of quality, review and production can also vary considerably. Grey literature is therefore often difficult to discover, access, and evaluate.”¹

Pisa Declaration on Policy Development for Grey Literature Resources

The Pisa Declaration on Grey Literature was developed at a forum held in Pisa, Italy in May 2014 organized by GreyNet² and the *National Research Council of Italy*. It is currently translated into 20 languages worldwide (English, Armenian, Bulgarian, Croatian, Czech, Dutch, French, German, Greek, Hindi, Hungarian, Italian, Japanese, Korean, Macedonian, Russian, Serbian Cyrillic, Serbian Latin, Spanish, Tagalog, Turkish).

*The Pisa Declaration*³ provides a 15-point roadmap that should serve as a guide for organizations involved with the production, publication, access and use of grey literature well into this 21st century.

Until now, the problem was the lack of cooperation and coordination between and among organizations dealing with grey literature. *The Pisa Declaration* marks an end to *ad hoc* policy and decision making with regard to grey literature resources.

The main points set out in *the Pisa Declaration* can be grouped into five categories:

1. Organizational commitment to open access and the sharing of open data standards;
2. Commitment to research and education, where recognition and reward is associated with quality grey literature, and where attention is given to good practices in the field;
3. Commitment to address and safeguard legal issues inherent to grey literature by exploring the various types of licensing agreements now available and by fostering constructive relations with commercial publishers;
4. Commitment to sustainability linked to a financial prerequisite. Identifying funding and grants for special collections and repositories, commitment to long-term preservation, and investments in new technologies;
5. Firm technical commitment to continued online services and further cross-linking of textual and non-textual content – a commitment that ranges from tackling broken links to

¹ https://en.wikipedia.org/wiki/Grey_literature

² <http://www.greynet.org/>

³ <http://greyguide.isti.cnr.it/>

facilitating interoperability regardless of the system or portal in which grey literature and its accompanying data are housed. (Farace, 2014)

“It is in this way that the Pisa Declaration can revel in the strengths and opportunities that grey literature offers, while at the same time exposing the weaknesses and threats facing our community. No longer are we resigned that grey literature is hard to find, but instead how can we best search and access it. No longer hold in question its worth and value, but instead set out the review process it has undergone. And, no longer hesitate as to whether it is published or not, but instead cite and reference grey literature – make it openly public isn’t that what published means?” (Farace, 2014: 3)

Comparative Analysis – *Slaveno-Serbian Magazin* and *Pisa Declaration*

As I have already highlighted, the four main postulates of *Slaveno-Serbian Magazin* are:

- Sharing knowledge among all social classes;
- Common good;
- Spreading knowledge through the new printed forms (journals, almanacs, manuals) instead of books, in lack of schools and official education;
- Relief from the clerical influences.

These postulates can be also found in *Pisa Declaration*, in the form of open access, knowledge transfer and innovation:

“In order to realize the benefits of research and information for scholarship, government, civil society, education and the economy, We, the signatories to this declaration, call for increased recognition of grey literature’s role and value by governments, academics and all stakeholders, particularly its importance for open access to research, open science, innovation, evidence-based policy, and knowledge transfer.”⁴

Sharing knowledge in the 18th century was one of the essential things that should be done in order to educate all social classes and bring the knowledge to everyone through different and innovative ways. Open access as one of the postulates given in *Pisa Declaration* points at sharing knowledge between scientists, knowledge for everyone, which take us back to the 18th centuries’ enlightenment idea of common good.

Both of these documents were published in Italy, with the time distance of almost 250 years, but the idea of spreading knowledge, innovative ways and forms of knowledge transfer remain the same till today. Mutual interaction of these postulates is presented in the Figure below.

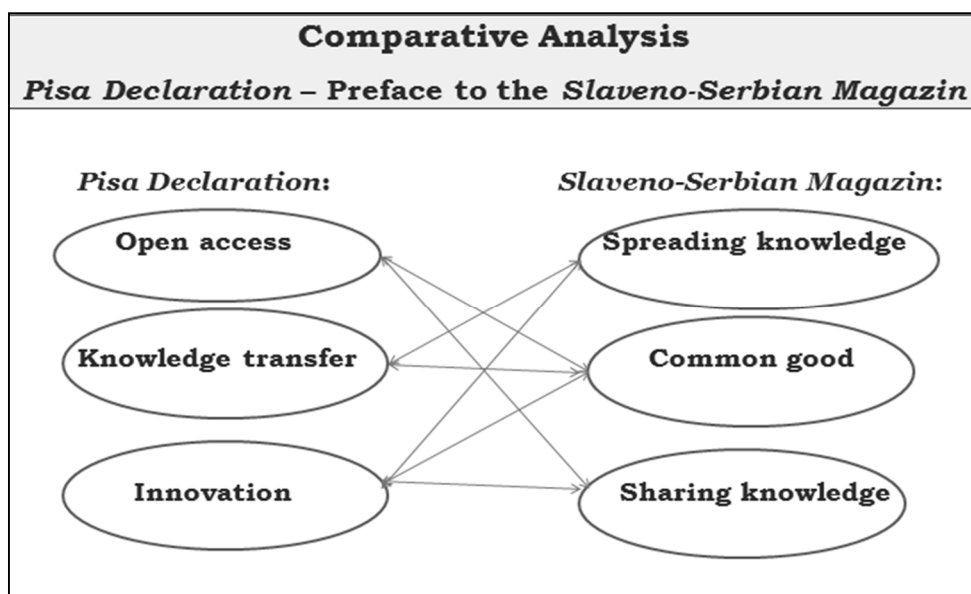


Fig. 1.: Comparative Analysis: *Pisa Declaration* and *Slaveno-Serbian Magazin*

⁴ http://www.greynet.org/images/Pisa_Declaration,_May_2014.pdf

Conclusion

Handwritten notes are important type of grey literature sources and in order to benefit from them, we need to:

- I. **Include handwritten notes in corpus with other grey literature sources** (such as: research and technical reports, briefings and reviews, evaluations, working papers, conference papers, theses, and multimedia content);
- III. **Develop standards for bibliographical description and implement them in applicable metadata structures;**
- IV. **Promote use of notes which are today mostly comments.**

Notes and comments as a source of grey literature went at least through four different development phases:

1. In the 18th century there were handwritten notes on the margins of books;
2. In the 20th century we had sticky notes in the books which we used to make highlights and comments, while protecting the books;
3. Today, in 21st century, handwritten notes developed into a new form of comments on eBooks. For example Ebrary, where the software gives us an opportunity to communicate with other users, and where all the notes, although still grey, are openly visible to the broad range of users, and not hidden in library treasures on margins of some old books, waiting to be discovered and tell us some new story. From grey to bright...forms are changing, but message remains.

References

- Blaaij, C. de (2007). The use of grey literature in historical journals and historical research: A bibliometric and qualitative approach. *GL9, Antwerp (Belgium)*. Retrieved from: <http://hdl.handle.net/10068/697875>
- Ćirković, S. (2013). *Biblioteka Zaharija Orfelina*. Beograd : Institut za teološka istraživanja, Pravoslavni bogoslovski fakultet Univerziteta u Beogradu.
- Ćurčić, L. (2002). *Knjiga o Zahariji Orfelinu*. Zagreb : Srpsko kulturno društvo Prosvjeta.
- Farace, D. (2014). The GreyGuide Repository and Web-Access Portal: GreyNet's Response to the Pisa Declaration. *Conference on Grey Literature and Repositories: proceedings 2014: the Value of Grey Literature in Repositories* [online]. Prague: National Library of Technology. Retrieved from: http://repozitar.techlib.cz/record/808/files/idr-808_4.pdf
- Orfelin, Z. (2000). Slaveno-serbski magazin: predgovor. *Ljetopis Srpskog kulturnog društva "Prosvjeta"*. Sv. 5., p. 189-195.
- Rucinski, T. L. (2015). The Elephant in the Room: Toward a Definition of Grey Legal Literature (December 2015). *Law Library Journal, Vol. 107:4, 2015*. Retrieved from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2708884
- Schöpfel, J. (2010). Towards a Prague Definition of Grey Literature. *Twelfth International Conference on Grey Literature: Transparency in Grey Literature. Grey Tech Approaches to High Tech Issues. Prague, 6-7 December 2010, Dec 2010, Czech Republic*. pp.11-26. Retrieved from: <http://hdl.handle.net/10068/700015>
- Todorović, J. (2006). *An Orthodox Festival Book in the Habsburg Empire – Zaharija Orfelin's Festive Greeting of Mojsije Putnik's in 1757*. London: Ashgate.

Grey Literature citations in the age of Digital Repositories and Open Access

Silvia Giannini and Stefania Biagioni, Institute of Information Science and Technologies ISTI-CNR;
Sara Goggi and Gabriella Pardelli, Istituto di Linguistica Computazionale, ILC-CNR, Italy

Abstract

The work measures grey citations in the years 2012, 2013 and 2014 and then describes the features of GL documents cited in different areas of knowledge: Computational Linguistics, Computer Science and Engineering. With the aim of surveying a wide and varied range of resources, we selected a sample data based on the bibliographical references of articles contained in four journals - all indexed by Scopus Citation Database and ISI Web of Science, with an Impact Factor (IF) over the last three years - and two proceedings of international conferences held in 2012 and 2014.

1. Introduction

About ten years ago we studied the impact of grey literature (GL) on conventional literature by observing the impact of grey citations in two different scientific fields, «after the growth in the use of the WWW» [Di Cesare et al. 2005]. Over the last decade the international scientific community and its players have undergone (and still undergo) essential changes with respect to the representation and dissemination of knowledge. The formalization of the Open Access (OA) model dates back to the beginning of the last decade: OA is a movement born in the 1990s with the purpose to contrast the monopoly of commercial publishing houses thus making knowledge accessible for free without violating the intellectual property rights or threaten the quality of a scientific work. Along with the success of OA, the prototype of the 90s e-print archives evolved into the current digital repositories (institutional or subject based). The new paradigms of scientific communication and the most advanced computer science technologies fostered the achievement and use of these novel infrastructures: in particular, institutional repositories have gained a specific and relevant role in storing, preserving and disseminating scientific information. The international Declarations of Budapest 2002, Bethesda 2003 and Berlin 2004 contributed to the definition of OA principles and opened the way to the movement from a normative point of view thus fostering the promulgation of national laws and European legislation as well (reference n° 2012/417/UE). Nowadays many academic and research institutions choose to adhere to the OA movement by issuing policies which compel to deposit/file their research products in OA repositories. Nevertheless the debate on this themes is extremely heated and many issues still need to be cleared up: from problems related to the quality of open access products, the peer-review processes, the integration between the OA world and the evaluation of research to – last but not least - the impact of citations of OA works [Guerrini 2010; De Bellis 2005; Eysenbach 2006]. In our current digital era bibliographical citations has gained a strategic role within the mechanisms of scientific communication, especially due to the implementation of the citation indexing services [Cassella 2011]. Citation has thus become «...the currency in scientific communication trade. It is a small denomination bill (quoting does not cost that much) but with a very high symbolic buying power...» [De Bellis 2014]¹. Besides it is undeniable that from a bibliometric point of view the two more widely recognized and used standards are those based on the number of citations per article and on the impact factor per journal.

In this scenario, it seemed interesting to investigate once again the “world” of scientific citations for proving if - and eventually to which extent – this “revolution” in the communication of knowledge might actually reflect on the GL approach to citations. Today institutional archives allow to store and make accessible any research “product”, whether “official” publications or various types of grey literature: from the more traditional technical reports or dissertations to the newest datasets, experiments, software, web sites, blogs.

Our hypothesis is then driven by the idea that grey literature, more easily identified and accessed, may have a greater visibility. Similarly it is possible to assume a greater impact of GL citations on the overall total of citations.

¹ From the Italian original work: «...moneta corrente nel commercio della comunicazione scientifica ufficiale. Moneta di piccolo taglio (costa poco citare), ma dal potere d'acquisto simbolico non indifferente...» (DeBellis 2014).

2. Materials and method

We analyzed a sample of journals indexed by the Science Citation Index (SCI) and included in the ISI-Journal Citation Report (JCR)². Two journals on four are Open Access journals: *EURASIP Journal on Advances in Signal Processing* and *Computational Linguistics* (Table 1). Moreover we analyzed a sample of two conference proceedings in order to evaluate any differences in the citation model (Table 2).

Journals Titles	IF* 2012	Rank* 2012	IF* 2013	Rank* 2013	IF 2014	Rank* 2014
ACM Transactions on Information Systems	1.070	59/132	1.300	53/135	1.021	70/139
EURASIP Journal on Advances in Signal Processing	0.807	155/243	0.808	164/248	0.777	170/249
Computational Linguistics	0.940	72/115	1.468	49/121	1.226	72/123
Language Resources and Evaluation	0.659	79/100	0.518	94/102	0.619	89/102

Table 1 – Sampled Journals

Proceedings Titles	Years
JCDL - ACM/IEEE-CS Joint Conference on Digital Libraries	2012 and 2014
EACL - Conference of the European Chapter of the Association for Computational Linguistics	2012 and 2014

Table 2 – Sampled Proceedings

The chosen journals are all indexed under the ISI-JCR subject category “Computer Science” (CS), except for the *EURASIP Journal on Advanced in Signal Processing* (EURASIP) which is under the subject category ENGINEERING, ELECTRICAL & ELECTRONIC (E&E). *ACM Transactions on Information Theory* (ACM TOIS) is under the sub-category “Information systems”; *Computational Linguistics* (CL) is under the sub-categories “Artificial Intelligence” and “Interdisciplinary Applications”; *Language Resources and Evaluation* (LR&E) is under the sub-category “Interdisciplinary Applications”. Table 1 compares the IF and rank in CS and E&E subject categories in the selected years.

Scopus citation database places the journals CL and LR&E in areas related also to the Humanities and Social Sciences: Language and Linguistics for CL; Language and Linguistics, Education and Library and Information Sciences for LR&E. Indeed, Computational Linguistics is a discipline that draws contributions from different fields of study such as linguistics, psychology, mathematics and statistics, in addition to computer science.

For all these reasons we considered the selected journals and conference proceedings as belonging to two different scientific communities: Computer science and Engineering and Computational Linguistics.

The journals present a value of IF substantially stable in the considered timespan, with the exclusion of CL showing a higher IF in 2013. This value also determines a significant shift of rank in the same year and makes CL the journal with the higher IF.

We extracted the information directly from primary sources, that is the bibliographical references of the articles in the selected journals and proceedings. The corpus was built by grouping the gathered data in six informative classes: year, issue number, bibliographical reference, kind of document – Grey (G) or Published (P), document type, standardized document type (Table 3).

Year	Issue	Reference	G/P	Doc. Type	SDType
2012	38(2)_2	Horn, Laurence R. 1972. On the Semantic Properties of Logical Operators in English. Ph.D. thesis, UCLA. Distributed by the Indiana University Linguistics Club, 1976.	G	PHD	Thesis

Table 3 – example extracted from *Computational Linguistics* (2012)

² ISI-JCR Science Edition 2012, 2013, 2014.

We analyzed 40.511 bibliographical references on 1.270 articles. For each journal and proceedings we counted:

1. the number of articles provided with references³;
2. the number of references in each article;
3. the number of GL references in each article;

For each GL reference we examined: document type; year of publication; GL linked references according to the following criteria:

- 1) Definition of GL starting from the York recommendations (1978) and the later integrations to its definition;
- 2) Classification of documentation produced by *no-profit* Associations, Institutions and Publishers (e.g. ACL Anthology, ISCA archive, OA journals) as grey literature;
- 3) Use of specialized indexes, catalogues and the Google search engine to clarify unclear or incomplete citations;
- 4) Categorization of GL document types as follows:
 - *ARTICLE* includes: journals, newspapers, newsletters and magazines articles;
 - *BLOG/FORUM*;
 - *BOOK/BOOK CHAPTER*;
 - *CONFERENCE PAPER* includes: papers presented at conferences, seminars, workshops, meeting;
 - *CORPORA* includes: downloadable linguistic resources;
 - *COURSE MATERIAL* includes: tutorials and teaching material;
 - *DATABASE*;
 - *DATASET*;
 - *DELIVERABLE*;
 - *GUIDELINES and NORMATIVE DOCUMENT* includes: standards, guidelines, protocols;
 - *PATENT*;
 - *PREPRINT/POSTPRINT* includes: documents "submitted-to", "to-be-published", "in press", "forthcoming"; "accepted"; "to appear";
 - *POSTER/PRESENTATION* includes: demo, poster and presentation;
 - *REPORT* includes: working notes, technical reports, white papers, working papers, research reports, project reports, discussion papers, occasional papers;
 - *SOFTWARE* includes: only downloadable software;
 - *TECHNICAL DOCUMENTATION* includes: user guides, manuals, technical specifications and technical documentation of computer programs and for statistical surveys;
 - *TERTIARY DOCUMENT* includes: dictionaries, catalogues and encyclopedia entries;
 - *THESIS* includes: PhD thesis, dissertations, master thesis;
 - *UNDEFINED* includes: all documents that could not be identified;
 - *WEBSITE* includes: simple URLs' or home pages.

We measured the different impact of GL on the different areas of knowledge, using the following indicators:

- 1) the frequency of GL citing (i.e. the proportion of GL references out of all the references examined);
- 2) the frequency of GL use (i.e. the proportion of articles with GL citation, out of all articles examined);
- 3) the intensity of GL use (i.e. the frequency of GL citing divided by the frequency of GL use).

3. Analysis of data and results

3.1 Frequency of GL citing

In our corpus the frequency of GL citing is 24% out of all references examined and it varies from a minimum of 7,8 to a maximum of 62,9.

³ Including: "editorial", "obituary", "squibs", "book review"; "report", "brief report", "project note", "editors' notes", "introduction" etc.

Title	2012			2013			2014		
	Number of references	Number of GL references	Frequency of GL citing (%)	Number of references	Number of GL references	Frequency of GL citing (%)	Number of references	Number of GL references	Frequency of GL citing (%)
ACM TOIS	1.413	285	20.2	1.097	135	12.3	1.096	150	13.7
Computational Linguistic	1.575	739	46.9	2.008	1.263	62.9	1.958	1.158	59.1
EURASIP Journal on Advances in SP	7.876	616	7.8	5.805	459	7.9	5.651	455	8.1
Language Resources and Evaluation	1.220	384	31.5	2.052	740	36.1	1.267	495	39.1
EACL Proceedings	2.307	884	38.3	ni	ni	ni	2.368	1.332	56.3
JCDL Proceedings	1.304	329	25.2	ni	ni	ni	1.514	384	25.4

Table 4 (ni=not included)

What is remarkable in Table 4 is:

- CL presents the highest frequency of GL citing as well as the highest IF;
- the frequency of GL citing in JCDL proceedings is stabler than in EACL proceedings;
- the frequency in journals alternates increase and decrease of the value over the years;
- the frequency of GL citing is higher for journals and proceedings belonging to the area of Computational Linguistics than for journals belonging to the area of Computer Science-Information Systems and the journal of E&E area;
- EURASIP presents the lowest number of GL citations: this journal shows a good stability of frequency of GL citing, followed by LR&E.

3.2 Frequency of GL use

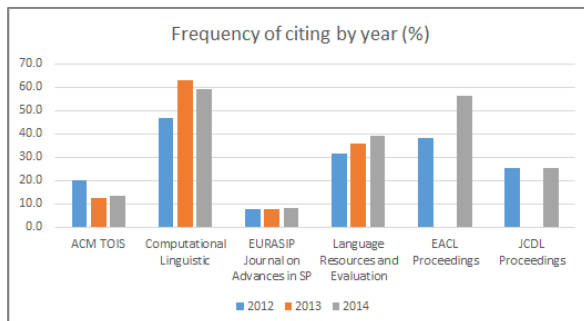
The frequency of GL use shows the percentage of articles with at least one GL citation out of the overall amount of articles: it is globally very high and varies from a minimum of 69,8% to a maximum of 100%.

Title	2012			2013			2014		
	Number of articles	Number of articles with GL references	Frequency of GL use (%)	Number of articles	Number of articles with GL references	Frequency of GL use (%)	Number of articles	Number of articles with GL references	Frequency of GL use (%)
ACM TOIS	25	25	100.0	22	22	100.0	21	18	85.7
Computational Linguistic	36	32	88.9	35	33	94.3	34	33	97.1
EURASIP Journal on Advances in SP	252	176	69.8	188	136	72.3	183	128	69.9
Language Resources and Evaluation	31	31	100.0	56	54	96.4	31	31	100.0
EACL Proceedings	85	82	96.5	ni	ni	ni	78	78	100.0
JCDL Proceedings	96	81	84.4	ni	ni	ni	97	75	77.3

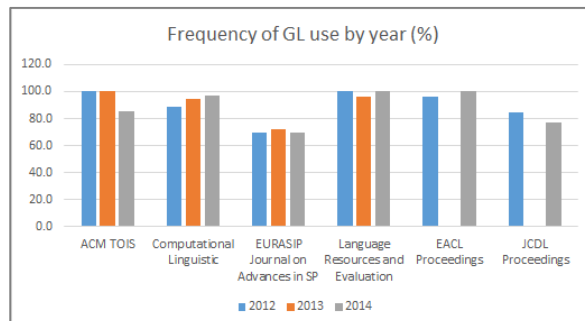
Table 5

Table 5 shows:

- 1) the frequency of GL use is very high for each journal and proceedings;
- 2) the frequency of GL use increases steadily for CL and for EACL proceedings, decreases for ACM TOIS in 2014 while increases for LR&E and EURASIP in 2013 but decreases again in 2014 (alternation). Finally the frequency for JCDL proceedings decreases from 2012 to 2014.



Graph 1



Graph 2

Graphs 1 and 2 show the variability of frequency of GL citing and use for each source over the years.

3.3 Frequency/Intensity of GL use

The intensity of GL use varies from a minimum of 10,9% (*EURASIP*) to a maximum of 66,7% (*Computational Linguistics*). Table 6 shows that generally the intensity of GL use increases for

both proceedings and *LR&E*, it alternates between increase and decrease of values in *ACM TOIS* and *CL* while is stable for *EURASIP*.

Title	2012			2013			2014		
	IF	Frequency of GL use (%)	Intensity of GL use (%)	IF	Frequency of GL use (%)	Intensity of GL use (%)	IF	Frequency of GL use (%)	Intensity of GL use (%)
ACM TOIS	1.070	100.0	20.2	1.300	100.0	12.3	1.021	85.7	16.0
Computational Linguistic	0.940	88.9	52.8	1.468	94.3	66.7	1.226	97.1	60.9
EURASIP Journal on Advances in SP	0.807	69.8	11.2	0.808	72.3	10.9	0.777	69.9	11.5
Language Resources and Evaluation	0.659	100.0	31.5	0.518	96.4	37.4	0.619	100.0	39.1

Table 6

The overview of the frequency and intensity in GL use related to the journals' IF does not allow to make any general consideration applicable to all journals.

From Table 6 we can note the following:

- 1) In *ACM TOIS* the IF value seems to affect more the intensity of use than the frequency (of use): if the IF increases the intensity of use decreases while the frequency remains stable; conversely, if the IF of *Computational Linguistics* increases even the frequency and intensity of GL use increase;
- 2) in *LR&E* the alternation of the IF value coincides with the stable growth of the intensity while the frequency alternates between increase and decrease;
- 3) in *EURASIP* both the IF value and the indicator values are substantially stable.

3.4 Frequency of use of GL-linked (with a link to a URL) references

The frequency of the use of GL-linked references varies from a minimum of 1,5% to a maximum of 50,8%.

Title	2012			2013			2014		
	Number of GL references	Number of linked-GL references	Frequency of linked-GL references (%)	Number of linked-GL references	Number of linked-GL references	Frequency of linked-GL references (%)	Number of linked-GL references	Number of linked-GL references	Frequency of linked-GL references (%)
ACM TOIS	285	32	11.2	135	16	11.9	150	34	22.7
Computational Linguistic	739	13	1.8	1.263	19	1.5	1.158	17	1.5
EURASIP Journal on Advances in SP	616	172	27.9	459	145	31.6	455	141	31.0
Language Resources and Evaluation	384	51	13.3	740	113	15.3	495	41	8.3
EACL Proceedings	884	27	3.1	ni	ni	ni	1.332	16	1.2
JCDL Proceedings	329	167	50.8	ni	ni	ni	384	133	34.6

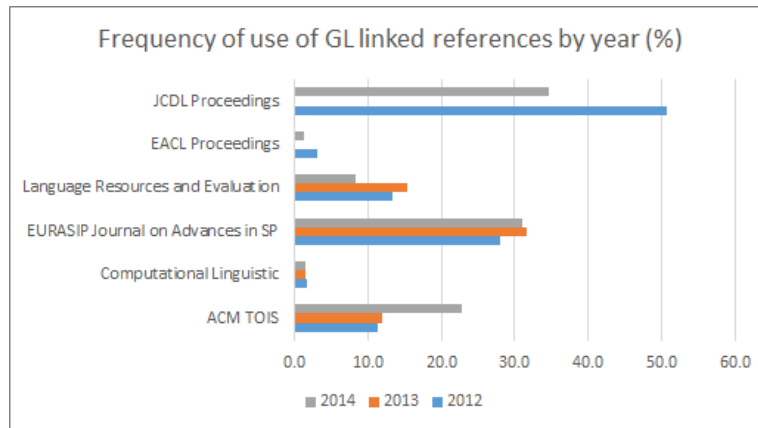
Table 7

Table 7 shows:

- 1) the highest percentage belongs to *JCDL* proceedings;
- 2) the lowest percentage belongs to *CL* (which has the highest IF among the journals chosen). The journal is *the premiere publication devoted exclusively to the design and analysis of natural language processing Systems*⁴.
- 3) The percentage is high (though with different values) for *EURASIP* (dedicated to the theoretical and practical aspects of signal processing) and for *LR&E*, belonging to the Computational Linguistics area but with strongly interdisciplinary features.

Graph 3 shows instead the variability of frequency of use of GL-linked references for each source over the years.

⁴ Cfr. <http://www.mitpressjournals.org/page/about/coli>.



Graph 3

As expected we found that the majority of GL-linked citations is concentrated in JCDL proceedings which is a publication dedicated to the study of multiple aspects of digital libraries as infrastructures metadata, contents, services, electronic publishing, multimedia etc. Conversely, we did not expect ACM TOIS, devoted to scholarly studies in all areas of information retrieval, would present a limited number of GL-linked citations. In the Computational Linguistics area, only LR&E seems to use GL linked citations.

4. GL Typology

Table 8 reports the distribution of GL documents by document type.

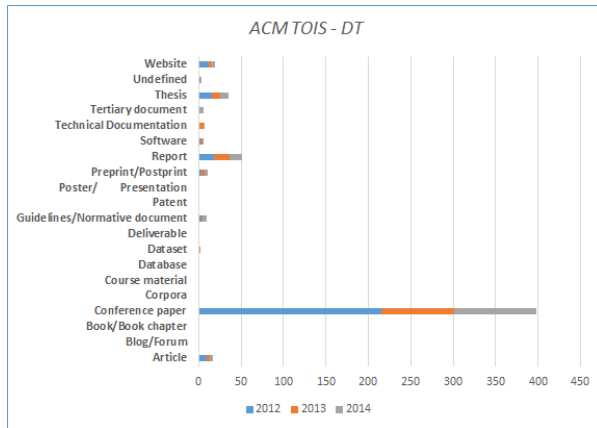
2012																				
Title	Article	Blog/Forum	Book/Book chapter	Conference paper	Corpora	Course material	Database	Dataset	Deliverable	Guidelines/Normative document	Patent	Poster/Presentation	Preprint/Postprint	Report	Software	Technical Documentation	Tertiary document	Thesis	Undefined	Website
ACM TOIS	10	1	0	215	0	0	0	0	1	4	1	0	4	18	2	1	0	16	1	11
EURASIP Journal on Advances in SP	6	1	5	213	0	2	16	2	2	32	19	3	33	85	16	45	10	80	3	43
Computational Linguistic	1	0	3	632	0	2	0	0	0	0	0	0	4	36	2	6	13	38	1	1
Language Resources and Evaluation	1	0	1	291	0	0	2	0	0	1	0	1	4	35	9	8	3	21	0	7
JCDL Proceedings	46	6	7	86	0	3	2	0	8	11	2	3	4	38	9	10	0	6	3	85
EAACL Proceedings	0	2	3	771	1	4	2	0	0	2	0	1	10	46	3	3	6	24	0	6
2013																				
Title	Article	Blog/Forum	Book/Book chapter	Conference paper	Corpora	Course material	Database	Dataset	Deliverable	Guidelines/Normative document	Patent	Poster/Presentation	Preprint/Postprint	Report	Software	Technical Documentation	Tertiary document	Thesis	Undefined	Website
ACM TOIS	3	0	0	86	0	0	0	2	0	1	0	0	3	18	3	5	1	10	0	3
EURASIP Journal on Advances in SP	11	1	2	123	0	2	7	3	1	45	8	0	23	58	19	49	2	50	2	53
Computational Linguistic	4	0	0	1107	0	0	0	0	1	0	0	9	2	40	2	14	10	71	0	3
Language Resources and Evaluation	11	2	0	567	22	0	0	1	4	17	0	1	7	32	6	7	5	50	3	5
2014																				
Title	Article	Blog/Forum	Book/Book chapter	Conference paper	Corpora	Course material	Database	Dataset	Deliverable	Guidelines/Normative document	Patent	Poster/Presentation	Preprint/Postprint	Report	Software	Technical Documentation	Tertiary document	Thesis	Undefined	Website
ACM TOIS	4	0	0	97	0	0	1	0	0	5	0	1	4	15	1	1	5	9	2	5
EURASIP Journal on Advances in SP	1	2	0	113	0	2	21	8	5	22	9	0	21	71	11	56	3	60	4	46
Computational Linguistic	5	0	4	1031	1	2	1	0	0	4	0	0	4	38	1	11	7	44	2	3
Language Resources and Evaluation	2	2	0	402	2	6	2	0	3	5	0	2	7	26	2	3	1	30	0	0
JCDL Proceedings	47	8	2	118	2	0	1	0	10	15	0	8	13	62	12	26	1	10	2	47
EAACL Proceedings	15	1	2	1206	0	0	0	0	0	2	1	0	12	41	4	7	6	32	0	3

Table 8

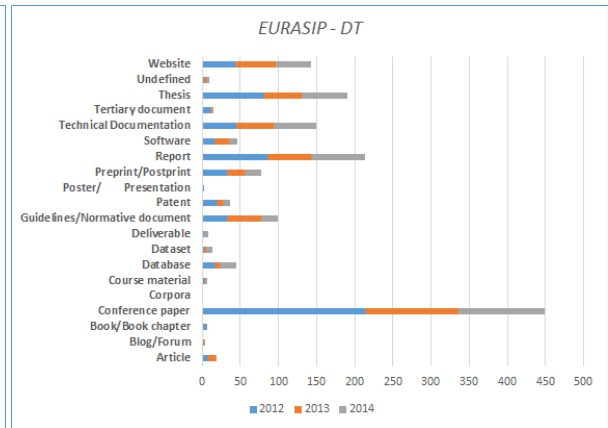
The most cited types of GL documents are: Conference papers (this group presents the highest number of GL citations), Reports, Thesis and Preprint/Postprint. These four types of document are the most cited regardless of the year, the nature of the products analyzed (journals and proceedings) and the area of knowledge to which they belong. The type Article, although less frequently, is present in each subject category and year (except for EAACL 2012).

As for other types, some peculiarities related primarily to the topic of the selected journals and proceedings emerged. For example, document types as Software/Tool, Technical Documentation, Database, Guideline/Normative document, Patent and Website are cited much more frequently in E&E area than in Computer Science and Computational Linguistics areas. In the same way the type Corpora is cited much more frequently in the Computational Linguistics and, in particular, in the LR&E Journal. None of the disciplines presents a considerable number of citations of Blog/Forum, Books, Dataset, Deliverable and Course material.

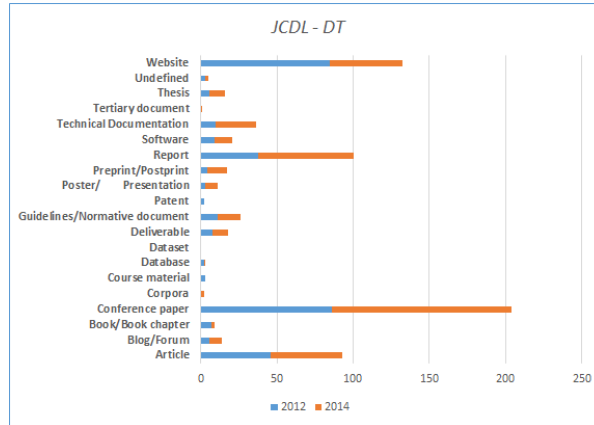
The Graphs below show the distribution of GL types for each source over the years: Graphs 4, 5, 6 show the document types in Computer Science and Engineering while Graphs 7, 8, 9 show the document types in Computational Linguistics. The graphical visualization allows us to better understand the intensity of use of the individual types of document by each resource analyzed.



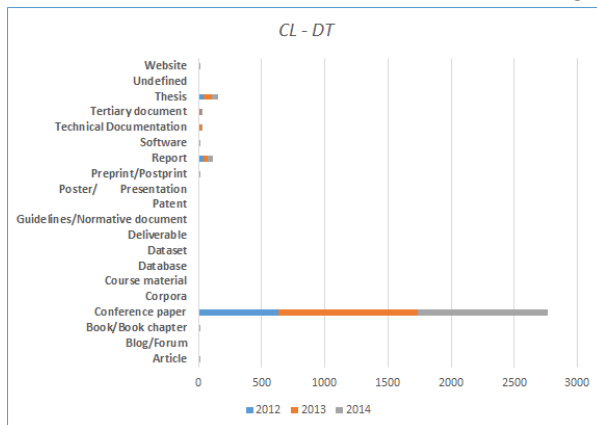
Graph 4



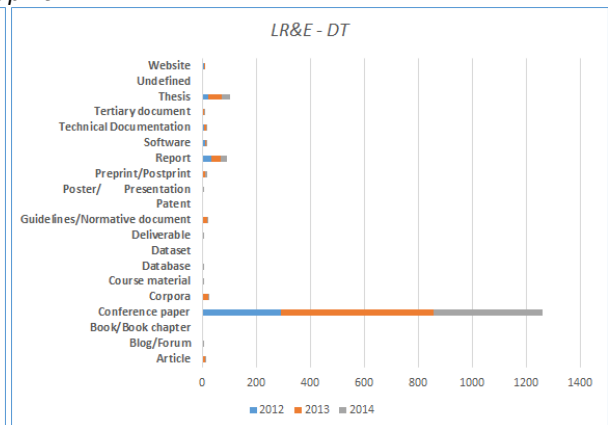
Graph 5



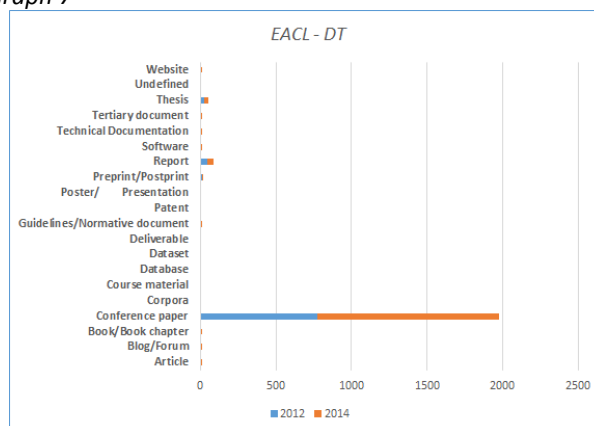
Graph 6



Graph 7



Graph 8



Graph 9

It seems clear that within ACM TOIS grey citations are limited to a small number of types while in EURASIP and JCDL the typology is more extensive. Both in EURASIP and in JCDL all document types are in use, although with little impact, except for Corpora in EURASIP and Dataset in JCDL, that are not present at all.

Let us note that both CL and EACL use only some types of document. CL, for example, never uses Patent, Guidelines/Normative document, Deliverable, Dataset, Database, Corpora and Blog/Forum while EACL never presents Poster/Presentation, Patent, Guidelines/Normative document, Deliverable, Dataset, Database, Course material and Corpora.

5. Conclusion

In 2004 we analyzed two sample data belonging to two very different scientific fields.

In this work the disciplinary boundaries of sample data are much less defined: but nevertheless, there are several significant differences, both in frequency and intensity of use of grey citations and about the cited type of documents, especially related to the specific field of study of each resource analyzed. The results obtained show that the Engineering domain has the least number of grey citations while the area of Computational Linguistics uses them most. In this respect, however, we must keep in mind that this result is strongly influenced by the presence of grey citations related to Conference papers published by the Association of Computational Linguistics. Even in the field of Computer Science, many “grey” Proceedings are collected and distributed freely by Associations, Institutions and no-profit services.

ACM TOIS is the only resource comparable with data analyzed in our work of 2004: the analysis indeed shows that GL frequency of citing and use remains in the range of 11.5 to 21.1 values identified for 1995 and 2003.

The quality of GL citations is still unclear and incomplete thus the analysis was difficult and time-consuming. We can conclude by saying that the traditional citation model - i.e. the habit to cite mainly conventional literature - is still very strong and leaves little room for alternative models. However, this survey returns percentages of frequency and intensity in use of GL substantially important, especially in the field of CL. It is increasingly clear the willingness of Associations and Organizations to collect, preserve and share the research results. The Repositories and the Open Access model have broken new ground and provided important tools for making these emerging communication needs come true.

Everything suggests, therefore, that the number of grey citations could increase in a very close future.

Bibliography

ACM Transactions on Information Systems. Retrieved January 12, 2016 from <http://tois.acm.org/>.

Budapest Open Access Initiative. 2002. Retrieved January 12, 2016 from <http://www.budapestopenaccessinitiative.org>.

Bethesda Statement on Open Access Publishing. 2003. Retrieved January 12, 2016 from <http://legacy.earlham.edu/~peters/fos/bethesda.htm>.

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. 2003. Retrieved January 12, 2016 from http://siba.unipv.it/biblioteche/banche_dati/berlin_declaration.pdf.

Cassella Maria, Bozzarelli Oriana. 2011. Nuovi scenari per la valutazione della ricerca tra indicatori bibliometrici citazionali e metriche alternative nel contesto digitale. *Biblioteche Oggi* 29, 2 (Feb. 211), 66-78.

Computational Linguistics. Retrieved January 12, 2016 from <http://www.mitpressjournals.org/loi/coli>.

De Bellis Nicola. 2009. *Bibliometrics and Citation Analysis: from the Science Citation Index to Cybermetrics*. Scarecrow Press, Lanham, MD, US.

De Bellis Nicola. 2005. La citazione bibliografica nell'epoca della sua riproducibilità tecnica - Bibliometria e analisi delle citazioni dallo Science Citation Index alla Cybermetrica. Retrieved January 12, 2016 from ...www.sba-old.unimore.it...⁵.

De Bellis Nicola. 2014. *Introduzione alla bibliometria: dalla teoria alla pratica*. AIB, Roma.

De Castro Paola, Salinetti Sandra (eds.) 2006. *La letteratura grigia nella comunicazione scientifica: il "Nancy style" per garantire la qualità editoriale dei rapporti tecnici*. Rapporti ITISAN 06/55. Istituto Superiore di Sanità, Roma.

Di Cesare Rosa, Ruggieri Roberta, Giannini Silvia, Biagioni Stefania. 2005. Trend evaluation and comparison of the use and value of GL in core demography and computer science journals. In *Proceedings of the 6th International Conference on Grey Literature (GL6)*, 41-49. Text Release, Amsterdam.

EACL - European Chapter of the Association for Computational Linguistics. Retrieved January 12, 2016 from <https://aclweb.org/anthology/docs/eacl.html>.

Eysenbach Gunther. 2006. Citation Advantage of Open Access Articles. Retrieved January 12, 2016 from <http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040157#pbio-0040157-t004>.

EURASIP Journal on Advances in Signal Processing. Retrieved January 12, 2016 from <http://www.asp.eurasipjournals.com/>.

Gazzetta ufficiale dell'Unione europea. 2012. Raccomandazione della Commissione del 17 luglio 2012 sull'accesso all'informazione scientifica e sulla sua conservazione (2012/417/UE). Retrieved January 12, 2016 from https://www.researchitaly.it/uploads/7309/rac_417.pdf?v=a901bf7.

Guerrini Mauro. 2010. *Gli archivi istituzionali. Open access, valutazione della ricerca e diritto d'autore*. Editrice Bibliografica, Milano.

JCDL - ACM/IEEE-CS Joint Conference on Digital Libraries. Retrieved January 12, 2016 from <http://www.icdl.org/>.

Language Resources and Evaluation. Retrieved January 12, 2016 from <http://link.springer.com/journal/10579>.

Pejšová Petra, Marcus Vaska. 2012. An Analysis of Current Grey Literature Document Typology. In *Proceedings of the 12th International Conference on Grey Literature (GL12)*, 39-47. Text Release, Amsterdam.

Schöpfel Joachim, Stock Christiane, Farace Dominic J., Frantzen Jerry. 2005. Citation analysis and grey literature: Stakeholders in the grey circuit. In *Proceedings of the 6th International Conference on Grey Literature (GL6)*, 55-63. Text Release, Amsterdam.

Schöpfel Joachim. 2010. Towards a Prague Definition of Grey Literature. In *Proceedings of the 12th International Conference on Grey Literature (GL12)*, 11-26. Text Release, Amsterdam.

Schöpfel Joachim, Hélène Prost. 2013. Degrees of Secrecy in an Open Environment. The Case of Electronic Theses and Dissertations. *ESSACHESS Journal for Communication Studies* 6, 2 (Dec. 2013), 66-85.

Serini Paola. 2003. Attualità della letteratura grigia. Il ruolo delle biblioteche nella sua valorizzazione. *Biblioteche Oggi* 21, 1 (Jan. 2003), 61-72.

⁵ Full URL:

https://www.google.it/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiVKn7nuabKAhWB0xQKHbJ2BBgQFggfMAA&url=http%3A%2F%2Fwww.sba-old.unimore.it%2FInside%2FBiblioteca%2520Digitale%2FDoc%2FLa%2520citazione%2520bibliografica%2520nell%27epoca%2520della%2520sua%2520riproducibilit%25C3%25A0%2520tecnica.pdf&usg=AFQjCNFaL-NEtFCMU_oTGAt9RXsnBNhByw&sig2=LNsOOD99NCx815rNggZqiw&bvm=bv.111677986,bs.2,d.ZWU&cad=rja



INIS

www.iaea.org/inis

The International Nuclear Information System

*The world's leading
source of nuclear
information since 1970*



IAEA

International Atomic Energy Agency