

Nineteenth International Conference on Grey Literature

Public Awareness and Access to Grey Literature

National Research Council of Italy, CNR Rome • October 23-24, 2017



Proceedings

ISSN 1386-2316

Conference Host and Sponsors:



SPRINGER NATURE



WILEY



GL19 Program and Conference Bureau

TextRelease

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



CIP

GL19 CONFERENCE PROCEEDINGS

Nineteenth International Conference on Grey Literature "Public Awareness and Access to Grey Literature" - CNR Rome, Italy, on October 23-24, 2017 / compiled by D. Farace and J. Frantzen ; GreyNet International, Grey Literature Network Service. - Amsterdam : TextRelease, 2018. - Author Index. - (GL Conference Series, ISSN 1386-2316; No. 19).

ISTI-CNR (IT), TIB Hannover (DE), DANS-KNAW (NL), FEDLINK; Library of Congress (USA), CVTISR (SK), EBSCO (USA), Inist-CNRS (FR), KISTI (KR), NIS-IAEA (AT), NTK (CZ), and the University of Florida; George A. Smathers Libraries (USA) are Corporate Authors and Associate Members of GreyNet International. These proceedings contain full text conference papers presented during the two days of plenary, panel, 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 and researchers along with their biographical notes. A list of more than 50 participating organizations as well as sponsored advertisements are likewise included.



Foreword

PUBLIC AWARENESS AND ACCESS TO GREY LITERATURE

Two of the most formidable problems that have faced information through the years are its overload on the one hand and its loss on the other. These are seen as interconnected with the supply and demand sides of grey literature.

A quarter century ago, the Grey Literature Network Service joined by research communities in library and information, physics, karst and marine sciences, biomedicine, nuclear energy, archeology, and many other scientific and technical fields set out to address this loss and overload of information.

In 1992, when the call for papers went out for the first conference in the GL-Series, the response was predominantly focused on the demand side of grey literature – that which was difficult to find and even more to access. The emphasis then lie in stemming the loss of grey literature. However, the outcome of that first conference also called attention to the equally important need for further research into the supply side of grey literature – namely its production, publication, and public awareness.

GL19 seeks to demonstrate how researchers and authors in the last 25 years have made significant inroads in responding to the loss and overload of grey literature. Likewise, this conference seeks to provide new directions in achieving public awareness and access to grey literature on an ever changing information landscape.

Dominic Farace
GREYNET INTERNATIONAL

Amsterdam,
FEBRUARY 2018



GL19 Conference Sponsors



 Consiglio Nazionale delle Ricerche

Biblioteca Centrale 'G. Marconi', Italy
National Research Council of Italy, CNR



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



The New York Academy of Medicine, USA



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



TIB, Germany
German National Library of Science and Technology –
Leibniz Information Centre for Science and
Technology University Library



GL19 Conference Sponsors



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



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



ACM Digital Library, USA
Association for Computing Machinery, Inc.



Springer-Verlag GmbH, Germany



EIPASS, Italy
European Informatics Passport



John Wiley & Sons, Inc., USA



School of Information Studies

University of Wisconsin, Milwaukee
School of Information Studies, USA



GL19 Program Committee



Alberto De Rosa ^{Chair}
National Research Council of Italy
CNR Central Library 'G. Marconi'
Italy



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



Margret Plank
German National Library
of Science and
Technology, Germany



Ján Turňa
Slovak Centre of Scientific
and Technical Information
CVTISR, Slovak Republic



Meg Tulloch
FEDLINK
Library of Congress
United States



Petra Pejšová
Wikimedia WMCZ
Czech Republic



Dobrica Savić
Nuclear Information
Section, International
Atomic Energy Agency,
Austria



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



Hana Vyčítalová
National Library of
Technology,
Czech Republic



YongJu Shin
Korea Institute of Science
and Technology
Information, KISTI
Korea



Deborah Rabina
Pratt Institute,
School of Information
United States



Tomas A. Lipinski
University of Wisconsin,
Milwaukee, UWM
United States



Joachim Schöpfel
University of Lille
France



Dominic Farace
GreyNet International
Grey Literature Network
Service, Netherlands



GL19 Program and Conference Bureau



Table of Contents

	Foreword.....	3
	Conference Sponsors.....	4
	Program Committee	6
	Program Chair and Conference Moderators	8
	Conference Program	9
<i>Program</i>	Opening Session	11
	Session One – Exposing Grey Literature to Wider Audiences.....	27
	Session Two – Overcoming Obstacles in Accessing Grey Literature.....	37
	Poster Session and Sponsor Showcase.....	93
	Panel Session – Innovations in Grey Literature Powered by Research Data.....	113
	Session Three – Impact of Emerging Technologies & Social Media on Grey Literature.....	127
<i>Advertisements</i>	ISTI-CNR, Institute of Information Science and Technologies.....	10
	DANS Your 7 steps to sustainable data.....	26
	CVTISR, Slovak Centre of Scientific and Technical Information.....	36
	LISTA-Full Text EBSCO.....	54
	PsycEXTRA EBSCO	60
	INIS, The International Nuclear Information System.....	92
	NTK, National Library of Technology, Czech Republic.....	100
	FEDLINK, The Federal Library and Information Network - Library of Congress	132
	KISTI, Korea Institute of Science and Technology Information.....	142
<i>Appendices</i>	List of Participating Organizations	139
	GL20 Conference Announcement.....	140
	GL20 Call for Papers.....	141
	Author information.....	143
	GreyNet’s Service Providers.....	148
	Index to Authors.....	149
	GL19 Publication Order Form	150



Moderator Day One

Margret Plank

Head of Development

German National Library of S&T

Margret Plank is currently the Head of Development and the Competence Centre for Non-Textual Materials at the German National Library of Science and Technology in Hannover, Germany. The aim of the Competence Centre is to develop emerging tools and services that actively support users in the scientific work process enabling non-textual material such as audiovisual media, 3D objects and research data to be published, found and made available on a permanent basis as easily as textual documents. Previously she was responsible for Information Competence and Usability at the TIB. She has also worked as a researcher at the Institute of Information Studies and Language Technology of the University of Hildesheim. She represents TIB on a number of boards including IFLA Steering Committee Audiovisual and Multimedia Section as well as the International Council for Scientific and Technical Information, ICSTI/ ITOC.

Email: margret.plank@tib.eu



Program Chair

Alberto De Rosa

Head Central Library

National Research Council of Italy

Alberto De Rosa is responsible for the National Research Council Central Library 'G. Marconi', the main Italian multidisciplinary library devoted to Science and Technics. He is involved in technological, administrative, and management activities related to scientific projects and Information Science in both National and European programs. He graduated in Economics at the Naples University 'Federico II'. He is a chartered accountant and statutory auditor (Register of the Italian Ministry of Economics and Finance). He obtained various Masters in Business and Management in the Public Administration and Research boards. From 1993 to 2013 he has been Administration Manager of CNR Research Institutes (e.g. the Institute of Biostructures and Bio-imaging); From 2002 to 2013 he has been Adjunct Professor of 'Company Structure' at the Naples University. And, from 2006 to 2014 he was Responsible for document management system, document flow and current archives at the CNR Institute of Biostructures and Bio-imaging.

Email: alberto.derosa@cnr.it



Moderator Day Two

Judith C. Russell

Dean of University Libraries

University of Florida

Judith C. Russell is the Dean of University Libraries at the University of Florida. She was formerly the Managing Director, Information Dissemination and Superintendent of Documents, at the U.S. Government Printing Office (GPO). Russell previously served as Deputy Director of the National Commission on Libraries and Information Science (NCLIS) and as director of the Office of Electronic Information Dissemination Services and Federal Depository Library Program at GPO. She worked for more than 10 years in the information industry in marketing and product development, as well as serving as a government-industry liaison. Her corporate experience includes Information Handling Services (IHS) and its parent company, the Information Technology Group; Disclosure Information Group; Lexis-Nexis (former Mead Data Central), and IDD Digital Alliances, a subsidiary of Investment Dealers Digest. She has an M.L.S. from Catholic University and a B.A. from Dunbarton College of the Holy Cross.

Email: jcrussell@ufl.edu

**Opening Session**

- Grey Literature and Research Assessment exercises: From the current criteria to the Open Science models** 11
Silvia Giannini, Rosaria Deluca, Anna Molino, and Stefania Biagioni, ISTI-CNR, Pisa, Italy

Session One – Exposing Grey Literature to Wider Audiences

- Data Papers are Witness to Trusted Resources in Grey Literature: Driving Access to Data thru Public Awareness** 27
Dominic Farace and Jerry Frantzen, GreyNet, Netherlands; Plato L. Smith, University of Florida, United States
- Public Access to the Dissertations in Russia** 33
Yuliya B. Balashova, Saint Petersburg State University, Russia

Session Two – Overcoming Obstacles in Accessing Grey Literature

- How open access policies affect access to GL in university digital repositories: A case study of iSchools** 37
Tomas A. Lipinski and Katie Chamberlain Kritikos, University of Wisconsin at Milwaukee, United States
- Law, Liability, and Grey Literature: Resolving Issues of Law and Compliance** 55
Daniel C. Mack, University of Maryland, United States
- Indexing grey multilingual literature in General Practice in the era of Semantic Web** 61
Marc Jamoulle, Department of General Practice, University of Liège, Belgium
Elena Cardillo, Institute of Informatics and Telematics, National Research Council, Italy
Ashwin Ittoo, HEC Management School, University of Liège, Belgium
Robert Vander Stichele, Heymans Institute of Pharmacology, University of Ghent, Belgium
Melissa P. Resnick, University of Texas, Health Science Center at Houston, United States
Julien Grosjean and Stk;efan Darmoni, D2IM, University of Rouen, France
Marc Vanmeerbeek, Department of General Practice, University of Liège, Belgium
- Preserving and accessing content stored on USB-flash-drives: A TIB workflow** 85
Oleg Nekhayenko, German National Library of Science and Technology, TIB, Germany

Poster Session and Sponsor Showcase

- Providing Access to Grey Literature: The CLARIN Infrastructure** 93
Sara Goggi, Gabriella Pardelli, Irene Russo, Roberto Bartolini, and Monica Monachini, CNR, Istituto di Linguistica Computazionale, "Antonio Zampolli", Italy
- Collecting Grey Literature – Institutional Repository versus National Aggregator** 101
Petra Černohlávková and Hana Vyčítalová, NTK, National Library of Technology, Czech Republic
- OpenAIRE : Advancing Open Science** 107
Paolo Manghi, Michele Artini, Claudio Atzori, Miriam Baglioni, Alessia Bardi, Sandro La Bruzzo, and Michele De Bonis; Institute of Information Science and technologies, Italian National Research Council, Italy;
Harry Dimitropoulos, Ioannis Foufoulas, Katerina Iatropoulou, Natalia Manola, and Stefania Martziou; Athena Research Center in Information, Communication and Knowledge Technologies, Greece;
Pedro Principe, University of Minho, Portugal

Panel Session – Innovations in Grey Literature Powered by Research Data

- A Facet-based Open and Extensible Resource Model for Research Data Infrastructures** 113
Luca Frosini and Pasquale Pagano, Istituto di Scienza e Tecnologie dell'Informazione (ISTI) "Alessandro Faedo" Italian National Research Council (CNR),
- D4Humanities: Deposit of Dissertation Data in Social Sciences & Humanities – A Project in Digital Humanities** 121
Joachim Schöpfel, GERiiCO Laboratory, University of Lille; Hélène Prost, CNRS, associate GERiiCO Laboratory, France

Session Three – Impact of Emerging Technologies & Social Media on Grey Literature

- Video is the new Grey** 127
Bastian Drees and Margret Plank, National Library of Science and Technology, Germany
- Apps & Codes: Making profiles for fluid publishing contents** 133
Flavia Cancedda and Luisa De Biagi, National Research Council of Italy, CNR Central Library, Italy



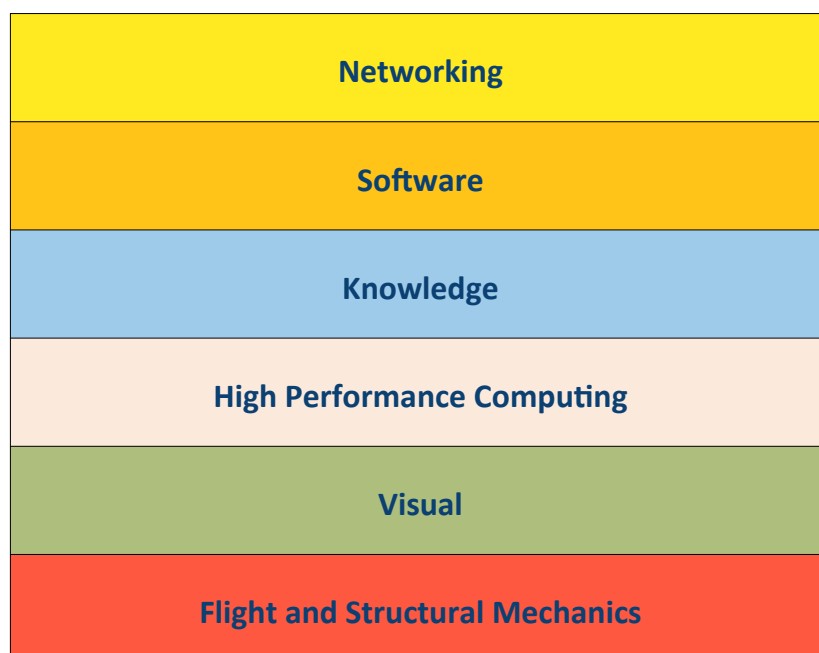
Institute of Information Science and Technologies “A. Faedo”

an Institute of the National Research Council of Italy CNR

***ISTI is committed to produce scientific excellence and to play an
active role in technology transfer.***

***The domain of competence covers Computer Science &
Technologies and a wide range of applications.***

***The research and development activity of the Institute can be
classified into 6 thematic areas***



**CNR-ISTI, Via G. Moruzzi 1
56124 Pisa (PI), Italy
Area della Ricerca del CNR**

**Contact: +39 050 315 2403
segreteria scientifica@isti.cnr.it
<http://www.isti.cnr.it>**



Grey Literature and Research Assessment exercises: From the current criteria to the Open Science models

Silvia Giannini, Rosaria Deluca, Anna Molino, Stefania Biagioni
CNR, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Italy

1. Introduction

In the recent years the application of strategies, procedures and tools to evaluate the work of researchers have become subject of interest and their application is currently matter of discussion.

This topic is of major importance and will probably have a greater influence, since this type of exercise has strong political implications and determines a significant economic impact on the future of Universities and Research Institutions. Research assessment is a complicated business, as the design of a practical, informative process requires making decisions about which methodology should be used, which indicators calculated, and which data collected (Moed 2011). The evaluation procedure is also time-consuming and expensive, since the comprehension of the mechanisms underlying the research activity might be quite complex, and the results are not wholly predictable in some cases (Bianco 2010). Moreover, finding the right balance between two different kinds of approach makes the assessment exercise more problematic. On the one hand, the *quantitative* approach deals with the impact of the research, measuring the degree of diffusion of a certain idea in the scientific community. On the other hand, the *qualitative* approach determines the value of the research in terms of authenticity, relevance and clarity in the exposition of the results (Baccini 2011). Finally, the social and economic impacts have to be taken into account. Indeed, the evaluation procedures has recently acquired a greater importance due to the shortage of economic resources, therefore becoming a strategic instrument for the quality assessment of Universities and Research bodies.

The assessment exercises are regulated at national level and are carried out in different European countries such as France, United Kingdom and The Netherlands. The English RAE - Research Assessment Exercise is the oldest performance-based research funding system (Rebora, Turri 2013). The Research Assessment Exercises (RAE) have been held in the UK since 1986.

The RAE has been a benchmark for the relatively recent Italian assessment exercise, as the first steps in this direction were taken in Italy at the beginning of the '90s. In 1993 the *Osservatorio per la valutazione del sistema universitario* was created, becoming active only in 1996. In 1998 the *Comitato di Indirizzo per la Valutazione della Ricerca* (CIVR) was set up, and in 1999 the *Comitato Nazionale per la Valutazione del Sistema Universitario* (CNVSU) was created as successor of the *Osservatorio*, becoming officially operative in 2000 (Rubele 2012). The first research assessment exercise has been legislated in 2003 and entrusted to CIVR. The Committee analyzed the research products of 2001-2003 in order to evaluate the scientific performance of Universities, as well as state and private Research Institutions. Three years later the CIVR and other committees have been replaced by a specific agency named National Agency for the Evaluation of Universities and Research Institutes [*Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca* (ANVUR)].

The Agency aims to « [...] rationalize the system of assessment of the quality of Universities, state and private Research Institutions beneficiary of public funds [...] » « The results of these activities managed by ANVUR represent one of the criteria to assign the state funds to Universities and Research Institutions ».

More in detail, the Agency evaluates the quality of the processes, the results, and the products released by Universities and Research Bodies and defines criteria and methods for the evaluation of the university branches and of the course of studies.

It cooperates with other scientific international committees operating in the field of research assessment and with the European Union.

At the present time, ANVUR has completed two evaluation exercises of the quality of the research named Evaluation of Research Quality [*Valutazione della Qualità della Ricerca* (VQR)]: the first one spans the years 2004 – 2010 (VQR1); the second from 2011 to 2014 (VQR2).

Despite the shared belief among the different scientific communities about the importance of the research assessment, there is no such agreement regarding the purposes and the procedures for its realization (Galimberti 2012). The international debate on methods and critical issues of the



research assessment practices has become more intense, registering an increased number of negative judgements about the procedures currently applied. Although the ultimate goal is obtaining excellence, quality and the greater impact on society, the parameters currently in use to evaluate the products of the research and the consequences of such measuring practices have a negative influence on the attitude of the researchers towards this topic (Galimberti 2017). Criticisms are mainly concentrated on quantitative measurements, because of their improper use of a range of commercial bibliometric indicators. The *ROARS – Return On Academic Research* association, whose aim is particularly focused on the policies for the evaluation of the research, dedicates large part of its blog to the major issues concerning the ANVUR evaluation procedures, promoting various initiatives that encourage fairer practices and more responsible behaviors in the research assessment. Proposals like *DORA – Declaration on Research Assessment*, as well as the *Leiden Manifesto for the Research Metrics* aim at defining criteria that would represent more widely the complexity of the research analysis.

The conceptual challenges taken on by the Open Science (OS) movement may be crucial for the evolution of these matters. The OS is multidimensional; approaches and skills of various kinds are necessary for its fulfilment and for the achievement of objectives such as openness, sharing, transparency and quality¹. As a matter of fact, the term Open Science indicates the opportunity to freely contribute to the knowledge production, sharing the outcomes and being inclined to the cooperation with the whole scientific community (Delfanti 2008). The expression encompasses subject matters like the *Open Access (OA)*, the free access to scientific publications; the *Open Data*, as it promotes the dissemination of the research data; the *Open-Notebook Science*, encouraging the online sharing of lab notes and raw data (Stafford 2010). Moreover, the OS introduces the ideas of *Open Learning*, a new, customizable teaching plan, independent of time and space; and the *Open Research and Citizen Science*, making science available to all the citizens of the *knowledge-society*, where they have the right to access the most advanced researches (Gioè 2016).

The work looks at the environment of VQR in order to understand the organizational set-up, the operational models, the scientific Areas involved in the process and the selection and evaluation criteria of the research products. More in detail, our work analyzes and compares the evaluation exercises conducted in Italy with the aim of verifying if and how *Grey Literature (GL)* is involved in the research evaluation processes. The article checked the types of products admissible for the research assessment and those actually presented by the researchers of Universities and Research Institutions. We measured the products from a quantitative point of view and observed their ramification in the different disciplinary fields rather than their transformation during the period of time taken into consideration. At the same time, we focused on the OS movement in order to understand what could be its role within the research assessment exercises and how it could affect the future of scholarly scientific communication.

2. The VQRs framework

2.1 Organization and methods

We consulted the public documentation provided by ANVUR, consisting in a set of preliminary documents and a set of documents produced as final reports for the two evaluations².

The two exercises done in Italy were addressed to the assessment of the research conducted in both state and private universities, as well as in public research bodies and other public and private subjects whose research activities are funded by the government. Researchers, assistant professors, associate professors, full professors, all being on duty at the time the evaluation

¹ <https://sites.google.com/site/scienzaapertaricercamigliore/programma>

² Announcement_VQR 2004-2010, 17 July 2011,

http://www.anvur.org/index.php?option=com_content&view=article&id=122&Itemid=305&lang=it.

Specific criteria by the expert groups selected by the Agency,

http://www.anvur.org/index.php?option=com_content&view=article&id=32&Itemid=372&lang=it.

Area Reports VQR_2004-2010, 30 June 2010, <http://www.anvur.org/rapporto/>.

Announcement_VQR 2011-2014, 11 November 2015,

http://www.anvur.org/index.php?option=com_content&view=article&id=825&Itemid=599&lang=it

Specific criteria by the expert groups selected by the Agency,

http://www.anvur.org/index.php?option=com_content&view=article&id=841&Itemid=601&lang=it

Area Reports VQR_2011-2014, 21 February 2017, <http://www.anvur.org/rapporto-2016/>.



started, have been appraised. The number of research products to be assessed was indicated with reference to each individual evaluated.

In both exercises, a taxonomy based on macro disciplinary areas (Tables 1-2) was used, each subdivided into Scientific Disciplinary Sectors [*Settori Scientifico-Disciplinari* (SSD)]³.

Macro-Areas VQR1

Area	Description
1	Computer science and Mathematics
2	Physics
3	Chemistry
4	Earth sciences
5	Biology
6	Medicine
7	Agricultural and veterinary sciences
8	Civil engineering and Architecture
9	Industrial and computer engineering
10	Antiquity, philological-literary and historical-artistic sciences
11	Historical, philosophical, psychological and pedagogical sciences
12	Legal sciences
13	Economics and Statistics sciences
14	Social and political sciences

Macro-Areas VQR2

Area	Description
1	Computer science and Mathematics
2	Physics
3	Chemistry
4	Earth sciences
5	Biology
6	Medicine
7	Agricultural and veterinary sciences
8a	Architecture
8b	Civil engineering
9	Industrial and computer engineering
10	Antiquity, philological-literary and historical-artistic sciences
11a	Historical, philosophical and pedagogical sciences
11b	Psychology
12	Legal science
13	Economics and Statistics sciences
14	Social and political sciences

Tables 1-2

From the comparison between the tables is clear the substantial overlapping between macro-areas in the two exercises, with the exception of Areas 8 and 11, split in two sub-categories in VQR2, the number of Areas going from 14 to 16.

The ANVUR constituted Groups of experts for the evaluation⁴ [*Gruppi di Esperti della Valutazione* (GEV)] for each macro-area, composed of both Italian and foreign qualified experts.

Sub-groups of specialists were created within those GEVs dedicated to highly heterogeneous disciplinary areas and characterized by a large number of products to be evaluated. Each GEV has a nominated Coordinator.

The judgment on the quality of the products was based on the following general criteria:

1. VQR 2004 – 2010: Relevance – Originality/Innovation – Internationalization
2. VQR 2011 – 2014: Originality – Methodological rigor – Impact (recognized or potential)

Among the general principles established by the Agency, each GEV set its own criteria, which allowed the Groups to define different quality steps, as illustrated in Table 3. It is evident that the quality steps and their related scores were partially revisited in the second exercise, where the class *Plagiarism/Fraud* was deleted.

VQR 2004-2010		VQR 2011-2014	
Class of merit	Score	Class of merit	Score
A. Excellent	1	A. Excellent	1
B. Good	0.8	B. High-level	0.7
C. Acceptable	0.5	C. Fair	0.4
D. Limited	0	D. Acceptable	0.1
E. Not evaluable	-1	E. Limited	0
F. Plagiarism/Fraud	-2	F. Not evaluable	0

Table 3 - Quality steps

The two assessments were based on an *informed peer-review*. In the technical areas, this technique was based on a combination of bibliometric parameters and peer-review. The papers

³Cfr. "Evaluation of Research Quality 2011 – 2014 (VQR 2011 – 2014). ANVUR final report".

For example: Area 1 - *Computer science, Logic, Algebra, Geometry, Complementary maths, Mathematical analysis...* Area 2 - *Experimental physics, Theoric physics, (mathematical models and methods), Physics of matter, Nuclear and subnuclear physics, Astronomy and Astrophysics, Physics for the Earth system...*

⁴cfr. "Evaluation of Research Quality 2011 – 2014 (VQR 2011 – 2014). ANVUR final report".



in journals were evaluated through indicators derived from commercial citation databases⁵. An algorithm taking into account the quality of the journal as well as the number of the article's citations was exploited during the assessment exercise. Other categories or more recent products with a lower number of citations were evaluated by *peer-review*. In the Humanities and Law disciplines the bibliometric analysis is less frequent, as databases such as WoS or Scopus do not have a widespread coverage. In these cases, the evaluation carried on by the GEVs was only peer-reviewed, or alternatively based on a technique named *informed peer review*, which relies on different evaluation methods that make use of various information tools, like editorial peculiarities, reviews, translations, awards.

The GEVs can entrust the *peer-review* to external reviewers or conduct it inside the group itself, and it must define specific assessment criteria to harmonize the different evaluation methods. Moreover, the GEV is responsible of the final result of the assessment.

2.2 The VQRs "objects"

The specifications given by ANVUR for the assessment of the research products by the GEVs were enhanced in the VQR2.

In VQR1 the categories listed below were defined for the classification of all identified research products, not providing further details on the characteristics and subject matters⁶:

Documentary categories in VQR1

- a) Papers in journals
- b) Books, chapters of books, and conference proceedings provided with ISBN
- c) Critical editions, translations, and scientific comments
- d) Patents
- e) Compositions, drawings, design, performances, exhibitions and organized expositions, handwork, prototypes, artworks and related designs, databases and software, thematic maps

Documentary categories in VQR2

1) *Scientific monographs*

Research monograph, Concordance, Scientific comment, Annotated bibliography, Critical editions of texts, Critical editions of excavations, Publication of unedited sources, Critical manuals (not for educational purpose only), Grammars and science dictionaries, Translations of books (upon GEV's decision).

2) *Articles in journals*

Scientific paper, Review essays, Letters, Contribution to a Forum upon invitation of the editorial staff, Case notes, Translations in journal.

3) *Contributions to books*

Scientific articles in peer-reviewed conference proceedings, Foreword and afterword in the form of essay, Curatorship of books with introductory essay, Catalogues with introductory essay, Critical entries in dictionaries or encyclopedias, Translations in book (upon GEV's decision), Catalographic records, bibliography or corpora.

4) *Other types of scientific products*

Compositions, Drawings, Architectonic projects, Performances, Exhibitions, Prototypes of art and related projects/designs, Database and software, Thematic maps, Psychological evaluations, Audiovisual material.

5) *Patents*

The category Patents is always considered as evaluable, but it may be attributed to class A or B only if internationally renowned or licensed.

Not admissible products in VQR1

- *Editorial and curatorial activities*
- *Conference abstracts (even if published in journals)*
- *Texts or software used for educational and dissemination purpose only*
- *Routine or laboratory tests*
- *Internal technical reports*

⁵ Web of Science (WoS) by Clarivate Analytics; Scopus by Elsevier; Mathematical Reviews on the web (MathSciNet) by The American Mathematical Society.

⁶ Cit. Announcement_VQR 2004-2010, 17 July 2011.

**Not admissible products in VQR2**

- *Manuals and texts for educational purpose only*
- *Review of a single article not showing any critical analysis of the literature on the topic*
- *Short, non-original encyclopedia or dictionary entries*
- *Short, non-original case notes*
- *Short catalographic records*

The merger of the research products in documentary classes highlights some differences in their own composition. The second assessment exercise was more inclined towards categories such as articles in journals and contributions in books, as alternative to the mere scientific articles.

Indeed, in the first class of VQR2 we find sub-categories like *review essays*, *letters*, *case notes*, *contribution to a forum*, and *translations*, while the second includes *essay collections*, *concordances*, *bibliographies*, *critical manuals*, *grammars*, *corpora*, and *catalographic records*, as well as the entries *critical editions*, *translations*, and *scientific comments*, considered as autonomous in the previous evaluation.

The category *e)* of VQR1 was merged into *Other types of scientific products*, which includes the additional entries *architectonic projects*, *psychological evaluations*, *audiovisual material*, while *design* and *handwork* were excluded.

3. Tracking the Grey Literature

The list of documentary typologies evaluated does not allow a thorough classification of the *grey* products. The Implementation Announcements of the two VQRs do not specify any requirement concerning the manners of publication; they only define the categories containing works circulated through the conventional distribution channels as well as products disseminated out of the traditional publishing chains. The fact that the documents have ISBN and/or ISSN code, or that they have been successfully peer-reviewed, does not grant their publication by a commercial publishing company. The procedures, the evaluation principles and the assessment tools only are able to determine if the majority of the products belong to the traditional literature.

The analysis of the GEVs' assessment criteria and of the FAQs published by ANVUR contributed to the understanding of some concepts and supported the interpretation of some results, facilitating only partially the process of identification of the *Grey Literature* inside the various documentary typologies. For these reasons, the process of identification of the Grey Literature was based on the following considerations:

- I. The evaluation exercises mainly founded their bibliometric analysis on the contents of the two commercial databases WoS and Scopus⁷. This is due to the fact that the international scientific community makes extensive use of them for the assessment of the scientific levels of the journals. Although the databases found their bibliometric indicators on different parameters, they are both based on the calculation of the number of citations. The majority of the literature indexed by the two databases is published by commercial publishing companies; only a small percentage of products ascribable to the *Grey Literature* are indexed in Scopus.
- II. The use of the databases restricts the contents only to the references indexed (based on ownership criteria). Especially for the *papers in journals*, the algorithm for the assessment takes into account the number of citations of a paper and the corresponding bibliometric indicator of the journal inside one or more disciplinary classes defined in the two databases⁸. The more commonly evaluated research products like *books*, *papers in books* and *papers in proceedings* are assessed taking into account their occurrence in the databases. Moreover, the GEVs reserve the *peer-review* only to the products not indexed in the databases.
- III. In both evaluation exercises, it is made reference frequently to the products published by commercial publishing companies, especially if they are renown at international levels. In some of the GEVs' criteria it is specified that self-published products are not evaluated. Moreover, it is clearly stated that products accepted but not yet published are not taken into consideration.
- IV. The GEVs' criteria specify that the products listed in *e)* in the VQR1 Announcement would be evaluated making reference to their characteristics, not to their formal publication. In VQR2 Announcement these products are grouped in *Other types of scientific products*. Therefore,

⁷ For the evaluation of Areas 12 and 13 in VQR1 a reference is made for the use of Google Scholar as assessment parameter for the evaluation of the journal, if not indexed in WoS or Scopus. This is not taken into account in VQR2.

⁸ WoS Subject Category (SC) and Scopus All Science Journal Classification (ASJC).



the eligibility conditions of these products are clearly expressed in the criteria used for specific disciplinary areas. For instance, in some cases it is specified that products like *drawings*, *prototypes of art*, or *architectonic projects* may be evaluated if they have been published or worthy of mention/winner of prizes in a competition. At the same time, the *thematic maps* may be assessed if their theme is evidenced using particular procedures and graphic adaptations, allowing the immediate understanding of the distribution, differentiations, and correlations of one or more phenomena. In the Area 02 (Physical sciences) the item *composition* was evaluated by the GEV with its products *handwork*, *devices* and *prototypes*, along with the entries *exhibition*, *database* and *software*, whereas the types *drawings*, *architectonic projects*, *performance*, *prototypes of art* and related projects, as well as the *thematic maps*, were not assessed. More details for the software products have been given in Area 9 (Industrial and computer engineering).

With specific reference to what listed above, we did not identify *grey products* within categories such as *papers in journals*, *books* or *proceedings* as well as products belonging to other assimilated categories such as *curatorship*, *critical editions*, and *translations*. On the other hand, we agreed on ascribing some groups of products to the *non-conventional literature*, including in this range also entries with a certain degree of uncertainty with respect to their arrangements for publication. This is the case of products measured in VQR2, such as: *concordance*, *publication of unedited sources*, *entries*, *catalographic record*. As a matter of fact, the indications given by the GEVs neither completely clarify the nature and the characteristics of these products, nor explicate their inclusion in the list of works conventionally published.

4. Analysis of data and results

Tables 4 and 5 show the whole range of products evaluated in the two exercises, sub-divided in the 14 disciplinary areas of reference⁹. Some of the original tables of VQR1 do not list the number of the products but the percentage calculated by the GEVs only, here reported. This is the case of Areas 2 (Physics), 4 (Earth sciences), 11 (Historical, philosophical, psychological and pedagogical sciences), and 12 (Legal sciences). Moreover, in the same exercise the data relative to the Area 11 are sub-divided between products evaluated through bibliometric analysis and those *peer-reviewed*. In VQR2 this sub-division is not present.

In both exercises and for each disciplinary area, the most significant numbers are referred to the entries *papers in journals*, *papers in books* e *papers in proceedings*.

VQR 2004-2010															
Categories	Area 01	Area 02	Area 03	Area 04	Area 05	Area 06	Area 07	Area 08	Area 09	Area 10	Area 11-nbib	Area 11-bib	Area 12	Area 13	Area 14
	%	%	%	%	%	%	%		%	%	%	%	%	%	%
Abstract (in journals or in proceedings)	0.06		0.09		0.21	0.07									
Case notes													0.56		
Composition								0.04							
Critical edition	0.03				0.01			0.07		0.92	0.56				
Curatorship	0.07		0.04		0.02		0.26	2.69		1.76	1.84	0.52	0.73		2.80
Database	0.01			0.20	0.01			0.01							
Design	0.02							0.13							
Entry (in dictionary o encyclopedia)					0.01			0.03					0.29		
Exhibition	0.01							0.02							
Foreword/Afterword								0.02							
Handwork	0.02	0.70	0.01					0.08							
Maps				0.20											
Monograph or scientific treaty	1.27	0.20	0.21	1.03	0.38	0.52	1.22	14.56	0.75	22.69	33.06	9.26	25.88	12.75	33.63
Other	0.13	0.30	0.01	2.28	0.04	0.07	0.82	1.09	0.44	0.80	0.61	0.22	0.64	1.01	0.46
Paper in books	3.27	0.60	0.39	5.06	1.26	1.53	4.65	23.60	2.46	32.80	32.86	11.57	36.00	19.88	32.59
Paper in journals	86.11	93.40	98.45	85.96	96.92	96.94	87.54	43.69	81.68	26.50	23.99	77.22	32.76	62.45	28.63
Paper in proceedings	8.84	4.70	0.40	4.80	0.89	0.74	5.51	13.82	14.16	14.19	7.09	1.21	3.13	3.91	1.90
Patent	0.10	0.20	0.39	0.18	0.26	0.13		0.14	0.51						
Prototype of art and related project								0.01							
Software	0.07		0.02	0.28	0.01										
Translation										0.33					
Total	10685	19773	11608	8433	16407	26713	10004	9533	16347	14073	9513	3639	11882	11941	4327

Table 4 - Research products in VQR1 1

In VQR1 the percentage indicates that the *papers in journals* prevail in almost all disciplinary areas, in some cases reaching nearly 100%.

⁹ The tables are a rework of those contained in the area reports produced by the GEV. These can be viewed at URLs <http://www.anvur.org/rapporto/> (VQR1) and <http://www.anvur.org/rapporto-2016/> (VQR2).



In Areas 10 (Antiquity, philological-literary and historical-artistic sciences), 11nb (Historical, philosophical, psychological and pedagogical sciences), 12 (Legal sciences), and 14 (Social and political sciences) only the *monographs* and the *papers in books* show significant percentages, proving the most widely used modalities in the scientific communication in these sectors. The *papers in proceedings* represent the largest number in Areas 8 (Civil engineering and Architecture), 9 (Industrial and computer engineering), and 10 (Antiquity, philological-literary and historical-artistic sciences).

VQR 2011-2014																
Categories	Area 01	Area 02	Area 03	Area 04	Area 05	Area 06	Area 07	Area 08a	Area 08b	Area 09	Area 10	Area 11a	Area 11b	Area 12	Area 13	Area 14
	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Abstract (in journals or in proceedings)														0.02		
Architectonic project								1.45								
Bibliographic/Catalographic record, corpus		0.09			0.06	0.03	0.03				0.09					
Bibliography							0.01				0.03	0.05				
Case notes					0.04	0.01		0.03					0.04	1.52		
Composition		0.75	0.01					0.03		0.05					0.01	
Concordance											0.05					
Critical edition		0.02						0.09			2.68	0.80		0.09	0.05	0.27
Curatorship	0.15		0.03	0.11	0.02	0.02	0.03	5.01		0.02	0.11	1.45	0.35	0.47	0.38	1.88
Database		0.04		0.18	0.03		0.08	0.03		0.04	0.05	0.02		0.02		
Design								0.12								
Entry (in dictionary or encyclopedia)					0.01					0.03	0.23	0.23	0.13	1.37	0.10	0.03
Exhibition		0.11		0.07	0.06	0.01	0.03	0.14	0.04	0.03	0.05					
Handwork																
Maps				0.23			0.04									
Monograph or scientific treaty	1.35	0.22	0.26	0.88	0.20	0.36	0.89	23.29	0.60	0.63	19.64	27.24	3.82	26.17	8.67	24.13
Other	0.21			0.18	0.01	0.01		0.52	0.04		0.17	0.07	0.09		0.55	
Paper in books	2.95	0.73	0.58	3.18	0.79	1.01	3.69	28.91	2.82	1.66	35.91	30.48	6.50	30.57	14.68	32.01
Paper in journals	87.92	96.85	98.12	91.78	97.23	97.92	91.31	26.42	88.63	88.71	32.12	34.79	88.09	38.04	72.71	40.83
Paper in proceedings	7.32	0.70	0.41	3.23	0.62	0.52	3.39	12.96	7.66	8.28	6.78	3.94	0.75	1.61	2.68	0.27
Patent	0.02	0.18	0.33	0.07	0.13	0.04	0.19	0.29	0.18	0.38					0.02	
Performance					0.03			0.03			0.07					0.07
Preface/Postface			0.01					0.20		0.01	0.37	0.38	0.04	0.05	0.04	0.30
Prototype of art and related project								0.12		0.03						
Publication of unedited sources								0.09			0.15	0.24		0.01	0.01	0.07
Review (in book or in journal)		0.12	0.25		0.76	0.07	0.32		0.04	0.10	0.10	0.15	0.09		0.06	0.03
Scientific comment											0.56	0.10			0.02	0.07
Software	0.07	0.21		0.09				0.09		0.04	0.01		0.09			
Translation					0.01	0.02					0.83	0.08		0.05	0.01	0.03
Totale	6062	10588	6897	4430	10986	16693	7541	3456	2832	11564	8744	6123	2276	8488	8385	2971

Table 5 - Research products in VQR2

In VQR2 the *papers in journals* still represents the more widely evaluated category, with the exception of the Areas 8a (Architecture) and 10 (Antiquity, philological-literary and historical-artistic sciences), where the *papers in books* dominate.

In the other documentary typologies, we find limited percentages of specific products for each disciplinary area, especially in VQR2. Among the products most frequently submitted for the evaluation are *curatorship* and *critical editions* in VQR1, *reviews* and *translations* in VQR2. However, the presence of products such as *curatorship*, *critical editions*, *abstracts*, *reviews*, *entries*, *catalographic record*, *translations*, *corpora*, etc. is determined by the criteria adopted by the GEVs. Indeed, each GEV had the possibility of defining more in detail the criteria determining the admission of the products to the evaluation, considering also the relevance in each research area, and the procedures applied to value their judgments. In some cases, the GEVs made different choices, including, for instance, in the entries *papers in journals* and *papers in books* products like *forwards/afterwards*, *lexicons*, *catalogues*, *guides*, *concordances*, *critical edition* and *publication of unedited sources* so.

Table 6 shows the frequency of GL by Area in VQR1 and VQR2.



VQR 2004-2010				VQR 2011-2014			
Areas	N. of products	N. of GL products	Frequency of GL products	Areas	N. of products	N. of GL products	Frequency of GL products
Area01	10685	38	0.36	Area01	6062	18	0.30
Area02		na		Area02	10588	145	1.37
Area03	11608	25	0.22	Area03	6897	24	0.35
Area04	na	na		Area04	4430	36	0.81
Area05	16407	51	0.31	Area05	10986	40	0.36
Area06	26713	54	0.20	Area06	16693	14	0.08
Area07	10004	82	0.82	Area07	7541	27	0.36
Area08	9533	148	1.55	Area08a	3456	105	3.04
Area09	16347	156	0.95	Area08b	2832	7	0.25
Area 10	14073	112	0.80	Area09	11564	69	0.60
Area 11-nbib		na		Area 10	8744	58	0.66
Area 11-bib		na		Area 11a	6123	19	0.31
Area 12		na		Area 11b	2276	8	0.35
Area 13	11941	121	1.01	Area 12	8488	247	2.91
Area 14	4327	20	0.46	Area 13	8385	57	0.68
				Area 14	2971	3	0.17
Total	131638	807	0.61	Total	118036	877	0.74

Table 6- Frequency by Areas

The frequency of the *Grey Literature* is 0.61 in VQR1 and 0.74 in VQR2. However, this is a rough estimation, as in the VQR1 calculations some Areas 2 (Physics), 4 (Earth sciences), 11 (Historical, philosophical, psychological and pedagogical sciences), and 12 (Legal sciences) had to be excluded, because the frequency of each product had not been stated in the GEVs' final reports.

Tables 7 and 8 show the products of GL by Area in VQR1 and VQR2.

VQR 2004-2010															
GL categories	Area 01	Area 02	Area 03	Area 04	Area 05	Area 06	Area 07	Area 08	Area 09	Area 10	Area 11-nbib	Area 11-bib	Area 12	Area 13	Area 14
	%	%	%	%	%	%	%		%	%	%	%	%	%	%
Case notes													0.56		
Composition								0.04							
Database	0.01			0.20	0.01			0.01							
Design	0.02							0.13							
Entry (in dictionary o encyclopedia)					0.01			0.03					0.29		
Exhibition	0.01							0.02							
Handwork	0.02	0.70	0.01					0.08							
Maps				0.20											
Other	0.13	0.30	0.01	2.28	0.04	0.07	0.82	1.09	0.44	0.80	0.61	0.22	0.64	1.01	0.46
Patent	0.10	0.20	0.39	0.18	0.26	0.13		0.14	0.51						
Prototype of art and related project								0.01							
Software	0.07		0.02	0.28	0.01										
Totale	10685	19773	11608	8433	16407	26713	10004	9533	16347	14073	9513	3639	11882	11941	4327

Table 7 - Grey products in VQR1 by Areas

In VQR1 the most relevant percentages are those referred to:

Area07 - Agricultural and veterinary science

Area08 – Civil engineering and architecture

Area10 – Antiquity, philological-literary and historical and artistic sciences

Area 12 – Legal Sciences

Area13 – Economics and statistics sciences



VQR 2011-2014																
GL categories	Area 01	Area 02	Area 03	Area 04	Area 05	Area 06	Area 07	Area 08a	Area 08b	Area 09	Area 10	Area 11a	Area 11b	Area 12	Area 13	Area 14
	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Architectonic project								1.45								
Bibliographic/Catalographic record, corpus		0.09			0.06	0.03	0.03				0.09					
Case notes					0.04	0.01		0.03					0.04	1.52		
Composition		0.75	0.01					0.03		0.05					0.01	
Database		0.04		0.18	0.03		0.08	0.03		0.04	0.05	0.02		0.02		
Design								0.12								
Entry (in dictionary or encyclopedia)					0.01			0.20		0.03	0.23	0.23	0.13	1.37	0.10	0.03
Exhibition		0.11		0.07	0.06	0.01	0.03	0.14	0.04	0.03	0.05					
Handwork																
Maps				0.23			0.04									
Other	0.21			0.18	0.01	0.01		0.52	0.04		0.17	0.07	0.09		0.55	
Patent	0.02	0.18	0.33	0.07	0.13	0.04	0.19	0.29	0.18	0.38					0.02	
Performance					0.03			0.03			0.07					0.07
Prototype of art and related project								0.12		0.03						
Software	0.07	0.21		0.09				0.09		0.04	0.01		0.09			
Totale	6062	10588	6897	4430	10986	16693	7541	3456	2832	11564	8744	6123	2276	8488	8385	2971

Table 8 - Grey products in VQR2 by Areas

In VQR2 the most relevant percentages are those referred to:

Area 2 – Physics

Area8a – Architecture

Area 12 – Legal Sciences

Area 13 – Economics and statistics sciences

A specific reference must be made to the category *other*, as it is not really clear which products includes. In GEVs' final reports, the entry is defined as the incorporation of a collection of different products, their number too small to be treated separately and impossible to merge into other categories. This item is present in all disciplinary areas in VQR1 and in multiple areas in VQR2, where the documentary categories are wider. Therefore, it is possible that some products have been classified more properly.

The following table shows the percentage distribution of GL by Areas and years.

Areas	VQR 2004-2010							VQR 2011-2014			
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Area01 - Computer science and Mathematics	0.15	0.36	0.45	0.31	0.31	0.61	0.25	0.22	0.27	0.44	0.25
Area 02 - Physics								0.88	1.24	1.03	2.65
Area03 - Chemistry	0.34	0.18		0.50	0.06	0.27	0.17	0.18	0.29	0.22	0.71
Area 04 - Earth sciences								0.58	0.95	0.78	0.91
Area05 - Biology	0.05	0.19	0.27	0.22	0.45	0.27	0.63	0.34	0.39	0.27	0.46
Area06 - Medicine	0.13	0.12	0.08	0.13	0.12	0.09	0.22	0.05	0.13	0.07	0.08
Area07 - Agricultural and veterinary sciences	1.43	0.81	0.58	1.19	0.68	0.45	0.84	0.35	0.32	0.15	0.63
Area 8 - Civil engineering and Architecture	1.39	2.04	1.47	2.25	1.05	0.96	1.95				
Area 8a - Architecture								5.42	2.62	3.54	2.09
Area 8b - Civil engineering								0.18	0.30	0.39	0.12
Area 09 - Industrial and computer engineering								0.56	0.51	0.50	0.81
Area 10 - Antiquity, philological-literary and historical-artistic sciences	0.64	0.64	0.61	0.99	0.86	0.74	0.93	0.69	0.71	0.87	1.12
Area 11 - Historical, philosophical, psychological and pedagogical sciences											
Area 11a - Historical, philosophical and pedagogical sciences								0.49	0.65	0.58	0.48
Area 11b - Psychology								0.40	0.36	0.33	0.32
Aea 12 - Legal sciences								3.19	2.43	3.08	3.16
Area 13 - Economics and Statistics sciences								0.78	0.47	0.76	0.77
Area 14 - Social and political sciences	0.85	0.20		0.15	0.29	0.93		0.17		0.13	0.38

Table 9 - GL percentages over the years

In VQR1 the annual trend is steadier for the following disciplinary Areas:

Area01 – Computer science and Mathematics

Area07 – Agricultural and veterinary sciences

Area08 – Civil engineering and Architecture

Area10 – Antiquity, philological-literary and historical-artistic sciences

For the other Areas, the annual trend is not steady because the values increase and decrease over the years.

In VQR2 the annual trend is steadier for most of the Areas.

Table 10 shows the different grey products and their annual distribution.



Grey categories	VQR 2004-2010							VQR 2011-2014			
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Architectonic project								38.00	24.00	20.00	18.00
Bibliographic/Catalographic record, corpus								25.81	32.26	19.35	22.58
Case notes								17.65	25.74	23.53	33.09
Composition	25.00	25.00		25.00	25.00			18.18	15.91	25.00	40.91
Database			33.33					23.53	20.59	20.59	35.29
Design	14.29	14.29	7.14	28.57		21.43	14.29	25.00	25.00	25.00	25.00
Entry (in dictionary or encyclopedia)		25.00			75.00			23.12	26.01	26.59	24.28
Exhibition		33.33					66.67	28.95	28.95	18.42	23.68
Handwork	9.09		9.09			27.27	27.27				
Map								7.69	38.46	30.77	23.08
Other	11.80	10.91	10.62	20.94	15.04	14.16	22.12	20.18	20.18	29.36	30.28
Patent	5.74	7.38	7.38	12.30	14.75	12.30	24.59	17.02	21.99	22.70	38.30
Performance									66.67	33.33	50.00
Prototype of art and related project							100.00	14.29	42.86		28.57
Software		10.00	20.00	20.00		40.00	10.00	9.76	39.02	21.95	29.27

Table 10 - Grey products in the VQRs by year

It seems evident that the extension of the documentary categories in VQR2 supported the presence of a wider range of research products, encouraging the submission of different typologies in each disciplinary area. This influenced also the incidence of the Grey categories in VQR2, where we find products that did not appear in the previous evaluation, such as *architectonic project*, *bibliographic/catalographic record* and *corpora*, *case notes*, *map*, *performance* and *publication of unedited sources*. The calculation of the total number of products per year in each category clearly shows that in VQR1 only *patents*¹⁰ and *other* are registered every year. Significant percentages are registered for *entries* (both *in dictionary* or *encyclopedia*) in 2008, *exhibition* in 2010 and *handwork*, permanent in 2004/2006 (9.09) and 2009/2010 (27.27). The *handwork* is completely absent in VQR2.

As far as the annual trend is concerned, some categories are more stable while others are more fluctuating through time. Only the percentages referred to the products in VQR2 are not conflicting. Two peaks for the entry *performance* are evident in 2012 and 2014, as well as a much higher value of the item *concordance* in 2012.

At the same time, some of the products present in both evaluations do not show significant annual variations, with the only exception of the category *entries* (both *in dictionary* or *encyclopedia*), presenting a very high value in 2008, as well as *exhibition* and *prototype* in 2010. In addition, there is a substantial increase of the entries *patents* and *other*, although not annually.

Among the non-admissible products, we find: *educational material*, *technical reports*, *commentary*, *obituary*, *erratum*. No mention has been made to entries such as *preliminary studies*, *progress reports*, *accounts*, *search results*, *dossier*, *market researches*, *normative documents*, *feasibility studies*, etc. The *genetic studies* or the *clinical trials* are admitted to the assessment only if the author is the person who carried on the work, whereas the participation of experimenters or collaborators in the study is not taken into account. Other indications inferred through the GEVs' criteria concern the *case notes*, admitted to the evaluation only if drafted in the form of an article. The entry *working papers* is not present, although it has been declared as admissible by all GEVs, with the exception of GEV13 – Economics and Statistics¹¹. It is not clear if they may have been included in the category *papers in journals* or *papers in books*.

5. Open Science and Grey Literature... a perfect marriage

As mentioned in the introduction, Open Science is a composite idea, promoting different approaches to research and scientific communication. We may represent OS using keywords such as *network*, *data*, *collaboration* and *transparency* (Adams 2015). The focus is on the cooperation and the distribution of information through advanced technology networks (Salmi 2015). We may also observe how the main contents of the OS are tightly related to *Grey Literature* (GL). Indeed, it is evident that the focal notions of the GL, from contents to production and distribution procedures, are shared by the OS, as it differs from traditional scientific methods of knowledge

¹⁰The category *patents* is always considered as evaluable, but it may be attributed to class A or B only if internationally renowned or licensed.

¹¹ The GEV13 considers these products as designed for future publication, therefore evaluable in forthcoming assessment exercises. For the other GEVs, if the *working papers* have a ISSN code, they are considered as articles; if they have ISBN code, and a) they are open access; b) the author divests the IP rights to the working papers' series; c) every manuscript is peer-reviewed, then they are related to *monographies*.



acquisition, generation and dissemination. Moreover, if we focus our attention on the principles of OS, we will identify numerous shared contents and goals between GL and Open Access. The latter certainly faces the drawbacks of the present publication system, proposing different rationales, opposed to those currently governing the editorial market. The open and unrestricted access to the research results aims at going beyond the logic of “publish or perish” and its difficulties, such as the substantial delays in the publication, the disconnection between what is scientifically relevant and what is relevant for the career of the researchers, with specific reference to the pressures on researchers to publish (Giglia 2017). GL and OA both promote the view of knowledge as a common good, as well as the immediate, detailed and in-depth circulation of the research products in every form, not exclusively those referable to scientific articles issued by publishing companies.

At the same time, the OS supports Open Data as a mean for sharing the so-called “computational products”, i.e. protocols, procedures and/or software used by researchers to analyse and share the data along with their experimental and practical context (Candela, Manghi 2017). The Open Notebook aims at voicing documents with a low degree of visibility and availability such as research notes and laboratory tests, encouraging new forms of open peer-review based on shared protocols for the evaluation of the products in the moment they are posted online.

Therefore, the common principle of GL and OA is distributing information not conventionally disseminated, like results of failed or ambiguous experiments (Stafford 2010). Moreover, the principle of the OA is based on the advantage of obtaining and co-generating new knowledge through the interaction with citizens and local communities (Reale 2017). The terms “Open Research” and “Citizen Science” make reference to the active civic participation to the collection of data, scientific experiments, and problem solving. The involvement of the population may bring to light urgent social needs and priorities, as well as drive the attention to issues of great interest, such as environmental protection, better health, more justice, a more equitable (not necessarily equal) distribution of income (Weber et al. 2015).

A further aspect of the OS is the Open Learning and Open Education, which consists in the introduction of innovative, interactive and collaborative education policies. Their main features are the openness, the digital dimension, and the innovative spirit (Reggiani 2017). In this way, the range of available resources, contents, and products of learning materials becomes wider and more shareable.

Likewise, in the relationship between OS and GL the availability of shared tools such as repositories is of crucial importance, particularly referring to those named *next generation repositories and infrastructures*. Indeed, institutional and disciplinary repositories are commonly accepted as the most suitable sites to disseminate and store the scientific production. They traditionally host different categories of *Grey Literature*, from dissertations to sketches, from conference objects to reports, etc. The OS encourages the creation of *next generation repositories*, archives based on a series of principles and requirements including value-added services.

The new repositories would be provided with a wider range of roles and functionalities, allowing the integration in an online distributed infrastructure. This view is sponsored by associations like *COAR – Towards a global knowledge commons* that promotes the implementation of next generation repositories provided with *11 new behaviours, as well as the technologies, standards and protocols that will facilitate the development of new services on top of the collective network, including social networking, peer review, notifications, and usage assessment*¹².

Another tool shared by both the Open Science movement and the “grey” community is the use of controlled vocabularies, as *the use of controlled vocabularies for bibliographic metadata “ensures that everyone is using the same word to mean the same thing”*¹³. The results published in institutional and thematic repositories are described by bibliographic metadata. The open access and metadata exchange requires a standardized description of specific properties concerning publications and research data. The review, updating, and curation of controlled vocabularies guarantee the semantic interoperability between repositories and linked archives.

Our study highlights the need of improving new-generation metrics, as this is one of the most debated topic in the OS movement. The results of our analysis outline two complementary paths to assess the research results: one is the *peer review*, a qualitative evaluation among scientists; the other consisting in a quantitative evaluation based on the use of bibliometric indicators that counts the number of publications and the number of citations received. As previously observed,

¹² Cfr. <https://www.coar-repositories.org/activities/advocacy-leadership/working-group-next-generation-repositories/>.

¹³ Cfr. <https://www.coar-repositories.org/activities/repository-interoperability/coar-vocabularies/>.



in Italy most of the documentary typologies undergoing evaluation are represented by *papers in journals*, *papers in books* and *papers in proceedings*, i.e. products that can be easily indexed in citation databases. The most important numbers generally refer to articles published in prestigious journals with high impact factors. Indeed, traditional metrics are based on indicators like Impact Factor and H-Index, this leading to the misapplication of principles born to count the number of citations, which are actually used to evaluate the value of researchers. Moreover, these indicators are the result of complex algorithms designed by commercial operators that select and organize information following criteria not always scientifically supported, often producing incorrect data (Galimberti 2017).

Bibliometrics has not been conceived for research evaluation, but to lead librarians in purchases and to measure progresses in scientific disciplines, in spite of the extensive use of its means.

Bibliometrists agree on the assumption that the indicators are misused, as the assessment of the research impact cannot be exclusively based on the calculation of citations, yet taking into account different aspects and dimensions. At the present time, *peer review* is the only and most effective type of qualitative analysis. However, although retrospective peer-review does not present the same level of criticality of perspective peer-review¹⁴, it is an onerous and subjective procedure, not completely free from inaccuracies and not completely objective.

The analysis of new research evaluation instruments and criteria is of basic importance for the fulfilment of the OS, as it is fundamental to transform the assessment models and methods along with the science itself (Cassella 2017). The realization of OS implies that the quality of a researcher and its publications would be evaluated through other parameters, such as: accuracy, reproducibility of the results, coherence of the methods, coherence with the ethical code, openness, and participation to editorial committees (Giglia 2015). Various international initiatives promote the application of new assessment standards for the evaluation of the scientific production¹⁵. The OS movement examines a series of alternative metrics to monitor the development of the scientific system and to measure both individual and group work. Among them we find the *usage metrics*, which are based on the number of views and downloads of a product. The usage measurement differs from the citations measurement since it involves a broader range of users who are potentially interested in reading and downloading articles and data. The metrics based on the usage appraise the interest and the degree of absorption of a work and may be quite relevant for the Open Science, not only for the use made of the publications, but also for the monitoring of non-traditional publications (e.g. posts, blogs) and for the reuse of open data and open software.

The *altmetrics* represents a further sub-system of new generation metrics mostly based on social media applications. More in detail, the *altmetrics* make reference to downloads, blog posts, social media interactions, citations, comments, tweets, opinions expressed by the users through means such as *likes* on Facebook. The *altmetrics* may contribute to the evaluation of the impact of a study. Indeed, nowadays researchers are increasingly exploiting the web in their studies; therefore, discussions among experts have shifted from laboratories' hallways to blogs and social networks, as well as the "raw science" (datasets, code, and experimental designs) finds place in blogging, microblogging, and annotations available and shareable online¹⁶.

The advantages of the use of *altmetrics* are quite clear: citations can be retrieved faster; authoritative, but not frequently cited works are not disregarded; the operating environment of

¹⁴ Among the most debated aspects, we find the fact that the reviewer is not always more competent than the author of the paper reviewed, in addition to potential conflicts of interest, the lack of accuracy, long times for the review process, implying consistent delays in the publication.

¹⁵ The *DORA Declaration* provides recommendations to funding agencies, institutions, publishers, organizations supplying metrics, and to researchers; it underlines the need of avoiding the use of indicators like the IF, instead considering the intrinsic value rather than the journal, and taking vantage of the opportunities provided by new digital indicators. The *Leiden Manifesto* contains 10 principles indicating possible solutions to the issues created by the inappropriate introduction of Bibliometrics, and suggests the usage of valid statistics along with a correct evaluation of the objectives and the nature of the research assessed. The *Science in Transition Position paper* highlights the continuous growth of exchanges in scientific information out of the traditional channels and documentation (e.g. journals and books), preferring more informal, fast and open modalities and self-production systems like blogs and microblogs. The *Metric Tide* examines the role of metrics in the evaluation and management of the research in the British system, emphasizing the limits of the quantitative measures and indicating *peer review* as the only possible yardstick. The report lists a series of recommendations for the design of metrics based on *robustness, humility, transparency, diversity, and reflexivity*.

¹⁶ Altmetrics: a manifesto, <http://altmetrics.org/manifesto/>.



the researcher is considered properly; different metrics for the evaluation of a study may be aggregated.

Although *altmetrics* are still in their experimental stage, the crisis of the traditional evaluation systems is so serious that the use of alternative indicators is expected to be promoted and improved in a very near future (Schöpfel, Prost 2017).

6. Conclusions

The Italian Research Assessment Processes do not completely exclude *Grey Literature*. However, they are almost exclusively based on the analysis of commercially distributed products.

This is due to:

- the non-eligibility of some research products (e.g. *educational material, technical reports, commentary, obituary, erratum...*);
- the lack of interest in items such as *preliminary studies, progress reports, accounts, search results, dossier, market researches, normative documents, feasibility studies*, etc.;
- the disadvantage in submitting scientific products other than articles in journals;
- the impact of the evaluation criteria on researchers leads to the philosophy of *Public or Perish*: the researchers publish only scientific articles in prestigious journals.

The combination of principles and tools of the OS may become a primary channel of dissemination for the GL. On the other hand, GL may evolve into a primary source for the OS. GL has a long-standing tradition, it is a mosaic of different documentary typologies including various areas of interest: from documentary research, to a wide range of materials produced by local, national and international private or public institutions, industry associations, foundations, private individuals, etc. Both GL and OS meet the need of faster, more efficient, economical, focused dissemination channels, insisting on the urgency of making available all the documentary forms currently not circulated and inaccessible. These documentary and procedural demands may be fulfilled by the tools of the OS. The new research scenarios offer considerable opportunities for the collection, description, identification, and dissemination of literature and data. Any kind of product may be identifiable and accessible using tools like repositories or new generation infrastructures, which supports all the components of the research activity: objects, people, technology, procedures (Candela, Manghi 2017). The use of these tools opens the status of scientific product to a wide range of documents: drafts, software, pictures, diagrams, tables, experimental protocols, then creating communication patterns of major interest either to those who would actively collaborate with researchers in their studies, or to those who would simply collect information or gain knowledge. GL, currently left out from traditional metrics, may be involved in the application of new typologies of metrics including aspects of scientific products usually labelled as *Grey Literature*. The *altmetrics* cross the borders of traditional *research results*, limited to databases like WoS and Scopus, taking into account non-traditional and non-commercial products. Exploiting such tools, the assessment exercises may finally turn into transparent, comprehensible, and shared processes.

The logic that moves the current scientific communication implies the risk of producing *fashionable* research rather than *quality* research. The concept of OS includes the necessary human skills, resources, standards, best practices and technical infrastructures necessary to realize an innovative *entire research enterprise*.

In this new ecosystem of the scientific communication *Grey Literature* might find its ideal collocation, but it is necessary that scientific institutions and politics exchange experiences and build networks across national borders in order to realize this new system, then allowing the growth of a new dialogue between science and society.

The developments in technology and the opportunities offered by the semantic web may have not led to the advancements expected about ten years ago. However, it has been widely demonstrated that the technologies available are ready to support substantial and complex goals. Cultural, political and economic changes are necessary in order to realize the Open Science. Europe plays a key role in supporting greater openness and in redefining the research processes.

**Bibliography**¹⁷

- 1) Adams J. (2015). *Impact of Open Science methods and practices on the economics of research and science. Case Studies from Life, Mathematical and Social Sciences*. European Commission, 2015.
- 2) Aliprandi S. (ed.) (2017). *Fare Open Access. La libera diffusione del sapere scientifico nell'era digitale*. LeEdizioni (ebook), <https://aliprandi.org/books/fare-openaccess/>.
- 3) *Altmetrics Status Quo*. OpenUP - OPENing UP new methods, indicators and tools for peer review, impact measurement and dissemination of research results, Deliverable D5.1. European Commission, 2016.
- 4) ANVUR – Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca, <http://www.anvur.org/>.
- 5) Baccini A. (2011). *Valutare la ricerca scientifica. Uso e abuso degli indicatori bibliometrici*. Il Mulino (e-book).
- 6) Biagetti M.T. (2017). Validità e limiti della *library catalog analysis* per la valutazione della ricerca nelle scienze umane e sociali. *AIB studi*, 57 (1)(gennaio/aprile 2017), 7-22.
- 7) Bianco L. (2010). *Riflessioni sulla valutazione della ricerca e l'istituzione dell'ANVUR*, <http://matematica.unibocconi.it/articoli/riflessioni-sulla-valutazione-della-ricerca-e-listituzione-dell'anvur>.
- 8) Candela L., Manghi P. (2017). La comunicazione scientifica i tempi della Open Science. *Paradoxa*, XI (1) 69-93.
- 9) Caso R. (ed.) (2009). *Atti del Convegno Pubblicazioni Scientifiche, Diritti d'autore e Open Access*. Università degli Studi di Trento.
- 10) Cassella M. (2013). Dati aperti e ricerca scientifica: aspetti gestionali e normativi nel contesto dell'e-science. *AIB studi*, 53 (3) 223-238.
- 11) Cassella M. (2014). Bibliometria sì, bibliometria no: la valutazione della ricerca nelle scienze umane e sociali al bivio. *AIB studi*, 54 (2/3) 295-304.
- 12) Cassella M. (2017). La valutazione alternativa: *altmetrics* e dintorni. *AIB studi*, 57 (1) 79-90.
- 13) Castellucci P. (2017). Formiche virtuali o virtuose? Verso un'etica dell'accesso. *AIB studi*, 57 (1) 51-62.
- 14) Ćirković S. (2017). Transition to Open Access and its Implications on Grey Literature Resources. *Proceedings of the Eighteen International Conference on Grey Literature (GL18)*, 81-90. TextRelease, Amsterdam.
- 15) CRUI – Commissione Biblioteche, Gruppo Open Access. *L'Open Access e la valutazione dei prodotti della ricerca scientifica. Raccomandazioni*. Roma, aprile 2009, <https://www.cru.it/oa.html>.
- 16) De Robbio A. (2007). Analisi citazionale e indicatori bibliometrici nel modello Open Access. *Bollettino AIB*, 3, 257-288.
- 17) Delfanti A. (2008). Collaborative Web between open and closed science. *Journal of Science Communication*, 7 (2), <https://jcom.sissa.it/archive/07/02/Jcom0702%282008%29C01>.
- 18) European Commission. *Next-generation metrics: Responsible metrics and evaluation for open science*. Report of the European Commission Expert Group on Altmetrics, <https://ec.europa.eu/research/openscience/pdf/report.pdf>.
- 19) Faggiolani C., Solimine G. (2012). La valutazione della ricerca, la bibliometria e l'albero di Bertoldo. *AIB Studi*, 52 (1) 57-63.
- 20) FORCE11 – The Future of Research Communications and e-Scholarship (2012). *Improving Future Research Communication and e-Scholarship. White Paper*. February 19, 2012. <https://www.force11.org/sites/default/files/files/Force11Manifesto20120219.pdf>
- 21) *The future of scholarly scientific communication*. Report, <https://royalsociety.org/science-events-and-lectures/2015/04/future-of-scholarly-scientific-communication-part-1/>.
- 22) Galimberti P. (2012). Qualità e quantità: stato dell'arte della valutazione della ricerca nelle scienze umane in Italia. *JLIS.it*, 3 (1), <https://www.jlis.it/issue/view/326>.
- 23) Galimberti P. (2017). Open Science: *altmetrics*, impatto e controllo della qualità. *Paradoxa*, XI (1) 29-38.
- 24) Galimberti P. (2017). Open access, open science. L'Italia, un paese in grave ritardo. *Intervento sul sito ROARS – Return On Academic Research*, <https://www.roars.it/online/?p=57268>.
- 25) *Interim Report on Researcher Impact*. OPENing UP new methods, indicators and tools for peer review, impact measurement and dissemination of research results – OpenUp Deliverable D5.3. European Commission, 2017.
- 26) Giglia, E. (2009). Più citazioni in Open Access? Panorama della letteratura con uno studio sull'Impact Factor delle riviste Open Access. *CIBER 1999-2009*, 125-145. Ledizioni, Milano.
- 27) Gioè L. (2016). Open Science. Per una democrazia della conoscenza. *Scienza in rete*, <http://www.scienzainrete.it/contenuto/partner/open-science-democrazia-della-conoscenza/marzo-2016>.
- 28) Guerrini M. (2009). Nuovi strumenti per la valutazione della ricerca scientifica. Il movimento dell'open access e gli archivi istituzionali. *Biblioteche oggi*, 27 (8) 7-17.
- 29) Leiden Manifesto for research metrics, <https://www.roars.it/online/leiden-manifesto-for-research-metrics/>.
- 30) The Metric Tide, <http://www.hefce.ac.uk/pubs/rereports/year/2015/metrictide/>.
- 31) Moed H.F. (2005). *Citation Analysis in Research Evaluation*. Springer, 2005.
- 32) Moed H.F. (2011). The multi-dimensional research assessment matrix. *Research Trends*, 23 (May 2011), <https://www.researchtrends.com/issue23-may-2011/the-multi-dimensional-research-assessment-matrix/>.
- 33) Pozzo R. (ed.) (2017). Scienziati, giù dalla torre d'avorio! *Paradoxa*, XI (1).
- 34) Reale E. (2017). Promuovere una scienza aperta: risorse, incentivi e valutazione. *Paradoxa*, XI (1) 15-27.
- 35) Rebora G., Turri M. (2013) - *The UK and Italian research assessment exercises face to face*. *Research Policies*, 42 (9) 1657– 1666.
- 36) ICSU – International Council for Science (2014). *Open access to scientific data and literature and the assessment of research by metrics*, <http://www.roars.it/online/wp-content/uploads/2014/09/ICSU-Report-on-Open-Access.pdf>.
- 37) Reggiani L. (2017). In vista d'alti cieli: l'Open Education tra conoscenza scientifica e società democratica. *Paradoxa*, XI (1) 95-106.

¹⁷ URL last access: December 2017.



- 38) ROARS – Return On Academic Research, <https://www.roars.it/>.
- 39) Rubele R. (2012). Appunti per una storia dell'ANVUR (I). *Intervento sul sito ROARS – Return On Academic Research*, <https://www.roars.it/online/appunti-per-una-storia-dellanvur-i/>.
- 40) Salmi J. (2015). *Study on Open Science. Impact, Implications and Policy Options*. European Commission, 2015.
- 41) San Francisco Declaration on Research Assessment, <http://www.ascb.org/dora/>.
- 42) Serini P. (2003). Attualità della letteratura grigia: *il ruolo delle biblioteche nella sua valorizzazione*. *Biblioteche oggi*, 21 (1) 61-72.
- 43) Stafford N. (2010). Science in the digital age. *Nature*, 467 (7317) S19-S21.
- 44) Scholarly Publishing and Academic Resources Coalition Advocating change in scholarly communications for the benefit of researchers and society. *Better ways to evaluate research and researchers A SPARC Europe Briefing Paper*, <https://sparceurope.org/new-sparc-europe-briefing-paper-better-ways-evaluate-research-researchers/>.
- 45) Schöpfel J., Prost H. (2017). Altmetrics and Grey Literature: Perspectives and Challenges. *The Grey Journal*, 13 (1) 5-11.
- 46) Turbanti S. (2016). La visibilità – e l'impatto? – nel Web ai tempi dei social: i principali strumenti di *altmetrics*. *AIB studi*, 56 (1) 41-58.
- 47) Weber M., André D., Llerena P. (2015). *A new role for EU Research and Innovation in the benefit of citizens: towards an open and transformative R&I policy. Policy Paper by the Research, Innovation, and Science Policy Experts (RISE)*. European Commission, 2015.
- 48) <https://opensource.com/resources/open-science>.
- 49) <http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/open-science-movement/>.
- 50) http://www.anvur.org/index.php?option=com_content&view=article&id=122&Itemid=305&lang=it.
- 51) http://www.anvur.org/index.php?option=com_content&view=article&id=32&Itemid=372&lang=it.
- 52) <http://www.anvur.org/rapporto/>.
- 53) http://www.anvur.org/index.php?option=com_content&view=article&id=825&Itemid=599&lang=it.
- 54) http://www.anvur.org/index.php?option=com_content&view=article&id=841&Itemid=601&lang=it.
- 55) <http://www.anvur.org/rapporto-2016/>.
- 56) <https://sites.google.com/site/scienzaapertaricercamigliore/programma>.
- 57) <http://www.anvur.org/rapporto/> (VQR1) and <http://www.anvur.org/rapporto-2016/> (VQR2).

Your 7 steps to sustainable data



1. **Prepare your data**

Select the relevant data files. Check them for privacy aspects and file format against the guidelines issued by DANS.



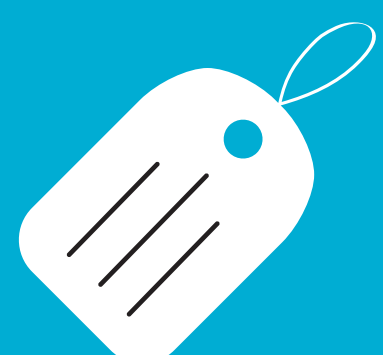
2. **Go to EASY**

Log in at <https://easy.dans.knaw.nl>. If you are new to EASY, you will have to register for an account first.



3. **Start the deposit procedure**

Go to 'deposit your data', select your discipline and click 'start deposit'.



4. **Documentation and access level**

Describe the dataset and indicate whether it is open access or whether access restrictions apply.



5. **Upload your data files**

Select your data files and click 'upload dataset'.



6. **Submit your data files**

Accept the licence agreement and send your dataset to DANS by clicking the 'submit' button.



7. **Publication by DANS**

DANS will verify the dataset and publish the description you made. Your data have now been sustainably archived and will be accessible to others on a permanent basis under the conditions you specified.



Newsletter:
[DataLinkDataLink](#)



Twitter:
[@DANSKNAW@DANSKNAW](#)



You Tube:
[DANSDataArchivingDANSDataAr](#)



E-mail:
info@dans.knaw.nlinfo@dans.knaw.nl



Website:
dans.knaw.nl



Data Papers are Witness to Trusted Resources in Grey Literature: A Project Use Case

Dominic Farace, GreyNet International, Netherlands
Jerry Frantzen, TextRelease, Netherlands
Plato L. Smith II, University of Florida Libraries, United States

Abstract

In 2011, GreyNet embarked on an Enhanced Publications Project with DANS, Data Archiving and Networked Services in an effort to circumvent the data versus document camps entrenched in grey literature communitiesⁱ. The results of that project served to incorporate in GreyNet's workflow the acquisition, indexing, and linking of research data with their full-text and accompanying metadata^{ii,iii}. Recently, GreyNet discussed with a data management librarian from the University of Florida a proposed follow-up project dealing with data papers – a new document type within grey literature. Data papers are defined as scholarly publications of a searchable metadata document describing a particular online accessible dataset or a group of datasets published in accordance to standard academic practices. As such, data papers represent a scholarly communication approach to data sharing^{iv}. The outcome of that discussion led to the formation of a project team with the twofold purpose of producing and publishing a set of data papers originating in the field of grey literature, and in so doing raise awareness to this new document type by demonstrating its value for library and information science. The method of approach includes the construction of an online standardized template that encompasses a data paper, defining the population asked to complete the template, instruction and further contact with the authors/researchers during the course of the project, along with an analysis of the results. The anticipated outcome of the project would provide a tested template that could be used by other grey literature communities in the production of data papers. It would demonstrate how OA, DSA^v and FAIR^{vi} principles are implemented and reinforced via data papers. And, it would further provide examples of how data citations can generate trusted bibliographic references.

Introduction

A few years ago, I heard the term data paper; and, as with all developments in the field of information, I began to think of how data papers are related to grey literature; and how GreyNet could incorporate them in and among its own resources? What sparked my initial interest was that data papers can be seen as a new document type in grey literature and that they have a place within GreyNet's ongoing Enhanced Publications Project dating back to 2011.

In early April of this year, I attended a meeting in Barcelona during the Research Data Alliance Conference. There a presentation on FAIR Data Principles was delivered; and, it was this that prompted me to explore the connection between data papers and FAIR Data principles. My position statement was and remains that 'Data papers are one of the most tangible means of implementing the FAIR-data Principles.'

The content of this paper is quite straightforward and reflects work over the past six months, or should I say the first 6 months of GreyNet's Data Papers Project.

Definition of Data Papers

The definition of data papers used for our project is borrowed from Wikipedia, where they are defined as "Scholarly publications of a searchable metadata document describing a particular online accessible dataset or a group of datasets published in accordance to standard academic practices. As such, data papers represent a scholarly communication approach to data sharing".

Background of the Project

Early in the project, it was understood that Data Papers was not a stand-alone project, but was very much linked to GreyNet's enhanced publications described by DRIVER II as "A publication that is enhanced with three categories of information: research data, extra materials, and post-publication data". Enhanced publications combine textual resources *i.e.* documents intended to be read by human beings, which contain an interpretation or analysis of primary data. Enhanced publications inherently contribute to the review process of grey literature as well as the replication of research and improved visibility of research results in the scholarly communication chain.



Developments in GreyNet's Data Collection

In 2011, GreyNet carried out a survey among its pool of authors and researchers in order to assess their attitudes towards data sharing and their willingness to submit research data for entry in the DANS Data Archive. The results of the survey were quite positive. While only 7% were unwilling to do so, 49% were, and 44% were as yet uncertain. In 2012, the year following that survey, the work of acquiring data retrospectively as well as integrating the acquisition of research data within GreyNet's workflow was undertaken. Since then, the tasks of publishing and linking GreyNet's research data with their full-texts account for 30 published datasets in the DANS Data Archive. This number of datasets was seen to provide the needed impetus and volume to undertake GreyNet's Data Papers Project. And, in early 2017, the compilation and design of a standardized template for data papers began.

The instrument of a data paper template

The data paper template implemented in this project is based on a compilation derived from three existing templates: RDJ, Research Data Journal for Humanities and Social Sciences^{vii}, JOHD, Journal of Open Humanities Data^{viii}, and JOPD, Journal of Open Psychology Data^{ix}. While these and other templates used for data papers contain similar fields, GreyNet opted to compile and adapt its own template, which consists of five main sections: Overview, Methods, Dataset Description, Potential Reuse, and References (see *Appendix*).

In drafting the template, each field in a section would provide a box note containing one or more examples. It was intended that once all 5 sections were completed, this would serve in drafting the data paper. Attention was further given to the overall design of the template in engaging would-be authors. Potential authors would not be confronted with just a sheet of instructions, but instead an instrument guiding them through to completion.

Population and Acquisition of Data Papers

As mentioned earlier, the number of datasets that formed the population for the project was thirty in total. It was determined that only first authors of GreyNet's datasets in the DANS Archive would receive an email request for participation in the Data Papers Project. This then amounted to 15 acquisition requests allowing for the fact that one author/researcher accounted for more than one dataset in the study. The text of the email read as follows:

Dear ... ,

GreyNet would like to enhance and promote its collection of datasets in the DANS Data Archive through the publication of a Data Paper corresponding to each dataset. According to our records, you have one or more datasets in the DANS Data Archive [[link](#)] that would benefit by an accompanying Data Paper.

Attached is a brief description of the Data Papers Project along with a standardized template to guide you in drafting a data paper. Also attached is a sample data paper that will appear published in the Autumn 2017 issue of The Grey Journal.

By drafting and submitting a data paper, you will also receive well deserved recognition through its publication in The Grey Journal as well as its preprint access in the DANS Data Archive. On behalf of the project team, I greatly appreciate your consideration in this request; and I look forward to fielding any questions you may have.

GreyNet' email signature

Review process for data papers

Upon receipt of a drafted data paper, the author received notification along with details of the review process that would be undertaken.

1. Submissions are first checked against the standardized template. The author(s) may be asked to provide additional information.
2. The (revised) data paper is then sent to the data management librarian on the project team for review.
3. Once reviewed the data paper is entered in the DANS Archive as a preprint, where it is processed by one of the data archive managers. It is assigned Creative Commons Licensing CC0, and provided the archive's Data Seal of Approval, DSA.



4. Having fulfilled the above, the data paper is considered to implement the FAIR-data Principles by making the related data(set) Findable, Accessible, Interoperable, and Reusable.

Publication of GreyNet's data papers

The data paper first published as a preprint alongside its corresponding dataset in the DANS Archive is further published as a preprint in the GreyGuide Repository, where via its assigned DOI becomes linked to its corresponding (data)set in the DANS Archive. The data paper is then scheduled for publication as an article in The Grey Journal (ISSN 1574-1796), where it receives coverage through multiple abstract and indexing services including Scopus and Thomson Reuters and is full-text available via EBSCO's LISTA-FT database.

Some early project results

Since the initial request to 15 first authors responsible for GreyNet's Collection of published datasets in the DANS Archive, results appear to show:

- The Data Paper Template compiled and edited for this Project proves efficient;
- Seven authors have indicated interest in drafting a Data Paper, five of which are now published as preprints in the DANS Data Archive and in the GreyGuide Repository. Two of the five data papers have appeared published as articles in the Autumn 2017 issue of The Grey Journal (TGJ) and the other three await journal publication in the Spring issue 2018;
- Data Papers are now listed in the GreySource index^x as a new document type in grey literature;
- And, Data Papers are shown to lend support to the Pisa Declaration on Policy Development for Grey Literature Resources^{xi} and specifically to Article 15 "Systems for linking data and other non-textual content to their grey literature publications together with interoperability standards for sharing grey literature."

Spin-off from the initial Data Papers Project

About six weeks after the initial round of acquisitions for data papers, it was brought to our attention that a number of published articles in The Grey Journal – with no tie to the GL-Conference Series – are based on research data. A request was then made to eight of these authors/researchers inquiring as to their willingness to participate in our project by submitting their data(sets) for entry in the DANS Archive followed by the completion of a Data Paper. Two of the eight authors have since responded with their intent to do so.

Another spin-off from our initial project is to draft a workshop that will instruct authors, researchers, and other information professionals and practitioners as to the benefits of data sharing and the instrument of a data paper in accomplishing this. The first of these workshops will be held at the University of Florida at Gainesville on March 20, 2018. It will be branded and offered within the GreyForum Series entitled "Data Papers, A Trusted Tool in Research and Data Sharing".

Follow-up of the Project

Just as the acquisition of data/datasets are incorporated in the workflow of the Annual Conference Series on Grey Literature, starting in 2018 data papers will also be included in the acquisition round. Furthermore, the request for research data and related data papers will likewise become incorporated in the workflow of The Grey Journal.

User statistics are a way of assessing the implementation of FAIR-data principles via data papers. While user statistics in 2018 may be limited, built into our project are the use of mechanisms that can capture the use of data papers. This will be accumulated from the following sources:

Source of Stats →	The Grey Journal LISTA-FT Database	GreyGuide Repository	DANS Data Archive
Data Paper (Article)	✓		
Data Paper (Preprint)		✓	CC0
Data/Dataset			CC0



Citations and references pertaining to data papers as well as datasets may be acquired via Abstract and Indexing Services. GreyNet maintains established agreements with CSA/PAIS International, LISA INDEX, MLA International, Scopus, and Thomson Reuters for The Grey Journal. For the Conference Proceedings on Grey Literature, GreyNet has standing agreements with Thomson Reuters and Curran's Scopus/Compendex. While these are not the only sources in which to turn, they are the most readily available for our project.

Also, a number ways to solicit and capture feedback from the authors and users of GreyNet's collections of published data(sets) and their concomitant data papers will be in place as follows:

Source of Feedback →	GreyNet Website	GreyGuide Repository	GreyForum Workshop
Data Papers	✓	✓	✓
Research Data	✓		✓

In fine, the tasks of our project team for the coming year involve the acquisition and publication of data papers along with the collection of user statistics, data citations, and references to the data papers. In turn, these results will be incorporated in subsequent workshops and trainings demonstrating use cases with Data Papers.

References

- ⁱ Linking full-text grey literature to underlying research and post-publication data: An Enhanced Publications Project 2011-2012 <https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:53456>
- ⁱⁱ GreyNet's Enhanced Publications Project: Tracking and Backtracking Data, 2012 <http://greyguide.isti.cnr.it/index.php/49-gl14/gl14-slide-share/416-gl14-farace-et-al-2>
- ⁱⁱⁱ Frequently Asked Questions (FAQ): Enhanced Publications Project (EPP), 2013 <http://www.greynet.org/images/FAQ-EPP.pdf>
- ^{iv} Data Papers definition https://en.wikipedia.org/wiki/Data_publishing#Data_papers
- ^v DSA, Data Seal of Approval <https://www.datasealofapproval.org/en/>
- ^{vi} FAIR Data Principles <https://www.force11.org/group/fairgroup/fairprinciples>
- ^{vii} Research Data Journal for Humanities and Social Sciences http://www.brill.com/sites/default/files/ftp/authors_instructions/RDJ.pdf
- ^{viii} Journal of Open Humanities Data <https://openhumanitiesdata.metajnl.com/about/submissions/>
- ^{ix} Journal of Open Psychology Data <https://openpsychologydata.metajnl.com/about/submissions/>
- ^x GreySource Index <http://www.greynet.org/greysourceindex/documenttypes.html>
- ^{xi} Pisa Declaration on Policy Development for Grey Literature Resources <http://greyguide.isti.cnr.it/pisa-declaration/>



APPENDIX – Data Paper Template

This template provides you with a standardized, accepted format for drafting a data paper. Please consult the note box found under each field name for brief explanations and examples. The Data Paper consists of five main sections: **1** Overview, **2** Methods, **3** Dataset Description, **4** Potential Reuse, and **5** References.

Once you have completed drafting your data paper, please submit to journal@greynet.org. Upon receipt, the preprint will be deposited in the [DANS Data Archive](#) where the dataset is stored. Further notice of its publication in [The Grey Journal](#) will likewise be provided. The above email address can be used for all correspondence pertaining to your data paper.

1 Overview

Title:

NOTE The title of the data paper should focus on the **data**. Since your data is closely linked to a specific research paper, precede the title of your paper with **Data from “followed by the Title of your paper”**.

Repository Location:

NOTE A link to the repository home page and a DOI (or other persistent identifier) that links directly to the dataset.

Data Paper Authors/Contributors, Affiliations, and Roles:

NOTE

1. ▪ First name, Last name;
▪ Author/Contributor unique ID (e.g. ORCID, ResearcherID)
▪ Organizational Affiliation;
▪ Author’s Role(s): Project Administration / Dataset Creator / Visualization / Draft Text / Review & Editing

2. ...

3. ...

Abstract:

NOTE A short (ca. 100 words) summary of the dataset being described: what the data covers, how it was collected, how it is stored, and a short description of its potential reuse.

Keywords:

NOTE A maximum of four keywords is allotted. Each keyword is separated by a semicolon followed by a space.

Subject Area:

NOTE Discipline or Community of Practice (e.g. Medicine, Engineering, Physics, etc.)

2 Methods

NOTE Describe the methods used to create the dataset (ca. 100-200 words), include the following sub-headings:

Steps – The series of procedures followed to produce the dataset. This should include any source data used, as well as software and instrumentation involved.

Sampling strategy – If relevant, outline the sampling strategy used to produce the data.

Quality Control – If applicable. Please list the methods used for quality control in the production of the data *i.e.* steps taken to normalize data.



3 Dataset Description

NOTE Enter your response after each sub-heading

File name:

The name of the file or file set in the repository/archive.

Format names and versions (if available):

Such as ASCII, CSV, Autocad, EPS, JPEG, Excel, SQL, etc.

Creation dates:

The start and end dates of when the data was created YYYY-MM-DD

Language(s):

Languages used in the dataset (i.e. variable names, etc.)

License

The open license under which the data has been deposited is Creative Commons, e.g. CC0.

Repository/Archive name:

The name of the repository/archive to which the data is uploaded: DANS EASY, Etc.

Publication date

The date the dataset was published in the repository/archive (YYYY-MM-DD)

4 Potential Reuse

NOTE Please describe the ways in which your data could be reused by other researchers both within and outside of your field. For example, this might include aggregation, further analysis, reference, validation, teaching or collaboration. This section should also include limitations to, or potential barriers for reuse. (maximum 800 words)

5 References

NOTE References cited here should be explicit to the Data Paper. When available, include a DOI, other persistent identifier, or link in each reference. See examples below:

1. Piwowar, H.A. 2011 Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS ONE* 6(7): e18657. DOI: <http://dx.doi.org/10.1371/journal.pone.0018657>.
2. Giannini, S., [et al] CNR Pisa, Italy, 2016 Grey Literature citations in the age of Digital Repositories and Open Access. Seventeenth International Conference on Grey Literature. Conference Preprint available from: <http://hdl.handle.net/10068/1024659>.
3. National Heart Lung and Blood Institute (US). What is a heart attack? [Internet]. Bethesda (MD): U S Department of Health and Human Services, National Institutes of Health; [cited 2009 Apr 3]. Available from: <http://www.nhlbi.nih.gov/actintime/aha/what.htm>.

Upon completion of your Data Paper, please submit to journal@greynet.org.
The email can also be used for all other correspondence pertaining to your Data Paper.

Thank You!



Public Access to the Dissertations in Russia

Yuliya B. Balashova

Saint Petersburg State University, Russia

Abstract

To date, a highly constructive situation has developed in the Russian public space in terms of access to dissertations. Mostly thanks to the efforts of official authorities, the procedure for defending dissertations becomes so public for the first time after a long period. At the same time, increasing publicity contributed to the consolidation of civil society and representatives of the scientific community efforts. Accordingly, open access resources containing information about the dissertation thesis, and its author are affiliated both with official and commercial and public organizations.

Keywords: *dissertation thesis, grey literature, science communication, Russian scientific public field.*

Introduction

In the modern Russia as a part of the global world science communication develop quite actively. Its growth is caused, among other things, by the task of wide audience familiarizing with the scientific thought achievements. New popular science media are being created, as well as aggregators of scientific news, scientific festivals are held, science promoted through the social networks.

At the same time, the system of training scientific staff and awarding academic degrees in Russia is still poorly integrated into the world community. The problem is that this system is hybrid, combining Soviet and modern Western properties. From the Soviet times, a complex two-level system of scientific degrees has been existing.

In the USSR, science was respected, partly in the ancient, magical sense, on one hand, from the state, and, on the other hand, from society. Obtaining a scientific degree opened many social possibilities.

The Russian degree of "Candidate of Sciences" corresponds to the generally accepted Ph.D. This degree is considered as the third level of a multilevel system of higher education and is awarded as a result of mastering the doctoral educational program accepted in the West, which in Russia is called "graduate school".

In some European countries awarded the highest scientific degree, named "doctor habilitatus" (Dr. habil. = habituated doctor). It gives the right to occupy a professorship at the university and corresponds to the Russian degree of the "Doctor of Sciences". In order to count on a scientific career in Russia, it's necessary to defend not one, but two dissertations.

In different countries, the state has a different power in regulating the procedure for awarding academic degrees, and accredited doctoral (Ph.D.) programs. In such European, and non-European countries, as France, Norway, Spain, Sweden, United Kingdom, New Zealand, state, and authorized with the state structures play a crucial role in this process. In the USA, and Canada, in many ways existing in a single space, this role belongs to the socio-professional organizations. Finally, in Netherlands the Academy of Sciences has the right to make decision. The right to assign academic degrees is different too. In Russia, and post-Soviet countries, this right is reserved for the state; in the rest of the world – for the universities themselves. Accordingly, in different countries the information support of the dissertation's defense and public access to them are formed in various ways. In Russia, the main role in this process belongs to the state.

Discussion

1. Obligatory stages of the dissertations publicity before defense.

1.1. Publication on the website of the organization, where the work was prepared. There are published the full text of the dissertation, the author's abstract of the dissertation, information about official opponents, and the leading organization, preparing review for the dissertation, as well as the place and time of dissertation's defense. From 2012, at Saint Petersburg State University all dissertations defenses are accompanied by online translation on the university website (Ph.D. SPbSU, 2017).



1.2. Publication on the website of High Attestation Commission under the Ministry of Education, and Science of the Russian Federation. Since 2006, High Attestation Commission publishes on its website the same information, only in a reduced amount, and with reference to the organization's website (High Attestation Commission, 2017)

1.3. The main results of the dissertation research should be published in the form of science papers in the list of peer-reviewed scientific journals. High Attestation Commission on a regular basis revises this list.

2. Obligatory stages of the dissertations publicity after defense

2.1. Obligatory dispatch of the dissertation abstract to the leading libraries of Russia, and also National Library of Belarus (National Library of Belarus, 2017).

2.2. The full written text of the dissertation is sent to the Russian State Library. The main library of the country provides public access to the digitized abstracts, and thesis themselves. Site navigation is available in English (Russian State Library, 2017).

2.3. A mandatory copy of both the thesis and the author's abstract is sent by the regular mail to the Russian Book Chamber (Russian Book Chamber, 2017). In accordance with law, a certain number of copies of any printed publication (books, brochures, periodicals, dissertations, musical notes, geographic maps and atlases) must be sent to the Russian Book Chamber.

3. Other (additional) online channels for dissertations distribution

3.1. Official state resources. The most important is Scientific Electronic Library (eLIBRARY.RU). Materials placed in this electronic library included into the Russian Scientific Citation Index. This Index is a National bibliographic database of scientific citations, accumulating more than 12 million publications of Russian scientists, as well as information on citing these publications from more than 6,000 Russian journals. Other resource, having a printed version, is network encyclopedia "Famous Scientists" (Famous Scientists, 2017). This is a biographical data of scientists, a project of the Russian Academy of Natural History. It also published the main results of the dissertation researches.

3.2. Commercial resources. To such belongs "Library of Dissertations" (DsLib.net). It provides both free and paid access to the thesis text, publishes information about upcoming dissertation defenses. At the same time, monetary deductions to authors of dissertations are declared, but in reality they do not work. We can give another example. "Cyber Leninka" (CyberLeninka, 2017) positions itself as a scientific open access electronic library, the main tasks of which are popularization of science and scientific activity, public control of the quality of scientific publications, the development of interdisciplinary research, the modern institute of scientific review, and the increasing citation of Russian scientists. "Cyber Leninka" is built on the basis of the open science paradigm.

The fundamental problem is that thesis and abstracts both are published "on the rights of the manuscript", in other words, they don't formally have copyright subject. So, private distributors don't pay in fact any royalty that violates the norms of morality, but not the law.

3.3. Public organizations resources. The main source of exposing falsified dissertations is network "Dissernet". It positions itself as "a free network of experts, researchers and reporters, who dedicate their work to exposing scammers, falsifiers and liars. The participants of the community are cooperating with the joint efforts based on the principles of the network distribution of labor and using modern computer technologies, to counteract illegal frauds and infringements in the field of scientific and educational activities, especially in the process of defending dissertations and appropriating academic degrees in Russia" (Dissernet, 2017). More than 10,000 dissertations have been analyzed by "Dissernet", since 2013. Based on "Dissernet" requests, High Attestation Commission made a decision on deprivation of academic degrees. The latest scandal was the case, initiated by "Dissernet", about deprivation of the Minister of Culture of the Russian Federation Vladimir Medinsky academic title of Doctor of Historical Sciences. His dissertation was devoted to the problem of objectivity in the Russian history coverage by foreigners. The main argument of the expert community was that Medinsky has considered the methodology of the historical research as corresponding to the national interests of the state. With this approach, historical science turns into a servant of the dominant ideology, as it happened in Soviet times. This approach is typical for historical pseudoscience. Nevertheless, at the final meeting on this



case, which took place on October 20, 2017, the presidium of the High Attestation Commission retained doctoral degree to Minister V. Medinsky. Such disclosures of the political establishment representatives often occur in Germany (thanks to “VroniPlag Wiki”, as well). And this is an indicator of the healthy society. In Russia, this case additionally indicates that the scientific reputation itself becomes irrelevant, which means a huge problem.

Conclusion

The most positive effect of the current situation is that never before science in Russia has been so public. This result was achieved due to the efforts both of the state (its scientific policy) and society. The next step will be integration this openness into the Western system.

But, in the present time, the procedure for awarding academic degrees begins to change. A number of leading Russian universities and scientific organizations in the near future will receive the right to self-award scientific degrees. On the one hand, thus, Russian universities become more integrated into the Western system. The supervising role of the state will be cancelled; the dissertational councils will not act on a permanent basis, but will be created in accordance with the subject matter of the concrete work. The defenses will take place in English, and – the most important think – the academic degree will be awarded directly. But such innovations cause a lot of criticism.

It is believed that in a democratic society decentralization is extremely constructive. In the scientific policy of modern Russia, this kind of decentralization causes great concern about maintaining a high level of publicity. Actually, there is a return to the closed Soviet system of leading universities.

One of the most serious concerns: will not the thesis be less public, and the thesis defense process more closed?

Acknowledgment

The author gratefully appreciates the support of the Russian Foundation for Basic Research (RFBR).

Funding Information


Article is prepared with financial support of the Russian Foundation for Basic Research. The project No.16-03-50128.

Ethics

This article is original and contains unpublished material. The authors confirm that there are no ethical issues involved.

References

- CyberLeninka. Accessed October 2017. <https://cyberleninka.ru/>
- Dissernet. Accessed October 2017. <https://www.dissernet.org/about/>
- Famous Scientists. Accessed October 2017. <https://www.famous-scientists.ru/about/>
- High Attestation Commission. Accessed October 2017. <http://vak.ed.gov.ru>.
- Library of Dissertations. Accessed October 2017. <http://www.dslib.net/>
- National Library of Belarus. Accessed October 2017. <http://www.nlb.by/portal/page/portal/index>.
- Ph. D. SPbSU. Accessed October 2017. <https://disser.spbu.ru/ph-d-spbsu.html>.
- Russian Book Chamber. Accessed October 2017. <http://www.bookchamber.ru>.
- Russian State Library. Accessed October 2017. <http://www.rsl.ru>.
- Scientific Electronic Library. Accessed October 2017. <https://elibrary.ru/defaultx.asp>.



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



How open access policies affect access to grey literature in university digital repositories: A case study of iSchools

Tomas A. Lipinski and Katie Chamberlain Kritikos

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

Abstract

Problem/Goal: An issue of interest to library and information science (LIS) scholars and practitioners is how open-access policies can affect the access and use of grey literature in university repositories. Open access (OA) refers to research placed online free from all price barriers and from most permission barriers (Suber, 2015), allowing unfettered access to scholarship and promoting open scholarly communication (Banach, 2011; Eysenbach, 2006). OA may apply research published traditionally, such as books (Schwartz, 2012) and academic articles (Suber, 2015), and non-traditionally published grey literature, such as electronic theses and dissertations (Schöpfel & Prost, 2013; Schöpfel & Lipinski, 2012). The treatment of grey literature in university repositories is of particular import due to “the ephemeral and changing nature of grey publication types, editions, and formats” (Rucinski, 2015, p. 548; see Farace & Schöpfel, 2010). The access and use of grey literature in these repositories is often executed through an OA policy. There is a gap in the literature, however, regarding best practices for drafting and implementing OA policies that promote unfettered access to grey literature.

Research Method/Procedure: This paper analyzes OA policies from a sample of U.S. iSchools, created by cross-referencing the Directory of North American iSchools (iSchools, 2017) with the top twenty-five best LIS programs as ranked by U.S. News and World Reports (U.S. News, 2017). Initial analysis shows that of the twenty-two iSchools in the sample, all schools have repositories, ten have OA policies in place, and three have proposed OA policies. This project maps five OA policies against variables drawn from the benchmark of open scholarly communication, the Harvard Open Access Project’s “Good Practices for University Open-Access Policies” (Shieber & Suber, 2017).

Results: The goal of this paper is to understand how OA policies at university repositories affect access to grey literature in an ever-changing information landscape. Based on the analysis of the sampled iSchool OA policies and the Harvard variables, it recommends best practices for drafting and implementing OA policies that provide unfettered access to grey literature in repositories.

Keywords: grey literature, open access, information policy, information access, university repository, best practice, scholarly communication, library science, information science

Introduction

An issue of interest to library and information science (LIS) scholars and practitioners is how open-access policies can affect the access and use of grey literature in university repositories. Open access (OA) refers to research placed online free from all price barriers and from most permission barriers (Suber, 2015), allowing unfettered access to scholarship and promoting open scholarly communication (Banach, 2011; Eysenbach, 2006). OA may apply research published traditionally, such as books (Schwartz, 2012) and academic articles (Suber, 2015), and non-traditionally published grey literature, such as student electronic theses and dissertations (ETDs) (Schöpfel & Prost, 2013; Schöpfel & Lipinski, 2012).

The treatment of grey literature in university repositories is of particular import due to “the ephemeral and changing nature of grey publication types, editions, and formats” (Rucinski, 2015, p. 548; see Farace & Schöpfel, 2010). Repositories are “digital collections capturing and preserving the intellectual output of a single or multi-university community” (Crow, 2002, p. 1; see also Lynch, 2003). In effect, they serve as “[o]pen archives” that a university or other institution hosts to “control and distribute” research and “become a significant part of scientific communication” (Schöpfel & Lipinski, 2012, p. 21). The access and use of grey literature in these repositories is often executed through an OA policy. There is a gap in the literature, however, regarding best practices for drafting and implementing OA policies that promote unfettered access to grey literature.

To study the impact of the OA phenomenon on LIS scholarly communication, this paper analyzes OA policies from a sample of U.S. iSchools, created by cross-referencing the Directory of North American iSchools (iSchools, 2017) with the top twenty-five best LIS programs ranked by U.S. News and World Reports (U.S. News, 2017). Initial analysis shows that of the twenty-two



schools in the sample, all have repositories, only ten have OA policies, and three have proposed policy drafts not yet adopted.

This project maps these policies against variables drawn from the benchmark for open scholarly communication, the Harvard Open Access Project's "Good Practices for University Open-Access Policies" (Shieber & Suber, 2017). The goal is to understand how OA policies at university repositories affects access to grey literature in an ever-changing information landscape. Based on the Harvard variables, it recommends best practices for drafting and implementing OA policies that balance copyright with provide unfettered access to grey literature in repositories.

Literature Review

Open Access and Scholarly Communication

As Margaret (2016) and Bohannon (2013) detail, backlash against the traditional publishing paradigm effected a shift to OA publishing (see also Armbruster, 2008). In 2002, the Budapest Open Access Initiative ("BOAI"), sponsored by the Open Society Institute (now the Open Society Foundations), coined the term "open access" (Ocholla and Ocholla, 2016; Rizer and Holley, 2014; Harnad, 2011). According to the BOAI, "By '[OA]' to this literature, we mean its *free availability on the public internet*, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, *without financial, legal, or technical barriers* other than those inseparable from gaining access to the internet itself" (Chan et al., 2002, emphasis added). Indeed, among the many factors that inspired the OA movement was the continuously increasing prices of journal and database subscriptions (Dawson & Yang, 2016; Nguyen, 2008).

Based on the BOAI, the OA movement created an online method for publishing scholarly, peer-reviewed journals with free access to full-text articles (Laasko et al., 2011; Harnad, 2011). Many scholars find that OA publishing promotes scholarly communication through the dissemination and use of research and also generates increased author citations and usage statistics (Dawson & Yang, 2016; Harnad et al., 2008; Björk, 2006; Antelman, 2004). In this way, the OA movement promotes unencumbered access to scholarship by promoting open scholarly communication (Eysenbach, 2006) and creating a comprehensive, efficient system for disseminating research findings (Margaret, 2016).

Routes to Open-Access Publishing in Scholarly Communication. It is typical for OA publications to take one of two route, Gold or Green (Dawson & Yang, 2016; Suber, 2015; see also Clobridge, 2014; Neugebauer and Murray, 2013; Willinsky, 2010; Harnad et al., 2008). The route to Gold OA refers to publishing in journals where the journal is itself OA (Harnad, 2011). Authors must often pay article processing charges to these publishers to publish their works openly, instead of behind the tridiagonal publishing paywall of subscriptions or licensing fees (Dawson & Yang, 2016; Harnad et al., 2008; Suber, 2005). Predatory journals and publishers, however, make it difficult for authors to determine their credibility (Al-Khatib, 2016; Dawson & Yang, 2016). Further, works published under Gold OA lacking proper peer review raise concerns about damage to authors' academic reputations and the increased likelihood of article theft or plagiarism (Yang and Li, 2015).

The route to Green OA, on the other hand, refers to self-archiving in open repositories by the *authors* (Harnad, 2011; Harnad et al., 2008). Green OA can avoid the above concerns with Gold OA because works are published in peer-reviewed journals and their authors have the publisher's permission to place them in repositories or on the authors' websites (Dawson & Yang, 2016; Suber, 2005). According to Harnad (2011), "the fastest and surest road to OA is the green road of OA self-archiving" (p. 88). While achievement of Green OA rests in the hands of the authors and can be mandated, it more directly benefits the interests of the research community, unlike Gold OA, which may rests in the hands of a commercial publishers (Harnad, 2011; Harnad et al., 2008).

Open-Access Movement Trends and Issues

OA publishing has disrupted the traditional subscription model in scholarly communication (Laakso and Bjork, 2012). A 2011 study by Laasko et al. found speedy growth in the OA journal publishing industry between 1993 and 2009. A second study by Laakso and Bjork (2012) of OA journal publishing trends between 2000 and 2011 found that mainly professional organizations or scientific societies published OA articles until 2005, after which commercial publishers significantly began publishing OA articles. Additionally, in 2015, Tenopir et al. found that social



science faculty increasingly seek, read, and use electronic resources for teaching and research in place of traditional print (Tenopir, King, Christian, and Volentine, 2015).

Despite the OA movement's traction in scholarly communication, it has many unresolved, often fraud-related issues, such as the proliferation of predatory journals and publishers whose sole purpose is to collect article processing charges (APCs) (Al-Khatib, 2016). These OA journals that levy APCs exploit unsuspecting authors who may not be able to determine whether these journals are legitimate (Al-Khatib, 2016) or what author's rights they retain after paying the fee (Carroll, 2011; Bloch; 2005). While most OA journals do not charge APCs, of those that do, the average APC of full OA journals is \$660 USD and that of hybrid OA journals is \$2,500 USD (Morrison, 2017). Troublingly, publishing giant Elsevier, one of the world's largest OA publishers in 2016, is considering the switch to charging APCs (Morrison, 2017).

Regardless of the OA movement's expansion over the last decade, its future role in LIS scholarly communication remains uncertain. While Harnad (2011) advocates Green OA publishing over Gold OA, arguing that, "The money to pay for gold OA publishing will only become available if universal green OA eventually makes subscriptions unsustainable. Paying for gold OA pre-emptively today, without first having mandated green OA, not only squanders scarce money, but it delays the attainment of universal OA" (p. 86), Rizer and Holley (2014) assert that in reality, Gold OA has fared better and has more potential for economic stability than Green OA. While commercial publishers have adapted to and even profited from OA, the movement has yet to actually reduce costs for libraries (Rizer and Holley, 2014).

Grey Literature and University Repositories

Copyright and Creative Commons Licensing. The OA movement creates new copyright and licensing issues for libraries and their repositories (Dawson & Yang, 2016). Because libraries could face copyright challenges when the repositories provide OA to full-text research publications, Dawson and Yang (2016) studied current practices that manage copyright permissions at repositories to help others update their own policies. They found that, "In spite of the enthusiasm for open and web-based access, copyright is one of the major deterrents for participation of faculty and students in repositories" (Dawson & Yang, 2016, p. 290). Thus, librarians play a critical role in educating authors and users about copyright and in obtaining copyright permissions for the repository (Dawson & Yang, 2016).

The BOAI explicitly states that OA applies to any "lawful purpose" (Chan et al., 2002) and does not advocate infringement, expropriation, or piracy of research outputs (Suber, 2005). Due to the OA movement's aspiration to provide free, unfettered access to research, copyright law and licensing should play a limited role: "The only constraint on reproduction and distribution, and the *only role for copyright in this domain*, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited" (Chan et al., 2002, emphasis added). Many OA journal author agreements that allow for Green OA require that subsequent deposit into an institutional repository require citation to the original journal. As Suber (2005; 2002) further explains, OA-published works do not inherently infringe United States copyright law because their legal basis comes from the voluntary consent of newer works' copyright holders or from the expiration of older works' copyright. For such older works in the public domain, there is no risk of copyright infringement and no consent from the copyright holder is needed.

Copyright holders may voluntarily consent to OA of newer works via, for example, a Creative Commons license (Suber, 2005; 2002). As Schöpfel and Lipinski (2012) explain: "The Creative Commons licenses, a 'some rights reserved' approach to copyright, applies to all kinds of creative, educational or scientific content created and owned by individuals, companies or institutions. The basic idea is that the creator keeps the copyright while allowing certain uses of his or her work in a standardized way. The condition is that the author ... owns the complete rights" (p. 20). Author consent through a Creative Commons license covers the "unrestricted reading, downloading, copying, sharing, storing, printing, searching, linking, and crawling of the full-text of the work" – essentially, "all the uses required by legitimate scholarship" (Suber, 2005). Copyright holders that distribute their works under a Creative Commons license may nevertheless place certain restrictions on the use of such works, usually to prevent plagiarism, misrepresentation, or commercial re-use (Suber, 2005). Creative Commons is also used by OA publishers as an alternative to journal distribution (Lipinski and Kritikos, 2017) and would likely be an element of Gold OA. While Creative Commons does not provide a repository service, it supports the creation of content that is readily sharable in networked information environments (Lipinski and Copeland, 2013).



Accessing Grey Literature in University Repositories. Repositories support the OA movement by providing access to the scholarly community's research outputs (Willinsky, 2006; Lynch, 2003; Crow, 2002). Deposited research outputs can include grey literature like ETDs, working papers, and datasets (Schöpfel & Lipinski, 2012; Juznic, 2010). Benefits of housing grey literature in a repository include increased usage of research outputs, showcasing the university's scholarship, workflow simplification, and control over metadata and embargoes (Banach, 2011).

When it comes to the effect of the OA movement and repositories on grey literature, a "crucial question will be the inclusion and use of scientific data in documents, and in particular the intellectual property of datasets" (Schöpfel & Lipinski, 2012, p. 4). According to Schöpfel and Lipinski (2012), the legal status of grey literature with regards to copyright depends on the type of research output (e.g., dissertation, thesis, working paper, data set, etc.) and the prevailing policies, laws, and customs of a particular country (e.g., the United States or France).

Because grey literature is published outside traditional commercial publication channels, publisher policies on self-archiving, embargoes, and other licensing aspects would seem to have little effect on grey literature, but the licensing and copyright of this scholarship can still be complex (Schöpfel & Lipinski, 2012, pp. 21-22). Thus, good OA policies at repositories are essential to manage the rights of authors and institutions while enable the access to and use of grey literature.

Open-Access Policy Best Practices

Harvard University provided the gold standard for OA best practices with its seminal guide "Good Practices for University Open-Access Policies" (Harvard Guidelines), a joint project between the Harvard Open Access Project and the Berkman Center for Internet and Society (now the Berkman Klein Center for Internet and Society) (Shieber & Suber, 2017, 2015). This guide established many OA best practices, ranging from drafting, adopting, and implementing good OA policies. They focus on how "an effective OA policy can build support for OA, as an academic and social good, into standard university policies" (Shieber & Suber, 2015, p. 6). Regarding "good practices" for drafting OA policies, considerations include policy goals, author embargoes, scope of research outputs, user rights, and policy implementation. The guide now lives online as a wiki.

The 2008 policy of the Harvard University Faculty of Arts and Sciences demonstrates the attempt to "build support for OA ... into its standard university policies" (Shieber & Suber, 2015, p. 6). This policy requires authors to deposit their works into open repositories as well as to grant non-exclusive copyright licenses to Harvard, allowing the university to archive and distribute the faculty's scholarly works (Priest, 2012, p. 391; see also Nguyen, 2008). Many other U.S. universities followed suit (Priest, 2012, pp. 396-397). As Nguyen (2008) describes, the Harvard OA mandate supports equitable access to research outputs and promotes the university as a steward of the dissemination of research and knowledge.

But not all scholars are enamored with it. For example, Priest (2012) reviewed the Harvard OA mandate and found that its legal effect is uncertain at best, complicated by copyright law factors like the work-made-for-hire doctrine (p. 398), policy overbreadth regarding the non-exclusivity of licenses, and transfer of copyright to journal publishers (p. 418), among others. And while the literature reflects attention to, for example, best practices for OA journal publication agreements (see, e.g., Crews, 2016), there is a gap regarding best practices for drafting and implementing OA policies that promote unfettered access to grey literature in university repositories.

Methodology

This case study explores the impact of the OA phenomenon LIS scholarly communication, namely, the access and use of grey literature in university repositories. Case studies are a useful research design because examining and understanding a distinct phenomenon may illustrate a more general problem (Flick, 2014; Creswell, 2013). To create a sample of universities, the authors cross-referenced the iSchool Directory (iSchools, 2017) with the top twenty-five best LIS programs as ranked by *U.S. News and World Reports* (U.S. News, 2017), resulting in an initial sample of twenty-two iSchools.



Table 1. Top ten U.S. iSchools (highlighted rows indicate final sample).

iSchool	Location	U.S. News 2017 Rank	OA Policy	Repository
University of Illinois at Urbana-Champaign	Champaign-Urbana, IL	1	Yes, 5/14/2016	Illinois Digital Environment for Access to Learning and Scholarship (IDEALS)
University of Washington	Seattle, WA	2	Proposed draft, 6/1/2016	ResearchWorks Archive
University of North Carolina-Chapel Hill	Chapel Hill, NC	3	Yes, 5/13/2016	Carolina Digital Repository
Syracuse University	Syracuse, NY	4	No, info about OA only	Syracuse University Research Facility and Collaborative Environment (SURFACE)
University of Michigan-Ann Arbor	Ann Arbor, MI	5	No, info about OA only	Deep Blue
University of Texas-Austin	Austin, TX	6	Yes, 10/31/2016	Texas ScholarWorks
Rutgers, The State University of New Jersey-New Brunswick	New Brunswick, NJ	7	Yes, 10/29/2014	Scholarly Open Access at Rutgers (SOAR); RUCore
University of Maryland-College Park	College Park, MD	8	No, info about OA only	DRUM
Indiana University-Bloomington	Bloomington, IN	9	Yes, n.d.	IUScholarWorks Repository
University of Pittsburgh	Pittsburgh, PA	10	Proposed draft, 9/13/2013	D-Scholarship@Pitt

All iSchools in the initial sample have open repositories. Only ten have OA policies in place and three have proposed policies not yet adopted. Due to considerations of space and based on the successful methodology used for the authors' recent related study of OA journal publication agreements (Lipinski and Kritikos, 2017), the final sample purposefully contains five iSchools (Table 1). All OA policies from the iSchools in the final sample were available online. After data collection, the authors mapped the similarities and differences in the OA policies based on variables drawn from the Harvard Guidelines (Shieber & Suber, 2017), discussed below.

Findings and Discussion

Reconciling Copyright and Open Access

Before discussing the five OA policies and the Harvard Guidelines, two related questions must be addressed: Does the university have a copyright policy that determines who holds the copyright for the works included in its repository? And if faculty hold the copyright of their works, are they instructed to retain their rights in the publication process? Without the alignment of who holds copyright, deposit into the repository may infringe copyright. When the full text of works protected by copyright are made available in repositories, copyright issues may arise that, depending on the format of the work, implicate the exclusive rights of reproduction and display or performance. Someone holds the copyright to those works and if they are deposited without proper permission, liability may result.

Who holds the copyright in works created by authors covered by an OA policy? All five of the OA policies contain a provision that implies that faculty hold copyright. Illinois states that its policy will not transfer ownership and UT-Austin states that authors retain copyright. Rutgers indicates its OA policy "does not alter ownership status." UNC ("where the faculty member holds the copyright") and Indiana ("you must ... confirm[] that you own the copyright") assume that faculty hold copyright. Each university covers copyright in a separate policy (Table 2). Only Illinois, however, references a separate copyright policy in its OA policy. The authors recommend that if a university's copyright policy indicates the copyright status of faculty works, then the OA policy should reference the copyright policy.



Table 2. University policies indicate faculty hold copyright.

	U-Illinois	UNC	UT-Austin	Rutgers	IU-Bloomington
Intellectual property or copyright policy	<p>Yes: "Unless subject to any of the exceptions specified below or in Section 4(c), creators retain rights to traditional academic copyrightable works as defined in Section 2(b) above."</p> <p>General Rules Concerning University Organization and Procedure, 13 (Jan. 19, 2017).</p>	<p>Yes: "The creator of such a work shall own the work unless ..." and "Traditional Works or Non-Directed Works are pedagogical, scholarly, literary, or aesthetic works resulting from non-directed effort."</p> <p>Copyright Policy of UNC at Chapel Hill, 6 (Jan. 1, 2009).</p>	<p>Yes: "... the Board of Regents will not assert an ownership interest in the copyright of scholarly or educational materials ... related to the author's academic or professional field ..."</p> <p>Board of Regents Rule 90101: Intellectual Property, 4 (May 8, 2017).</p>	<p>Yes: "This policy reaffirms the faculty's rights to retain copyright ownership to the scholarly and artistic works they create ..."</p> <p>Policy Section 50.3.7: Copyright Policy (Jan. 18, 2007).</p>	<p>Yes: "Except as provided in section 2.C.v, the University shall assert no claims to copyright ownership in or to distribution of revenue from Traditional Works of Scholarship."</p> <p>Intellectual Property Policy 3 (May 2, 2008).</p>

While each separate copyright policy examined confirms that faculty hold the copyright in the scholarly work each creates, two policies are located in jurisdictions where case law also supports this position (*Weinstein v. University of Illinois*, 1987, p. 1094; *Hays v. Sony Corporation of America*, 1988, pp. 417-418; *Bosch v. Ball-Kell*, 2006, *7). And in the absence of judicial pronouncements, copyright policies are necessary because the U.S. copyright law states that the employer holds the copyright in works created by employees in the course of their employment: "A 'work made for hire' is a work prepared by an employee within the scope of his or her employment" (17 U.S.C. § 101). Faculty retaining copyright for scholarly work product operates as an exception to the work-for-hire rules. Without case law or a copyright policy to set faculty as the holders of copyright, the university holds the copyright to all faculty work product by default.

This default setting raises another question: Are statements in copyright policies that the university "will not assert" or "shall assert no" (UT-Austin and Indiana), that faculty "retain" (Illinois and Rutgers), or that "creator of such a work shall own the work" (UNC) sufficient to alter the default? Perhaps not. The work-for-hire doctrine in the copyright statute dictates the default: Employers hold the copyright in works that employees create in the course of their employment. While the copyright holder cannot waive copyright, it can transfer copyright to another. According to 17 U.S.C. § 204(a), a "transfer of copyright ownership, other than by operation of law, is not valid unless an instrument of conveyance, or a note or memorandum of the transfer, is in writing and signed by the owner of the rights conveyed or such owner's duly authorized agent."

A university's copyright policy, however well-intentioned, probably does not meet the writing requirements for a valid transfer. Scant case law confirms this conclusion: "There must be a sufficient writing to rebut the presumption that the employer retains the copyright in a work made for hire ... It is the employee's burden to show the existence of a writing granting the employee the copyright in any work made for ... Unwritten understandings or writings not containing the signatures of both parties are insufficient to rebut the presumption" (*Manning v. Board of Trustees*, 2000, p. 980). Transfers of copyright ownership are insufficient.

Assuming that faculty do hold the copyright of their university work product, either by operation of law (statutory or case law) or by valid transfer (policy or legal instrument), are faculty careful when publishing to retain rights sufficient to make the work available in an open repository? Or have they inadvertently signed their rights away via a publication agreement? For example, under Indiana's policy, faculty submitting an item into the repository are required to affirm that the faculty member holds copyright. But are the faculty aware that they may transfer away important rights as part of the publication process? It is typical for universities to warn that some publication agreements divest faculty of their copyright (Table 3).



Table 3. Faculty awareness and retention of copyright in publication.

	U-Illinois	UNC	UT-Austin	Rutgers	IU-Bloomington
Policy or instructions alerting faculty to loss of copyright in publication process	Yes. Example: "Some publishers require you to sign away your rights to your intellectual property in order to have your research published. In such cases, you may lose all control over further reproduction or distribution of your work." http://www.library.illinois.edu/sc/services/scholarly_communications/your_rights.html .	Yes. Example: "Keep Your Copyrights: A Resource for Creators. Designed to help creators hold on to their copyrights and to license their rights on author-friendly terms" http://guides.lib.unc.edu/open-access-and-scholarly-communications/managing-rights	Yes. Example: "Depending on the agreement, you may no longer be able to use your work in future publications or teaching, distribute your work to colleagues, or post your work in an online repository." http://libguides.uta.edu/copyright/authors .	Yes. Example: "Retaining copyright rather than transferring to a publisher may leave the author with more flexibility with respect to future uses, but even if copyright is transferred to a publisher, significant flexibility may be built into the publication agreement ..." https://www.libraries.rutgers.edu/copyright/copyright-academic-research-and-publication .	Yes. Example: "... add to any copyright license (or assignment for scholarly articles) an addendum stating that the agreement is subject to this prior license. That way, you will avoid agreeing to give the publisher rights that are inconsistent with the prior license to IUScholarWorks that permits open-access distribution." https://openscholarship.indiana.edu/policy-faq
Sample language or addendum	Yes: CIC/BTAA or SPARC.	Yes: SPARC.	Yes: SPARC.	Yes: BTAA.	Yes: BTAA.

In addition, the separate copyright policies reviewed recommend the use of an addendum agreement to secure the continued use of the work by faculty and/or the university (Table 3). Two common addendums are SPARC and CIC/BTAA. The SPARC Author Addendum is "a legal instrument that [authors] can use to modify [their] copyright transfer agreements with non-[OA] journal publishers" that allows authors "to select which individual rights out of the bundle of copyrights [they] want to keep, such as: ... Posting the article on a personal or institutional Web site ..." (SPARC, 2017). The CIC/BTAA is the Big Ten Academic Alliance Statement on Publishing Agreements, which includes the Addendum to Publication Agreements for BTAA Authors (BTAA, 2016). The authors recommend that the university's OA webpage, not its separate copyright policy, should include instructions and addendums for authors' rights.

Open-Access Policy Goals and Types

The Harvard Guidelines classifies OA policy into six "types" depending on three variables (Table 4). The first is the disposition of non-exclusive rights (the copyright) in the work. The second is whether deposit in an open repository is required. The final variable is whether, if deposit is mandatory, the OA policy allows faculty to exclude a work via opt-out or waiver. Likewise, if deposit is not mandatory, the policy allows faculty to opt-in.



Table 4. Taxonomy of open-access policy types.

	Type I	Type II	Type III	Type IV	Type V	Type VI
Non-exclusive rights	Granted to university.	Faculty retain (in order that rights can be granted).	No.	No.	No.	Granted to university.
Opt-out/waiver or opt-in for authors	Opt-out.	No.	N/A	N/A	N/A	Opt-in.
Deposit in repository	Required.	Required.	Required (OA or dark).	If publisher permits.	Encouraged.	Not required.

Indiana does not require deposit (i.e., the repository is opt-in but does require faculty to grant a non-exclusive license to the university), which the guidelines classify as Type VI. The other four policies require deposit through a grant of a non-exclusive license to the university and allow a waiver or opt-out, or Type I. These four also allow for an embargo period (Illinois, UT-Austin, and Rutgers for a “specified” time; UNC for a “reasonable period of time”). The embargoed work would still be available in the repository’s metadata, such as in the online catalog. According to the guidelines, the embargo is a “period of dark deposit” (Shieber & Suber, 2015, p. 8) characteristic of Type III. The dominant elements of the four policies remain within Type I, however, with the caveat that there is the possibility for embargo. Rutgers adds that the grant of non-exclusive license is conditioned upon the work not being sold for profit.

Assessment of Open-Access Policy Elements

The framework for assessment of the OA policy elements draws on the variables in the Harvard Guidelines (Table 5). Each policy mentions the repository’s intent or goal. For example, an introductory “whereas” clauses in Illinois expresses that the “Faculty ... committed to disseminating its research and scholarship as widely as possible.” UNC states that reason for its policy is to “disseminate the fruits of its research and scholarship as widely as possible.” Likewise, UT-Austin is “committed to disseminating the fruits of its research and scholarship.” Rutgers (“solely for the purpose of provided and maintaining public access”) and Indiana (“with the goal of providing perpetual access to deposited materials”) focus on the permanency of the repository.

Table 5. Assessment of open-access policy elements.

	U-Illinois	UNC	UT-Austin	Rutgers	IU-Bloom
Date of adoption	October 19, 2015	May 13, 2016	October 31, 2016	October 29, 2014	Not dated
Policy goal / mission statement	Yes: several of the 8 “whereas” clauses expresses intent and purpose.	Yes: “Policy Statement” indicates the reason for grant of license; a later section (“Reason for Policy”) appears to operate as a goals statement	Yes: statement of intent indicates “committed to disseminating the fruits of its research and scholarship”	Yes: statement of intent indicates “solely for the purpose of provided and maintaining public access to them”	Yes: Removal Policy indicates “IUScholar Works has been established as a permanent archive with the goal of providing perpetual access to deposited materials”
Policy type (see Table 1)	Type I	Type I	Type I	Type I	Type VI
Copyright with author	Yes: “This policy does not transfer copyright ownership, which generally remains with Faculty authors under existing University of Illinois General Rules ...”	Implied: “...where the Faculty member holds the copyright under University policy ...”	Yes: “staff members retain their copyright” and “author retains all rights to their work”	Implied?: “Policy does not alter the copyright ownership rights determined in accordance with law or with Rutgers University ... copyright policies”	Required: “you must agree to the ... license, which confirms that you own the copyright to the items”



Includes student research	Not explicit, but theses included via General Rules.	Policy does not indicate.	Policy does not indicate.	Yes: "as well as Rutgers graduate and postdoctoral students, while they are employed or enrolled at, Rutgers"	Yes: "dissertation writers" and "students with authorization from a sponsoring department or faculty member"
Grants university non-exclusive rights Rights are assignable	Yes. Assignable: "and authorize others to do the same"	Yes. Policy does not indicate assignability.	Yes. Assignable: "and authorize others to do the same" for non-commercial "educational, research and personal"	Yes: "provided that the articles are not sold for a profit" Policy does not indicate assignability.	Yes: "grants Indiana University permission" Assignability implied?: "grants ... permission to distribute the items worldwide ... and ... to preserve"
Requires deposit in repository	Yes: "in an [OA] repository"	Yes: "for the purposes of making those articles freely and widely available in an [OA] repository"	Not explicit: "make available [] online"	Not explicit: "grants to ... in any medium ..."	No, at option of scholar: "In order to place your work(s) in IUScholar Works..." and supported by Removal Policy
Deposit version: author accepted manuscript version preferred / final version only if same rights	Yes: "final author's version post peer review or the final published version."	Yes: "ordinarily the author's final edited version"	Yes: "author's final version" and "final revised version ...after peer-review comments have been incorporated.	Yes: "author's final version" and "the version that follows formal peer review...as distinguished from the publisher's branded PDF version"	No: included submitted, accepted and published versions.
Deposit timing: time of journal acceptance preferred	Policy does not indicate.	Policy does not indicate.	Yes: "no later than the date of its publication"	Yes: "no later than the date of its publication"	No: accepts submitted, accepted and published versions.
Author may opt-out / waive policy	Yes: "Upon express direction ... waive application ..."	Yes: "Provost ... will waive application of this policy for a particular article ..."	Yes: "waive application of the license for a particular article" and "functions as an opt-out policy, rather than an opt-in policy"	Yes: "will grant a waiver"	No (deposit is not required so an opt-out or waiver is not necessary).
Allows embargo period / dark deposit	"Yes: "Upon express direction ... or delay access for a specified time"	Yes: "... or delay public access to an article for a reasonable period of time"	Yes: "or delay access for a specified period of time"	Yes: "or delay access for a specified period of time (an 'embargo') upon express direction ..."	Yes: "Embargoes may be set for any period up to five (5) years from the date of deposit"
Includes peer-reviewed articles and conference proceedings	Yes: "scholarly articles"	Yes: scholarly articles are "typically published in scholarly journals"	Yes: "scholarly articles and conference papers"	Yes: "generally refers to peer-reviewed journal articles or conference proceedings created without expectation of payment"	No: "Submitted manuscripts (as sent to journals for peer-review), accepted versions, published versions, supplementary files, including multimedia or datasets, gray literature (conference papers,



					working drafts, primary evidence), dissertations and theses, negative results or work that will not be finished."
Excludes royalty-producing research	Policy does not indicate other than "scholarly articles that fall outside the scope of copyrightable works described in General Rules III, Sections 4a and 4c"	Yes: "do not include classroom pedagogical material or books sold for profit"	Yes: "does this policy apply to books or book chapters? No ... encourages staff to deposit their book chapters, conference posters and other scholarly output into Texas Scholar Works."	Yes: "not include books, commissioned articles, artworks, popular writings, fiction, poetry or pedagogical materials such as lecture notes and videos, case studies and the like"	Yes: "It is not equipped to support the archiving and/or accessibility of dynamic resources like open web sites, interactive applications, files with complex metadata requirements, authoring tools, or dynamic learning objects."
Excludes research restricted by law (contract or otherwise)	Yes: "any articles for which the Faculty member entered into an incompatible licensing agreement before the adoption of this policy"	Yes: "or subject to a conflicting agreement formed before the adoption of this Policy"	Policy does not indicate.	Yes: "any articles for which the scholar entered into an incompatible licensing or assignment agreement before the adoption of this policy"	Yes: items removed in accordance with "Journal publishers' requirements", or in instances of copyright infringement or plagiarism, libel or invasion of privacy, falsified research.
Rights granted and deposits required are not retroactive	Yes: "outside the scope ... any articles published before adoption of this policy"	Yes: "except for articles authored or co-authored before adoption of this policy"	Yes: "except for any works completed before the adoption of this policy"	Yes: "except for articles completed before the adoption of this Policy"	Policy does not indicate.
Rights granted to university are transferable, i.e., assignable	Implied: "and to authorize others to do the same"	Policy does not indicate.	Implied: "and authorize others to do the same"	Policy does not indicate.	Policy does not indicate.
Allows but does not require open licenses	Yes: OA repository or link to publisher website.	Policy does not indicate.	Policy does not indicate.	Policy does not indicate.	Suggested: "Authors may also consider licensing their works with a Creative Commons License."
Assigns responsibility	Yes: "Campus Senate and Office of the Provost"	Yes: "Scholarly Communications Offices of the University Library or other Office designated by the Provost."	Yes: University Library Director or designee "Scholarly Communications Librarian serves as the Director's designee"	Yes: "Executive Vice President for Academic Affairs (or designee)"	None indicated but, "IU Libraries and Indiana University retain the rights to withdraw any item ... deem such action necessary"
Green OA or Gold OA	Green.	Green.	Green.	Green.	Green.

Works Included in Repositories. As for the types of research that the OA policies include in the university repositories, all but Indiana restrict content to scholarly journal publications (Illinois: "scholarly articles"; UNC: "typically published in scholarly" journals) and, in two instances, to proceedings (UT-Austin: "scholarly articles and conference papers"; Rutgers: "peer-reviewed journal articles or conference proceedings"). As Indiana collects a wide range of



materials, there is no restriction limiting submissions to scholarly journals or proceedings. Indiana does, however, exclude the “archiving and/or accessibility of dynamic resources like open web sites, interactive applications, files with complex metadata requirements, authoring tools, or dynamic learning objects.” Illinois excludes “scholarly articles that fall outside the scope of copyrightable works” per section III.4. of its General Rules policy.

UNC does “not include classroom pedagogical material or books sold for profit,” but the Q&A following its policy contains unclear statements. In reply to “What kinds of scholarship are covered under the policy, it states that replies submissions are limited to “scholarly articles.” But it replies affirmatively to two other questions: “I’ve created scholarly material other than journal articles. May I deposit it in the CDR?” and “May I upload supplementary material that goes with my article?” The response to the latter question notes that “supplementary material, such as additional figures, datasets, or video” can be included and adds that “if you have large files over 500 MB, contact us,” which may have more to do with file size than with content. The context of the statement implies that the supplementary material is related to the article itself. Additionally, the Q&A states that, “the CDR will handle and preserve a wide variety of formats and file types,” but this may not indicate different kinds of content.

While UT-Austin applies to “scholarly articles and conference papers,” it nonetheless encourages faculty to deposit their “book chapters, conference posters and other scholarly output into Texas Scholar Works.” Rutgers excludes “books, commissioned articles, artworks, popular writings, fiction, poetry or pedagogical materials such as lecture notes and videos, case studies and the like.” Other than Indiana, there does not appear to be attention to grey literature in the OA policies, though the Illinois and Rutgers policies imply that their repositories accept ETDs. Grey literature in repositories is discussed at length below.

Works Excluded from Repositories. All OA policies exclude research published before the effective date of the policy. Illinois excludes “articles published before the adoption of this policy.” UNC takes a slightly broader formulation that focuses on the date of authorship: “except for articles authored or co-authored before adoption of this Policy.” UT-Austin and Rutgers, respectively, take a similar approach and use near-identical language: “except for any works completed before adoption of this policy” and “except for any articles completed before the adoption of this policy.” Indiana does not indicate a date-based restriction, likely because it reflects an intention to collect anything representing the “research, scholarship and intellectual output of the Indiana University community” regardless of when the work was published or first created or authored.

Using almost identical language, Illinois, UNC, and Rutgers indicate exclusion where there is an “incompatible” or “conflicting” agreement entered into before the date of OA policy adoption. In other words, the author signed a publication agreement that transferred the exclusive rights to the work to the publisher or limited the author’s exclusive rights. UT-Austin is silent on this issue, and Indiana will remove articles from its repository in accordance with “Journal publishers’ requirements.” The Removal Policy section of IUScholarWorks indicates content that “files will be removed ... only under extraordinary circumstances.” In addition to publisher requirements, it will remove files that infringe copyright, contain falsified research or libelous material, or are plagiarizing or an invasion of privacy.

Transfer and Assignability of Author Rights. The Harvard Guidelines suggest that an institution transfer rights back to the authors, but only Indiana and UT-Austin include such language (“authorize others to do the same”). The purpose is “to allow the institution to grant rights back to the author. The effect is that authors retain or regain certain rights to their work, including rights that they might have transferred away in their publishing contracts” (Shieber & Suber, 2015, p. 14). But rights transferred in a valid publication agreement (contract) between author and publisher cannot not be returned (or “regained,” to use the Guide’s terminology) by a third party not privy to the original agreement. Such a return of rights would only be possible where the author granted the publisher non-exclusive rights. If the author granted exclusive rights to the publisher, then the author would have no rights remaining to grant to a third party (university) that could then be granted back to the author.

Assuming that copyright resides with the faculty who grants non-exclusive right to use the work to the university, does this right include assigning rights to another? Illinois and UT-Austin require the assignability of the non-exclusive use right. UT-Austin restricts assignment to non-commercial “educational, research and personal” uses. UNC and Rutgers do not indicate assignability, and Indiana implies a right of assignment: “grants ... permission to distribute the items worldwide ... and ... to preserve.”



Waiver or Opt-Out of Open-Access Policy. Allowing faculty to waive or opt-out of an OA policy does, to an extent, defeat the goal of open repositories. Opting out would exclude works published, authored, or created after the date of OA policy adoption. Unless impractical to discipline, the authors recommend that universities require by policy that faculty publish only in journals that allow, even with an embargo, open repository submission. This mandate might be conditioned on the faculty using one of the suggested addendums like SPARC or CIC/BTAA. In contract, Indiana allows authors to opt in to its OA policy because its repository accepts all types of research for deposit.

Applicable Authors: Faculty, Students, and Staff? All OA policies use the word “faculty.” Illinois applies to works authored “while the creator is a member of the faculty,” as does UNC (“Each Faculty member”). Rutgers includes “faculty” as well as employed or enrolled “graduate and postdoctoral students.” UT-Austin more broadly includes “Each staff member.” Indiana, the more inclusive, includes works authored or published by any “research unit, institute, center, department, or university partner” as well as faculty, staff, dissertators, and students if authorized by a “sponsoring department or faculty member.”

The Harvard Guidelines do not address whether an OA policy should apply to students. Several OA policies, however, specifically include student work. Illinois includes “theses” in its General Rules, and the website for the Illinois IDEALS repository allows deposit of the “research and scholarship” of graduate students. Rutgers’ OA policy applies to “graduate and postdoctoral students” if enrolled or employed by the university. The website for the Rutgers RUCore repository also references “scholarly articles deposited by ... doctoral students and postdoctoral scholars” and “dissertations and theses,” which includes works by masters students. UNC’s Carolina Digital Repository also includes “electronic theses and dissertations,” and the TexasScholarWorks repository at UT-Austin references “electronic theses and dissertations” (Table 6). Indiana does not include theses but does include “dissertation writers.” Oddly, the broader IUScholarWorks repository website states that all “graduate students may submit their thesis or dissertation.” Moreover, Indiana’s repository includes a broader array of student work when authorized by a “sponsoring department or faculty member.”

Version and Timing of Scholarly Submissions. Illinois, UNC, UT-Austin, and Rutgers accept only the post-peer-review, final pre-publication version of a work for deposit in the repository. Indiana accepts the submitted, accepted, and published versions of articles. The versions submitted are thus unpublished, rather than a PDF or link to the final published article. Most OA journals that allow repository submission, however, require citation to the original journal publication (Lipinski and Kritikos, 2017). The citation information would not be on the final pre-publication version, so adding citation information to the metadata and institutional repository copy is necessary. Furthermore, if Gold OA is desired, policy or practice need to accommodate the addition of citation information to repository submissions. Only UT-Austin and Rutgers indicate the timing of the submission to the repository (“no later than the date of its publication”).

Responsibility for Open-Access Policy. The final element discussed in the Harvard Guidelines is whether the policy indicates responsibility for administering the OA policy. Illinois (Provost and Faculty Senate) and Rutgers (Vice-President) place responsibility at a very high level, signifying OA’s importance on campus. UNC and UT-Austin both logically place the responsibility with the Scholarly Communications Office or librarian. The Indiana policy is not explicit, but since the “IU Libraries and Indiana University retain the rights to withdraw any item ... deem such action necessary,” it is logical that the responsibility for administering the OA policy also resides in the campus library.

Accessing Grey Literature in University Repositories

As mentioned above, except for Indiana, the OA policies do not specifically refer to grey literature. Illinois applies to “scholarly articles,” which could include grey literature such as unpublished papers; however, the policy further states that the articles submitted to the repository should be the “final version of the article (i.e., the final author’s version post peer-review or the final published version where possible).” Elsewhere, the policy indicates that if submission is not possible, “a Faculty member may instead notify the University of Illinois that the article will be ... made available via a link to public access versions of those articles on publisher websites.”

Likewise, UNC refers to “scholarly articles,” defined as “articles that are typically published in scholarly journals where the Faculty member holds the copyright under University policy and where members of the Faculty are authors or coauthors.” Again, the focus is on the published



version. Oddly, the purpose of UNC's policy is "to disseminate the fruits of its research and scholarship as widely as possible," which suggests a broader range of works (not all research and scholarship is published). It also suggests that the repository only includes the final versions of articles accepted for publication (not working papers). As some disciplines rely on grey materials, it does not appear that Illinois and UNC repositories are reliable sources of grey literature.

UT-Austin states a "commitment to disseminate the fruits of its research and scholarship as widely as possible." The TexasScholarWorks repository includes "scholarly articles" and "conference papers—also called conference proceedings," though the repository submission should be the "final version of each work" or the "final version of their article or conference paper (in PDF format)" – earlier versions are not accepted.

Only Indiana specially mentions grey literature and accepts a broad array of submissions to the IUScholarWorks repository, which allows for variation across disciplines as "academic units decide what content to put into their respective communities." Specifically, the OA policy indicates that the repository can include "[g]rey literature (conference papers, working drafts, primary evidence)," as well as earlier versions of articles ("[s]ubmitted manuscripts (as sent to journals for peer-review)"); incomplete works ("[n]egative results or work that will not be finished"); "[d]issertations and theses" (see also Schöpfel & Prost, 2013; Schöpfel & Lipinski, 2012); and "[s]upplementary files, including multimedia or datasets."

Notably, the Revisions Policy for IUScholarWorks further suggests that revisions of works can be included ("[r]evisions to submitted files may only be made to correct typographical, grammatical and spelling errors"). If there needs to be "substantial revisions of findings, facts, etc." and "authors have substantially reworked the content and wish to make a newer version available, they are encouraged to submit the new or revised version as a new item." Thus the repository may house many iterations of a single work, an important consideration for the work's historical record of development.

As discussed above, three OA policies include theses (Illinois) and dissertations (Indiana and Rutgers' graduate and postdoctoral students). Examining the repositories' websites shows that all actually house ETDs and other types of grey literature (Table 6).

Table 6. Treatment of grey literature in university repositories.

	U-Illinois	UNC	UT-Austin	Rutgers	IU-Bloom
ETDs	<p>IDEALS: "... graduate students can deposit their research and scholarship—unpublished and, in many cases, published—directly into IDEALS."</p> <p>Link: https://www.ideals.illinois.edu/</p>	<p>Carolina Digital Repository: "... Electronic Theses and Dissertations ..."</p> <p>http://blogs.lib.unc.edu/cdr/index.php/about/policies-guidelines/</p>	<p>Texas ScholarWorks: "UT Electronic Theses and Dissertations. Content: Electronic theses and dissertations (ETD)"</p> <p>https://repositories.lib.utexas.edu/pages/policies_collections</p>	<p>RUcore: "dissertations and theses ... are ... added to RUCore ..."</p> <p>https://rucore.libraries.rutgers.edu/etd/</p> <p>Scholarly Open Access at Rutgers: "SOAR gathers, and makes available globally via the internet, scholarly articles deposited by ... doctoral students, and postdoctoral scholars."</p> <p>http://soar.libraries.rutgers.edu/about-soar</p>	<p>IUScholarWorks Repository: "graduate students may submit their thesis or dissertation to IUScholarWorks ..."</p> <p>https://scholarworks.iu.edu/deposit#etd</p>



Datasets	<p>IDEALS: “distribute ... other research material.”</p> <p>https://www.ideals.illinois.edu/</p>	<p>Carolina Digital Repository: “... Datasets ...”</p> <p>http://blogs.lib.unc.edu/cdr/index.php/about/policies-guidelines/</p>	<p>Texas ScholarWorks: “UT Faculty / Researcher Work. Content: ... collections of digitized data ...”</p> <p>https://repositories.lib.utexas.edu/pages/policies_collections</p>	<p>RUcore: “RUcore... will include: ... data sets ...”</p> <p>https://rucore.libraries.rutgers.edu/about/include.php</p>	<p>IUScholarWorks Repository: “individual research files (... data set ...)”</p> <p>https://scholarworks.iu.edu/deposit#research</p> <p>Knowledge Base: “primary evidence”</p> <p>https://kb.iu.edu/d/aujg</p>
Working papers	<p>IDEALS: “distribute their working papers ...”</p> <p>https://www.ideals.illinois.edu/</p>	<p>Carolina Digital Repository: “... Unpublished Scholarly Works ...”</p> <p>http://blogs.lib.unc.edu/cdr/index.php/about/policies-guidelines/</p>	<p>Texas ScholarWorks: “UT Faculty / Researcher Work. Content: ... white papers ...”</p> <p>https://repositories.lib.utexas.edu/pages/policies_collections</p>	<p>Scholarly Open Access at Rutgers: “‘scholarly articles’ generally refers to ... working papers ...”</p> <p>http://soar.libraries.rutgers.edu/about-soar</p>	<p>Knowledge Base: “working papers”</p> <p>https://kb.iu.edu/d/aujg</p>
Other outputs	<p>IDEALS: “distribute ... technical reports, or other research material.”</p> <p>https://www.ideals.illinois.edu/</p>	<p>Carolina Digital Repository: “... Research Materials Learning Materials Institutional Records ... Audio or Visual Files”</p> <p>http://blogs.lib.unc.edu/cdr/index.php/about/policies-guidelines/</p>	<p>Texas ScholarWorks: “UT Faculty / Researcher Work Content: Peer-reviewed pre-print articles ... technical reports ... presentations ... field notes, etc.”</p> <p>https://repositories.lib.utexas.edu/pages/policies_collections</p>	<p>RUcore: “RUcore ... will include: ... photographs ... digital video and audio, preprints, technical reports, grant reports, etc. ...”</p> <p>https://rucore.libraries.rutgers.edu/about/include.php</p>	<p>IUScholarWorks Repository: “thematic collection (i.e. workshop series presentations, lab publications, entire conference proceedings, etc.)”</p> <p>https://scholarworks.iu.edu/deposit#research</p>

Table 6 indicates that the repositories collect a wide array of grey literature related to underlying research, such as “primary evidence” (Indiana), “other research material” (Illinois), “works of a scholarly nature” (UNC), datasets (UNC, Rutgers, Indiana), “collections of digitized data” (UT-Austin), other “research materials” (Illinois, UNC), “field notes” (UT-Austin), and lab publications (Indiana). Additionally, most repositories house works falling outside mainstream serial publications, such as working papers (Illinois, Rutgers, Indiana), white papers (UT-Austin), technical reports (Illinois, UT-Austin, Rutgers), and grant reports (Rutgers). Some repositories also include broader educational materials and other formats, such as “learning materials” (UNC), “audio or visual files” (UNC), “digital video and audio” (Rutgers), photographs (Rutgers), pre-print articles (UT-Austin), presentations (UT-Austin), “workshop series presentations (Indiana), entire conference proceedings (Indiana), and even “institutional records (UNC).

While free dissemination and access to all scholarship is the goal of the OA movement, the OA policies examined in this case often restrict deposits to the final versions of published scholarly articles, perhaps because both the Harvard model OA policy (Harvard OSC, 2015) and Harvard Guidelines refer to the “final version” of “the accepted author manuscript” (Shieber & Suber, 2015, p.10). Several OA policies do include student works, but ETDs form only a small part of the grey taxonomy. In contrast, the websites for the general institutional repositories do include a broad array of grey literature.



Recommendations and Conclusion

An OA policy should apply to all persons within the university community engaging in scholarship, not just to faculty. Many OA policies also restrict submissions to university repositories to the final versions of works published in scholarly publications, though some include conference proceedings. These policies also restrict submissions to works created by faculty or, in some cases, doctoral students, but in some disciplines and institutions, graduate and even undergraduate students engage in publishable research (see, e.g., UWM, 2017).

The university repositories themselves, however, accept a broader range of works and accommodate grey literature, providing access to robust and diverse collections of scholarship published outside of traditional channels. Often a separate repository or policy relates to ETDs. The authors recommend that OA policies should refer to the repositories and attendant policies. Further, OA policies should refer to a university's copyright policy where it indicates the copyright status of works created by faculty or students. The authors recommend that OA policies include copyright policy information on the retention of authors' rights sufficient to allow inclusion in an open repository, referring to or providing examples of addendums like SPARC or CIC/BTAA. Depending on the circumstances, a university could require that faculty publish only in journals that allow, even with an embargo period, for OA repository inclusion.

The access and use of grey literature in university repositories play a vital role in LIS education, research, and scholarship. Alongside a review of the literature on the OA movement, this case study mapped the OA policies in current use in at five U.S. iSchools against variables drawn from the Harvard Guidelines. Analysis of the sampled agreements reveals best practices for OA policies that balance copyright with unfettered access to grey literature.

References

- Al-Khatib, A. (2016). [Protecting authors from predatory journals and publishers](#). *Publishing Research Quarterly*, 32(4), 281-285. doi:10.1007/s12109-016-9474-3.
- Antelman, K. (2004). Do open-access articles have a greater research impact? *College and Research Libraries*, 65(5), 372-382. doi:10.5860/crl.65.5.372.
- Armbruster, C. (2008). Cyberscience and the knowledge based economy. Open access and trade publishing: from contradiction to compatibility with non-exclusive copyright licensing. *Policy Futures in Education*, 6(4), 439-452. doi:10.2304/pfie.2008.6.4.439.
- Banach, M. (2011). *The benefits of managing and publishing ETDs "in house" using an open access repository*. Presentation at United States Electronic Thesis and Dissertation Association (USETDA) Conference 2011, Orlando, Florida, May 18-20, 2011.
- Bernard, H. R. (2013). *Social research methods: Qualitative and quantitative approaches*. Thousand Oaks, CA: Sage Publications.
- Big Ten Academic Alliance [BTAA] (2016). Statement on Publishing Agreements and Addendum to Publication Agreements for BTAA Authors. Retrieved from <http://www.btaa.org/docs/default-source/library/authorsrights.pdf?sfvrsn=>.
- Björk, B.-C. (2006). Open access: Maximizing research impact in the Internet age. *Journal of Computing in Civil Engineering*, 20(4), 225-226. doi:http://dx.doi.org/10.1061/(ASCE)0887-3801(2006)20:4(225)#sthash.svoJbz1w.dpuf.
- Bloch, G. (2005). Transformation in publishing: Modeling the effect of new media. *Berkeley Technology Law Journal*, 20(1), 647-671. doi:10.15779/Z387T2C.
- Bohannon, J. (2013). Who's afraid of peer review? *Science*, 342(6154), 60-65. doi:10.1126/science.342.6154.60.
- Bosch v. Ball-Kell*, 2006 WL 2548053 (C.D. Ill. 2006).
- Bowen, G. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27-40. Retrieved from <http://connection.ebscohost.com/c/articles/47652758/document-analysis-as-qualitative-research-method>.
- Carroll, M. W. (2011). Why full open access matters. *PLoS Biology*, 9(11), e1001210, 1-3. doi:10.1371/journal.pbio.1001210.
- Centivany, A. (2011). Paper tigers: Rethinking the relationship between copyright and scholarly publishing. *Michigan Telecommunications and Technology Law Review*, 17(2), 385-416. Retrieved from <http://repository.law.umich.edu/mttlr/vol17/iss2/2>.
- Chan, L., et al. (2002). *Budapest Open Access Initiative*. Retrieved from <http://www.budapestopenaccessinitiative.org/read>.
- Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches*. 3rd edition. Thousand Oaks, CA: Sage Publications.
- Crews, K. D. (2016). Principles of the JCEL publication agreement: Stakeholders, copyright, and policy positions. *Journal of Copyright in Education and Librarianship*, 1(1), 1-7. doi:10.17161/jcel.v1i1.5913.
- Crow, R. (2002). The case for institutional repositories: A SPARC position paper. *ARL Bimonthly Report* 223. SPARC. Retrieved from https://uta-ir.tdl.org/uta-ir/bitstream/handle/10106/24350/Case%20for%20IRs_SPARC.pdf?sequence=1.
- Dawson, P. H., & Yang, S. Q. (2016). Institutional repositories, open access and copyright: What are the practices and implications? *Science & Technology Libraries*, 35(4), 279-294.



- Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology*, 4(5), e157, 0692-0698. doi:10.1371/journal.pbio.0040157.
- Farace, D. J. & Schöpfel, J. (2010). *Grey literature in library and information studies*. Berlin: DeGruyter Saur.
- Flick, U. (2014). *An introduction to qualitative research*. 5th edition. Thousand Oaks, CA: Sage Publications.
- Harnad, S. (2011). Gold open access publishing must not be allowed to retard the progress of green open access self-archiving. *Logos*, 21(3-4), 86-93. doi:10.1163/095796511X559972.
- Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C., & Hilf, E. R. (2008). The access/impact problem and the green and gold roads to open access: An update. *Serials Review*, 34(1), 36-40. doi:10.1016/j.serrev.2007.12.005.
- Harvard Office for Scholarly Communication [Harvard OSC] (2015). Model open access policy. *Harvard Library Office for Scholarly Communication*. Retrieved from <https://osc.hul.harvard.edu/modelpolicy/>.
- Hays v. Sony Corporation of America*, 847 F.2d 412 (7th Cir. 1988).
- iSchools (2017). North American Directory. *iSchools.org*. Retrieved from <http://ischools.org/regions/north-american-ischools/north-american-directory/>.
- Juznic, P. (2010). Grey literature produced and published by universities: A case for ETDs. In Farace, D. J. & Schöpfel, J., eds., *Grey literature in library and information studies* (pp. 39-51). Berlin: DeGruyter Saur.
- Laakso, M., & Björk, B.-C. (2012). Anatomy of open access publishing: A study of longitudinal development and internal structure. *BMC Medicine*, 10(125), 1-9. doi:10.1186/1741-7015-10-124.
- Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C., & Hedlund, T. (2011). The development of open access journal publishing from 1993 to 2009. *PLoS ONE*, 6(6), e20961, 1-10. doi:10.1371/journal.pone.0020961.
- Lipinski, T. A. & Copeland, A. J. (2013). Look before you license: The use of public sharing websites in building patron initiated public library repositories. *Preservation, Digital Technology, and Culture*, 42(4), 174-198. doi:10.1515/pdtc-2013-0028.
- Lipinski, T. A. & Kritikos, K. C. (2017). *Legal and policy implications of licenses between LIS open access journal publishers and authors: A qualitative case study*. Presentation at the 9th Qualitative and Quantitative Methods in Libraries International Conference (QQML2017), May 23-26, 2017, Limerick, Ireland.
- Lynch, C. A. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *Libraries and the Academy*, 3(2), 327-336. doi:10.1353/pla.2003.0039.
- Manning v. Board of Trustees of Community College District No. 505 (Parkland College)*, 109 F.Supp.2d 976 (C.D. Ill. 2000).
- Margaret, R. (2016). *An expanded approach to evaluating open access journals*. *Journal of Scholarly Publishing*, 47(4), 307-327. doi:10.3138/jsp.47.4.307.
- Maxwell, J. A. (2012). "What will you actually do?" (Chapter 5). In *Qualitative research design: An interactive approach*. Thousand Oaks, CA: Sage Publications.
- Morrison, H. (2017). *Elsevier: Among the world's largest open access publishers as of 2016*. *Charleston Advisor*, 18(3), 53-59. doi:10.5260/chara.18.3.53.
- Nguyen, T. (2008). *Open doors and open minds: What faculty authors can do to ensure open access to their work through their institution*. SPARC/Science Commons White Paper, 1-15. Retrieved from https://sparcopen.org/wp-content/uploads/2016/01/opendoors_v1_0.pdf.
- Ocholla, D. N. & Ocholla, L. (2016). *Does open access prevent plagiarism in higher education?* *African Journal of Library, Archives & Information Science*, 26(2), 187-200.
- Priest, E. (2012). Copyright and the Harvard open access mandate. *Northwestern Journal of Technology and Intellectual Property*, 10(7), 377-440. Retrieved from <http://scholarlycommons.law.northwestern.edu/njtip/vol10/iss7/1>.
- Rizor, S. L., & Holley, R. P. (2014). *Open access goals revisited: How green and gold open access are meeting (or not) their original goals*. *Journal of Scholarly Publishing*, 45(4), 321-335. doi:10.3138/jsp.45.4.01.
- Rucinski, T. L. (2015). The elephant in the room: Toward a definition of grey legal literature. *Law Library Journal*, 107(4), 543-559. Retrieved from <http://www.aallnet.org/mm/Publications/llj/LLJ-Archives/Vol-107/no-4/2015-26.pdf>.
- Schöpfel, J. & Lipinski, T. A. (2012). Legal aspects of grey literature. *The Grey Journal*, 8(3), 137-153.
- 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), 65-86. Retrieved from <http://www.essachess.com/index.php/jcs/article/view/214>.
- Schwartz, M. (2012, Apr. 13). Directory of open access books goes live. *Library Journal*. Retrieved from <http://lj.libraryjournal.com/2012/04/academic-libraries/directory-of-open-access-books-goes-live/#>.
- Scimago Lab (2017). Library and information science journal rank. *Scimagojr*. Retrieved from <http://www.scimagojr.com/journalrank.php?category=3309>.
- Solomon, D. J. (2008). *Developing open access to journals: A practical guide*. Oxford: Chandos Publishing. Retrieved from <https://books.google.com/books?isbn=1780632150>.
- Steele, C. (2014). *Scholarly communication, scholarly publishing and university libraries. Plus ça change?* *Australian Academic & Research Libraries*, 45(4), 241-261. doi:10.1080/00048623.2014.950042.
- Suber, P. (2015). Open access overview: Focusing on open access to peer-reviewed research articles and their preprints. *Earlham.edu*. Retrieved from <http://legacy.earlham.edu/~peters/fos/overview.htm>.
- Suber, P. (2012). *Open access*. Cambridge, MA: MIT Press. Retrieved from bit.ly/oa-book.
- Shieber, S. & Suber, P., eds. (2017). *Good practices for university open-access policies* (wiki). Harvard Open Access Project. Retrieved from bit.ly/goodoa.
- Shieber, S. & Suber, P., eds. (2015). *Good practices for university open-access policies*. Harvard Open Access Project. Retrieved from <https://cyber.harvard.edu/hoap/sites/hoap/images/Bestpracticesguide-2015.pdf>.
- SPARC (2017). SPARC Author Addendum. Retrieved from <https://sparcopen.org/our-work/author-rights/#addendum>.



- Tenopir, C., King, D. W., Christian, L., & Volentine, R. (2015). Scholarly article seeking, reading, and use: A continuing evolution from print to electronic in the sciences and social sciences. *Learned Publishing*, 28(2), 93-105.
- Willinsky, J. (2006). *The access principle: The case for open access to research and scholarship*. Cambridge, MA: Massachusetts Institute of Technology. Retrieved from https://kuramoto.files.wordpress.com/2008/09/theaccessprinciple_themitpress_0262232421.pdf.
- Yang, Z. L., & Li, Y. (2015). University faculty awareness and attitudes towards open access publishing and the institutional repository: A case study. *Journal of Librarianship and Scholarly Communication*, 3(1), eP1210. doi:10.7710/2162-3309.1210.
- U.S. Copyright Act, 17 U.S.C. § 101. Definitions.
- U.S. Copyright Act, 17 U.S.C. § 204(a). Execution of transfers of copyright ownership.
- U.S. News and World Report [U.S. News] (2017). Best grad schools rankings: Library and information studies. *USNews.com*. Retrieved from <https://www.usnews.com/best-graduate-schools/top-library-informa>
- University of Wisconsin Milwaukee [UWM] (2017). Office of Undergraduate Research. *UWM.edu*. Retrieved from <http://uwm.edu/our>.
- Weinstein v. University of Illinois*, 811 F.2d 1091 (7th Cir. 1987).

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:

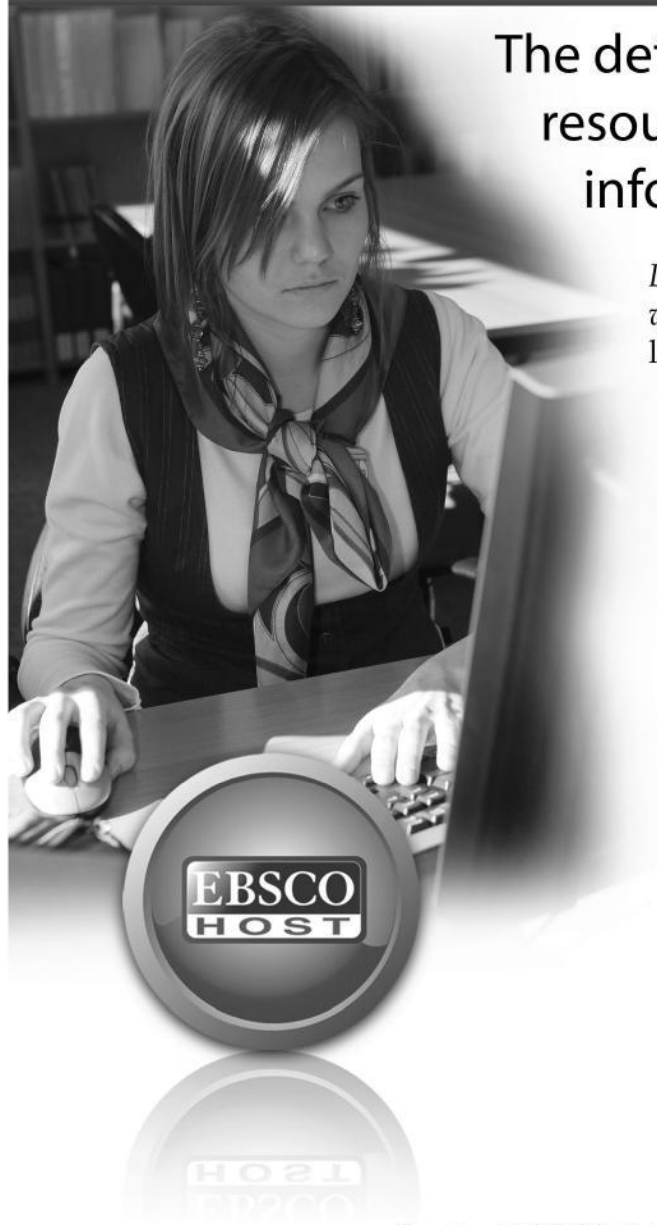
- Bibliometrics
- Cataloging
- Classification
- Information Management
- Librarianship
- Online Information Retrieval
- And much more...

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





Law, Liability, and Grey Literature: Resolving Issues of Law and Compliance

Daniel C. Mack, University of Maryland, United States

Abstract:

Grey literature faces a number of unique challenges when facing issues of liability and compliance with local, national, and international law. Conference and symposium proceedings, white papers, reports, newsletters, and other forms of grey literature often do not have the same legal support as journals, monographs, and other, more formal publications. This can create problems of compliance and liability. Grey publications often include works authored, edited, translated, compiled, or otherwise modified by many people affiliated with multiple institutions, sometimes in several countries. In the process of producing and distributing grey literature, its producers may be confused about, or even unaware of, legal issues such as copyright ownership, authors' and editors' rights, official affiliation, licensing, and distribution. As a result, grey publications may often be at odds with local, national, and international laws and regulations. This in turn may have the unfortunate and unintended consequence of making the producers of grey literature legally liable for damages. This paper proposes a model for identifying potential legal problems, obtaining proper legal counsel, and limiting liability for authors, editors, and distributors of grey literature. Like other publications, grey literature exists within multiple systems of law and regulation. These systems may include institutional policies and regulations, laws at every level, and the formal and informal networks of researchers, authors, editors, distributors, and others affiliated with grey publications. This paper presents a model based on systems theory to approach the problem. The model offers practical means by which producers of grey literature can identify and address these issues before they become problems, and thus minimize noncompliance and avoid liability. The resources necessary for implementing this model are usually minimal. The primary expense may be the use of human resources affiliated with sponsoring institutions. This may include administrators, legal counsel, and other personnel not directly involved with the production of grey literature. Other resources, such as Creative Commons licenses, are available at no cost. The result of implementing the model will be publication of grey literature that is in compliance with local, national, and international law, as well as with institutional policies and regulations. This will minimize or eliminate liability for violation of such laws and regulations by researchers.

Compliance Challenges to Grey Literature

Despite, or perhaps because of, its prevalence, grey literature faces some challenges that differ from those of works published through more traditional venues. One important, but often neglected, challenge to the dissemination of grey literature is the issue of compliance. Because of the nature of both its production and its dissemination, grey literature can face complex issues of compliance due to regulatory protocols, publication type and format, relationships between authors and editors, and concerns regarding licensing and distribution. Researchers whose works appear in grey literature, and distributors of grey literature, often are unfamiliar with the legal and regulatory requirements necessary to remain in compliance with law and policy.

Compliance requires knowledge of a variety of factors, many of which vary with location, type of publication, and institutional affiliation. Local, national, and international law are certainly to most obvious set of regulations influencing compliance. Schöpfel and Lipinski¹ provide an excellent summary of the legal issues surrounding the production and dissemination of grey literature. Law, however, is only one factor for authors and distributors to consider. Institutional policy can have a significant impact on this issue as well. Many universities, research centers, and other institutions have requirements for researchers, authors, and distributors. For example, an institution may require authors to deposit a copy of any published research in the institutional repository regardless of where or by whom it is published. Standards of various types also influence the compliance of grey literature. These may take the form of International Organization for Standardization (ISO) standards, as well as standards specific to a disciplinary area, institution, country or region, professional organization, or other body.



Laws and other regulations are not the only factor influencing regulatory compliance for grey literature. What we call “grey literature” now exists in multiple formats and types of publications. These include works containing images, sounds, video, and other media, as well as datasets.² In addition, because of the increasing internationalization and interdisciplinarity of research, grey literature involves multiple types of authors. These may include sole, group, and corporate authors, often from multiple institutions, nations, and legal jurisdictions. Some of the other legal and regulatory considerations include copyright and permissions, obtaining official affiliation from an institution or organization, licensing, and distribution. These factors influence each other and form a complex system of relationships among regulatory requirements, researchers and other associated personnel, and the content and format of the information in question (please see Diagram 1).

Diagram 1: Challenges to Compliance



Consequences of noncompliance

Noncompliance with legal and regulatory requirements surrounding grey literature may have several negative consequences for researchers, publishers, and distributors. The most obvious negative impact of noncompliance is being unable to disseminate information legally. As a consequence, personnel and institutions involved could find themselves liable for civil and even criminal penalties. This legal liability has the potential to lead to lawsuits as well as financial and other penalties. Noncompliance with institutional policy may likewise negatively impact researchers, authors, and distributors by placing them at risk of disciplinary action. Finally, a very real negative consequence of noncompliance with legal and regulatory obligations is that such action can lead to mistrust among collaborators. Individual contributors may each have unique institutional requirements. Not accommodating these obligations has the potential to foster hard feelings and mistrust among the researchers involved in a project.

Resources for Compliance

Human Resources

Fortunately, a wide range of resources exist to assist in the creation and dissemination of grey literature in compliance with law and policy. The first system to consider are human resources. This includes of course the researchers, authors, and editors involved with a work. Depending on the content and format, other personnel involved with a work may include those involved with data visualization, website development, video production, and other technological support. These personnel are often an excellent resource not only for their direct roles within a research project, but also as a source of information about the legal and regulatory requirements surrounding their contributions.

Two other groups of personnel can also provide important and useful assistance in creating compliant grey literature: institutional administrators and legal counsel. Administrators are not only familiar with policy, but also are usually able to interpret or even create policy. In this role, such administrators can be extremely useful to identify the institutional obligations of researchers, authors, editors, and other contributors. If an administrator is unfamiliar with a specific policy, for example any requirement to deposit research in an institutional repository, the administrator is usually at least able to locate any relevant policy and to make a referral to the person or office who can provide the best assistance. Likewise, institutional legal counsel is often available to provide advice on compliance. Many institutions and organizations retain permanent



legal counsel to provide such advice. Depending on the nature of the specific law or regulation, authors and editors may wish to consult with an institutional attorney for specific legal advice.

Advice and other assistance from institutional personnel can be a valuable resource for compliance. This assistance may not incur a cost to the research project, although this is not always the case. Sometimes there may be costs associated with use of human resources. It is best to find this out when first planning a research project. Determining if this assistance will require a cost in advance will permit researchers to calculate the time and cost of consultation, and to include these costs in the budget of a research project. In some cases, costs might be included in a grant application or similar proposal for funding.

Copyright and Standards

Fortunately, many resources for legal compliance are freely and readily available on the Internet. National and international copyright law are the most obvious of these resources. Many jurisdictions provide not only laws and regulations governing copyright and permissions, but also include accompanying information, as well as any required forms. The European Union, the United Kingdom, and the United States are good examples of jurisdictions with considerable relevant content freely available on the Internet. Legal information from other jurisdictions, including copyright law for most countries, is readily available and easily discoverable through internet search engines. In addition to law, another extremely useful international resource for compliance is information regarding standards. The International Organization for Standardization (ISO) offers information relevant to a wide range of scholarly and technological activities (please see Appendix for a list of resources with ULRs).

Creative Commons

Creative Commons offers authors and editors a wide range of pre-written agreements that both allow creators to control how their content is used, as well as permit distributors of content to remain in compliance with copyright law. These free resources provide a simple solution for sharing information according to whatever conditions may be required or desired (please see Appendix for a list of resources with ULRs). Myška presents researchers with a useful overview of the use, benefits, and limits of recent Creative Commons licenses in her 2015 article.³

SPARC

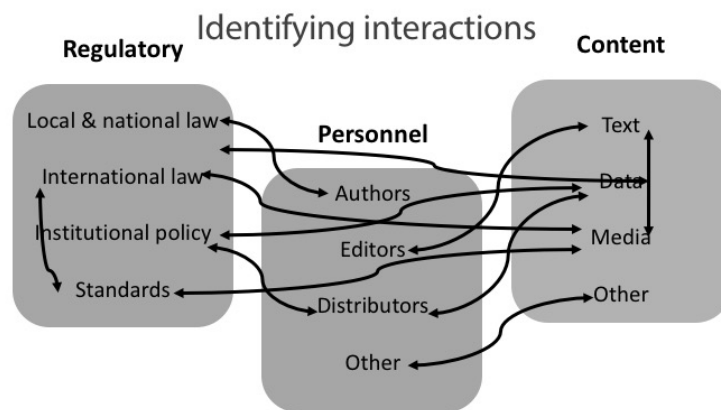
The Scholarly Publishing and Academic Resources Coalition (SPARC) is a key player in the Open Access (OA) movement. SPARC's goal is "to enable the open sharing of research outputs and educational materials in order to democratize access to knowledge, accelerate discovery, and increase the return on our investment in research and education."⁴ In addition to advocacy for OA, SPARC offers valuable free resources for creating and distributing compliant information. The Author Addendum to maintain OA rights is useful for authors whose home institutions require deposit of research in institutional repositories, as well as anyone who wishes to retain OA rights. SPARC also provides resources for alternative PA publishing models that are highly relevant for the dissemination of grey literature (please see Appendix for list of resources with ULRs). Researchers will find that Schöpfel offers a useful analysis of the impact of the OA movement on grey literature.⁵

Creating Compliance: Identifying System Interactions & Solutions

How can researchers best create and distribute grey literature that meets statutory and regulatory requirements? The most effective way of doing this is to identify the interactions of a project, and then to identify the solutions to potential obstacles posed by those interactions. These interactions are most easily visualized as an open system in which elements of multiple subsystems interact (please see Diagram 2).



Diagram 2: System Interactions



This model includes three subsystems. The first is regulatory, including local, national, and international law; institutional and organizational policy; and professional, industrial, and technical standards. The second subsystem consists of personnel involved in a work. This includes authors, editors, and other researchers, as well as personnel involved with distribution, technical support, and consultation. The third subsystem involves the content of a work, including not only its intellectual content, but also its format, including whether the work contains text, media, data, or other material.

This systems model permits us to easily view possible interactions among these various components of the system. Some theoretical examples demonstrate how this model operates. For example, an author (Personnel subsystem) may be affiliated with an institution whose policy (Regulatory subsystem) requires deposit of research in an institutional repository. As another example, a work may be in the form of a video recording (Content subsystem). The distributor of this work may require that it comply with specific technical standards for quality (Regulatory subsystem). A third example involves distribution of a work including data (Content subsystem). If this data was the result of research funded by a national granting agency, there could be a statutory legal requirement that it be made available in an OA repository (Regulatory subsystem). In one last example, a work contains audio material (Content subsystem). Both institutional policy of its supporting organization and national law could require that audio content also include text captions (Regulatory subsystem). The arrows in Diagram 2 demonstrate a few of the many possible interactions that this model can present.

Identifying Compliance Solutions

The systems model of interactions between the Regulatory, Personnel and Content subsystems permits authors and distributors of grey literature with an effective means of identifying both the interactions that require compliance, as well as solutions for achieving compliance. Several of these solutions have already been mentioned. One set of solutions are found in statutory and regulatory law and policy governing copyright, permissions, licensing, and standards; much of this is readily available on the Internet (please see Appendix). Services and projects such as Creative Commons and SPARC offer free, convenient, and easy-to-use solutions for copyright, licensing, permissions, and author agreements. Another group of compliance solutions includes consultation with personnel. This may include not only conferring with researchers, authors, and editors involved with a project, but also with personnel providing technical support such as data visualization, website development, and media production. Finally, institutional administrators and legal counsel may also provide valuable feedback; in some cases, their support or advice may be required.

The best time to identify potential problems with legal and regulatory liability is in the early planning stage of a research project. Unfortunately, all too often researchers wait until a project is complete, the final report is written, and they are ready to distribute the report, only to discover that the project is missing one or more elements necessary for compliance. This can result in a last minute scramble to secure author permissions, to meet legal or institutional requirements for compliance with standards, to consult with counsel or administrative authorities, or to secure an appropriate copyright license. Early planning using the systems model for identifying interactions and potential compliance issues will permit researchers to incorporate solutions into projects from the beginning. As a consequence, the necessary elements for legal compliance will already be in place by the time a work is ready for dissemination.

**Conclusion: results of compliance**

Researchers, authors, editors, and distributors of grey literature have an obligation to disseminate works that meet a wide number of legal, institutional, and professional requirements. Addressing regulatory, personnel, content, and format challenges will result in distribution of grey literature that is:

- In compliance with local, national, and international law
- Meets requirements of institutional, organizational, and professional policies and regulations
- Supports the needs of all personnel involved

This compliance will minimize or eliminate liability for researchers, distributors, and others involved in production and distribution of grey literature. The key factor for producing compliant information is an awareness of the range of possible legal and regulatory issues surrounding a research project. The systems model of interactions between legal and regulatory issues, personnel, and content offers scholars a useful and effective tool for identifying both potential problems as well as practical solutions.

Appendix:**Copyright and Standards**

- European Union: <https://ec.europa.eu/digital-single-market/en/policies/copyright>
- International Standards Organization (ISO): [iso.org](https://www.iso.org)
- Italy: <http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1941-04-22;633!vig>
- United Kingdom: https://www.copyrightservice.co.uk/copyright/p01_uk_copyright_law
- United States: copyright.gov

Creative Commons

- Creative Commons: <https://creativecommons.org>
- License options for use and commercialization: <https://creativecommons.org/share-your-work/>

SPARC

- SPARC: <https://sparcopen.org>
- Author Addendum to maintain Open Access rights: <https://sparcopen.org/our-work/author-rights/>
- Alternative OA Publishing Models: <https://sparcopen.org/our-work/alternative-publishing-models/>

¹ Joachim Schöpfel and Tomas Lipinski, "Legal Aspects of Grey Literature." *The Grey Journal* 8.3 (2012): 135-153.

² Tomas Lipinski and Katie Chamberlain Kritikos, "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." *The Grey Journal* 12.2 (2016): 67-81.

³ Matěj Myška, "The New Creative Commons 4.0 Licenses." *The Grey Journal* 1, Special Winter Issue (2015): 58-62.

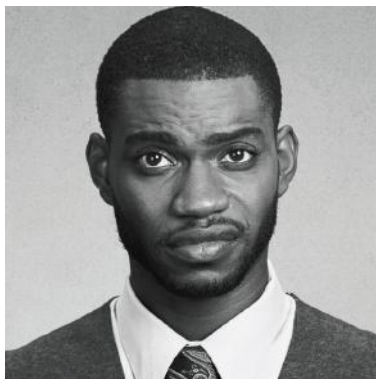
⁴ "SPARC: Who We Are" (<https://sparcopen.org/who-we-are/>), accessed 10/1/2017.

⁵ Joachim Schöpfel, "Document Supply of Grey Literature and Open Access: Ten Years Later." *Internet and Document Supply* 43.2 (2015): 84-93.



EXPAND BEHAVIORAL SCIENCE RESEARCH WITH THE PREMIER RESOURCE FOR GRAY LITERATURE

Devoted to curating and indexing hard-to-find content from authoritative sources, PsycEXTRA® allows researchers to go beyond traditional peer-reviewed research materials. This high-quality and relevant database combines bibliographic records with unique full-text materials, focusing on the latest conference presentations and papers, newsletters, reports, patient-oriented factsheets and brochures, magazines, monographs, and standards and guidelines relevant to the needs of students, faculty, and clinicians alike.



Explore original, cutting-edge, high-quality research

- ▶ Updated biweekly
- ▶ More than 300,000 records and growing
- ▶ No coverage overlap with PsycINFO®, creating an ideal companion database
- ▶ Full text for more than 70% of records
- ▶ Ongoing updates of select professional literature from multiple state and government entities, associations, foundations, and more
- ▶ Average of 25,000 records or more added each year

Available via
EBSCO



Indexing grey multilingual literature in General Practice: Family Medicine in the Era of Semantic Web

Marc Jamoulle, Department of General Practice, University of Liège, Belgium

Elena Cardillo, Institute of Informatics and Telematics, National Research Council, Italy

Ashwin Ittoo, HEC Management School, University of Liège, Belgium

Robert Vander Stichele, Heymans Institute of Pharmacology, University of Ghent, Belgium

Melissa P. Resnick, University of Texas, Health Science Center at Houston, United States

Julien Grosjean and Stk;efan Darmoni, D2IM, University of Rouen, France

Marc Vanmeerbeek, Department of General Practice, University of Liège, Belgium

Problem/Goal: *Sharing the results of research with General Practitioners (GPs) is crucial for the survival of the discipline of General Practice / Family Medicine (GP/FM). The production of abstracts in GP/FM probably exceeds 15,000 per year worldwide. Each abstract often represents two years of work for its authors and is expressed in local languages. Only 45% of them are published in indexed medical journals. Usual indexing systems like MeSH are not multilingual nor adapted to the particular field of GP/FM. Consequently, these abstracts are lacking bibliographic control and more than half of the research presented by GPs at congresses is lost. Considering the absence of appropriate domain-specific terminologies or classification systems, we propose a new multilingual indexing system. The existing International Classification of Primary Care (ICPC) is currently used for clinical purposes and has now been expanded with a taxonomy related to contextual aspects (called Q-Codes) such as education, research, practice organization, ethics or policy in GP/FM, currently not captured. The set is proposed under the name Core Content Classification in General Practice (3CGP). The aim is to facilitate indexing of GP/FM specific scientific work and to improve performance in information storage and retrieval for research purposes in this field.*

Research Method/Procedure: *Using qualitative analysis, a corpus of 1,702 abstracts from six GP/FM- related European congresses was analyzed to identify main themes discussed by GPs (e.g., continuity, accessibility or medical ethics), handled in a domain-specific taxonomy called Q-Codes and translated in 8 languages. In addition, a methodology for building a lightweight ontology (in OWL-2) was applied to Q-Codes, adding object and datatype properties to the hierarchical relations, including mapping to the MeSH thesaurus, Babelnet (www.babelnet.org) and Dbpedia. Finally, the Q-Codes in 8 languages have been integrated healthcare terminology service (www.hetop.eu/q) with a companion website (<http://3cgp.docpatient.net>).*

Anticipated Results of the Research: *The creation and the on-line publication of this multilingual terminological resource, for indexing abstracts and for facilitating Medline searches, could reduce loss of knowledge in the domain. In addition, through better indexing of the grey literature (congress abstracts, master's and doctoral thesis), we hope to enhance the accessibility of research results of GP/FM domain and promote the emergence of networks of researchers. First result of experimental implementations of the new indexing system will be presented.*

Indication of costs related to the project: *This project has not been funded. 3CGP is placed under Attribution-Non-Commercial-Share-Alike 4.0 International (CC BY-NC-SA 4.0). ICPC is copyrighted by WONCA.*

Keywords *General practice, Terminology, Electronic publishing, Repository, Grey Literature.*

1 BACKGROUND

Need for information in family medicine

In the cycle of patient centered information (Jamoulle et al., 2015), the General Practitioner (GP) is simultaneously a heavy user and producer of published/unpublished data. Data could be clinical, (i.e. dealing with symptoms, processes and diseases) or contextual. Contextual data can address particular issues concerning the patient, which may influence the process of care (Schrans et al., 2016). However, contextual data can also deal with issues concerning the doctor, the managerial aspects of care. In particular, it could address the position of GPs within the health care system, the general concepts used in Primary Health Care (PHC), or the delivery services. In this work, the focus will remain on these last contextual medical features of General Practice / Family Medicine (GP/FM), as its tools for training, research, ethics, inquiry, environmental issues, infrastructure and principle of care. These features are central to this field, and family doctors are



used to exchange information over them when they meet in training sessions or during congresses.

The realm of GP/FM differs from mainstream health care, as Family Physicians (FPs) address biological, technological, behavioral, sociological and anthropological domains. All of these have a deep impact on the terminologies needed (Helman, 2008; Thompson et al., 2014). As the creation of already available terminologies was focused on specialized domains, the biological and technological fields of medical terminologies are now almost complete (Jonquet et al., 2016; Lelong et al., 2016). However, they sometimes fall short when applied to the field of GP/FM, which relies intensely on complexity and timeline issues (Liang et al., 2014; Madkour, Benhaddou, and Tao, 2016). Albeit well documented clinical issues, professional contextual issues, like management, teaching, research, and ethics are documented in a fragmented way for the first level of care (Jamoulle et al., 2017a).

1.1.1 GP/FM, a profession without clear limits

Despite elaborate definitions of GP/FM Allen et al. (2011) and Primary Care Physicians (PCPs) (AAFP, 2011), the manner in which the profession of GP/FM or PCP is defined and structured varies greatly across family medicine textbooks (Casado Vicente, 2012; Gusso and Lopes, 2012; Kochen, 2012; Murtagh, 2011; Druais et al., 2009; David et al., 2013; Lakhani, 2003; McWhinney, 1997). This is especially true in regard to managerial and contextual features. These textbooks have offered a top-down expert view of the profession, as the authors of those textbooks themselves chose the subjects addressed. In this research, we rely on what practicing doctors are interested in. In this sense, one can speak of a bottom-up approach.

If one examines the table of contents of these cited works, as far as the general management and the contextual background are concerned, the quoted books are absolutely different. One focuses on communication, the other on the systemic approach, and the third one on the ethics of relationships. None give a similar view of the scope and contextual scope of family medicine. Also, technology is often absent. Only one author or another approaches current technical processes in family medicine.

An extensive review of the vocational training programs in the specialty of General Practice was not done. The author, however, knows from experience and the many contacts he has in family medicine in Europe/the world that these programs have no homogeneity, despite the recommendations of EURACT, the WONCA Europe working group on education (Heyrman, 2005). Also, when considering Continuous Medical Education, the drug industry's influence on the choice of subjects is decisive, which creates multiple conflicts of interests (Davis, 2004). Therefore, it seemed wise to develop an index of concepts dealing with family medicine by listening to practicing GPs, and to develop a bottom-up approach rid of conflicts of interest.

1.1.2 Published and unpublished in GP/FM, what's the meaning?

Medical Subject Headings (MeSH) are normalized keywords directed to enter queries in Medline, the bibliographic data base of the National Library of Medicine (NLM) through the interface PubMed (Lowe and Barnett, 1994). Currently the PubMed interface gives access to 27,3 millions of citations. A search with 7 GP/FM relevant MeSH in June 2017 brings a little less than 200.000 citations. The same interrogation with 8 specific MeSH descriptors of Primary Health Care (PHC) gives 480.000 citations. (PHC and GP/FM related MeSH are listed in Fig.1). Together the 15 specific descriptors of the first line of care gather less than 2% of Medline content.



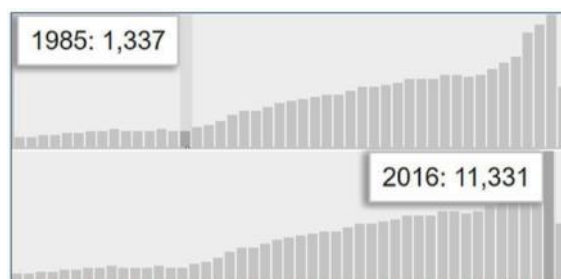
Figure 1. GP/FM & PHC related MeSH in PubMed

Types	MeSH	Year of introduction
Primary Health care (PHC) related MeSH	"Community Health Service" [MeSH:Noexp]	1967
	"Community Mental Health Services" [MeSH:NoExp]	1967
	"Home Care Services" [MeSH:NoExp]	1967
	"Primary Health Care" [MeSH:NoExp]	1974
	"Community Health Centers"[MeSH:NoExp]	1979
	"Community Mental Health Centers" [MeSH:NoExp]	1979
	"Home Care Agencies" [MeSH]	1995
	"Rural Health Services" [MeSH:NoExp]	1996
General Practice / Family Medicine (GP/FM) MeSH	"Physician, Family" [MeSH]	1974
	"Community Medicine" [MeSH]	1977
	"Family Practice" [MeSH]	1978-2010
	"Gatekeeping" [MeSH]	2000
	"General Practice" [MeSH]	2011
	"General Practitioners" [MeSH]	2011
	"Physician, Primary care" [MeSH]	2011

The rise in publications in GP/FM has been inevitable since 1985 (see Fig. 2). The quoted publications considered relevant to GP/FM are obtained through the use of a search strategy composed of seven MeSH concepts related to GP/FM (see Fig. 1). The citations obtained could be certainly appraised as *published*

data and claimed white literature. This raises question about the Pisa definition of grey literature, considered as *not controlled by commercial publishing* (GreyNet, 2014), as one can hardly pretend that papers published by the NLM are always commercially controlled.

Figure 2. Evolution of publications in GP/FM (7 MeSH) from 1985 to 2016 on PubMed



Unpublished data in the GP/FM field are numerous. Yet, GP/FM organizations are heavy producers of continuous medical education (VanNieuwenborg et al., 2016). They contribute greatly to training sessions and organize local, regional and national level medical conferences, as well as research meetings (Buono et al., 2013), virtual conferences (Cavadas, Villanueva, and Gervas, 2010), websites, and blogs. They are also active on social networking (Veuillette et al., 2015). With so many heavy producers of various nationalities, it is important to note that at local and national events, the local language is the rule.

It is not a secret that knowledge translation between doctor and patient is highly controlled by pharmaceutical companies (Moynihan, 2003) (Moynihan and Bero, 2017). Despite the activities of GP organizations, in some countries, domestic papers as medical newspapers remain the main sources of information for practicing doctors (Tabatabaei-Malazy, Nedjat, and Majdzadeh, 2012). Usually, they are edited within an atmosphere of heavy, silent corruption (Angell, 2017). The translation of information by drug representatives is also a determining factor (Greenway and Ross, 2017) as well as predatory open access publications (Shen and Björk, 2015) or pay for publishing process (Quan, Chen, and Shu, 2017) which dismisses whole sectors of publication. This implies that knowledge management tools in GP/FM must be controlled by dedicated, unbiased GPs. The movement of free lunch doctors (<http://www.nofreelunch.org/> (USA)), the no-gracias one (<http://www.nogracias.eu/> (Spain)), the Medico Sin Marca (Chile), (<http://www.medicossinmarca.cl/>), the Therapeutic initiative (Canada)



(<http://www.ti.ubc.ca/>) or, more generally, the members of the International Society of Drug Bulletins (ISDB) (<http://www.isdbweb.org>) refuse to adhere the pharmaceutical industry influence. Their work merits distributing its grey literature in a professional multilingual indexing system that is accessible by all. Finding a way to publish and share information outside the default language of English is attractive to many GPs. However, sharing the results of research with General Practitioners is crucial for the survival of the discipline of GP/FM, which means a universal system must be created (McIntyre et al., 2016).

As an example of hidden grey literature, the website of the World Family Doctor Association in Europe (www.woncaeurope.org) edits more than 30,000 non-indexed abstracts of European or world conferences in English. Each abstract often represents two years of work (Master's theses are included in this). If not published, they become lost work. This also represents missed opportunities to develop networks between authors that share similar interests. Only half of this production has the chance to be published in indexed medical journals (Van Royen et al., 2010; Hummers-pradier, 2007).

Producing Information at the point of care

It could be believed that a simple terminological subset may be sufficient to meet the needs of GP/FM computer systems. Lack of visibility of the complexity of the work of family doctors has allowed for

such biased vision. Moreover, special tools specifically dedicated to primary health care, such as the International Classification of Diseases, tenth revision, for primary care (ICD-10 PC) (Ustün et al., 1995) or the Statistical Manual of Mental Disorders, 4th ed., primary care version (DSM IV PC) (Pingitore and Sansone, 1998) have been developed and nicknamed quickly ICD-10 or DSM-IV for dummies.

Of course, the family physician often sees simple problems. But he is accompanying a set of patients throughout a lifetime. The family physician knows a patient more extensively than most specialists do. He also becomes a specialist in patients bearing rare diseases and of various cultural background. He will, therefore, have extensive terminological needs, even more extensive than many specialists.

1.2.1 Clinical information

The adjective clinical deals here with patient related data, such as: reasons for encounter, symptoms, acts performed or requested and diagnosis. Terminology for clinical information is a highly specialized and difficult field of current medicinal research (Jamoulle et al., 2014). Clinical information is accumulated in Electronic Medical Records (EMRs) and, if well organized, transferred to study centers where huge database may be used to teach medicine, analyze epidemiological data or be used for secondary searches (Charlton et al., 2010; Carey et al., 2004; Britt et al., 2003). Among others, studies worth citing are, mostly with good validity (Khan, Harrison, and Rose, 2010), produced by the Dutch Transition Project (Soler et al., 2012) (<http://www.transhis.nl>), data produced by the Belgian Intego project (Bartholomeeusen, Buntinx, and Heyman, 2002) (<https://intego.be/en>), the Beach project in Australia (Britt et al., 2016) or at a larger scale the UK General Practice Research Database (<http://gprd.com>).

1.2.2 Professional contextual information

The term *contextual* applies as a generic term for the name of the taxonomic product presented here. A concept like *uncertainty*, the usual companion of the doctor or the concept of *quality assurance* or *environmental health*, are all essential elements of professional practice. These are not clinical terms as they do not always deal with current patient problems. The term *contextual* appeared the most relevant, as it was defined in the Meriam-Webster Dictionary as: *the interrelated conditions in which something exists or occurs*.

Answering the question; *What are they discussing?* in a meeting of two, twelve or several hundreds or thousands of GPs may give insight into the details of this well-defined (Jamoulle et al., 2017b) but not limited profession. As stated by Cimino (1998) : *Part of the difficulty with using a standard controlled vocabulary is that the vocabulary was created independent of the specific contexts in which it is to be used*. Adding a contextual supplement to ICPC gives birth to an extended *Controlled vocabulary*, able to take in account the extension and the complexity of the domain covered by GPs. Controlled vocabulary *is a general term for a list of standardized terms*



used for indexing and information retrieval usually in a defined information domain (Library and Archives Canada, 2017). In this case the Controlled vocabulary is also a *Vocabulary coding scheme* as defined in Dublin core (see further).

This approach to the doctor's professional context should not be confused with the contextual approach of the patient's universe, as developed by Schrans et al. (2016) who studied the elements of the patient's life context that influence his or her state of health and the health problems he shares with the doctor.

Consuming information at the point of care

As stated by James (2016), the Internet has triggered a transformational change in the dissemination of science in the form of a global transition to open access (OA) publishing. GPs, the rank and file (Chinitz and Rodwin, 2014) workforce in medicine are using those resources extensively despite sometimes huge material difficulties to access the sources when working in rural or remote areas (Salman Bin Naeem, Shamshad, and Amjid, 2013).

Finding information is sometimes difficult for researchers, even though they may not work with patients and remain mostly in academic laboratories with access to expensive medical journals. So, what about *rank and file* physicians who have only a few seconds to check information and source relevance? (Hubbard, 2008). Availability in the real world, at the point of care, is often clashing with the economic model (paid access) or with copyright issues (Myška and Savelka, 2013).

Open access (OA) to researchable and usable information (Heilman, 2015) at the point of care is of utmost importance in GP/FM. It is necessary to maintain open access, point of care resources at the high level of quality that patient care demands (PLOS Medicine Editors, 2015). A potential issue with this is that GPs should consider that daily medical journals as well as major papers can also be manipulated by the pharmaceutical industry (Schwitzer, 2017; Dowden, 2015).

1.2.3 Sources of information at the point of care in GP/FM

On-line, directly accessible major documentary databases in open access and local language are not numerous nor specific to the GP/FM field (Hubbard, 2008). The US NLM PubMed database (<https://www.ncbi.nlm.nih.gov/pubmed/>) gives access to millions of paid and open access publications, mostly in English. The Pan American Health Organization sponsors Scientific Electronic Library on-line (<http://www.scielo.org>). SciELO is the South American champion of free access journals, mostly in Spanish (PAHO Bireme Sao Paulo, 2016). The World Health Organization (WHO) supports the Global Index Medicus (<http://www.globalhealthlibrary.net>). In France, LiSSa (<http://www.lissa.fr>) gives access to more than 10^6 citations of French literature (Cabot et al., 2017a).

To say nothing of Google Scholar, the web is a considerable resource of information in medicine, especially in gray literature. Janamian et al. (2016) have identified and searched for 260 web sites as GP/FM sources. As stated by González-González et al. (2007) "*In primary care, each practitioner encounters more than 500 different clinical topics in any year*". Ten years ago Internet base accounted for only 5% (ibidem) for search of information by Spanish GPs. This percentage has risen to 59% for German GPs in 2016 (Eberbach et al., 2016). Anyway the use of Internet is now so generalized that studies are performed on influence of Internet the patient-physician relationship (Tan and Goonawardene, 2017).

1.2.4 Use of Medical Subject Headings (MeSH) descriptors in GP/FM

For retrieval of scientific bibliographic information in GP/FM, MeSH descriptors are used by all doctors and researchers. The MeSH is a huge thesaurus of 27,000 hierarchically managed descriptors- i.e., normalized terms, intended to index medical documents and growing yearly (Jamoulle 2016).. Currently, the use of indexing civil data in the health care domain is being studied (Marc et al., 2015).

General Practice is a profession generally regarded as part of the first line of care, PHC, a form of care organization. General Practice and Primary Health Care concepts share the same extension, but not the same intension. The first describes the duty of a profession, the second the management of a service (Jamoulle et al., 2017b).

The confusion could be found in the MeSH thesaurus. Gill et al. (2014) state that: *Constructing a highly efficient search filter to identify primary care relevant articles is challenging, particularly due to the inadequate and ambiguous description of the clinical research setting in title, abstract and*



MeSH keywords. Huang, Névél, and Lu (2011) observes that: *manually assigning MeSH terms to biomedical articles is a complex, subjective, and time-consuming task*. Shultz (2007) argues that: *terminology was observed to be a major factor affecting retrieval and the ability of both systems to obtain unique items*. However, not all aspects of the broad field of GP/FM are covered in a specific area (Sladek et al., 2006). Despite this, interesting advance in MeSH indexation for GP/FM have been proposed (Mendis and Solangaarachchi, 2005) (Jelercic et al., 2010). Theme of interest searched for by GPs have also been studied (Hong et al., 2016). The future looks prepared as MeSH are now available in RDF, ready for Semantic web (Bushman, Anderson, and Fu, 2015).

1.2.5 Use of Health descriptors (DeCS)

The controlled vocabulary of Health Sciences Descriptors (DeCS), initially a translation of MeSH into Spanish and Portuguese, has been expanded with new categories. It was also adopted into the indexing and multilingual search of the scientific and technical literature in South America (<http://decs.bvs.br/l/homepagei.htm>).

It's a by-product of the Latin American and Caribbean System on Health Sciences Information (BIREME), the Pan American Health organization (PAHO) network of libraries and documentation centers (Neghme, 1975). Doctors and students in South America are using DeCS as a standard controlled vocabulary for indexing scientific and technical health-related documents in knowledge databases like the Virtual Health Library (<http://bvsalud.org/>) or SciELO Both sources generally give open access to documents.

1.2.6 ICPC for Classifying Clinical Issues in GP/FM

The International Classification of Primary Care (ICPC) (Soler, Jamoulle, and Schattner, 2015) is routinely used by physicians around the world to categorize the problems encountered in their practice with patients, i.e. their clinical activity. We will see that ICPC has proved effective in many other uses and show that it is adapted for collecting clinical problems that doctors discuss at congresses.

Primary care isn't specialized care. It encompasses specialized care and biotechnological vocabularies along with anthropological, family-based/personal information. A medical record in primary care must encompass all facets of health care, alongside personal and family environment knowledge. This is why, WONCA, the world organization of family doctors, through the ongoing work of its International Classification Committee (WICC), has developed ICPC (Wonca, 1987) (WONCA, 2005) (Lusignan, 2005).

Grey Literature as source of information

The price of access to international high-level journals, mostly exclusively in English, is prohibitive. The economic aspect is therefore a major obstacle to the spread of knowledge outside academic circles. However, the change of economic model is under way in the world of publishing, the cost of which is largely reflected on the author and not the reader. In the meantime, appeals for public access of research data continue to proliferate (Lin and Strasser, 2014). McKenzie (2017) considers that the explosion of the use of Sci-Hub facilities (<http://sci-hub.cc>) is the beginning of the end of the scholarly publishing.

Whatever the case, the open access grey literature in medicine is in full development (Swan, 2012). According to Schöpfel (2015), *The term gray literature remains ill-defined, imprecise, with fuzzy outlines. Its two handicaps are part of its definition: identification, access and acquisition are often difficult, and quality and reliability are not always assured*. Ferreras Fernández (2016, p.211-216) has done an exhaustive review of the definitions of grey literature. Those definitions share the negation of commercial involvement like Pisa's (GreyNet, 2014). Grey literature reviews are not always free of access like the Grey Journal itself (www.greynet.org).

Practically, when addressing the case of grey literature, authors exchange more pragmatic definitions than the Pisa's one as; *difficult to locate or retrieve* (Moher et al., 2000), or; *has not been formally published* (Hopewell et al., 2007) or; *there is no such peer review or passage through quality filters* (Silva, Garcia, and Cássia, 2009).

Interestingly, Hoffmann et al. (2011) points out that the *grey literature yields more substantial information* (than white literature) *on the content of interest*. This could be understandable,



partly as white medical literature is not free from the influence of the industry (Gotzsche, 2013; Schwitzer, 2017).

We propose to consider grey literature in GP/FM publications that share the following characteristics:

- For the background:
 - Sharing knowledge specific to the field of GP/FM, no matter the format (papers, articles, memo, master thesis, PhD thesis, leaflet, abstracts of presentation, web pages, video, images, YouTube, Facebook, Twitter, Google+, LinkedIn, dataset).
 - Being unreferenced in well-known local or international medical databases (PubMed, LiSSa, Scielo, Lilacs, ORBI, etc.).
 - Being submitted to a scientific quality assurance process (in anthropology or bio-sciences).
- For the format:
 - Being freely accessible in an Open access model.
 - Using a systematic multilingual vocabulary encoding scheme (indexing system).
 - Relying to Dublin Core Metadata Initiative or equivalent standardization process.
 - Being ready for machine use in the semantic web.

Metadata and Vocabulary Coding Scheme

Metadata consists of statements we make about resources to help us find, identify, use, manage, evaluate, and preserve them (Sutton, 2007). Metadata may be interpreted by machines and people. Dublin Core Metadata Initiative (DCMI) (<http://dublincore.org/>) provides simple standards to facilitate the finding, sharing and management of information. Metadata are basic description mechanism for digital information that, can be used in all domains, for any type of resource, simple, yet powerful, can be extended and can work with specific solutions, making it easier to find information on the Web as it develops. DCMI participates in the development of the “new Web”, the Semantic Web and Linked Data (Dekkers, 2009). Allen (2016) states that *The emergence of machine intelligence and machine reading in the second machine age will make it even easier to automate the production of metadata to help people find, filter and organize information.*

Quality of search results is dependent on the quality of the metadata in the original repositories of which high quality structured metadata are more accessible (Farace and Schöpfel, 2010).

We are thus dealing with knowledge identification process by humans and by machine through well formalized denominations. So we are addressing here the concept of *Controlled Vocabulary*. The reader has to be conscious that the same concept could bear different names. For instance the Australian Metadata Online Registry (MeteOR) uses the term *Classification scheme* for pointing the same issue. *A classification scheme is an official terminological system, recognized and endorsed by a national or international body, that is used to classify data.* (<http://meteor.aihw.gov.au>).

Grey literature and Semantic web opportunities

Metadata allow the retrieval from data from dedicated repositories. Nevertheless, as stated by Goggi et al. (2015), *documents may contain important information that has not been encoded in the metadata*. Extracting key concepts from unstructured texts is the following step, done by semantic annotators, by-product of research in Natural Language Processing (Cabot et al., 2017b). Key concepts could be added to indexing facilities or tagged as identifiable information for use in Linked Open Data (LOD). This opens 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* (Goggi et al., 2015).

As stated by (Cardillo, 2015) : *During the last ten years ontologies and the use of Semantic Web technologies has been seen as a better solution to semantic interoperability because this allows describing the semantics of information sources and makes its contents explicit by providing a shared comprehension of a given domain of knowledge [.] Unfortunately, ontologies and their structure are not really familiar and natural to most healthcare providers and their use raises heterogeneity problems to a higher level.*



2 AIM OF OUR RESEARCH: PROPOSAL FOR A NEW CODING SCHEME IN GP/FM

Our work aims to identify the themes in knowledge production by GPs in a new Vocabulary Coding Scheme called Core Classification of General Practice Family Medicine (3CGP). This program encompasses clinical and contextual situations in the GP/FM practice. Simultaneously, we hope to develop our system in such a way that machines (i.e., computer), could deal with that data and reason about it using the Semantic web technologies.

As mentioned above, classifications and terminologies for patient data retrieval are numerous. However, one must know if they could be used to index and retrieve documents in a specific manner. Indeed, units of knowledge managed in historically different terminologies and classifications are interlinked and address the same reality seen by various eyes and interests (Bowker and Star, 1999).

The absence of adapted concepts and descriptors for contextual aspects of GP/FM is one of the reasons why the scientific work of family physicians is hard to retrieve from mainstream bibliographic systems. In addition, more than 50% of the scientific output of GPs at conferences is never published (Van Royen et al., 2010). There are no dedicated indexes of grey literature (Mahood, Eerd, and Irvin, 2014), and abstracts or collections of dissertation titles are often not properly indexed in this field (Lawrence et al., 2014).

This work presents a new taxonomy of contextual aspects of GM/FM, in hopes of helping to improve the situation surrounding GM/FM grey literature. Taxonomies provide schemes to help classify entities and define the relationships between them (Dixon, Zafar, and McGowan, 2007). The purpose of this development is also to provide tools to exploit modern technology - in terms of terminology for information storage and retrieval systems (Vanopstal et al., 2011), such as: machine learning, semantic web techniques, natural language processing (NLP) and linking data. This kind of system is already in use in clinical settings for patient data (Colliers et al., 2016) and one hopes to apply such techniques to an indexing system in a near future for the communication of family doctors in congresses and related grey literature.

In brief, our aims are triple:

- To improve annotation of grey literature in primary care.
- To facilitate indexing of congress abstracts and theses.
- To improve the searchability of repositories for these information artefacts.

3 METHODS

Referring to METHONTOLOGY steps for the development of the project

The phases of development of the project are shown on Fig. 3 along the time line. Qualitative analysis of communications of GPs during congresses has induced the creation of a controlled vocabulary organized in a taxonomy. To develop a domain-oriented taxonomy (the simplest form of an ontology- i.e., a light- weight ontology), methodology for ontology construction was included (Gómez-Pérez, Fernández-López, and Corcho, 2003). The four main phases of the METHONTOLOGY process are shown Vertically :

1. Knowledge Acquisition and formalization;
2. Integration process;
3. Implementation;
4. Publication and Dissemination

Knowledge acquisition, formalization and integration were added in 2005. The implementation phase in the online Hetop server began in 2014. We have added a dissemination phase through Internet and publications.

Knowledge acquisition & formalization; Qualitative analysis of GPs' communications by a Computer-Assisted Qualitative Data Analysis Software (CAQDAS)

Using qualitative analysis, a corpus of 1,702 abstracts from six GP/FM-related European congresses was analyzed to identify 182 themes discussed by GPs (e.g., continuity, accessibility or medical ethics), handled in a domain-specific taxonomy called Q-Codes and translated into 8 languages. To identify key concepts in a domain-specific taxonomy, data is analyzed in a grounded theory

approach (Glaser and Strauss, 1999). This approach is often used in disciplines such as: economics, law, and medicine (Wells, 1995; Denzin and Lincoln, 2000). It involves the construction of a hypothesis or discovery of concepts through data analysis (Faggiolani, 2011; Martin and Turner, 2016). After a careful study of existing products, the qualitative analysis software (ATLAS.ti® <http://atlasti.com/>) was used, as it enabled the required analyses to be executed at a relatively low cost. ATLAS.ti enabled the ability to map specific words to already-defined ICPC-2 and to find new concepts to feed the new Q-Codes taxonomy. The same theme could not reappear in the same abstract more than once, and (generally) no more than six themes were identified in each abstract. The analysis performed by EGPRN in 2010 on 614 abstracts, using a similar approach, has been used to control the QR (Research) domain and check the consistency of the Q-Codes proposal.

Two additional steps are required to complete the lightweight ontology (taxonomy) construction process according to METHONTOLOGY, namely: Integration and Implementation.

Integration phase, birth of 3CGP

The Core Content Classification in General Practice/Family medicine (3CGP) is formed by the addition of ICPC-2 for clinical issues and Q-Codes for professional contextual issues, both discussed during meetings between GPs. The Q-Codes taxonomy was elaborated on the model of ICPC, using the letter Q to categorize the contextual elements, for the letter Q was unemployed in ICPC-2.

$$\text{ICPC-2} + \text{Q-Codes} = \text{3CGP}$$

Figure 3. The phases of development of the project on a time-line. The four main phases of the METHONTOLOGY process

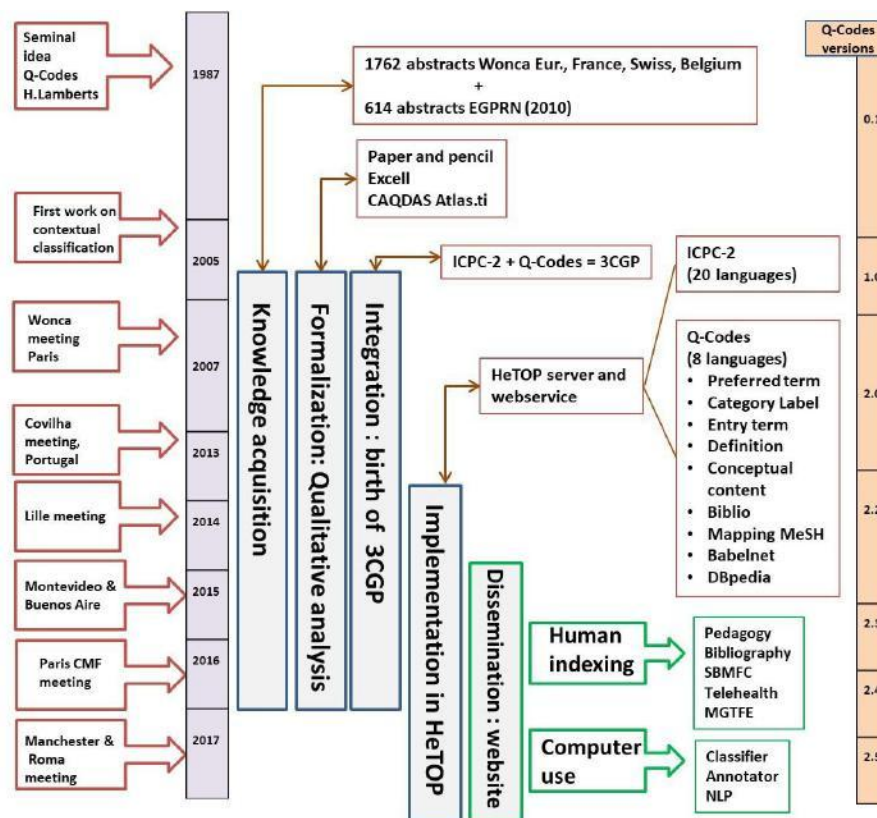
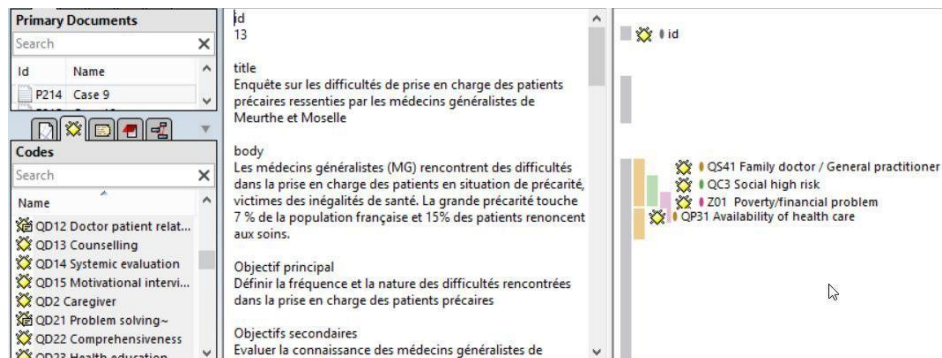


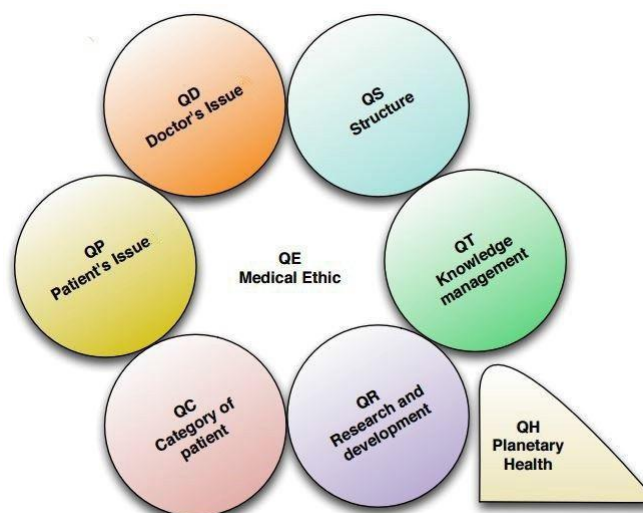


Figure 4. Congress CNGE 2013: After being scanned, coded themes appear on the right column: Q-Codes (QS41, QC3, QP31) and one ICPC code (Z01). Here, the theme identified is the offer of family medicinal services in vulnerable populations. At the bottom left, the software proposes the list of pre-registered Q-Codes. (ATLAS.ti Software)



Q-Codes are born from the analysis of 1,702 abstracts of conferences. Naturally, we hope that future conferences will allow for a surge in new concepts and new entries in the Q-Codes classification. Q- Codes are divided into 8 domains. The taxonomy starts with the QC domain, which represents *Patient's category*, and covers topics such as age, gender issues, and victimhood. The second one is the QD domain, representing *Family doctor's issue*, which covers issues such as disease management, communication, clinical prevention, and medico legal issues. QE represents *Medical Ethics*. This domain covers bioethics, professional ethics, and info-ethics. The fourth domain is QH, representing "Planetary Health", which deals with such areas as environmental health, biological hazards, and nuclear hazards. The fifth domain is QP, *Patient Issue*, which includes patient safety, patient centeredness, and quality of care. The QR domain is *Research & Development*, covering research methods, research tools, and epidemiology of primary care. QS is the Structure of Practice domain. It covers topics such as primary care settings, primary care providers, and practice relationships. Finally, the QT domain is Knowledge management. This domain deals with teaching, training, and knowledge dissemination. Each domain of Q-Codes is divided in Categories, Sub-Categories, and Sub-Sub-Categories. A ninth domain, QO for Other will be used in the abstract coding process for not precise descriptions or for a concept worth to be considered as a potential candidate for a new theme.

Figure 5. The Q-Codes matrix in the shape of a Q-Letter.



The presentation of the Q-codes under a matrix format is shown on Fig. 5. The matrix takes the shape of the letter Q, representing the 8 domains of the Q-Codes. On the left, the people related domains - Doctor's issue, Patient's issue and Category of patients; On the right, the managerial related domains - Structure, Knowledge management including Teaching and Training, and Research and development; Hazards are the underlying Planetary health conditions



represented by the downward oblique tail stylized as a triangle but which are in reality the background of the GPs work; in the center, joining all, Medical Ethics. Note that the Q's tail, which is the Planetary Health, prevents the wheel from turning endlessly. This is a nice demonstration of the importance of the environment on health issues (Graphic design Patrick Ouvrard).

Implementation ; Organizing the concepts of the taxonomy following a Data Structure Diagram on the HeTOP server

Integration and implementation came to fruition in the meantime. The ICPC-2 classification was edited on the HeTOP web site in 22 languages and the Q-Codes in 9 languages; French, English, Dutch, Spanish, Portuguese, Vietnamese, Turkish, Georgian and Korean. More are coming (Greek, German, Italian, Ukrainian). The Data Structure Diagram, a graphic technique, based on a type of notation dealing with classes of entities and the classes of their relationships (Bachman, 1969) has been used to organize the mappings. In Fig.7, the central concept (here, Overmedicalisation), is linked by its relations (is a - consider - has a definition, conceptually related to) to other formally defined fields of knowledge. This kind of structure is machine readable and forms the basic structure of our taxonomy. It is presented in Excel format in Fig.6.

Dissemination; The HeTOP server as a GP/FM knowledge resource

The HeTOP server, produced by the Department of Medical Information and Informatics (D2IM) of Rouen University (France) is edited in the Web Ontology Language (OWL), which allows for the linking of data with other data (McGuinness and Harmelen, 2004). HeTOP is based on a multi-terminology meta-model that integrates all terminologies and ontologies into its data core. It is cross-lingual since terminologies and ontologies are often available in several languages. The web site can be used by both humans and machines via a dedicated web service (<http://www.hetop.eu>). HeTOP currently contains 71 health terminologies and ontologies (only 17 are included in UMLS as most of them are French terminologies),

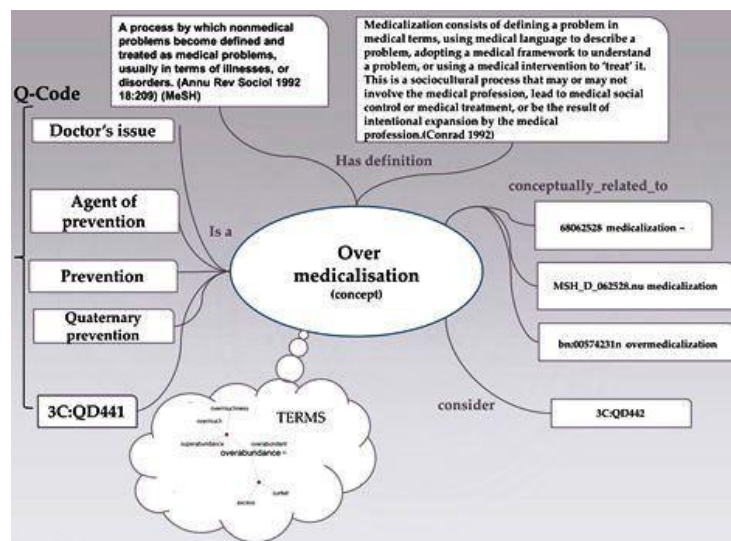
Figure 6. The 14 fields of each Q-Code in the HeTOP interface of which conceptual links are described in Fig.7.

HeTop field	Signification	Remark
Category ID	Alphanumeric identifier of the Q-code	Letter Q followed by one letter (C,D,P,S,T,R,H,O) followed by maximum four numeric digits
Category label	Full title of the category	First letter capital, plural possible, can be compound words
Preferred term (PT)	Normalized title of the category	English, masculine, singular, lower case
Preferred term, other language	Translated (es, pt, nl, vn, tr, fr) version of the PT (more allowed)	Will be used as text word [TW] search term in PubMed equation (in English)
Supplementary entry term(s)	Additional search term(s)	Will be used as text word [TW] search term in PubMed equation (in English)
Category definition	Definition of the title of the category	Reflecting the corporate culture of GP/FM
Category conceptual content	Set of definitions or descriptions from online nomenclature, dictionaries or thesauri	Shows the extension of the concepts, includes the curated MeSH definitions
Automatic HeTop interterminologic relations	Automatic mapping of the PT by the HeTop embedded terminologies	Each proposal could be checked accepted or refused
Terminological features	Broader Than Narrower Term (BTNT) or Narrower Than Broader Term (NTBT)	Establish the hierarchical position of the PT mapping and links it to the HeTop semantic network
Curated MeSH	Accepted MeSH(s) which meaning correspond to the defined content of the PT	Will be used as [MH] in the PubMed equation automatically generated by the HeTop system
Refused links	Manually terminological links judged non-convenient	Will not be used in the search equation
Bibliographic free full text links	URL of citations of free full texts highlighting the content of the Q-Code	Generally chosen in PubMed but also through Google scholar
Babelnet.org link	Link to the URI of the corresponding babelnet.org entry(ies)	Map the Q-Code to an extensive semantic network of multilingual knowledge
DBpedia.org link	Link to the URL of the corresponding DBpedia entry(ies).	Map the Q-Code to a major knowledge semantic database. Could be Wikipedia if DBpedia missing

2,538,595 concepts, 9,982,113 terms, 10,120,417 relations and 32 managed languages. This cross-lingual terminology server is dedicated to various usages by different types of users: translators, students, teachers, researchers, librarians, physicians, etc. HeTOP allows users to search and browse Health terminologies and ontologies in a second (Grosjean et al., 2012). Cross-linguality allows matrix navigation: among terminologies, but also among languages. HeTOP is a multilingual terminology server that not only allows one to search for concepts in several terminologies (and several languages) at the same time, but represents their interoperability. (Friedman et al., 1999). The content of the HeTOP database could be downloaded in CSV (EXCEL), RDF, SKOS or OWL format.

Thanks to contacts all around the world, the collaboration with D2IM team allowed for the completion of the online publication of ICPC-2 in 20 languages (Schuers et al., 2015), Ukrainian and Greek being the last ones. Some colleagues went a step further and translated the HeTOP interface, allowing health professionals to use it in their native language to access ICPC-2. Short after, Q-Codes took place as the last born of terminologies on the HeTOP server. Adding Q-Codes to ICPC-2, it was then possible to develop a complete terminology adapted to GP/FM and PHC needs. Although WONCA retains copyright on the use of ICPC, every effort should be made to disseminate it. The HeTOP database is freely accessible by means of a simple registration process. The Q-Codes belong to the author, but they are licensed under a Creative Commons Attribution Non Commercial (CC-BY-NC) licence.

Figure 7. Data structure diagram (DSD) of a Q-Code, showing the map of concepts and their relationships (conceptual data model)



ICPC-2 & Q-Codes available under URI format

Each HeTOP rubric could be also expressed under an Unique Resource Identifier (URI) format (see Fig.8) A Uniform Resource Identifier (URI) is *a compact sequence of characters that identifies an abstract or physical resource* (Berners-Lee, Fielding, and Masinter, 1998). It is a string of characters used to identify a resource (Miller, 1998) A URI identifies a resource by either location, name, or both. In addition to identifying a web resource, a URI specifies the means of acting upon or obtaining the representation of it. Each entries of ICPC-2 and Q-Codes individual rubrics are available under URI format on the HeTOP server. The chain of character is stable. Languages are expressed under the ISO 639-3 Codes for the representation of names of languages and ICPC-2 or Q-Codes rubrics by their respective codes (see fig. 8). The following URIs are giving access to the hierarchies and rubrics of the corresponding classification. Note that each entry give access to a detailed terminological description, mappings to other terminologies and to automatic queries on resources like PubMed.

- URIs to reach the hierarchy of ICPC and Q-Codes
 - ICPC-2 http://www.hetop.org/hetop/?la=en&rr=CIP_C_ARBO&tab=1
 - ICPC-2 Process http://www.hetop.org/hetop/?la=en&rr=CIP_C_ARBOPROC&tab=1
 - Q-Codes http://www.hetop.eu/hetop/Q/?la=en&rr=CGP_CO_Q&tab=1
- URIs to reach each rubrics of ICPC and Q-Codes



- ICPC RFE and diagnosis: http://www.hetop.org/hetop/?la=en&rr=CIP_D_A01
- ICPCProcess: http://www.hetop.org/hetop/?la=en&rr=CIP_P_30
- Q-Codes: http://www.hetop.eu/hetop/Q?la=en&rr=CGP_QC_QC1
- To change the language; change the ISO 639 for the language; Ex.: =en for =pt for Portuguese (en,fr,es,pt,tr,vi,ko,nl, ge allowed - more in progress)
- To change the rubric; change the code at the end. Examples :
 - ICPC process code #33 in English: http://www.hetop.org/hetop/?la=en&rr=CIP_P_33
 - ICPC-2 S chapter in Japanese: http://www.hetop.org/hetop/?la=ja&rr=CIP_C_S&tab=1
 - Q-Code QC Patient category in English: http://www.hetop.eu/hetop/Q?la=en&rr=CGP_QC_QC
 - Q-Code QD323 Shared decision making in Spanish: http://www.hetop.eu/hetop/Q?la=es&rr=CGP_QC_QD323.

Figure 8. The URI for the code ICPC-2 A04 (Tiredness) in English



RESULTS

The tools developed to carry out this research are presented in the methods portion of this paper but can also be considered results. The provision of ICPC-2 and the Q-Codes in the form of Unique Resource Identifiers (URIs) was completed by a support web site (<http://3cgp.docpatient.net/>) and a Q-Code working group (<https://tinyurl.com/Q-codesWG>). All are technical, communicational or human realizations aimed at the achievement of this endeavor. 3CGP has been designed to be used by both humans and by machine.

3CGP use by humans

3.1.1 Pedagogical use

The ICPC is used worldwide as the main data producing system in Primary Care. It is incorporated into Health Information Systems and used in Electronic Medical Records in numerous countries. Availability of ICPC-2 in multilingual URIs is a must for teaching ICPC worldwide.

The eight domains addressed by the Q-codes are the embryo of what could become the table of contents of GP/FM. Teaching GP/FM is a must when referring to a rarely taught although so frequent as *Medically unexplained symptoms* or *Indoor pollution*.

The terms and definitions of the 182 Q-Codes are available in multiple languages, stressing the international interest surrounding this database. The terms and definitions have been edited in book format in 6 languages (es, pt, fr, en, nl, vi). All versions minus Vietnamese are available at the printing office (<https://www.publier-un-livre.com/en/>). All terminologies are available online on <http://3cgp.docpatient.net/>.

3.1.2 Bibliographic use

GP/FM has no specific indexing system. Twenty per cent of the ICPC-2 codes and all the Q-Codes are mapped automatically to MeSH and each mapping curated manually to MeSH of the National Library of Medicine. Q-Codes are a wonderful tool for teaching specific fields of GP/FM. They are also a useful resource of knowledge for students, researchers and working practitioners at the point of care. Automatic specific citations retrieval system allows access to dedicated bibliography on PubMed but also to LiSSa, the French resources base in medicine (see Fig. 9).



Figure 9. The HeTOP query interface for the Q-Codes *Medically Unexplained Symptom* proposes automatic queries to PubMed and LiSSa bibliographic bases.

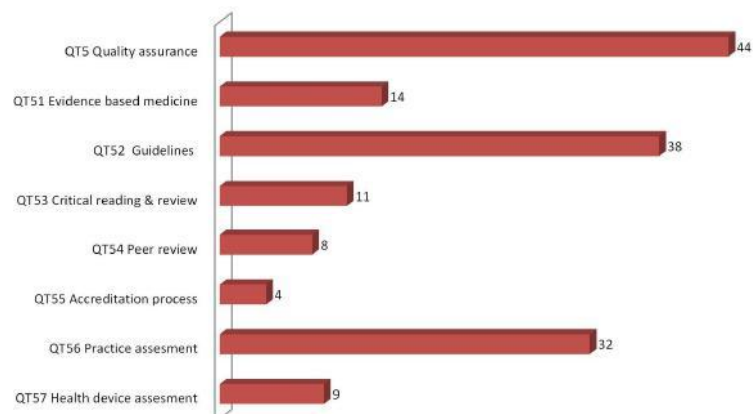
3.1.3 Indexing master theses

The figure obtained by manual indexing of conferences to develop the Q-Codes taxonomy are still a source of information on the interest of participating GPs. In Fig.10 we show the distribution of the rubrics of the QT domain in a French congress of GPs.

In this domain, several experiments are ongoing. The work of doctors in general and family medicine training is often of high quality, requires considerable investment from the authors and sometimes represents little-explored research areas. The leaders of the association of the three Belgian francophone universities (CCFFMG) (see <http://www.mgtfe.be/le-guide-dindexation-dun-tfe/>) decided to give visibility to this work by publishing the best of them online. The 3CGP indexing system was chosen to index work by authors At the time of registration. This guide, also available online

in English (<https://tinyurl.com/Q-Codes-guide>), has been incorporated into the online instructions for authors.

Figure 10. Distribution of QT (Training & Teaching) codes in a congress of French GPs showing the main domains of Interest of participating doctors (number of codes used on 212 abstracts) (Q-Codes version 2.3).



3.1.4 Indexing of congresses

The system developed to index master's theses is reused by the Brazilian Society of Family and Community Medicine (SBMFC). The 3CGP coding system is in use at the deposit page of their 14th Congress. The participants have to choose at least two codes and a maximum of 4 codes of ICPC-2 and Q-Codes. This experiment is currently underway (see <http://3cgp.docpatient.net/codificacao-do-congresso-sbmfc>).

So far 1,746 reviewed and coded abstracts with ICPC and Q-Code have been retrieved and will be analyzed.



3.1.5 Indexing question-answer pairs

This approach is now used by researchers on data from Pernambuco, Brazil, for initially manually indexing a sample of 550 questions; with an ultimate goal of semi-automated indexing of larger data sets, measured in the tens of thousands of question-answer pairs. These question-answer pairs originate from the Brazilian Telehealth system, representing communication between rural health care providers and nurses and doctors in the urban Telehealth centers. (Resnick et al. 2013)

3.8 3CGP use by machines

3.8.1 Automated classifier

To find an automated method capable of analyzing the content of non-clinical General Practice articles and predicting the corresponding Q-Codes categories is not an easy task. A classifier was already developed at the department of Information Systems at the University of Liège (HEC, Professor Ashwin Ittoo). The main difficulties arose from the small amount of sample data available, the large number of categories to be identified, and the high specificity of the scope of Q-Codes making categories difficult to discern (Rigaux, 2015). The classifiers also use filtered lemmatization, and they obtain a modest F1-score of 0.452 and 0.344 respectively. The full work is available in French on; <https://tinyurl.com/yc5ej2bw>.

3.8.2 Automated annotator

The tool *Extracting Concepts with Multiple Terminologies* (ECMT) is a web service developed at D2IM, Rouen. It aims to fully, automatically identify clinically relevant entities in medical texts in French with several types of documents: abstracts titles, documents about marketed drugs and death certificates (Cabot, 2016). The extraction is performed at the phrase level of the text. ECMT has also a user-friendly interface accessible after authentication (<http://ecmt.chu-rouen.fr/>).

Figure 11. Automatic annotation of concepts by ECMT v3 by MeSH (MSH), National Cancer Institute (NCI), MedDRA (MDR), SNOMED (SNO) etc. The red arrow shows the automated identification of concepts in Q-Codes (CGP); QD4 Prevention and QD44 Quaternary prevention (in French).

Extracteur de Concepts Multi-Terminologique (ECMT v3)			
How-to - Contact - © 2017 CHU de Rouen - CISMeF.			
La prévention Quaternaire (P4) est l'ensemble des activités de santé qui atténuent ou empêchent les conséquences des interventions inutiles ou excessives du système de santé.			
Effacer 1 phrases annotées en 525 ms. 44 codes distincts identifiés.			
Terme	Ter. Code	Prévention	TSP 009460
154.24 - Activités	DEW 154.24	Prévention	MDR 10036654
320.101 1 - Systèmes	DEW 320.101 1	Prévention	NCI C15843
540.113 - Systèmes	DEW 540.113	prévention et contrôle	MSH Q000517
551.79 - Quaternaire	DEW 551.79	Prévention santé	TSP 009471
Activité	TSP 000206	procédure	SCT 71388002
activité	IUP A00113	procédure préventive (procédure)	SCT 169443000
activité	NCI C43431	QD4 prévention clinique	CGP QD4
activité	SCT 257733005	QD44 prévention quaternaire	CGP QD44
Analyse systémique	TSP 000743	S70B301 PREVENTION	CLA S70B301
Conséquence	TSP 002943	S72EA PREVENTION	CLA S72EA
conséquence	NCI C74555	Santé	MSH M0009825
Ensembl	NCI C45763	Santé	ICN 10008711
ensemble	NCI C63802	santé	SCT 263775005
ensemble	NCI C47894	santé	MSH D006262
excessif	NCI C73992	santé	NCI C25178
excessif	SCT 260378005	Système	NCI C25700
Intervention	RAD RID10381	système	SCT 246333005
Intervention	NCI C25218	système	NCI C40568
LE systémique	MDR 10024067	système	IUP S06234
médecine préventive	CIS MT21	système	SCT 31099001
médecine préventive	MSH D011315	système	SNO G-A572
		système	NCI C13310

The pertinent terms are retained based on the HeTOP resources. In Fig. 11 an example is given for the processing of the phrase: *La prévention Quaternaire (P4) est l'ensemble des activités de santé qui atténuent ou empêchent les conséquences des interventions inutiles ou excessives du système de santé.* [Quaternary Prevention (P4) is the set of health activities that mitigate or prevent



the consequences of unnecessary or excessive health system interventions.] ECMT extracts the terms in French from terminologies in HeTOP like MeSH or the National Cancer Institute terminology (NCI) as well as from Q-Codes terminology (CGP). As observable in Fig.11, QD4 Clinical Prevention and QD44 Quaternary Prevention have been identified.

3.8.3 Using Q-Codes in an e-learning program, Vietnam

Dr. Thành Liêm Võ, from the family medicine unit of the Pham Ngoc Thach Medical University, Ho Chi Minh City, Vietnam (<http://www.pnt.edu.vn/vi/>) has incorporated the Q-codes in Vietnamese in an e-learning system for medical students. The glossary of Family Medicine terminology helps one to understand and standardize the complex concepts of the discipline. It reduces the variety of these interpretations. Vietnamese versions of Q codes are used as a source of reference. For now, Q-Codes has been integrated into the format of glossary in 3 months FM training at Pham Ngoc Thach Medical University.

4 DISCUSSION

4.1 Main findings

Three areas of knowledge are at stake in this study: (i) Family Medicine as a pillar of primary care, (ii) Computational linguistics, and (iii) Information systems. The association of ICPC, in its three components Symptoms, Procedures and Diagnostics, with the Q-Codes forms an indexing system. This system therefore covers clinical and contextual elements specific to General Practice and Family Medicine. This system allows us to identify patients' symptoms and complaints, diagnosis or disease hypotheses, processes used by physicians, either by themselves or by third parties, and, finally, the context of application given by Q-Codes.

The Q-Codes represent a form of controlled medical, multipurpose vocabulary that is subject to further additions. As stated by Cimino *the unit of symbolic processing is the concept - an embodiment of a particular meaning*. Q-Codes can be seen as a medical subject authority list, including medical subject headings, a comprehensive series of mutually exclusive terms. According to guidelines set by Cimino, we have tried to gather a set of non-redundant, shareable, multipurpose, high-quality permanent concepts, in a mono-hierarchical organization, identified by a set of definitions and linked to existing terminologies. This study proposes a system of Knowledge Management (KM) in GP/FM which could potentially fill a major gap in KM of GP/FM. Conceived as a lightweight, multilingual ontology that is fit for new Internet technologies, NLP, and Semantic Web, 3CGP gives the opportunity to unravel GP/FM productivity and establish GP/FM as a professional discipline aiming at an extended range of specific knowledge.

4.1.1 Filling in a major gap in GP/FM and PHC

To the best of our knowledge, there is nothing similar available in GP/FM that has been developed for both human and machine use. There is also not anything of this measure that demonstrates the complexity of GP/FM. Due to the overlap GP/FM with the first line of health service, this tool could also be useful in Primary Care. All doctors and health managers, for whom proximity and health management are of utmost importance, could potentially reuse Q-Codes for their clinical needs, for teaching and for indexing.

4.1.2 Paving the way for an ontology in GP/FM and PHC

Though this project took years of work, it acts only as a base from which future researchers may expand upon. As it was designed according to terminological/ontological concepts, is available in OWL and is ready for use with Linked Data. The set of ICPC-2 and Q-Codes is a lightweight ontology; however, because it adapts NLP and automatic and semiautomatic coding (Cabot et al., 2017b), it could serve as the basis for the development of a real ontology in GP/FM. The path to a real ontology is still a long time in the making.

4.1.3 Opening the gates for multilingualism in GP/FM and PHC

English has always been the fall-back language of GP/FM. However, family doctors speak to their patients and with one another in their own language, which leads to confusion in translation and varied context of vocabularies. This study has given a potential solution to this issue, by



allowing for ICPC-2 to be published in 20 languages and Q-Codes in 8 languages. Having tools that facilitate various languages, while simultaneously communicating the same concept without variation in context or understanding, is incredibly important and useful/necessary for GPs. Having a tool that accommodates so many different mother-tongues may explain the enthusiasm of so many international colleagues that wished to participate in this multilingual edition.

Study limitations

4.2.1 A Single-Researcher Study

An important issue to address is that there was a seven-year hiatus in this research, shown by the dates of the conference abstracts analyzed. This was due to an extended illness by the main author. Despite this hiatus, research was eventually able to move forward. Any negative effects resulting from this hiatus may be offset by the fact that only one researcher analyzed the abstracts. Bradley, Curry, and Devers (2007), qualitative data analysis experts, argue that *a single researcher conducting all the coding is both sufficient and preferred.[...]. In such cases, the researcher is the instrument; data collection and analysis are so intertwined that they should be integrated in a single person who is the choreographer of his/her own dance[...]* However, bias of said researcher could have influence over the collection of data and its analysis. Therefore, disclosure of the researcher's biases and philosophical approaches is essential. In this case, the main researcher is a male of Occidental origin who has been practicing as family doctor for more than 40 years with an expertise in Public health and taxonomy. An evaluation for the appropriateness of the selected terms could be in as future work the identification of the terms in a big sample of documents using semi-automatic term extraction or key phrases tools, to see the coverage with respect to the one selected by the one researcher. Of course a term extraction process needs in any case a further clinical review by one or more (better) domain experts.

4.2.2 An empirical move

The Q-Codes form the initial building blocks of classification in the GP/FM field. However, this approach has been filled with the personal experience of the main researcher, which may lead to unintentional biases. One can argue that the qualitative approach to the coding process is both inductive and deductive, an approach sometimes called abductive (Silver and Lewins, 2014). As an empirical document, one has tried to change, fill in the gaps and modify content of classifications using GP/FM publications, pair experience, critiques, and application to real work. MeSH's corresponding descriptors, searching, and indexation exercises on published documents have been also a good way to verify the applicability of the classifications. For safety, we've chosen to distribute the concept over all the classifications when adequate, rather than creating a special category. Each conference affirmed this thought and allowed for the addition of new elements. The fact that new concepts have emerged within Q-Codes has two reasons. First of all, the issue was addressed several times in conference abstracts. Secondly, the expertise in the field confirmed that the discussed issue was important.

4.2.3 Potentially Eurocentric

Another limitation of this study is that the data is Eurocentric. This is due to the fact that the conference abstracts analyzed present work done mainly by European GPs. Thus, the Q-Codes concepts might not be fully representative of other geographical areas- i.e., North America, South America, and Asia. This, in turn, may limit worldwide usability. Nevertheless, the fact that the Q-Codes have been translated into three non-European languages (Turkish, Vietnamese and Korean) implies that the translators have found points of connection to their own culture in the proposed concepts. However, this illustrates that the globalization of GP/FM concepts are strongly influenced by its Occidental, Anglo-Saxon origins (Simon, 2009; Gutierrez and Scheid, 2002).

4.2.4 Validity and reliability

Another potential issue is the validity of identification and concepts generated. Validity is concerned with *whether a variable measures what it is supposed to measure* (Bollen 1984 cited by Adcock and Collier, 2001). Here, we deal with the identification of concepts in texts. Yet, how can we measure that the same text will generate the same concepts accurately? Adcock and Collier (2001) also state: *Because background concepts routinely include a variety of meanings,*



the formation of systematized concepts often involves choosing among them. They distinguish between a *consensual concept* and a *contested concept*. It is supposed that a text about *gender violence* will be identified with the corresponding concept *gender violence* by a reader. But, for more ambiguous terms, like *continuity* which is often confused with *permanence*, or more contestable concepts like *disease mongering* and *deprescription* which some colleagues may have no knowledge of, how does one proceed? This ambiguity may pose issue in the execution of this project.

There are as many definitions of *validity* in qualitative research as there are authors. *Face validity*, in quantitative research, is defined as *the extent to which a test is subjectively viewed as covering the concept it purports to measure*. (Holden, 2010). Noble and Smith (2015) propose a new terminology and criterion to evaluate the credibility of research findings. Usual terms used in quantitative research such as *validity*, *reliability* or *generalisability* are replaced with *Truth value*, *Consistency* and *Applicability*. Evaluating *Truth Value* - *Face Validity* - *Descriptive Validity* is recognizing that the interpretation bias, the particular way in which researcher view reality, corresponds to the reality in his/her colleague's world of reference. Many participants offered to translate and contribute to the development of the tool. This made a good argument in favor of the *Truth Value* of these findings. On the other hand, we have seen that the tool could be applied to very different situations in different countries in different languages. These two last points can bear witness to good *Truth Value* but also to good *Applicability*.

Evaluating *Consistency* - *Reliability* - *Interpretive Validity* is referring to whether these Q-Codes could be tested. It was imperative that the Q-Codes could be evaluated through extensive use GP/FM grey literature indexation before being considered a valid construct. One measure used for testing was inter-indexer reliability. But, according to Funk and Reid (1983), who have studied the PubMed data-base for consistency in indexing, the quality of indexing cannot be directly measured, as there is no right or wrong way to index an article or abstract. In turn, the issue of holding abstracts to ambiguous standards of correctness is a potential downfall.

4.2.5 A searcher bias in need of discussion

We are not proposing a standard; however, we are proposing a searcher bias in need of discussion. The main aim of this research is to facilitate the management of information produced by family doctors and to prepare it for further computerized development/reuse. The current version of this program is named Q-Codes, in honor of its creator Professor Lamberts, but it is still only a preliminary version (ver. 2.5). It will obviously evolve, and names will most likely change. But the need to manage GP/FM information in a structured and standardized way must remain a substantial facet of research. The future of the profession is at stake.

It is important to recognize that Q-Codes have been created from a limited number of abstracts. If a concept was not present in the read abstracts, it will have no place in the Q-codes. This emphasizes that the current program is limited to a small number of abstracts within the GP/FM field. Q-Codes would need to integrate much more information to be considered a fully applicable program to the field of GP/FM. Further conferences will contribute new concepts to this, while simultaneously helping GP/FM to evolve. We hope that the structure of this proposed taxonomy will remain enough strong to support the introduction of new items, but it must be taken into account that as more information is added, the basis could potentially not be strong enough to accommodate all.

One issue with the Q-codes ontology involves the unique identifiers. Cimino (1998) notes that when building an ontology, there is an *irresistible temptation to make the unique identifier a hierarchical code which reflects the concept's position in the hierarchy*. However, there are inherent disadvantages to using unique identifiers. The first issue, which we have encountered here, is that the coding system runs out of room to grow (Cimino, 1996; Cimino, 1998). This can be due to limited depth, limited breadth, or both of the unique identifier. For instance, when the code has a limited number of positions (digits), the depth of the hierarchy is limited.

Further ontological research is needed to determine whether the two main rules of taxonomic thinking have been respected: completeness (all identified) and exclusivity (a place for each concept) (Ittoo and Bouma, 2013).



4.2.6 Advantages and limits of The Semantic Web

The Q-Codes bases, like all the terminologies edited on HeTOP server, are fit for The Semantic Web. Semantic Web technologies promote common data formats and exchange protocols on the Web, like the Resource Description Framework (RDF), the cited OWL (now available OWL-2) and the query language SPARQL. We have seen that Linked Open Vocabulary (LOV) (<http://lov.okfn.org/dataset/lov/>), a lightweight ontology, differentiates from other ontologies through its characteristics that enable reuse and integration of other vocabularies-i.e., (i) small size, (ii) low formal constraints, (iii) few instances except for examples, (iv) rich user documentation. Labels, comments, definition, description, etc. are all characteristics of Q-Codes and ICPC-2 on the HeTOP server. We hope that in the near future, ICPC-2 and Q-Codes could find their place in Linked Open Vocabularies, that are published and used by actors in diverse media corporations like BBC, national administrations like INSEE, the European Community, universities and research projects. Or perhaps, we hope to see them reused and published by individuals and put on the community table. (Vandenbussche and Vatan, 2011).

Nevertheless semantic difficulties may arise from the supposed simplicity of the language. The relation *is a* is not an simple relation. Aristotelian logic, which decomposes the proposition into subject and predicate (Younes, 2016), is not sufficient in rendering reality. Following Wittgenstein, the relation *is a* has at least three semantic interpretations. As stated by Wittgenstein (1922) in *Tractatus Logico Philosophicus* (TLP 3.323): *In the language of everyday life it very often happens that the same word signifies in two different ways – and therefore belongs to two different symbols – or that two words, which signify in different ways, are apparently applied in the same way in the proposition. Thus the word "is" appears as the copula, as the sign of equality, and as the expression of existence [.....] In the proposition "Green is green" – where the first word is a proper name as the last an adjective – these words have not merely different meanings but they are different symbols.* Language could be more complex than its use in Health Information systems. As stated by Elish and Boyd (2017): *Because computational systems require precise definitions and mathematically sound logics, sociocultural phenomena that are typically nuanced and fuzzy are rendered in coarse ways when implemented into code.* Again, the last word will be given to Wittgenstein (TLP 4.002): *Language disguises the thought. So that from the external form of the clothes one cannot infer the form of the thought.*

5 CONCLUSION

Constructed on the basis of Semantic web technologies, Q-Codes could be considered as a lightweight ontology ready to be used in the semantic web domain, to be extracted in OWL. The multilingual classes of the classification could be individually reached through Unique Resource Identifiers (URIs). Note that each entry gives access to a detailed terminological description, mappings to other terminologies like Babelnet and DBpedia and to automatic queries on resources like PubMed.

We have created a terminology that highlights the vastness of GP/FM contributions to medical knowledge. We hope, by doing this, to contribute to the recognition of GP/FM as a professional entity within the scientific community that contributes heavily to all fields of medicine. Several question remain unsolved. Does the current extent of the knowledge base efficiently cover the GP/FM domain? How will this resist the hierarchical structure proposed to the introduction of new themes? Will this system retain enough inter-observers reliability? Nevertheless, given the number of contributions by volunteer translators, such an indexing system seems largely expected by the profession. We hope to transform it into a validated tool for its development.



REFERENCES

- AAFP (2011). *Primary Care definitions – AAFP Policies*. URL: <http://www.aafp.org/about/policies/all/primary-care.html> (visited on 02/07/2016).
- Adcock, Robert and David Collier (2001). "A Shared Standard for Qualitative and Quantitative Research". In: *American Political Science Review* 95.3, pp. 529–546. ISSN: 00030554. DOI: 10.1017/S0003055401003100. arXiv: arXiv:1011.1669v3.
- Allen, Bradley P (2016). "The role of metadata in the second machine age". In: *Second International Conference Establishment Surveys*. ISBN: 9788578110796. URL: <http://www.amstat.org/meetings/ices/2000/proceedings/S57.pdf>.
- Allen, Justin et al. (2011). "The European Definition Of General Practice / Family Medicine". In: URL: <http://www.woncaeurope.org>.
- Angell, Marcia (2017). "Drug Companies & Doctors: A Story of Corruption". In: Bachman, CW (1969). "Data structure diagrams". In: *SIGMIS Newsletter* 1.2, pp. 4–10. DOI: 10.1145/1017466.1017467. URL: <http://dl.acm.org/citation.cfm?id=1017467>.
- Bartholomeeusen, Stefaan, Frank Buntinx, and Jan Heyrman (2002). "Ziekten in de huisartspraktijk: methode en eerste resultaten van het Intego-netwerk". In: *Tijdschrift voor Geneeskunde* 58.863–871.
- Bentzen N.(ed) (2003). *WONCA Dictionary of general/family practice*. Ed. by Niels Bentzen. Maanedsskr. Copenhagen. URL: <http://www.ph3c.org/PH3C/docs/27/000092/0000052.pdf>.
- Berners-Lee, T, R Fielding, and Larry Masinter (1998). "Uniform Resource Identifiers (URI): Generic Syntax". In: *Request For Comments* 2396.
- Bowker, GC and SL Star (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.
- Bradley, Elizabeth H, Leslie A Curry, and Kelly J Devers (2007). "Qualitative data analysis for health services research: developing taxonomy, themes, and theory." In: *Health services research* 4, pp. 1758– 72. ISSN: 0017-9124. DOI: 10.1111/j.1475-6773.2006.00684.x.
- Britt, H et al. (2003). "Bettering the Evaluation And Care of Health 2001-2002 (summary of results)". In: *Australian Family Physician* 32.1/2, pp. 59–63. URL: <http://www.racgp.org.au/document.asp?id=8998>.
- Britt, Helena et al. (2016). *A decade of Australian general practice activity 2006–07 to 2015–16. Bettering the Evaluation and Care of Health (BEACH)*. Report. Family Medicine Research Centre. Sydney School of Public Health. University of Sydney. URL: https://ses.library.usyd.edu.au/bitstream/2123/15482/5/9781743325162{ }_ONLINE.pdf.
- Buono, Nicola et al. (2013). "40 years of biannual family medicine research meetings – The European General Practice Research Network (EGPRN)". EN. In: *Scandinavian Journal of Primary Health Care*. URL: <http://www.tandfonline.com/doi/full/10.3109/02813432.2013.847594>.
- Bushman, B., D. Anderson, and G. Fu (2015). "Transforming the Medical Subject Headings into Linked Data: Creating the Authorized Version of MeSH in RDF". In: *J Libr Metadata* 15.3-4, pp. 157–176. ISSN: 1938-6389 (Print)1937-5034. DOI: 10.1080/19386389.2015.1099967.
- Cabot, Chloé et al. (2017a). "Evaluation of the Terminology Coverage in the French Corpus LiSSa." In: *Studies in health technology and informatics* 235, pp. 126–130.
- Cardillo, Elena (2015). "Mapping between international medical terminologies to SHN Work Package 3". In: *SemanticHealthNet*. Chap. Deliverable 3.3, 18p.
- Carey, Iain M et al. (2004). "Developing a large electronic primary care database (Doctors' Independent Network) for research". In: *International Journal of Medical Informatics* 5, pp. 443–453. ISSN: 1386-5056. DOI: 10.1016/j.ijmedinf.2004.02.002.
- Casado Vicente, Verónica (2012). *Tratado de medicina de familia y comunitaria*. Ed. by Verónica Casado Vicente. Médica Panamericana, p. 2563. ISBN: 8498355850.
- Cavadas, L F, T Villanueva, and J Gervas (2010). "General practice innovation: a Portuguese virtual conference". In: *Med Educ* 44.5, pp. 514–515. ISSN: 0308-0110. DOI: 10.1111/j.1365-2923.2010.03649.x.
- Charlton, Rachel A et al. (2010). "Identifying major congenital malformations in the UK General Practice Research Database (GPRD)". In: *Drug Safety: An International Journal of Medical Toxicology and Drug Experience* 9, pp. 741–750. ISSN: 0114-5916. DOI: 10.2165/11536820-000000000-00000.
- Chinitz, David P and Victor G Rodwin (2014). "Perspective On Health Policy and Management (HPAM): mind the theory-policy-practice gap". In: *International Journal of Health Policy Management* 3.x, pp. 1–3. DOI: 10.15171/ijhpm.2014.122.
- Cimino, JJ (1996). "Review paper: codingsystems in health care." In: *Methods of information in medicine* 35.4-5, pp. 273–84. ISSN: 0026-1270. URL: <http://www.ncbi.nlm.nih.gov/pubmed/9019091>. — (1998). "Desiderata for controlled medical vocabularies in the twenty-first century". In: *Methods of Information in Medicine* 37.4-5, pp. 394–403.
- Colliers, Annelies et al. (2016). "Improving Care And Research Electronic Data Trust Antwerp (iCARE- data): a research database of linked data on out-of-hours primary care". In: *BMC Research Notes* 9.1, p. 259. ISSN: 1756-0500. DOI: 10.1186/s13104-016-2055-x.
- David, A.K. et al. (2013). *Family Medicine: Principles and Practice*. Springer Science & Business, p. 1240. ISBN: 0387217444.
- Davis, D. A. (2004). "CME and the pharmaceutical industry: two worlds, three views, four steps". In: *CMAJ*. Vol. 171, pp. 149–50. ISBN: 0820-3946 (Print)1488-2329 (Electronic). DOI: 10.1503/cmaj.1040361.
- Dekkers, Makx (2009). "History, objectives and approaches of the Dublin Core Metadata Initiative". In: December, pp. 1–24.



- Denzin, Norman and Yvonna S Lincoln (2000). "The Sage handbook of qualitative research (2nd ed.)" In: *Sage Publications*. Thousand Oaks: Sage., p. 784. ISBN: 9781412974172. DOI: Doi10.1177/ 1354067x07080505.
- Dixon, Brian E, Atif Zafar, and Julie J McGowan (2007). "Development of a taxonomy for health information technology." In: *Studies in health technology and informatics* 129.Pt 1, pp. 616–620. ISSN: 0926-9630.
- Dowden, John (2015). "Conflict of interest in medical journals". In: *Australian Prescriber* 38.1, pp. 2–3. DOI: 10.18773/austprescr.2015.001.
- Druais, PL et al. (2009). *Médecine générale*. Ed. by D Pouchain. Masson SA, p. 460.
- Eberbach, Andreas et al. (2016). "A simple heuristic for Internet-based evidence search in primary care: a randomized controlled trial". In: *Advances in Medical Education and Practice*, pp. 433–441. DOI: 10.2147/AMEP.S78385.
- Elish, M. C. and Danah Boyd (2017). "Situating methods in the magic of Big Data and AI". In: *Communication Monographs*, pp. 1–24. ISSN: 0363-7751. DOI: 10.1080/03637751.2017.1375130.
- Faggiolani, Chiara (2011). *Perceived Identity: applying Grounded Theory in Libraries*. it. DOI: 10.4403/jlis.it-4592.
- Farace, Dominic John and Joachim Schöpfel. (2010). "Collection building with special regards to Report Literature" In: *Grey Literature in Library and Information Studies*. Ed. by Walter de Gruyter,
- Ferreras Fernández, Tránsito (2016). "Visibilidad e impacto de la literatura gris científica en repositorios institucionales de acceso abierto. Estudio de caso bibliométrico del repositorio Gredos de la Universidad de Salamanca". PhD thesis. URL: <https://gredos.usal.es/jspui/handle/10366/132444>.
- Friedman, Carol et al. (1999). "Representing Information in Patient Reports Using Natural Language Processing and the Extensible Markup Language". In: *Journal of the American Medical Informatics Association* 6.1, pp. 76–87. URL: <http://www.jamia.org/cgi/content/abstract/6/1/76>.
- Funk, M E and C A Reid (1983). "Indexing consistency in MEDLINE." In: *Bulletin of the Medical Library Association* 71.2, pp. 176–83. ISSN: 0025-7338. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC227138/>.
- Gill, P. J. et al. (2014). "Development of a search filter for identifying studies completed in primary care". In: *Fam Pract* 31.6, pp. 739–45. ISSN: 0263-2136. DOI: 10.1093/fampra/cmu066. URL: <http://dx.doi.org/10.1093/fampra/cmu066>.
- Glaser, Barney G. and Anselm L. Strauss (1999). *Discovery of Grounded Theory: Strategies for Qualitative Research -- Livres*. AldineTransaction, p. 284. ISBN: 0202302601.
- Goggi, S. et al. (2015). "A semantic engine for grey literature retrieval in the oceanography domain". In: *Seventeenth International Conference on Grey Literature*. Amsterdam, pp. 76–77. URL: <http://greyguide.isti.cnr.it/wp-content/uploads/2017/04/GL17\Program\Book-min.pdf>.
- González-González, A I et al. (2007). "Information Needs and Information-Seeking Behavior of Primary Care Physicians". In: *Ann Fam Med* 4, pp. 345–352. ISSN: 1544-1709 (Print)1544-1717 (Electronic). DOI: 10.1370/afm.681.
- Gotzsche, Peter G (2013). *Deadly Medicines and Organised Crime: How Big Pharma Has Corrupted Healthcare*. Radcliffe Publishing Ltd.
- Greenway, Tyler and Joseph S Ross (2017). "US drug marketing: how does promotion correspond with health value?" In: DOI: <https://doi.org/10.1136/bmj.j1855>.
- GreyNet (2014). *Pisa Declaration on Policy Development for Grey Literature Resources*. URL: <http://greyguiderep.isti.cnr.it/Pisadeclapdf/Pisa-Declaration-May-2014.pdf> (visited on 06/16/2017).
- Grosjean, Julien et al. (2012). "Teaching medicine with a terminology/ontology portal." eng. In: *Studies in health technology and informatics* 180, pp. 949–53. URL: <http://europepmc.org/abstract/MED/22874333>.
- Gusso, Gustavo and José Mauro Ceratti Lopes (2012). *Tratado de Medicina de Família e Comunidade: 2 Volumes: Princípios, Formação e Prática*, p. 2180. ISBN: 8536327979.
- Gutierrez, Cecilia and Peter Scheid (2002). "The History of Family Medicine and Its Impact in US Health Care Delivery". In: *AAFP Foundation*, pp. 1–31.
- Gómez-Pérez, A., M. Fernández-López, and O. Corcho (2003). "Ontological Engineering and the Semantic Web". URL: http://www.exa.unicen.edu.ar/escuelapav/cursos/corcho/01_introduction.pdf.
- Heilman, J (2015). "Point of care Information in Open Access: A Time to Sow?" In: *PLOS Medicine* 12.8, e1001870. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001870.
- Helman, Cecil G (2008). *Medical Anthropology*. Ashgate, p. 580. ISBN: 978-0-7546-2655-8.
- Heyrman, J (ed.) (2005). *EURACT Educational Agenda, European Academy of Teachers in General Practice EURACT*, Leuven.
- Hoffmann, Kathryn et al. (2011). "Antibiotic resistance in primary care in Austria - a systematic review of scientific and grey literature". In: *BMC Infectious Diseases* 11.1, p. 330. ISSN: 1471-2334. DOI: 10.1186/1471-2334-11-330.
- Holden, Ronald B. (2010). "Face validity". In: *The Corsini encyclopedia of psychology*. Ed. by Irving B. Weiner and W. Edward Craighead. Wiley, pp. 637–638. ISBN: 9780470170267.
- Hong, Y. et al. (2016). "Knowledge structure and theme trends analysis on general practitioner research: A Co-word perspective". In: *BMC Fam Pract* 17, p. 10. ISSN: 1471-2296. DOI: 10.1186/s12875-016-0403-5.
- Hopewell, S et al. (2007). "Grey literature in meta-analyses of randomized trials of health care interventions." In: *Cochrane database of systematic reviews (Online)* 2, MR000010. ISSN: 1469-493X. DOI: 10.1002/14651858.MR000010.pub3.
- Huang, Minlie, Aurélie Névéal, and Zhiyong Lu (2011). "Recommending MeSH terms for annotating



- biomedical articles." In: *Journal of the American Medical Informatics Association : JAMIA* 18.5, pp. 660–7. ISSN: 1527-974X. DOI: 10.1136/amiajnl-2010-000055.
- Hubbard, Derek (2008). *How to Find Clinical Information Quickly at the Point of Care*. Vol. 15. 6. American Academy of Family Physicians, p. 23.
- Hummers-pradier, Eva (2007). "Which Abstracts Do Get Published ? – Output Of German Gp Research 1999-2003". In: *Wonca Europe Paris 2007*.
- Ittoo, Ashwin and Gosse Bouma (2013). "Term extraction from sparse, ungrammatical domain-specific documents". In: *Expert Systems with Applications* 40, pp. 2530–2540.
- James, Jack E (2016). "Free-to-publish, free-to-read, or both? Cost, equality of access, and integrity in science publishing". In: *Journal of the Association for Information Science and Technology* 68.6, pp. 1584–1589. ISSN: 2330-1643. DOI: 10.1002/asi.23757.
- Jamoulle, M et al. (2014). "Mapping French terms in a Belgian guideline on heart failure to international classifications and nomenclatures: the devil is in the detail". eng. In: *Inform Prim Care* 21.4, pp. 189–198. DOI: 10.14236/jhi.v21i4.66.
- Jamoulle, Marc (2015). "Quaternary prevention, an answer of family doctors to overmedicalization". In: *International Journal of Health Policy and Management* 4.2, pp. 61–64. ISSN: 2322-5939. DOI: 10.15171/ijhpm.2015.24. URL: http://ijhpm.com/article/{_}2950{_}0.html.
- Jamoulle, Marc, Julien Grosjean, and Stefan Darmoni (2017). "Access to multilingual individual rubrics in URI format for ICP-2 and the Q-Codes". In: URL: <http://orbi.ulg.ac.be/handle/2268/211268>.
- Jamoulle, Marc et al. (2015). "Semantic Web and the Future of Health Care Data in Family Practice". In: *Merit Research Journal of Medicine and Medical Sciences* 3.12, pp. 586–594. URL: <http://orbi.ulg.ac.be/handle/2268/189292>.
- Jamoulle, Marc et al. (2017a). "A terminology in General Practice / Family Medicine to represent non-clinical aspects for various usages : the Q-Codes". In: *Medical Informatics Europe (MIE2017) Informatics for Health 2017 / April*, pp. 1–5. URL: <http://orbi.ulg.ac.be/handle/2268/206527>.
- Jamoulle, Marc et al. (2017b). "Analysis of definitions of General Practice/Family Medicine and Primary Health Care". In: *British Journal of General Practice - Open*, 050 ISSN: 0960-1643. URL: <http://orbi.ulg.ac.be/handle/2268/210049>.
- Janamian, T et al. (2016). "Quality tools and resources to support organisational improvement integral to high-quality primary care: a systematic review of published and grey literature". In: *Med J Aust* 204.7 Suppl, S22–8. ISSN: 0025-729x.
- Jelercic, S. et al. (2010). "Assessment of publication output in the field of general practice and family medicine and by general practitioners and general practice institutions". In: *Fam Pract* 27.5, pp. 582–9. DOI: 10.1093/fampra/cm032.
- Jonquet, Clement et al. (2016). "SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique". In: *16e Journées Francophones d'Informatique Médicale(JFIM)* June.
- Khan, Nada F, Sian E Harrison, and Peter W Rose (2010). "Validity of diagnostic coding within the General Practice Research Database: a systematic review". In: 60.572, e128–e136. ISSN: 0960-1643. DOI: 10.3399/bjgp10X483562.
- Kochen, Michael M. (2012). *Allgemeinmedizin und Familienmedizin*. Thieme; Auflage: 4., vollständig überarbeitete und erweiterte Auflage, p. 652. ISBN: 3131413840.
- Lakhani, Mayur K. (2003). *A Celebration of General Practice*. Radcliffe Publishing, p. 203. ISBN: 1857759230.
- Lawrence, Amanda et al. (2014). "Where is the evidence: realising the value of grey literature for public policy and practice". In: *Australian Policy Online*. DOI: 10.4225/50/5580B1E02DAF9. URL: <http://apo.org.au/node/42299>.
- Lelong, R et al. (2016). "Semantic Search Engine to Query into Electronic Health Records with a Multiple-Layer Query Language". In: *MEDIR workshop*. URL: http://medir2016.imag.fr/data/MEDIR_2016_paper_8.pdf.
- Liang, S. F. et al. (2014). "Semi Automated Transformation to OWL Formatted Files as an Approach to Data Integration". In: *Methods of Information in Medicine* 54.1, pp. 32–40. ISSN: 0026-1270. DOI: 10.3414/ME13-02-0029.
- Library and Archives Canada (2017). *Canadian Subject Headings*. URL: <http://www.bac-lac.gc/>.
- Lin, Jennifer and Carly Strasser (2014). "Recommendations for the role of publishers in access to data." In: *PLoS biology* 12.10, e1001975. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001975.
- Lowe, H J and G O Barnett (1994). "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches." In: *JAMA* 271.14, pp. 1103–8. URL: <http://www.ncbi.nlm.nih.gov/pubmed/8151853>.
- Lusignan, Simon de (2005). "Codes, classifications, terminologies and nomenclatures: definition, development and application in practice". In: *Informatics in primary care* 13.1, pp. 65–70. URL: <http://www.ncbi.nlm.nih.gov/pubmed/15949178>.
- Madkour, Mohcine, Driss Benhaddou, and Cui Tao (2016). "Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain". In: *Computer Methods and Programs in Biomedicine*, pp. 52–68. DOI: 10.1016/j.cmpb.2016.02.007.
- Mahood, Quenby, Dwayne Van Eerd, and Emma Irvin (2014). "Searching for grey literature for systematic reviews: challenges and benefits". In: *Research synthesis methods* 5.3, pp. 221–34. DOI: 10.1002/jrsm.1106.
- Marc, D. T. et al. (2015). "Indexing Publicly Available Health Data with Medical Subject Headings (MeSH): An Evaluation of Term Coverage". In: *Stud Health Technol Inform* 216, pp. 529–33. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26262107>.



- Martin, Patricia and Barry A. Turner (2016). "Grounded Theory and Organizational Research". In: <http://dx.doi.org/10.1177/002188638602200207>. DOI: 10.1177_002188638602200207.
- McGuinness, DL and Frank van Harmelen (2004). *Web Ontology Language*. URL: <http://www.w3.org/TR/owl-features/>.
- McIntyre, Ellen et al. (2016). "The contribution of a knowledge exchange organisation in primary healthcare." In: *Australian family physician* 45.9, pp. 684–7. ISSN: 0300-8495. URL: <http://www.ncbi.nlm.nih.gov/pubmed/27606374>.
- McKenzie, Lindsay (2017). "Sci-Hub's cache of pirated papers is so big, subscription journals are doomed, data analyst suggests". In: *Science*. ISSN: 0036-8075. DOI: 10.1126/science.aan7164.
- McWhinney, Ian R. (1997). *A Textbook of Family Medicine*. Oxford University Press, p. 448. ISBN:019511518X.
- Mendis, Kumara and Indragit Solangaarachchi (2005). "PubMed perspective of family medicine research: Where does it stand?" In: *Family Practice* 22.5, pp. 570–575. ISSN: 02632136. DOI: 10.1093/fampra/cmi085.
- Miller, Eric (1998). "An Introduction to the Resource Description Framework". In: *Bulletin of the Association for Information Science and Technology* 25.1, pp. 15–19. ISSN: 1550-8366. DOI: 10.1002/bult.105. URL: <http://onlinelibrary.wiley.com/doi/10.1002/bult.105/abstract>.
- Moher, David et al. (2000). "Mejora de la calidad de los informes de los metaanálisis de los ensayos clínicos controlados: el acuerdo QUOROM". In: *Rev. Esp. Salud Pública* 74.2. URL: <http://scielo.isciii.es/pdf/resp/v74n2/mejora.pdf>.
- Moynihan, R (2003). "Who pays for the pizza? Redefining the relationships between doctors and drug companies. 1: entanglement". In: *Bmj* 7400, pp. 1189–1192. ISSN: 0959-535x. DOI: 10.1136/bmj.326.7400.1189.
- Moynihan, R and L Bero (2017). "Toward a Healthier Patient Voice: More Independence, Less Industry Funding". In: *JAMA Intern Med* 177.3, pp. 350–351. ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2016.9179.
- Murtagh, John (2011). *John Murtagh's General Practice*. McGraw-Hill Medical Publishing Division, p. 1508. ISBN: 0070285381.
- Myška, Matěj and Jaromír Šavelka (2013). "A Model Framework for Publishing Grey Literature in Open Access Defining Grey Literature". In: *Open Access 4.JIPITEC 2*, para 104, pp. 1–12.
- Neghme, A. (1975). "Operations of the Biblioteca Regional de Medicina (BIREME)". In: *Bull Med Libr Assoc* 63.2, pp. 173–9. ISSN: 0025-7338 (Print).
- Noble, Helen and Joanna Smith (2015). "Issues of validity and reliability in qualitative research". In: *Evidence-Based Nursing* 18.2, pp. 34–35. ISSN: 1367-6539. DOI: 10.1136/eb-2015-102054. URL: <http://ebn.bmj.com/cgi/doi/10.1136/eb-2015-102054>.
- PAHO Bireme Sao Paulo (2016). *Virtual Health Library Search Portal of the Regional library of Medicine*. URL: <http://regional.bvsalud.org/http://www.paho.org/bireme/>.
- Pingitore, D and R A Sansone (1998). "Using DSM-IV primary care version: a guide to psychiatric diagnosis in primary care." In: *American family physician* 58.6, pp. 1347–52. URL: <http://www.ncbi.nlm.nih.gov/pubmed/9803199>.
- PLoS Medicine Editors (2015). "Point of care Information in Open Access: A Time to Sow?" In: *PLoS medicine* 12.8, e1001870. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001870.
- Quan, Wei, Bikun Chen, and Fei Shu (2017). "Publish or impoverish: An investigation of the monetary reward system of science in China (1999-2016)". In: *ArXiv e-prints*. arXiv: 1707.01162. URL: <http://arxiv.org/abs/1707.01162>.
- Resnick, M. P., Santana, F., de Araujo Novaes, M., Shamenek, F. S., Frieden, L., & Iyengar, M. S. (2013). Representing second opinion requests from primary care within the Brazilian tele-health program: international classification of primary care, second edition. *Studies in Health Technology and Informatics*, 192, 1190. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23920964>
- Rigaux, Sébastien (2015). "Classification automatisée de résumés médicaux–Belgique 2015 [Multi-Label Text Classification of Medical Abstracts]". Master thesis.
- Salman Bin Naeem, Salman, Ahmed Shamshad, and Khan Amjid (2013). "Information seeking in primary care: a survey of doctors working in remote government health facilities in Pakistan". In: *Library Philosophy and Practice (e-journal)* Paper 1009. URL: <http://digitalcommons.unl.edu/libphilprac/1009>.
- Schöpfel, Joachim (2015). "Littérature << grise >> : de l'ombre à la lumière". In: *I2D – Information, données & documents*. Vol. Volume 52. 1. Chap. Introduction, pp. 28–29. URL: <https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-1-page-28.htm>.
- Schrans, D. et al. (2016). "The search for person-related information in general practice: a qualitative study". In: *Fam Pract* 33.1, pp. 95–9. ISSN: 0263-2136. DOI: 10.1093/fampra/cmv099.
- Schuers, Matthieu et al. (2015). "Mise en ligne de la CISP en près de 20 langues au sein d'un portail terminologique de santé". In: *Congrès de la Médecine Générale, Paris, 26 et 27 mars 2015*, poster. URL: <http://orbi.ulg.ac.be/handle/2268/207421>.
- Schwitzer, Gary (2017). *Conflicts of interest in health care journalism. Who's watching the watchdogs? We are. Part 1 of 3 - HealthNewsReview.org*. URL: <https://www.healthnewsreview.org/2017/06/conflicts-of-interest-in-health-care-journalism-1-of-3/> (visited on 06/16/2017).
- Shen, Cenyu and Bo-Christer Björk (2015). "'Predatory' open access: a longitudinal study of article volumes and market characteristics". In: *BMC Medicine* 13.1, p. 230. ISSN: 1741-7015. DOI: 10.1186/s12916-015-0469-2.
- Shultz, M. (2007). "Comparing test searches in PubMed and Google Scholar". In: *J Med Libr Assoc* 95.4, pp. 442–5. ISSN: 1536-5050 (Print) 1558-9439 (Electronic). DOI: 10.3163/1536-5050.95.4.442.



- Silva, Caio da, Regina Garcia, and Rita Bonadio Inacio de Cássia (2009). "Literatura Cinzenta : teses , eventos e relatórios". Thesis. URL: <http://rabci.org>.
- Silver, Christina and Ann Lewins (2014). *Using Software in Qualitative Research*. SAGE Companion. URL: <https://study.sagepub.com/using-software-in-qualitative-research>.
- Simon, C. (2009). "From generalism to specialty—a short history of General Practice". In: *InnovAiT* 2.1, pp. 2–9. ISSN: 1755-7380. DOI: 10.1093/innovait/inn171.
- Sladek, Ruth et al. (2006). "Development of a subject search filter to find information relevant to palliative care in the general medical literature." In: *Journal of the Medical Library Association : JMLA* 94.4, pp. 394–401. URL: <http://www.ncbi.nlm.nih.gov/pubmed/17082830>.
- Soler, Jean karl, Marc Jamoulle, and Peter Schattner (2015). "The International Classification of Primary Care". en. In: *The World Book of Family Medicine – European Edition 2015*. Ljubljana. ISBN: 978-961-281-983-5. URL: <http://orbi.ulg.ac.be/handle/2268/187050>.
- Soler, Jean K. et al. (2012). "The interpretation of the reasons for encounter 'cough' and 'sadness' in four international family medicine populations". In: *Informatics in Primary Care* 20, pp. 25–39. ISSN: 14760320.
- Sutton, Stuart A (2007). "Tutorial 1: Basic Semantics". In: August. URL: <http://dublincore.org/resources/training/dc-2007/T1-BasicSemantics.pdf>.
- Swan, Alma. (2012). *Policy guidelines for the development and promotion of open access*. United Nations Educational, Scientific, and Cultural Organization, p. 76. ISBN: 9789230010522.
- Tabatabaei-Malazy, O, S Nedjat, and R Majdzadeh (2012). "Which information resources are used by general practitioners for updating knowledge regarding diabetes?" In: *Arch Iran Med* 4, pp. 223–227. ISSN: 1029-2977. DOI: 012154/aim.0010.
- Tan, Sharon Swee-Lin and Nadee Goonawardene (2017). "Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review". In: *Journal of Medical Internet Research* 19.1, e9. ISSN: 1438-8871. DOI: 10.2196/jmir.5729.
- Thompson, C. A. et al. (2014). "Patient and provider characteristics associated with colorectal, breast, and cervical cancer screening among Asian Americans". In: *Cancer Epidemiol Biomarkers Prev*. Vol. 23. United States: (c)2014 American Association for Cancer Research., pp. 2208–17. ISBN: 1538-7755 (Electronic)1055-9965 (Linking). DOI: 10.1158/1055-9965.epi-14-0487.
- Ustün, T B et al. (1995). "New classification for mental disorders with management guidelines for use in primary care: ICD-10 PHC chapter five." In: *The British journal of general practice : the journal of the Royal College of General Practitioners* 45.393, pp. 211–5. ISSN: 0960-1643. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1239204/>.
- Van Royen, Paul et al. (2010). "Are presentations of abstracts at EGPRN meetings followed by publication?" EN. In: *The European journal of general practice* 16.2, pp. 100–5. ISSN: 1751-1402. DOI: 10.3109/13814788.2010.482582.
- Vandenbussche, Py and Bernard Vatan (2011). "Metadata recommendations for linked open data vocabularies". In: *Version*. URL: http://lov.okfn.org/dataset/lov/Recommendations{_}Vocabulary{_}Design.pdf.
- VanNieuwenborg, L et al. (2016). "Continuing medical education for general practitioners: a practice format". In: *Postgrad Med J* 92.1086, pp. 217–222. ISSN: 0032-5473. DOI: 10.1136/postgradmedj-2015-133662.
- Vanopstal, Klaar et al. (2011). "Vocabularies and retrieval tools in biomedicine: disentangling the terminological knot." In: *Journal of medical systems* 35.4, pp. 527–43. ISSN: 0148-5598. DOI: 10.1007/s10916-009-9389-z.
- Veuillotte, I et al. (2015). "General practice and the Internet revolution. Use of an Internet social network to communicate information on prevention in France". In: *Health Informatics J* 1, pp. 3–9. ISSN: 1460-4582. DOI: 10.1177/1460458213494905.
- Wells, K (1995). "The strategy of grounded theory: possibilities and problems." In: *Social work research* 19.1, pp. 33–7. ISSN: 1070-5309. URL: <http://www.ncbi.nlm.nih.gov/pubmed/10140997>.
- Wittgenstein, Ludwig (1922). *Tractatus logico-philosophicus*. London: Kegan Paul, p. 108. ISBN: 1602064512. URL: <http://people.umass.edu/klement/tlp/>.
- WONCA (1987). *ICPC: International Classification of Primary Care*. Ed. by H Lamberts and M Wood. Oxford: Oxford University Press. ISBN: 0 19 261633 1.
- WONCA (2005). *ICPC-2-R: International Classification of Primary Care (Oxford Medical Publications)*. Oxford University Press, USA, p. 204. ISBN: 019262802X.
- Younes, Nora (2016). *Introduction à Wittgenstein*. Paris: Editions La Découverte, p. 11



Preserving and accessing content stored on USB-flash-drives: A TIB workflow

Oleg Nekhayenko

German National Library of Science and Technology, TIB, Germany

Abstract

Grey Literature has become a corner stone of the supply of scientific and academic literature. TIB is the German National Library of Science and Technology with a strong focus on Grey Literature. The acquisition of this kind of literature is not a simple process due to the unconventional distribution channels and the cost. For this purpose TIB has its own grey literature acquisition team. In order to keep the data readable and accessible in the long term TIB hosts, administrates and runs a digital preservation system which manages different kind of objects and file formats. Each group of objects needs its own workflow and ingest preparation (pre-ingest). This also applies to grey literature.

Some of the grey literature publications are delivered on USB-flash-drives. In order to catalog them the USB-flash-drive will be accessed by librarians several times. As USB-flash-drives are "write-many" storage devices, accidental write or erase processes could already take place during the access by initially connecting with a PC. This fact leads to the risk of loss or overwriting of rare and expensive data. Furthermore, the lifespan of this data storage and the interface dependency are two remaining basic concerns of their long term accessibility in libraries. The minimization of risks of accidental manipulation by direct access to the USB-flash-drives can be achieved through the use of a USB write blocker. This device prevents all write instructions and ensures the safe forensic data transfer. The secure storage of the data from the USB-flash-drive can be done through the separation of data from the storage device. For this purpose TIB has developed a reusable, semi-automated workflow with a strong focus on usability for librarians. Using a script and the write blocker, that prevents the accidental write accesses, the workflow saves the data from USB flash devices via imaging or copying.

This project consists of a market survey (4 person-days), implementation (28 person-days), tests (4 person-days) and launch in team (1 person-days). The price of the write blocker is 500\$.

The main focus of the paper is the detailed description of the project including its technical implementation, discussion of the results and further improvements.

1. Introduction

Nowadays data delivery and storage of grey literature on USB-flash-drives isn't unusual in many libraries anymore. However, USB-Sticks show some risks regarding their usage and storage in the libraries over time. Because the life expectancy of this data carrier is limited and the USB interface underlies technological development the digital objects on this carrier could be sometimes no longer accessible in their original condition after some years. The hardware dependency of USB-Sticks and the software dependency of file formats stored on it are the main obstacles to make valuable scientific information accessible in the long term (Scott, 2013). Recent developments of standards for new USB-Interfaces such as USB-C can replace the old standards (Pullen, 2015). This hardware change leads to obsolescence of the USB-Port of type A and as a consequence the data on the old USB-Sticks becomes unreadable because of lack of the old interface. Furthermore, new data transfer technologies such as Bluetooth, NFC, Wi-Fi Direct and AirDrop that do what USB does, but without wires, will be the main obstacle for the further maintaining of USB-Interface (Cunningham, 2014). As for the data on the USB-Stick itself, the file system and format dependency on specific operating system or software leads to obsolescence of these files if the original operating system or software doesn't exist anymore. Due to the fact that there is no write protection on USB-Sticks delivered to the library the probability of accidental erase of data during the connection to PC exists. Considering all these possible risks of data loss the only way for data rescue is it's separation from the storage device. The project described in this paper deals with a development of a content preservation strategy for USB-Sticks and focuses on the first step of data preparation as a pre-ingest process before transferring the data into a digital preservation system. A special workflow was developed for librarians from the grey literature team at TIB, which is aimed to separate the data from USB-Stick guaranteeing safe forensic data transfer. With the help of the developed script the data will be analyzed, transferred, structured and prepared for ingest into digital preservation system.



Organizational and technical processes implemented in the digital preservation system, which is hosted, operated and administered at the TIB, keep the data readable and accessible in the long term. TIB received the Data Seal of Approval in 2015 and therefore is certified as a trustworthy digital archive. The data files transferred into this system will be identified, validated, run through integrity checks and saved with their technical descriptive and administrative metadata in the trustworthy archive. Via ongoing preservation planning which includes monitoring of technology changes and community watch, the system checks regularly if the file formats remain accessible. When required the data can be migrated in a new format or emulated¹.

2. Background and related work

According to the current knowledge of the author there are no related projects that deal with separation of data from USB-flash-drives. However, some projects such as by the British Library (Dappert, Jackson & Kimura, 2013), Library of Congress (Lazorchak, 2012) and German National Library of Science and Technology (TIB) are dedicated to rescue the data from CD-ROMs. Optical carrier migration projects differ from the USB-Stick project in amount of the processed data carriers. In the case of CDs and DVDs a manual handling process is extremely time-consuming. For that reason a special disc copying robot to change the CD-ROMs and appropriate copying software are used to automate the workflow in such projects. A robot, which allows to queue up to 100 CD-ROMs at once, is used for the CD-Imaging project at the TIB. The main difficulties occur when the automated process encounters defective CD-ROMs. Defective optical carriers can cause the copying software to freeze. In order to check the copying process and guarantee the correct allocation of CD-ROMs to their catalogue number an additional control script was written. The British Library conducted a similar optical carrier migrating project and also had to create self-written additional software for the identification of problematic situations where the disc copying robot's software couldn't automatically recognize the CD types and process different stabilization activities depending on them (Dappert, Jackson & Kimura, 2013).

Regarding the USB-flash-drives standard software for data transfer from USB-sticks also exists. However, customization possibilities of this software for the needs of librarians are very poor or not available at all. The current project tries to close this gap and develops a data transfer script for USB-Sticks with a strong focus on usability for librarians. Guided user dialog will support the user during the entire copying process.

3. USB write blocker

The first access to a delivered USB-Stick from a PC takes place during the common file check and cataloging process by a librarian. As previously mentioned the USB-Sticks are commonly not write-protected. For that reason the risk of accidental and unnoticed write or erase processes exists. In order to minimize these risks a USB write blocker will be attached between the USB-Stick and the PC. USB write blockers are commonly known within the IT forensic context. "The intent of the write-blocker is to prevent the forensic workstation's software or operating system from making any inadvertent changes to the original media, including adding, deleting, or modifying any information" (Kessler & Karlton, 2014, p. 51). In other words it blocks the write access to the storage device and allows the access only in Read-Only Mode. In this way not a single bit will be changed on the USB storage device.

There are two different types of USB write blockers available: hardware and software. Both of them are intended to serve the same purpose. A software write blocker is a layer of software that is located between the operating system and the device driver for the disc. A hardware device is plugged in between the disc controller and the physical disc (Solomon, Rudolph, Tittel, Broom, Barret, 2011). In the described project the author conducted a short market survey of the available write blockers to find the most appropriate one for applying in the library context. The main features of four hardware and four software write blockers were compared and one of them which fulfills all the requirements was chosen. The list of requirements consists of the following: easy handling without additional configuration effort, common model in the IT-forensic, quick data transfer via USB 3.0, compatibility with old USB interfaces and most operating systems. Literature research showed that software write blockers aren't always reliable and stable. Occasionally they need additional control to check if the protection is still on. „System updates by OS vendor, configuration tweaks by the examiner, and additionally installed software all create a risk of disabling, overwriting, bypassing, or causing the failure of write-blocking

¹ <http://www.exlibrisgroup.com/de/category/Rosetta>



functionality implemented in software“ (Nikkel, 2016, p. 97). Table 1 contains a list of pros and cons of hardware and software write blockers.

Table 1: Pros and cons of software and hardware write blockers

	Hardware write blocker		Software write blocker
+	“Hardware devices that write block also provide visual indication of function through LEDs and switches. This makes them easy to use and makes functionality clear to users” ² .	-	“Software methods are not as simple, repeatable and idiot-proof as the hardware solution” (Smith, 2014)
+	“Some courts view Hardware write blockers as more secure than software write blockers because a physical connection blocks any other paths to the disc” (Solomon, Rudolph, Tittel, Broom, Barret, 2011)	-	“Although software approach is generally safe some software write blockers allow direct disc access in some cases” (Solomon, Rudolph, Tittel, Broom, Barret, 2011)
+	“Most hardware write blockers are software independent” ³	-	“Some software write blockers are designed for a specific operating system. One designed for Windows will not work on Linux” ³
-	“Disk imaging using hardware write blockers is slowed considerably due to protocol translations that the device must perform” ⁴	-	“For software write blockers you need to ensure that your tool of choice is updated to the latest version” (Solomon, Rudolph, Tittel, Broom, Barret, 2011)
-	Expensive	+	sometimes freeware, but not always reliable

Two freeware software write blockers⁵ can’t absolutely guarantee their forensic write blocking ability and another two commercial write blockers⁶ support many interfaces and have many settings which affect the software usability. Considering the mentioned negative aspects of software write blockers only hardware write blockers were shortlisted (table 2).

Table 2: Main features of hardware write blockers

Name/Model	Price	OS Windows	OS Linux	USB 1.1	USB 2.0	USB 3.0	Common model in the IT-forensic	Easy handling
CRU WiebeTech USB WriteBlocker ⁷	198 euro	yes (from XP)	yes	yes	yes	yes	yes	yes
Tableau Forensic USB 3.0 Bridge ⁸	459\$	yes (from 7)	yes	yes	yes	yes	yes	yes
CRU WiebeTech USB 3.0 WriteBlocker ⁹	398 euro	yes (from 7)	no	No information	No information	yes	yes	yes
EPOS BadDrive Adapter USB ¹⁰	550\$	yes (from XP)	yes	No information	yes	No Information	no	yes

Finally after weighting pros and cons the decision was made for the simplest hardware write blocker „CRU WiebeTech USB“. Due to the fact that this device couldn’t be delivered at that time, it was decided to purchase another write blocker „Tableau T8u Forensic USB 3.0 Bridge“. This device fulfils all necessary requirements and additionally has a small screen on which the main information about USB-Stick is shown. The price of this write blocker is 459\$. Figure 1 shows „Tableau T8u Forensic USB 3.0 Bridge“ with attached USB-Stick on the right side. The USB connection to the PC is located on the left side and the power supply is on the top.

² <https://www.cru-inc.com/data-protection-topics/write-blockers/>

³ http://www.forensicswiki.org/wiki/Write_Blockers

⁴ <https://dfcsc.uri.edu/research/swb>

⁵ USB Write Blocker for ALL Windows; Thumbscrew: Software USB Write Blocker

⁶ DRPU USB data theft protection tool; Safe Block XP

⁷ https://www.cru-inc.com/products/wiebetech/usb_writeblocker/

⁸ <https://www.guidancesoftware.com/tableau/hardware/t8u>

⁹ <https://www.cru-inc.com/products/wiebetech/usb-3-0-writeblocker/>

¹⁰ http://www.epos.ua/view.php/en/products_epos_baddrive_usb



Figure 1: Tableau T8u Forensic USB 3.0 Bridge

4. Technical implementation of the script

A reusable, semi-automated workflow using a script and the write blocker was developed in the project to save the data from USB flash devices via imaging or/and copying. The script is written in the Batch programming language for Windows and uses two Unix commands `dd` and `rsync` respectively for making images and copying all the files from USB flash device. A USB-Stick image is a file containing the contents and structure of the original USB-Stick. The command `dd` creates raw images with `.img` filename extension which are basically a bit for bit copy of the raw data of the USB-Stick without adding or deleting anything. The only requirement to use the script is the installation of Cygwin- a Unix-like environment which contains these commands and allows using them under Windows. The execution of the imaging or copying process is dependent on the initial analysis of the content on USB-Stick. After launching the script automatically recognizes which USB port the USB-Stick is connected to and searches for executable files with the extension `.exe`. The existence of `.exe` files on the USB-Stick indicates that there are possible connections between files on it. In order to ensure the integrity of these connections the imaging process will be started. Additionally after that the file copying process will be launched. This is needed to let the digital preservation system afterwards identify and validate each file from the USB-Stick, because the individual files from the raw image can't be recognized. If no executable files were detected, only the file copying process will take place. There's no need to create an image if only some independent PDF files exist on the USB-Stick. In this case the imaging of the entire USB-Stick leads to the waste of memory space as the imaging process will always image the entire stick, regardless of how little of the memory is effectively used. For example, if on a 4 GB size stick only 500 MB are filled with data such as PDF files, the image will still be 4 GB large. It should be noticed, that the identification of executable files based on the file extensions isn't very effective due to the fact, that file extensions not always provide the correct information about the file format. Furthermore, `.exe` files are not the only ones which hint at dependencies of the files on each other – `html` objects with `css` files are another example. Format identification tool such as Jhove can be built into the next version of the script to provide the trustworthy identification of the execution files. Also, further types of file formats will be included in the script.

4.1 Error control

Three control mechanisms are implemented in the script. If any errors occur during the imaging or the copying processes an appropriate entry with the catalogue ID of the USB-Stick will be made in a log file. In order to ensure that the image isn't different from the original USB-Stick, their md5 hashes will be calculated and compared. The md5 hash is a digital fingerprint of a file and is used to verify if any changes caused by inaccurate imaging occurred to the file. If the md5 hashes are not matching, the appropriate entry with the catalogue ID of the USB-Stick will be captured in a log file. All copied files are also listed in a log file.

5. Workflow

The author will now describe the workflow from the user view.

The user attaches the USB-Stick to the write blocker and randomly controls the files on it in order to exclude a defective USB-Stick in advance. In the second step he or she starts the script and is asked to enter the catalogue ID of the USB-Stick. The script creates a folder which is named after the entered catalogue ID. Later by ingesting the data into the digital preservation system it will be automatically enriched with descriptive metadata from the library catalogue on the basis of this catalogue ID. If some executable files were found, the script informs the user about their amount

and creates an image in the folder %catalogue ID%/Image. If the imaging process was successfully finished, the script informs the user with a message “The image was successfully created”. After the successful imaging process the script calculates the md5 hashes of the image and the original and informs the user about the result. Finally a file copying process starts and copies all files into the folder %catalogue ID%/Files. In the case of no existence of executable files the script starts with a copying process immediately. If imaging, file copying or md5 hash check was finished unsuccessfully, the user is asked to label the problematic USB-Stick appropriately and continue with the next USB-Stick. As already mentioned, every unsuccessful operation is logged. All described operations are illustrated in the figure 2.

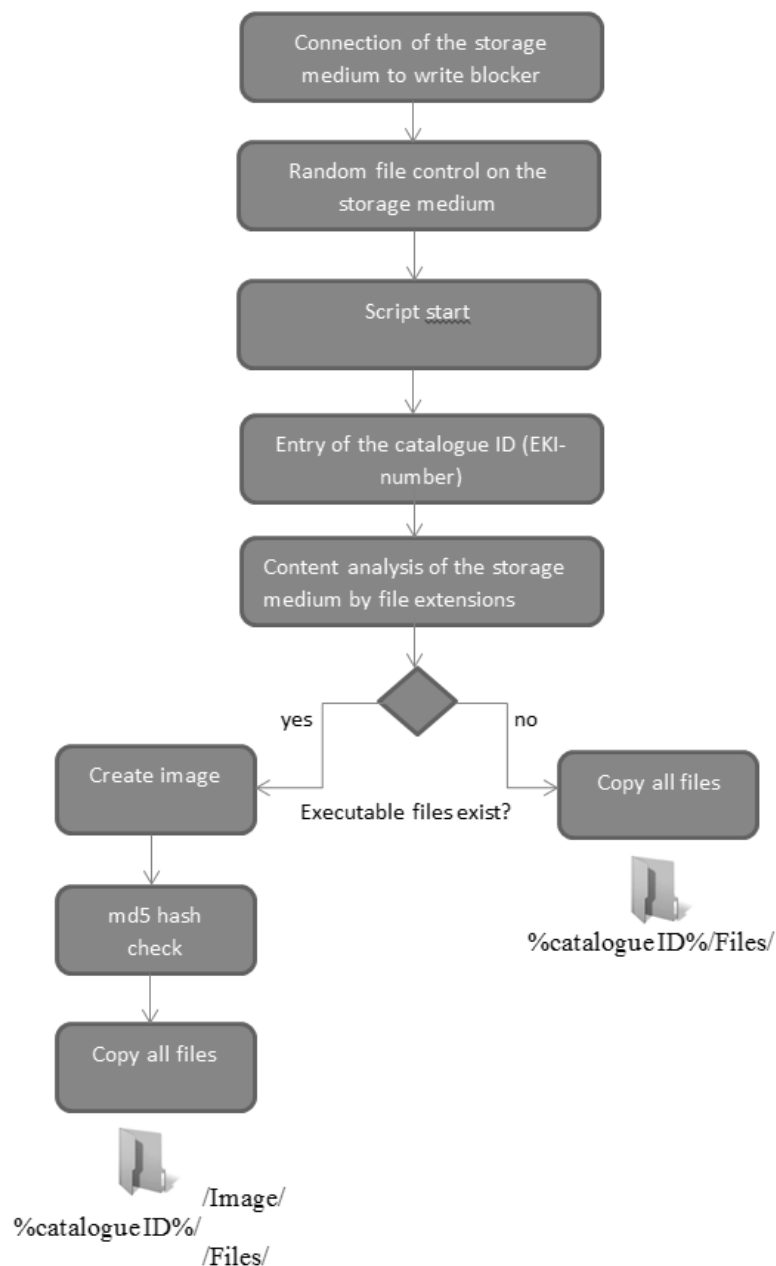


Figure 2: USB-Stick workflow

6. Results and error handling

The launch of the workflow in the grey literature team at the TIB was supported with the preparatory training for the staff. In total 334 USB-Sticks were processed on the basis of the developed workflow. No complications by using the script were reported from the grey literature team. The average process time for one USB-Stick was approximately 13 minutes. The analysis of the log file showed the following results:

- 9 errors by comparisons of md5 hash values
- 6 errors by creating an image



The final report from the grey literature team confirmed the number of errors in the log file. Moreover 11 USB-Sticks that couldn't be processed and were not found in the log file were reported. The diagram in the figure 3 illustrates the total distribution of all problem cases.

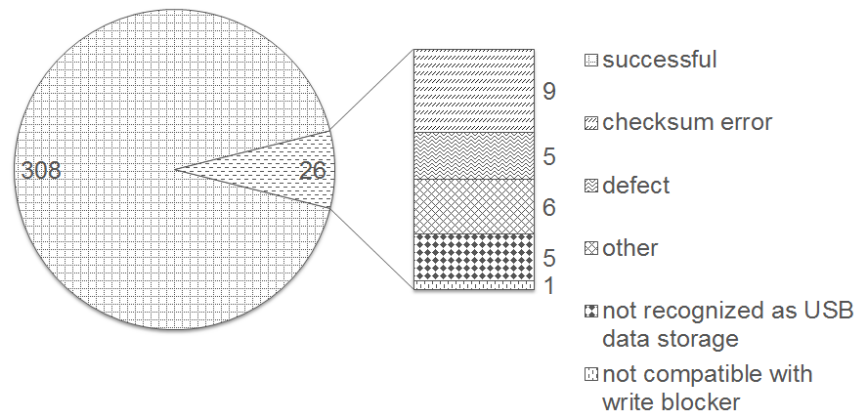


Figure 3: Distribution of the problem cases

From 334 processed USB-Sticks 26 (8%) have shown different kinds of errors. It was tried to reproduce all the errors in order to find out their reasons.

- checksum error
Since the md5 hash values were calculated for the entire storage device and image, it makes sense to find out, if the md5 hash values of individual files on the USB-Stick and files within the image are different. The comparison for 9 USB-Sticks with checksum error has confirmed that the md5 hash values of individual files are identical. This fact means that comparison of the md5 hash values of the entire USB-Stick and image isn't always reliable and should be replaced by comparison of hash values of individual files.
- other
6 errors by creating an image (other) were caused through the wrong parameter of the imaging command dd. The solution is to adjust the parameters to prevent these errors.
- not recognized as USB data storage
Since the script can only process the data storage USB drives, 5 USB-Sticks that were not detected as such, couldn't be processed. The reason for this may be the fact that the manufacturer could initially configure a drive type of the USB-Stick as a fixed storage or local disk instead of removable disk. The statement of the manufacturer SanDisk regarding this was that Windows 8 certification requires flash drive manufacturers to configure flash drives as Hard Disc Drives (Vandewerdt, 2013). The solution for these 5 USB-Sticks is to separate the data without the script.
- not compatible with write blocker
1 USB-Stick attached to the write blocker wasn't recognized. The solution is to separate the data without write blocker.
- defect
4 USB-Sticks acquired between 2009-2016 lost their data completely, while stored in the library. 1 USB-Stick was already defective at the time of acquisition. Data from 5 defective USB-Sticks can't be saved.

7. Summary

In the project presented in this paper, the author has created a reusable, semi-automated workflow using a script and the write blocker for secure data transfer from USB-Sticks. The workflow was applied in the grey literature team at the TIB as a pre-ingest process to prepare the data for transfer into the digital preservation system. As a USB-Stick isn't safe data storage, the access to the valuable information on it can't be guaranteed in the long term. Initially errors during data separation occurred for 8% of 334 USB-Sticks. However, 6,5% of problem cases were caused by another reasons such as unrecognized USB data storage or wrong parameters of the imaging command. As described in section 6 the data from these USB-Sticks can be transferred by another method or after some adjustments of the script. The conducted project has shown a failure rate of 1,5% for 334 USB-Sticks acquired in the years between 2009 and 2016. For these 5 sticks the loss/data corruption occurred while they were stored in the library. This is a clear indicator that separation of data from the carrier should take place as early as possible.



The developed workflow has proven its reliability and can continue to be used in the grey literature team. Future improvements in the workflow may concern the implementation of more trustworthy file format identification system, checksum comparison of individual files instead of entire storage and adjustment of parameters of the imaging command.

References:

- Cunningham, A. (2014). *A brief history of USB, what it replaced, and what has failed to replace it*. Retrieved on 04.10.17 from <https://arstechnica.com/gadgets/2014/08/a-brief-history-of-usb-what-it-replaced-and-what-has-failed-to-replace-it/2/>
- Dappert, A., Jackson, A. N., & Kimura, A. (2013). Developing a Robust Migration Workflow for Preserving and Curating Hand-held Media. *arXiv preprint arXiv:1309.4932*.
- EPOS BadDrive Adapter USB. http://www.epos.ua/view.php/en/products_epos_baddrive_usb retrieved on 20.09.17
- Fast forensic access to USB 3.0 storage devices. <https://www.cru-inc.com/products/wiebetech/usb-3-0-writeblocker/> retrieved on 20.09.17
- Forensic in-line USB Write Blocker. https://www.cru-inc.com/products/wiebetech/usb_writeblocker/ retrieved on 20.09.17
- Kessler, G. C., & Carlton, G. H. (2014). A Study of Forensic Imaging in the Absence of Write-Blockers. *The Journal of Digital Forensics, Security and Law: JDFSL*, 9(3), 51.
- Lazorchak, B. (2012). *Rescuing the Tangible From the Intangible*. Retrieved on 05.10.17 from <http://blogs.loc.gov/thesignal/2012/07/rescuing-the-tangible-from-the-intangible/>
- Nikkel, B. (2016). *Practical Forensic Imaging: Securing Digital Evidence with Linux Tools*. No Starch Press.
- Pullen, K. (2015). *How Your USB Cables Are About to Change Forever*. Retrieved on 05.10.17 from <http://time.com/3745070/usb-c-macbook/>
- [Rosetta overview]. <http://www.exlibrisgroup.com/de/category/Rosetta> retrieved on 01.10.17
- Scott, J. (2013). *Long-term Digital Storage: Simple Steps to Get Started*. Retrieved on 04.10.17 from <https://www.historyassociates.com/resources/blog/long-term-digital-storage-simple-steps-to-get-started/>
- Smith, T. (2014). *Hardware Write Blockers: No Worry, No Write*. Retrieved on 04.10.17 from <http://tapeop.com/tutorials/100/using-hardware-write-blockers/>
- Solomon, M. G., Rudolph, K., Tittel, E., Broom, N., & Barrett, D. (2011). *Computer forensics jumpstart*. John Wiley & Sons.
- Software write blocking. <https://dfcsc.uri.edu/research/swb> retrieved on 02.10.17
- Tableau Forensic USB 3.0 Bridge. <https://www.guidancesoftware.com/tableau/hardware/t8u> retrieved on 20.09.17
- Vandewerdt, A. (2013). *Your SanDisk USB Stick is no longer removable – and it's Microsoft's fault! [Blogpost]*. Retrieved from <https://aussiestorageblog.wordpress.com/2013/11/11/your-sandisk-usb-stick-is-no-longer-removable-and-its-microsofts-fault/> on 29.09.17
- Write Blockers. http://www.forensicswiki.org/wiki/Write_Blockers retrieved on 01.10.17

International Nuclear Information System **INIS**

*organizing the world's information
on nuclear science and technology
and making it universally accessible
for peaceful uses*

over 150 Member States and
international organizations

millions of citations and
abstracts published worldwide

hundreds of thousands of full text
non-conventional 'grey' literature

multilingual thesaurus in Arabic,
Chinese, English, French, German,
Japanese, Russian, Spanish



www.iaea.org/inis

IAEA

International Atomic Energy Agency



Providing Access to Grey Literature: The CLARIN Infrastructure

Sara Goggi, Gabriella Pardelli, Irene Russo, Roberto Bartolini, and Monica Monachini

CNR, Istituto di Linguistica Computazionale, "Antonio Zampolli", Italy

1. Introduction

"In the electronic age, the World Wide Web has played a major role in making scientific information accessible to a wide audience more rapidly and efficiently. This democratic approach to information dissemination in science is changing the way science is perceived and implemented in our daily lives" (Weintraub, 2000).

Technological process, in particular in the field of computer science, has thus eased access, retrieval and use of information as a consequence of the radical transformation which formats underwent: from papers organized on shelves to electronic files archived on the web. "The Internet has thus had the paradoxical result of making grey literature far easier to access and retrieve than once was the case, but simultaneously making so much available that it is often much harder to find or identify relevant material in the first place" (Hartman, 2006). While in its first days technology, as Hartman states, kept a very quick – and somehow wild – pace in publishing any type of information on line, nowadays there is the need of more sophisticated core technologies and technological building blocks in order to better exploit the huge amount of digital content available on the web. Therefore there is this blossoming of infrastructures, large technological shells which host documentary repositories intended to meet the expectations of a well-educated and demanding audience. The strengthening of these infrastructures at different levels (academic, national, trans-national, community, disciplinary, commercial, industrial, etc.) implies a further step in the process of gathering, organizing, managing, preserving and spreading a huge amount of relevant information. "The official definition of 'research infrastructure' refers to structures, resources and services used by a scientific community for carrying out a high-level research in several fields, from the from astronomy, physics, biology, archaeology, to the humanities. At a European level the scientific communities get together in a consortium thus creating infrastructures accessible to all their members and sharing the same resources" (Monachini, Frontini, 2016). Infrastructures stimulate new research avenues, relying on the comparison, re-use and integration into current research of the outcomes of past and on-going field and laboratory activity. Such data are scattered amongst diverse digital collections and datasets, unpublished reports (grey literature), and in publications.

Given this scenario, the authors – who deal with documentation, digitalization and language technologies for the Humanities since years now – focus on an important European research infrastructure called CLARIN (Common Language Resources and Technology Infrastructure) for assessing the traceability of grey literature within it. This work will provide a map of the documentation archived in the CLARIN infrastructure, whose purpose is to share language resources¹ produced and managed in the various European countries but finally merged into the CLARIN data centers for allowing access, interoperability, reuse and preservation of scientific documentation as well as Grey Literature.

¹ "The term 'Language Resources' refers to (usually large) sets of language data and descriptions in machine readable form, to be used in building, improving or evaluating natural language, speech or multimodal algorithms or systems. Typical examples of LRs are written, spoken, multimodal corpora, lexicons, grammars, terminologies, multimodal resources, ontologies, translation memories, but the term may also be extended to include basic software tools for their acquisition, preparation, annotation, collection, management and use. The creation and use of these resources span several related but relatively isolated disciplines, including NLP, information retrieval, machine translation, speech, and multimodality. There is the need today to broaden the definition of LRs, i.e. to re-define the "extension" of the term and recast its definition in the light of recent scientific, methodological-epistemological, technological, social and organisational developments in the application fields of content processing/access/understanding/creation, Human-Machine, Human-Human & Machine-Machine communication, and the corresponding areas from which the theoretical underpinnings of these application fields emerge (linguistics, cognitive science, AI, robotics). The extension of LRs seems to be indispensable to ensure long-lasting credibility. To achieve this, the LRs community must "liaise" with the "new" communities hinted at by the new fields above and draw a new evolving picture of existing/available/future resources following the extended definition" (FLaReNet, Fostering Language Resources Network, 2007).



2. About CLARIN

On 1st October 2015 Italy became the 16th Full Member of CLARIN ERIC (European Research Infrastructure Consortium). CLARIN-IT is the Italian node of CLARIN, whose grand vision it shares (<http://www.clarin-it.it/en/content/about>).

The Virtual Language Observatory VLO (<https://vlo.clarin.eu/search?0>) is the central repository of the CLARIN infrastructure which allows to discover language resources with a facet browser (that is, filtered by semantic categories, such as *Language*, *Collection*, *Resource type*, *Modality*, *Format*, *Keyword*, *Availability*, *Search options*), and advanced query syntax.

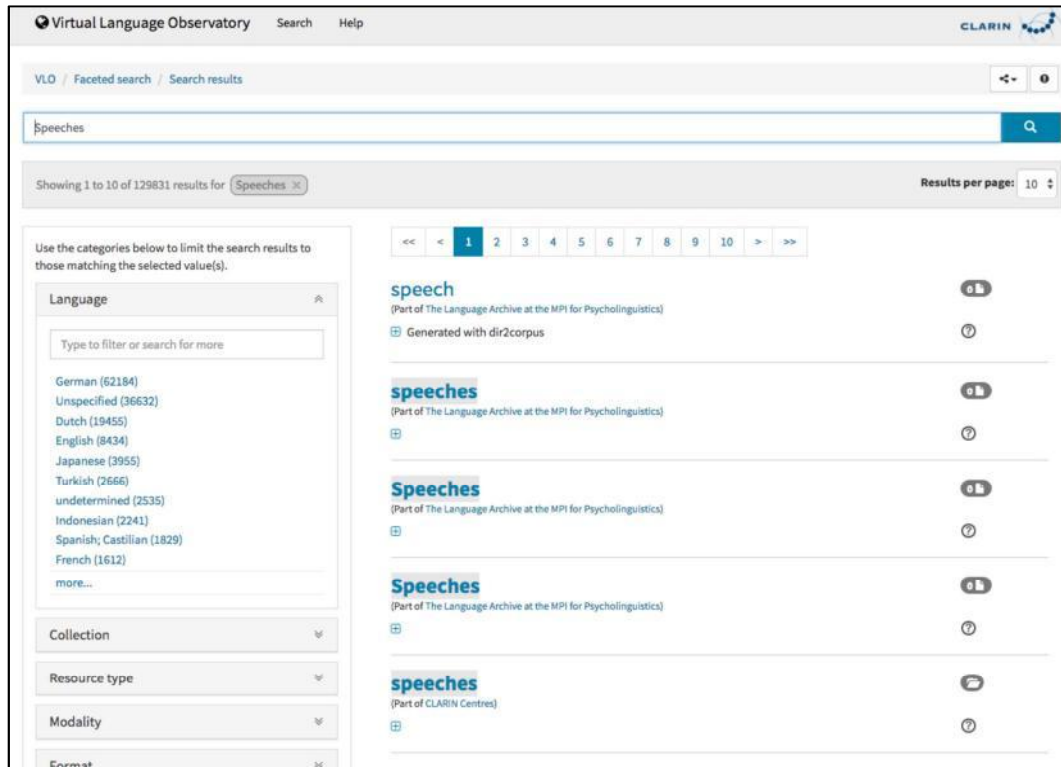


Fig. 1: Virtual Language Observatory

By simply clicking on the term *Language*, the total amount of documents in the various languages is immediately visible from the VLO: the engine has been actually conceived for providing a real time information on the quantity of available resources for that language.

All the other categories (*Language*, *Collection*, *Resource type*, *Modality*, *Format*, *Keyword*, *Availability*, *Search options*) are searchable in cascade (Fig 1).

2.1 Grey Literature in CLARIN

Grey literature that can be found in CLARIN is about language resources, and in particular about corpora, lexical resources, ontologies, treebanks, parsers, metadata, annotation tools and automatic translation.

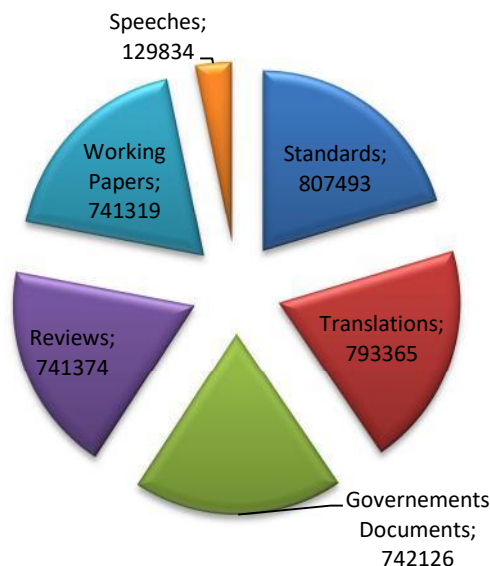
The idea is to use the VLO for checking the quantity of grey literature which can actually be found in CLARIN. A mapping between terms of the types in Grey Literature contained in the *GreyNet International 1992-2017* website [<http://www.greynet.org/>] - and in particular in the *GreySource Index > Document Types in Grey Literature* (Fig.2) – and those retrievable from the VLO, has been performed.

The data about Grey Literature contained in the CLARIN VLO is visualized in a graphical form. The three graphs below represent, respectively, the three sets of high, medium and low frequency and display the number of document found for each type.



Fig.2. Document Types in Grey Literature

2.1.1 High frequency GL documents

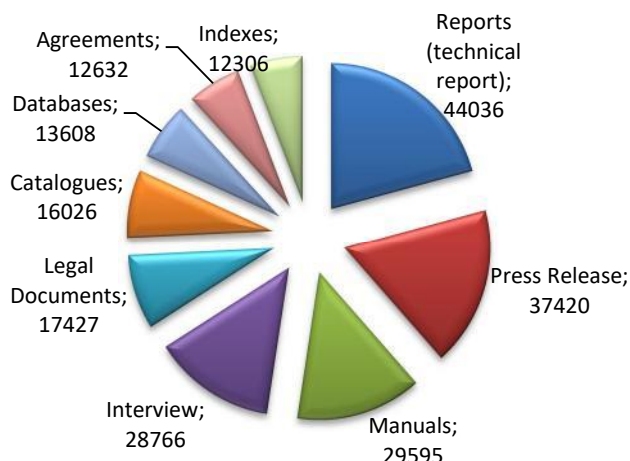


Graph.1. High frequency documents (from 1000000 to 100000)

Graph 1 highlights the types of grey literature present in the VLO in the high-frequency range. Compared to the overall quantity of information contained in the VLO, only six GL types are present and, not surprisingly, the documents about standardization appear in the spectrum of the most frequent types: standards are crucial in the infrastructure, thus allowing interoperability and sharing of data and good practices.

2.1.2 Medium frequency GL documents

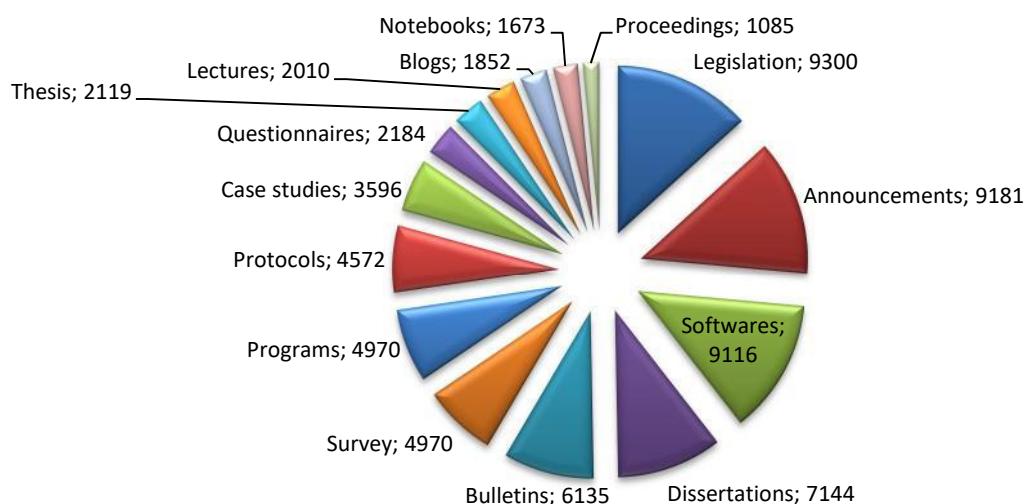
Graph 2 displays the nine types of GL documents contained in the VLO in the medium frequency range; the most frequent type is 'Reports', in particular technical reports describing resources and tools produced by various stakeholders (Business, Governments, Academia, etc.) and a high number of reports concern the documentation of treebanks.



Graph.2 Medium frequency documents (from 100000 to 10000)

2.1.3 Low Frequency GL documents

Graph 3 shows document types with low frequency in the VLO: most of the GreyNet types are in this range. In this case, the most represented document type being 'Legislation'.



Graph.3. Low frequency documents (from 10000 to 1000)

2.2 Grey Literature not retrieved in CLARIN Infrastructure

The three graphs above show the GL documents in the VLO up to the frequency of 1085. Below, we provide the document types that do not seem to be present in the central repository. In particular, compound terms have not been found, as in the case of <Research Memoranda> which apparently is not present in the VLO². Table 1 shows the list of GL compound terms which are missing in the VLO and could be further investigated.

3. Languages in CLARIN

A key feature of CLARIN is the quantity of multilingual documents stored in the infrastructure by the European communities involved in the project. The user can select the language (see Fig. 3) and access information about document types classified according to this parameter.

The Grey literature presented in Graphs 1, 2 and 3 has been further analyzed by language. Graph 4 shows how varied is the landscape of the VLO when languages are taken into consideration. In addition to the most frequent languages, it is also interesting to monitor document types attested for less frequent languages. This magnifying lens on languages shows the richness and complexity of data contained in the CLARIN infrastructure (Table 2). Fig. 3 shows the documents available for the different varieties of Occitan and constitutes a further example of the wide spectrum of languages, including the less-attested or ancient languages.

² However, if we would restrict the search to only one term, e.g. <Memoranda>, 7 occurrences could be found.



- | | | |
|---------------------------|-----------------------------|-------------------------------|
| 1. Bank Reports | 24. Grey Journals | 47. Research Proposals |
| 2. Business Reports | 25. Guidebooks | 48. Research Registers |
| 3. Call for Papers | 26. House Journals | 49. Research Reports |
| 4. Codebooks | 27. Image Directories | 50. Risk Analyses |
| 5. Committee Reports | 28. Inaugural Lectures | 51. Satellite Data |
| 6. Compliance Reports | 29. Intelligence Reports | 52. Scientific Protocols |
| 7. Country Profiles | 30. Interactive Posters | 53. Scientific Visualizations |
| 8. Country Reports | 31. Internal Reports | 54. Site Reports |
| 9. Data Papers | 32. Internet Reviews | 55. Source Document |
| 10. Deposited Papers | 33. K-blogs | 56. Specifications |
| 11. Discussion Papers | 34. Lib Guides | 57. State of the Art |
| 12. Draft Reports | 35. Non-commercial Journals | 58. Statistical Surveys |
| 13. Enhanced Publications | 36. Policy Documents | 59. Stockbroker Reports |
| 14. E-Prints | 37. Policy Reports | 60. Supplements |
| 15. Theses/ Dissertations | 38. Policy Statements | 61. Tenders |
| 16. E-texts | 39. Précis Articles | 62. Timelines |
| 17. Exchange Agreements | 40. Product Data | 63. Trade Directories |
| 18. Fact Sheets | 41. Progress Reports | 64. Treatises |
| 19. Feasibility Reports | 42. Readers | 65. White Books |
| 20. Feasibility Studies | 43. Registers | 66. White Papers |
| 21. Flyers | 44. Regulatory Reports | 67. White Papers |
| 22. Folders | 45. Research Memoranda | 68. Work Packages |
| 23. Green Papers | 46. Research Notes | 69. Working Documents |

Table 1. List of GL compound terms

DOCUMENT TYPE	LANGUAGE	HAPAX
Chronicles	Old English	1
Conference posters	Franco-Provençal	1
Conference posters	Occitan; Provençal	1
Conference posters	Romanian; Moldavian; Moldovan	1
Conference posters	Alpine Provençal O El Norte-occitano	1
Conference posters	Alpine Provençal Or North-Occitan	1
Conference posters	Dialecto De Valbonnais	1
Conference posters	Francoprovençal	1
Leaflets	Achumawi	1
Leaflets	Bengali; Bangla	1
Leaflets	Finnish	1
Leaflets	Taiap	1
Leaflets	Gujarati	1
Leaflets	Hindi	1
Leaflets	Kuanua	1
Newsgroups	Baharna Arabic	1
Newsgroups	Finnish	1
Newsgroups	Urdu	1
Newsgroups	Chino	1
Newsgroups	Chinois	1
Newsletters	Czech	1
Newsletters	Mandarin Chinese	1
Newsletters	Zenzontepec Chatino	1
Orations	Ancient Greek	1
Orations	Northwest Maidu	1
Patents	Japanese	1
Patents	Piemontese	1
Preprints	English	1
Preprints	French	1
Preprints	Yuki	1

Table 2 Languages with frequency 1

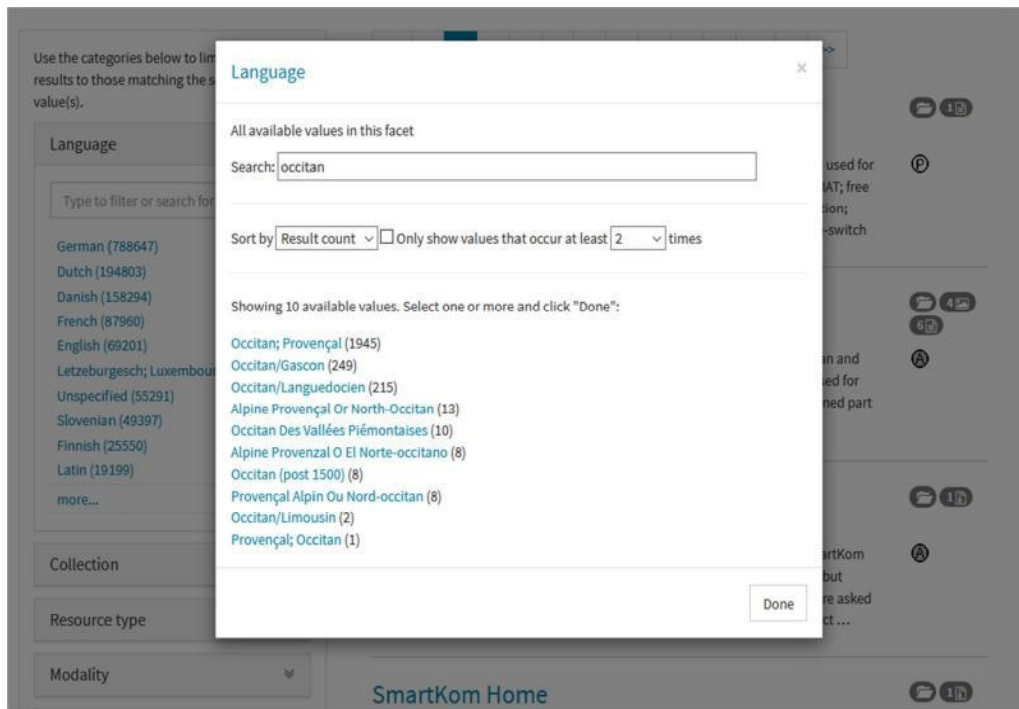
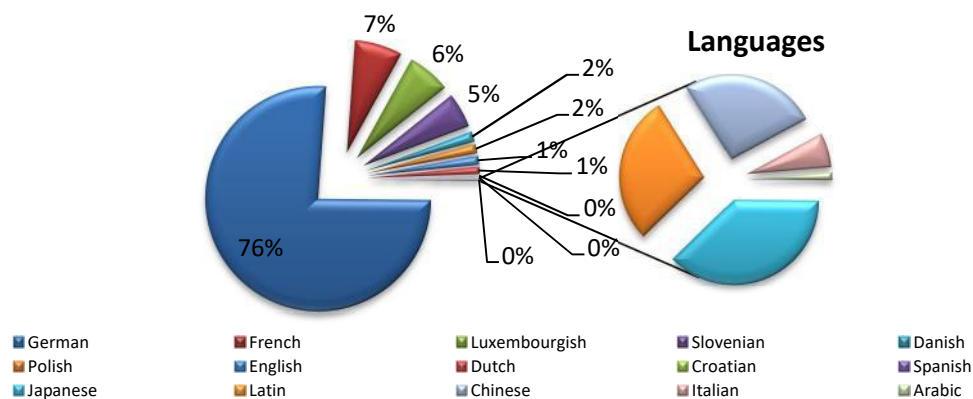


Fig.3. Languages VS GL documents



Graph. 4 Relationship between languages and documents

4. Conclusions and Remarks

The data analyzed in this paper have been extracted for the first time in April 2017 and then updated just before the GL19 Conference in October 2017. The reason for the update depends on the organization of CLARIN, which collects information constantly provided by all the European national centers.

The work presented here was aimed at verifying how a federation of data centers, not specific to bibliographic documentations but focused on the documentation of language resources, can be a mine of grey literature documentation. The 50% of the information extracted from the VLO of CLARIN has been found using key terms for grey literature document types.

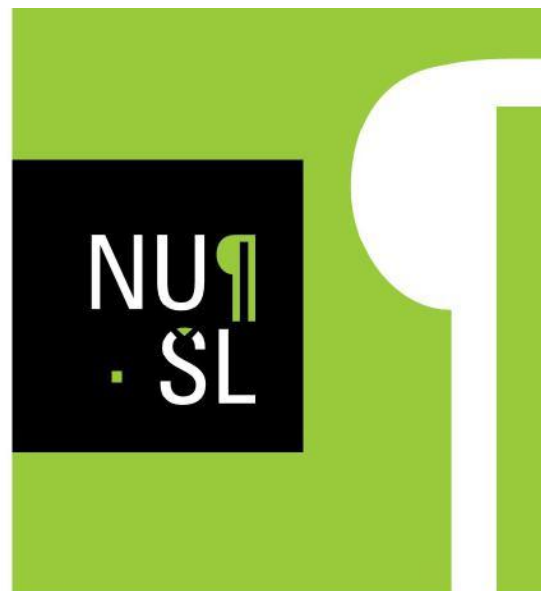
This monitoring activity has pointed out the existence of new types of Grey Literature documents in the field of language resources: these types have indeed the typical features of the unconventional documentation, firstly not being necessarily conveyed through traditional editorial channels and being open access in many cases. For the most part this documentation is constituted of monolingual and multilingual computational lexicons, syntactic databases (treebanks), collection of oral and written texts (corpora), terminologies (terminological resources); in addition, also other types of documents labeled as <storytelling> have been retrieved: they could represent a new generation of grey documentation.

The enormous outcome of the national and international research projects need to be stored in adequate infrastructures for becoming available to everyone: CLARIN is a precise example of this trend.

**Essential References**

- Banks, M.A. (2006). Towards a continuum of scholarship: The eventual collapse of the distinction between grey and non-grey literature, *Pub Res Q* (2006) 22: 4. doi:10.1007/s12109-006-0002-8.
- Calzolari N., Monachini M., Quochi V., Soria C., Goggi S., Baroni P. (2007). *FLaReNet, Fostering Language Resources Network*. ECP-2007-LANG-617001. Annex 1.Description of Work (Thematic Network). Grant Agreement n° 617001, eContentPlus.
- Carroll, B.C. & Cotter, G.A. (1997). A new generation of grey literature: the impact of advanced information technologies, *Pub Res Q* (1997) 13: 5. doi:10.1007/s12109-997-0014-z.
- Hartman, Karen A. (2006). Social Policy Resources for Social Work: Grey Literature and the Internet, *Behavioral & Social Sciences Librarian*, 25:1. Pages 1-11. doi = {10.1300/J103v25n01_01}.
- Monachini M., Frontini F. (2016). CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT. *IJCoL - Italian Journal of Computational Linguistics* vol. 2, n. 2, December 2016 "Special Issue on Digital Humanities and Computational Linguistics". Pages 11-30.
<http://www.aaccademia.it/scheda-libro?aaref=895>
- Rudasill, Lynne M. (2015). Changing dimensions of libraries in the internet era: Grey literature and scholarly communication. In Sanjay Kataria John Paul Anbu K Shri Ram Richard Gartner, *4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services*, Proceedings. Institute of Electrical and Electronics Engineers, Inc. SBN: 978-1-4799-7999-8 (PRINT) . Pages 53-57. doi={10.1109/ETTLIS.2015.7048171}.
- Weintraub, Irwin (2000) "The impact of alternative presses on scientific communication", *International Journal on Grey Literature*, Vol. 1 Issue: 2, pp.54-59, doi: 10.1108/14666180010327195.

National Repository of Grey Literature (NRGL)



NRGL is

digital
repository
for grey
literature

Free

online
access

Features

Provider:

National Library of Technology
Prague, Czech Republic

Records:

over 400,000 records

Collection provenance:

Czech Republic

Partners:

over 130 organizations (Academy of Science,
Public Research Institutions, Universities, State
Offices, Libraries, NGOs etc.)

International Cooperation:

OpenGrey, OpenAire, ROAR, OpenDOAR, BASE

Goals

- Central access to grey literature and the results of research and development in the Czech Republic
- Support of science, research and education
- Systematic collection of metadata and digital documents
- Long-term archiving and preservation
- Cooperation with foreign repositories

What else?

Conference on Grey Literature
and Repositories

<http://nrgl.techlib.cz/conference/>

Informative Web pages

<http://nrgl.techlib.cz>

www.nusl.cz

NTK
4x 2,5x 4,5x
Národní technická knihovna
National Library of Technology

NUŠL
národní
úložiště
šedé
literatury



Collecting Grey Literature – Institutional Repository versus National Aggregator

Petra Černohlávková and Hana Vyčítalová,
NTK, National Library of Technology, Czech Republic

Abstract

The Czech National Library of Technology (NTK) provides two digital repositories – the National Repository of Grey Literature (NRGL) and the NTK Institutional Digital Repository (IDR). NRGL's primary is providing access to grey literature as well as long-term archiving and preservation of grey literature from various institutions in the Czech Republic. The IDR was created for collecting and archiving of employee-generated content and other documents, including grey literature, connected with the library and its services. Our poster highlights the differences between collecting grey literature at the institutional level (through the institutional repository) and at the national level. What commonalities and differences do they have? What problems do they solve? Differences include not only overall conceptions and document types, but also methods for collecting, legal issues and standards as well as functionality and options. Thanks to our experiences in managing both types of repositories, we define general differences, obstacles, and development possibilities. Information presented here, including a mode for cooperating at the institutional or national level, is useful for all institutions planning to start collecting (not only) grey literature at the institutional or the national level even at cooperating institutional model/level.

Introduction

The poster represents the two most common streams for how the grey literature in electronic form is presented nowadays - by institutional repositories and aggregators. The contribution highlights the differences between collecting grey literature at the institutional level and at the national (or international) level. The major difference is that an institutional repository usually provides access to and preservation of the institutional publications, whereas an aggregator gathers results from multiple resources such as databases, repositories, digital libraries, or webpages.

The Czech National Library of Technology (NTK) runs two repositories – the National Repository of Grey Literature (NRGL¹, since 2009) and the NTK Institutional Digital Repository (IDR², since 2011). NRGL's primary aim is providing access to grey literature at the national level as well as long-term archiving and preservation of grey literature from various institutions in the Czech Republic (NRGL Project, 2017). The IDR was created for collecting and archiving of employee-generated content and other documents, including grey literature, related to the library and its services.

Based on our experiences, we define seven of the most important topics necessitating discussion when establishing a new repository – general conception, document types, collecting methods, participation, legal issues, functionalities, and accessing. The topics are briefly discussed in the rest of the contribution.

General conception is the starting point for all the topics.

General conception

The general conception of the institutional repository is quite clear compared to an aggregator. Institutional repositories usually collect publishing activities of the institution and internal documents such as directives or reports of business trips. Nowadays there is increasing number of repositories, including also datasets (131 out of 2952 institutional repositories³). The general conception of aggregators differs from one to the other. For example, the conception of the NRGL is to collect all grey literature in the Czech Republic. The aggregators very often are multi-disciplinary or have multiple purposes. Based on the data from OpenDoar, we have created a chart dividing aggregators by their focus/profile (Chart 1).

¹ <http://nusi.cz/?language=en>

² <http://repozitar.techlib.cz/?ln=en>

³ OpenDoar, up to 17th October 2017

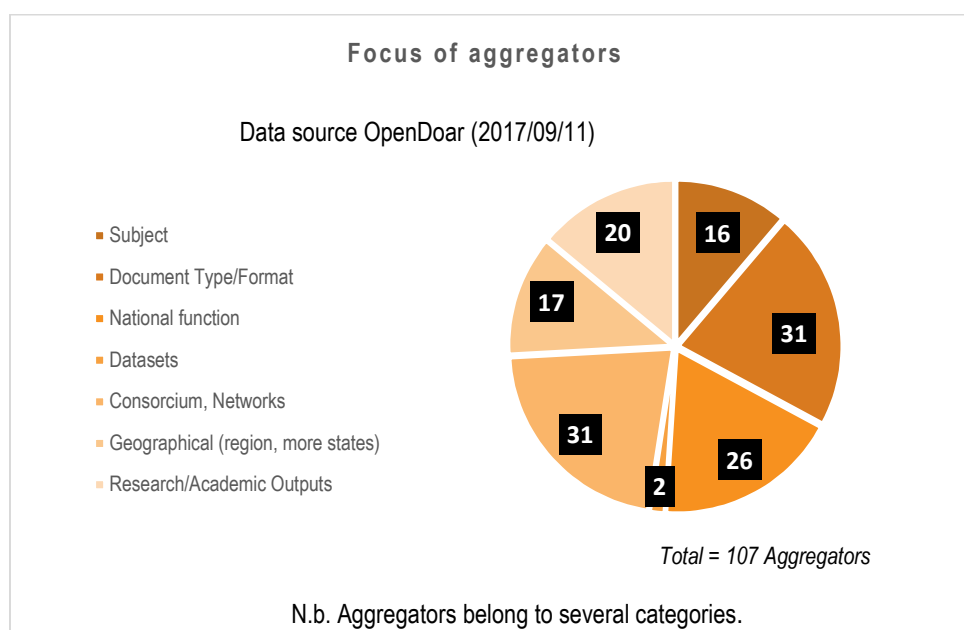


Chart 1: Focus of aggregators (data source: OpenDOAR <http://www.opendoar.org/>)

Document types

There are plenty of document types that could be collected by any repository. At the beginning, it is necessary to analyse or map which types of documents are published in the institution(s) and in which formats. Regarding formats, it is important to verify which are appropriate for long-term preservation, and which are more suitable for access to end users. Any repository must be flexible enough to deal with the addition of another type of documents than have been determined at the beginning and fulfil needs of the institution(s). Because each repository and institution has different needs and publication activities, we distinguish different document types for the IDR and the NRGL (see Table 1; types in *italics* are common for both typologies). These typologies also have been changing on account of changing needs of the institution(s).

Table 1: Comparison of NRGL typology and IDR typology

NRGL	IDR
Academic theses (ETDs)	Academic theses (ETDs)
<i>Bachelor's theses</i>	<i>Bachelors thesis</i>
<i>Master's theses</i>	<i>Master's thesis</i>
Doctoral theses	Others
Habilitation theses	
Rigorous theses	
Analytical and methodological materials	
<i>Analysis</i>	
<i>Methods</i>	
Studies	
Author works	Author works
<i>Monographs</i>	<i>Monographs</i>
<i>Preprints</i>	Scholarly Articles
Reviews	Post-prints
Thematic collections	<i>Preprints</i>
Conference materials	Conference Materials
<i>Posters</i>	<i>Posters</i>
<i>Programs</i>	<i>Programs</i>
<i>Papers</i>	<i>Papers</i>



<i>Proceedings</i>	<i>Proceedings</i>
Promotional and educational materials	Promotional Materials
Brochures	Photographs
Flyers	Leaflets
Exhibition catalogues	Monitoring
Exhibition guides	Help
<i>Press releases</i>	Awards
	Invitations
	PF
	Videos
	<i>Press releases</i>
Reports	Reports
<i>Business reports</i>	<i>Business reports</i>
Grant reports	<i>Progress reports of the project</i>
<i>Progress reports of the project</i>	<i>Annual reports</i>
Statistical reports	<i>Final report of the project</i>
Technical reports	Work placement reports
Research reports	
<i>Annual reports</i>	
<i>Final reports of the project</i>	
Status reports	
Survey reports	
Trade literature	Informative Documents
Trade print	<i>Analysis</i>
Product catalogues	Conceptions
Gazettes	<i>Methods</i>
	Normative documents
	Statistics
	Legal Documents
	Study Materials
	Presentations of training

Table 1: Comparison of NRGL typology and IDR typology

The OpenDoar differentiates following 12 content types: Journal articles, Theses and dissertations, Unpublished reports and working papers, Books, chapters and sections, Conference and workshop papers, Multimedia and audio-visual materials, Learning Objects, Bibliographic references, Datasets, Other special item types, Software, and Patents (OpenDOAR, 2016). You can see in Chart 2 that the collected documents are very similar but the intensity and the order is slightly different.

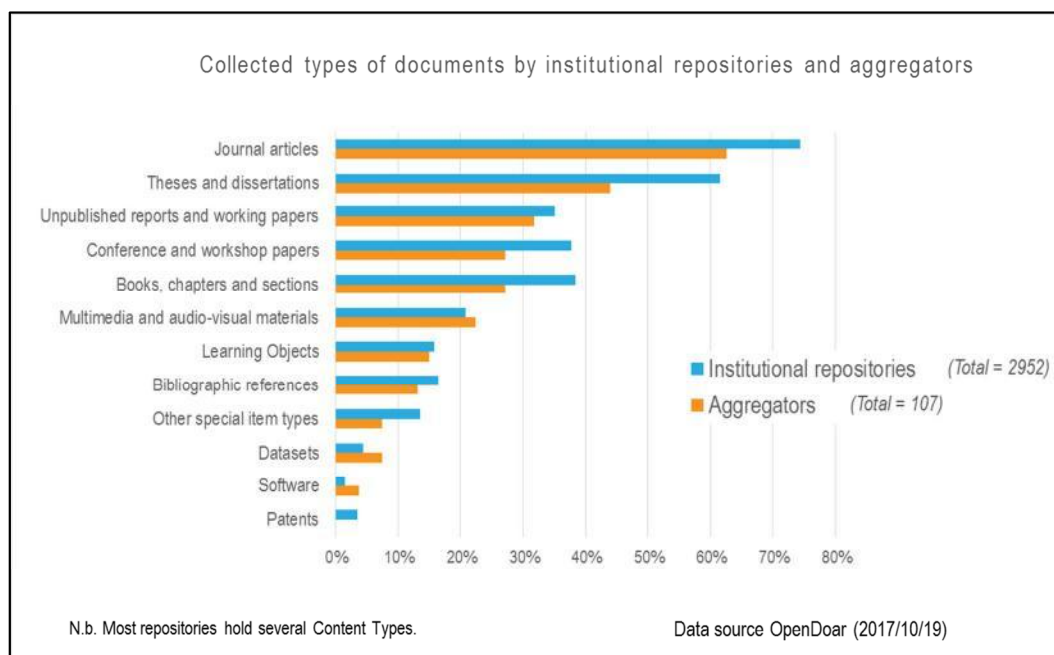


Chart 2: Types of documents collected by institutional repositories and aggregators (data source: OpenDOAR <http://www.opendoar.org/>)

Collecting methods

An institutional repository can use the same collection methods as an aggregator. However, they can have very often different priorities and possibilities. Nevertheless, they should agree on one strategy – get the records and full-text quickly and easily.

This strategy could be trickier for the institutional repositories if there is not any other system(s) from which they could harvest the documents via OAI-PMH or export and then import to the repository. Exclusion of these solutions moves the possibilities to the self-archiving or manual submission by a person in charge of the repository, very often a library staff member.

Aggregators often combine the aforementioned methods. The most suitable is harvesting via OAI-PMH which allows collecting the data automatically from other repositories or library catalogues at regular intervals; for example, once a week. There are many metadata formats that can be used, therefore it increases the importance of the metadata schema selection for your repository. Based on analysis of the resources that could be harvested, keep in mind that selection of the same schema as other institutions use means less extra work. The most common metadata format within repositories is Dublin Core and, within library catalogues, modifications of MARC.

Participation

At the institutional level, all employees should participate and submit their employee works to the repository. It is usually determined by institutional mandates or norms signed by an organizational director. In cases of universities, these mandates mostly include also theses and dissertations. Nevertheless, for implementing mandates, authors, all departments with publishing activities, and the repository team (often part of the library) must participate.

Participation in the aggregator's network is usually voluntary. It is related to the willingness of the document/data producers such as universities or research centres to share their publishing outputs, grey literature, or data. This willingness can be affected, in case of manual submissions, by the fact that the depositor often does the submissions as an extra additional task with lower priority. There could be some exceptions when the participation is determined by the law – e.g. results of the research.

It could seem that an institutional repository will more easily reach high numbers of submissions, but this is not always the case. If there is no penalty for missing submissions, the publishing activity of the institution or the content collected, the repository is never going to be completed. On the other hand, it cannot be forgotten that copyright legislation will always provide some limits and specific conditions.



Legal issues

Legal issues are very important topic and sometimes very specific in the context of grey literature. Even grey documents are authorial works and are protected by copyright legislation in many countries. Managers of all repositories must be aware of this and respect this.

In the case of institutional repositories, there are usually employees' works and emerge out of the status the employer has rights to archive these works in its repository. It is not necessary to ask the authors for their agreement, but many institutions do so to be safe⁴.

Aggregators are the opposite – since participation in an aggregator is voluntary, repositories cannot collect anything without previous agreement or contract with the collected organizations or all authors. It can be classical paper contract or only an electronic confirmation checked during submitting of the document. For our repositories we prefer paper contracts. Then there is an option of using any type of open licenses, e.g. Creative Commons. Documents marked with Creative Commons' symbol are possible to archive or share freely under the terms of the license.

It is very important to keep in mind that copyright regulations are different in every country, especially if part of planned management is the involvement of repositories in other countries. See Chart 3 to notice the stratification of the repositories around the world.

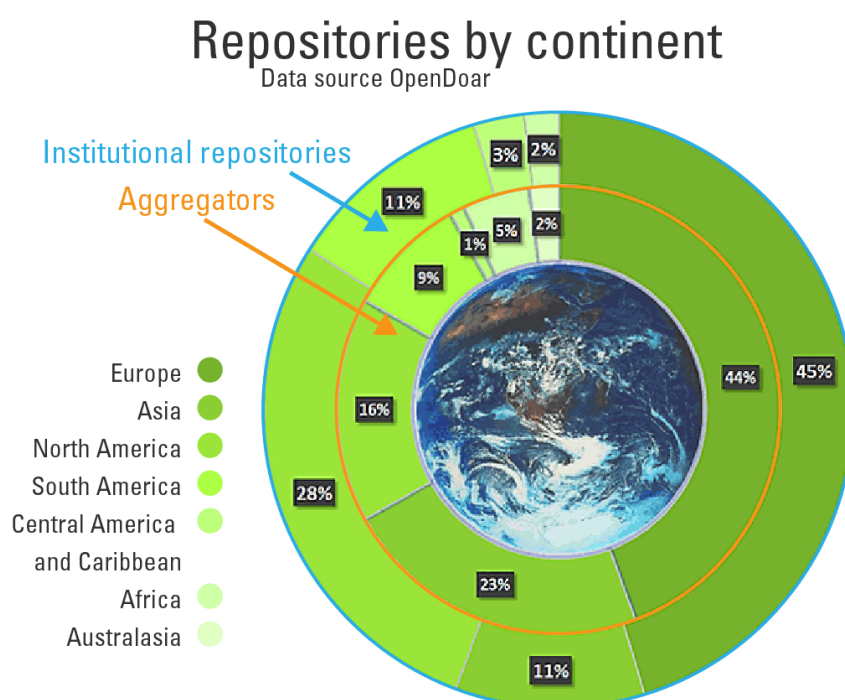


Chart 3: Repositories by continent (data source: OpenDOAR <http://www.opendoar.org/>)

Functionalities

The primary function of an institutional repository is long term preservation of its documents. Secondary functions could be considered a digest of publishing activities of the employees or a presentation of the institution to the public (Charvátová, 2016). However, it is expected that the employees will use it the most often. Search functionality could be limited if there is higher emphasis on archiving. As mentioned above, the main user groups are employees and - at universities - professors and students, based on this fact the easy management of electronic documents is required to simplify their work.

The aggregator's most important functionality is searching over many different sources from one access point together with very robust advanced searching options. The aim is save time to the end user who would otherwise have to search all these sources one by one. The aggregators usually have access to less full-texts, but they capable provide information on full-text availability and link end users directly to the full-text. Secondly, the goal is to collect documents with similar aspects or topics and, in some cases, this could have equal or even higher priority than search functionality. For example, at the NRGL case, goals - to collect all Czech grey literature and to enable search over many resources –are going side by side.

⁴ For further reading: "Breaking It Down: A Brief Exploration of Institutional Repository Submission Agreements" by Amanda Rinehart and Jim Cunningham, available from <https://doi.org/10.1016/j.acalib.2016.10.002>

Accessing

The institutional repository serves mainly to the particular institution and its internal policy decides about access to documents. An institutional repository can be open access or only for internal use, or a mix of both options. An embargo period also very often appears regarding to the journal articles, documents from research area, or project documents. In the IDR, access to internal documents is limited to the IP addresses in the building, but access management could also be managed using employee user accounts.

The aggregator's provider usually cannot decide easily about accessing of the documents. He or she must respect copyright regulations and particularly licenses and conditions given by the producers of the documents (authors or institutions) in some agreements. There could be some exceptions too; e.g., when the existence of the aggregator is established by the law or by some authority (Ministry, European Commission). However, in many cases the aggregator provides access to metadata and only selectively to full-texts (e.g., if there is some agreement between the producer and the provider of the aggregator or if the full-text is open access or under any open licence). Metadata records without full-text should always contain the information about full-text availability. See the schema below for a detailed comparison of institutional repositories and aggregators in point of view of accessibility.

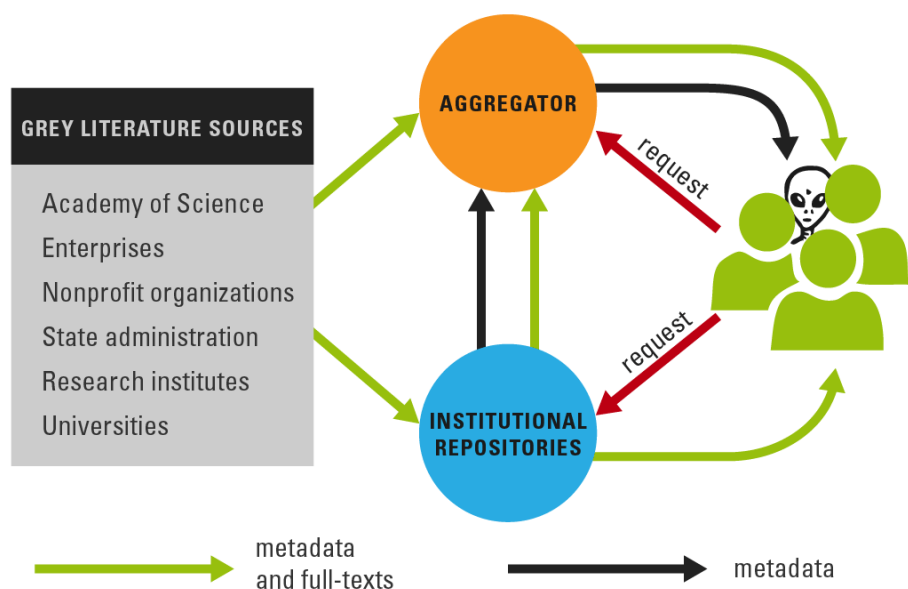


Chart 4: Schema of aggregator workflow (source: The National Library of Technology)

Conclusion

Merging all the discussed points and factors – the general conception, the collected document types, the collection and preservation methods, participation and legal points of view, manners of presentation of results and dissemination tools – could help in the selection of the best system for a new repository (institutional or aggregating type) and even at the beginning of a project or when rethinking the current situation of your repository.

References

- CHARVÁTOVÁ, Michaela, 2016. Institucionální digitální repozitář NTK. *Čtenář* [online]. Středočeská vědecká knihovna v Kladně, 2016, **68**(11). ISSN 1805-4064. Available from: <http://ctenar.svkkl.cz/clanky/2016-roc-68/11-2016/tema-institucionalni-digitalni-repozitar-ntk-163-2372.htm>.
- NRGL Project, 2017. National Repository of Grey Literature [online]. Praha: National Library of Technology [cit. 2017-12-01]. Available from: <https://nrgl.techlib.cz/nrgl/nrgl-project/>
- OpenDOAR: *The Directory of Open Access Repositories*, 2016 [online]. Nottingham, UK: University of Nottingham, UK, 2014 – 2016 [cit. 2017-12-01]. Available from: http://www.open_doar.org/



OpenAIRE: Advancing Open Science

Paolo Manghi, Michele Artini, Claudio Atzori, Miriam Baglioni, Alessia Bardi, Sandro La Bruzzo, and Michele De Bonis, Institute of Information Science and technologies, Italian National Research Council, Italy

Harry Dimitropoulos, Ioannis Foufoulas, Katerina Iatropoulou, Natalia Manola, Stefania Martziou, Athena Research Center in Information, Communication and Knowledge Technologies, Greece

Pedro Principe, University of Minho, Portugal

Abstract

OpenAIRE¹, the point of reference for Open Access in Europe, is now addressing the problem of enabling the Open Science paradigm. To this aim it will provide services to: (i) overcome the limits of today's scientific communication landscape, by allowing research communities and the relative e-infrastructures to fully publish, interlink, package and reuse their research artefacts (e.g. literature, data, and software) and their funding grants within the European and global ecosystem as supported/promoted by OpenAIRE, (ii) enable end-users (e.g. researchers, funder officers) to search and consult a rich and up-to-date knowledge graph of research results and (iii) enable scientific and educational information repositories and publishers to subscribe and be notified of changes in the OpenAIRE knowledge graph. These combined actions will bring long-term and immediate benefits to research communities, research organisations, repository managers, and funders by affecting the way research results are disseminated and reused. On the one hand, publishing the interlinked and packaged research literature, data and software via OpenAIRE drives research communities to an Open Science transition in a consistent and interoperable fashion. On the other hand, the resulting infrastructure concretely enables the construction of Open Science oriented services, supporting practices such as machine-assisted research reproducibility and evaluation.

1. Introduction

Open Science is frequently defined as an umbrella term that involves various movements aiming to remove the barriers for sharing any kind of output, resources, methods or tools, at any stage of the research process⁹. This is intended as a means for accelerating research by enhancing transparency and collaboration, and fostering innovation and reproducibility. Scientists and organizations see Open Science as a way to speed up, improve quality, and more effectively reward research activities, while funders and ministries see it as a means to optimize cost of science and leverage innovation. Open Science is an emerging vision, a way of thinking, whose challenges always gaze beyond its actual achievements. Today, the effective implementation of Open Science calls for a scientific communication ecosystem capable of enabling Open Science publishing principles. The ecosystem should allow research communities to share (for discovery and transparent evaluation) and re-use (for reproducibility²¹³) their scientific results by publishing all intermediary and final research artefacts, beyond scientific literature. Artefacts can be research data, software and research methods (e.g. workflows, protocols, algorithms, etc.), which should be deposited in repositories for scientific communication (e.g. institutional repositories, data archives, software repositories, CRIS systems), and should be published together with the semantic links between them. To complete the picture, such ecosystem should support publishing of packages of artefacts (e.g. research objects²¹⁴, enhanced publications⁵¹⁵, RMap¹⁶) to allow discovery, evaluation, and reproducibility of science (e.g. workflows or experiments with input datasets).

Today's scientific communication landscape is far from supporting this vision, mainly due to its inability to:

1. *Support publishing of all kinds of research artefacts.* For example, research methods publishing workflows are generally not best practice, i.e. no research method repositories, no persistent identifiers for methods, no citation practices and, therefore, no scientific reward;
2. *Keep a complete and up-to-date record of research artefacts relationships.* For example, publication, data, software repositories and publishers do not keep bi-lateral links between each other's artefacts, and the links they keep are not in-sync with the updates of the artefacts (e.g. links to new versions of the data, obsolete links);

¹ OpenAIRE www.openaire.eu



3. *Find agreements on how to share and publish packages of artefacts.* Solutions exist (e.g. research objects, enhanced publications, RMap) but are specific to rather small communities of scientists, implying that research packages, as well as research methods, are not regarded as first-class citizens in the scientific communication domain.

De facto, today's scientific communication ecosystem lacks tools and practices for engaging research communities at adopting the aforementioned novel Open Science publishing principles, even when researchers are already in the position of publishing interlinked artefacts and/or packages.

OpenAIRE fosters transparent evaluation of results and facilitates reproducibility of science for research communities by enabling a scientific communication ecosystem where artefacts, packages of artefacts, and links between them can be exchanged across communities and across content providers. To this aim, OpenAIRE, via the OpenAIRE-Connect project, introduces and implements the concept of Open Science as a Service (OSaaS) on top of the existing OpenAIRE infrastructure 1, by delivering services in support of Open Science. Following the NIST definition of *aaS* service models², the service model "Open Science as a Service" provides the consumers with the capability of accessing tools that implement Open Science principles, transparently with respect to the underlying technical infrastructure. Tools are accessible through either a thin client interface, such as a web browser, or an application program interface (API).

OpenAIRE-Connect³ will realize and operate two services for Open Science. The first, Research Community Dashboard, will serve research communities to publish research artefacts, packages, and links, and to monitor their research impact. The second, Catch-All Notification Broker Service, will engage and mobilize content providers, and serve them with services enabling notification-based exchange of research artefacts, to leverage their transition towards the Open Science paradigm. Both services will be served on-demand according to the OSaaS paradigm, hence be re-usable by different disciplines and providers, each with different practices and maturity levels, so as to favor a shift towards a uniform cross-community and cross-content provider scientific communication ecosystem.

The adoption of these services, eased by the OSaaS approach, aims at incepting Open Science publishing within the existing scholarly communication landscape. By introducing an OSaaS approach, OpenAIRE-Connect will deliver on-demand Open Science publishing services to research communities and content providers, aligning practices and mechanisms that address transparent evaluation and reproducibility (see Figure 1). By complementing the technological efforts with networking activities that will strengthen the emerging Open Science social environment, OpenAIRE-Connect will facilitate a cultural and technological shift towards common Open Science publishing practices.

To achieve its objectives, OpenAIRE-Connect involves key stakeholders of scientific communication: a pool of forward-looking research communities, today publishing or in the need of publishing research data and methods, international representatives of Open Access publishers (Jisc 3), publication repositories (COAR 9), and data archives (ICSU World Data Systems/WDS 11), willing to support and benefit from such a change.

² Peter Mell and Timothy Grance (September 2011). The NIST Definition of Cloud Computing (Technical report). National Institute of Standards and Technology: U.S. Department of Commerce. Special publication 800-145.
<http://doi.org/10.6028/NIST.SP.800-145>.

³ OpenAIRE-Connect <https://www.openaire.eu/connect>

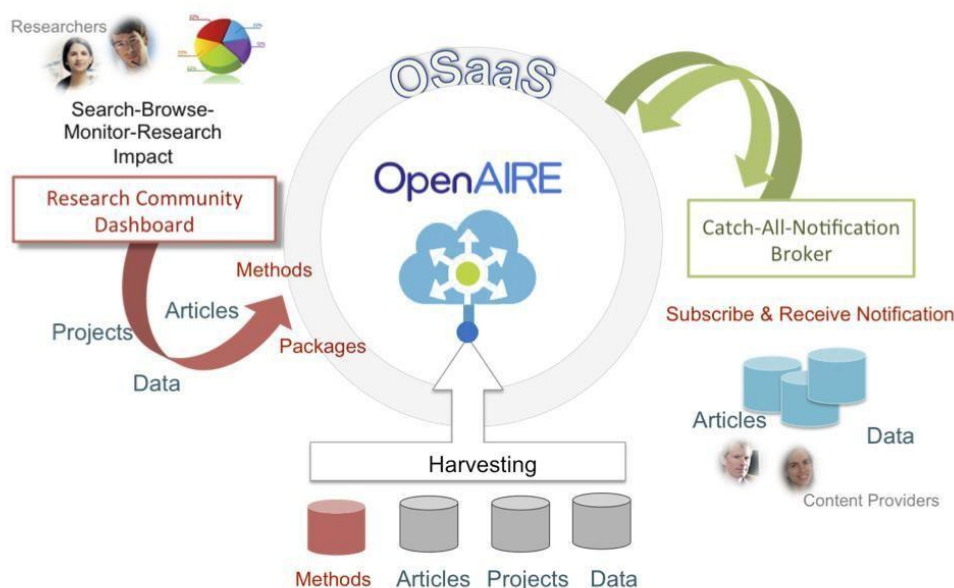


Figure 1. Research Community Dashboard and Catch-All Notification Broker Service

2. OpenAIRE services in support of Open Science

OpenAIRE-Connect extends the technological services today offered by the OpenAIRE infrastructure in order to foster the adoption of Open Science publishing practices and facilitate the emergence of shared solutions. Specifically, OpenAIRE-Connect introduces two classes of new services:

1. *Research community services* offering support for a uniform transition of research communities towards Open Science publishing via the Research Community Dashboard;
2. *Content provider services* leveraging the transition of content providers towards Open Science publishing via the Catch-All Notification Broker Service.

Continuing and building on OpenAIRE's openness and sharing of content, services and practices, OpenAIRE-Connect will develop a *uniform, common strategy* for approaching and engaging with research communities (especially Research Infrastructures targeting ESFRI 8 programmes), which will be a major outcome of the project. This strategy will be used in due course by the wider OpenAIRE constituency, i.e. the National Open Access Desks (NOADs), in its strategic synergies within the emerging European Open Science Cloud (EOSC 12) ecosystem to increase awareness on the Open Science topics and to promote the adoption and uptake of the new services.

In synergy with EOSC, towards the widely discussed need of the gluing social aspect (support and human infrastructure), OpenAIRE-Connect aims to design and deliver a *targeted support and training programme for research communities* and relevant stakeholders. This will inform them about the benefits and use of the OpenAIRE-Connect services, to pass on best practices and to lower the barriers of participation in the Open Science ecosystem and, particularly, in the OpenAIRE infrastructure.

2.1. Research Community Dashboard

OpenAIRE-Connect will support the evolution of today's fragmented scientific communication landscape by providing researchers of specific communities with services giving access to facilities for collaboratively maintain an up-to-date knowledge graph of their interlinked or packaged research artefacts, e.g. literature, data, software and methods. This research community graph will be built as an extension of the broader knowledge graph today populated by the OpenAIRE infrastructure by adding software and other research artefacts (all the artefacts that are different from literature, dataset, and software). It will therefore integrate and inherit links to funders, projects, literature and datasets as inferred from article full-texts (today more than 4 millions) or harvested by OpenAIRE from content providers (today more than 1000). These facilities will be provided by a new infrastructural service, the OpenAIRE Research Community Dashboard, that each community will be able to request and configure according to its specific needs. Each dashboard will serve two types of community users (researchers and research community operators, who have an "administrative role" for the community and can configure the dashboard) with a suite of value added functionalities:



- *Connect and Link*: researchers authoritatively provide and curate links between artefacts related to their scientific community, a process moderated by the research community operators.
- *Deposit*: researchers who have not a repository of reference can deposit in OpenAIRE's Zenodo 67 files and metadata relative to their research literature, data, software, methods, and packages, and obtain a DOI.
- *Enrich Content* (configurable inference): research community operators authoritatively tune the configuration of OpenAIRE text mining algorithms with community specific rules to identify artefacts relevant for the research community;
- *Learn your Impact*: the research community can reliably monitor and report the research impact of their scientific production w.r.t. several European (and beyond) funders, visualize trends, classifications resulting from OpenAIRE's analytics services; the knowledge graph inclusive of all community artefacts, with links between them and relative attribution of work, enables transparent evaluation of science.
- *Discovery and reuse*: researchers can restrict their search, browse, and navigation focus to the subpart of the OpenAIRE information space associated to the community; discovering and accessing packages of artefacts fosters reproducibility of science.

OpenAIRE-Connect will also develop, in collaboration with other international initiatives, interoperability guidelines to enable the exchange of research artefacts and packages and will offer APIs enabling third-party services to bulk-feed research artefacts into the OpenAIRE knowledge graph according to the established guidelines.

OpenAIRE-Connect involves its onset researchers from a wide range of communities looking into pragmatic solutions to research data and methods publishing in Open Science settings. With different levels of maturity, touching upon interdisciplinary, they will deliver end-user requirements for the realization of the *Dashboard* service, and engage in pilots to test, assess, and adopt the services:

- *Earth and Environmental Sciences (UniHB)*: the ATLAS⁴ community relies on thematic data archives (Pangaea) whose datasets are packages of datasets and related literature, aiming to link to different scientific domains.
- *Neuroinformatics (CNRS)*: the France Life Imaging national infrastructure⁵ produces data images, links them with methods (software and services), and produces packages.
- *Fisheries and aquaculture management (CNR, IDR)*: the BlueBridge⁶ and MARBEC⁷ infrastructures are moving towards collaborative editing of "dynamic publications", looking for Open Science solutions.
- *Humanities and Cultural Heritage (PIN)*: PARTHENOS⁸, a cluster of research infrastructures from Linguistics, Humanities, Cultural Heritage, History, Archaeology, with different types of data to interconnect.
- *Environment & Economics (ICRE8)*: the national/EU node of the United Nations Sustainable Development Solutions Network⁹ sets out to build an infrastructure to gather all publications and data available in repositories and in Public Sector Information portals, and link them to European and national funding.

2.2. Catch-All Notification Broker Service

Research artefacts repositories, a.k.a. content providers, (e.g. institutional and thematic repositories, aggregators, data archives) are key in serving research communities towards their Open Science goals. To ease the adoption of Open Science principles among researchers it is important to lower the barriers to Open Science publishing.

The "publish/deposit once" practice is a decisive means in the overall Open Science roadmap, and will only be achieved when content providers seamlessly connect to the wider open scientific communication ecosystem. This will allow them to pro-actively exchange information about research artefacts and links of value to all interested communities or stakeholders, without the

⁴ ATLAS: <https://www.eu-atlas.org/>

⁵ France Life Imaging: <https://www.francelifeimaging.fr/>

⁶ BlueBridge: <http://www.bluebridge-vres.eu/>

⁷ MARBEC: <http://www.umar-marbec.fr/en/?lang=en>

⁸ PARTHENOS: <http://www.parthenos-project.eu/>

⁹ United Nations Sustainable Development Solutions Network: <http://unsdsn.org/>

researchers having to worry *where*, *when*, and *how* to publish in order to fulfill the numerous mandates.

As part of the OSaaS portfolio, OpenAIRE-Connect will develop and deploy a Catch-All Notification Broker Service that connects all types of content providers (institutional repositories, publishers, data repositories, and CRIS systems). The Catch-All Notification Broker will extend OpenAIRE's notification brokering service 4, which serves literature repositories, and will broaden the content provider base with the ones that serve specific research communities.

Thanks to its functionality, providers can be notified of metadata records relative to artefacts that are "of interest to them" (i.e. metadata records that should be in the content provider's data base), or "linked to them" (i.e. a scholarly link exists between one of the provider's artefact and the identified artefact).

Notifications are sent only to subscribed providers, following a subscription and notification pattern, and can be delivered by mail, OAI-PMH interfaces, or, currently under investigation, via push APIs (e.g. SWORD protocol), FTP and ResourceSync.

This will effectively allow content providers to complete or enrich their collection of artefacts with up-to-date information from the wider OpenAIRE ecosystem, and research communities or infrastructures to have a direct communication line with content providers via OpenAIRE.

The idea behind the service is to disseminate and advocate the principle that scholarly communication data sources are not a passive component of the scholarly communication ecosystem, but rather active and interactive part of it. They should not consider themselves as thematic silos of products, but rather as hubs of products semantically interlinked with any kinds of research artefacts and, more broadly, up-to-date with the evolving research ecosystem.

OpenAIRE-Connect brings on board leading representatives of institutional repositories (e.g. COAR), data repositories (UniHB/Pangaea, UniHB/WDS), publishers (via Jisc) that are already moving towards Open Science-oriented publishing and are committed to provide requirements and engage in the experimentation of the brokering services. Beyond the ones on board to the project, i.e. Pangaea (UniHB) and Zenodo (CERN), a number of content providers have already indicated their interest in the Catch-All Notification Broker Service and are ready to engage in a number of pilots for testing and adopting the service: the German GESIS Datorium repository for datasets and scripts in the Social Sciences, the FCT-FCCN network of Portuguese institutional repositories, Australian ANDS data archive, and Open Access publishers, such as eLife, Frontiers, EuropePMC.

In addition, OpenAIRE-Connect will implement a pilot for exchanging subscriptions and notifications between the Catch-All Broker Service and the Jisc UK Notification Router, in order to serve each other's "customers" with a wider range of subscriptions and notifications (see Figure 2). This cooperation goes in the direction of a scientific communication ecosystem where brokers conforming to common requirements for the exchange of subscriptions and notifications can interoperate to collaboratively (by granting or delegating subscriptions to the network) channel the information they collect from producers of events (OpenAIRE for the Catch-All Broker Service and publishers for the Jisc Notification Router) to interested consumers (e.g. any content provider in OpenAIRE and UK repositories for Jisc Notification Router).

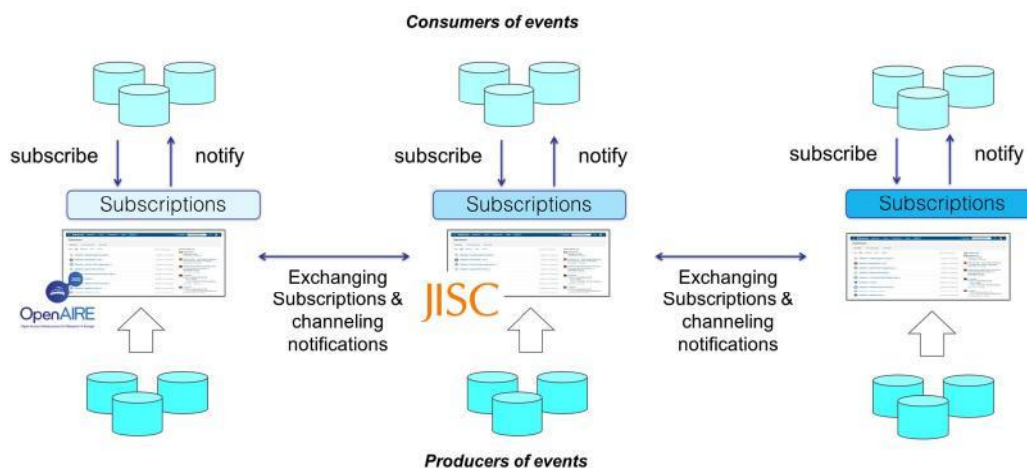


Figure 2. Interoperability between the Broker services



3. Conclusions

The effective implementation of Open Science calls for a scientific communication ecosystem capable of enabling the “Open Science publishing principles” of transparency and reproducibility. Such ecosystem should provide tools, policies, and trust needed by scientists for sharing and interlinking (for “discovery” and “transparent evaluation”) and re-using (for “reproducibility”) all research artefacts produced during the scientific process, e.g. literature, research data, methods, software, workflows, protocols.

OpenAIRE fosters Open Science by advocating its publishing principles across Europe and research communities and by offering technical services in support of Open Access monitoring, research impact monitoring, and Open Science publishing. Its aim is to provide Research Infrastructures (RIs) with the services required to bridge the research life-cycle they support (where scientists produce research artefacts) with the scholarly communication infrastructure (where scientists publish research artefacts) in such a way science is reusable, reproducible, and transparently assessable. OpenAIRE is fostering the establishment of reliable, trusted, and long lasting RIs by compensating the lack of Open Science publishing solutions and by providing the support required by RIs to upgrade existing solutions to meet Open Science publishing needs (e.g. technical guidelines, best practices, Open Access mandates). To this aim, OpenAIRE is working closely with existing RIs to extend its portfolio by implementing the concept of “Open Science as a Service” (OSaaS) and offer two new services: the Research Community Dashboard and the Catch-All Notification Broker Service.

The first beta release of the services is planned on March 2018. A set of testing sessions will be conducted by five research communities (for the Research Community Dashboard) and a number of content providers (for the Catch-All Notification Broker Service) before the first production public release, expected in June 2019.

Acknowledgments

This research was supported by EU funded project OpenAIRE-Connect (grant agreement: 731011; Call: H2020-EINFRA-2016-1). We thank our colleague Stefania Biagioni for her help during the writing of this paper.

References

1. Manghi, P., Bolikowski, L., Manold, N., Schirrwagen, J., & Smith, T. (2012). OpenairePlus: the european scholarly communication data infrastructure. *D-Lib Magazine*, 18(9), 1. <http://doi.org/10.1045/september2012-manghi>
2. Bechhofer, Sean, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch et al. "Why linked data is not enough for scientists." *Future Generation Computer Systems* 29, no. 2 (2013): 599-611. <https://doi.org/10.1016/j.future.2011.08.004>
3. Jisc Publication Router <http://www.jisc.ac.uk>
4. Artini, Michele, Claudio Atzori, Alessia Bardi, Sandro La Bruzzo, Paolo Manghi, and Andrea Mannocci. "The OpenAIRE literature broker service for institutional repositories." *D-Lib Magazine* 21, no. 11/12 (2015). <http://doi.org/10.1045/november2015-artini>
5. Hoogerwerf, Maarten, Mathias Löscher, Jochen Schirrwagen, Sarah Callaghan, Paolo Manghi, Katerina Iatropoulou, Dimitra Keramida, and Najla Rettberg. "Linking data and publications: towards a cross-disciplinary approach." *International Journal of Digital Curation* 8, no. 1 (2013). <http://doi.org/10.2218/ijdc.v8i1.257>
6. Zenodo. <https://zenodo.org>
7. Potter, M., & Nielsen, L. H. (2015, September 1). Zenodo Information Architecture and Usability. <http://doi.org/10.5281/zenodo.31870>
8. ESFRI https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri
9. COAR repository <https://www.coar-repositories.org>
10. FOSTER <https://www.fosteropenscience.eu/node/1420>
11. WDS <https://www.icsu-wds.org>
12. EOSC <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>
13. Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean?. *Science translational medicine*, 8(341). <http://doi.org/10.1126/scitranslmed.aaf5027>
14. Research Object <http://www.researchobject.org>
15. Bardi, A., & Manghi, P. (2014). Enhanced Publications: Data Models and Information Systems. *LIBER Quarterly*, 23(4), 240–273. <http://doi.org/10.18352/lq.8445>
16. RMap. <http://rmap-project.info>



A Facet-based Open and Extensible Resource Model for Research Data Infrastructures

Luca Frosini and Pasquale Pagano,

Istituto di Scienza e Tecnologie dell'Informazione (ISTI) "Alessandro Faedo"

National Research Council (CNR), Italy

Abstract

Research Data represent a valuable assets in science as well as in our society. Their management requires the development and operation of Research Data Infrastructures, i.e. complex and distributed systems specifically conceived to address the needs arising in Research Data collection, collation, processing and publishing. The development of such systems require a shared model for describing the existing "resources", namely the datasets as well as the rest of services and entities worth being considered to properly deal with the datasets. In this paper it is presented an open and extensible model based on two basic notions: Resource to describe the entities, Facet to characterize a feature of a Resource. The model enables its users to instantiate these basic concepts and define context-specific relationships among the typologies of defined resources.

Keywords: Research Data Infrastructure, Resource Model, Resource, Facet.

1. Introduction

Research Data play a key role in our society. They include both "primary dataset", i.e. data genuinely produced, as well as "derived datasets", i.e. datasets resulting by processing existing datasets. Their management requires dedicated Research Infrastructure and a description of the entire set of "resources" surrounding each dataset (e.g. other datasets, services the dataset has been produced with or suitable for "consuming" the dataset, entities responsible for the dataset).

Research infrastructure (RI) are "facilities, resources and services used by the science community to conduct research and foster innovation"¹. Grid and Cloud computing Infrastructure provides an excellent support to address Data/Computation intensive paradigm but in a certain sense they can be seen as facilities to implement such a paradigm.

Bardi and Frosini [1] highlighted the needs of researcher for digital services to realize Digital Research Infrastructures (henceforth e-Infrastructure) and highlighting the Data e-Infrastructure as one of the relevant category (in this paper we will use Data e-Infrastructure and Research Data Infrastructure interchangeably). According to Candela et al. [2] Data e-Infrastructures "integrates several technologies, including Grid and Cloud, and promises to offer the necessary management and usage capabilities required to implement the 'Big Data' enabled scientific paradigm" and promoting data sharing and consumption.

To realise this, it is key to implement an Information System (IS) capable to represent datasets as well as the rest of resources associated with it such as Services, Hardware, Actors, Facilities, and Policies. Such an IS acts as a registry offering global and partial view of the Infrastructure resources, their current status, their relationships with other resources, and the policies governing their exploitation. The "descriptions" attached to such "entities" should not be prescribed a priori, rather they should be open and extensible thus to enable diverse actors (being them publishers or consumers) to annotate each entity with specific features.

In particular, we need a model to deal with heterogeneity with respect to:

- Open-ended set of manageable resources;
- Open-ended model for describing resources;
- Diverse workflows governing registration and update of resources;

Due to such an heterogeneity the IS should have the ability to evolve with the evolving needs of the infrastructure at no cost for its clients by (a) supporting new resource types, (b) supporting evolution in the way a resource is described, (c) supporting the same resource type described by using different models.

In Section 2 it is introduced a core model (Information System Model) defining the building blocks for developing a resource model with the envisaged characteristics. In Section 3 it is presented the resource model obtained by relying on the core model to capture the entities of interest

¹ https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=about

identified in D4Science.org², a Research Data Infrastructure conceived to support several communities and data management scenarios arising in fields including biological sciences, earth and environmental sciences, agricultural sciences, social sciences and humanities.

2. The Core Model

The proposed “core model” is conceived to provide its users with the building blocks needed to develop an information system suitable for Data e-Infrastructure. This model is based on a graph model having **Entities** as nodes and **Relations** as edges.

Two typologies of **Entities** are envisaged (cf. Fig. 1): **Resources**, i.e. entities representing a description of a “thing” to be managed; **Facets**, i.e. entities contributing to “build” a description of a Resource. Every Resource is characterised by a number of Facets. Every facet, once attached to a Resource profile captures a certain aspect / characterization of the resource. Every facet is characterised by a number of properties.

Two typologies of **Relations** are envisaged: **isRelatedTo**, i.e. a relation linking any two Resources; **consistsOf**, i.e. a relation connecting each Resource with each one of the Facets characterizing it.

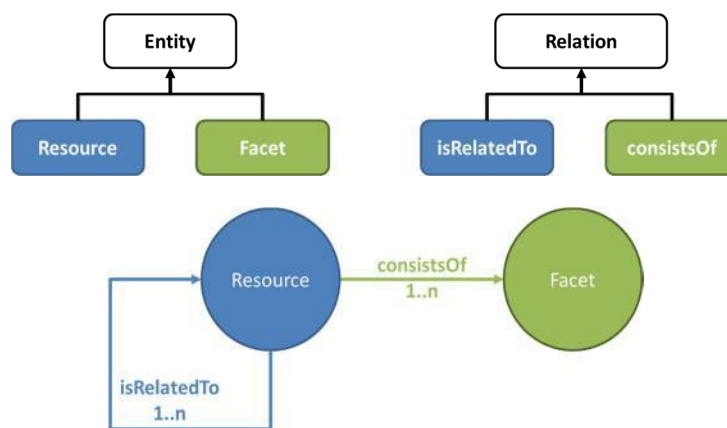


Fig. 1 represent the main concepts of the model.

Fig. 1a evidences the inheritance model of entities and relations.

Fig. 1b evidences the conceptual graph model that is realized by using entities and relations.

Each Entity and Relation: has an **Header** automatically generated for the sake of identification and provenance of the specific information and can be **specialized**. A number of Entity and Relation are expected to be defined when defining a specific information model (cf. Sec. 3). Facet and Relation instances can have additional properties which are not defined in the schema (henceforth schema-mixed mode).

Every Relation relation has - apart the Header - zero or more properties. One of these properties is the **Propagation Constraint** (see dedicated section). Any Relations has a direction, i.e. a “source” and a “target”. When inspecting the graph (e.g. at query time) relations can be navigated in both directions, i.e. from source to target and from target to source.

Facet describe a characteristic of a Resource definition, for such a reason, it is not permitted to define a Relation having a Facet as “source”. In other words: it is not permitted to define a Relation connecting a Facet with another one or Relation connecting a Facet with a Resource (as target).

² www.d4science.org

**Property**

As stated any entity and relation is characterised by some properties. A property can be described by the attributes in Table 1.

Table 1. Property attributes

Attribute name	Attribute description
Name	The name of the property
Type	The type of the property
Description	A textual description of the property.
Mandatory	A boolean flag describing the mandatory nature of the property.
ReadOnly	A boolean flag describing the alter-ability nature of the property.
NotNull	A boolean flag describing whether the value of the property must be filled with values other than null or not.
Max	The maximum acceptable value. It is significant for Numbers, Strings (intended as maximum length of the String) and Dates.
Min	The minimum acceptable value. It is significant for Numbers, Strings (intended as minimum length of the String) and Dates.
Regexpr	A Regular Expression used to validate the property value, i.e. precisely characterising the allowed values.

Types can vary from *Basic Type* (typical language programming types i.e. Boolean, Integer, Short, Long, Float, Double, Date, String, Byte and Binary (any values as byte array)); to *Embedded Types* (i.e. complex objects defined from clients and composed of two or more properties belonging to one of the Basic Types). Moreover List, Set (a list with no duplicates) and Map (a key-value pairs having a String as key and an Embedded instance as value) of Embedded can be used.

Defined Embedded Types

Embedded types are mainly composed by two or more properties belonging to *Basic Types* or to another *Embedded Type* (for clear reason recursion is not allowed).

Header

As already stated, every Entity and Relation has an header automatically created and updated by the System. The header is composed by the following properties:

- **uuid** (*String*) [Mandatory=true, NotNull=true, ReadOnly=true] *Regex*=`^[a-zA-Z0-9]{8}-[a-zA-Z0-9]{4}-[a-zA-Z0-9]{4}-[a-zA-Z0-9]{4}-[a-zA-Z0-9]{12}}{1}$`;): this uuid can be used to univocally identify the Entity or the Relation;
- **creator** (*String*) [Mandatory=true, NotNull=true, ReadOnly=true] : the individual or service which create the Entity of Relation;
- **modifiedBy** (*String*) [Mandatory=true, NotNull=true] : the individual or service which modified the last time the Entity of Relation. At creation time it assumes the same value of creator;
- **creationTime** (*Date*) [Mandatory=true, NotNull=true, ReadOnly=true] : creation time in milliseconds. Represent the difference, measured in milliseconds, between the creation time and midnight, January 1, 1970 UTC;
- **lastUpdateTime** (*Date*) [Mandatory=true, NotNull=true] : last Update time in milliseconds. Represent the difference, measured in milliseconds, between the last update time and midnight, January 1, 1970 UTC.

PropagationConstraint

As already stated, each Relation has a propagation constraint which indicates the behavior to be held on a target entity when an event occur in the source entity (please note that the source entity of a relation is always a Resource by Relation definition). The following two are envisaged:

- **remove** (Enum) *Regex*=(*cascadeWhenOrphan*/*cascade*/*keep*) : i.e. indicate the behaviour to implement for the target Entity when a remove action is performed on the source Resource.
- **add** (Enum) *Regex*=(*propagate*/*unpropagate*) : i.e. indicate the behaviour to implement for the target Entity when a add action is performed on the source Resource. The default values of the propagation constraints for the basic relations are the following:
- **consistsOf**: remove=cascadeWhenOrphan, add=propagate;
- **isRelatedTo**: remove=keep, add=unpropagate.

isIdentifiedBy

This **consistsOf** specialization is a relation connecting each Resource with one of the Facet which can be used to identify the Resource. Each Resource must have at least one **isIdentifiedBy** relation. Moreover every Resource can decide to define the type of target Facet for such a relation.

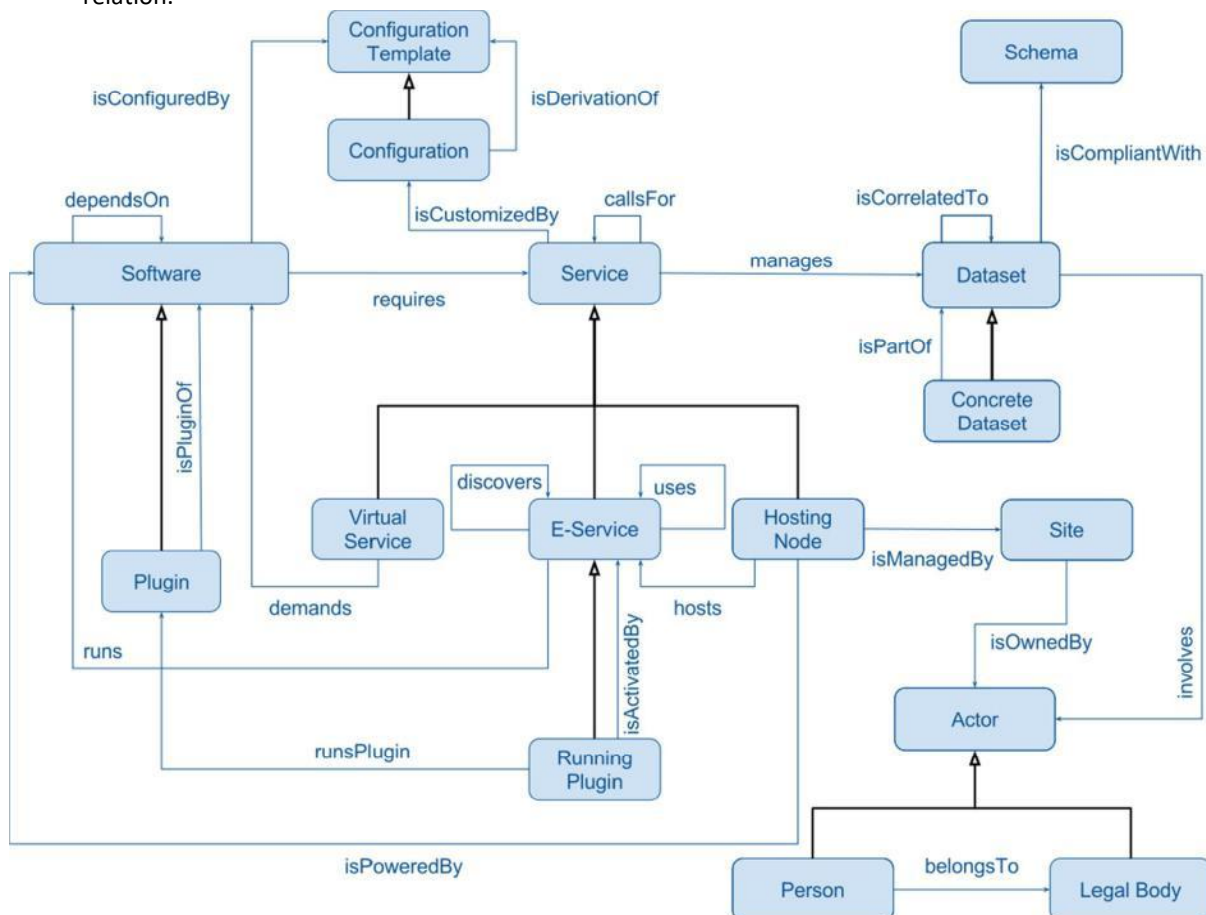


Fig. 2 Overview of gCube Model Resources and (isRelatedTo) Relations

3. The Resource Model

An extended resource model built on top of IS Model which aim to capture the different aspect of the most common resources needed in a Research Data Infrastructure has been defined. The model get the name of gCube Model. gCube is an open-source software toolkit used for building and operating data infrastructures [2] enabling the data sharing and reuse [3]. gCube is developed and maintained by CNR-ISTI.

3.1 Resources

A Resource Type can be identified as Abstract. This means that cannot be instantiated. It is expected that one of its specializations are instantiated. It is not required that an Abstract class establishes an **isIdentifiedBy** relation with a Facet.

Seven resource typologies have been identified and defined (cf. Fig. 2), namely *Dataset*, *Actor*, *Schema*, *Configuration Template*, *Site*, *Service*, and *Software*. In some cases these typologies have



been further specialized to capture specific entities, e.g. Concrete Dataset is a sub-type of Dataset, E-Servcie is a subtype of Service. In the reminder of this section the defined Resources are described.

Dataset

A **DataSet** is any set of digital objects representing data and treated collectively as a unit. It is not only the key resource of a Research Data Infrastructure but we could affirm that is the reason why a Research Data Infrastructure is created. It is characterized by the following facets:

- **Identifier Facet** (associated with *isIdentifiedBy*) : this facet captures information on Identifiers (other than the ones automatically generated by the system) that can be attached to the dataset, e.g. for discovery purpose;
- **Contact Facet** : this facet captures information on a point of contact for the dataset, i.e. a person or a department serving as the coordinator or focal point of information concerning the dataset. There are diverse points of contact that can be associated to the dataset and the role of the association is captured by using a specific *consistsOf* relation e.g. to represent the owner, the responsible, the creator, the curator, the maintainers and any contributors of the dataset;
- **Access Point Facet** : this facet captures information on an “access point” of a dataset, i.e. any web-based endpoint to programmatically interact with the resource via a known protocol. It represent the access point to use for having access to the dataset. The embargo state can be modeled through access policy defined in the *consistsOf* relation;
- **License Facet** : this facet captures information on any license associated with the dataset to capture the policies governing its exploitation and use. The duration of license (if any) can be captured expiry date property defined in the *consistsOf* relation;
- **Provenance Facet** : this facet captures information on provenance/lineage of associated with the dataset;
- **Coverage Facet** this facet captures information on the *extent* of the dataset, i.e. any aspect aiming at capturing an indicator of the amount/area the resource covers be it a geospatial area, a temporal area, or any other "area". Any temporal coverage information characterising the content of the dataset, e.g. the time-span covered by the dataset can be described with this facet associated to the dataset with a specific *consistsOf* relation. Any geospatial coverage information characterising the content of the dataset, e.g. the area covered by the dataset can be described with this facet associated to the dataset with a specific *consistsOf* relation;
- **Descriptive Metadata Facet** : this facet captures information on descriptive metadata to be associated with the dataset, e.g. for discovery purposes;
- **Subject Facet** : this facet captures information on subjects associated with the dataset for descriptive and discovery purposes.

A **Dataset** can be correlated to another **Dataset** (see *isCorrelatedTo* relation Fig. 2).

Concrete Dataset

Concrete Dataset (specialization of *Dataset*) is any incarnation/manifestation of a dataset or part of it. It is characterized by the following facets:

- **Identifier Facet** (associated with *isIdentifiedBy*) : the set of Identifiers associated with the concrete dataset instance;
- **Contact Facet** : the contact information of the entity responsible for the maintenance of the concrete dataset;
- **Access Point Facet** : the access point to use for having access to the concrete dataset.

A **Concrete Dataset** is part of a **Dataset** (see *isPartOf* relation Fig. 2).

Actor

Actor (*Abstract*) is any entity (human or machine) playing an active role in a Research Data Infrastructure. It is characterized by the following facets:

- **Contact Facet** (associated with *isIdentifiedBy*) : an Actor has at least a Contact Facet which permit to identify the Actor per se. An Actor can have other Contact Facets which provide secondary contact information;
- **Contact Reference Facet** : this facet captures information on the primary and authoritative contact for the resource it is associated with.

This Resource is used from **Dataset** to indicates the involved **Actors** by using specialization (see *involves* relation in Fig. 2).



Legal Body

A **Legal Body** (specialization of **Actor**) is any legal entity playing the role of an Actor.

Person

A **Person** (specialization of **Actor**) is any human playing the role of Actor.

Please note that a person can belongs to a legal body (see **belongsTo** relation Fig. 2).

Schema

Schema any reference schema to be used to specify values compliant with it. Examples include controlled vocabularies, ontologies, etc. It is characterized by the following facets:

- **Schema Facet** (associated with **isIdentifiedBy**) : this facet captures information on any schema associated with a resource. There are diverse type of schema that can be associated to the schema each one is capture by a dedicated schema facet specialization i.e. **JSON Schema Facet**, **XML Schema Facet**;
- **Contact Facet** : this facet captures information on a *point of contact* for the Schema;
- **Descriptive Metadata Facet** : this facet captures information on descriptive metadata to be associated with the schema, e.g. for discovery purposes;
- **Subject Facet** : this facet captures information on subjects associated with the schema for descriptive and discovery purposes.

This resource is mainly used by **Dataset** to evidence that is compliant with a **Schema** (see **isCompliantWith** relation Fig. 2).

Configuration Template

Configuration Template represents a template for a configuration. It describe how a configuration has to be realized. E.g. Used to define the accounting configuration parameters template. It is characterized by the following facets:

- **Identifier Facet** (associated with **isIdentifiedBy**) : the set of Identifiers associated with the configuration template instance;
- **Simple Property Facet** : This facet captures information on any property by a simple name-value pair.

Configuration

Configuration (specialization of **Configuration Template**) an instance of a configuration template characterising the behaviour and shape of the resource it is attached to.

The Configuration can be related to the template it derives to (see **isDerivationOf** relation Fig. 2).

Site

Site is an entity representing the location (physical or virtual) hosting and providing the resources associated with it. It is characterized by the following facets:

- **Identifier Facet** (associated with **isIdentifiedBy**) : the set of Identifiers associated with the site instance;
- **Contact Facet** : There are diverse points of contact that can be associated to the site and the role of the association is captured by using a specific *consistsOf* relation e.g. to represent the manager and the maintainers of the site;
- **Location Facet** : this facet captures information on a physical area characterising the resource it is associated with e.g. the gps coordinates of the site or the geographical address of the site. This should not be confused with Coverage Facet.
- **Networking Facet** : this facet captures information on any (computer) network interface/access point associated with the resource. A site has one or more ip subnet to address the machines in the site.

Any Site is owned by an Actor (see **isOwnedBy** relation Fig. 2).

Service

Service (*Abstract*) represents any typology of Service worth registering in the infrastructure. It is characterized by the following facets:

- **Descriptive Metadata Facet** : any descriptive information associated with the service, e.g. for discovery purposes;
- **Subject Facet** : any subject / tag associated with the service for descriptive, cataloguing and discovery purposes;



- **Capability Facet** : this facet captures a defined facility for performing a specified task supported/offered by a given Service.

Any specializations of **Service** can: manage a dataset or its specialization such as concrete dataset (see **manages** relation Fig. 2); be customized from a **Configuration** (see **isCustomizedBy** relation Fig. 2); require another **Service** to properly operates (see **callsFor** relation Fig. 2);

E-Service

E-Service (specialization of **Service**) is any working service that is registered in the infrastructure and made available by an Access Point. It is characterized by the following facets:

- **Software Facet** (associated with **isIdentifiedBy**) : this facet captures information on any software associated with the resource. The one associated with **isIdentifiedBy** represent the main software enabling the E-Service capabilities (this facet is the one identifying the E-Service);
- **Software Facet** : software available in the E-Service environment that characterizes the specific E-Service instance;
- **Access Point Facet** : identify the endpoints of the E-Service;
- **Event Facet** : this facet captures information on a certain event / happening characterising the current status and the life cycle of the E-Service events (e.g. Activation Time, Deployment Time);
- **Service State Facet** : this facet captures information on the current operational state of the E-Service it is associated with (e.g. started, ready, down, failed);
- **License Facet** : this facet captures information on any license associated with the E-Service to capture the policies governing its exploitation and use.

Any E-Service or its specializations can be related with other E-Service because it: discovers another E-Service for example to check the availability (see **discovers** relation Fig. 2); uses another E-Service to accomplish its tasks (see **uses** relation Fig. 2).

Please note that both relations are specializations of **callsFor** relation.

Running Plugin

Running Plugin (specialization of **E-Service**) is any instance of a Plugin deployed and running by an E-Service. This knowledge is expressed by **isActivatedBy** relation (see Fig. 2).

Hosting Node

Hosting Node (specialization of **Service**) is any server\machine playing the role of "Hosting Node", i.e., being capable to host and operate an E-Service. This knowledge is expressed by **hosts** relations (see Fig. 2). Hosting Node is characterized by the following facets:

- **Networking Facet** (associated with **isIdentifiedBy**) : this facet captures information on any (computer) network interface/access point associated with the resource. It define the Network ID characterising the Hosting Node;
- **CPU Facet** : this facet captures information on the Central Processing Unit of the resource it is associated;
- **Memory Facet** : this facet captures information on computer memory equipping the resource and its usage such as the persistent memory (i.e. the Disk Space Capacity of the Hosting Node) or the volatile memory (the RAM Capacity of the Hosting Node);
- **Event Facet** : this facet captures information on a certain event / happening characterising the life cycle of the Hosting Node, e.g. the activation time;
- **Container State Facet** : this facet captures information on the operational status of the Hosting Node (e.g. started, ready, certified, down, failed);
- **Simple Property Facet** : this facet captures information by a simple <key, value> pair property worth associating with the Hosting Node, e.g. Environment Variables;
- **Software Facet** : this facet captures information on any software associated with the Hosting Node. Useful to report the hosted software such as the operating system.

Any hosting node is located in a site which provides the management facilities to create, maintains and dismiss it. This knowledge is expressed by **isManagedBy** relation (see Fig. 2).

Virtual Service

Virtual Service (specialization of **Service**) is an abstract service (non physically existing service) worth being represented as an existing Service for management purposes. Examples of usage include cases where classes or set of services are to be managed like an existing unit.

Software



Any **Software** entity worth being represented for management purposes. It is characterized by the following facets:

- **Software Facet** (associated with **isIdentifiedBy**) : Software coordinates which identify the Software per se;
- **Software Facet** : Apart the one connected by the *isIdentifiedBy* relation the others identify the sw in other way e.g. (Maven coordinates);
- **Access Point Facet** : identify endpoint useful for software download, documentation, source code etc e.g. links to maven artifact on public maven repositories, javadoc, wiki, svn;
- **License Facet** : the Software License characterizing its possible exploitation and use eg EUPL, LGPL, GPL, Apache2;
- **State Facet** : This facet captures information on state to be associated with the resource. State is captured by any controlled vocabulary which is an integral part of the facet e.g. Deprecated, Active, Obsolete;
- **Capability Facet** : any facility supported/offered by the Software.

Any Service or its specializations can : depends on other software (see **dependsOn** relation Fig. 2); be configured by a configuration template (see **isConfiguredBy** relation Fig. 2); Moreover: Any E-Service runs a certain software (see **runs** relation Fig. 2); Any hosting node provides its capabilities thank to a certain software (see **isPoweredBy** relation Fig. 2).

Plugin

A piece of Software extending the capabilities of another Software (main) and requiring the main Software to be executed. The relation between main software and plugin is expressed by **isPluginOf** relation (see Fig. 2).

4. Conclusion

Research Data Infrastructures are complex systems called to offer services for Research Data Management. In order to meet this goal their developers and managers as well as their constituents (systems on its own) need to be provided with a constantly update and comprehensive description of the datasets to be managed and the associated resources (e.g. other datasets, services, people, machines) that are “available” at a given point in time. Capturing this information need poses a number of challenges including to deal with the heterogeneous and evolving nature of both the typologies and descriptions of the resources of interest. This paper described both a core model and a comprehensive model to capture the information needs arising in a Research Data Infrastructure when managing Research Data.

Such a model has been used by D4Science infrastructure in the context of BlueBRIDGE and PARTHENOS european projects.

Acknowledgments This work has received funding from the European Union’s Horizon 2020 research and innovation programme under the BlueBRIDGE project (Grant agreement No. 675680) and the PARTHENOS project (Grant agreement No. 654119).

References

- [1] A. Bardi, L. Frosini : Building a Federation of Digital Humanities Infrastructures. ERCIM News 111. October 2017.
- [2] L. Candela et al.: “Managing Big Data through Hybrid Data Infrastructures”, in ERCIM News, Issue 89, April 2012.
- [3] L. Candela, P. Pagano: “Cross-disciplinary data sharing and reuse via gCube”, in: ERCIM News, Issue 100, January 2015.
<https://kwz.me/h07>



D4Humanities: Deposit of Dissertation Data in Social Sciences & Humanities – A Project in Digital Humanities

Joachim Schöpfel, GERiCO Laboratory, University of Lille
Hélène Prost, CNRS, associate GERiCO Laboratory, France

Abstract

Following our work on research data and electronic theses and dissertations since 2013, we are conducting a new research project between 2017 and 2018 called D4Humanities with three objectives – to develop the research data management and stewardship on our campus, to gain better insight into the nature of research data in social sciences and humanities and to produce empirical evidence on the development of dissertations. In particular, the project contains three components:

- 1. Qualitative survey on behaviours and knowledge in the field of research data with 50 scientists from the University of Lille Social Sciences and Humanities Department, with a special focus on the FAIR guiding principles of scientific data management and stewardship.*
- 2. The creation of a workflow for the submission of research data related to PhD dissertations (deposit, preservation and dissemination of data via the NAKALA service Huma-Num)*
- 3. Two conceptual studies on the definition and typology of research data in SSH and on the development of dissertations in the environment of e-Science and Open Science (content, format, structure, requirements).*

In the following we present some preliminary results, in particular from the survey and from the conceptual studies, in order to enhance the understanding of research data in SSH and of the development of dissertations.

Acknowledgment: The project receives funding from the European Institute of Social Sciences and Humanities (MESHS Lille) and from the Regional Council (Conseil Régional Hauts-de-France). We would like to thank the D4Humanities project team for their contribution to the research underlying this paper, in particular Cécile Malleret, Eric Kergosien and Leslie Hyacinthe.

Keywords PhD dissertations, research data, digital humanities, open access, open science, social sciences and humanities

Introduction

For more than ten years, one part of our professional and scientific work has been focused on PhD dissertations as one major document type of academic grey literature. We started with research on their production and findability (Paillassard et al. 2005) and then moved on to questions related to their accessibility, especially in the environment of electronic theses and dissertations (ETD), open access (OA) and institutional repositories (Schöpfel 2013, Schöpfel & Prost 2013, Schöpfel et al. 2015c). Three years ago, we began to study the research data produced by PhD students and submitted as complementary material together with the dissertation (Schöpfel et al. 2014), trying to establish the link between grey literature and e-Science in the field of ETDs. Our first questions were (are) operational: what could and should be done with this material, how can it be stored and preserved for a longer time, what are the conditions for sharing, publishing and reuse? However, these practical questions always included conceptual elements, about the definition and typology of data, about their identification and description, about their relationship with dissertations, and about the development of dissertations themselves and their potential for reuse with content mining tools. Because of the complexity of the field, we limited our research to the disciplines of social sciences and humanities (SSH).

In 2017 we launched a two-year project called *D4Humanities*¹ in order to transform our research work into operational service development on the campus and to enhance our knowledge of data and dissertation. The basic question is how to enable the exploration of research data in social sciences and humanities (textual or oral corpus, raw data, images...) with digital technologies (text and data mining, mapping, visualization ...) to convey a new meaning? The project *D4Humanities* is part of the Digital Humanities and a continuation of the recent research of the GERiCO laboratory and its partners at the University of Lille Humanities and Social Sciences (academic library, SSH graduate school, digitization centre ANRT...) with the objective of

¹ Deposit of Dissertation Data in Social Sciences and Humanities. A Project in Digital Humanities <https://d4h.meshs.fr/>



accelerating the research data management project in particular for PhD students and young researchers, and of fostering the preparation of an international research project.

We started our project in March 2017, and it will continue until fall 2018. So what we will do here is deliver some preliminary results on data behaviour and data management, including the development of a workflow for ETD related datasets, and first conceptual work on data and dissertations. This will be followed by an invitation to join our research consortium.

Data literacy (survey)

In 2015, we conducted a campus-wide survey at the University of Lille on research data management in social sciences and humanities. The survey received 270 responses, equivalent to 15% of all scientists, scholars, PhD students and administrative and technical staff; all disciplines were represented. The responses showed a wide variety of data, practice and usage; some differences seem related to job status and disciplines. Generally, 20-25% of the sample can be considered as pioneers in data management and sharing, and 25-30% are motivated; only 5-10% appear reluctant to make their data available (Schöpfel & Prost 2016).

On the basis of the results of this first survey, we prepared a small qualitative survey with academic “volunteers” on the Lille SSH campus, among researchers and PhD students from various disciplines. We wanted to gain more insight in personal research data management behaviour and data literacy, in particular those contributing to the compliance with the FAIR principles for data management (Wilkinson et al. 2016). The investigation is not over; for the moment, we have conducted 27 interviews with researchers from history, archaeology, literature and language studies, psychology and information sciences. First results and comments:

Interest and motivation: finding volunteers on the campus was not easy this time; obviously, for many colleagues RDM is not a “hot topic” to spend one hour or more in a semi-directive interview on data practice and literacy. At least, it does not appear as priority or relevant.

Funding agencies: one half of the volunteering respondents (14) has conducted or participated in one or more research projects funded by the European Commission (H2020 program) and/or the French National Research Agency (ANR programs). But only 10 have knowledge of requirements (such as of the H2020 program), guidelines or recommendations for RDM.

Privacy: 13 respondents use or produce personal data as defined by the French CNIL commission, or confidential data. 6 submitted a research protocol to the university's ethics committee.

Standards, description: 8 participants reported assigning codes to their data, 9 people have already drafted a data management plan, and 5 participants follow standards for describing their data.

Dissemination and sharing: data collection, analysis and storage are often carried out by the researcher him/herself or together with the research team. 16 participants agree to share their data with others, which means above all with other colleagues from the project team. 10 participants have already submitted their data to an online server, 2 others intend to do so; only one refuses for security reasons.

Need for advice: generally, the respondents need advice on querying databases, formatting and naming data; they seek advice on licensing and legal protection of sensitive data; they want to know more about the services offered by the deposit platforms. So far, they have been seeking advice on RDM not at the library but with people from the IT department (system security, storage) and from the ethics committee.

Need for data services: the services requested by the researchers relate mainly to the different aspects related to data storage: to know what data to store, under which formats, on which server, with which guarantees of duration and security. They want to encourage exchanges between researchers and information professionals.

So far, we have observed very large differences between disciplines and research domains, but also between research methods and tools in the same field. Some scientists have a long experience with RDM and apply standard and transparent data procedure, even if they do not always call it RDM. This data literacy can mainly be explained by legal issues (privacy laws, especially in psychology, education, sociology, and projects in public health) or ethics rules, less (up to now) by requirements from funding agencies. However, application of standards in RDM remains exceptional, such as data publishing and sharing. We did not encounter significant reluctance or even opposition to RDM and data sharing, but rather ignorance or lack of interest.

Data workflow

Similar to other ETD projects² we are developing a local workflow for the deposit of research data by PhD students. The main characteristics of this workflow are:

- Data and dissertations are submitted on different servers,
- The local deposit is interconnected with existing infrastructures, in particular with the French SSH data platform NAKALA,
- Data and dissertations are stored and preserved on various platforms but linked via their metadata and identifiers.

Figure 1 shows the workflow and the separation of data and dissertations from the beginning on (deposit). The guiding principle was to provide an interface (with technical assistance) on our campus for the deposit of research data on the NAKALA platform of the national infrastructure for SSH communities. For a detailed description, see Schöpfel et al. (2017b).

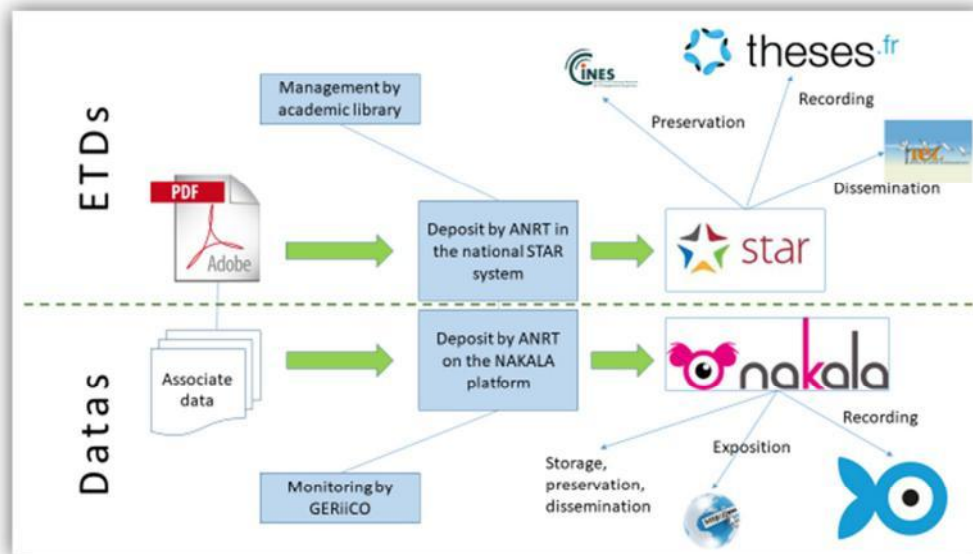


Figure 1: Local ETD/data workflow

Our intention is to offer young scientists a “default solution”, complementary to existing disciplinary data repositories, accompanied by technical assistance and a PhD training program for research data management delivered by our Graduate School.

The preparation and development of the workflow raised several issues, some of them familiar to the grey community:

- Granularity: what exactly should be defined as a dataset for deposit? We have discussed this question in two communications (Schöpfel et al. 2016, 2017a). There are no clear rules or guidelines. The pragmatic solution is to accept datasets on a granularity level which makes sense for understanding (validation) and reuse, and to allow deposit of dataset collections with a hierarchical structure.
- Data structure and description: how are data to be described and structured? Our option is to apply the Metadata Encoding & Transmission Standard of the Library of Congress.
- Identifier: which unique identifier should be used for the datasets? Even if France is part of the DataCite consortium for the assignment of DOIs, we opt for the handle system which is applied by the Huma-Num infrastructure but remain open for future adoption of the DOI.
- Legal aspects: we anticipate legal issues like copyright, third party rights, privacy etc. Our approach is twofold: we provide legal advice as part of the library's data service, and we ask the students to provide a declaration (template) that they have the permission to upload the datasets on NAKALA.
- Quality: the question was raised about the quality of datasets. Should all datasets provided by PhD students be accepted? Should we set up a kind of validation procedure? If so, which criteria should be applied? Who should evaluate? For the moment, we will not filter submitted data files otherwise than by formal criteria (size, format...), similar to other projects and data repositories. But the question remains open.

² For instance, the ETDplus project funded by Educopia <https://educopia.org/research/grants/etdplus> and the workflow at the University of Bielefeld, see Vompras & Schirrwagen (2015)

The tests of the new workflow started end of September. The workflow will be operational in 2018.

Data definition

But what exactly are data and datasets? The issue was raised during the preparation of the data workflow. Therefore, we carry out a conceptual analysis of the meaning and content of the term of research data as a vital complement to the workflow development and survey. The first results were presented during a workshop at the University of Toulouse in May 2017 (Schöpfel et al. 2017a). Figure 2 resumes the main characteristics of our approach which is based on a synthesis of recent French and international reviews and definitions.

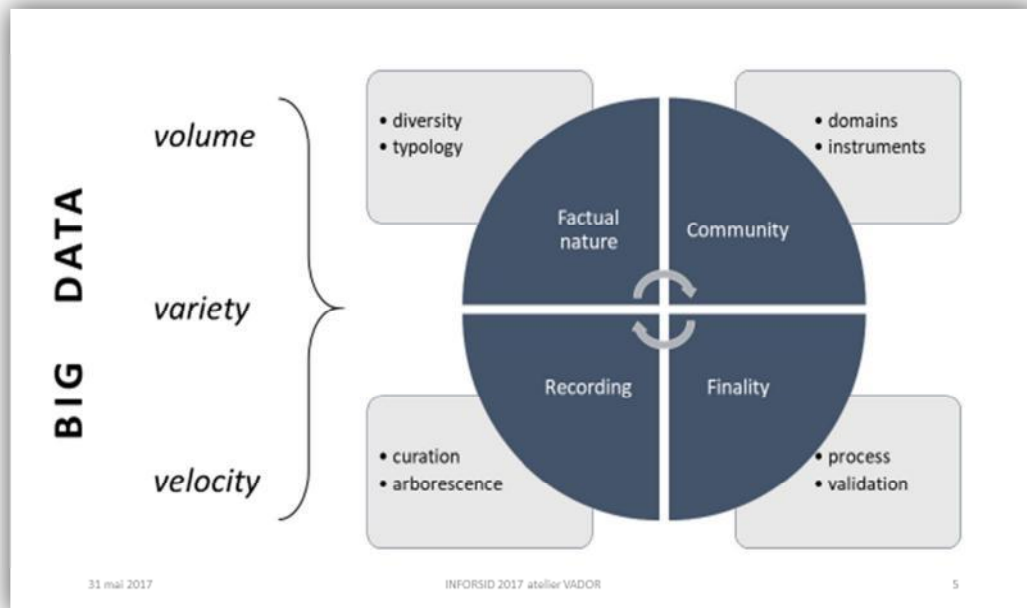


Figure 2: Elements of a data definition

We can identify five key elements of research data:

1. Link to the concept of big data: even if one part of research data is considered as small or "smart" data, the link with the "3 V's" of the big data is always present³, in particular the diversity of data, their large number and size and the continuous stream of data input and output.
2. Factual nature: definitions of research data often insist on their factual nature, at least implicit, as primary material in need for processing, analysis and interpretation. This often implies a more or less detailed typology of data.
3. The link to community: research data are embedded in disciplinary and institutional context, are specific to large instruments, research infrastructures and scientific domains.
4. Finality: research data are also embedded in the research process (cycle), are dynamic, with different functions and requirements. The most basic distinction is between input and output, primary and secondary data, data as resources and data as results of scientific work. Among the various functions, the most important (in a mainly STI and library perspective) are the validation of results and hypotheses (replication) and their preservation along with publications.
5. Recording: the need for recording (curation, preservation) is the last key element of research data definitions. Part of the research data management, data curation raises issues like granularity, identification and data arborescence (hierarchical structure of data and datasets).

³ See the consensual definition on big data by De Mauro et al. 2016: "Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value".



Actually, we complete this synthesis with an assessment of the re3data⁴ typology, their distribution and definition especially in the field of social sciences and humanities. Special attention is paid to the content of large data types (raw data, images) and the “other” categories of the more than 500 repositories in SSH.⁵

Data impact: evolution of PhD dissertations

The fourth and last work package of the D4Humanities project is intimately associated with the research and debates in the grey community. Our question is: how does the new environment of research data management and text and data mining impact the characteristics and requirements of ETDs? The discussion is open whether or not PhD dissertations should still be considered as grey literature in the digital age and how (Schöpfel & Rasuli 2017); but it seems obvious that the potential of text and data mining and the availability of datasets related to dissertations will have (or already have) substantial effects on the writing, content, format, length and submission and processing of dissertations, perhaps even on their legal status and licensing.

In the past years, we tried to assess which kind of data are related to PhD dissertations, especially in social sciences and humanities, how they are linked to the dissertation and how they should be curated (Prost et al. 2015, Schöpfel et al. 2015a, b); furthermore, we started to re-examine the meaning of dissertations in the light of text and data mining, considering dissertations as data (Schöpfel et al. 2016). Content mining tends to make the borders between text and data increasingly blurred, even insignificant, and revives the discussion on the distinction between publications (documents) and data.

The D4Humanities project contributes to this research field from a special perspective, i.e. the guidelines, prescriptions and laws ruling the writing and submission of digital PhD dissertations. In 2018, the project team will conduct a landscape study together with academic and corporate partners, including a state of the art on recent research and papers on dissertations and data and a small-scale survey on the development of PhD prescriptions.

Perspectives

This last work package is just a beginning. In fact, its objective is threefold:

1. An overview on ongoing research in order to define questions and hypotheses for further research.
2. The setting up of a scientific consortium around a core project team (GERiiCO laboratory at Lille and Institute of Scientific Networking at Oldenburg).
3. And third, the preparation of an international research project on new forms of PhD dissertations, with European (H2020) or French-German funding (ANR/DFG). For the time being, the project's code name is *xDiss*, for “Special Dissertations”.

Therefore our conclusion is an appeal to the members of the grey community: if you are interested, contact us and join our consortium.

⁴ The international registry of research data repositories, available at <http://www.re3data.org/>

⁵ See Kindling et al. (2017) for some general elements of these repositories. Our own results are stored on a wiki and available on request at <http://d4hdata.pbworks.com>



References

- 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.
- Jacquemin, B., Prost, H., Schöpfel, J., Severo, M., Thiault, F., 2013. Ouvrir les données de la recherche pour la veille scientifique. Le cas des thèses électroniques. In: *VSST'2013*, Nancy, 23-25 octobre 2013.
- Kindling, M., et al., 2017. The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine* 23 (3/4).
- De Mauro, A., Greco, M., Grimaldi, M., Apr. 2016. A formal definition of big data based on its essential features. *Library Review* 65 (3), 122-135.
- Paillassard, P., Schöpfel, J., Stock, C., 2005. How to get a French doctoral thesis, especially when you aren't French. *Publishing Research Quarterly* 21 (1), 73-93.
- 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+.
- 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.
- Schöpfel, J., 2013. Adding value to electronic theses and dissertations in institutional repositories. *D-Lib Magazine* 19 (3/4).
- 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)).
- 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.
- Schöpfel, J., Juznic, P., Prost, H., Malleret, C., Cesarek, A., Koler-Povh, T., 2015a. Dissertations and data (keynote address). In: *GL17 International Conference on Grey Literature*, 1-2 December 2015, Amsterdam.
- Schöpfel, J., Prost, H., Malleret, C., 2015b. 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.
- Schöpfel, J., Prost, H., Piotrowski, M., Hilf, E. R., Severiens, T., Grabbe, P., 2015c. A French-German survey of electronic theses and dissertations: Access and restrictions. *D-Lib Magazine* 21 (3/4).
- Schöpfel, J., Kergosien, E., Chaudiron, S., Jacquemin, B., 2016. Dissertations as data. In: *ETD2016*, Lille 11-13 July 2016.
- Schöpfel, J., Prost, H., 2016. Research data management in social sciences and humanities: A survey at the University of Lille 3 (France). *LIBREAS. Library Ideas* 29, 98-112.
- Schöpfel, J., Prost, H., Rebouillat, V., 2016. Research data in current research information systems. In: *CRIS 2016*, St Andrews, 8-11 June 2016.
- Schöpfel, J., Kergosien, E., Prost, H., 2017a. « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse. In: *Atelier VADOR : Valorisation et Analyse des Données de la Recherche, INFORSID 2017*, 31 mai 2017 Toulouse (France).
- Schöpfel, J., Prost, H., Malleret, C., 2017b. Research and development in the field of research data and dissertations. The D4Humanities project at the University of Lille (France). In: *10th Conference on Grey Literature and Repositories*, National Library of Technology (NTK), 19 October 2017, Prague, Czech Republic.
- Schöpfel, J., Rasuli, B., 2017. Are electronic theses and dissertations (still) grey literature in a digital age? a FAIR debate. *The Electronic Library* 35 (4).
- 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.
- Wilkinson, M. D., et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, sdata201618+.



Video is the new Grey

Bastian Drees and Margret Plank,
National Library of Science and Technology, Germany

Abstract

Conference reports are a crucial source of information as they document the current state of research in a scientific community. Additionally, it is becoming more and more common practice to record and publish conference talks. These videos are an important element of contemporary scientific output. However, no sustainable standard has yet been established for handling these media types. While libraries have well-established procedures for collecting textual conference reports as part of the grey literature, comparable procedures for audio-visual conference recordings have not yet been established.

1. Introduction

Conference reports and conference proceedings are a crucial source of information as they document the current state of research in a scientific community. Moreover, these documents are mostly grey literature. However, in addition to printed conference proceedings, it has become more and more common practice to record and subsequently publish conference talks. As such, these videos are also an important element of contemporary scientific output and thus part of the cultural heritage. However, no sustainable standard has yet been established for handling these documents. Essentially, all of these videos are either published on commercial platforms like YouTube or Vimeo, on the conference webpage or not at all. Therefore, they are truly grey material.

While libraries have well-established procedures for collecting textual conference reports as part of the grey literature, comparable procedures for audio-visual conference recordings have not yet been established. On the other hand, according to the Meeting & EventBarometer 2016 [GCB 2016], more than half of the hosts and organizers express a need for action in setting up virtual platforms that complement the real life event.

Against this backdrop, we conducted an analysis of the needs and demands of conference hosts, organizers and service providers regarding audiovisual recordings. Qualitative interviews were conducted among 36 respondents to find out how widespread conference recordings are in engineering and the natural sciences. The aim was to determine the status quo and future demand concerning the production, publishing and re-use of conference videos. Furthermore, we wanted to obtain information on the use of these materials and about potential needs for support.

Here we present the results of the interviews (section 2) and report how these findings are used to improve the existing services of TIB and to extend the range of services offered (section 3). Regarding the first aspect, the results of this study are especially used to improve the workflows and services offered in the TIB AV-Portal (<https://av.tib.eu>), a web-based platform for quality-tested scientific videos such as conference recordings. Regarding the second aspect, TIB has decided to establish a conference recording service which will start operating in 2018.

2. Results of the Study¹

Building on a large scale study performed by TIB regarding the *information procurement and publishing behaviour in science and technology* [Einbock 2017], we conducted an analysis of the needs and demands of conference hosts, organizers and service providers regarding audiovisual recordings. For this purpose, we prepared a questionnaire as an interview guide that was used in the interviews. The questionnaire asks for the status quo in production and publication of conference recordings, as well as for problems and requirements. Qualitative interviews were conducted among 36 respondents to find out how widespread conference recordings are in engineering and the natural sciences. Furthermore, we wanted to obtain information on the use of these materials and about potential needs for support. The 36 respondents can be divided into three different groups, namely conference hosts, conference organizers and service providers for audiovisual recordings, i.e. non-profit or commercial video production services. Among the respondents, 20 were in the first group (hosts), ten in the second (organizers) and six in the third group (AV service providers; cf. Fig.1).

¹ Survey results available at: <https://doi.org/10.22000/64>

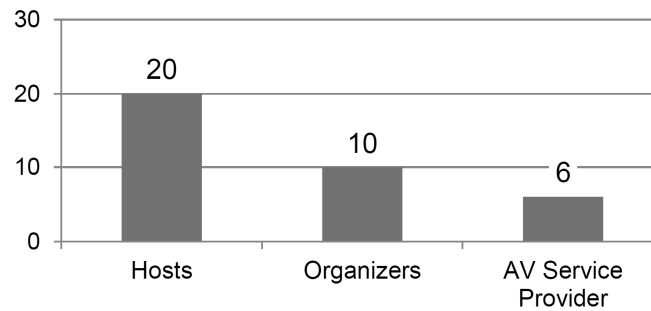


Fig. 1: The 36 respondents can be divided into three different groups, namely conference organizers, conference hosts and service providers for audiovisual recordings. Among the 36 respondents, 20 are conference hosts, ten are conference organizers and six service providers for audiovisual recordings.

2.1 Status quo of conference recordings

We asked the 30 conference hosts and organizers whether they already recorded the conferences in the past or if recordings were planned for the future (note: AV service providers were not asked this question as all of them produce video recordings). Almost half (47%) of the respondents stated that they already produced conference recordings (40%) or are planning to do so in the future (7%); see Fig.2 (top). The results differ largely depending on the subject areas of the conference. Ranging from 33% in chemistry to 89% in computer science; cf. Fig.2 (bottom).

Interestingly, these numbers show a similar trend as can be observed in the level of dissemination of open access publications in the respective field [Köhler 2017, Piwowski 2017, Archambault 2014]. Köhler 2017 reports that biology and medicine have the widest coverage of open access publications followed by mathematics and physics and chemistry being the discipline with the fewest open access articles. While this agrees with the ranking order observed in our survey, two disciplines differ from this order. Engineering and computer science are disciplines in which we found conference recordings to be widespread while Köhler reports that these disciplines have fewer Open Access publications. This correlation may be explained by the existence of different scientific cultures regarding the free sharing of research results. The two outliers may be explained in the way that both disciplines have a strong focus on conferences and therefore also have a larger number of conference recordings. However, the reasons for the subject-specific differences in both open access publications and conference recordings need to be examined more closely in order to be able to make reliable statements about the topic.

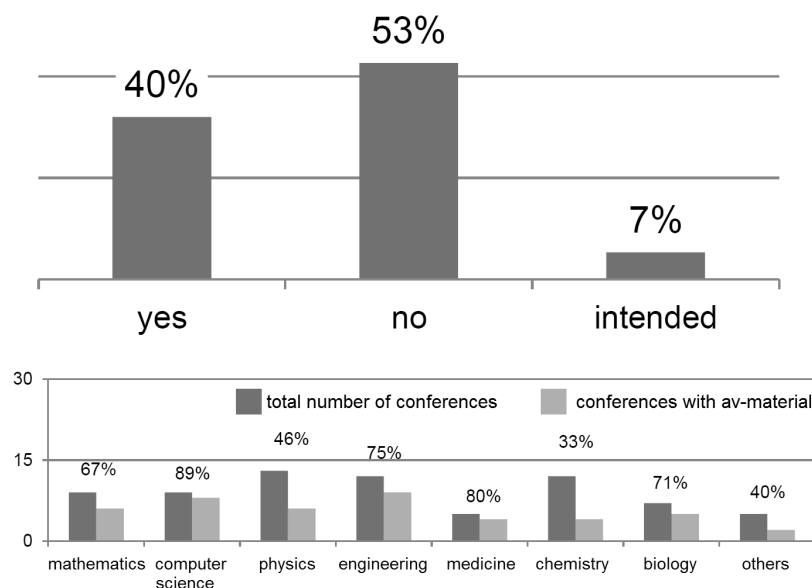


Fig. 2: Almost half (47%) of the respondents stated that they already produced conference recordings or are planning to do so in the future (top). The results differ largely between subject areas, ranging from 33% in chemistry to 89% in computer science (bottom)



According to the survey, opening events, workshops and panel discussions are recorded in addition to classic conference talks (see Fig. 3). The number of talks at scientific conferences varies between 10 and 300. Many organizers reported that they did not have sufficient financial resources regarding personnel and technology for production and post-production of the recordings. For cost reasons they often decide to have only the most important plenary talks recorded in full and to have them edited. 54% of the respondents that produce recordings stated that they also provide livestreams. Further respondents would like to create livestreams as well, but for cost reasons decide against it.

As reported by the respondents, no common standards have yet been established for the production process itself. Standards can be, for example, how many cameras and microphones are typically recorded and how the cameras are placed. Sometimes the camera only records the speaker and sometimes with an auditorium in the picture. In some cases cameras are positioned in a static manner, while in other cases they move along with the speaker. There are also respondents who film slides and presenters with the same camera, while other record the speaker and the presentation signal separately. For this purpose, a video signal (lecturer) and an HDMI signal are combined.

The most common publishing platforms are the corresponding conference website (92%) and YouTube (46%); cf. Fig.4 (top, left). Some of those interviewed said that they tried to develop a business model which provided the videos onto a password-protected member area, offering packages for a certain amount of money per talk. However, this business model was not accepted by the users and therefore they decided against this model.

If a licence is assigned at all, the freely accessible AV recordings use Creative Commons licenses². It is taken for granted by most of the respondents that the speaker's permission to publish the video must be obtained. When it comes to formats, a standard Full HD video format (e. g.. mp4,. flv/. swf,. mov) is used in most cases.

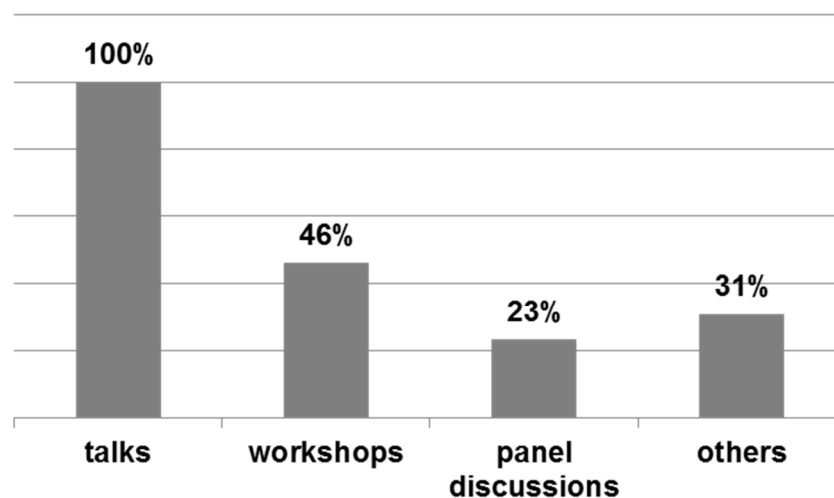


Fig. 3: If conference recordings are produced, conference talks are always among the events recorded. Additionally, opening events, workshops and panel discussions are also recorded in many cases.

None of the respondents said they used persistent identifiers (DOI); cf. Fig.4 (top, right). The addition of Metadata and subtitles is only common practice among conference hosts (80%) and not among conference organisers. Also none of the previously mentioned groups said they published the data as linked-(open)-data in e.g. standard RDF (Resource description framework), so that they are machine readable and can be used by third parties. The interviewed conference recording service provider reported that they carried out post-production to a certain extent such as description of the AV Media in standard RDF (33%) and tagging (50%). Moreover it is implicitly assumed by all groups that publishing the videos on commercial platforms like YouTube or Vimeo means that they are digitally preserved for a long period of time.

² <https://creativecommons.org/>



2.2 Problems and requirements

After the respondents were asked about the current status, we asked the interviewees about problems and requirements. While most respondents (69%) stated that they are satisfied with the current situation, the biggest potential for improvement is seen in speed and efficiency (62%) concerning the production and publication process followed by financial costs (38%). The most important aspects during video production and publication are rapid, cost-efficient and simple procedures that are at the same time sustainable; cf. Fig.4 (bottom, left). After publication visibility, long-term accessibility and quality of the material are the most important concerns; cf. Fig.4 (bottom, right).

It may seem unexpected that 69% of the respondents name sustainability and long-term accessibility among the most important aspects regarding the fact that no persistent identifiers are used and the conference website or YouTube are the main publishing platforms. Publishing conference recordings on commercial platforms like YouTube or Vimeo, where videos are provided with little or no metadata is not sustainable and therefore inadequate for scientific results. It is unclear how long content will be archived on those platforms and how it can be cited consistently. This makes the search for and the re-use of conference recordings difficult and valuable scientific information remains hidden or gets lost [Drees 2016].

Asked why they don't use persistent identifiers, just under half said they hadn't heard of it yet. Among the rest, lack of infrastructure and high costs were mentioned as reasons for non-use. However, all respondents regard the use of DOIs as useful.

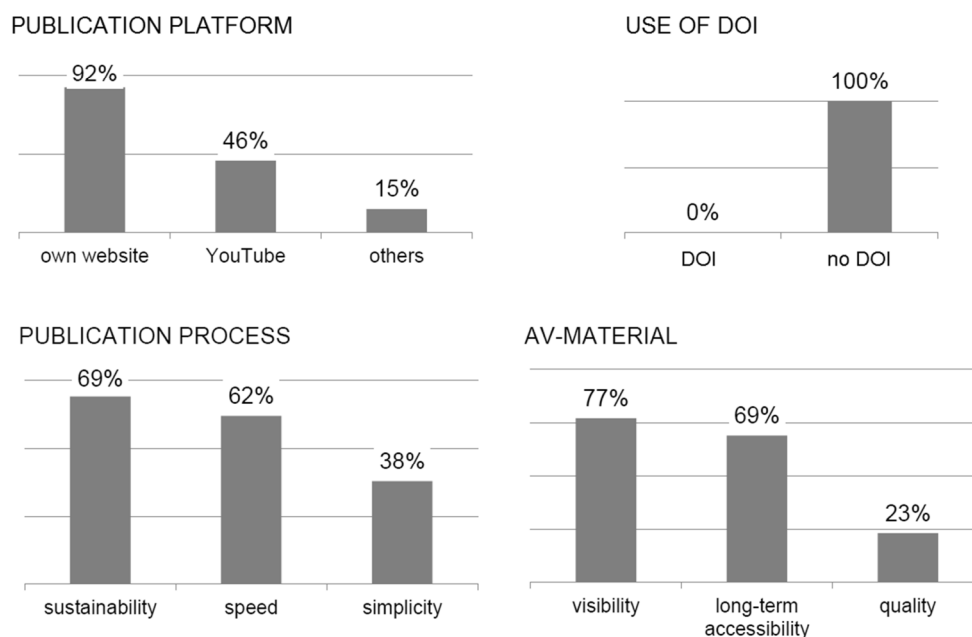


Fig. 4: *Top row:* In most cases conference recordings are published on the conference website or YouTube (left). Accordingly none of the respondents assigns DOIs (or other persistent identifiers) to the videos (right). *Bottom row:* Most important aspects named by respondents regarding the audiovisual material (left) and the publication process (right).

3. Comprehensive Conference Recording Service

Based on the results of the survey, TIB will continue to expand its service portfolio and establish a conference recording service, which includes on-demand recording and publication in the TIB AV portal (av.tib.eu). The TIB AV-Portal [Plank 2016] was developed in 2014 and provides an open access platform for sharing scientific videos, computer visualisations, simulations, experiments and conference recordings under CC licences. The automatic video analysis of the portal includes not only structural analysis (shot boundary detection) but also text, speech, and image recognition and semantic annotation. Automatic indexing describes the videos at the segment level, enabling pinpoint searches in the videos. All videos are allocated a digital object identifier (DOI) so that they can be referenced clearly even by the second (with a media fragment identifier). Moreover, videos in the TIB AV-Portal are digitally preserved.



The TIB AV-Portal provides the ideal infrastructure to host, find and reuse conference recordings. It's a single access point for videos from different conferences and years. Additionally, conference videos are linked to the corresponding proceedings [Drees 2017].

A pilot project of the Conference Recording Service will take place in 2018, after which the service is to be gradually expanded in line with demand. The service packages of the first phase include:

- Basic offer package: Conference video service (on-demand recording and publication in the TIB AV portal) without customization (graphic elements)
- Standard offer package: conference video service (on-demand recording and delivery in the AV portal) with customization
- Complete package: Conference video service (on-demand recording and publication in the AV portal) with two cameras, e. g. for recordings with higher technical and creative demands.
- A live streaming of the contributions will be offered in the second phase

4. Conclusion

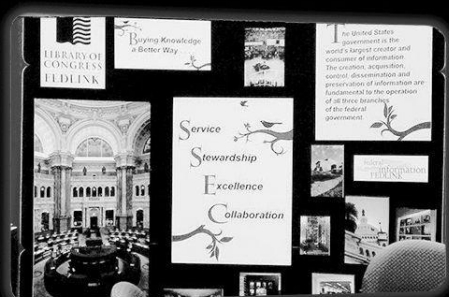
Video recordings of conference talks are becoming more and more common in the scientific communities (although there are huge differences between subjects). While aspects such as long-term accessibility and sustainability are considered very important, persistent identifiers, like DOIs, are scarcely used and videos mainly uploaded on the conference website or YouTube. Libraries should start here and provide reliable, free and open infrastructures for audiovisual media. The TIB AV-Portal (av.tib.eu) is such an infrastructure that guarantees the digital preservation of videos and uses persistent identifiers.

5. References

- [Archambault 2014] Archambault, É., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L. and Roberge, G. (2014), *Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996–2013*. http://science-metrix.com/sites/default/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf; Retrieved on 2017-12-15
- [Drees 2016] Drees, B. (2016). 404: Video not found - The TIB AV-Portal stops scientific videos from disappearing. TIB Blog. <https://blogs.tib.eu/wp/tib/2016/11/16/404-video-not-found-the-tib-av-portal-stops-scientific-videos-from-disappearing/>; Retrieved on 2017-12-15
- [Drees 2017] Drees B. and Plank M. (2017), *TIB AV-Portal: A Trusted Home for Conference Recordings*. Proceedings of the IATUL Conferences 2017, <http://docs.lib.purdue.edu/iatul/2017/analytics/1>; Retrieved on 2017-12-15
- [Einbock 2017] Einbock, J., Dreyer, B., Heller, L., Kraft, A., Niemeyer, S., Plank, M., Schrenk, P., Sens, I., Struß, J. and Tullney, M. and engage AG (2017), *The information procurement and publishing behaviour of researchers in the natural sciences and engineering*; <https://tib.eu/tibsurveyinformationprocurementandpublishingbehaviour>; Retrieved on 2017-12-15
- [GCB 2016] GCB German Convention Bureau e.V. (2016), *Meeting- & Event-Barometer 2016: Trends-Modern technologies are in demand in the meetings industry*. https://www.gcb.de/fileadmin/GCB/News/Press_Releases/Downloads/Meeting_EventBaromter2016_presskit.pdf; Retrieved on 2017-12-15
- [Köhler 2017] Köhler, M. (2017), *Open Access in den MINT Fächern*. In: Söllner, K. and Mittermaier, B., *Praxishandbuch Open Access*, Berlin: De Gruyter Saur, 274-280 <https://doi.org/10.3204/PUBDB-2017-04494>
- [Piowar 2017] Piowar H., Priem J., Larivière V., Alperin J.P., Matthias L., Norlander B., Farley A., West J., Haustein S. (2017), *The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles*. PeerJ Preprints 5:e3119v1 <https://doi.org/10.7287/peerj.preprints.3119v1>
- [Plank 2016] Plank M. (2016), *Reliable infrastructure for scientific videos*. Proceedings of the Conference on Grey Literature and Repositories 9:16-23, http://repozitar.techlib.cz/record/1033/files/idr-1033_4.pdf; Retrieved on 2017-12-15

FUNDING

The study was financed within the “Professionalisation and continuation of the concept for exploitation of research results at the German National Library of Science and Technology, Hannover (TIB)” project (funding ID:03IO1441) funded by the Federal Ministry of Education and Research (BMBF). Within this project, TIB’s knowledge and technology transfer concept is further developed, adapted to new structures, and extended to include new topic areas. TIB seeks to increase the number and quality of utilisation projects, to better market its offerings, and to implement new, innovative services.



Strategic Sourcing



Currently, more than 20 federal agencies, both military and civilian – including the Department of Defense – participate in the Federal Strategic Sourcing Initiative (FSSI). FSSI was created in 2005 by the Department 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 to streamline efficiency and reduce costs to the federal government.

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.



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



Apps & Codes: Making profiles for fluid publishing contents

Flavia Cancedda, CNR, National Research Council - Central Library, ISSN National Centre
Luisa De Biagi, CNR, National Research Council - Central Library, OpenGREY Network, Italy

Abstract

Apps are one of the most used digital publishing tools and convey intellectual contents for innumerable functionalities. App programs present technical characteristics different from each other, depending on the intellectual content for which they are profiled and made available. In publishing field, apps that convey traditional editorial products disclose the maximum of their innovative potential in representing the interface of continuously updating products (such as newsletters, magazines, newspapers, guides and tourist maps, blogs, open/e-gov/research data systems and related data-bases, clinical trials, forums and websites with specific matter or content, etc.). If apps are the interface software – thus, the communication interface - of publishing products, could it be possible to make them identifiable (and manageable) through the same bibliographic codes used for the corresponding traditional publishing products (e.g., ISBN, ISSN, DOI, etc.)? We think that when apps show and preserve the essential bibliographic identifying data, the answer could be no less than positive.

The goal of this study is to show, through a comparative analysis of National and European case-studies and best practices examples, how the potential public usefulness (not only commercial but also informative) of a bibliographic identification for apps is similar to the usefulness recognized in case of traditional publishing products, opening new scenarios and results also for grey products: a new traceability skill could be established for apps containing permanent references, traditionally considered necessary for the identification of editorial objects (title, publisher, year, updating mode or frequency, etc.); traceability through numerical codes would allow also easier dissemination, marketing, indexing processes by search engines, portals and sales store, up to make indexing tools for bibliographic services and librarians more specific and professional.

Apps&codes

Apps are one of the most used digital publishing tools... and convey intellectual contents for innumerable functionalities. We would like to start from this statement – basic and almost obvious - to introduce a subject that is increasingly focusing in publishing environment. Preventively, it's better to emphasize that we are here conceptually moving within publishing and media area or environment: we do not mean to refer to apps outside of this context or to apps providing non-editorial services (e.g.: games, advertising or promotional tools, information on transport timetables, personal physical functions monitoring, weather forecast etc.), neither to apps whose function is only to facilitate the connections of mobile devices to traditional websites. Moving within the publishing industry, we hereby present the knowledge and therefore the practical experience of the ISSN Italian Center (member of the International ISSN Network¹): for its function, it occupies a privileged point of observation on the panorama of publishing innovations.

ISSN National Centers deal with registering and identifying serial publications (magazines, newspapers, monographic series, websites or updating databases, and so on...). Centers operate in a field of activity undoubtedly relevant to librarianship, also strongly focused on the *techniques of bibliographic identification* (more than on bibliographic description, subject indexing or other specific library activities). ISSN Centers are also particularly involved in the *application policies* of those techniques, keeping constantly in touch with the changing international phenomena of publishing. For many years, there were questions in identifying digital (online) editorial/intellectual products with features different from those typical of a magazine, a book, a monographic series published or marketed online. A traditional online magazine, or book, or monographic series keeps several editorial indicators – a kind of signposting, the paratextual areas - borrowed from paper publishing: a homepage acts as a cover even when there is not digital cover; one or more webpages dedicated to presentation, contacts, acting as a colophon; another web page dedicated to the index, etc. It is not particularly difficult to assimilate cataloging and identification of these products to that used for paper products: the physical support difference makes it necessary to add or modify some descriptive details, but does not

¹ The International Standard Serial Number Network is composed by 89 national Centers, and coordinated by the International Center in Paris, www.issn.org.



affect the possibility of circumscribing, and thus identifying these digital products, for all the purposes for which an identification is useful.

But now we are not talking about an online magazine, which can also be searched by smartphones or about an e-book, which can also be read in a tablet. Instead, we would like to point out that the apps we frequently download on our devices are something more and different from a "traditional" digital editorial product, even when they aim to convey a publishing product that has a pretty traditional source. Apps are born as programs, not as "contents"; they arise as a tool or vector, and not as "editorial form". In short, they allow you to use intellectual content on the latest generation of devices: the programs that allow this type of use (via tablet, e-book reader, iphone, smartphone...), with technological and usage parameters defined by the type of device involved, are synthetically called "apps". There are, in fact, apps with a traditional interface, which allow them to be optimally visualized on those devices but do not significantly affect the editorial product they deliver: many e-books apps have these features. But there are also "less-traditional" apps (those we are currently reflecting on) for *app-native editorial products* where the intellectual or informational content has been profiled - we could say: editorially imprinted, constructed - to exploit the technological potential of the app. Our work experience leads us to meet more and more traditional editorial products that are substantially re-designed to be used through apps and to exploit the undoubted technological advantages of apps: e.g. newspapers or weekly magazines, scientific databases, thesauri, interactive manuals, guides and tourist maps, personal directories or address books, commercial catalogs for sales, newsletters focused about specific events (as fairs, meetings, exhibitions...), clinical case-studies directories, multiplatform related to info services etc. Many of these types - and others - start from a traditional editorial idea, but their content is assembled and reassembled through the cooperation of different software components for managing interactions between the publisher, other related content to which the publisher decides to allow access, and the operations of interactive use by the end-user.

Especially regarding the continuous updating and the connectivity with other related digital products or environments, the technology under the app is so incisive on the intellectual content that it is legitimate to ask: are these "editorial" products actually so? and, if so, are they still identifiable as such (i.e., by same tools and methodologies applied to more traditional products)?

If the answer were "no" we should admit that we are facing software: that is, we are facing technological tools, probably intended to change in a short time, as well as all many other technologies: software would not fall into our field of publishing experts; it would not be useful to discuss whether and how to identify software using bibliographic methods.

However, we believe that the answer may be "yes": publishing products transmitted through apps are editorial products, and it is possible - and useful - to describe/identify them using bibliographic methods and tools. From our point of view, we can use bibliographic identification codes (ISBN, ISSN, DOI etc.) to uniquely identify these products, provided these products present the minimum bibliographic characteristics through which they can be identified.

According to the traditional cataloging praxis and bibliographic descriptive literature, the minimum characteristics to identify with sufficient certainty an editorial product are: the product has a title (words, phrases, characters or character groups on the product, intent on identifying it; the title has to be detectable and constant even in following updates or serial outputs of the same product); the product shows a copyright or an editorial responsibility; the product has one or more dates of publication/release/update dates... and so on. If these bibliographic indicators can easily be found in the editorial product delivered via app - e.g. an interactive tour guide that helps us move around in a city, find places and services for us, memorize our preferences, connect us to other information sources or allow chat and messaging - we are faced with an editorial product recognizable and bibliographically identifiable through traditional tools as international identification codes: an interactive tourist guide can be identified through ISBN and also through DOI; and through ISSN too, if the guide explicitly declares its future updates (i.e. if it assumes the characteristics of a serial publication).

We have just mentioned three international reference systems for bibliographic identification, ISBN, ISSN, and DOI: the ISO standards underlying these systems indicate their very extensive application without restrictions concerning digital environments. For example, the ISBN Manual, par. 6.1 / 6.2 explicitly states: "*Where a publication is available electronically (e.g., an e-book, e-book app, CD-ROM or publication available on the Internet), it will qualify for an ISBN provided*



that it contains text and is available to the public ... An ISBN may be used to identify a specific software product that is intended for educational and / or instructional purposes, such as a computer-based training product, provided that it is neither customizable nor requires data in order to function"². The ISO 3297 standard which regulates the ISSN system provides a very large scenario: in fact, ISSN code can be assigned to continuous (=ongoing) integrating resources: "Continuing resource that is added to or changed by means of updates that do not remain discrete and are integrated into the whole"³.

More specifically and explicitly in the DOI Foundation's *Factsheets*⁴, several definitions and statements confirm that DOI can be easily applied to editorial apps as well:

"DOI® is applicable to any object (= any entity or thing: physical, digital, or abstract; resources, parties, licenses, etc.);

- ▶ *is a digital Identifier of an object = network actionable identifier ("click on it and do something")*
- ▶ *initial focus on entities was documents/media e.g., articles, data sets; now moving into parties, licenses and other sectors*
DOI® provides an actionable, interoperable, persistent link
- ▶ *Actionable – through use of identifier syntax and network resolution mechanism (Handle System®)*
- ▶ *Persistent – through combination of supporting improved handle infrastructure (registry, proxy, etc.) and social infrastructure (obligations by Registration Agencies)*
- ▶ *Interoperable – through use of a data model providing semantic interoperability and grouping mechanisms".*

After this quick overview on international bibliographic identification codes/tools - and therefore on the feasibility of international identification of publishing apps – two questions are necessary, which will affect the future of bibliographic treatment activities by international identification Agencies:

Question 1): why an app should be bibliographically identified (i.e.: why to assign an ID code such as ISBN or ISSN)?

Question 2): ... and, once answered to the previous question, is it practically advantageous to identify an app by a numeric code?

Let us briefly answer the questions.

- 1) The bibliographic identification of an app - that has been explicitly requested by some publishers - means allowing the publisher/distributor/provider to recognize their product unequivocally for all bibliographic and commercial transactions, just like for a traditional editorial product.

Of course, the unique identification can also be used as a prerequisite for attesting the validity and/or the level of editorial quality or accuracy (quite similar as in the field of scientific publishing when an editorial product to be considered for public funding should be identified by one of the existing international bibliographic codes).

- 2) But why should a publisher, practically, want to identify an app (in the absence of legal obligations or common and shared media practices that make it necessary or at least very appropriate)? Why, for example, did a publisher ask us to identify with ISSN code a travel guide that could be downloaded on smartphone, continuously updated and interactively managed through apps? What would be a bibliographic identification code for, since an interactive app would improbably have registered into a library catalog?

The reasons driving a publisher are connected to commercial and digital distribution procedures: therefore, an international identification Agency would not identify an app for library purposes or others purposes related to dissemination at cultural institutions such as libraries or documentation centers, but because the unique identifier code given to the app would be used as fil-rouge (or maybe Ariadne's thread) for each business publishing procedure: connecting first

² *ISBN User's Manual*, London, International ISBN Agency, 2012 (International ed.; sixth ed.), p. 13 <<https://www.isbn-international.org/sites/default/files/ISBN%20Manual%202012%20-corr.pdf>>

³ *ISO 3297:2017, Information and documentation – International standard serial number*, Geneva, International Organization for Standardization, 2017, par. 3.3 (partially available at <https://www.iso.org/obp/ui/#iso:std:iso:3297:ed-5:v1:en>).

⁴ <https://www.doi.org/factsheets/DOIKeyFacts.html>



release and following releases of the product/app from the IT company to the publisher; delivering the product by publisher to providers and different sales platforms; tracking sale or release of the app and its updates for each buyer or user, and for each connected device; facilitating credits access management and related interactive services; recognizing different service levels according to purchasing/subscription levels; managing subscription procedures, subsequent releases, "patches"; accounting the number of downloading; identifying the commercial product/app for tax purposes, for distribution of revenues in connection with the different industrial or intellectual property rights quotes; etc.

Consequently, considering the app as commercial object, the software and the conveyed intellectual content become much easier to manage by the publisher (which is not an IT company, and does not normally have all technological skills) if he use a single identifier code - better one of the bibliographic codes already known - that accompanies the product through all transactions.

The publisher, as well as most of his customers, keeps a "conservative side" in his mind: he certainly appreciates he can act on technologically innovative elusive products with usual and well-known tested tools (the identifier codes). He appreciates much more if that procedure allows him to incardinate without trauma his innovative app within his usual editorial catalog: there, he will incorporate new apps alongside traditional products, transmitting to customers the idea that the purchase or the enjoyment of both will not lead to tedious changes of habit or too awkward access procedures.

So, for the publisher it's very convenient - both for practical and promotional reasons - that apps and books, apps and magazine, apps and articles give the idea they are commercially the same things, so they are managed by users with the same procedures. The use of identified, approved, and standardized bibliographic codes - such as ISBN, ISSN, DOI, and so on - presents for the media several benefits: the codes are already well-known and recognizable (we must remember that distinguishing the type of code means recognizing the type of publishing product); they are internationally reliable and guaranteed as they are maintained by institutionally solid Agencies; the codes have transparent, consolidated and quite easy acquisition procedures; have a fairly low cost (or even no cost at all); they authenticate the identified product, and "welcoming" it in their sphere of competence they sanction identification, nature and, in some respects, cultural profile as well.

For all these reasons, we believe that traditional identifying systems - far from being an obstacle - can be useful for authors and publishers who want to entrust intellectual and informational content to apps, ensuring it is capillary distributed: technologically advanced and more liquid apps, increasingly popular for their plain use and eye-catching graphic presentation - but even more for easy and immediate recognizability.

Software codes, releases, software repositories and DOI interactive use: applications and trends

ISBN and apps

Regarding ISBN, the Italian national Centre provided by AIE (Italian Publisher Association)⁵, reported us that it's impossible to know to which product and typology the code requested will be assigned. The ISBN user/applicant gets a block (lot) of ISBN codes to be applied in publications, usually books or e-books.

For ISBN Italian Centre, surely app can't be identified with ISBN standard... but it's equally possible to identify an educational and didactic software with its user manual and technical instructions if the manual is fundamental for making the software work and its usefulness is strictly conditioned to the match with the software itself.

A new development has been done with OCR applied to ISBN.

ISBN Scan is an ISBN (International Standard Book Number) reader application available on Google Play. ISBN "RealCodeScan" engine can be easily integrated into various smart-phone applications, including iOS systems. This app can read not only the ISBN bar-code but also the ISBN number itself by the OCR (Optical Character Recognition) powered functions.

There is no even need to press the shutter when reading because ISBN Scan reads snap shot images reflected in the camera just by simply waving the device. It's capable to read the ISBN at

⁵ <http://www.aie.it/>



high accuracy with super-fast speed, “modifying the camera configuration (in-app) for BEST Reading performance”.

One of the main benefits of this app is also the possibility of getting and viewing immediately the detailed information sheet of the book (author, title, publisher, etc.), as well as share them via SNS (Social Network Services) or e-mail.

A similar app is ‘My Library’: a simple application helping users to manage a sort of private library at books.google.com. ‘My Library’ is a totally free app (with no advertising) to keep track of all personal books and wishlist books.

In ‘My library’ it’s possible to add a book searching its ISBN code or even scanning its barcode. In those cases informations (author, title, publisher, publication date etc.) are given from Amazon, Google Books and Open Library.

My Library’s main features are⁶:

- possibility of using phone’s camera to scan a barcode
- adding /removing books from books.google.com shelves
- downloading pdf of books in public domain
- creating a personal shelf with personalized recommendations/alerts
- support tablet devices.

Though, there are some criticalities reported by users of this app, starting from the impossibility of cataloguing books by matter or a non-complete recognition of all ISBN codes. As ISBN scan ‘My library’ – runs both on Android (Google play) and iOS (Apple I-Tunes)⁷.

Finally **The US DOI Agency is experimenting a new** qr-encoded Crossref DOI, inspired by Google recent promotion of QR codes: it’s possible to generate a QR Code for any given Crossref DOI (even media gadgets)⁸

Making code citable: Zenodo and GitHub

For Open Science it is important to cite the software used in research studies. It ought to be cited any software making a significant impact. Modern research relies constantly on data analysis and the main mission might be to preserve and cite software in a permanent/sustainable, identifiable and simple way. Innovative Digital repositories like Zenodo can really help us reaching this goal.

DOIs are persistent identifiers obtained only by an agency committed to maintain a reliable level of consistency and preservation of a digital resource. So, as a digital repository, Zenodo registers DOIs through DataCite and preserves these submissions using trusted foundation of CERN’s data centre, alongside the LHC’s 100PB Big Data store (biggest scientific dataset worldwide). The code preserved in Zenodo will be accessible in the future to supply long-term digital storage and preservation. The DOIs will function as perpetual links to the resources. DOI based citations remain even if URL and protocol change, because DOI currently direct to URIs (Uniform Resource Identifier). DOIs has also search engines and indexing services, to support software usage through multiple citations.

Long-term digital storage and preservation. Although it is possible to identify softwares uploaded to platforms like GitHub⁹, these platforms do not issue DOIs and there is not a perpetual guarantee of access to older software: we might have to face with preservation and versioning issues problems (e.g. about which version of the code is being referenced and if is it still available on the hosting website).

That’s why it’s important submitting the software/code in a digital repository like Zenodo. Born within the OpenAIRE project Zenodo is a free, open-access research repository¹⁰ which profits from and contributes to provide CERN important initiatives as Open Data services (CERN opendata portal)¹¹.

Launched at the CERN Data Centre in May 2013 with a grant from the European Commission, Zenodo has a special commitment to sharing, citing and preserving data and code.

⁶This app is only a book tracking append it’s not a book reader app

⁷<https://itunes.apple.com/it/app/mylibrary/id1141437096?mt=8>

⁸<https://www.crossref.org/display-guidelines/>

⁹<https://github.com/github>

¹⁰Based on the Invenio open-source software

¹¹<http://opendata.cern.ch/>



Submitting the code to Zenodo and receiving a DOI has become easier since Zenodo's integration to the platform GitHub has been done. Moreover, if preservation is based on releases, as the software changes each release can be cited with its own DOI, giving precise traceability and awareness of the exact code used in a data analysis. Code releases are both archived and 'published' to become public, so that they can be described with rich metadata, an explanatory abstract and a significant author list. This means that software developers could theoretically skip the journal publication step to announce and make their creation immediately visible.

Zenodo allows also to sign up only one time by choosing the GitHub or the Zenodo account, in order to advance the immediate interlinking. Finally, if a research is funded by an EU grant, it's even possible to directly connect the code to the grant by updating the grant section of the metadata on the Zenodo record.

References

- ▶ ISO 2108:2005, International standard book number (ISBN)
- ▶ ISO 26324:2012, Digital object identifier system [DOI]
- ▶ ISO 17316:2015, International standard link identifier (ISLI)
- ▶ ISO 3297:2017, International standard serial number (ISSN)

Past Internet standards:

- ▶ RFC 1630 (1994), <https://tools.ietf.org/html/rfc1630>
- ▶ RFC 2141 (1997), <https://tools.ietf.org/html/rfc2141>
- ▶ RFC 3986 (2005), <https://tools.ietf.org/html/rfc3986>
- ▶ <https://ec.europa.eu/research/openscience/index.cfm>



List of Participating Organizations

Access Innovations, Inc.	United States
ACM Digital Library	United States
Alberta Health Services	Canada
Biblioteca Centrale, G. Marconi; CNR	Italy
Centre National de Recherche Scientifique, CNRS	France
Data Archiving and Networked Services, DANS-KNAW	Netherlands
Deanet Media Company	Italy
EBSCO	United States
EiPass	Italy
Federal Library Information Network, FEDLINK	United States
Food and Agriculture Organization of the United Nations, FAO	Italy
Georgetown University	United States
GERiiCO laboratory	France
German National Library of Science and Technology, TIB	Germany
Grey Literature Network Service, GreyNet International	Netherlands
Institute for Applied Mathematics and Information Technologies, IMATI-CNR	Italy
Institut de l'Information Scientifique et Technique, Inist-CNRS	France
Institute of Clinical Physiology, IFC-CNR	Italy
Institute of Computational Linguistics, ILC-CNR	Italy
Institute of Informatics and Telematics, National Research Council	Italy
Institute of Information Science and Technologies, ISTI-CNR	Italy
Irvine Valley College	United States
Istituto Motori; IM-CNR	Italy
Korea Institute of Science & Technology Information, KISTI	Korea
Kumoh National Institute of Technology	Korea
Library of Congress, LoC	United States
National Civil Service Project, UNSC	Italy
National Library of Technology, NTK	Czech Republic
National Research Council of Italy, CNR Central Library	Italy
New York Academy of Medicine	United States
New York University; School of Medicine	United States
Nuclear Information Section; International Atomic Energy Agency, NIS-IAEA	Austria
Pratt Institute, School of Information	United States
PricewaterhouseCoopers, PwC	Netherlands
PUSAN National University	Korea
Saint Petersburg State University	Russia
Slovak Centre of Scientific and Technical Information, CVTISR	Slovakia
Springer Nature	Italy
TextRelease, Program and Conference Bureau	Netherlands
University of Arizona, College of Medicine Phoenix	United States
University of British Columbia, UBC	Canada
University of California, Irvine Libraries, UCI	United States
University of Florida; George A. Smathers Libraries	United States
University of Ghent, Heymans Institute of Pharmacology	Belgium
University of Liège, Department of General Practice	Belgium
University of Lille 3	France
University of Maryland; University Libraries	United States
University of Rouen, Department of Information and Medical Informatics	France
University of Texas Health Science Center at Houston, School of Biomedical Informatics	United States
University of Toronto, Institute of Health Policy, Management and Evaluation	Canada
University of Wisconsin, Milwaukee	United States
Wikimedia	Czech Republic
Wiley	Italy
WorldWide ScienceAlliance	United States

Twentieth International Conference on Grey Literature Research Data Fuels and Sustains Grey Literature



Loyola University, New Orleans, Louisiana, USA

December 3-4, 2018



Conference Announcement

The definition of research data is as encompassing as the field of grey literature. What should be included and what should be excluded is and remains an issue of concern. Research data can be defined as factual materials collected by diverse communities of practice required to validate findings. While the majority of research data is created in digital format, research data in other formats cannot be excluded. The formats in which research data appear are multiple and the types of research data are diverse. This also holds for the numerous document types in which grey literature appear published.

Today, while emphasis is placed on big data, the fact that the majority of research projects are small to medium size is overlooked. This is but another characteristic that holds true for grey literature. Nonetheless, one should be aware that research publications are not research data, for they are often managed separately from one another. Just as there are a number of stakeholders involved in the production, access, and preservation of grey publications, so too are there stakeholders tasked with the creation and management of research data. Libraries and data management librarians have the responsibility for the curation of the data they collect and preserve. And, it is important to stress the need to maintain appropriate metadata related to research data in order to facilitate their interpretation and further reuse.

Over the past quarter century, grey literature communities have worked diligently to demonstrate how their documents are produced, published, reviewed, indexed, accessed, and further used, applied, and preserved. Today, these communities are now challenged to demonstrate how research data fuels and sustains their grey literature. These communities of dedicated researchers and authors maintain a strong conviction in the uses and applications of grey literature for science and society. Through the years, they have proved willing to share the results of scholarly work well beyond their own institutions. Hence, one can assume they are aware that innovation forfeits with the loss of data as with the loss of information. This 20th International Conference in the GL-Series seeks to address key issues and topics related to grey literature and its underlying research data.

Conference Topics

- Long Tail Research Data and Grey Literature
- Metadata on the Frontline of Research
- Data Management and the Role of Librarians
- Current Research Trends in Grey Literature
- Research Data and Open Access Compliance
- Data Policies Reflect the Political Will of Organizations

Dateline 2018

● April 15	● May 1	● May 8	● May 15	● Oct 15	● Oct. 30	● Nov 15	● Dec. 3-4
Close, Call for Papers	Program Committee Meeting	Authors Notified	Open, Call for Posters	Close, Early Conference Registration	Close, Call for Posters	Submission Conference Papers	GL20 Conference in New Orleans

TextRelease

GL20 Program and Conference Bureau

Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands

www.textrelease.com • conference@textrelease.com

Tel. +31-20-331.2420

Twentieth International Conference on Grey Literature Research Data Fuels and Sustains Grey Literature



Loyola University, New Orleans, Louisiana, USA

December 3-4, 2018



Call for Papers

Title of Paper:

Conference Topic(s):

Author Name(s):

Phone:

Organization(s):

Email:

Postal Address:

URL:

Postal/Zip Code – City – Country:

Guidelines for Abstracts

Participants who seek to present a paper dealing with grey literature are invited to submit an English language abstract between 350-400 words. The abstract should address the problem/goal, the research method/procedure, an indication of costs related to the project, as well as the anticipated results of the research. The abstract should likewise include the title of the proposed paper, conference topic(s) most suited to the paper, name(s) of the author(s), and full address information. Abstracts are the only tangible source that allows the Program Committee to guarantee the content and balance in the conference program. Every effort should be made to reflect the content of your work in the abstract submitted. Abstracts not in compliance with the guidelines will be returned to the author for revision.

Related Conference Topics

☐ Long Tail Research Data and Grey Literature

☐ Metadata on the Frontline of Research

☐ Data Management and the Role of Librarians

☐ Current Research Trends in Grey Literature

☐ Research Data and Open Access Compliance

☐ Data Policies Reflect the Political Will of Organizations

☐ Other related topic:

Due Date and Format for Submission

Abstracts in MS Word must be emailed to conference@textrelease.com on or before **April 15, 2018**. The author will receive verification upon its receipt. By early May, shortly after the Program Committee meets, authors will be notified of their place on the conference program. This notice will be accompanied by further guidelines for submission of full text papers, biographical information, accompanying research data, PowerPoint slides, and required Author Registration.

TextRelease

GL20 Program and Conference Bureau

Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands

www.textrelease.com • conference@textrelease.com

Tel. +31-20-331.2420

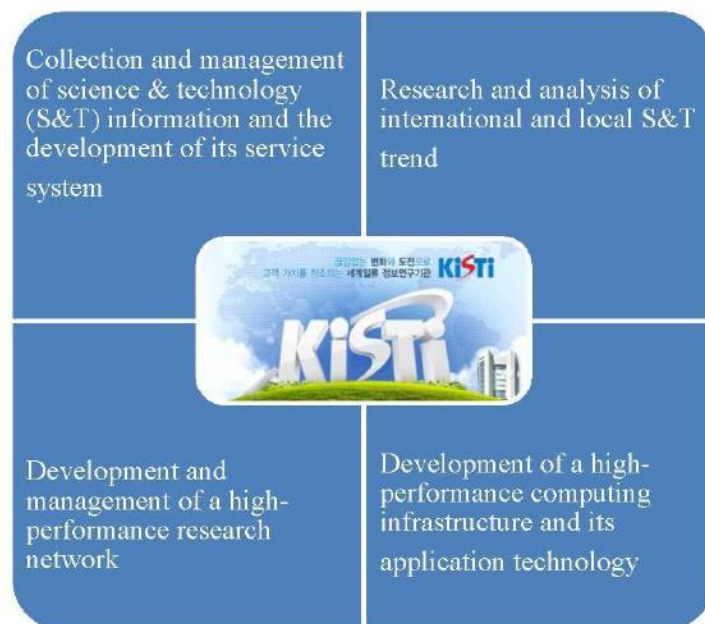
Korea Institute of Science and Technology Information (KISTI)

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



Author Information

Artini, Michele**107**

Michele Artini is a research fellow at the Networked Multimedia Information Systems laboratory of the "Istituto di Scienza e Tecnologie dell'Informazione", Consiglio Nazionale delle Ricerche, Pisa, Italy. Since 2005 he is involved in EC funded projects for the realisation of aggregative data infrastructures like DRIVER, DRIVER II, BELIEF, HOPE, EFG, EFG1914, OpenAIRE, OpenAIREPlus and Openaire2020. He is interested in digital libraries, service-oriented infrastructures, database systems and workflow management systems. Email: michele.artini@isti.cnr.it

Atzori, Claudio**107**

Claudio Atzori received his MSc in "Information Technology" in 2009 at the University of Cagliari and completed his PhD at the Information Engineering at the Engineering School "Leonardo da Vinci" of the University of Pisa in 2016. He works as research fellow in the InfraScience research group, part of the Multimedia Networked Information System Laboratory (NeMIS), at the "Istituto di Scienza e Tecnologie dell'Informazione" (ISTI), National Research Council (CNR), in Pisa, Italy. He works on the realisation of aggregative data infrastructures for the Open Science and Scholarly Communication. He has also participated to the EC funded R&D projects: DRIVER-II, EFG, EFG1914, HOPE, EAGLE, OpenAIRE, OpenAIRE-Plus, OpenAIRE2020, OpenAIRE-Connect. Email: claudio.atzori@isti.cnr.it

Baglioni, Miriam**107**

Miriam Baglioni works as a research fellow at Networked Multimedia Information Systems (NeMIS) Laboratory of the Italian National Research Council - Institute of Information Science and Technologies (CNR-ISTI). She graduated in Computer Science with honors, in 2001, from the University of Pisa and received her PhD in Computer Science, in July 2005 from the University of Pisa. She was involved in the EU funded projects BRITE, MUSING, and GeoPKDD, and in the projects ClickWorld, Dantex, TetraModel and BiNet. She is currently collaborating at the EU funded projects EFG1914, OpenAIRE2020 and OpenAire-connect. She has worked on Data Mining, Knowledge Discovery, ontologies, social networks and bioinformatics. Her current research interests include data e-infrastructure for science, and science reproducibility. Email: miriam.baglioni@isti.cnr.it

Balashova, Yuliya B.**33**

Professor Yuliya B. Balashova, is Doctor of Philology, journalist, Associate Professor at St Petersburg State University, Russia. Email: j.balashova@spbu.ru

Bardi, Alessia**107**

Alessia Bardi is a researcher at Networked Multimedia Information Systems (NeMIS) Laboratory of the Italian National Research Council - Institute of Information Science and Technologies (CNR-ISTI). She graduated in Computer Science in 2009 at University of Pisa and completed a PhD in Information Engineering at the Engineering Ph.D. School "Leonardo da Vinci" of the University of Pisa in 2016. She is involved in EC funded projects for the realisation aggregative data infrastructures. Her research interests include service-oriented architectures and data infrastructures for e-science and scholarly communication. Email: alessia.bardi@isti.cnr.it

Bartolini, Roberto**93**

Roberto Bartolini - Expertise on design and development of compilers of finite state grammars for functional analysis (macro-textual and syntactic) of Italian texts. Expertise on design and implementation of compilers of finite state grammars for analysis of natural language texts producing not recursive syntactic constituents (chunking) with specialization for Italian and English languages. Skills on acquiring and extracting domain terminology from unstructured text. Skills on semi-automatic acquisition of ontologies from texts to support advanced document management for the dynamic creation of ontologies starting from the linguistic analysis of documents. Email: roberto.bartolini@ilc.cnr.it

Biagioni, Stefania**11**

Stefania Biagioni graduated in Italian Language and Literature at the University of Pisa and specialized in Data Processing and DBMS. She is currently a member of the research staff at the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), an institute of the Italian National Research Council (CNR) located in Pisa. She is head librarian of the Multidisciplinary Library of the CNR Campus in Pisa and member of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She has been the responsible of ERCIM Technical Reference Digital Library (ETRD) and currently of the PUMA (Publication Management) & MetaPub, a service oriented and user focused infrastructure for institutional and thematic Open Access repositories looking at the DRIVER/OpenAire vision, <http://puma.isti.cnr.it>. She has coauthored a number of publications dealing with digital libraries. Her activities include integration of grey literature into library collections and web access to the library's digital resources, including electronic journals and databases. She is a member of GreyNet since 2005. Since 2013 she is involved on the GreyGuide Project. Email: stefania.biagioni@isti.cnr.it

Cancedda, Flavia**133**

Flavia Cancedda, Technologist, librarian and information specialist. Head of the Italian ISSN Centre (Central Library CNR); Component of "Documentation and Information" Committee of the UNI, National standardization organism (Chairman of the Subcommittee "Presentation, identification and description of documents"; component of the Working Group on National Standard about Professional Requirements for Librarians); Component of the ACNP Libraries Committee (Archivio Collettivo Nazionale Periodici) and of the ACNP Cataloguing Rules Revision Working Group. Since 1991 to 2000 librarian in two different Italian universities and in the National Library of Florence (section: National Bibliography); since 2001 librarian in the Central Library of the Italian National Research Council (CNR). Master's degree in Libraries Management and Direction. Post-graduate degree as Manuscripts conservator at the Special School for Archivists and Librarians (La Sapienza University in Rome). Author of books and essays about the history of libraries and historical bibliography. Email: flavia.cancedda@cnr.it

**Černohlávková, Petra****101**

Petra Černohlávková studied Information science and librarianship at the Charles University in Prague. She is currently working in the National Library of Technology (NTK) in Prague in Digital National Library of Technology Department. She is a content coordinator of the National Repository of Grey Literature and of the Institutional Repository of NTK. Email: petra.cernohlavkova@techlib.cz

Chamberlain Kritikos, Katie**37**

Katie Chamberlain Kritikos is a PhD student at the University of Wisconsin-Milwaukee School of Information Studies. She graduated from the University of Illinois at Urbana-Champaign with a JD (2009) and MLIS (2010) and received her BA, summa cum laude, in English (2006) from the University of Alabama. Katie researches law and information policy with interests in free speech, privacy, information policy, and scholarly communication. She is a member of the Board of Trustees of the Freedom to Read Foundation. Email: kritikos@uwm.edu

De Biagi, Luisa**133**

Luisa De Biagi got her Laurea Degree in Literature and Philosophy at 'La Sapienza' Univ. of Rome (Art history and Cultural heritage). With a Specialization in 'Archivist-Palaeographer' (Vatican School of Palaeography, Diplomatics and Archivistics at the Vatican Secret Archive) as well as a Specialization Degree in Archivistics, Palaeography and Diplomatics (Archivio di Stato, Rome) and a Degree from the Vatican School of Library Sciences. De Biagi further holds a Master in 'Business Publishing' (LUISS Management – Rome). She's been working for the SIGLE Network (System for Information on Grey Literature in Europe) since 2002. Since 2010 she's responsible for the Italian National Referring Centre of Grey Literature at CNR Central Library 'G. Marconi' as representative of the European Network and Openarchive OpenGrey (System for Information on Grey Literature in Europe). She's taken part in 3 editions of the Annual International Congress on Grey Literature – GL (GL5, Amsterdam, GL13, Washington D.C. GL14, Rome and GL15 at Bratislava). She's also a member of the CNR Working Group for Cedefop-Refernet Project (Consortium for Professional Education and Training coordinated by ISFOL), the Committee for Legal Deposit Acquisition at CNR Central Library, and a member of the European Association of Health Information and Libraries (EAHIL). She's also responsible for the Library Functional Units 'Education and Training' and 'Cultural Activities Management', organizing didactics laboratories for students, professional training courses and teaching in professional trainings for librarians, students and users. Email: luisa.debiagi@cnr.it

De Bonis, Michele**107**

Michele De Bonis is a research fellow at Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. He received his MSc in Computer Engineering in the year 2017 at the University of Pisa, Italy. He is currently researching in the areas of clustering methods. He is currently working for the developing of digital libraries and deduplication of the authors for the OpenAIRE project. Email: michele.debonis@isti.cnr.it

Deluca, Rosaria**11**

Rosaria Deluca graduated in Law at the University of Pisa in 1995. Since 2003 she has been working in Pisa at the Institute for the Science and Technologies of Information "A. Faedo" of the Italian National Council of Research (ISTI-CNR). Currently she works at the Library and Scientific Documentation Center of the CNR in Pisa. She is interested in Library Management, Open Access repositories, EU and national law on data protection. Email: rosaria.deluca@isti.cnr.it

Drees, Bastian**127**

Bastian Drees works in the Competence Centre for non-textual Materials in the Research and Development Department of the German National Library of Science and Technology (TIB). He holds a PhD in physics and did his traineeship as a librarian at TIB and the Bavarian State Library (BSB) from 2014-2016. Email: Bastian.Drees@tib.eu

Farace Dominic**27**

Dominic Farace is Head of GreyNet International and Director of TextRelease, an independent information bureau specializing in grey literature and networked information. He holds degrees in sociology from Creighton University (BA) and the University of New Orleans (MA). His doctoral dissertation in social sciences is from the University of Utrecht, The Netherlands, where he has lived and worked since 1976. After six years heading the Department of Documentary Information at the Royal Netherlands Academy of Arts and Sciences (SWIDOC/KNAW), Farace founded GreyNet, Grey Literature Network Service in 1992. He has since been responsible for the International Conference Series on Grey Literature (1993-2013). In this capacity, he also serves as Program and Conference Director as well as managing editor of the Conference Proceedings. He is editor of The Grey Journal and provides workshops and training in the field of grey literature. Email: info@grey.net.org

Frantzen, Jerry**27**

Jerry Frantzen graduated in 1999 from the Amsterdam University of Applied Sciences/Hogeschool van Amsterdam (HvA) in Library and Information Science. Frantzen is the technical editor of The Grey Journal (TGJ). And, since 1996, he is affiliated with GreyNet, Grey Literature Network Service, as a freelance technical consultant. Email: info@grey.net.org

Frosini, Luca**113**

Luca Frosini is Researcher at Networked Multimedia Information Systems (NeMIS) Laboratory of the Italian National Research Council - Institute of Information Science and Technologies (CNR-ISTI). He graduated in Computer Science in 2006 at University of Pisa. Luca was involved in various EU-funded projects (e.g. DILIGENT, D4Science, BlueBRIDGE, PARTHENOS, SoBigData). His research interests include Data Infrastructures, Virtual Research Environments, Software Repositories, Accounting systems, Distributed Information System and Grid and Cloud Computing. He is one of the top most contributor of gCube software, the open-source platform for the management and operation of scientific data infrastructures. Email: luca.frosini@isti.cnr.it

**Giannini, Silvia****11**

Silvia Giannini graduated and specialized in library sciences. Since 1987 she has been working in Pisa at the Institute for the Science and Technologies of Information "A. Faedo" of the Italian National Council of Research (ISTI-CNR) as a librarian. She is a member of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She is responsible of the library automation software "Libero" in use at the CNR Research Area in Pisa and coordinates the bibliographic and managing activities of the ISTI library team. She cooperates in the design and development of the PUMA (PUBlication MANagement) & MetaPub, an infrastructure software for institutional and thematic Open Access repositories of published and grey literature produced by CNR. Email: silvia.giannini@isti.cnr.it

Goggi, Sara**93**

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

La Bruzzo, Sandro**107**

Sandro La Bruzzo is a research fellow at Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. He received his MSc in Information Technologies in the year 2010 at the University of Pisa, Italy. Today he is a member of the InfraScience research group, part of the Multimedia Networked Information System Laboratory (NeMIS). His current research interests are in the areas of Service-Oriented Infrastructures for Digital Libraries, protocols for metadata exchanging, Database, Index. He is currently working for the development of the Digital Library and Data infrastructures for the European Commission projects OpenAIRE, OpenAIREplus, OpenAIRE2020, EFG1914, HOPE, and EAGLE. Email: sandro.labruzzo@isti.cnr.it

Lipinski, Tomas A.**37**

Professor Tomas A. Lipinski completed his Juris Doctor (J.D.) from Marquette University Law School, Milwaukee, Wisconsin, received the Master of Laws (LL.M.) from The John Marshall Law School, Chicago, IL, and the Ph.D. from the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Dr. Lipinski has worked in a variety of legal settings including the private, public and non-profit sectors. He is the author of numerous articles and book chapters; his monographs include, *The Library's Legal Answer Book* co-authored with Mary Minow (2003); *The Copyright Law In The Distance Education Classroom* (2005), *The Complete Copyright Liability Handbook For Librarians And Educators* (2006), and *The Librarian's Legal Companion For Licensing Information Resources And Services* (2012). Recent articles and chapters

include, *Click Here to Cloud: End User Issues in Cloud Computing Terms of Service Agreements*, in *Challenges Of Information Management Beyond The Cloud: 4th International Symposium On Information Management In A Changing World*, Imcw 2013 (Revised Selected Papers.), with Kathrine Henderson, *Hate Speech: Legal and Philosophical Aspects*, in *The Handbook Of Intellectual Freedom Concepts* (2014), in 2013 with Andrea Copeland, *Look before you License: The Use of Public Sharing Websites in building Patron Initiated Public Library Repositories*, *Preservation, Digital Technology & Culture* and in 2012, *Law vs. Ethics, Conflict and Contrast in Laws Affecting the Role of Libraries, Schools and other Information Intermediaries*, *Journal Of Information Ethics*. He has been a visiting professor in summers at the University of Pretoria-School of Information Technology (Pretoria, South Africa) and at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Lipinski was the first named member of the Global Law Faculty, Faculty of Law, University of Leuven, Belgium, in 2006 where he continues to lecture annually at its Centers for Intellectual Property Rights and Interdisciplinary Center for Law and ICT. He is active in copyright education and policy-making, chairing the ACRL Copyright Discussion Group, a member of the ALA OITP Committee on Legislation Copyright Subcommittee, a member of the Copyright and Other Legal Matters Committee of IFLA and serves as an IFLA delegate to the World Intellectual Property Organization's Standing Committee on Copyright and Other Rights. In October of 2014 he returned to the University of Wisconsin to serve as Professor and Dean of its i-School, the School of Information Studies. Email: tlipinsk@uwm.edu

Mack, Daniel C.**55**

Daniel C. Mack is Associate Dean for Collection Strategies and Services, University Libraries at the University of Maryland in College Park, where he provides leadership in policy creation and implementation, strategic planning, program development, and assessment for library collections. He is also responsible for coordinating copyright and licensing issues for faculty produced publications. His previous positions include Tombros Librarian for Classics and Ancient Mediterranean Studies and Head of the Arts and Humanities Library at Penn State, and Library Director at the Dauphin County (PA) Prison. Mack has advanced degrees in library science and ancient history and has taught college courses in ancient history, Roman archaeology, classical literature and Latin grammar. Recent publications include work as co-editor of the Association of College and Research Libraries' monograph *Interdisciplinarity and Academic Libraries*, as consulting editor for Brill's *New Pauly: Encyclopaedia of the Ancient World* and as author of the "Language, Linguistics and Philology" section of the American Library Association's *Guide to Reference Sources*. Mack's current research interests include interdisciplinarity in the twenty-first century academy, assessment of library collections and services, and Roman civilization in the age of Caesar Augustus. When he has time, Mack plays the viola da gamba and cello. Email: dmack@umd.edu

**Manghi, Paolo****107**

Paolo Manghi is a (PhD) Researcher in computer science at Istituto di Scienza e Tecnologie dell'Informazione (ISTI) of Consiglio Nazionale delle Ricerche (CNR), in Pisa, Italy. He is the technical director of the OpenAIRE infrastructure, technical manager and researcher for the EU-H2020 infrastructure projects OpenAIRE2020, SoBigData.eu, PARTHENOS, EOSC, and RDA Europe, and he is the scientific coordinator of the OpenAIRE-Connect project. He is active member of a number of Data Citation and Data Publishing Working groups of the Research Data Alliance; and invited member of the advisory boards of EC project and the Research Object initiative. His research areas of interest are today data e-infrastructures for science and scholarly communication infrastructures, with a focus on technologies supporting open science publishing, i.e. computational reproducibility and transparent evaluation of science. Email: paolo.manghi@isti.cnr.it

Molino, Anna**11**

Anna Molino graduated in Linguistics at the University of Pisa in 2010. Since 2013, she works at CNR - ISTI ("Istituto di Scienza e Tecnologie dell'Informazione - A. Faedo") as member of the Networked Multimedia Information Systems Lab. (NeMIS). She has worked as project assistant and financial manager in various EU funded and national research projects for the Digital Libraries group of the NeMIS lab. She contributes in the language revision and translation of scientific papers. Email: anna.molino@isti.cnr.it

Monachini, Monica**93**

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

Nekhayenko, Oleg**85**

Oleg Nekhayenko is research assistant at the German National Library of Science and Technology in the department of digital preservation. He received a bachelor's degree in Information Science and Language Technologies from Heinrich-Heine University Duesseldorf and a master's degree in International Information Management from University of Hildesheim. Email: Oleg.Nekhayenko@tib.eu

Pagano, Pasquale**113**

Pasquale Pagano is Senior Researcher at the Networked Multimedia Information Systems Laboratory of the "Istituto di Scienza e Tecnologie della Informazione A. Faedo" (ISTI) of the Italian National Research Council (CNR). He received his M.Sc. in Information Systems Technologies from the Department of Computer Science of the University of Pisa (1998), and the Ph.D. degree in Information Engineering

from the Department of Information Engineering: Electronics, Information Theory, Telecommunications of the same university (2006). "The aim of my research is the study and experimentation of models, methodologies and techniques for the design and development of distributed virtual research environments (VREs) which require the handling of heterogeneous computational and storage resources, provided by Grid and Cloud based e-Infrastructures, for the management of heterogenous data sources. I have a strong background on distributed architectures. I participated to the design of the most relevant distributed systems and e-Infrastructure enabling middleware developed by ISTI - CNR. I am currently the Technical Director of D4Science, the Hybrid Data Infrastructure serving scientists in 37 countries, and chief manager of gCube software, the open-source platform for the management and operation of scientific data infrastructures. I am collaborating with the Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources (iMarine). I am serving the BlueBRIDGE European Project as Technical Director, the Parthenos European Project as Service Operation Manager, and the SoBigData Research Infrastructure as Infrastructure Manager." Email: pasquale.pagano@isti.cnr.it

Pardelli, Gabriella**93**

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

Plank, Margret**127**

Margret Plank is currently the Head of the Competence Centre for Non-Textual Materials at the German National Library of Science and Technology in Hannover (Germany). The aim of the Competence Centre for Non-Textual Materials is to develop emerging tools and services that actively support users in the scientific work process enabling non-textual material such as audiovisual media, 3D objects and research data to be published, found and made available on a permanent basis as easily as textual documents. Previously she was responsible for Information Competence and Usability at the TIB. She has also worked as a researcher at the Institute of Information Studies and Language Technology at the University of Hildesheim. She represents TIB on a number of boards including IFLA Steering Committee Audiovisual and Multimedia Section as well as ICSTI / ITOC. Margret Plank holds a Master degree in information science and media studies from the University of Hildesheim, Germany. Email: margret.plank@tib.uni-hannover.de



Prost, H  l  ne

121

H  l  ne Prost is information professional at the Institute of Scientific and Technical Information (CNRS) and associate member of the GERiICO research laboratory (University of Lille 3). She is interested in empirical library and information sciences and statistical data analysis. She participates in research projects on evaluation of collections, document delivery, usage analysis, grey literature and open access, and she is author of several publications.

Email: helene.prost@inist.fr

Sch  pfel, Joachim

121

Joachim Sch  pfel is senior lecturer at the Department of Information and Library Sciences at the Charles de Gaulle University of Lille 3 and Researcher at the GERiICO laboratory. He is interested in scientific information, academic publishing, open access, grey literature and eScience. He is a member of GreyNet and euroCRIS. He is also the Director of the National Digitization Centre for PhD Theses (ANRT) in Lille, France.

Email: joachim.schopfel@univ-lille3.fr

Smith, Plato L.

27

Plato Smith is the Data Management Librarian at the University of Florida with experience in academic research libraries, digital libraries, and data management. He received his doctorate in the field of Information Science from the School of Information within the College of Communication and Information at Florida State University, Florida's iSchool, Summer 2014. From 2005 to 2012, he was Department Head for the FSU Libraries' Digital Library where he developed, populated, and managed digital collections in the FSU Libraries' digital content management system, DigiNole Repository, and electronic theses and dissertations (ETDs) institutional repository. Email: plato.smith@ufl.edu

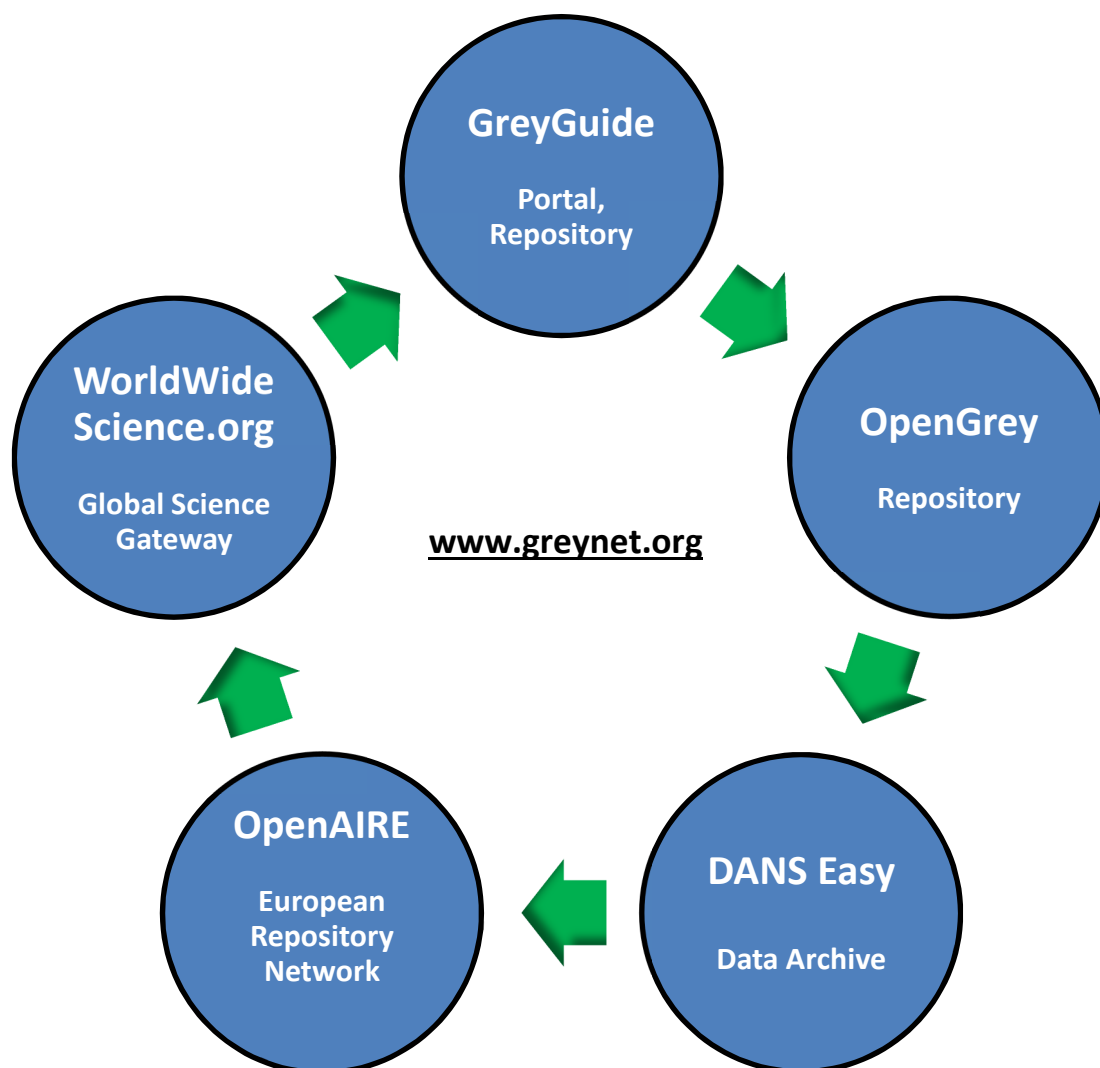
Vy  italov  , Hana

101

Hana Vy  italov   studied Information science and librarianship at the Charles University in Prague. Since 2012 she works in the National Library of Technology in Prague (Czech Republic) in Digital National Library of Technology Department. Currently she is partnership network manager of the National Repository of Grey Literature. She is interested in grey literature, open access, research data, enhanced publications and free licences. She is coordinator of the Conference on Grey Literature and Repositories in Czech Republic. Email: hana.vycitalova@techlib.cz

GreyNet's Service Providers

GreyNet is fully open access compliant. Authors and Researchers in grey literature communities worldwide know that their metadata, full-texts, slide presentations, research data, and other outputs are preserved and made openly accessible to the broader public.



1992  2017



Index to Authors

A-B

Artini, Michele	107
Atzori, Claudio	107
Baglioni, Miriam	107
Balashova, Yuliya B.	33
Bardi, Alessia	107
Bartolini, Roberto	93
Biagioni, Stefania	11

C-D

Cancedda, Flavia	133
Cardillo, Elena	61
Černohlávková, Petra	101
Chamberlain Kritikos, Katie	37
Darmoni, Stzkefan	61
De Biagi, Luisa	133
De Bonis, Michele	107
Deluca, Rosaria	11
Dimitropoulos, Harry	107
Drees, Bastian	127

F-G

Farace Dominic	27
Foufoulas, Ioannis	107
Frantzen, Jerry	27
Frosini, Luca	113
Giannini, Silvia	11
Goggi, Sara	93
Grosjean Julien	61

I-J

Iatropoulou, Katerina	107
Ittoo, Ashwin	61
Jamoulle, Marc	61

L-M

La Bruzzo, Sandro	107
Lipinski, Tomas A.	37
Mack, Daniel C.	55
Manghi, Paolo	107
Manola, Natalia	107
Martziou, Stefania	107
Molino, Anna	11
Monachini, Monica	93

N-O-P

Nekhayenko, Oleg	85
Pagano, Pasquale	113
Pardelli, Gabriella	93
Plank, Margret	127
Principe, Pedro	107
Prost, Hélène	121

R-S

Resnick, Melissa P.	61
Russo, Irene	93
Schöpfel, Joachim	121
Smith, Plato L.	27

V-Z

Vander Stichele, Robert	61
Vanmeerbeek, Marc	61
Vyčítalová, Hana	101

Forthcoming
February 2018


'Public Awareness and Access to Grey literature'

National Research Council of Italy, CNR Rome * October 23-24, 2017

Publication Order Form

NINETEENTH INTERNATIONAL CONFERENCE ON GREY LITERATURE

Publication(s):	No. of Copies	x	Amount in Euros	Subtotal
GL19 CONFERENCE PROCEEDINGS - Printed Edition ISBN 978-90-77484-31-9 ISSN 1386-2316 <i>Postage and Handling excluded*)</i>		x	109.00 = €	
GL19 CONFERENCE PROCEEDINGS - PDF Edition ISBN 978-90-77484-31-9 ISSN 1386-2316 <i>Forwarded via email</i>		x	109.00 = €	
GL19 Conference Proceedings - Online Edition ISBN 978-90-77484-31-9 ISSN 2211-7199 <i>Password Protected Access</i>		x	109.00 = €	



*POSTAGE AND HANDLING PER PRINTED COPY *)*

Holland	<input type="text"/>	x	5.00	€
Other	<input type="text"/>	x	15.00	€
TOTAL EURO =				€

Customer Name:	
Organisation:	
Postal Address:	
City/Code/Country:	
E-mail Address:	

☐ Direct transfer to TextRelease, Rabobank Amsterdam
BIC: RABONL2U IBAN: NL70 RABO 0313 5853 42, with reference to "GL19 Publication Order"

☐ MasterCard/Eurocard ☐ Visa Card ☐ American Express

Card No. _____ Expiration Date: _____

Print the name that appears on the credit card, here _____

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

Place: _____ Date: _____

NOTE: CREDIT CARD TRANSACTIONS WILL BE AUTHORIZED VIA OGONE/INGENICO DESIGNATED PAYMENT SERVICES

TextRelease
www.textrelease.com

GL19 Program and Conference Bureau
Javastraat 194-HS, 1095 CP Amsterdam, Netherlands
T/F +31-(0) 20-331.2420 Email: info@textrelease.com