

An International Journal on Grey Literature



Summer 2018, Volume 14, Number 2

‘BREAKING NEW GROUND IN GREY LITERATURE’

GreyNet

The Grey Journal

An International Journal on Grey Literature

COLOPHON

Journal Editor:

Dr. Dominic Farace
GreyNet International,
Grey Literature Network Service
Netherlands
journal@grey.net

Associate Editors:

Julia Gelfand
University of California, Irvine
United States

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

Dr. Joachim Schöpfel
Université Charles de Gaulle Lille 3
France

Prof. Gretta E. Siegel (Emeritus)
Portland State University
United States

Dr. Plato L. Smith
University of Florida
George A. Smathers Libraries
United States

Technical Editor:

Jerry Frantzen, TextRelease

CIP

The Grey Journal (TGJ) : An international journal on grey literature / Dominic Farace (Journal Editor); Jerry Frantzen (Technical Editor) ; GreyNet International, Grey Literature Network Service. - Amsterdam: TextRelease, Volume 14, Number 2, Summer 2018. - *TIB (DE), DANS-KNAW (NL), FEDLINK-Library of Congress (USA), CVTISR (SK), EBSCO (USA), Inist-CNRS (FR), ISTI-CNR (IT), KISTI (KR), NIS-IAEA (AT), NTK (CZ), and the University of Florida (USA) are Corporate Authors and Associate Members of GreyNet International.* ISSN 1574-1796 (Print)
ISSN 1574-180X (PDF)
ISSN 1574-9320 (CD-Rom)

Subscription Rates:

€240 institutional, including postage & handling

Contact Address:

Back Issues, Document Delivery, Advertising, Inserts, Subscriptions:

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

About TGJ

The Grey Journal is a flagship journal for the international grey literature community. It crosses continents, disciplines, and sectors both public and private.

The Grey Journal not only deals with the topic of grey literature but is itself a document type classified as grey literature. It is akin to other grey serial publications, such as conference proceedings, reports, working papers, etc.



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

About GreyNet

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

Full-Text License Agreement

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

© 2018 TextRelease

Copyright, all rights reserved. No part of this publication may be reproduced, stored in or introduced into a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise without prior permission of the publisher.

Contents

'BREAKING NEW GROUND IN GREY LITERATURE'

Grey Literature and Research Assessment Exercises: From the Current Criteria to the Open Science models	55
<i>Silvia Giannini, Rosaria Deluca, Anna Molino, and Stefania Biagioni (Italy)</i>	
A Facet-based Open and Extensible Resource Model for Research Data Infrastructures	69
<i>Luca Frosini and Pasquale Pagano (Italy)</i>	
Impact of Disruptive Technologies on Grey Literature Management	77
<i>Dobrica Savić (Austria)</i>	
Collecting Grey Literature – Institutional Repository versus National Aggregator	81
<i>Petra Černohlávková and Hana Vyčítalová (Czech Republic)</i>	
Providing Access to Grey Literature: The CLARIN Infrastructure	87
<i>Sara Goggi, Gabriella Pardelli, Irene Russo, Roberto Bartolini, and Monica Monachini (Italy)</i>	
Preserving and accessing content stored on USB-flash-drives: A TIB workflow	95
<i>Oleg Nekhayenko (Germany)</i>	

Colophon.....	50
Editor's Note.....	53
On The Newsfront	
GL20 Conference Announcement - Twentieth International Conference on Grey Literature "Research Data Fuels and Sustains Grey Literature" Loyola University, New Orleans, Louisiana, USA • December 3-4, 2018	102
GL20 Call for Posters - Twentieth International Conference on Grey Literature Loyola University, New Orleans, Louisiana, USA • December 3-4, 2018	103
Open Data Forum - Open Data for Research and Enhanced Publication GreyForum Series Center for National Research, Pisa, Italy • May 25, 2018	104
Advertisements	
GreyGuide Portal and Repositories "Sharing knowledge as early as possible"	52
ISTI-CNR, Institute of Information Science and Technologies.....	54
CVTISR, Slovak Centre of Scientific and Technical Information.....	94
Author Information	105
Notes for Contributors.....	107

GreyGuide Portal and Repositories

<https://doi.org/10.26089/greynet-2025-000.531-1qj>

"sharing knowledge as early as possible"



<http://greyguide.isti.cnr.it>

EDITOR'S NOTE

In 2011, GreyNet embarked on an Enhanced Publications Project (EPP) in order to link its collection of full text conference papers with accompanying research data. The initial phase in the study dealt with the design and implementation of an online questionnaire among authors, who were published in the International Conference Series on Grey Literature. From 2012 onwards, subsequent phases in the project dealt with the acquisition, submission, indexing, and archiving of GreyNet's collection of published datasets now housed in the DANS EASY data archive.

In 2017, GreyNet's Enhanced Publications Project was further broadened to include a Data Papers Project. Here, emphasis focused on describing the data rather than analyzing it. As such, the data paper signals data sharing and in this way promotes both data citation and the potential reuse of research data in line with the FAIR Guiding Principles for scientific data management and stewardship.

Available results from the Data Papers Project presented last year at GL19 concludes where this study commences entitled: *"Open Data engages Citation and Reuse: A Follow-up Study on Enhanced Publication"*. Here, GreyNet now seeks to demonstrate the reuse of survey data collected in 2011 combined with survey data that will be newly collected via an online questionnaire. The survey population will be drawn from among GreyNet's author base; and, a selection of questions from the 2011 Survey will be joined by newly formulated questions in constructing the questionnaire. Furthermore, GreyNet relying upon available use and usage statistics compiled from various sources will seek to provide evidence of data citation and referencing.

The results of this study are expected to demonstrate an increased willingness among GreyNet authors to share their research data – this in part due to GreyNet's program of enhanced publication embedded in its workflow over the past six years. The study will provide an example of the reuse and further comparison of the results of survey data, which can be incorporated in GreyNet's program of training and instruction. However, statistics on data citation and referencing is expected to prove a formidable challenge.

Dominic Farace,
Journal Editor



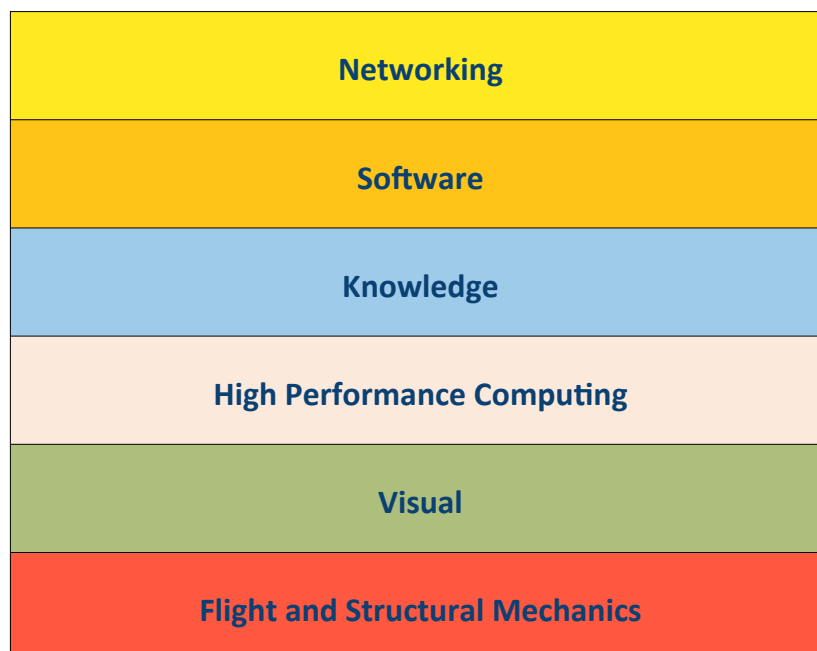
Institute of Information Science and Technologies “A. Faedo”

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 (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

* First published in the GL19 Conference Proceedings, February 2018.

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 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 sub-divided into Scientific Disciplinary Sectors [*Settori Scientifico-Disciplinari (SSD)*]³.

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/>.

³ Cfr. "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...*

Macro-Areas VQR1

Macro-Areas VQR2

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

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

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

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*

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.

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

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 *catalogographic 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, 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).

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).

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%.

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).

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).

Categories	VQR 2011-2014															
	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			0.20		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									
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 or 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

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

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 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).

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*.

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, 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

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/>.

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.

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 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,

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/>.

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.

Bibliography¹⁷

- 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.crui.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.

¹⁷ URL last access: December 2017.

- 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.
- 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/ \(VQR2\)](http://www.anvur.org/rapporto/ (VQR1) and http://www.anvur.org/rapporto-2016/ (VQR2)).

A Facet-based Open and Extensible Resource Model for Research Data Infrastructures^{*}

Luca Frosini and Pasquale Pagano (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 Date 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 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.

^{*} First published in the GL19 Conference Proceedings, February 2018.

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

² www.d4science.org

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.

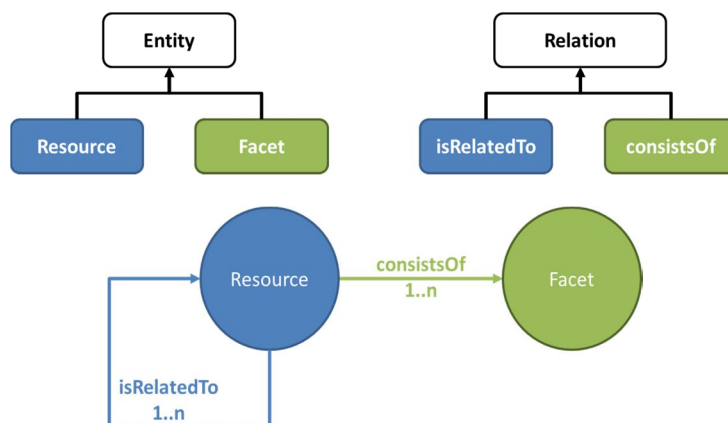


Fig. 1 represent the main concepts of the model.

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).

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-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{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.

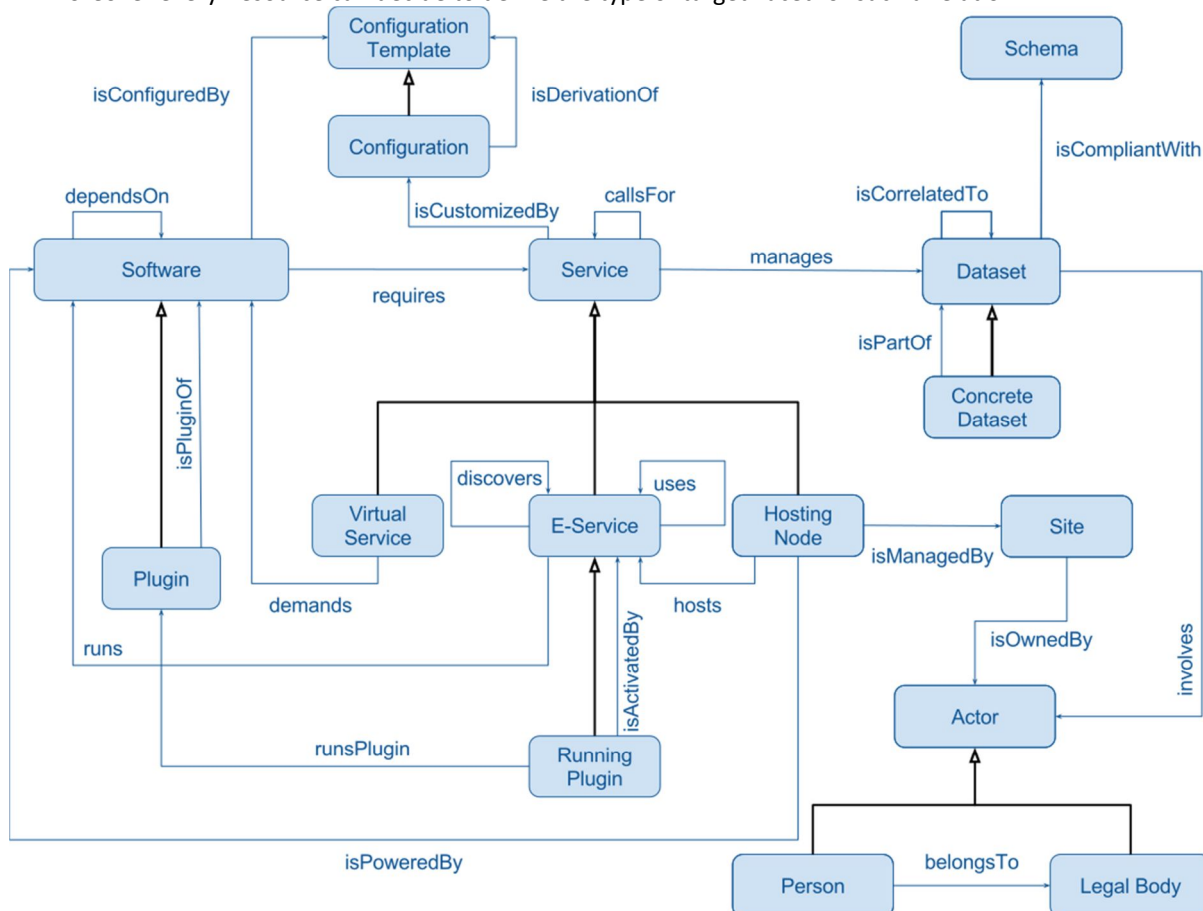


Fig. 2 Overview of gCube Model Resources and (isRelatedTo) Relations

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-Service* 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 represents 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>

Impact of Disruptive Technologies on Grey Literature Management *

Dobrica Savić (Austria)

Abstract

Analysis of disruptive information technologies (IT) is based on five 2017 reports, issued by several leading international consulting companies with a special interest in researching the impact of IT on business and work environments. Four technologies were identified with a potentially high impact on information and grey literature management. The major disruptors identified are artificial intelligence and machine learning, virtual and augmented reality, internet of things and big data.

Introduction

Information technology has become a driving force for change and innovation of our work processes and the services we offer. It is particularly evident in the way we create, distribute and use information products and services. Management of grey literature is a good example of this. The emerging new digital information environment has the potential to disrupt all aspects of our work and the world around us. In order to meet expectations and benefit from this challenging opportunity, information managers need to be aware of new IT trends, the disruptors they bring and the disruptions they might cause. At the same time, these changes also offer opportunities to improve our work.

Grey Literature Defined

At the 12th International Conference on Grey Literature (GL12), held in Prague in 2010, the most widely currently used definition of grey literature (GL) was drafted. It states that ***"Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body"*** (Farace, D. and Schöpfel, J., 2010).

An attempt to add an additional twist to the definition of grey literature was a suggestion to look at it from the perspective of traditional publishing, which usually goes through a peer-review process. This revised definition is regarded as: ***"the diverse and heterogeneous body of material that is made public outside, and not subject to, traditional academic peer-review processes"*** (Adams et al. 2016).

The current concept of GL, as stated in the Prague definition, faces some challenges. The main challenge relates to multiple types of originators; humans and machines, volume and type, and the speed of GL creation. In order to overcome a possibility of the definition becoming obsolete due to its inability to differentiate GL from other types of literature, I have proposed a new definition, which might help meet some of the above-mentioned challenges. Accordingly, ***grey literature is any recorded, referable and sustainable data or information resource of current or future value, made publically available without a traditional peer-review process.***

Disruptive Information Technologies

The term ***'disruptive technology'*** was coined by Harvard Business School professor Clayton M. Christensen in his 1997 best-selling book - "The Innovator's Dilemma" (Christensen, 1997). There, he separated new technology into two categories: sustaining and disruptive. Sustaining technology relies on incremental improvements to an already established technology, while disruptive technology lacks refinement, often has performance problems because it is new, appeals to a limited audience and may not yet have a proven practical application.

In another words, a disruptive technology displaces an established technology and shakes up the industry, or introduces a ground-breaking product that creates a completely new industry. During the last few decades, information technology played that role by introducing new solutions that shook up the world, especially the world of business. IT continues doing so by introducing newer and newer products and solutions which impact our lives and the way we organize our work and conduct business.

Many leading international consulting companies have a special interest in researching the impact of IT on business and work environments. For this review, five reports created by leading IT consulting companies were selected (Table 1). Although somewhat different in their approach and the way they look at the future, the IT trends identified offer a good starting point to determine major IT influences and potential disruptors that will define the near future.

* First published in ITlib.Informacne tehnologije a kniznice. 4/2017. – Bratislava : SCSTI, 2017

Gartner	Forbes	Forrester	Deloitte	Accenture
1.AI and advanced machine learning 2.Intelligent apps 3.Intelligent things 4.Virtual and augmented reality 5.Digital twin 6.Blockchain and Distributed Ledgers 7.Conversational System 8.Mesh App and Service Architecture 9.Digital Technology Platforms 10.Adaptive Security Architecture Gartner's Top 10 Strategic Technology Trends for 2017	1.IoT and smart home tech 2.AR and VR 3.Machine learning 4.Automation 5.Humanized Big Data. (visual, empathetic, qualitative) 6.Physical-Digital Integrations 7.Everything On-Demand 7 Technology Trends That Will Dominate 2017	Engagement technologies 1.IoT software and solutions 2.Intelligent agents 3.Personal identity and data management 4.Real-time interaction management 5.Augmented and virtual reality Insight technologies 1.AI/cognitive 2.Customer journey analytics 3.Insight platforms 4.IoT analytics 5.Spatial analytics Supporting technologies 1.Security automation and orchestration 2.Containers and container management 3.Edge computing 4.Cloud native application platforms 5.Hybrid wireless The Top Technology Trends To Watch: 2017 To 2021	1.IT unbounded 2.Dark analytics 3.Machine intelligence 4.Mixed reality 5.Inevitable architecture 6.Everything as a service 7.Blockchain: Trust economy 8.Exponentials watch list Tech Trends 2017: The kinetic enterprise	1.AI is the new UI 2.Ecosystems power play 3.Workforce marketplace 4.Design for humans 5.The uncharted Technology Vision 2017 – Technology for People: The Era of the Intelligent Enterprise

Table 1: 2017 Information technology trends

After looking at the reports, four major potential disruptors of grey literature management were identified. They are:

- Artificial intelligence and machine learning
- Virtual and augmented reality
- Internet of things
- Big data



Artificial intelligence and machine learning

Artificial intelligence (AI) and machine learning applications are systems that can think and act rationally, almost like humans. They are usually very costly and complex to develop maintain and deploy. Their power comes from a combination of many technologies and techniques, such as deep learning, neural networks, and natural-language processing (NLP). Initially they appeared as rule-based or expert systems, but today's algorithms can understand, learn, predict, adapt and potentially operate autonomously. They are often built into physical devices (e.g., robots, cars, consumer electronics, and security systems); into apps and services (e.g., virtual personal assistants, smart advisors, voice recognition, computer vision, translation, and finance). Applied in the area of information and knowledge management they become a powerful help in processing, organizing and disseminating data and information.

Virtual and augmented reality

Virtual reality (VR) systems take us out of our reality and bring us to some other place, while augmented reality (AR) takes our current reality and adds something to it. A simple way to understand the difference between VR and AR is to compare scuba diving to the aquarium. For example, VR can bring us to a construction site where we can walk in any direction and see every detail, while AR could be helpful for a client who can't visualize something. The idea is that a designer, an architect and a homeowner could sit around a table and look at the same 3D model on the table instead of a 2D plan. The power and the danger of this type of technology lies in the fact that the human mind is not able to tell the

difference between computer-generated images and the real world. This technology offers many possibilities for developing applications for the military, medicine, science, manufacturing, real estate, fashion, navigation, sightseeing, advertising and promotion, and games like Pokémon Go.

AR's ability to enhance what already exists is what makes it an excellent fit for document repositories, libraries, museums and similar institutions.

Internet of things

The Internet of Things (IoT) is a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction. IoT systems require an object, often referred to as a 'thing', a unique identifier (IP), Wi-Fi, sensors, and electronic circuits for control. A 'thing' can be goods, objects, machines, appliances, buildings, vehicles, animals, people, plants, or even soil. Some examples include a person with a heart monitor implant, a farm animal with a biochip transponder, cars with built-in sensors to alert the driver when tire pressure is low, supply monitors, search and location, city management, controllers for electricity use, game immersion, etc. Something to keep in mind is that data creation and capture is moved from people to computers. Some of the challenges with IoT include its huge complexity and non-existent privacy.

Big data

There is wide spread and growing interest in 'Big Data.' This popular term is often used to describe the huge growth and availability of both structured and unstructured data. Generally, five V's are highlighted when talking about big data – Volume, Variety, Veracity, Velocity and Value.

The 5 Vs of Big Data



Currently, we are facing an exponential growth in the volume of data. Virtually all devices are now generating data in some way, shape, or form. A modern car has over 100 sensors built in. This translates into approximately 11 billion sensors used just in cars around the world. Think of the volume of data being generated by your TV cable subscription today versus just 20 years ago.

With big data, there are problems with collection, privacy, security, long-term preservation, technical and financial sustainability, standardization and future usability.

Impact on Grey Literature Management

The impact of the four disruptors mentioned above is already being felt across the wider information management sector, and it will continue to be of major concern, bringing many challenges and requiring well-planned, well-established, and well-financed actions. As a part of general information management, grey literature will be impacted by these technologies.

Based on the above brief summary of disruptors, four different groups of trends and impacts dealing with various aspects of grey literature management are identified. They are:

Technology -More difficult access to GL due to security constraints -Higher level of IT expertise required to access and process GL -More dynamic documents, less GL -New tech-driven/based forms -Tremendous increase of big data and related documents	Customers -High expectations (e.g. comprehensiveness, relevance, aggregation, added value) -Interconnectivity -Top of the line finding tools -Web 2.0 features (e.g. social networking, collaboration, user generated content) -Tools to analyse and exploit big data -Mobile expectations of the new generation -Lack of understanding of GL value and importance
Products/Services -Self generated data/information/docs. -Competition with 'big players' -Lack of interest to collect, preserve and make GL widely available -Difficulty with going beyond local repositories -Intellectual property protection -Disappearance of e-archives and older web-based materials	Employees -Availability of HR and financial resources -Lack of proper training and education -Limited career development -Frequent change of jobs and interests (i.e. lack of continuity and long-term planning) -Changing technical requirements -Business focus -Culture of preservation missing -Multitasking and rapid delivery expectations

Table 2: Key areas of grey literature management

The above table indicates that key areas of grey literature management, namely technology, customers, products/services, and employees will be impacted in the future by those disruptive technologies that have been identified. Sustainability of grey literature management requires many planned efforts to overcome all the anticipated challenges in order to continue making it valuable, viable and usable as future information resource.

Conclusions

In the last few decades, developments in information technology have had an immense impact on the way we manage information in general, and on the way we create, disseminate and use grey literature. Based on this review of the main disruptive information technologies, it can be concluded that grey literature will face considerable challenges in the future. Due to the impact of big data and the Internet of Things, grey literature volume will experience exponential growth, while due to the use of artificial intelligence and virtual and augmented reality, the management of grey literature will become more complex and more difficult.

In order to meet the challenges of disruptive technologies on the future of information and grey literature, it is necessary to increase knowledge of information specialists and improve cooperation among data and information specialists, librarians and archivists. Promotion and increased efforts to demonstrate the value of properly managed grey literature collections could bring about the necessary recognition of this important field of information management.

References

- Accenture. Technology Vision 2017 – Technology for People: The Era of the Intelligent Enterprise.
- Adams et al., 2016. Shades of Grey: Guidelines for Working with the Grey Literature in Systematic Reviews for Management and Organizational Studies. International Journal of Management Reviews. 2016. (<http://onlinelibrary.wiley.com/doi/10.1111/ijmr.12102/full>).
- Christensen, Clayton M. 1997. The Innovator's Dilemma. Harvard Business Review Press. USA.
- Deloitte. Tech Trends 2017: The Kinetic Enterprise.
- Gartner. Gartner's Top 10 Strategic Technology Trends for 2017.
- Farace, Dominic J. and Schöpfel, Joachim (Eds.), 2010. Grey Literature In Library and Information Studies. De Gruyter Saur, Germany.
- Forbes. 7 Technology Trends That Will Dominate 2017.
- Forrester. The Top Technology Trends To Watch: 2017 To 2021.

Collecting Grey Literature – Institutional Repository versus National Aggregator*

Petra Černohlávková and Hana Vyčítalová (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).

* First published in the GL19 Conference Proceedings, February 2018.

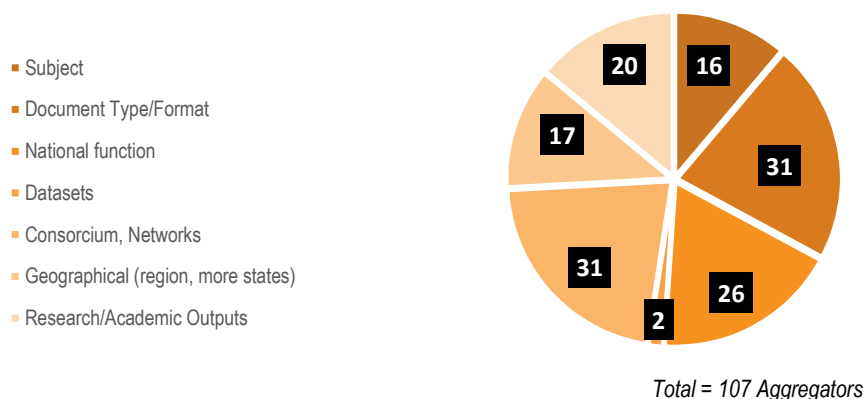
¹ <http://nusi.cz/?language=en>

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

³ OpenDoar, up to 17th October 2017

Focus of aggregators

Data source OpenDoar (2017/09/11)



N.b. Aggregators belong to several categories.

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.

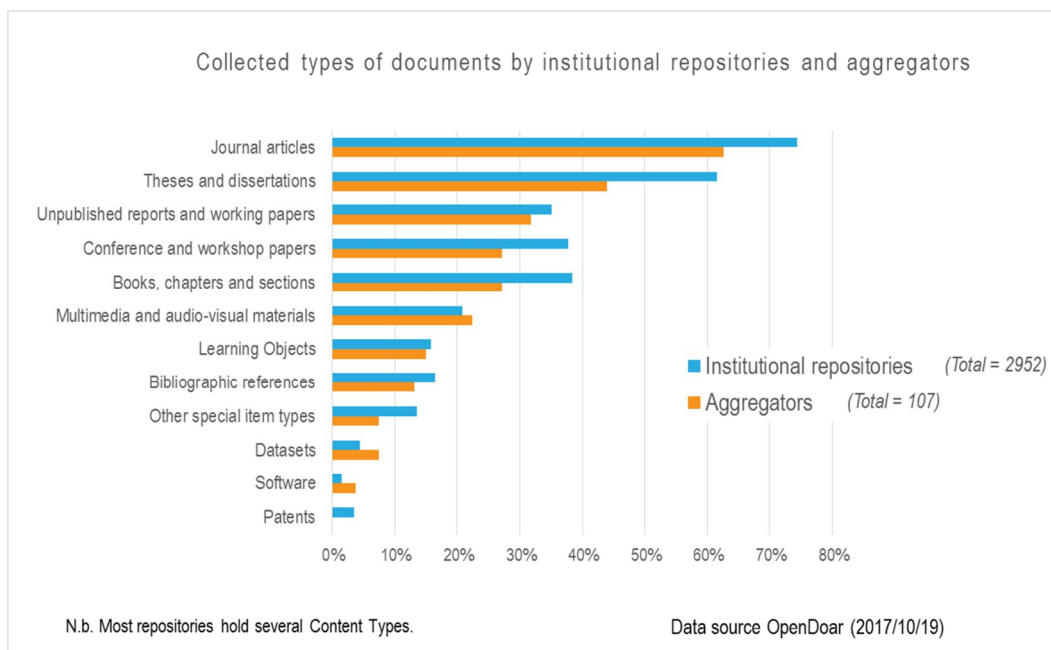


Chart 2: Types of documents collected by institutional repositories and aggregators (data source: OpenDOAR http://www.open_doar.org/)

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.

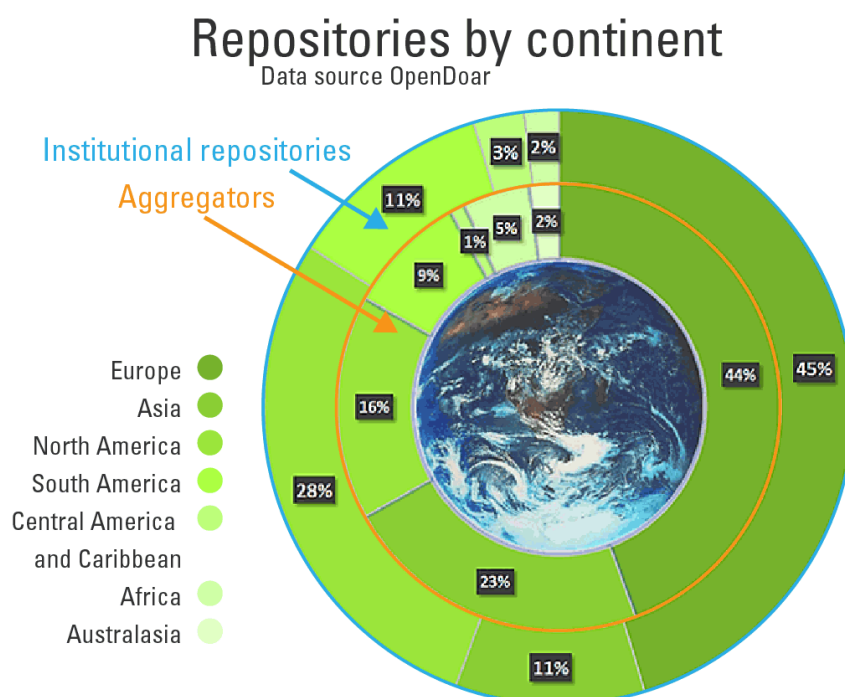


Chart 3: Repositories by continent (data source: OpenDOAR <http://www.opendoar.org/>)

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.

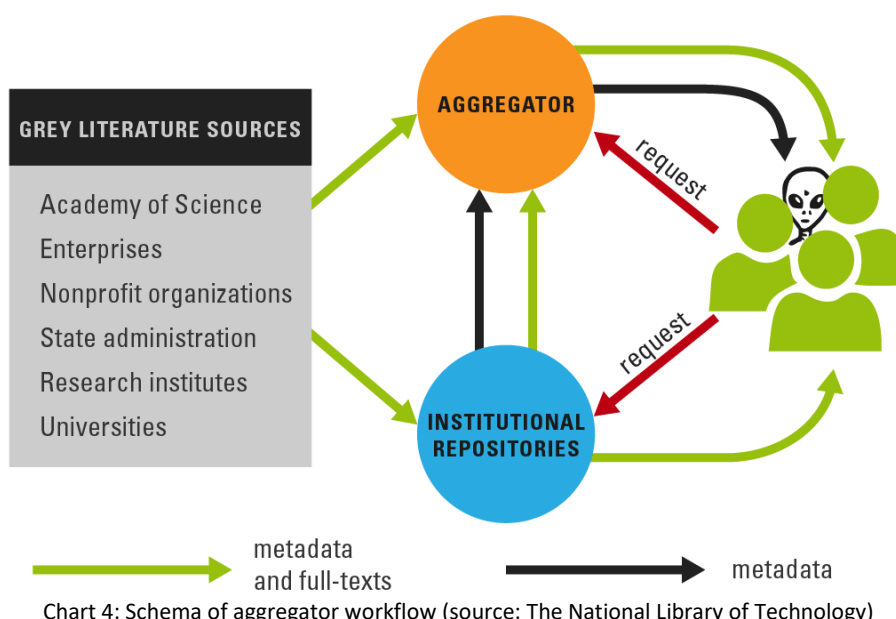


Chart 4: Schema of aggregator workflow (source: The National Library of Technology)

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.svkkk.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/

Providing Access to Grey Literature: The CLARIN Infrastructure*

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

* First published in the GL19 Conference Proceedings, February 2018.

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/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.

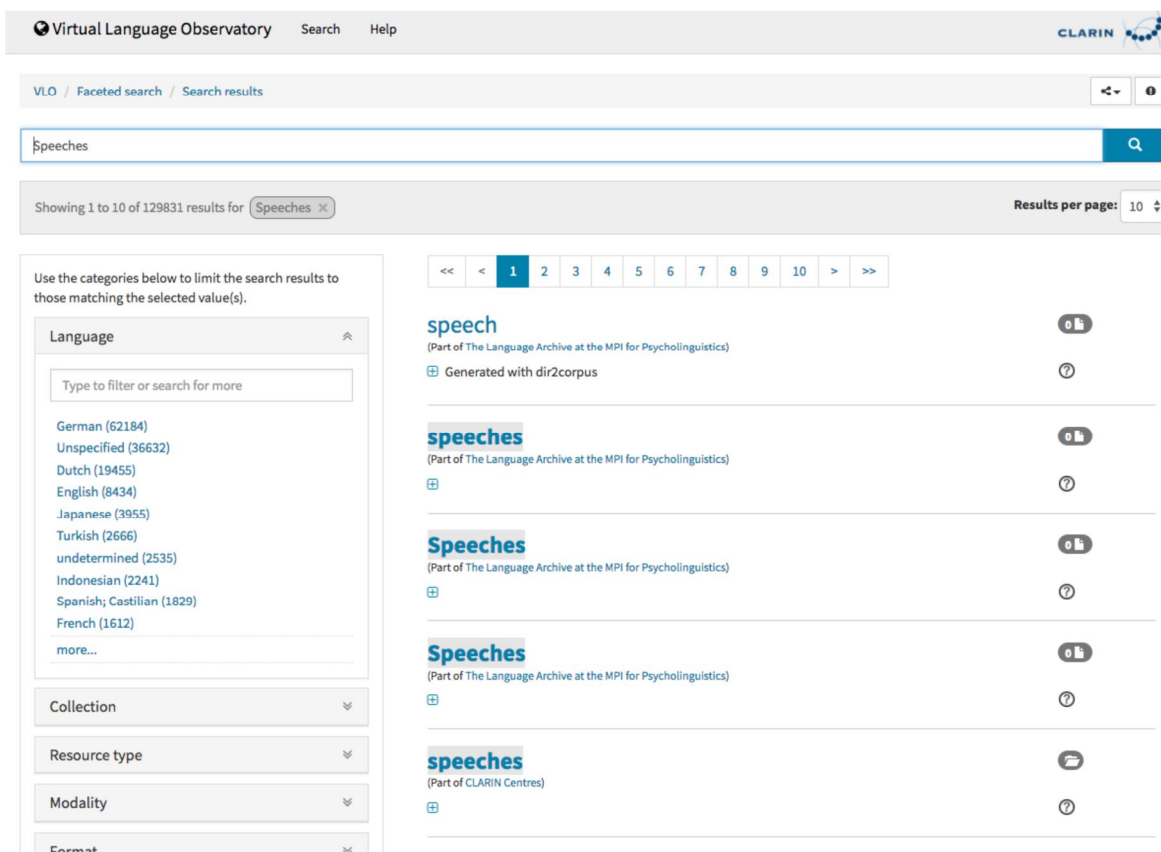


Fig. 1: Virtual Language Observatory

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.

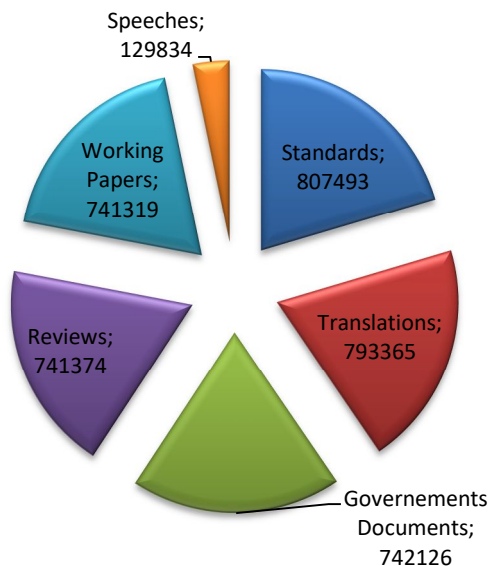
The idea is to use the VLO for checking the quantity of grey literature which can actually be found in CLARIN. A mapping between terms of the types in Grey Literature contained in the *GreyNet International 1992-2017* website (<http://www.greynet.org/>) - and in particular in the *GreySource Index > 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

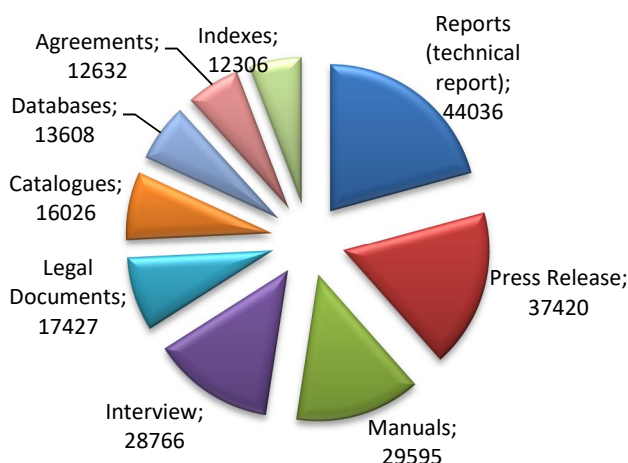


Graph.1. High frequency documents (from 1000000 to 100000)

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

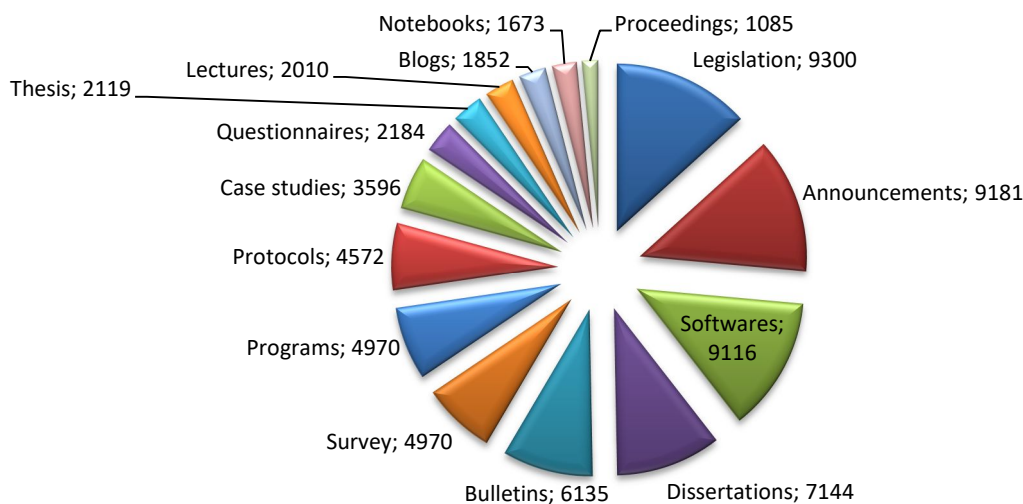
Graph 2 displays the nine types of GL documents contained in the VLO in the medium frequency range; the most frequent type is 'Reports', in particular technical reports describing resources and tools produced by various stakeholders (Business, Governments, Academia, etc.) and a high number of reports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

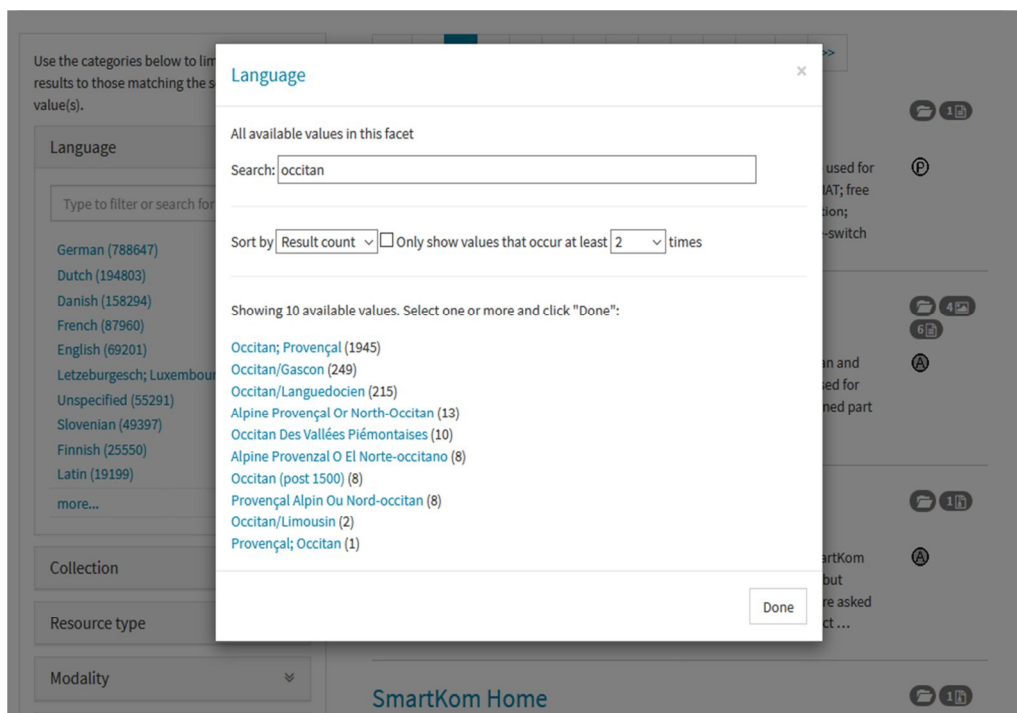
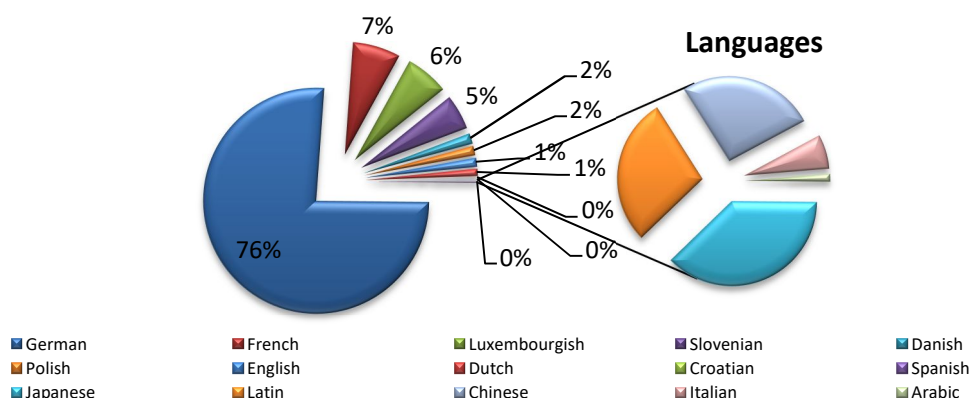


Fig.3. Languages VS GL documents



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.

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

Preserving and accessing content stored on USB-flash-drives: A TIB workflow^{*}

Oleg Nekhayenko (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

^{*} First published in the GL19 Conference Proceedings, February 2018.

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

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

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.

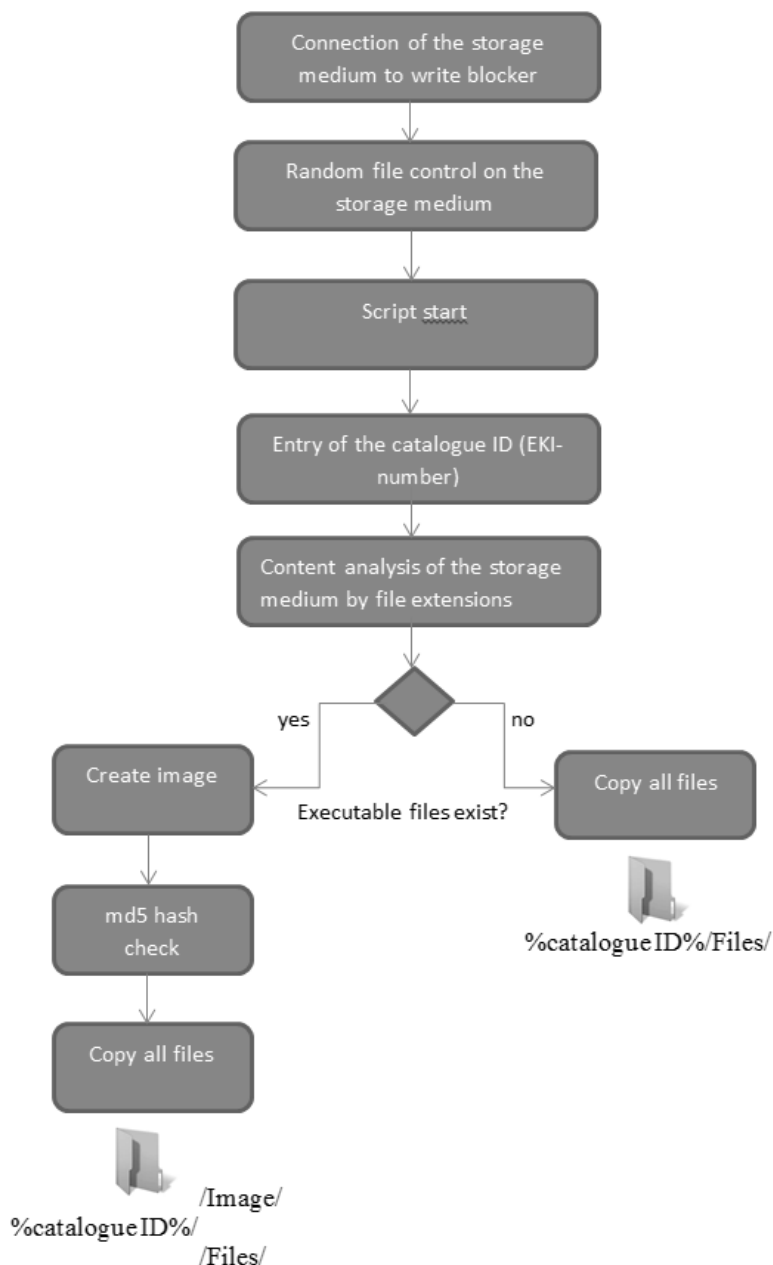


Figure 2: USB-Stick workflow

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.

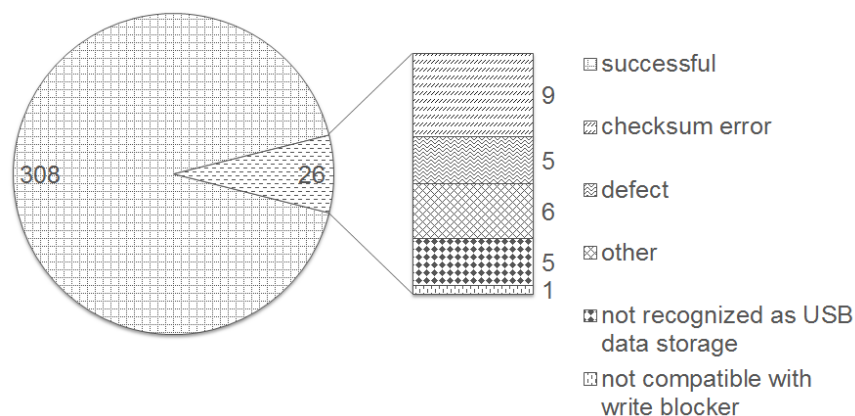


Figure 3: Distribution of the problem cases

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:

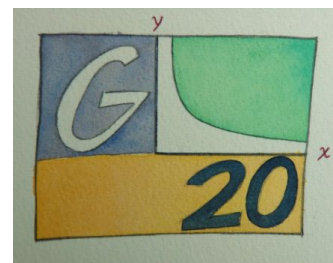
- Cunningham, A. (2014). A brief history of USB, what it replaced, and what has failed to replace it. Retrieved on 04.10.17 from <https://arstechnica.com/gadgets/2014/08/a-brief-history-of-usb-what-it-replaced-and-what-has-failed-to-replace-it/2/>
- Dappert, A., Jackson, A. N., & Kimura, A. (2013). Developing 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

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

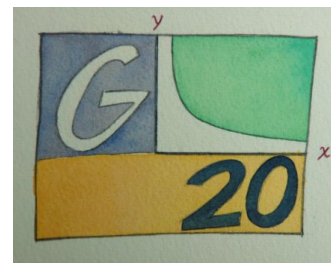
Tel. +31-20-331.2420

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 Posters

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

Participants who seek to present a poster during GL20 are invited to submit an English abstract between 250-300 words. The abstract should describe the project and/or related information product or service. The abstract should likewise include a title, name(s) of the presenter(s) and their full address information. Abstracts are an important source of information available prior to the conference that is accessible to both the delegates and conference participants.

Related 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, Format, and Poster Prize 2018

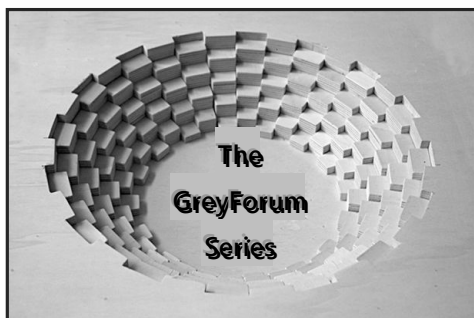
Timely registration is a guarantee of your placement on the conference program. The conference venue is able to accommodate a limited number of physical posters (20 inches x 30 inches / 50,8 cm x 76.2 cm), which will be mounted on foam board.

Abstracts in MSWord should be emailed to conference@textrelease.com on or before **October 30th 2018**. Those submitting poster abstracts will receive verification upon receipt and this will be accompanied by further guidelines for posters, submission of PowerPoint slides, Conference Registration, etc.

Those presenting conference posters are eligible for the Poster Prize 2018 that will be awarded during the GL20 Closing Session. Posters will be judged on their innovative content, relevance to the conference, graphic design, and presentation.

"The GreyForum is a series of onsite and online courses, seminars, and workshops where grey literature provides common ground for information professionals in the process of knowledge transfer"

Open Data for Research and Enhanced Publication



This forum is organized by GreyNet's Resource Policy Committee and the Institute of Information Science and Technologies (ISTI-CNR) in cooperation with Research Data Alliance, OpenAIRE, and DANS-KNAW, the Netherlands Data Archiving and Networked Services

Open Data for Research and Enhanced Publication



GreyForum Series 7.1

Workshop

Open Data for Research
and Enhanced Publication

Date

Friday, 25 May 2018

Time

10:30 – 16:30 hours

Location

National Research Council, CNR
Via Giuseppe Moruzzi 1
Room 27
56124 Pisa, Italy

Host and Organizers

Institute of Information Science
and Technologies (ISTI-CNR);
GreyNet's Resource Policy
Committee (RPC)

Participant Quota

The workshop can accommodate
80 participants

Online Registration

<http://greyforum.isti.cnr.it/RegistraGF>

Contact Information

GreyNet International
Javastraat 194-HS, 1095 CP
Amsterdam, Netherlands

info@grey.net.org
www.grey.net.org
Phone +31 (0) 20 331 2420

Program Outline

Welcome, Introductions

Research Data Alliance contribution to Open Science
Dr. Donatella Castelli, ISTI-CNR (Italy)

OpenAIRE: The e-infrastructure to collect and enhance OA
publications and data

Dr. Paola Gargiulo, OpenAIRE (Italy)

DANS Online Archiving and Core Trust Seal Certification

Dr. Henk Harmsen, Data Archiving and Networked Services (NL)

Data Papers – Grey Literature that implements FAIR Principles

Dr. Dominic Farace, GreyNet International

The data forum is geared to librarians, archivists, researchers, and other information professionals in government, academics, and business involved in the creation and/or management of research data. Emphasis lies in raising awareness to data access and its potential reuse for science and society. Instructor presentations are delivered in a plenary setting.

10:30 Registration Desk Open	13:00 Lunch
10:45 Welcome, Introductions	14:00 DANS Online Archiving ...
11:00 Research Data Alliance ...	15:00 Data Papers ... FAIR
12:00 OpenAIRE e-infrastructure	16:00 Closing Round

About the GreyForum Series www.grey.net.org/greyforumseries.html

GreyNet

Author Information

Bartolini, Roberto
87

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
55

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 (NeMIS). 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

Deluca, Rosaria
55

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

Frosini, Luca
69

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
55

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
87

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. 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

Molino, Anna
55

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
87

Monica Monachini is a Senior Researcher at CNR-ILC. Field of expertise: computational linguistics, computational lexicography, semantics, lexical

Author Information *CONTINUED*

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

95

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

69

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). 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

87

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

Savić, Dobrica

77

Dr Dobrica Savic is Head of the Nuclear Information Section (NIS) of the IAEA. He holds a PhD degree from Middlesex University in London, an MPhil degree in Library and Information Science from Loughborough University, UK, an MA in International Relations from the University of Belgrade, Serbia, as well as a Graduate Diploma in Public Administration, Concordia University, Montreal, Canada. He has extensive experience in the management and operations of web, library, information and knowledge management, as well as records management and archives services across various United Nations Agencies. Email: d.savic@iaea.org

Černohlávková, Petra

81

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

Vyčítalová, Hana

81

Hana Vyčítalová 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

Notes for Contributors

Non-Exclusive Rights Agreement

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

Submission Requirements

Manuscripts should not exceed 15 double-spaced typed pages. The size of the page can be either A-4 or 8½x11 inches. Allow 4cm or 1½ inch from the top of each page. Provide the title, author(s) and affiliation(s) followed by your abstract, suggested keywords, and a brief biographical note.

A printout or PDF of the full text of your manuscript should be forwarded to the office of TextRelease. A corresponding MS Word file should either accompany the printed copy or be sent as an attachment by email. Both text and graphics are required in black and white.

REFERENCE GUIDELINES

General

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

Specific

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

Examples

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

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

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

Review Process

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

Journal Publication and Article Deposit

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

An International Journal on Grey Literature **'BREAKING NEW GROUND IN GREY LITERATURE'**

Grey Literature and Research Assessment Exercises: From the Current Criteria to the Open Science models	55
<i>Silvia Giannini, Rosaria Deluca, Anna Molino, and Stefania Biagioni (Italy)</i>	
A Facet-based Open and Extensible Resource Model for Research Data Infrastructures	69
<i>Luca Frosini and Pasquale Pagano (Italy)</i>	
Impact of Disruptive Technologies on Grey Literature Management	77
<i>Dobrica Savić (Austria)</i>	
Collecting Grey Literature – Institutional Repository versus National Aggregator	81
<i>Petra Černohlávková and Hana Vyčítalová (Czech Republic)</i>	
Providing Access to Grey Literature: The CLARIN Infrastructure	87
<i>Sara Goggi, Gabriella Pardelli, Irene Russo, Roberto Bartolini, and Monica Monachini (Italy)</i>	
Preserving and accessing content stored on USB-flash-drives: A TIB workflow	95
<i>Oleg Nekhayenko (Germany)</i>	

Colophon	50
Editor's Note	53
On The Newsfront	
GL20 Conference Announcement - Twentieth International Conference on Grey Literature "Research Data Fuels and Sustains Grey Literature" Loyola University, New Orleans, Louisiana, USA • December 3-4, 2018.....	102
GL20 Call for Posters - Twentieth International Conference on Grey Literature Loyola University, New Orleans, Louisiana, USA • December 3-4, 2018.....	103
Open Data Forum - Open Data for Research and Enhanced Publication GreyForum Series Center for National Research, Pisa, Italy • May 25, 2018.....	104
Advertisements	
GreyGuide Portal and Repositories "Sharing knowledge as early as possible"	52
ISTI-CNR, Institute of Information Science and Technologies.....	54
CVTISR, Slovak Centre of Scientific and Technical Information.....	94
Author Information	105
Notes for Contributors.....	107