https://doi.org/10.26069/greynet-2025-000.534
ISSN 1574-1796

An International Journal on Grey Literature

Spring 2017, TGJ Volume 13, Number 1

'Grey Literature Domain in Broad View'

1992 25 YEARS 2017
GreyNet
www.textrelease.com Grey Literature Network Service www.greynet.org

https://doi.org/10.26069/greynet-2025-000.534

The Grey Journal

An International Journal on Grey Literature

COLOPHON

Journal Editor:

Dr. Dominic Farace
GreyNet International,
Grey Literature Network Service
Netherlands
journal@greynet.org

Associate Editors:

Julia Gelfand
University of California, Irvine
UCI, 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

Gretta E. Siegel
Portland State University
PSU, United States

Kate Wittenberg
Project Director, Client and Partnership
Development at Ithaka, United States

Technical Editor:

Jerry Frantzen, TextRelease

About TGJ

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

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

CIP
The Grey Journal (TGJ) : An international journal on grey literature / Dominic Farace (Journal Editor); Jerry Frantzen (Technical Editor) ; GreyNet International, Grey Literature Network Service. - Amsterdam: TextRelease, Volume 13, Number 1, Spring 2017. – *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 New York Academy of Medicine (USA) are Corporate Authors and Associate Members of GreyNet International.*
ISSN 1574-1796 (Print)
ISSN 1574-180X (PDF)
ISSN 1574-9320 (CD-Rom)

Subscription Rates:
€240 institutional, including postage & handling

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

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

About *GreyNet*

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

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

© 2017 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
'GREY LITERATURE DOMAIN IN BROAD VIEW'

Altmetrics and Grey Literature: Perspectives and Challenges 5
Joachim Schöpfel and Hélène Prost (France)

Grey Literature citations in the age of Digital Repositories and Open Access 23
Silvia Giannini, Stefania Biagioni, Sara Goggi and Gabriella Pardelli (Italy)

International identification and 'white and grey literature' : Identities, retrieval, reuse and the certainty of knowledge while sharing and connecting information 32
Flavia Cancedda and Luisa De Biagi (Italy)

Developing, linking, and providing access to supplemental genetics dataset vcf files 37
Plato L. Smith II and Lauren McIntyre (United States)

A terminological "journey" in the Grey Literature domain 41
Roberto Bartolini, Gabriella Pardelli, Sara Goggi, Silvia Giannini, and Stefania Biagioni (Italy)

Managing Diversity in the International Nuclear Information System 55
Dobrica Savić (Austria)

Colophon 2

Editor's Note 4

On The Newsfront

GL19 Announcement – Nineteenth International Conference on Grey Literature, October 23-24, 2017 in Rome, Italy 63

GL19 Call for Papers – 'Public Awareness and Access to Grey Literature' Submissions close on April 15, 2017 64

Advertisement

INIS-IAEA, International Nuclear Information System 54

Author Information 65

Notes for Contributors 67

https://doi.org/10.26069/greynet-2025-000.534

EDITOR'S NOTE

GreyNet Celebrates 25 Years of Research Support

GreyNet, an international grey literature network service, celebrates its 25th anniversary this year of providing research and reference support in the fields of academia, government, business and industry.

Based in Amsterdam, the company was founded in 1992 as an independent program and conference bureau specializing in the field of grey literature, which includes papers, abstracts, government reports, PowerPoint presentations, biographical notes, and a host of other document types in electronic and print formats not contained in traditional journal publishing.

Over the past 25 years, GreyNet has provided a valuable vehicle for a worldwide network of literature produced by various government agencies, universities and other organizations. GreyNet is fulfilling its mission to research and education, publication and public access to information and data important for libraries, research communities, and net citizens.

In addition to providing workshops and referral services, GreyNet organizes an annual international conference on grey literature held in various locations hosted by leading organizations in the United States, Italy, Belgium, Luxembourg, Czech Republic, Slovakia, France, Luxembourg, and the United Kingdom. GreyNet follows each conference with a publication of the complete proceedings as well as "The Grey Journal" now in its 13th volume. GreyNet likewise publishes a quarterly newsletter as a means of keeping its members and the general public informed of developments linked to its information resources and research output.

While primarily serving institutions in Europe, North America, and a number of countries in the Asia Pacific, representatives from recognized organizations worldwide have over the past 25 years been involved in significant ways.

GreyNet currently includes 11 organizational members, affiliate contacts in 40 nations, over 350 contributing researches and authors with 280 organizational contacts worldwide. GreyNet's newsletter currently boasts one thousand recipients.

The past 25 years have seen significant growth in GreyNet, but there exists additional roles in training and education as well as opportunities for greater cooperation with the open access community. Likewise, there are untapped possibilities for real cooperation with commercial publishers and licensing agents as well as greater use of available social media for public awareness to the wealth of knowledge produced by scientific and technical communities alongside that of commercial publishing.

For additional information on GreyNet International and its information products and services, visit www.greynet.org or contact via email info@greynet.org.

Dominic Farace
journal editor

https://doi.org/10.26069/greynet-2025-000.534

Altmetrics and Grey Literature: Perspectives and Challenges*

Joachim Schöpfel and Hélène Prost (France)

Abstract

Traditional metrics largely overlook grey literature. The new altmetrics introduced in 2010 as "new, online scholarly tools (that allow) to make new filters" (Altmetrics Manifesto), can include all kinds of scholarly output which makes them interesting for grey literature. The topic of our paper is the connection between altmetrics and grey literature. Do altmetrics offer new opportunities for the development and impact of grey literature? In particular, the paper explores how altmetrics could add value to grey literature, in particular how reference managers, repositories, academic search engines and social networks can produce altmetrics of dissertations, reports, conference papers etc. We explore, too, how new altmetric tools incorporate grey literature as source for impact assessment, and if they do. The discussion analyses the potential but also the limits of the actual application of altmetrics to grey literatures and highlights the importance of unique identifiers, above all the DOI. For the moment, grey literature missed the opportunity to get on board of the new movement. However, getting grey literature into the heart of the coming mainstream adoption of altmetrics is not only essential for the future of grey literature in open science but also for academic and institutional control of research output and societal impact. This can be a special mission for academic librarians.

Introduction

Traditional metrics largely overlook grey literature. Worse, they basically disregard grey literature as irrelevant for the evaluation of research. Established metrics for individuals and organisations are journal-centric. Measuring the performance and popularity of scientists or research structures means counting the number of articles citing other articles, resulting in journal impact factors, normalized citation rates and the h-index. Even those rare studies including conference papers are limited to published proceedings[1]. Grey literature remains out of scope. The most important reason is the way these metrics are produced – they rely on bibliographic tools like the Web of Sciences (WoS) and Scopus which from the beginning on were (nearly) exclusively journal and monograph A&I services, dismissing other vectors of scientific communication outside of the academic publishing market[2].

The emergence of webometrics, i.e. the "study of the quantitative aspects of the construction and use of information resources, structures and technologies on the web drawing on bibliometric and informetric approaches" (Björneborn & Ingwersen 2004, p. 1217), change the situation. As many scholarly activities today are web-based, the field of webometrics is partially covered by scientometrics (figure 1). These new or alternative metrics are not limited to journals but apply to academic content (scholarly work) at large, insofar and as long as this content is available on the web, in particular on the social web (Galligan & Dyas-Correia 2013). They are sometimes called scholarly metrics or social media metrics, and most often defined as altmetrics.

The fact that these new metrics can include all kinds of scholarly output makes them interesting for grey literature. In a draft on altmetrics definitions and use cases, the National Information Standards Organization describes scholarly output as "a product created or executed by scholars and investigators in the course of their academic and/or research efforts. Scholarly output may include but is not limited to journal articles, conference proceedings, books and book chapters, reports, theses and dissertations, edited volumes, working papers, scholarly editions, oral presentations, performances, artifacts, exhibitions, online events, software and multimedia, composition, designs, online publications, and other forms of intellectual property" (NISO 2016, p.9). One part of this output clearly belongs to grey literature, especially when citable and accessible[3].

* First published in the GL18 Conference Proceedings, February 2017

[1] See for instance Ingwersen et al. 2014, also for similar, older studies

[2] The methodological problems to identify theses in bibliographic databases in Larivière et al. (2008) confirm the situation

[3] See the definition of "acceptable products" by the National Science Foundation, *Grant Proposal Guide II-12 NSF 14-1*, November 2013 http://www.nsf.gov/pubs/policydocs/pappguide/nsf14001/gpgprint.pdf

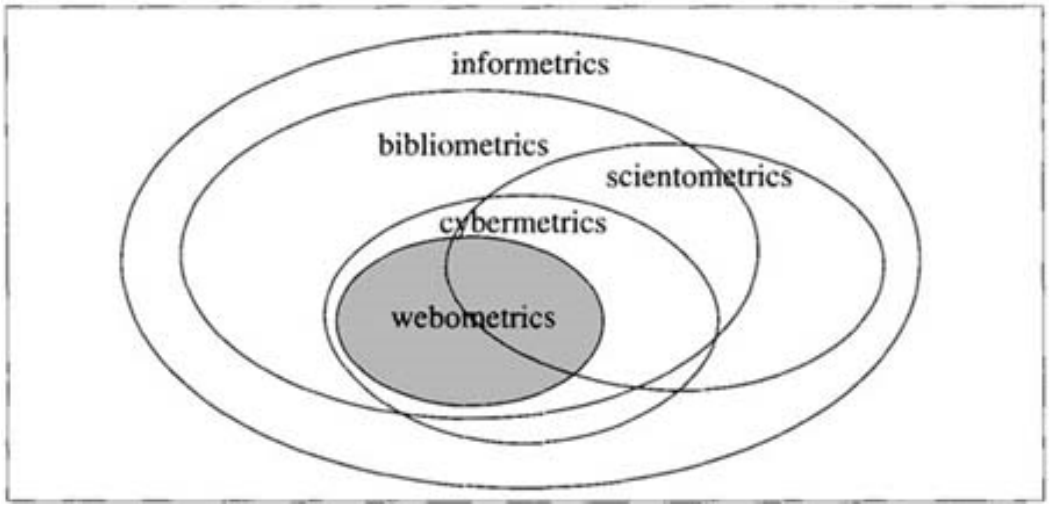


Figure 1: Webometrics in the field of library and information sciences (source: Björneborn & Ingwersen 2004)

The topic of our paper is the connection between altmetrics and conference proceedings, reports, theses and dissertations, and working papers. Do altmetrics offer new opportunities for the development and impact of grey literature? Are there already examples of good practice? Are there any barriers? However, before we outline the potential of altmetrics for grey literature, we will provide some elements for a better understanding of this concept.

A short history of altmetrics

Altmetrics have a short history[4]. The term was introduced by Jason Priem from Chapel Hill in 2010, in a tweet published on the 29th September 2010: "I like the term #articlelevelmetrics, but it fails to imply *diversity* of measures. Lately, I'm liking #altmetrics"[5]. This came after the global success of the web 2.0 tools and media, such as Facebook, Twitter etc., and it became popular as a kind of marketing umbrella for a broad range of new metrics of scholarly impact on the social web (Priem & Hemminger 2010).

The Altmetrics Manifesto[6] from 26 October 2010 merges article-level metrics and distributed scientific evaluation with social media into research on altmetrics and defines them as fast and open filters to relevant and significant scholarly sources, not in continuity but in disruption with webometrics or citations; "given the crisis facing existing filters and the rapid evolution of scholarly communication, the speed, richness, and breadth of altmetrics make them worth investing in" (Priem et al. 2010).

From that moment on, the interest for altmetrics increased steadily to join and finally exceed scientometrics, according to Google Trends (figure 2). Two years after the Manifesto, the San Francisco Declaration on Research Assessment (DORA), initiated by the American Society for Cell Biology (ASCB), recognizes the need to improve the ways in which the outputs of scientific research are evaluated and suggests the "use of a range of article metrics and indicators on personal/supporting statements, as evidence of the impact of individual published articles and other research" (DORA 2012). Signed by nearly 12,500 individuals and 800+ organizations[7], DORA fostered the awareness for altmetrics and became a reference for the debate, research and development in the field.

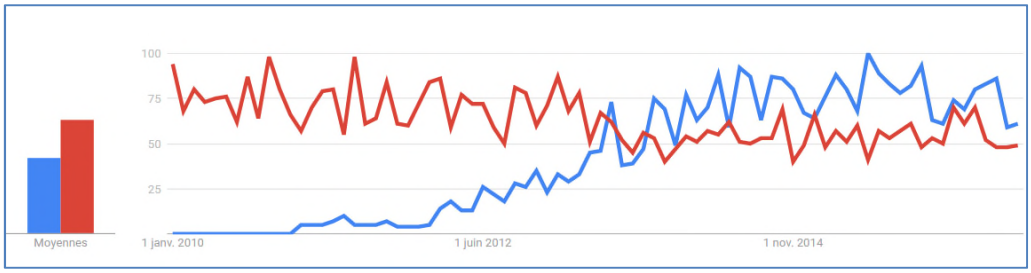


Figure 2: Altmetrics (blue) on Google Trends, compared to scientometrics (red) (2010-2016)[8]

[4] See comprehensive reviews by Erdt et al. (2016) and Sugimoto et al. (2016)
[5] https://twitter.com/jasonpriem/status/25844968813 by @jasonpriem
[6] http://altmetrics.org/manifesto/
[7] http://www.ascb.org/dora/ accessed 7 September 2016
[8] Data source: Google Trends www.google.com/trends accessed Sept 3, 2016

The increasing number of scholarly work dedicated to altmetrics reveals the same trend (figure 3). No study on altmetrics before 2010, and then a steadily growth from 8 references in 2010 to 122 in 2015.

The Google Scholar statistics confirm the Google Trend figures – the interest for scientometrics remains relatively stable, with 50-70 publications per year, but is exceeded by works on altmetrics from 2013 on.

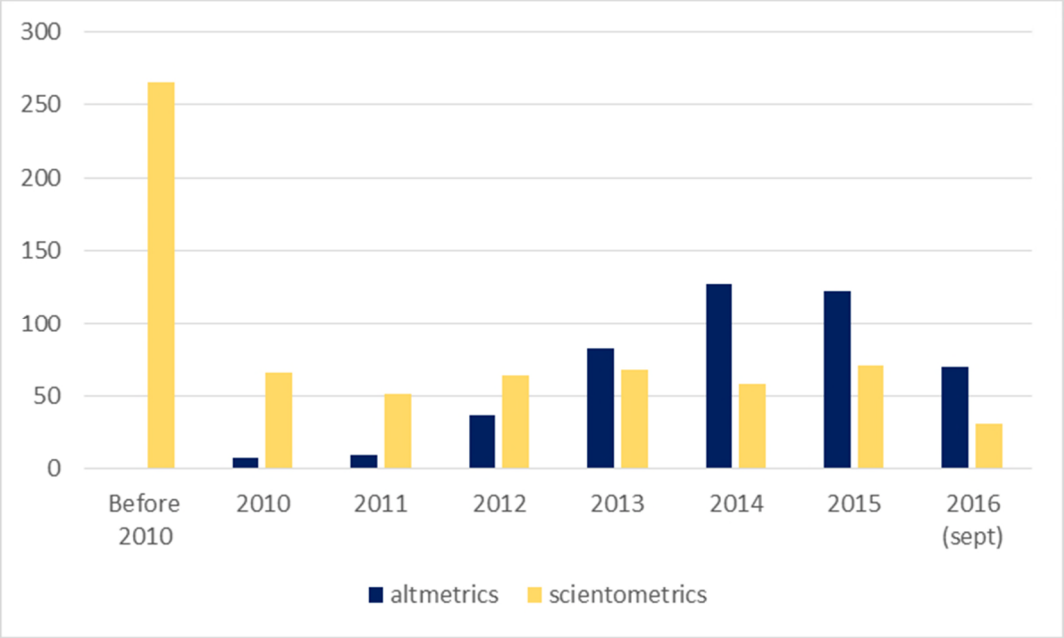


Figure 3: Publications on altmetrics and scientometrics[9]

Basically, altmetrics are "social web metrics for academic publications" (Sud & Thelwall 2014, p.1131) and particularly interesting for measuring societal impact, beyond the academic community (Piwowar 2013), through the count of views, downloads, clicks, likes, tags, posts (blogging) and tweets (micro-blogging), shares, discussions etc. The term "usually describes metrics that are alternative to the established citation counts and usage stats—and/or metrics about alternative research outputs, as opposed to journal articles" (NISO 2014, p.4).

Variety is one main feature of altmetrics, a class of indicators measuring attention, dissemination and influence[10], even if the distinction between attention, dissemination and influence is not self-evident. The main areas of altmetrics are shown in figure 4. Impact on the (social) web can be assessed through the count of PDF or HTML downloads (viewed), the creation of references in online reference managers like CiteULike, Zotero or Mendeley (saved), the number of posts in blogs and micro-blogs, on Facebook or Wikipedia (discussed), the number of mentions in editorials or tools like F1000 (recommended) or as usual, simply via the number of citations in the WoS, Scopus, PubMed Central or CrossRef (cited).

The NISO Alternative Assessment Metrics Initiative (2016) defines altmetrics as a broad concept that includes "multiple forms of assessment that are derived from activity and engagement among diverse stakeholders and scholarly outputs in the research ecosystem".

Today, a clear, common, widely accepted definition is not in sight. Altmetrics comprise many different types of metrics in a constantly changing landscape and "refer to a heterogeneous subset of scholarly metrics and are a proper subset of informetrics, scientometrics and webometrics" (Haustein 2016, p.416). Perhaps a pragmatic approach like Altmetric's recent definition will fit best, for the moment: "Altmetrics are attention data from the social web that can help librarians understand which articles, journals, books, datasets, or other scholarly outputs are being discussed, shared, recommended, saved, or otherwise used online. They can be reported at the item-, journal-, or author-level"[11].

[9] Data source: Google Scholar https://scholar.google.fr allintitle: altmetrics (or scientometrics), accessed Sept 5, 2016
[10] See https://www.altmetric.com/about-altmetrics/what-are-altmetrics/
[11] https://www.altmetric.com/blog/altmetrics-collection-development/

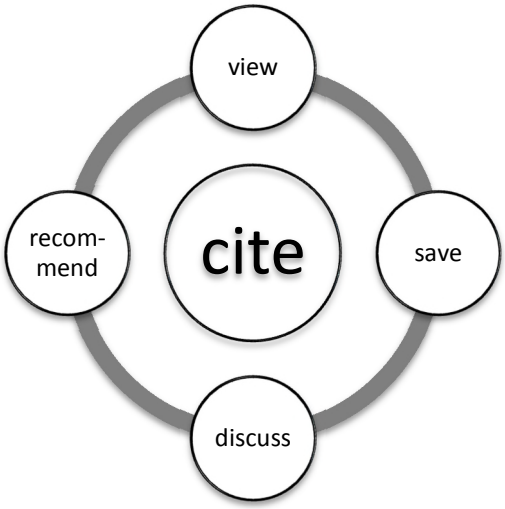


Figure 4: Altmetrics areas of assessment

Six years after the Manifesto, however, it is not quite clear if altmetrics are "an alternative or enhancement to the use of journal impact factors and click-through rate analysis to measure the impact and value of scholarly work" (Galligan & Dyas-Correia 2013, p.56). But they are already relevant for research evaluation. The European Commission DG Research and Innovation has established an Expert Group on Altmetrics which describes the emergence of altmetrics as part of the "transition to a more accountable and transparent research system"[12], more efficient, open to society, and expects "robust, responsible, transparent and interoperable uses of metrics and altmetrics in open science". Altmetrics are levers in support of open science. Up to now, including altmetrics in decisions on grants, hiring and tenure still requires careful consideration but they may soon become a normal part of a CV (Kwok 2013).

What does this mean for grey literature? What is the potential of altmetrics for grey literature? The next section tries to provide a global answer.

The potential

Bornmann (2014) mentions four benefits of altmetrics compared to traditional metrics: they measure impact beyond science, they can include scholarly products other than papers (articles), they allow impact to be measured shortly after the output, and as a rule, it is easy to obtain altmetric data (figure 5).

Compared to traditional, citation-based metrics, altmetrics endorse two different developments: "Widening the definition of research outputs to include more than just books and journal articles, and looking beyond citations for a quantitative way of assessing or discovering them" (Adie 2016, p.67). Thus, at least in theory, altmetrics are not limited to a coverage similar to the WoS or Scopus. As stated by Andy Tattersall, "altmetrics focuses on research artefact level metrics that are not exclusive to traditional papers but also extend to book chapters, posters and data sets among other items" (2016, p.1). "Among other items" – this could or should bring in non-traditional, non-commercial items, like working papers, dissertations, conference papers, reports etc.

[12] Next-generation altmetrics: responsible metrics and evaluation for open science, available at https://ec.europa.eu/research/openscience/index.cfm?pg=altmetrics_eg

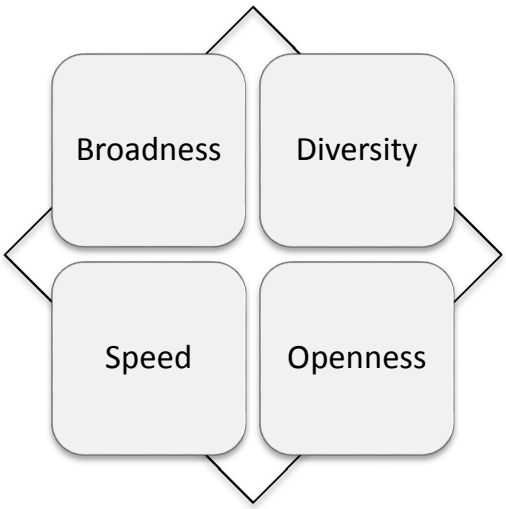


Figure 5: Benefits of altmetrics (source: Bornmann 2014)

And even if altmetrics still focus primarily on practices relating to research articles as "central research output that informs research assessment (they) can and should be extended by recognizing additional products, such as datasets (…)" (DORA 2012). Therefore, their potential for grey literature is twofold (figure 6):

- Diversity: Impact assessment on article level such as download counts also applies to grey literature. "Altmetrics (…) allow for evaluation of a greater diversity of products, i.e., not just publications (…). These products might be datasets, software, copyrights, algorithms, grey literature, and slides (...). Altmetrics now offer the opportunity to determine the impact of these products both in science (…) and beyond science" (Bornmann 2014, p.898). Diversity, as said above, is considered as one crucial advantage of altmetrics, and this includes grey literature.

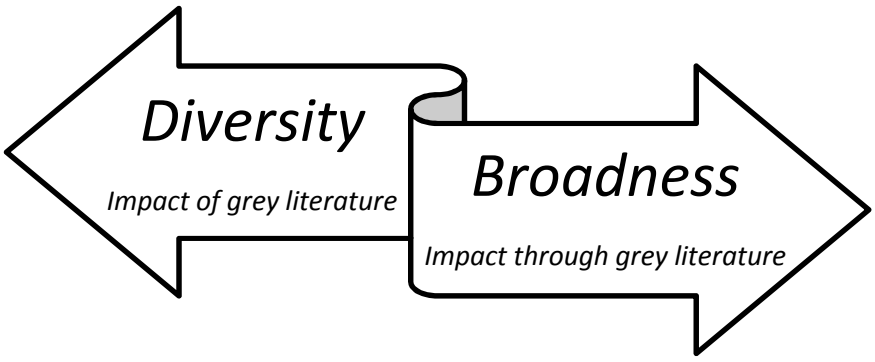


Figure 6: Double potential of altmetrics for grey literature

- Broadness: In contrast to traditional metrics which usually exclude "documents such as technical reports or professional papers which some label as 'grey literature' (…) due to lack of indexing" (Moed & Halevi 2014), altmetrics are not limited to scientometric databases; also their impact assessment based on citations, links and referrals can take account of a broader range of scientific information, including citations in dissertations, reports, white papers etc. Thus, grey literature can serve as material to measure impact of scientific output.

Beyond these two major perspectives for grey literature, recent studies on altmetrics mention two other potential benefits:

1. Dissemination of reports: Altmetrics may be a way to foster more efficient information practices by research organizations and funding bodies (foundations); one obstacle to disseminate reports etc. via open and shared systems is that often "organizations aren't sure people are even reading this stuff (…) Altmetrics (…) hopefully can better inform our expectations and measures of readership" (Brooks & Fitz 2015, p.43; see also Dinsmore et al. 2014);

2. Scientific information in developing countries: Neylon et al. (2014) insist on the application of altmetrics, especially social media usage metrics, for grey literature in "a developing country context, such as in sub-Saharan Africa (where) the importance of 'grey literature' – policy briefs, working papers, media articles and other scholarship aimed at lay audiences – is massive, satisfying both the need for social engagement as well as scholars' professional expectations" (p.2).

In the following, we will address the first two issues, diversity ("altmetrics for grey literature") and broadness ("altmetrics through grey literature"), with some examples and a focus on special conditions and prerequisites.

Altmetrics for grey literature

Our first issue is about impact assessment of grey literature. As said above, traditional metrics have largely overlooked grey literature. Altmetrics can offer new and unique opportunities for the web-based impact measurement of reports, conferences, dissertations etc. But do they really? And if so, how? To what extent?

In 2012, a study funded by the Dutch SURF-Foundation assessed fifteen "novel impact monitors", such as reference managers, academic search engines and new altmetrics tools (Wouters & Costas 2012). At least nine out of the fifteen "monitors" can produce impact data for grey literature[13]. These seem rather favourable and promising conditions. Let's get some empirical insight for a better understanding.

Repositories

In the GreyNet community, repositories, especially institutional repositories, are generally considered as "natural home" for grey literature, as institution-based platforms for the dissemination and preservation of the institutional scientific output (Banks & de Blaaij 2006). Most of the open repositories contain one or more categories of grey literature, often theses and dissertations (particularly in university repositories), but also conference and workshop papers, unpublished reports, working papers or other "special items"[14].

All repository servers produce log files of views and downloads which can be transformed into statistics and metrics, useful as well for institutions and hosting organisations as for authors and readers. However, a couple of years ago only few repositories made these metrics freely available on their website, along with the metadata and deposited files, and even less did so in a standard, interoperable way (Schöpfel & Prost 2009, Prost et al. 2010).

The debate on new metrics accelerated the movement, and following the Altmetrics Manifesto and the DORA Declaration, repository hosts and managers started to improve the availability of web analytics and to implement new altmetrics tools. As a result, today "institutional repositories are (…) embracing altmetrics as a means of both tracking and encouraging engagement with the resources, and the ability to track and measure engagement with grey literature can be a good source of evidence of the role these outputs play in the research and publication life-cycle" (Priego 2014). Four very different examples may show the potential but also the limits of this development.

HAL[15]

The French national repository HAL contains 400,000+ documents and 1,1m records. Nearly half of the full text deposits (46%) are grey literature, with nearly 60,000 dissertations and more than 80,000 conference papers. For all these documents, HAL produces usage statistics on the item level, of full text downloads and retrievals (views) of the records (metadata). Since the launch of HAL in 2001, authors as well as collection managers have access to detailed and customizable usage statistics for each item or, cumulated, for a collection, an institution, an author etc.

On the public interface, HAL displays for each record two metrics, cumulated metadata views and full text downloads. In our example (figure 7), HAL shows that the Lille 3 White Paper on research data in PhD theses received so far 3,591 record views and 2,150 successful download requests[16]. But HAL does not offer comparative metrics (average statistics per document type and/or domain etc.).

Since 2015, HAL displays an Altmetric badge with metrics based on the unique identifiers DOI, arXiv-id and PubMed ID. So far, HAL does not allocate DOIs to deposits without identifiers and does not use its own identifier HAL Id or other identifiers like the French national dissertation number (NNT) for the

[13] GS, MAS, ArnetMiner, Mendeley, CiteULike, Zotero, ReaderMeter, ImpactStory, SURE2

[14] See the Directory of Open Access Repositories OpenDOAR, available at http://www.opendoar.org/

[15] HAL = Hyper articles en ligne https://hal.archives-ouvertes.fr/

[16] Accessed 9 September 2016

assessment of altmetrics. Thus, the only conference papers with Altmetric badges we could find in HAL are those published by Springer, IEEE or other commercial publishers specialised in proceedings and members of CrossRef. Probably, this means that while all grey literature in HAL is displayed with usage statistics, no grey item has received an Altmetric badge up to now.

Figure 7: Display of usage statistics in a HAL record

figshare[17]

The online digital repository figshare where researchers can preserve and share their research outputs contains above all figures, datasets and filesets but also some papers, dissertations, posters and presentations. Figshare has "three basic functions: it acts (1) as a personal repository for yet unpublished materials, (2) as a platform for newly published research materials, and (3) as an archive for PLOS" (Kraker et al. 2015). In fact, almost 90% of the input comes from PLOS – mostly figures, while text files represent less than 2% of all entries, and the part of dissertations (all kinds of short or long unpublished written texts), posters and presentations is extremely low (0.3%).

Figshare exhibits view and download counts for all deposits. In April 2016 figshare implemented Altmetric badges to showcase attention surrounding research output (figure 8).

Figure 8: USAGE statistics and altmetrics in figshare

The reader can click through to detailed impact information on the Altmetric server. All figshare entries have a DOI; 89% of the DOIs are provided by PLOS, the other 11% are allocated by figshare (with DataCite) as soon as a user makes an uploaded material publicly available. This means that dissertations may get a DataCite DOI[18]. The systematic allocation of DataCite DOIs facilitates the generation of attentions scores with the Altmetric tool. However, obviously not all figshare entries have an Altmetric badge.

[17] https://figshare.com/

[18] See the following example, a two-page dissertation: Harper, Danny (2016): Plagiarism in college essays and assignments.docx. figshare. https://dx.doi.org/10.6084/m9.figshare.3528812.v1 Retrieved: 10 47, Sep 09, 2016 (GMT)

IRUS-UK[19]

IRUS-UK is a Jisc-funded national aggregation service which collects raw usage data from 113 institutional repositories and transforms them into COUNTER-compliant statistics. Insofar as these repositories contain unpublished grey items, IRUS-UK aggregates usage statistics from more than 200,000 conference papers, reports, dissertations and working papers which represent 34% of the repositories' content (figure 9). All these deposits received nearly 30m successful download requests, or 43% of all aggregated IRUS-UK downloads. These statistics are interesting for three reasons.

- Often grey literature usage statistics are not standardized. Here, as the IRUS-UK statistics are COUNTER-compliant, the data are comparable, authoritative, and standard-based.
- For this reason, they can be compared to download figures from other document types, in particular with article statistics. This direct comparison reveals for instance that in early September 2016, the average downloads of journal articles are similar to conference items, proceedings or reports, but two times lower than the usage statistics of dissertations and working papers. Taken together, the average usage of grey literature is one third higher than for articles or books.
- The aggregated usage statistics should allow for further standardization, e.g. for document- and/or domain-specific average download figures that could be used as a kind of reference set for individual items, like the PLOS metrics.

Document type	Total number	Total downloads
Conference or Workshop Item - Other	50 892	4 536 836
Conference Papers /Posters	7 823	389 402
Conference Proceedings	5 256	252 382
Report	14 265	1 704 870
Thesis or dissertation	123 014	21 409 166
Working Paper	5 915	795 653

Figure 9: Grey literature in IRUS-UK[20]

CORE[21]

For five years now, the CORE project aggregated and enriched content from nearly 1 000 repositories from all over the world, in order to increase the discoverability and reusability of open access papers (Pontika et al. 2016). As CORE harvests not only the repository metadata but also the full-text and caches this PDF version in its own database, it can provide IRUS usage statistics on full-text downloads. Among the 37m harvested items (called "articles" or "manuscripts"), CORE also contains reports, conference papers and other unpublished documents. However, the CORE portal does not allow for document-type specific browsing or search, a fact which reduces its interest for our purpose. -

To sum up, these four examples confirm the potential of repositories for the production of altmetrics, on a continuum from usage statistics (views, downloads) to impact measures based on social media and the possibility to display standard-based data and reference sets. The limits or pre-requisites are the need for rich metadata, including the document type, and the allocation of an established unique identifier.

Social networks

Much has been said about academic social networks like ResearchGate and Academia, about their functionalities, their uptake by the research communities[22] and their impact on scientific communication[23], their competitive strategy challenging above all institutional repositories, and their business model. Because of the increasing number of users, records, documents and other material shared via these networks, they are part of these "novel impact monitors" mentioned above. At least three different aspects are relevant for altmetrics: the creation of metadata, the deposit of the document and the mention of a document in a debate or in an answer to a question.

[19] Institutional Repository Usage Statistics UK http://www.irus.mimas.ac.uk/

[20] All IRUS-UK figures and statistics accessed 7 September 2016

[21] COnnecting REpositories https://core.ac.uk/

[22] 41m accounts in Academia, 10m accounts in ResearchGate

[23] 14.7m papers in Academia, 100m papers and other items in ResearchGate (24% papers with full-text), mostly STM

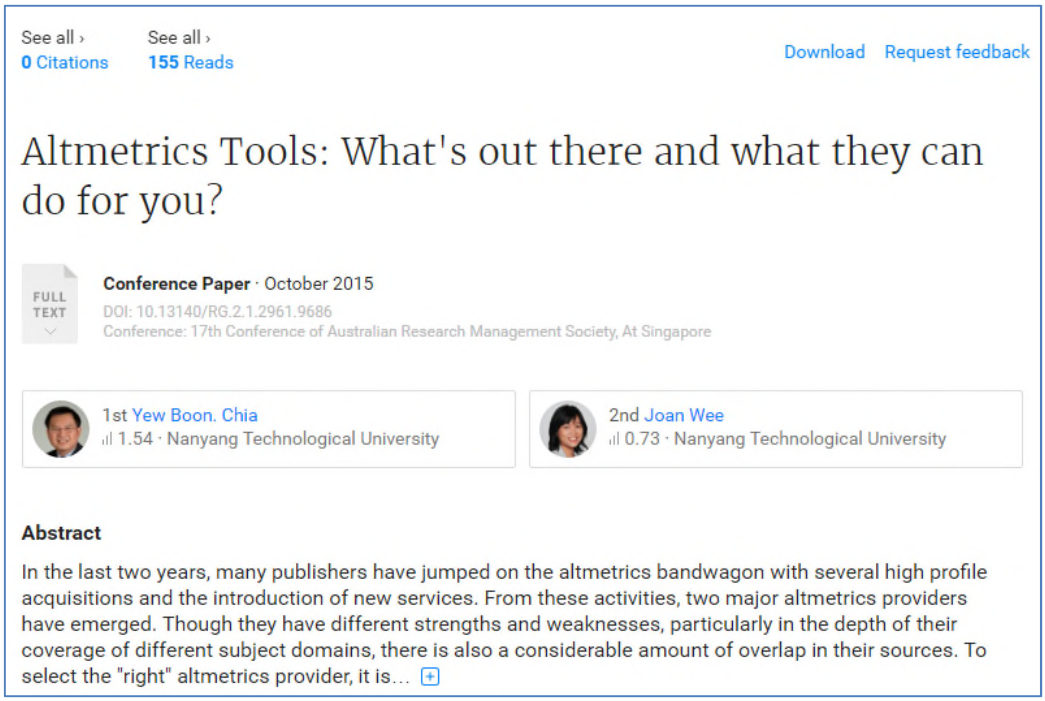


Figure 10: A shared conference paper with DOI (ResearchGate)

Basically, academic social networks invite researchers to share their results, without imposing limits or specific items, i.e. all major types of grey literature can be deposited in social networks. ResearchGate for instance suggests 18 types of "publications", including conference papers, posters, presentations, technical reports, theses and working papers; but also unpublished articles (preprints) and working copies, datasets, negative results and raw data. Also, if necessary, a new format (category) can be created for a specific deposit. Clearly they have become large reservoirs for all kinds of unpublished, grey literature, with the potential to make them available for impact measurement. However, this potential is conditioned by the quality of the metadata, in particular of an identifier. The social networks do not allocate unique identifiers but invite to add or import existing DOIs to the deposit (Figure *10).*

Recently, a study was conducted on the effectiveness of six ResearchGate metrics on the author level (ResearchGate score, impact points, number of downloads, number of publication views, number of citations, and number of profile views), concluding that in a small sample and a specific field "the ResearchGate score can be an effective indicator for measuring an individual researcher's performance" (Yu et al. 2016, p.1005).

Beyond academic networks, scientists share and discuss results also on Facebook and Google+; yet, on the one hand these networks are not designed for documentary metadata; on the other hand, the coverage of scientific documents still seems low, producing unreliable metrics (Haustein 2015).

Reference managers

Reference managers like CiteULike, Zotero and Mendeley can provide relevant information for altmetrics, in particular about the number of copies of a given reference. Mendeley for instance is a large database of "white papers, conference proceedings, book and journal references, and other kinds of grey literature that is searchable by other Mendeley users (…)" (Tattersall 2016, p.114). Mendeley provides how many users have a copy for each item.

Figure 11: Number of bookmarks in CiteULike

CiteULike is said to contain 8.3m references and proposes 17 item types, including conference papers, technical reports, Master's and PhD theses, unpublished work and "miscellaneous". There is no available reliable data on the actual number of references for each of these categories. Like Mendeley, CiteULike inform about the number of copies (bookmarks) for each reference (figure 11). CiteULike allocates its own identifier and supports DOI and Pubmed ID, for importing, creating (generating) and searching references.

Bookmarks can be used as a complement to citation metrics. Traditional citation-based indicators, in particular the journal impact factor and author mean citation per paper, are correlated with bookmark-based indicators (altmetrics), such as journal mean bookmarks per paper, the percentage of bookmarked articles in the journal and author mean bookmarks per paper; an analysis of data from the WoS and CiteULike reports the correlations slightly higher for journals than for authors (Sotudeh et al. 2015). Zoller et al. (2016) conclude that a bookmarking system's most inherent feature – tagging – is suitable for identifying topic subsets of publications where usage and future citations exhibit higher correlations. Yet, apparently no data has been published on bookmark-based metrics of grey literature.

Academic search engines

The SURF-study on altmetrics tools mentions the academic search engines Google Scholar and Microsoft Academic Search because they include "cited by" data for some items, whenever they can identify citations (figure 12). This data can be analysed and interpreted as an indicator for impact on the web.

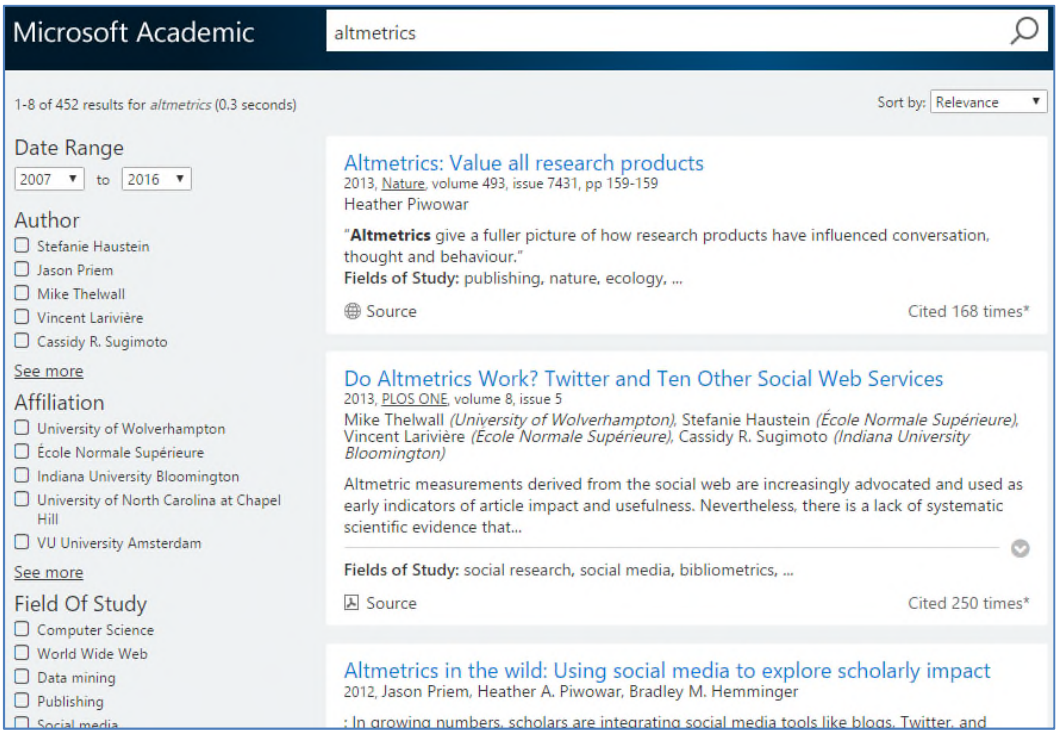


Figure 12: "Cited by" data in Microsoft Academic Search

As these search engines cover a large part of the academic web, in particular institutional repositories and other non-commercial platforms, their crawling and indexing include preprints, dissertations, reports, conference papers etc. For example, figure 13 shows citation data for a workshop paper available on figshare and not published elsewhere, without an allocated DOI.

Obviously, the academic search engine are able to produce impact data for all kinds of scientific papers, as long as they are made available on referenced and indexed platforms, in particular institutional and other repositories. Unique identifiers like the DOI are not indispensable but may improve the reliability of the search results.

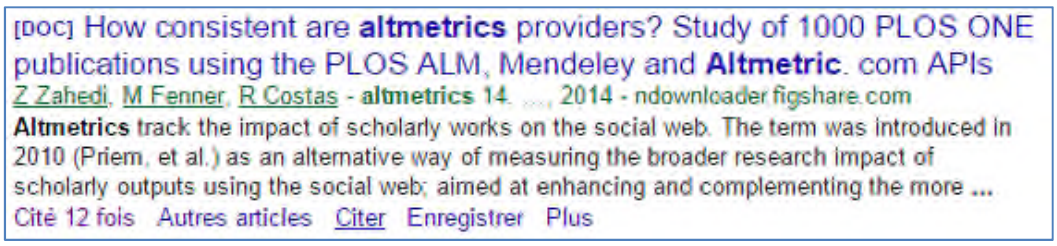


Figure 13: "Cited by" data for a workshop paper on figshare, in Google Scholar

New altmetrics tools

Following Kraker et al. (2015), the most important data provider for altmetrics are not reference managers, academic social networks or search engines but Twitter: "In the altmetrics analysis, we found that Twitter was the social media service where research data gained most attention". A growing number of new tools and platforms aggregate these online events (tweets, likes, comments, downloads etc.) as well as derived metrics from repositories, reference managers etc. (NISO 2016).

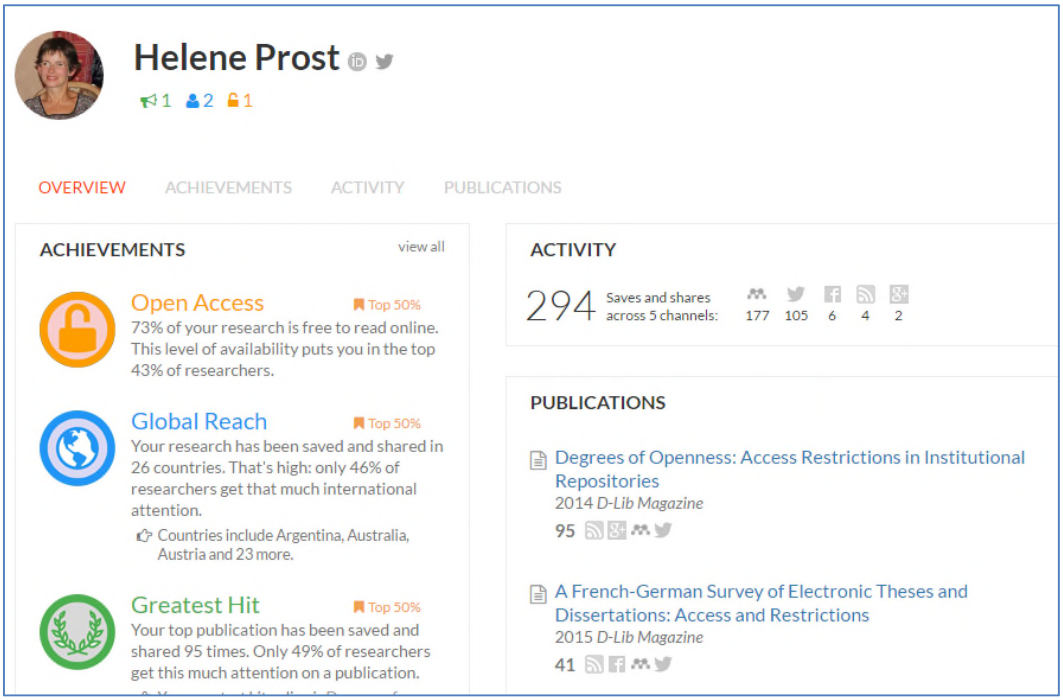


Figure 14: Author's page on ImpactStory

Among these data aggregators one can find Altmetric.com, PlumX, ImpactStory etc. (figure 14). They do not measure the same aspects, and they do not generate the same metrics. While PlumX from PlumAnalytics fits more with libraries' and institutions' needs, especially for repositories (Lindsay 2016), ImpactStory is aimed at individual researchers, and Altmetric offers services for individual researchers, institutions and funders but/and above all for commercial publishers (Konkiel 2012). PlumX detects considerably more items in social media and also finds higher altmetric scores than ImpactStory; but comparison of altmetrics tools is difficult due to differences in assignments to categories, which result in different counts (Kraker et al. 2015).

Basically, these tools can produce social impact metrics for working and conference papers, dissertations and other grey items. However, only few studies have been published on aggregated altmetrics data incorporating grey literature. Altmetric does not track non-traditional outputs. Wilkinson et al. (2014) made use of the Web Impact Report (WIRe) as a novel solution to assess the impact of organisational reports, especially when in open access. WIRe consists of a "range of web-derived statistics about the frequency and geographic location of online mentions of an organisation's reports (…)" (p. 797), such as online citations, site domain and genre of the citing site (blog, governmental sources etc.). Nevertheless, this case study with a small corpus of 20 reports reveals two major issues, i.e. a relatively high percentage of incorrect matches and a time-consuming human workload (content analysis).

ArnetMiner[24] aims to provide comprehensive search and mining services for academic social networks, with a special focus on 6,000+ conferences, mostly in computer sciences, and with a ranking based on the H5-Index, top-cited authors and papers, and data on the social network and semantics for each conference (figure 15).

[24] http://www.aminer.org

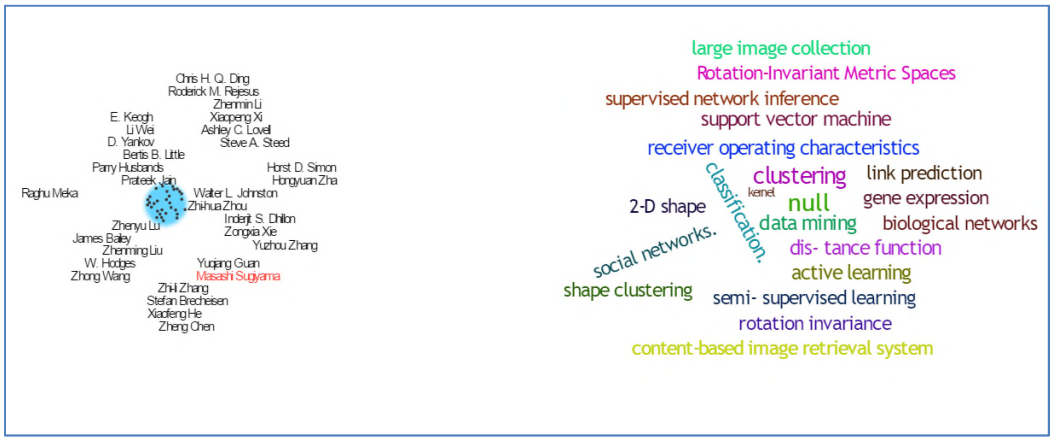


Figure 15: Conference analysis of SIAM International Conference on Data Mining on ArnetMiner

Literature about altmetrics mentions other tools like PaperCritic, PeerEvaluation or ReaderMeter. Some are operational, others are not; some are part of PLOS, Mendeley etc. Basically they are all open for grey and other items, primary data etc. But we could not find any reliable information about their real value and interest for grey literature, in terms of specificity and impact.

Altmetrics through grey literature

"Broadness", the possibility to take account of a broader range of scientific information is one of the four major benefits of altmetrics. They could be "more diverse in kinds of data and accordingly numbers of data sources (whereas for traditional citations only the cited references in journals serve as data source)" (Bornmann 2014, p.898), thus revealing more diverse and nuanced forms of impact than traditional indicators. So, how can grey literature contribute to altmetrics? Do altmetrics tools make use of grey literature? For Euan Adie, CEO of Altmetric, grey literature[25] "presents great opportunities for alternative metrics, providing data and indicators that cannot be found anywhere else" but also drew attention to the specific characteristics and challenges, e.g. missing identifiers, no "canonical metadata", lack of long term preservation and availability (Adie 2014). Five examples may illustrate potential benefits and limits of this "broadness".

1. Web-based grey literature can serve as source to increase impact of other grey literature. Wilkinson et al. (2014) conducted their study on WIRe with a small sample of 20 research reports. Their results showed that most of them (17) had been cited by other reports, conference papers, white papers, MA and PhD theses and speeches and/or dissertations available on the web. But without standard or automated procedures, including grey literature involves a lot of human work.
2. In some fields, grey literature may be more relevant than in others. Working on subfields of sustainable energy research, Ingwersen et al. (2014) insist that such analyses "should include proceedings papers – because this document type does have significant (…) influence on the overall citation impact of a research field, in particular in proceedings-dominant fields" (p.1290). The same observation would probably apply to economics (working papers), physics (preprints) and computer science (conference papers).
3. New content mining tools improve the efficiency and broadness of data aggregators. Thus, Altmetric has developed a text-mining solution (*Altmetric Policy Miner*) to discover mentions of publications in policy documents on selected websites[26]. Due to this APM-software, Bornmann et al. (2016) were able to assess the societal impact of climate change publications mainly through grey literature from governmental agencies, international organizations and NGOs.
4. Human knowledge, manual searching and browsing are the price of "broadness" and inclusion of grey literature in alternative impact assessment. Urquhart & Dunn (2013) evaluated the impact of the National Minimum Dataset for Social Care (NMDS-SC). References to the dataset (citations) were identified in 175 separate publications, with 50% policy and practice reports, 35% media communications and only 15% academic journal articles. Google Scholar fits more with this kind of analysis than the Web of Science, because of a greater range of included material. Other relevant sources are field-specific databases, aggregators' and publishers' platforms[27]. The procedure of

[25] Described as "theses, posters, preprints, patents and policy documents and similar"

[26] Such as European Food Safety Authority (EFSA), GOV.UK–Policy papers, Research & Analysis, Intergovernmental Panel on Climate Change (IPCC), International Committee of the Red Cross (ICRC), World Health Organization (WHO), International Monetary Fund (IMF), Médicins sans Frontières (MSF), NICE Evidence, Oxfam Policy & Practice, UNESCO and World Bank

[27] In this study NHS Evidence, LG Search, Hein Law Online, EBSCO Business Source Complete, Nexis and Emerald Journals

such an extensive targeted grey literature search is rather complex, far from automated and quick processing of large amounts of bibliographic information: "Organisations considered likely to be publishing materials drawing on NMDS-SC (…) were identified from existing knowledge of the sector, by the client and ourselves, and the initial findings from the bibliometric survey. A total of 24 organisational websites were manually searched and browsed, including UK government departments, sectoral bodies, knowledge intermediary organisations such as independent research organisations (…), campaigning organisations, think tanks, trade/employer organisations and the professional and mainstream press. We also conducted a limited search of social media, using social media aggregator sites" (p.297).

5. Altmetrics with grey literature produce more content but are time-consuming. Sibbald et al. (2015) conducted a case study on the inclusion of grey literature in citation analysis, based on one published article in the field of violence against women. Google Scholar and the Web of Science produced eighty journal articles citing the paper. The grey literature searches[28] found 29 other sources (27% of all results). But "this method requires additional resources. The much broader range of potential search venues demands more time and expertise. Delving into gray literature is a challenging task and requires planning and coordination, including consideration-specific inclusion/exclusion searching. Unlike database searching, common nomenclatures rarely exist for searching diverse gray literature sources; therefore, the concept of consistency in search terms across sources is difficult to achieve."

6.

These examples confirm the potential of grey literature as source for altmetric impact assessment, with significant and complementary results based on citations, links and referrals from a broader range of scientific information, including dissertations, reports, white papers etc. But they also show that this approach is more complicated and time-consuming than the usual WoS or Scopus-based work. In contrast to traditional metrics which usually exclude grey literature, altmetrics are not limited to scientometric databases. But when it comes to larger empirical studies, this exploitation of grey literature remains an exception and is sometimes limited to science blogs, while proceedings, dissertations etc. are dismissed (see for instance Thelwall et al. 2013 or Costas et al. 2015).

Discussion

Our objective was to clarify the connection between altmetrics and grey literature. Traditional metrics have largely overlooked grey literature. Do altmetrics offer new opportunities for the development and impact of grey literature? In fact, we explored two different issues:

1. Impact assessment of grey literature – do altmetrics offer new and unique opportunities for the web-based impact measurement of reports, conferences, dissertations etc.? Do they contribute to improved visibility and impact? And if so, how?
2. Impact assessment through grey literature - how can grey literature contribute to altmetrics? Do altmetrics tools make use of grey literature?

3.

These two issues have been described in terms of diversity and broadness, as specific benefits of altmetrics compared to traditional indicators.

Our review of recent publications, together with some altmetrics tools presents a contradictory situation:

- The potential of altmetrics for grey literature is real. Altmetric data providers like Twitter, Mendeley, Facebook or figshare but also reference managers and institutional repositories are tailored for grey literature, and they already contain significant amounts of unpublished documents.
- Assessment studies on the grey literature's web-based impact show partly higher impact than journal articles or books. Apparently, altmetrics offer a unique opportunity to exhibit the real impact of unpublished research results in conference papers, dissertations, working papers etc. and to contribute to improved visibility of these documents.
- But this work remains more or less exceptional. Most studies on F1000, Mendeley etc. include only journal articles (see for instance Mohammadi & Thelwall 2013, 2014). The main reason is that altmetrics tools need unique identifiers, standard metadata and good availability. "One of the critical issues is that these aggregators concentrate on documents that have a unique object identifier, which inevitably neglects certain document types (…) For example, Altmetric.com (…)

[28] Searches were conducted with Google and in Scopus, MedlinePlus, MDConsult, UpToDate, Factiva, Lexis Nexis, Google News, and Proquest Canadian Major Dailies. Major health care associations and professional organisations likely to include related content were identified, and their websites were individually searched.

focuses its data collection on DOIs, which has led to a de facto reduction of altmetrics studies to journal articles, excluding many types of documents and journals" (Sugimoto et al 2016).

- Impact assessment with grey literature is difficult, time-consuming and manual work, and requests expert knowledge of the scientific information landscape, especially when the grey resources are not available on open repositories but somewhere in the dark web, e.g. on less-referenced, personal or other websites.
- And then there may be other reasons to dismiss grey literature. In Hammarfelt's (2014) study research impact in the humanities, all grey items - 1,006 conference papers, dissertations and reports (20%) - were skipped from the initial corpus of 5,091 scholarly works[29] because of the "scarcity" of altmetrics data in particular for the Swedish language documents.

No identifier, lack of bibliographic control and no standard metadata, unsatisfying availability – all this is not new in the field of grey literature, and Adie's (2014) suggestions to improve the situation is only too familiar for the grey community: minimum standards for metadata (PRISM[30], DC), persistent identifiers (handle, DOI), discoverability (index, repository). His suggestion: "An open, central index of scholarly grey literature that enforced a minimum level of metadata for each item (…) An alternative would be to maintain a central index of grey literature repositories (…) and to allow harvesting from each (…)".

A central index of grey literature – this sounds like utopia. Probably the main issue is that altmetrics need DOI (Adie 2016); and the DOI appears to be the only realistic option for the assignment of permanent and citable identifiers to grey literature when it comes to prepare academic output in repositories for alternative metrics (Gerritsma 2015[31], see also Brooks & Fitz 2015). But given the history of failed initiatives for standard identifiers and metadata, we must admit that this may be just another missed opportunity.

Perspectives

Are altmetrics the future of scientometrics? For the moment, they are still "in infancy" (Erdt et al. 2016), and for many researchers, impact factor and large citation databases are still preferred for determining impact, with 'pure' altmetrics tools scoring much lower, especially in physical sciences, engineering and technology[32]. Likewise, because of not-yet achieved critical mass, lack of theory, lack of quality control mechanisms, inconsistencies and multiplicity of social web sources, data, tools and methods, Sotudeh et al. (2015) speak of "immaturity of the field" and call for cautious application and interpretation, even as a complement to traditional metrics. The risk of misuse and rankings based on such arbitrary information is real.

Are download counts really a metric of scholarship or only of computer activity? Is popularity an indicator of quality? How does one deal with multiple versions of the same item? For these and other reasons, Booth (2016) condemns the limited validity of the new generation of altmetrics and suspects that they follow a logic of easiness to get the data; "(they are) neither a more accurate representation of academic 'quality' nor immune to critics" (p.41). In particular, the composite, "all-in-one" Altmetric Score has been critically appraised, because of lacking of transparency, reproducibility and stability, questionable validity and significance, and problems with data sources, consistency and completeness (see Gumpenberger et al. 2016).

The "pressure of various stakeholders" and the dependency on aggregators and social media as data providers may explain one part of the criticisms (Haustein 2016). Lack of transparency and conceptual deficit are at the opposite of the purpose of the Leiden Manifesto for Research Metrics (Hicks et al. 2015) but may be related to the increasing commercial take-over of these new tools and services by those who already dominate the scientific information market.

According to Gartner's famous Hype Cycle model[33], new technology go through a typical five-phase life cycle (figure 16). After a potential technology breakthrough kicks things off ("technology trigger") and a growing number of success stories ("peak of inflated expectations") comes the "trough of disillusionment", with growing criticisms, failures and dissatisfaction.

[29] Extraction from the SwePub database of academic publications at Swedish universities http://swepub.kb.se/

[30] Publisher Requirements for Industry Standard Metadata, see http://www.idealliance.org/specifications/prism-metadata-initiative

[31] Theses, working papers, reports, conference contribution – in Gerritsma's example (VU Amsterdam) grey items represents 14% of the whole output

[32] Innovations in Scholarly Communication Survey, http://altmetricsconference.com/who-is-using-altmetrics-tools/

[33] http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

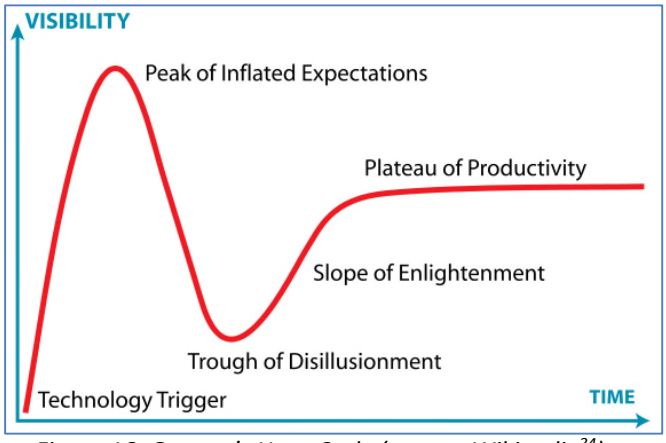


Figure 16: Gartner's Hype Cycle (source: Wikipedia[34])

On their own technology life cycle, altmetrics probably have passed by the peak of inflated expectations and are moving forward to this "trough of disillusionment", which is a necessary and salutary transition to a more realistic and satisfying situation where this new generation of metrics is no longer considered as the one and only alternative to traditional performance assessment but as new and interesting methods to assess impact of research output, complementary to traditional metrics.

Metrics shape the science, said Paul Wouters from the Centre for Science and Technology Studies at Leiden University, and we can reasonably expect that altmetrics will be part of the game. Altmetrics are already a major topic of the European Open Science Agenda and will contribute to a new rewarding and funding system.

To come back to our initial question – what is the role of grey literature in this emerging world of new assessment tools? When second- and third-generation products will appear from technology providers and later, when mainstream adoption will take off, will grey literature be part of the game or remain out of scope, just as before? For the moment, grey literature missed the opportunity to get on board. Since the Altmetrics Manifesto 2010, no real effort has been made to adapt the new assessment tools to grey literature or to make this literature suitable for altmetrics. Publications in the field of scientometrics show that journal (and sometimes book) publishing is still at the heart of research and development, not only for traditional metrics but also for alternative metrics. For instance, most of the contributions to the last Altmetrics Conference in Bucharest[35] are about journal publishing, and the rare exceptions deal with datasets and software, not with grey literature. Today, the future development of this new technology bears the risk of dismissing large parts of scientific literature – those parts not controlled by commercial publishers. Just as before, it is business as usual. Sometimes you don't get a second chance. But you have to be at the station when the train arrives. To get to the station means to:

- Contribute to research on altmetrics for or with grey literature, for instance in the fields of economics (working papers) or computer science (conference papers).
- Cooperate with altmetrics companies and teams for the development of appropriate tools that fit with grey literature.
- Accelerate the allocation of unique identifiers for grey literature and their authors and why not their institutions, above all this means partnership with DOI, ORCID and CASRAI[36], in particular for electronic theses and dissertations and for scientific reports.
- Contribute to further standardization of grey literature metadata.
- Contribute to increasing availability of grey literature in institutional repositories.

Getting grey literature into the heart of the coming mainstream adoption of altmetrics is essential not only for the future of grey literature in open science but also for academic and institutional control of research output and societal impact. This can be a special mission for academic librarians. Grey literature has always been a library-driven concept (Schöpfel 2010); today, as a recent survey shows, academic librarians demonstrate a higher awareness for altmetrics tools than researchers[37]. Perhaps this convergence or happy coincidence may be helpful.

[34] https://en.wikipedia.org/wiki/File:Gartner_Hype_Cycle.svg
[35] 3:AM Conference, Bucharest 28-29th September 2016, http://altmetricsconference.com/schedule/
[36] See the Jisc CASRAI-UK pilot on organisational identifiers https://jisccasraipilot.jiscinvolve.org/wp/2015/03/06/organisational-identifiers-working-group-outputs-and-update/
[37] Innovations in Scholarly Communication Survey, http://altmetricsconference.com/who-is-using-altmetrics-tools/ See also Malone & Burke (2016).

References

Adie, E., 2014. The grey literature from an altmetrics perspective – opportunity and challenges. *Research Trends* (37).

Adie, E., 2016. The rise of altmetrics. In: Tattersall, A. (Ed.), pp. 67-82.

Banks, M. A., de Blaaij, C., 2006. Implications of copyright evolution for the future of scholarly communication and grey literature. In: GL8 *Eighth International Conference on Grey Literature*, New Orleans, 4-5 December 2006.

Björneborn, L., Ingwersen, P., 2004. Towards a basic framework of webometrics. *Journal of the American Society for Information Science and Technology* 55 (14), 1216-1227.

Booth, A., 2016. 'Metrics of the trade': where have we come from? In: Tattersall, A. (Ed.), pp. 21-47.

Bornmann, L., 2014. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics* 8 (4), 895-903.

Bornmann, L., Haunschild, R., Marx, W., 2016. Policy documents as sources for measuring societal impact: how often is climate change research mentioned in policy-related documents? *Scientometrics,* 1-19.

Brooks, L., Fitz, G., 2015. Grey matter(s): Embracing the publisher within. *The Foundation Review* 7 (2), 38-50.

Costas, R., Zahedi, Z., Wouters, P., 2015. Do "altmetrics" correlate with citations? Extensive comparison of Altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology* 66 (10), 2003-2019.

Dinsmore, A., Allen, L., Dolby, K., 2014. Alternative perspectives on impact: The potential of ALMs and altmetrics to inform funders about research impact. *PLOS Biol* 12 (11), e1002003+.

DORA (2012). *`San Francisco Declaration on Research Assessment'*. American Society for Cell Biology (ASCB), San Francisco.

Erdt, M., Nagarajan, A., Sin, S.-C. J., Theng, Y.-L., 2016. Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics* (first online 10 August 2016).

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

Gerritsma, W., 2015. Altmetric opportunities for libraries. In: *2:AM Conference (Altmetrics Conference),* University Amsterdam, 8th November 2015.

Gumpenberger, C., Glänzel, W., Gorraiz, J., 2016. The ecstasy and the agony of the Altmetric score. *Scientometrics* 108 (2), 977-982.

Hammarfelt, B., 2014. Using altmetrics for assessing research impact in the humanities. *Scientometrics* 101 (2), 1419-1430.

Haustein, S., Costas, R., Larivière, V., 2015. Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLoS ONE* 10 (3), e0120495+.

Haustein, S., 2016. Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics* 108 (1), 413-423.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., Rafols, I., 2015. Bibliometrics: The Leiden Manifesto for research metrics. *Nature* 520 (7548), 429-431.

Ingwersen, P., Larsen, B., Carlos Garcia-Zorita, J., Serrano-López, A., Sanz-Casado, E., 2014. Influence of proceedings papers on citation impact in seven sub-fields of sustainable energy research 2005–2011. *Scientometrics* 101 (2), 1273-1292.

Konkiel, S., 2012. Altmetrics: An app review. In: *OCLC Innovation in Libraries, post-conference event, LITA Forum 2012*, Columbus, OH, USA, October 8, 2012.

Kraker, P., Lex, E., Gorraiz, J., Gumpenberger, C., Peters, I., 2015. Research data explored II: the anatomy and reception of figshare. In: *STI2015 20th International Conference on Science and Technology Indicators*, Lugano, 4 September 2015.

Kwok, R., 2013. Research impact: Altmetrics make their mark. *Nature* 500 (7463), 491-493.

Larivière, V., Zuccala, A., Archambault, E., 2008. The declining scientific impact of theses: Implications for electronic thesis and dissertation repositories and graduate studies. *Scientometrics* 74 (1), 109-121.

Lindsay, J. M., 2016. PlumX from plum analytics: Not just altmetrics. *Journal of Electronic Resources in Medical Libraries* 13 (1), 8-17.

Malone, T., Burke, S., 2016. Academic librarians' knowledge of bibliometrics and altmetrics. *Evidence Based Library and Information Practice* 11 (3), 34+.

Moed, H. F., Halevi, G., 2014. The multidimensional assessment of scholarly research impact. *Journal of the Association for Information Science and Technology* 66 (10), 1988-2002.

Mohammadi, E., Thelwall, M., 2013. Assessing non-standard article impact using f1000 labels. *Scientometrics* 97 (2), 383-395.

Mohammadi, E., Thelwall, M., 2014. Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology* 65 (8), 1627-1638.

Neylon, C., Willmers, M., King, T., 2014. *Rethinking impact: Applying altmetrics to Southern African research.* Report, University of Cape Town. Scholarly Communication in Africa Programme, Paper 1.

NISO, 2014. *Alternative metrics initiative. Phase 1 White Paper*. National Information Standards Organization, Baltimore MD.

NISO, 2016. *Altmetrics definitions and use cases. A recommended practice of the National Information Standards Organization*. Draft. NISO RP-25-201x-1, National Information Standards Organization, Baltimore MD.

Piwowar, H., 2013. Altmetrics: Value all research products. *Nature* 493 (7431), 159.

Pontika, N., Knoth, P., Cancellieri, M., Pearce, S., 2016. Developing infrastructure to support closer collaboration of aggregators with open repositories. *LIBER Quaterly* 25 (4).

Priego, E., 2014. *On metrics and research assessment*.

Priem, J., Taraborelli, D., Groth, P., Neylon, C., 2010. *Altmetrics: A Manifesto*.

Priem, J., Hemminger, B. H., 2010. Scientometrics 2.0: New metrics of scholarly impact on the social web. *First Monday* 15 (7).

Prost, H., Le Bescond, I., Schöpfel, J., 2010. Usage assessment of an institutional repository: a case study. In: *GL12 Twelfth International Conference on Grey Literature*, Prague, 6-7 December 2010.

Schöpfel, J., 2010. Towards a Prague definition of grey literature. In: *GL12 Twelfth International Conference on Grey Literature*. Prague, 6-7 December 2010

Schöpfel, J., Prost, H., 2009. Usage of grey literature in open archives: state of the art and empirical results. In: *GL11 Eleventh International Conference on Grey Literature*. Washington D.C., 14-15 December 2009.

Sibbald, S. L., MacGregor, J. C. D., Surmacz, M., Wathen, C. N., 2015. Into the gray: a modified approach to citation analysis to better understand research impact. *Journal of the Medical Library Association: JMLA* 103 (1), 49-54.

Sotudeh, H., Mazarei, Z., Mirzabeigi, M., 2015. CiteULike bookmarks are correlated to citations at journal and author levels in library and information science. *Scientometrics* 105 (3), 2237-2248.

Sud, P., Thelwall, M., 2014. Evaluating altmetrics. *Scientometrics* 98 (2), 1131-1143.

Sugimoto, C. R., Work, S., Larivière, V., Haustein, S., submitted. Scholarly use of social media and altmetrics: a review of the literature. *Journal of the American Society for Information Science and Technology*.

Tattersall, A. (Ed.), 2016. *Altmetrics: a practical guide for librarians, researchers and academics*. Facet Publishing, London.

Thelwall, M., Haustein, S., Larivière, V., Sugimoto, C. R., 2013. Do altmetrics work? Twitter and ten other social web services. *PLOS ONE* 8 (5), e64841+.

Urquhart, C., Dunn, S., 2013. A bibliometric approach demonstrates the impact of a social care data set on research and policy. *Health Information & Libraries Journal* 30 (4), 294-302.

Wilkinson, D., Sud, P., Thelwall, M., 2014. Substance without citation: evaluating the online impact of grey literature. *Scientometrics* 98 (2), 797-806.

Wouters, P., Costas, R., 2012. Users, narcissism and Control—Tracking the impact of scholarly publications in the 21st century. In: *STI2012 17th International Conference on Science and Technology Indicators (STI)*, 5-8 September, 2012, Montreal, Quebec, Canada.

Yu, M.-C., Wu, Y.-C. J., Alhalabi, W., Kao, H.-Y., Wu, W.-H., 2016. ResearchGate: An effective altmetric indicator for active researchers? *Computers in Human Behavior* 55, 1001-1006.

Zoller, D., Doerfel, S., Jäschke, R., Stumme, G., Hotho, A., 2016. Posted, visited, exported: Altmetrics in the social tagging system BibSonomy. *Journal of Informetrics* 10 (3), 732-749.

The complete bibliography with links and further readings is available at http://www.citeulike.org/user/Schopfel/tag/gl18

All websites visited in August and September 2016.

Acknowledgment to Paul Needham, JISC, for helpful advice with the IRUS-UK project.

https://doi.org/10.26069/greynet-2025-000.534

Grey Literature Citations in the age of Digital Repositories and Open Access*

Silvia Giannini Stefania Biagioni, Sara Goggi and Gabriella Pardelli (Italy)

Abstract

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

1. Introduction

About ten years ago we studied the impact of grey literature (GL) on conventional literature by observing the impact of grey citations in two different scientific fields, «after the growth in the use of the WWW» [Di Cesare et al. 2005]. Over the last decade the international scientific community and its players have undergone (and still undergo) essential changes with respect to the representation and dissemination of knowledge. The formalization of the Open Access (OA) model dates back to the beginning of the last decade: OA is a movement born in the 1990s with the purpose to contrast the monopoly of commercial publishing houses thus making knowledge accessible for free without violating the intellectual property rights or threaten the quality of a scientific work. Along with the success of OA, the prototype of the 90s e-print archives evolved into the current digital repositories (institutional or subject based). The new paradigms of scientific communication and the most advanced computer science technologies fostered the achievement and use of these novel infrastructures: in particular, institutional repositories have gained a specific and relevant role in storing, preserving and disseminating scientific information.

The international Declarations of Budapest 2002, Bethesda 2003 and Berlin 2004 contributed to the definition of OA principles and opened the way to the movement from a normative point of view thus fostering the promulgation of national laws and European legislation as well (reference n° 2012/417/UE). Nowadays many academic and research institutions choose to adhere to the OA movement by issuing policies which compel to deposit/file their research products in OA repositories. Nevertheless the debate on this themes is extremely heated and many issues still need to be cleared up: from problems related to the quality of open access products, the peer-review processes, the integration between the OA world and the evaluation of research to – last but not least - the impact of citations of OA works [Guerrini 2010; De Bellis 2005; Eysenbach 2006]. In our current digital era bibliographical citations has gained a strategic role within the mechanisms of scientific communication, especially due to the implementation of the citation indexing services [Cassella 2011]. Citation has thus become «….the currency in scientific communication trade. It is a small denomination bill (quoting does not cost that much) but with a very high symbolic buying power…» [De Bellis 2014][1]. Besides it is undeniable that from a bibliometric point of view the two more widely recognized and used standards are those based on the number of citations per article and on the impact factor per journal.

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

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

* First published in the GL17 Conference Proceedings, February 2016

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

2. Materials and method

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

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

Table 1 – Sampled Journals

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

Table 2 – Sampled Proceedings

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

Scopus citation database places the journals CL and LR&E in areas related also to the Humanities and Social Sciences: Language and Linguistics for CL; Language and Linguistics, Education and Library and Information Sciences for LR&E. Indeed, Computational Linguistics is a discipline that draws contributions from different fields of study such as linguistics, psychology, mathematics and statistics, in addition to computer science.
For all these reasons we considered the selected journals and conference proceedings as belonging to two different scientific communities: Computer science and Engineering and Computational Linguistics.

The journals present a value of IF substantially stable in the considered timespan, with the exclusion of CL showing a higher IF in 2013. This value also determines a significant shift of rank in the same year and makes CL the journal with the higher IF.
We extracted the information directly from primary sources, that is the bibliographical references of the articles in the selected journals and proceedings. The corpus was built by grouping the gathered data in six informative classes: year, issue number, bibliographical reference, kind of document – Grey (G) or Published (P), document type, standardized document type (Table 3).

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

Table 3 – example extracted from Computational Linguistics (2012)

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

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

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

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

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

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

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

3. Analysis of data and results

3.1 Frequency of GL citing

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

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

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

Table 4 (ni=not included)

What is remarkable in Table 4 is:

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

3.2 Frequency of GL use

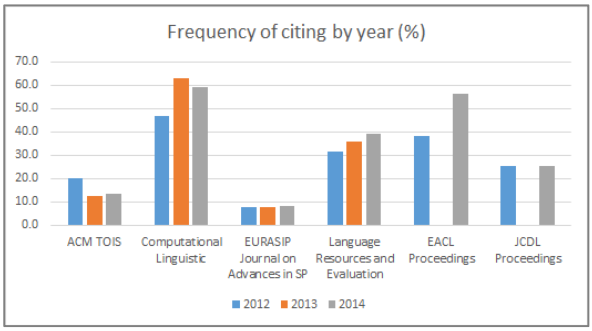
The frequency of GL use shows the percentage of articles with at least one GL citation out of the overall amount of articles: it is globally very high and varies from a minimum of 69,8% to a maximum of 100%.

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

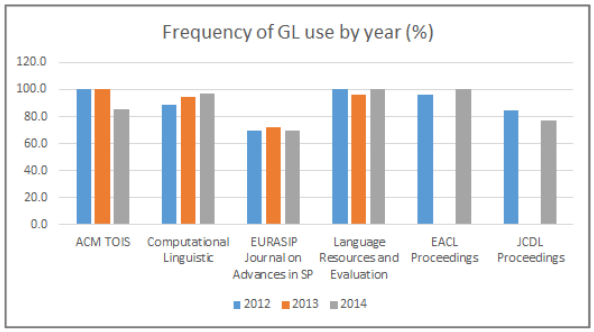
Table 5

Table 5 shows:

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



Graph 1



Graph 2

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

3.3 Frequency/Intensity of GL use

The intensity of GL use varies from a minimum of 10,9% (*EURASIP*) to a maximum of 66,7% (*Computational Linguistics*). Table 6 shows that generally the intensity of GL use increases for both proceedings and *LR&E*, it alternates between increase and decrease of values in *ACM TOIS* and *CL* while is stable for *EURASIP.*

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

Table 6

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

From Table 6 we can note the following:

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

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

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

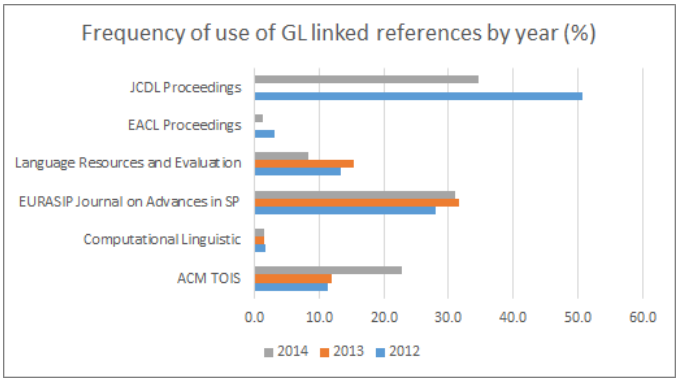
Title	2012			2013			2014		
	Number of GL references	Number of *linked*-GL references	Frequency of *linked*-GL references (%)	Number of *linked*-GL references	Number of *linked*-GL references	Frequency of *linked*-GL references (%)	Number of *linked*-GL references	Number of *linked*-GL references	Frequency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

Table 7

Table 7 shows:

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

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



Graph 3

[4] Cfr. http://www.mitpressjournals.org/page/about/coli.

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

4. GL Typology

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

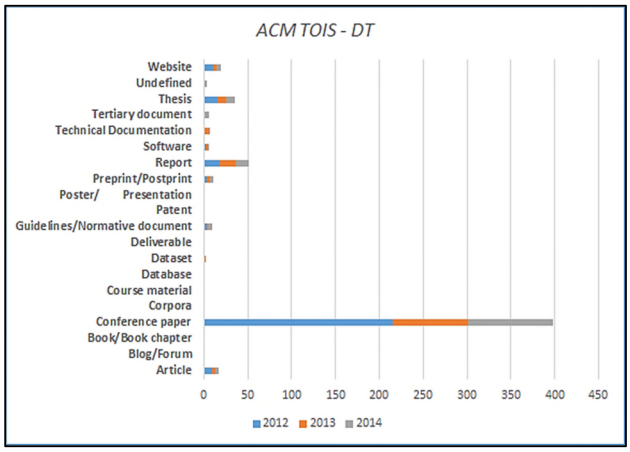
2012																				
Title	Article	Blog/Forum	Book/Book chapter	Conference paper	Corpora	Course material	Database	Dataset	Deliverable	Guidelines/ Normative document	Patent	Poster/ Presentation	Preprint/ Postprint	Report	Software	Technical Documentation	Tertiary document	Thesis	Undefined	Website
ACM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

Table 8

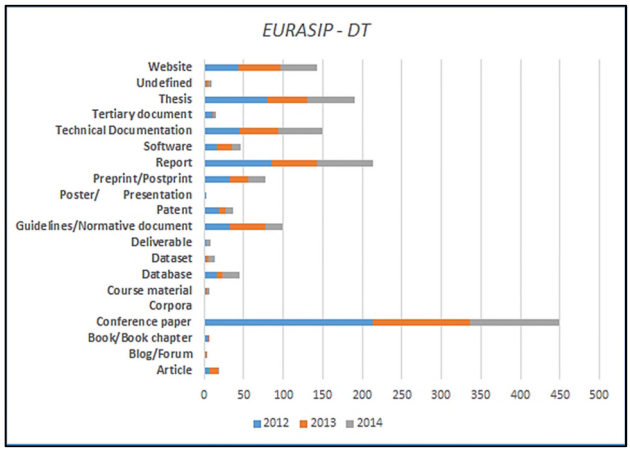
The most cited types of GL documents are: Conference papers (this group presents the highest number of GL citations), Reports, Thesis and Preprint/Postprint. These four types of document are the most cited regardless of the year, the nature of the products analyzed (journals and proceedings) and the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

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

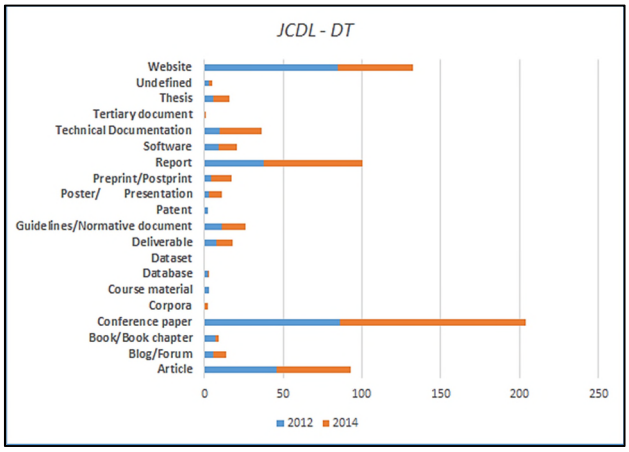
The Graphs below show the distribution of GL types for each source over the years: Graphs 4, 5, 6 show the document types in Computer Science and Engineering while Graphs 7, 8, 9 show the document types in Computational Linguistics. The graphical visualization allows us to better understand the intensity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



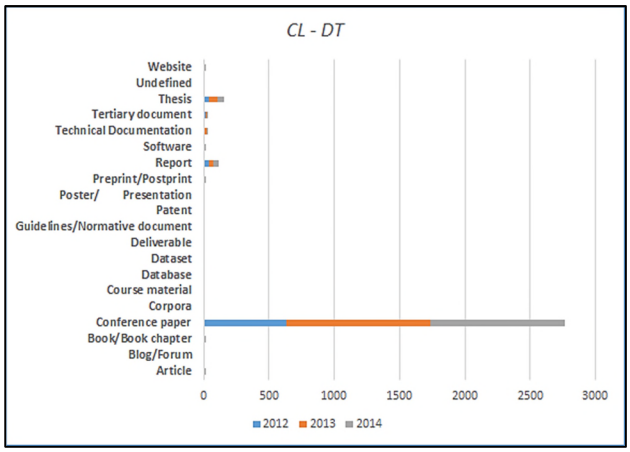
Graph 4



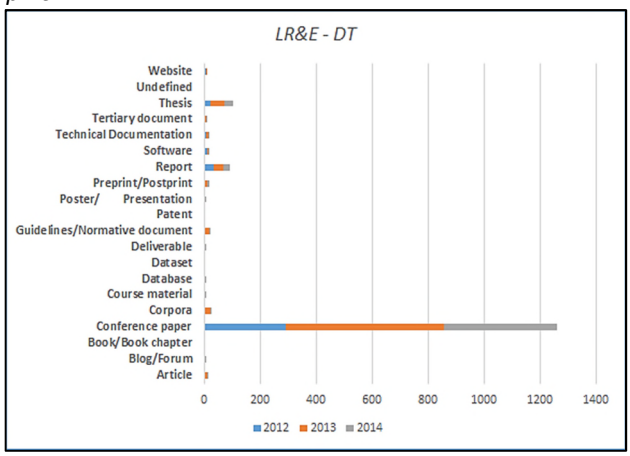
Graph 5



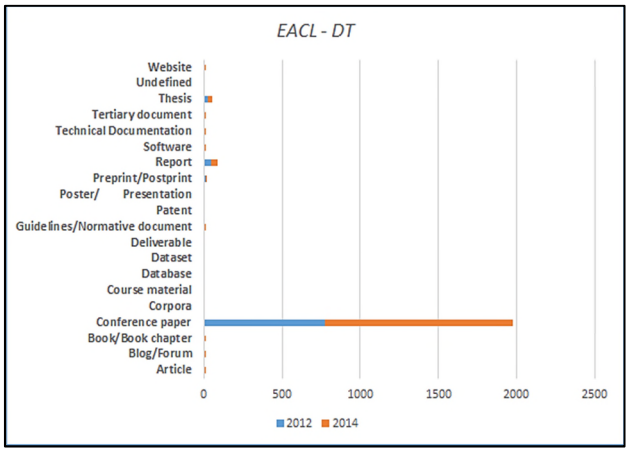
Graph 6



Graph 7



Graph 8



Graph 9

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

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

5. Conclusion

In 2004 we analyzed two sample data belonging to two very different scientific fields. In this work the disciplinary boundaries of sample data are much less defined: but nevertheless, there are several significant differences, both in frequency and intensity of use of grey citations and about the cited type of documents, especially related to the specific field of study of each resource analyzed. The results obtained show that the Engineering domain has the least number of grey citations while the area of Computational Linguistics uses them most. In this respect, however, we must keep in mind that this result is strongly influenced by the presence of grey citations related to Conference papers published by the Association of Computational Linguistics. Even in the field of Computer Science, many "grey" Proceedings are collected and distributed freely by Associations, Institutions and no-profit services.

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

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

Bibliography

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

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

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

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

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

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

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

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

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

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

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

[5] Full URL:
https://www.google.it/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiVkN7nuabKAhWB0xQKHbJ2BBgQFggfMAA&url=http%3A%2F%2Fwww.sba-old.unimore.it%2FInside%2FBiblioteca%2520Digitale%2FDoc%2FLa%2520citazione%2520bibliografica%2520nell%27epoca%2520della%2520sua%2520riproducibilit%25C3%25A0%2520tecnica.pdf&usg=AFQjCNFaL-NETfCMU_oTGAt9RXsnBNhByw&sig2=LNsOOd99NCx815rNggZqiw&bvm=bv.111677986,bs.2,d.ZWU&cad=rja

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

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

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

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

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

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

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

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

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

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

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

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

https://doi.org/10.26069/greynet-2025-000.534

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

Flavia Cancedda and Luisa De Biagi (Italy)

Abstract

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

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

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

Flexibility of grey literature: *pleasure & pain.*

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

In fact, Grey literature is made by researchers and 'feeded' by Research, but isn't still usually viewed in the 'Upper Class' of the scholarly literature because, for example, it's difficult to cite in academic journals. Grey documents are not totally and constantly considered in citation indexes like Web of Science or Scopus: some editors and publishers even discourage any formal citation of preprints and similar (e.g. American Chemical Society[1], American Association for Cancer Research - AACR - , the *Journal of Clinical Investigation* by American Society for Clinical Investigation, *The Journal of Experimental Biology* by the Company of Biologists[2].

The wide panorama rewarding GL is still without a 'sustainable ecosystem', though good initiatives like greylit.org are to be considered a significant effort[3]. So, important challenges and processes have previously to be done:

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

* First published in the GL17 Conference Proceedings, February 2016

[1] Most ACS journals do not explicitly allow preprints.

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

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

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

In particular NBN, defined by RFC 3188 standard (http://www.rfc-editor.org/rfc/rfc3188.txt) is a persistent identifier based on URN (Uniform Resource Name), and identifying in an unambiguous way a publication. In Italy NBN:IT is linked to legal deposit and digital preservation. The project, coordinated by the National Libraries of Florence, Rome, and Venice, by the Foundation Rinascimento Digitale and by the Conference the Italian University Rectors, completes the PI current offer[5].

Past and present: identifier codes in bibliographic fields

During the 20th century the development of identifier codes - in most cases internationally spread under ISO[6] auspices - encouraged the idea that few things were yet "unidentifiable" and "uncontrollable" in documentary fields. It could be useful to remember that "bibliographic identifier code" means a sequence of characters - mostly numbers, but not only - preceded by an alphabetic acronym that indicates the type of code. The codes are used to identify uniquely publishing/intellectual works or identities (persons, bodies etc.) in the world of communication and media. Each type of code is issued by a responsible agency - guarantor of integrity and uniqueness of it -; each code matches to a set of descriptive data and metadata managed by the agency (e.g.: name, title, author, year of publication, date of code assignment, etc.). The bibliographic codes are widely used in the whole publishing and media market for commercial distribution. They are also necessary for tracking the publication/work/entity in whatever step of the circulation chain, from the producer to the end-user and to other agencies which process bibliographic data for providing advanced information services (e.g.: bibliographic or informative archives of scientific publications or scholars/scientists). The most common and known identifiers? ISBN for books and ISSN for periodicals, but also, e. g., ISAN for audiovisuals, ISRC for recorded music, ORCID for scholars.

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

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

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

[4] Conference of Italian University Rectors

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

[6] International organization for standardization, iso.org.

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

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

Dynamic model for re-using and disseminating P.I. ***(the public control and re-use of bibliographic warranties)***
In the world there is a sizeable quantity of bibliographic identifier codes (the ISO family alone contains ten different codes), implemented and continuously managed by several different international agencies, often spread widely by country. The overall mass of bibliographic and informative data they collect/update/maintain is impressive[7].
It is now urgent to establish common guidelines - about description, metadata collecting systems, cross-identification - that could be widely shared and implemented by international and national agencies, or by any other component of the information chain. In fact, by now the different agencies – even in cases when they are established and promoted by bodies having similar organization profiles, and even though those agencies share similar socio-cultural backgrounds, aims and objectives – do not usually cooperate in common policies; mainly they do not collaborate in gathering the bibliographic information, which is the basis activity for building the metadata grids necessary to ensure a reliable correspondence between publishing objects and identifier codes.
How to draw an information cross-exchange now? First of all, it is crucial that international and national agencies acknowledge that a cross-cooperation is needed: the exchange of data is useful for each actor of the informative chain, and particularly for agencies themselves. The informative data common to more than one system identification (i. e., the main part of bibliographic data, as names and address of publishers, trademarks, names and address of issuing bodies, names of intellectual contributors, titles of works edited in different media, title of collections…) could be easily pooled by different registration agencies. This operation should not be implemented collecting the data in a new *nth* database – obviously uncontrollable because of its size and because of the different policy needs of the co-owners -, but using the cross-reference system allowed by linked data technology[8].
The agencies responsible for assignment of codes could choose this kind of data-sharing, agreeing about a minimum set of data and metadata acknowledged as "common" and shareable; so, same agencies would evidently save substantially their own human and technological resources, avoiding millions of recording operations, often identically repeated in different identifying systems. Moreover, this system could be implemented in open access regime, so the results would become immediately available for the whole Net public.
This big saving would make available the same H&T resources for other kinds of advanced services; the loss of profits depending from the public availability of bibliographic data (now mostly priced) could be balanced by the pricing of new customized services, particularly when they are addressed to publishers, providers, data aggregators, institutional bodies, etc.

[7] Example: at present the ISBN network collects essential public data of more than 1 million of publishers (the enormous number of the books identified by ISBN code is actually not known…); the ISSN network records the bibliographic data of about 1.9 million of serial publications, etc.
[8] General info about the world of linked data in http://www.w3.org/standards/semanticweb/data.

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

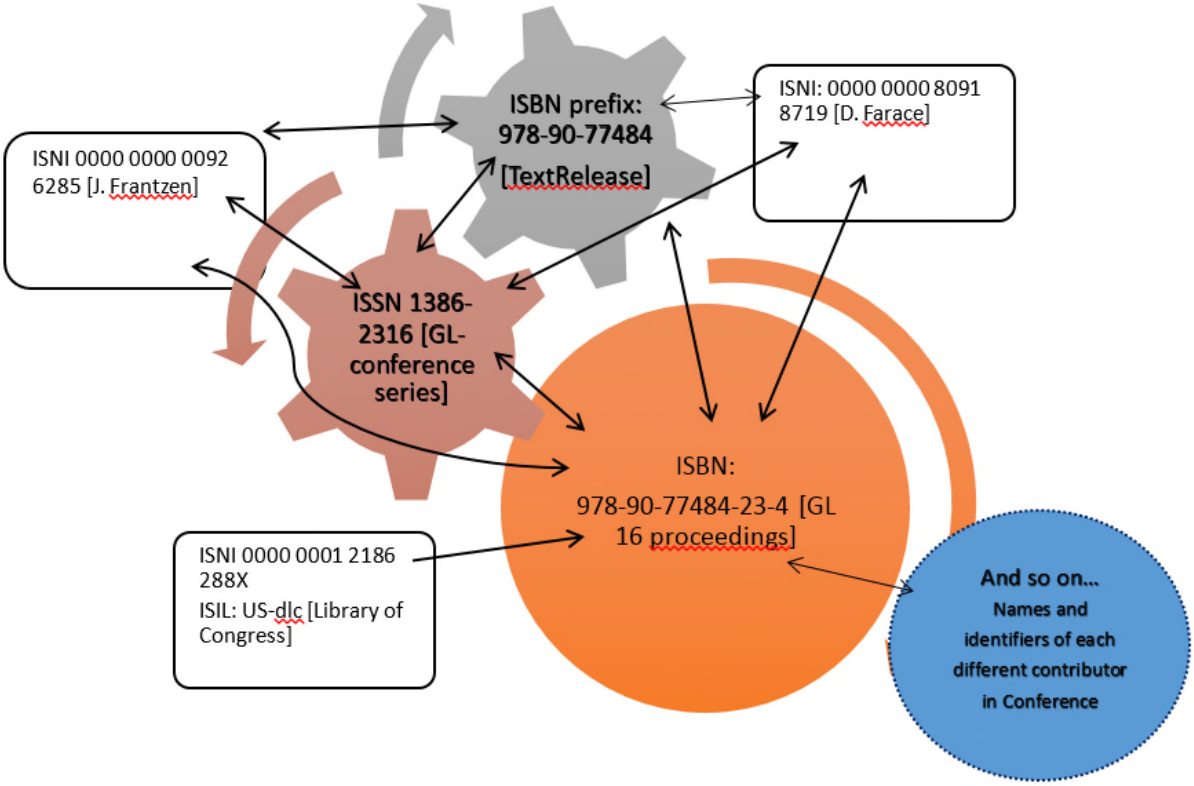
cataloguing record (as in the homepage http://www.greynet.org/)

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

Several components of this description are already identified by codes:

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

Each of the above identifier codes (ISBN, ISNI, ISIL, ISSN) has been issued and granted by a recording system where personal and bibliographical data – according to the specific internal rules of each identifying system - has been previously checked and stored.
It is not difficult to imagine that this traditional list of elements, codes and related metadata could be easily make available for a linked data approach, where the ontologies (i.e.: roles and functions) are the digital glue linking chains of digits (the codes) which, in their turn, drag behind other metadata – id est: descriptive textual information or URI – already validated and not to be re-written or for the umpteenth-time repeated.



References:

RFC 2288, Feb 1998, C. Lynch, C. Preston, R. Daniel, *Using Existing Bibliographic Identifiers as Uniform Resource Names;*
RFC 3044:2001, S. Rozenfeld, *Using The ISSN (International Serial Standard Number) as URN (Uniform Resource Names) within an ISSN-URN Namespace*;
RFC 3187:2001, J. Hakala, H. Walravens, *Using International Standard Book Numbers as Uniform Resource Names*;
RFC 3188:2001, J. Hakala, *Using National Bibliography Numbers as Uniform Resource Names;*
I-D IEFT-URNBIS-RFC3044BIS-ISSN-URN-01, 7 Nov 2012, P. Godefroy, *Using International Standard Serial Numbers as Uniform Resource Names;*
I-D IEFT-URNBIS-RFC3187BIS-ISBN-URN, 19 Oct 2012, M. Huttunen, J. Hakala, A. Hoenes eds., *Using International Standard Book Numbers as Uniform Resource*

ISO standards (last or current editions):

ISO 15707:2001, *International Standard Musical Work Code (ISWC)*
ISO 3901:2001, *International standard recording code (ISRC)*
ISO 15706-1:2002 *International Standard Audiovisual Number (ISAN), Part 1: Audiovisual work identifier*
ISO 2108:2005 *International standard book number (ISBN)*
ISO 3297:2007 *International standard serial number (ISSN)*
ISO 15706-2:2007 *International Standard Audiovisual Number (ISAN),- Part 2: Version identifier*
ISO 10957:2009 *International standard music number (ISMN)*
ISO 21047:2009 *International standard text code*
ISO 26324:2012 *Digital object identifier system*
ISO 27729:2012 *International standard name identifier (ISNI)*
ISO 17316:2015 *International standard link identifier (ISLI)*

Auger, C.P., Ed. (1989) *Information Sources in Grey Literature* (2nd ed.). London: Bowker-Saur. ISBN 0862918715

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

https://doi.org/10.26069/greynet-2025-000.534

Developing, linking, and providing access to supplemental genetics dataset vcf files*

Plato L. Smith II and Lauren McIntyre (United States)

Abstract

This conference proceeding paper is the written version component of the data panel discussion on developing a dataset collection using Zenodo for a professor in the Department of Molecular Genetics & Microbiology at the University of Florida.

An internal University of Florida George A. Smathers Libraries Strategic Opportunities Program (SOP) grant award provided support for the creation and development of an initial supplemental datasets digital collection of large, static variant call format (vcf) in zenodo. The "Documenting a Genomics Variant Files Data Management: Developing Research Data management (RDM) workflows and providing research data access via HPC" project inspired this paper. The large vcf datasets used for this project ranged from 34 megabytes to 43 gigabytes. The researcher needed to (1) develop a data repository for supplemental datasets vcf files too large for attachment as supplemental data files for journal submissions, (2) provide digital object identifiers (DOIs) for all vcf dataset files, and (3) link the supplemental vcf dataset files to the journal article via the vcf doi. These three outcomes were accomplished during phase 1 (June 2016 – December 2016) of this project and presented at the GL18. Phase 2 (January 2017 – June 2017) of this project includes performing (1) a dataset reproducibility interview, (2) an open archival initiative protocol for metadata harvest (OAI-PMH) from Zenodo to the University of Florida institutional repository (IR@UF), and (3) developing a similar use case project for researchers in UF/IFAS Nature Coast Biological Station (NCBS).

Introduction

Data continuously extended, stored, and consulted is useful to science (*As We May Think*, Bush, 1945). Data is useful to science if data is accessible, discoverable, and reproducible. This project allows researchers to access, discover, share, and cite large variant call format (VCF) datasets from a data repository. This project can be as a use case scenario for articulating, demonstrating, and detailing the data lifecycle.

Data generation, raw data files are often text files of large sizes. These raw data- effectively what comes off the machine have been a focus for many in terms of collection and preservation. While there has been a huge effort expended in this area for gene expression data, there has been much less effort for capturing and storing DNA based variant data, and mass spectrometry based metabolomics and proteomic data. As our capacity for data generation continues to expand, thoughtful discussion on what to keep, for how long, and where is an ongoing important public dialogue.

The gene expression and data collection as an example, array images (GEO), or sequence fastq files in the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) are the raw data. The linking of signal intensity, or number of reads to an identifier is the basis then for all subsequent analysis in the Gene Expression Omnibus (GEO). "Geo is a public functional genomics data repository supporting MIAME-compliant data submissions" (NCBI Resources, GEO). Updates, deposition of quantified data, and annotation of the features for the snapshot in time are possible. The quantified processed file is the basis for further analysis and is an intermediate file. Public access and long-term storage of primary data in GEO are important to the research community. The full list of features and the result of whatever analysis done by an individual scientist is then a results file. The results file is often relatively small and it can be included with publication. While not consistently encouraged among all scientific journals, the GSA journals do try to capture this final file and include as supplementary material. Individual scientists should think more carefully about how the inclusion of a full supplement helps their own research long term. Complete data capture, access, storage and reuse enable current and future researchers to start/build research. It obviates the necessity of identifying exactly which file in the folder was the full and final result file. Multiple versions, trainee turnover, and manuscript submission contribute to subtle differences among files on the main analysis system. The final supplement is an opportunity to capture the full information about the analytical results of a particular experiment. This is particularly beneficial if there is any break in the work on any one project, where trainees may not directly overlap. There are, of course, more altruistic reasons for including such files

* First published in the GL18 Conference Proceedings, February 2017

with a paper. However, the main beneficiaries of such preservation are often the lab that generated the results and close collaborators of the project.

As well developed, though imperfectly executed, this data preservation, processing, and sharing paradigm is for gene expression data, for other omics data these steps are significantly more challenging. In attempting to solve these challenges, we present one promising option for scientists to consider, partnerships with their institutional libraries. We describe one scenario here and invite the community to dialogue with us about these ideas.

For some population studies such as the 1000 genomes, special repositories are searchable have been created (). There is no consensus or centralized repository for all pieces and variant information. The affordability of sequencing efforts increases other populations sequencing. SRA (Sequence Read Archive) allows deposits of raw data in the form of sequence reads. Theoretically, full genomes can be stored at NCBI. Sequencing information, data quality, and the deposit workflow of genomes in NCBI needs enhancement. One of the most useful pieces of intermediate data is the list of variants, usually in a standard .vcf format. Depending on the population size, the .vcf files can be quite large. They are certainly cumbersome for inclusion in a journal supplement. It is not at all clear that these are good targets for either NCBI or journal supplements- for one compelling reason- the utility of these files is time limited. After the passage of some time, perhaps ten years, our algorithms will be better, we would likely prefer to reanalyse the raw data. The argument for semi-permanent storage of raw data is compelling, for intermediate files, it is less so. In some cases, the computational time to regenerate files is negligible. In the case of a vcf file, this is less true as creating these files often takes weeks if not months of computational effort, as well as considerable human effort along the way. There is therefore a good reason to keep these files in the short to medium term. How to keep them and then make them accessible for both the initial research group and the larger public has some current possibilities.

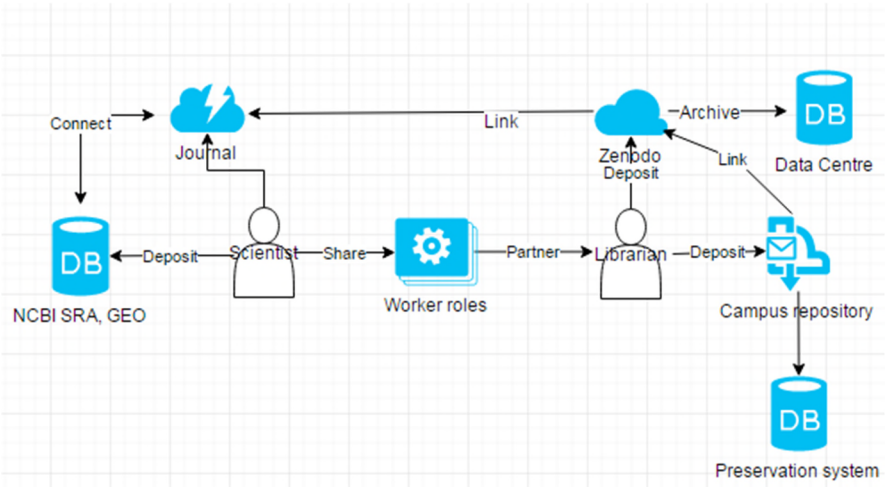


Figure 1. Raw data to SRA, results to journal, vcf to zenodo and IR@UF projected workflows

Researcher need for a vcf data repository

"We can generate terabytes of data about the genetics of population from a myriad of species. As a community, we have even made progress in developing unified data formats for reporting on observed genetic variation (vcf files). Yet, there is no national repository for this information. The libraries represent a transparent, public venue for sharing information on variants.

As a test project, the UFL library has collaborated with investigator Lauren McIntyre on a project in Drosophila (fruit fly). A population study of D. simulans funded by the NIH where ~200 different simulans genotypes were sequenced with 10x-20x coverage were analysed for variants. The resulting variant files have been deposited in the UFL library and links are included in the manuscripts currently under review." - UF Molecular Genetics & Microbiology Scientist

Using Zenodo as a vcf data repository

Zenodo is a public general research data repository that supports multiple forms of data from publication, poster, presentation, dataset, to image, video/audio, software, and lesson deposits up to 50 Gigabytes per record. The "Large 'static' processed data files" collection in zenodo enables the aggregation, representation, dissemination, and preservation of large, vcf datasets that exceeded the file size limits for supplemental data files for publications. These vcf datasets have digital object

identifiers (DOI); and BibTex, CSL, DataCite, Dublin Core, JSON, MARCXML and Mendeley export features. The vcf datasets collection are harvestable via Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) using the Zenodo Harvesting API for vcf datasets collection. The datasets in zenodo are guaranteed access and preservation for at least 20 years. Zenodo support provided an excellent response to request for information on the access, storage, and preservation of the vcf datasets stored in zenodo. The following preservation information about datasets stored in zenodo is useful to the GL18 and DataONE Users Group communities.

Zenodo stores its data as part of CERN's disk-storage service EOS (see http://information-technology.web.cern.ch/services/eos-service). This same service is being used for High Energy Physics data storage that is being obtained from CERN's LHC. The data is stored in CERN Data Centre (see general overview of CERN's Data Center here: http://information-technology.web.cern.ch/about/computer-centre as well as and more detailed report here on bit-preservation practices: https://cds.cern.ch/record/2195937/files/iPRES2016-CERN_July3.pdf). Additionally on the legal note of data hosting: CERN premises (including the Data Centre) are located on an intergovernmental territory, which is exempt from the host country's local jurisdiction (the host countries for the data centre is Switzerland and France, but as said their legislation does not apply since CERN has a status similar to United Nations).

As for Zenodo's internal workflow for data preservation - we are partially implementing the OAIS model for data archiving (ISO 14721), and are working towards being fully compliant with the model later this year or early next year. In parallel to that we will be working in the near future on fulfilling the requirements to be compliant with the Data Seal of Approval. All records in Zenodo (in all Zenodo collections) are treated equally and are undergoing the same procedures and workflows.

We compute MD5 checksums for all data that is uploaded - this information is visible to the user after upload, as well as returned in the response from the REST API for verification by the user. The checksum comes from the aforementioned CERN's EOS service, which is also used by EOS (alongside other measures) to prevent "bit-rot". For the time being we only generate and extract metadata from software repositories published through our GitHub integration. The metadata is normalized in a sense that its structure and data types are conforming to a pre-defined JSON schema, which is consistent across all records.' – Zenodo Technical Support, 2016

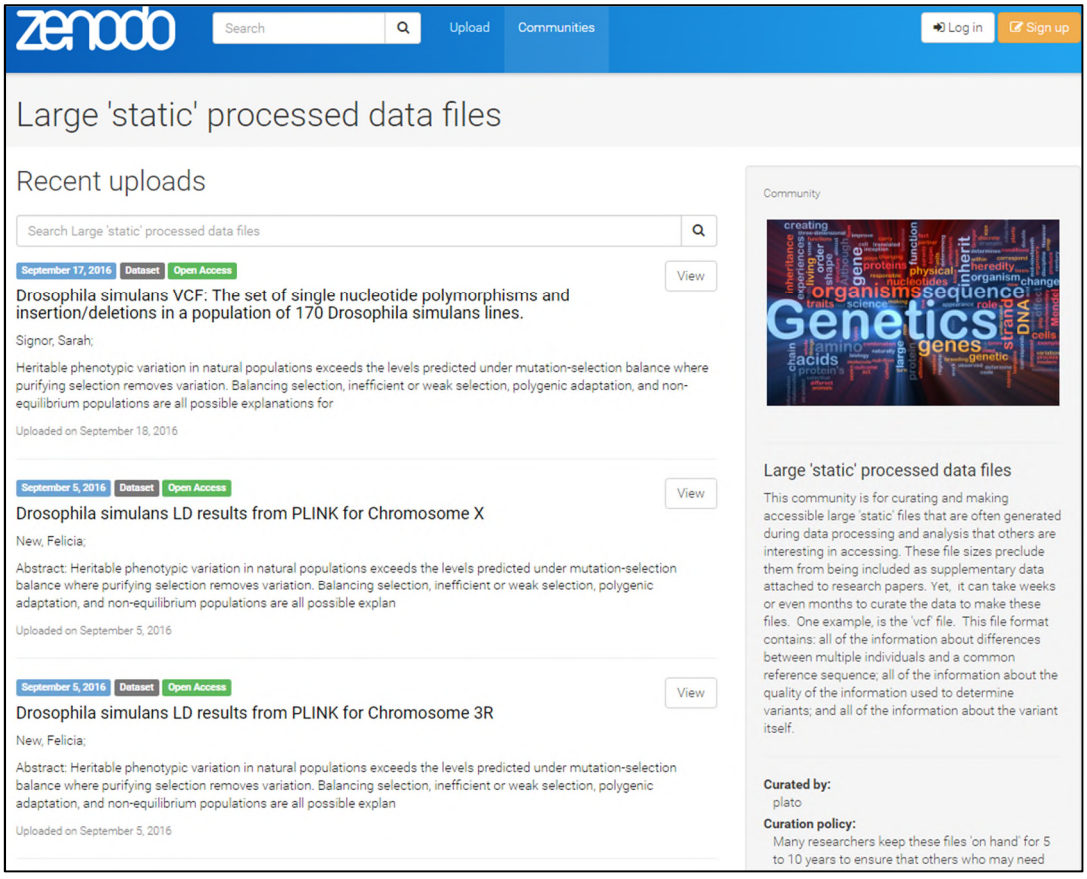


Figure 2. Large 'static' processed data files (vcf) repository for access, linking, and sharing

Benefits, features, and outcomes from this project

Some benefits and features from this project include but not limited to the following:

- Genetics project website in Open Science Framework (OSF) – Phase 1
- Genetic vcf datasets collection in zenodo - (Phase 1)
- DOI for vcf dataset files by DataCite
- Export vcf datasets in multiple formats (e.g. BibTeX, CSL (Citation Style Language) JSON Export, DataCite, Dublin Core, MARCXML, Mendeley)
- vcf datasets data collection harvestable via OAI-PMH API
- MD5 checksum performed on all uploaded data
- Zenodo partially implementing the OAIS model for data archiving (ISO 14721:2012) – Space data and information transfer systems - OAIS Reference Model – full compliance later this year or early next year
- Zenodo working in near future on fulfilling the requirements to be complaint with the Data Seal of Approval (DSA)
- Data access/preservation guaranteed for at least 20 years - FAQ
- Zenodo is US Department of Transportation (DOT) Public Access Plan conformant http://ntl.bts.gov/publicaccess/repositories.html

Acknowledgements

The authors acknowledge Allison Meryl Mores, Felicia New, Patrick Reakes, Ruth Isaacson, Genetics Editorial Office, and GreyNet/GL18.

References

[report] Consultative Committee for Space Data Systems. (2012). *Reference model for an Open Archival Information System (OAIS)* (Magenta Book CCSDS 650.0-B-1). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

[website] IGSR: The International Genome Sample Resource. (2016). Retrieved from http://www.internationalgenome.org/wiki/Analysis/vcf4.0/

[journal article] Kurmangaliyev, Y. Z., Favorov, A. V., Osman, N. M., Lehmann, K., Campo, D., Salomon, M. P., Tower, J., Gelfand, M. S., & Nuzhdin, S. V. (2015). Natural variation of gene models in Drosophilia melanogaster. *BMC Genomics.* doi: 10.1186/s12864-015-1415-6

[report] National Institutes of Health. (2015). *National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research*. Retrieved from https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf

[report] National Science Foundation Public Access Plan. (2015). *Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation.* Retrieved from https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

[website] NCBI. (2016). SRA, Sequence Read Archive. Retrieved from https://www.ncbi.nlm.nih.gov/sra

https://doi.org/10.26069/greynet-2025-000.534

A terminological "journey" in the Grey Literature domain*

Roberto Bartolini, Gabriella Pardelli, Sara Goggi, Silvia Giannini, and Stefania Biagioni (Italy)

1. Introduction

"It is by means of terms that the expert usually transfer their knowledge and again through terms scientific communication reaches the highest effectiveness. Therefore we can assert that terminology - in the sense of a set of representative and domain-specific units - is necessary for representing and connecting specialized fields as well as any attempt to represent and/or transfer scientific knowledge requires, more or less extensively, the use of terminology." (Cabré, 2000). "When we read the articles or papers of a particular domain, we can recognize some lexical items in the texts as technical terms. In a domain where new knowledge is generated, new terms are constantly created to fulfill the needs of the domain, while others become obsolete. In addition, existing terms may undergo changes of meaning…" (Kageura K., 1998/1999). Specialized lexicons are made up of the terms which are specific to each field of knowledge, «a subset which is distinct but not separated from the common language» (Cassese, 1992): it is usually difficult to extract the relevant domain-specific terminology, meaning to discern terms which belong to a specialized glossary from those belonging to the common dictionary.

The interest in the study of terminology and the "truth" contained in the above definitions has led us to make a "journey" in the Grey Literature (GL) domain in order to offer an overall vision on the terms used and the links between them. Within this scenario, the work analyzes a corpus constituted of the entire amount of full research papers published in the GL conference series over a time-span of more than one decade (2003-2014) with the aim of creating a terminological map of relevant words in the various GL research topics. "… corpora used to extract terminological units can be further investigated to find semantic and conceptual information on terms or to represent conceptual relationships between terms. (Bourigault D. et al., 2001). Another interesting inquiry is the terminology used in the GL conferences for describing the types of documents which can be detected (Pejšová P. et al., 2012).

2. GL Corpus and method

The work is split up in four sections: creation of the corpus by acquiring the digital papers of GL conference proceedings (GL5 – GL16)[1]; data cleaning; data processing using the described NLP pipeline; terminological analysis and comparison. The corpus - made up of 231 research papers (for a total amount of 785.042 tokens) - was processed using a Natural Language Processing (NLP) tool for term extraction developed at the Institute of Computational Linguistics "Antonio Zampolli" of CNR[2] (Goggi et al. 2015; 2016).

This tool is what is called a "pipeline" - that is, a sequence of different tools - which extracts lexical knowledge from texts: in short, this is a rule-based system tool for knowledge extraction and document indexing that combines Natural Language Processing (NLP) technologies for term extraction.

The NLP pipeline analyzes textual data thanks to generic tools and its result is an annotated text that allows for terminological extraction of relevant concepts.

More in details, these are the steps which it follows:

- transformation of the original document, in our case in Word format, in plain utf-8 format text;
- use of some pre-existing software tools for:
    1. sentence splitting: dividing the text into sentences
    2. word tokenization: splitting sentences into words
    3. lemmatization and morphological analysis (part of speech tagging)
    4. basic syntactic analysis (chunking: dividing the sentence into non recursive constituents)
    5. parsing with the **Ideal** dependency parser, a rule-based system whose specific rules were developed for both Italian and English. This tool was developed specifically for the MAPS project, being the most important part of the NLP pipeline.

6.

The output of the chunking phase produces an intermediate annotated document preliminary to terminology extraction performed by the **Ideal** parser, which relies on rules, written in an ad-hoc language, designed to extract all simple and complex noun phrases in the text[3].

---

* First published in the GL18 Conference Proceedings, February 2017

[1] Kindly authorised from Greynet International, http://www.greynet.org/.

[2] CNR stands for National Research Council, Italy, https://www.cnr.it/.

[3] The extractor works on the chunked text searching for patterns such as nominal phrases (monograms) and nominal phrases followed by one or more adjectival or prepositional phrases (bigrams and trigrams).

Terminological extraction is in turn necessary in order to be able to correctly index the document in the document base to be later semantically searched. The **Ideal** extraction tool takes the chunked text as input, containing all the required morpho-syntactic information.
The output of this terminology extraction pipeline is a set of terms in a standardised *Jason* format.
Within our corpus made of GL articles, this NLP tool – already used as semantic engine within the MAPS project (GL16 and GL17 papers) - extracts a list of single (monograms) and multi-word terms (bigrams and trigrams) ordered by frequency with respect to the context.

3. Terminological analysis

The terminological analysis started with the identification of the monograms of high, medium and low frequency within the glossaries provided by the extraction. This first step gave us an overview of single-terms used in the papers. The study of the terms grouped according to their frequency allowed us to: a) select some of the terms most frequently used; b) examine their co-occurrences; c) determine the variations between them. We continued the terminological analysis with the observation of fragments of taxonomic chains in order to shed light on the usage of specific terms within the topics of the various GL conferences. Through these steps it was possible to monitor the terminological flow and to indicate the resulting lexical trend within the GL domain.

3.1 High, medium and low frequency

For frequency segment of vocabularies we mean the organization of words by decreasing frequencies, starting from the word with $freq_{max}$ and coming to those with $freq_{min}$, usually with only one occurrence (hapax). The occurrences can then be divided into three groups (high, medium and low frequency): in the high segment each word has a different number of occurrences, the limit between the high and medium frequencies being placed immediately above the first parity, that is, the first pair of words that occur the same number of times. To determine the $freq_{min}$ segment and separate it from the mid-range, it is necessary to start from the bottom, i.e. from the hapax, and consider the first gap in the consecutive number of increasing occurrences. After having organized the terms it results that the highest percentage of terms is to be found in the lowest frequency segment: this applies to all GLs'. The GL16 and GL6 glossaries stand out for the substantial amount of terms in the highest segment while the medium segment can be allocated to GL5 followed by GL14.

In Table 1 and Table 2 (Appendix 1) we grouped, respectively, the terms of the highest and medium segment of each GL corpus. The following categories have been excluded from the visualization: adjectives with a semantically low relevance with respect to the context (such as "new", "coastal", "public", etc.); acronyms and generic nomenclatures of bodies, proper names of individuals and institutions.
It was not possible to representing in a table the data with a low frequency given their huge extension; however a section of the lexicon of this segment has been analyzed because it was considered as much relevant.
In Appendix 1, we can read the terms occurring most frequently in the high segment: the only two monograms which consistently remain in this segment and in all GL glossaries are "Literature" and "Research".

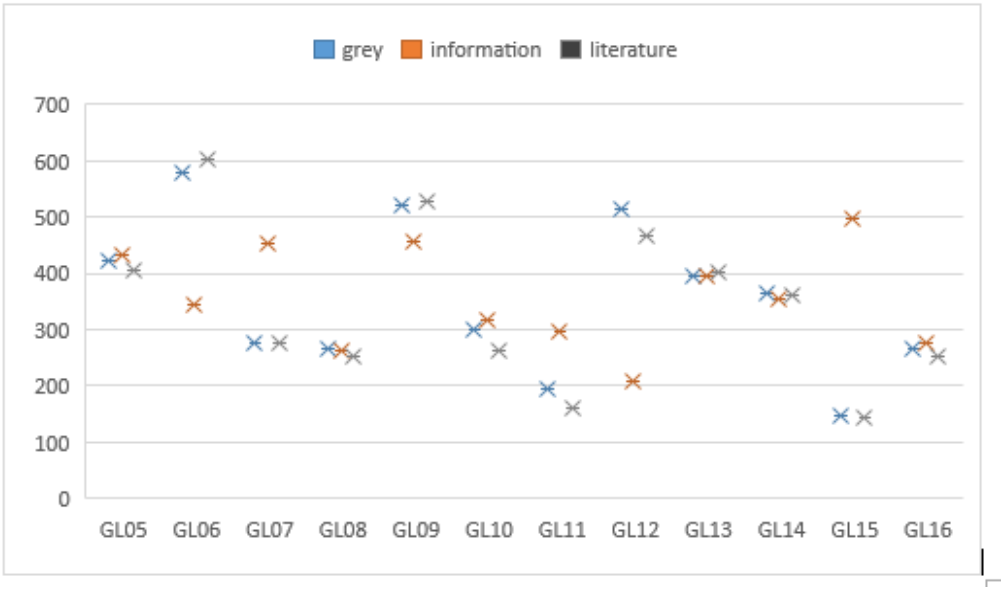
We retrieved words such as "information" or "document" that have a very wide semantic content as well as words closely connected with the specific domain of Grey Literature such as "literature" and "grey". There are also some terms linked to specific documentary categories such as "report", "journal" and "thesis".

Although "Information" is the monogram with the highest frequency in the entire corpus (4302 occurrences) it occupies the second place in the table representing the high segment terms : the first position belongs to "literature", one of the more content-related terms, while "grey" shows 3851 occurrences in the high segment out of a total of 4298 in the whole corpus.
"Grey" appears as monogram also in the following forms: "e-grey", "metagrey", "non-grey", "opengrey". The acronym "GL" occurs 1025 times in the entire corpus; the word "information" appears as monograms also in the following form: "Bioinformation", "Cs-Information", "Cultural/Information", "Data/Information", "Librarians/Information", "Library-Information", "Meta-Information", "Misinformation", "Novel-Information"; and "literature" appears as monogram also in the following forms: "grey-literature-typology" and "sub-literature".

3.2 Mapping

We started the terminological mapping from observing the term that occurs most frequently in the entire corpus: "information" and the two terms more closely related to the context, "grey" and "literature".

Graph 1 shows that the terms "grey" and "literature" have the highest frequency in GL6 (2004) and the lowest in GL15 (2013) while the term "information" has the peak in GL15 (2013) and the bottom in GL12 (2010).



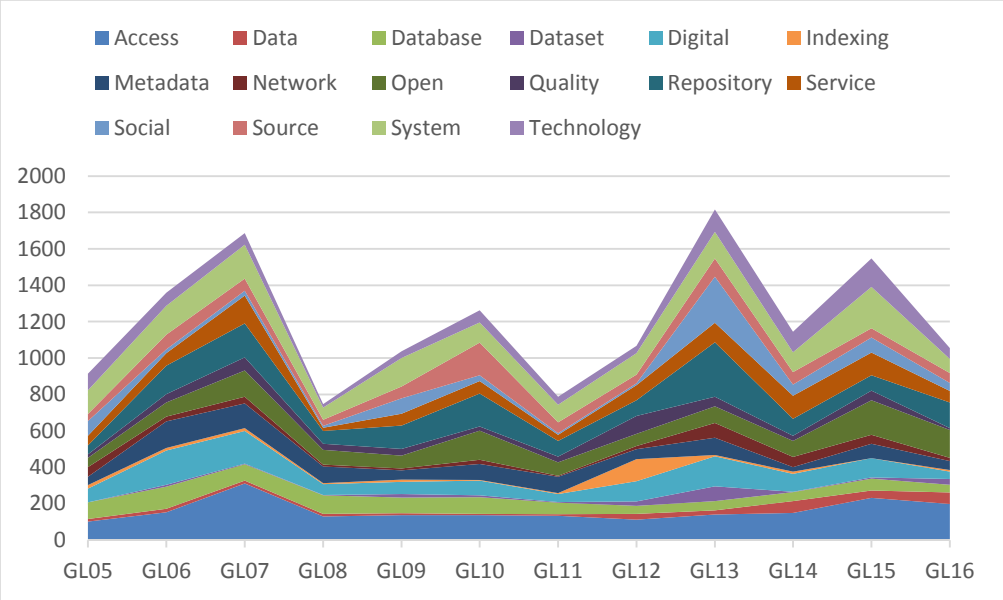
Graph 1 – "Grey", "Information", "Literature" – Trend over the years

As expected, the bigram "grey literature" is the most used with 2816 frequencies in the entire corpus while the bigrams "grey material" (66 occurrences) and "grey document" (98 occurrences) are not present in all GL proceedings and their frequencies are much lower. The bigram "grey documentation" only appears in GL5, GL9 and GL16. Among the other bigrams we find: "grey medical", "grey document", "digital grey", "grey publisher", "grey content", "grey object", "grey resource", "grey collection". Amongst the trigrams we have: "grey literature collection", "grey medical literature", "grey literature community", "grey literature document", "grey literature repository", "grey literature resource", "grey literature material", "grey literature typology", "grey literature report", "digital grey literature", "grey literature field", "grey literature problem".
In addition to the pair "grey literature" the term "literature" appears in the following bigrams: "medical literature", "conventional literature", "type literature", "literature repository", "literature collection", "repository literature", "use literature", "access literature", "journal literature", "literature collection", "conference literature", "trade literature", "definition literature", "literature document", "topic literature", "literature repository", "literature review". The trigram "non conventional literature" is only used in GL7 and GL14 terminology. Excluding the already mentioned trigrams in which "literature" appears associated with grey, there are: "bibliographic control literature", "new generation literature", "scholar information literature", "digital curation literature", "digital repository literature", "strategic collection literature", "literature network service", "web-based dissemination literature", "digital library literature".
The most common bigrams with the term "information" are in GL15: "Information object" is the top term (39 occurrences) while the bottom one is "Information retrieval" (17 occurrences) in GL14. Amongst the others we find: "information system", "source information", "scientific information", "information interaction", "information system", "internet information", "access information".

Looking at trigrams, "Open Source Information" is the top term with 228 occurrences and "Heterogeneous Information Object" the bottom one with 56 occurrences. Others are: "Research Information System", "Information Distribution System", "Public Health Information", "Source Information Product", "Carbon Dioxide Information", "Grey Literature Information", "Scientific Information System".

All the given lists of terms are ordered by descending frequencies.

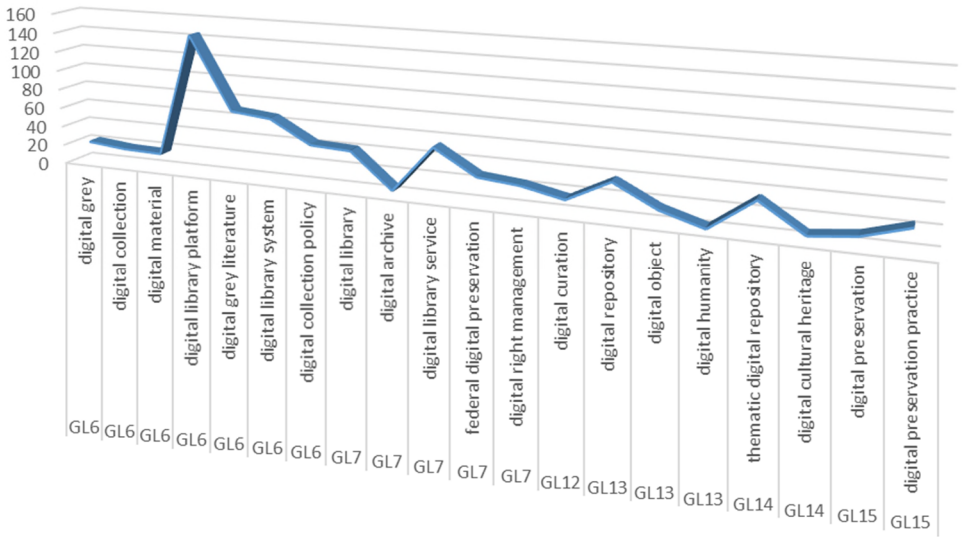
Hereafter the analysis focused on some terms traceable in the three segments: given the dimension of the corpus and the long time-span taken into exam, the terms have been chosen according to their technical connotation with respect both to the context where they are placed and to a very dynamic and cross field, Information and Communication Technology (ICT): "access", "data", "database", "dataset", "digital", "indexing", "metadata", "network", "open", "quality", "repository", "service", "social", "source", "system", "technology".



Graph 2. – Selected terms

Graph 2 shows the trend of the selected terms over the years: it is clear that - with the exemption of "indexing" and "dataset" – all of them are occurring in each GL glossary. Generally, there are monograms which seem to be constantly used and therefore their trend over the time is stable (e.g. access, database and digital) while the vast majority of terms alternate high and low frequency peaks.
The monogram "access" has the highest number of occurrences (1928) and "dataset" the lowest (196); amongst the most frequent terms, also "system", "repository", "open" and "digital" can be found.

Let us start our investigation from one of the most versatile adjectives of the corpus: "digital".
Graph 3 shows the bigrams and trigrams this term form with the several nouns: "digital library" and "digital library platform" are the most recurring Multi-Word Expressions (MWE). The overview provided by the list of selected terms also points out some nouns and verbs which combined with the adjective "digital" - though with relatively low frequencies - disclose the technological nature of the GL community: infrastructure, platform, system, software, network. The MWE "digital humanity" and "cultural heritage" represent entire branches of knowledge whose activities require an expertise crossing from computer science to social and human sciences.
Among bigrams: "digital library" appears in 2005 (GL7). The community does not neglect relevant contents such as "digital preservation" which appears in 2013 (GL15) and even uses the trigram "digital preservation practice". Among trigrams: "digital library platform" has the highest frequency in 2004 (GL6). In most recent years (from 2013 onwards) s such as "digital repository" and "thematic digital repository" replace others like "digital library" and "digital library service" thus revealing new demands for identification, accessibility, interoperability and reuse of the scientific data they host, as well as the need of ad-hoc services for those specific contents.



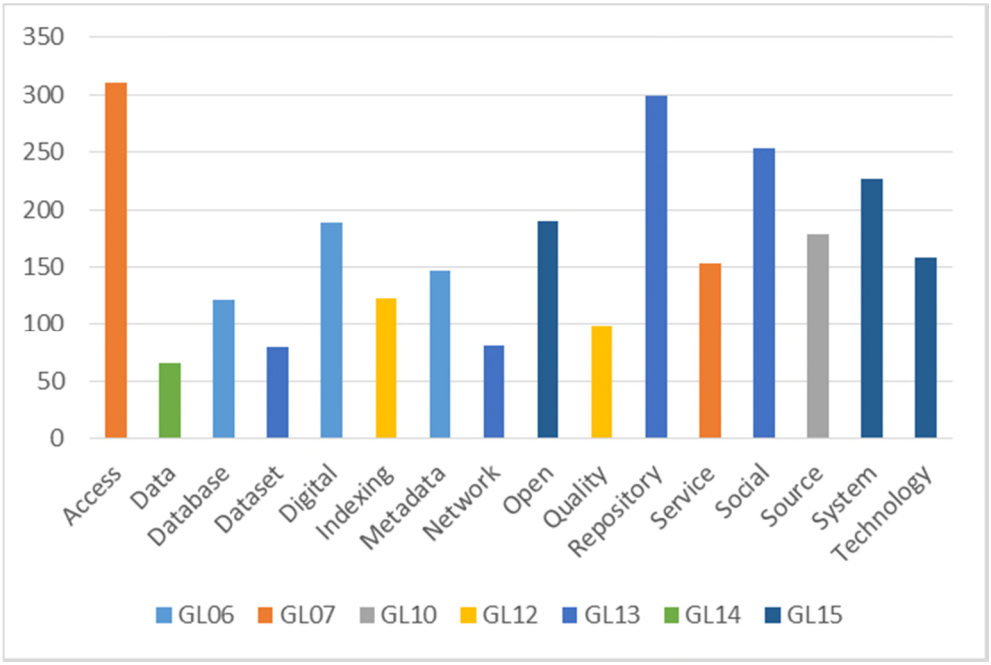
Graph 3. "Digital" – bigrams & trigrams

The term “data” shows the highest frequency as a bigram, “big data”: it introduces the set of problems about gathering, managing, representing and accessing huge volumes of data which are dynamically generated from various sources. The bigram first appears in GL14 (2012) and then again the next year thus witnessing the community’s immediate appeal for the subject; as a trigram is mostly in combination with terms like “discovery”, “service” and “product”.
The term “database” cannot be neglected too: it is used and reused in different contexts as a synonym of an archive of structured and connected data and occurs in the entire time-span associated with various semantic values: "citation databases", "technological databases", "grey literature database". “Database” and “metadata” register the highest number of occurrences in the papers of the GL6 (2004) conference, exactly like “digital”.
The exam of the term “metadata” points out its presence in all GLs in the mid-range segment and already in GL6 (2004) and GL7 (2005) in the high frequency one: these are years when there is a considerable discussion over themes as standards, document management and fast information retrieval systems and the term is often found in association with nouns and adjectives which highlight the importance of properly describing and organizing documentation in the field of digital libraries: “metadata schema”, “navigational metadata”, “descriptive metadata”, “metadata format”, “metadata harvesting” and again “Dublin Core metadata” (the highest percentage), “right management metadata”, “standardized metadata schema” and “metadata quality control”. In GL7 and GL10 “metadata” is often combined with standards and schemes such as Dublin Core and Cerif.
The term “dataset”, in the two variations “dataset” and “data set”, appears in 2005 and remains constantly present in the following editions forming the most frequent bigrams “scientific dataset” and “dataset archive” while is more occasionally associated with “accessibility”, “collection” and “management”.
Already in GL6 (2004) the GL community faces the need to examine the quality of information available on the web: the term “quality” is repeatedly associated with “assessment” or “control”, in particular in the forms “metadata quality control”, “quality assessment metadata”, “quality information”, “quality performance”, “high-quality information” and “metadata quality certification”.
Another interesting term is the adjective “social”. Although we found the topic “Social Networking” only in GL13 (2011), this bigram is in use since GL7 (2005) and the monogram “social” is steadily used in the GL lexicon since GL5 (2003). The adjective “social” is combined with a large number of nouns to form bigrams, trigrams and strings of words with a strong semantic impact. In GL8 the multi-word expression "social network" appears, as a “neologism”, in the GL lexicon. Other linguistic forms emerging from the terminology are linked to the same concept: "virtual social networking”, "social networking tools", "social networking sites", "new social networking technologies". The MWE “social media” was “born” instead in the GL9 conference (2011).

The bigram “open access” which represents one of the most studied research fields in recent years, is a constant feature in the grey literature lexicon. It is in fact used since the far GL5 (2003) in the two graphic variations “open access” and “open-access” that coexist in some GLs’. From the separate analysis of the bigrams formed by “open” and “access” it can be noted that the most frequent is anyway the one which combines them; the monogram “access” then constitutes other bigrams (amongst the others “access information”, “access literature” and “access model”) and trigrams, once again with “open”: “open access model”, “open access repository” and “open access movement”. In order to avoid “open” from the lexical forms taken into exam, the lowest frequencies should be analyzed for finding forms like “sustained access information”, “access datum repository”, “public access resource”. But “open” often creates MWE also with other terms: “open archive”, “open source”, “open repository”, “open model”.
In our context the term “technology” is related to telematics and computer science applications to the documentary field: the single term is paired with “information technology” while the trigram is "technology information system". Information management is represented by nouns such as “system” and “source”: both words are also retrieved in the lexical forms "information system ", "information system database", "electronic sources", "open source repository". A special case is the word “service” which is very frequently used for defining activities for the users of the Internet: “information services", "integrative web services".

3.3 GL Conference topics

The flow of themes discussed in these years at GL conferences is represented by the topics appearing in the twelve Call for Papers (Appendix 2).
Therefore the previous selected terms have been analyzed in relation to the topics of all GL conferences by retrieving the frequency peaks of the chosen terms and then verifying when they occurred.



Graph. 4 – Terms and Topics

From Graph 4 it is clear that the peaks of frequency are limited to certain years: 2004, 2005, 2008, 2009, 2010, 2011, 2012 while the other editions are lacking. The highest frequencies occur in GL7 with the term "access" and in GL13 with "repository" and "social. The word "repository" is never found amongst the topics in its singular form but rather diffusely as "repositories" since GL6 (2004) and then again in 2005, 2006, 2008 and 2009 combined with "collection", "metadata" and "grey literature" for creating "Institutional Repository "and "Grey Literature repository". Again "repository", together with "dataset", "network" and the already mentioned "social", counts the highest number of occurrences in the GL13 papers, where some of the topics were "Social Networking", "Special Collections", "Open Access and Wealth Creation", "Data Frontiers".

The maximum number of occurrences of the terms "digital", "database" and "metadata" dates back to the GL6 (2004) conference which introduced the following topics: "Institutional Repositories", "Use Analysis", "IT & Research Initiative", "Knowledge Management and Dissemination", "Collection Development and Resource Discovery".

In the same year the adjective "digital" registers the highest frequencies with the two forms "digital library" and "digital library platform". It is curious to note that the bigram "digital library" never appears amongst the GLs' topics notwithstanding it is widespread within the articles and, even more curios, the monogram "digital" is never used either. The same for "database" while "metadata" appears only once, in the GL8 Call for Papers.

In GL14 (2012) "data" and "indexing" register their peaks: in this year the chosen conference topics were "Tracing the Research Life Cycle", "Tracking Methods for Grey Literature", "Adapting New Technologies", "Repurposing Grey Literature".

Finally three topics are dedicated to "open access" in GL conferences: "Open Access to Grey Resources", "Open Access and Wealth Creation" and "Open Access to Research Data" (GL16 - 2014).

3.4 Types of documents

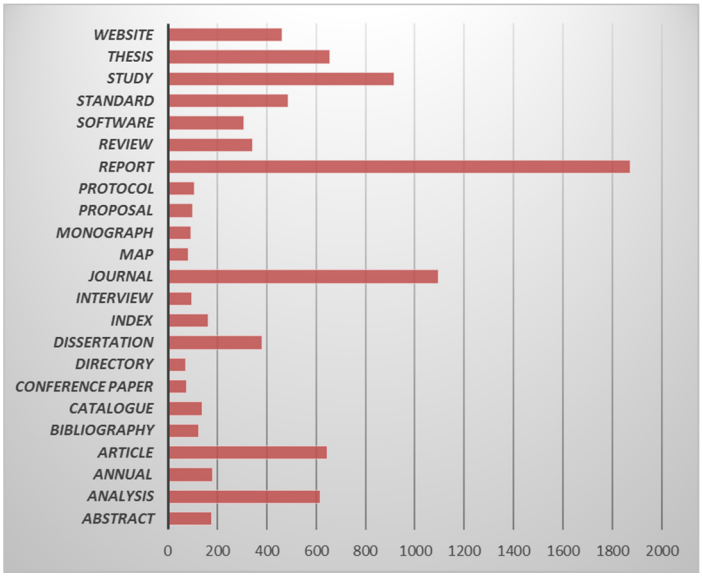
This last chapter is dedicated to the terminology used for describing the types of documents occurring in the corpus.

The analysis of terminology adopted for describing the types of documents started from the entries of the *Vocabulary of the types of Grey Literature (*2011) which has been considered as the reference model. It is though important to take into account the possibility that the terms extracted from the corpus do not necessarily describe the type of GL documents because it was not possible to verify automatically the actual correspondence between the term and its context. An outstanding example is "journal" which can easily refer to the title of a publication.

From this perspective, the presence of the *Vocabulary* terminology within our corpus has been verified: the table in Appendix 3 lists the terms appearing in the various GLs and their quantitative consistency. This table is ordered by frequency and the results – in terms of the most occurring terms - are therefore very clear.

In the attempt of making a partition of this list – however arbitrary - we can circumscribe a first area where the frequencies decrease from 1871 to 307 and the terms retrieved are: "report", "journal,

"study", "thesis", "article", "analysis", "standard, "website, "dissertation, "review", "software". In the intermediate area where frequencies decrease from 196 to 30 (with a remarkable gap between the last occurrence of the first zone and the first occurrence of the last zone) the terms found are the following: "dataset", "annual", "abstract", "questionnaire", "index", "patent", "catalogue", "bibliography", "annual report", "protocol", "proposal", "interview", "monograph", "map", "conference paper", "preprint", "directory", "newsletter", "manual", "bulletin", "curriculum", "poster".
The terms used with the lowest frequency (from 24 to 1) for describing the types of documents are: "brochure", "proceedings", "government document", "glossary", "memorandum", "handbook", "timeline", "announcement", "conference program", "essay", "press release", "chronicle", "leaflet", "course material", "informative material", "normative document", "anthology", "research plan", "syllabus", "tertiary source", "corporate literature", "habilitation thesis", "image material", "legal document", "guidebook", "technical documentation".



Graph 5. – Types of documents retrieved in all GLs

In Graph 5 we can observe that a significant percentage of entries of the vocabulary is found in all GL lexicons as well: "abstract", "analysis", "annual", "article", "bibliography", "catalogue", "conference paper", "directory", "dissertation", "index", "interview", "journal", "map", "monograph", "proposal", "protocol", "report", "review", "software", "standard", "study", "thesis", "website".
At the end of this terminological overview based on the *Vocabulary of the types of Grey Literature* these are the entries of the dictionary which cannot be found in our corpus: "bachelor's thesis"; "call for papers"; "codebook; "conference materials"; "conference proceedings"; "course text"; "exam topics"; "green paper"; "house journal"; "master's thesis"; "minutes"; "product catalogue".

Conclusions

To conclude, this survey on the results of the information extraction process performed by the described NLP tool has been a sort of linguistic path in the past and present of the terminology used in GL proceedings with the goal of drawing a picture of the lexicon used by the GL community and thus contributing to get a deeper knowledge of the GL domain.
Many of the terms encountered cannot have synonyms because they reflect specific concepts devoid of the ambiguities peculiar to the common language. Some expressions such as "grey resources" and "open access" or nouns as "library" and "repository" refer straight and univocally to the "documentary science", that is they belong to a specific semantic field.
By adopting a diachronic point of view, a significant terminological stability can be noticed. However some terms have been pointed out as obsolete while others emerged as very up-to-date, the latter are those chosen for assembling studies in the same domain or even for labeling emerging fields of knowledge. This is the case, for example, of the bigram "electronic dataset" retrieved in 2004 and 2007 glossaries and then substituted by the bigram "digital dataset" in 2010 and 2014.
Examples could be endless but the size of the corpus had made necessary to delimit the study to a sample by choosing some of its parts and pertaining taxonomies.
In these last twelve years we have witnessed the establishment of new paradigms of scientific communication, the stunning development of information technology and the creation of new infrastructures for storing, preserving and disseminating scientific information. A fact clearly comes to

light from this analysis: the grey literature field has a dynamic and cross nature, its community is sensible to technological innovation and proves to be able of keeping pace with the changes.
The lexicon adopted in the GLs' scientific papers has confirmed that the "grey" community soon paid specific attention to topics like "open access", "repository" , "digital objects" and "preservation", just to cite a few. At the same time the almost stable use of a technical and specialized terminology over the time indicates the interest and the willingness to deepen the knowledge of some themes by reporting updates and novelties.
Lastly, this work must be considered a preliminary analysis of the GL corpus, a linguistic resources to be further investigated with different purposes and different tools.

Essential References

1. Bourigault D., Jacquemin C., L'Homme Marie C. (2001). Introduction. In *Recent Advances in Computational Terminology*, VIII-XVIII. Amsterdam, John Benjamins.
2. Cabré M.T. (2000). La terminologia tra lessicologia e documentazione: aspetti storici e importanza sociale, http://web.tiscali.it/assiterm91/cabreita.htm.
3. Cassese S. (1992). Introduzione allo studio della formazione. In «Rivista trimestrale di diritto pubblico» 2, 307-330.
4. Goggi S., Pardelli G., Sassi M., Giannini S., Biagioni S. (2015). A Terminological Survey on the Titles of the Seventh Framework Programme (FP7). In *Proceedings of the Fourteenth International Symposium on Comunicación Social: retos y perspectivas*, vol. I, 223-227. Centro de Lingüística Aplicada, Ministero de Ciencia, Tecnología y Medio Ambiente, Cuba.
5. Goggi S., Monachini M., Frontini F., Bartolini R., Pardelli G., De Mattei M., Bustaffa F., Manzella G. (2015). Marine Planning and Service Platform (MAPS): An Advanced Research Engine for Grey Literature in Marine Science. In *Proceedings of the Sixteenth International Conference on Grey Literature (GL16)*, 108-115. TextRelease, Amsterdam.
6. Goggi S., Pardelli G., Bartolini R., Frontini F., Monachini M., Manzella G., De Mattei M., Bustaffa F. (2016). A semantic engine for grey literature retrieval in the oceanography domain. In *Proceedings of the Seventeenth International Conference on Grey Literature (GL17)*, 104-111. TextRelease, Amsterdam.
7. Kageura K. (1998/1999). Theories of terminology: a quest for a framework for the study of term formation. *Terminology* 5 (1) 21-40.
8. Megerdoomian K. (2003). Text Mining, Corpus building, and testing. In *Handbook for Language Engineers,* 213-268. CSLI Publications, Stanford.
9. Pardelli G., Goggi S., Giannini S., Biagioni S. (2016). Two decades of terminology: European framework programmes titles. In *Proceedings of Tenth International Conference on Language Resources and Evaluation*. *(LREC 2016)*, 373-378. ELRA - European Language Resources Association, 2016.
10. Park Y., Byrd Roy J., Boguraev Branimir K. (2002). Automatic Glossary Extraction: Beyond Terminology Identification. In *Proceedings of the 19th international conference on Computational linguistics (COLING '02)*, http://aclweb.org/anthology/C02-1142.
11. Pazienza M.T, Virdigni M. (2003). Agent based ontological mediation in IE systems. In *Information Extraction in the WEB Era. Natural Language Communication for Knowledge Acquisition and Intelligent Information Agents*, 92-128. Springer, Heidelberg.
12. Pejšová P., Vaska M. (2010). An Analysis of Current Grey Literature Document Typology. In *Book of Abstracts of the 12th International Conference on Grey Literature (GL12)*, 39-47. Text Release, Amsterdam.
13. Pejšová P., Simandlová T., Mynarz J. (2012). A linked data Vocabulary of the Types of Grey Literature: Version 1.0. In *Proceedings of the Thirteen International Conference on Grey Literature (GL13)*, 170-173. TextRelease, Amsterdam.

Appendix 1 – Frequency

High segment													
Term	**GL5**	**GL6**	**GL7**	**GL8**	**GL9**	**GL10**	**GL11**	**GL12**	**GL13**	**GL14**	**GL15**	**GL16**	**Total**
Literature	405	604	277	252	527	263	160	466	403	363	143	254	4117
Information	433	344	455	264	456	317	298	210		355	497	277	3906
Grey	421	579	275	267	520	299	196	515		366	146	267	3851
Research	294	266	314	153	269	250	193	192	403	532	508	223	3597
Document	260	360	392	118	332	143	201	168		155	168	115	2412
Library		299	276	152	188	312	123	267		153	73	91	1934
Access		152	310	130	136	137	133	112		148	231	198	1687
Report	315	193	165	94			161	197				184	1309
Datum								144		358	367	257	1126
System		158	186		156			117			227	76	920
Publication	230	131		107	233						213		914
Repository		157	187		129	181						142	796
Project		183	164	168							271		786
Open			144	80		159					190	153	726
Collection		213	152	96			102	155					718
Journal		139			176			98			153		566
Science					129					141	201	84	555
Digital		188	180					110					478
Material		146				126	109						381
Metadata		147	137	92									376
User		140						114				73	327
Thesis			141			152							293
Citation		153			134								287
Policy		121										116	237
Database		121		102									223
Source		179											179
Technology											158		158
Service			153										153
Development			130										130
Indexing								122					122
Resource				122									122
Quality								98					98
License												91	91

Table 1

	Medium segment												
Term	**GL5**	**GL6**	**GL7**	**GL8**	**GL9**	**GL10**	**GL11**	**GL12**	**GL13**	**GL14**	**GL15**	**GL16**	**Total**
Datum	80	119	125	65	106	106	88		229				918
Project	130				121	76	95	64	139	129		67	821
User	72		90		104	69	86		136	104	122		783
Repository	48			70			86	85	299	94	84		766
Service	54	69			65	69		84	106	126	125	63	761
Development	93	95		62	61	70	47	63	87	79	101		758
Digital	75			60	66	80	44		166	99	106		696
Collection	97				48	109			198	75	67	69	663
System	130			68		112	96		146	108			660
Science	141	85	64	46		63	53	96	107				655
Resource	36	83	130			60	60		87	112	82		650
Technology	91	73	64			66	45		124	113		61	637
Web	124	84		51	51	97	43		55	87			592
Database	92		90		86	91	64		51	50	65		589
Social	81				85				254	62	82		564
Report					95	106			116	117	128		562
Material	107		82	66	95				56	77	75		558
Process	57	63	110					60	57	107	88		542
Source	39	80	68		65		60		100	69		55	536
Knowledge	62		51	51			39		87	107	138		535
Open	51	78			70		74	67	92	88			520
Community	38		68		78		40		97	109	85		515
Management	52			64				67	54	87	104	69	497
Publication			100			94	48	60	66	120			488
Archive		67	94			116	56		93	43			469
Article		86	87	53	97					52	88		463
Library	158								221		82		461
Format	55	82	68			47			62	44	99		457
Electronic	66	65	85		60	68				44	66		454
Metadata	46				51	88	91		95		79		450
Journal	92		92	67						115		61	427
Institutional	53	69	90			48			101			57	418
Grey									394				394
Survey	71			53	53	69			50	74			370
Academic	72	72	58		60	48						55	365
Communication	65				94				108			54	321
Policy	69		73	68		48				53			311
Online	50	51	51						56	42	54		304
Standard	46	60	68							57	58		289
Citation	143		54	63									260
Access	101								140				241
Health		63	110	58									231
Education	55					65				107			227
Dissertation	45				48	107							200
Model	59		74								63		196
Environment	71								54	70			195
Network	53								81		55		189
Thesis	39	65			85								189
Product	56									129			185
Government		64				56	63						183
Production	90					48				43			181
Website	43									57	81		181
Life					121					58			179
Quality			72						53		54		179
Document									176				176
Book	49		93										142
Documentation	140												140
Data										65		64	129
Practice										47	77		124
Copyright						122							122
Bibliographic	60						53						113
Innovation										113			113

Term	GL5	GL6	GL7	GL8	GL9	GL10	GL11	GL12	GL13	GL14	GL15	GL16	Total
Right							49					62	111
Internet	62									45			107
History				104									104
Security			54	46									100
Discipline					49					46			95
Tool	38			57									95
Review								93					93
Risk									83				83
Patent											82		82
Dataset									80				80
Blog										79			79
Guideline											74		74
Evaluation			73										73
Reactor		66											66
Commercial	62												62
Environmental	62												62
Networking									60				60
Multimedia						59							59
European	58												58
Law							58						58
Application											57		57
Structure			56										56
Corpus										55			55
Training	49												49
Workflow										48			48
Traditional						47							47
Conventional					45								45
Several	44												44
Dissemination	42												42
Engineering	41												41
Magazine	41												41
Protection	41												41
World	41												41

Table 2

Appendix 2 – GL Conference topics

GL	Conference topics
GL5	Models for Academic Grey, Part I: Specific Approaches
GL5	Research is Grey Dependent
GL5	The Economy of Grey
GL5	Strategies for Academic Grey, Part II: General Approaches
GL5	Search Engines are Growing Grey
GL5	Roadmap of Grey Literature Systems and Services
GL5	Alternative Issues in Grey Literature
GL5	Product and Service Reviews
GL6	Institutional Repositories
GL6	Use Analysis
GL6	IT & Research Initiative
GL6	Knowledge Management and Dissemination
GL6	Collection Development and Resource Discovery
GL7	Curriculum Development and Research On Grey Literature
GL7	Theses and Dissertations
GL7	Repositories and Collections of Grey Literature
GL7	Quality Assessment of Grey Literature
GL8	Collection Development, Collection Policies, and Collection Rescue
GL8	Metadata Schemes, Repositories, Software, and Standards
GL8	Curriculum Development and Grey Literature
GL8	Metadata Schemes and Repositories for GL
GL8	Quality Assessment of Grey Literature
GL8	Economic and Legal Aspects of Grey
GL8	Mapping Grey Resources for Costal and Aquatic Environments
GL9	Grey Foundations in Information Landscape

GL	Conference topics
GL9	Tools for Publishing, Archiving, and Accessing GL
GL9	Use and Impact of GL in Scholarly Communication
GL9	Information Walk-Thru, Poster Presentations & Product and Service
GL9	Grey Literature in Central and Eastern Europe
GL9	New Discoveries in GL for Research Communities'
GL9	Education and Grey Literature
GL9	Information Walk-Thru Poster Presentations, P&S Review
GL10	Institutional Repositories and Grey Literature
GL10	Grey Literature in Biomedical Communities
GL10	Legal Aspects, Intelligence, and Text Mining In Grey Literature
GL10	Grey Literature in Research
GL11	Impact of Grey Literature on Net Citizens
GL11	Uses and Applications of Subject Based Grey Literature
GL11	Grey Literature Repositories
GL11	Open Access to Grey Resources
GL12	Redefining Grey Literature
GL12	New Stakeholders in Grey Literature
GL12	Standardization in Grey Literature
GL12	New Frontiers in Grey Literature
GL13	Social Networking
GL13	Special Collections
GL13	Open Access and Wealth Creation
GL13	Data Frontiers
GL14	Tracing the Research Life Cicle
GL14	Tracking Methods for Grey Literature
GL14	Adapting new Technologies
GL14	Repurposing Grey Literature
GL15	Technology Assessment
GL15	Sustaining Good Practices
GL15	Research and Data
GL15	Towards Informed Policies
GL16	Public Awareness of Grey Literature
GL16	Publishing and Licensing Grey Literature
GL16	Open Access to Research Data
GL16	Managing Change in Grey Literature

Table 3

Appendix 3 - Types of documents

Vocabulary	GL5	GL6	GL7	GL8	GL9	GL10	GL11	GL12	GL13	GL14	GL15	GL16	Total
Report	315	193	165	94	95	106	161	197	116	117	128	184	1871
Journal	92	139	92	67	176	44	24	98	35	115	153	61	1096
Study	111	96	67	23	102	74	33	66	83	75	110	77	917
Thesis	39	65	141	9	85	152	12	31	33	14	51	25	657
Article	29	86	87	53	97	24	30	42	26	52	88	32	646
Analysis	57	65	33	27	46	38	45	75	58	77	47	48	616
Standard	46	60	68	19	37	24	21	33	48	57	58	16	487
Website	43	29	20	14	31	38	30	46	41	57	81	32	462
Dissertation	45	21	36	4	48	107	7	25	35	12	25	17	382
Review	21	47	32	18	17	10	4	93	35	30	19	15	341
Software	15	30	48	22	15	35	13	21	28	25	49	6	307
Dataset		10	3	2	20	11	4	25	80	2	7	32	196
Annual	21	2	7	6	8	16	19	76	11	8	5	3	182
Abstract	18	24	24	6	13	11	22	9	12	18	9	10	176
Questionnaire	1	15		19	16	34	13	8	15	24	19	1	165
Index	32	32	16	11	16	8	4	7	12	8	6	11	163
Patent	16	3	9	2	6	5		3	7	14	82	11	158
Catalogue	21	18	20	5	23	22	5	5	3	2	9	4	137
Bibliography	4	15	8	2	10	30	28	1	1	6	18	3	126
Annual Report	7		6	3	3	12	4	66	1	4	1		107
Protocol	13	28	15	3	4	6	12	3	3	5	12	2	106
Proposal	26	16	10	5	7	7	1	6	5	6	5	7	101
Interview	9	16	6	5	5	4	12	2	4	14	15	3	95

Vocabulary	GL5	GL6	GL7	GL8	GL9	GL10	GL11	GL12	GL13	GL14	GL15	GL16	Total
Monograph	23	3	9	2	9	3	2	2	3	25	8	4	93
Map	7	8	2	7	2	8	12	3	9	5	13	6	82
Conference	4	17	9	1	7	7	12	5	3	4	4	3	76
Preprint	12	10	10		9	6		4	16		2	4	73
Directory	4	6	10	7	6	7	2	1	12	7	1	7	70
Newsletter	22	15	4	5	9	3			5	4	1		68
Manual	5	4	1	5		9	7	15	2	6	2	1	57
Bulletin	13	3	2	3	3	4			1	1	2	2	34
Curriculum	3		4	2	6	2		7	2	4		1	31
Poster	1		1	2	1	1	5	7	2	1	5	4	30
Brochure	11		2	2	4	1	1	2			1		24
Proceedings	1	6	1		1	1		1	6	4	1		22
Government	1	5		2	4	1	3	2				1	19
Glossary	1	1	1			1	1		2		6	3	16
Memorandum			2			1	5	1	1	1	1		12
Handbook	2		2	3	1			1		1	1		11
Timeline			3	2					2	1		3	11
Announcement	1				3			1	2		2	1	10
Conference					1			1	5		1	1	9
Essay	2				1	2	2		1	1			9
Press Release			1			5					1	1	8
Chronicle			2						2	1		1	6
Leaflet					3	1		1				1	6
Course	2						1	1	1				5
Informative									1		3	1	5
Normative					2	1					2		5
Anthology						1	1		1	1			4
Research Plan									1	1	2		4
Syllabus					3								3
Tertiary	1				1			1					3
Corporate											2		2
Habilitation									1		1		2
Image Material											2		2
Legal							1				1		2
Guidebook		1											1
Technical		1											1

Table 4

International Nuclear Information System
INIS
organizing the world's information on nuclear science and technology and making it universally accessible for peaceful uses
over 150 Member States and international organizations
millions of citations and abstracts published worldwide
hundreds of thousands of full text non-conventional 'grey' literature
multilingual thesaurus in Arabic, Chinese, English, French, German, Japanese, Russian, Spanish
www.iaea.org/inis
IAEA
International Atomic Energy Agency

https://doi.org/10.26069/greynet-2025-000.534

Managing Diversity in the International Nuclear Information System*

Dobrica Savić (Austria)

Abstract

Although diversity is defined as the state of having many different forms, types, ideas or properties, most often it is associated with cultural diversity or different ethnic backgrounds. Biology, religion, and political science researchers have their own view of diversity, such as biodiversity (i.e. variety of life on earth); religious pluralism (i.e. multi-confessionalism, multi-faith), or various shades of democracy (i.e. freedom of opinion or expression). Diversity is also applicable to information management, usually through the diversity of information resource formats and the variety of information users (i.e. customers or clientèle).

This paper adopts a holistic approach to information management, in particular to the management of scientific and technical information (STI), as carried out by an international repository, and attempts to identify various intrinsic and extrinsic properties which include elements of apparent diversity. In addition to the above mentioned diverse information formats and the variety of information users, intended and actual, this paper also examines the diversity of subject content, temporal distribution, geographic coverage, variety of distribution channels, search paths, and composition of the staff directly involved with the running and management of an information repository.

The data used in this research comes from the International Nuclear Information System (INIS), which has collected, processed and provided access to more than 4 million bibliographic references of publications, documents, technical reports, non-copyrighted materials, and other grey literature, as well as over a million full-texts. The INIS repository represents one of the world's largest collections of published information on the peaceful uses of nuclear science and technology. The history of INIS throughout the last 47 years represents a wealth of information on the successful management of diversity, starting with 154 member states and international organization who share and allow access to their valuable nuclear information resources while preserving them for future generations, offering nuclear information repository used annually by millions of scientists, researchers, engineers, technicians, students, managers, and government employees.

***Keywords**: nuclear information; information management; document repository; diversity; INIS; IAEA*

Introduction

The main topic of this paper is leveraging diversity in grey literature. However, as we can see from the title, there are two major concepts here put into one topic. The first one is ***diversity*** and the second one is ***grey literature***. Taking this into consideration, the first part of the paper concentrates on the review of conventional diversity, while the second part focuses on information diversity, more specifically on diversity related to grey literature and the way it is being managed in the International Nuclear Information System (INIS)[1].

Conventional diversity will be presented as a concept, followed by a review of challenges, and culminating with a look at values and benefits. The review of diversity in INIS will start with identifying various aspects of diversity, followed by diversity-based results, and concluding with ways that diversity is being managed.

In closing, some general observations and conclusions will be drawn. It is expected that examples given throughout will demonstrate potential strength and benefits that can be achieved through diversity.

***In diversity there is beauty and there is strength** - Maya Angelou*

Grey literature

There are several definitions of grey literature, the most common being the so-called "Luxembourg definition," which was discussed and approved during the Third International Conference on Grey

* First published in the GL18 Conference Proceedings, February 2017

[1] The International Nuclear Information System (INIS) hosts one of the world's largest collections of published information on the peaceful uses of nuclear science and technology. INIS is operated by the International Atomic Energy Agency (IAEA) in collaboration with 154 Member States and international organizations. There are over 4 million bibliographic references to publications, documents, technical reports, non-copyrighted documentation and other 'grey literature', as well as 350,000 full-texts. INIS offers free and open online access to this unique collection of non-conventional literature through its search application (http://inis.iaea.org/search/).

Literature in 1997: "[Grey literature is] that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers." In 2004, at the 6th Grey Literature conference in New York, a postscript was added for purposes of clarification "...not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body" (Schöpfel, 2010). As a synonym for grey literature, INIS uses the term 'non-conventional literature' or NCL, which will be mentioned later in this paper.

Diversity concept

Diversity is usually defined as the state of having many different forms, types, ideas or properties[2]. It encompasses acceptance and respect.

Diversity is most often associated with cultural diversity or different ethnic backgrounds. Keeping in mind, diversity also refers to human qualities that are different from our own and those of groups to which we belong; but that are manifested in other individuals and groups.

Biology, religion, and political science researchers have their own view of diversity, such as biodiversity (i.e. variety of life on earth); religious pluralism (i.e. multi-confessionalism, multi-faith), or various shades of democracy (i.e. freedom of opinion or expression).

There are many papers written on these types of diversity, but not so many on 'information diversity'. For example, Google Search offers only 42,000 hits for 'information diversity' and over 11 million hits for 'cultural diversity'.

Diversity is also applicable to information management, usually through the diversity of information resource formats, the variety of information users (i.e. customers or clientèle), or a number of different information services.

Diversity challenges

Diversity comes with a number of inherited challenges. The most common ones are stereotypes, bias, generalizations, beliefs and resistance to change.

A ***stereotype*** is a generalized view or popular belief about a particular group of people, usually inaccurate and sometimes offensive (Kanahara, 2006). As the picture on the right shows, we can easily assume the nationality of the character, its characteristics, and other attributes. However, that might be completely wrong and coincidental. A very important lesson here is to connect with the individual, not with the stereotype and not by applying a label, whether positive or negative. Stereotyping is most often related to people or nations, but it can also be related to ideas and things. It should be kept in mind that stereotyping is not necessarily something wrong. It could be quite useful, since it allows us to condense information and make quicker decisions.

Another example of stereotypes is age. This is a big challenge, especially when dealing with and trying to benefit from diversity in the work place. Some research suggests that age related stereotypes bring particularly unfavourable consequences for women and minorities (Shultz & Adams, 2007). A common stereotype related to age is that older generations have poor IT skills, which is often not correct. Being 'too young' or 'too old' is often seen as negative, rather than an opportunity to bridge a gap between generations. Gender is another example of stereotypes, reflected in unequal pay, different career development opportunities, female vs. male jobs, etc.

Age stereotype

- Less motivated
- Not interested in training
- Resistant to change
- Not trusting
- Having health problems
- Vulnerable to work-family conflicts

(Ng & Feldman, 2012)

Bias and stereotype are somewhat related. While stereotype is a preconceived idea that attributes certain characteristics to all the members of a class or set, bias is an individual or personal preference[3]. For example, a statement such as - I think that grey literature has no academic or scientific merit - is a stereotype. However, if we have in front of us two equally valid documents, one published in a journal and the other in a personal blog, and we decide to use only the one from the journal – that is bias.

[2] Merrian-Webster Online Dictionary. https://www.merriam-webster.com/dictionary/diversity

[3] A list of cognitive biases can be found at http://rationalwiki.org/wiki/List_of_cognitive_biases

A ***belief*** is an internal feeling that something is true, even though that belief may be unproven or irrational. For example, I believe that tweeting is the best way to quickly exchange information, although that is probably far from the truth. However, changing beliefs is a very difficult task. Often, in order to change beliefs, people need to change their experiences first.

Resistance to change and resistance to diversity are attributed, in part, to the natural fear of change, which often brings with uncertainty, discomfort and loss of control. There are some who outright oppose the concept of diversity (e.g. We've always done it this way), but in practice, this resistance usually manifests itself as a challenge expressed at the very beginning of a diversity based project or program. Some examples of diversity resistance include:

- discrediting information provided by people with diverse backgrounds;
- unwillingness to acknowledge and recognize the contributions of people with diverse backgrounds;
- delaying consideration/implementation of diversity issues;
- rejecting diversity as being too time consuming or complex.

In addition, the frequency of change undergone by a person at the work place is connected to the frequency of the negative emotions reported, which leads to a lowering of trust, lack of engagement, and a decrease in performance (Rafferty & Griffin, 2008). In order to achieve efficient change, all involved need to fulfil the commitments undertaken and not have hidden interests (Grama & Todericiu, 2016).

Diversity values and benefits

It is evident that diversity has very strong ethical, personal, business and social implications.

Ethics – a code of moral principles that set standards of conduct about what is right and wrong, moral and immoral, fair and unfair, and proper and improper in one's behaviour (Seid & Venkataram, 2016). Globalization and an increasingly interconnected world have created tension between universal ethics and local values and norms. What one culture perceives as an ethical action, another culture may not. Organizations need to invest resources and time to train their staff and leaders to properly manage diversity and make ethical decisions. Diversity training is often used as a tool to improve current situations in an organization.

From a ***personal*** aspect, diversity can create curiosity. By being exposed to different people and cultures, we may want to learn more about a particular group which can give us insight into how and why things work as they do and possibly a way to change them for the better.

In ***business***, employing a diverse workforce enables the of use a wider range of talents and skills. This leads to creativity and innovation. Businesses need to mirror the communities and cultures in which they work in order to understand and anticipate the diverse needs of their customers. In ***society***, diversity brings richness and variety. There are always new and interesting things to learn from each other.

Diversity at INIS

When dealing with aspects of diversity within INIS, a number of facets come to mind. Namely, diversity can be related to subject content, information formats, temporal distribution, geographic coverage, distribution channels, and also to the representation of INIS staff. Striving on its diversity, INIS throughout its history has achieved a number of remarkable milestones, marking its commitment to the efficient and effective collection, processing and dissemination of nuclear information, as well as its contribution and support for open access to scientific information, with global benefits.

Subject content of the INIS repository covers around 50 well defined categories (IAEA, 2010) which are regularly maintained by INIS, and provides scope descriptions used by national and regional centres to categorize nuclear literature for INIS input. The INIS Joint Reference Series publications are also available on the INIS website.

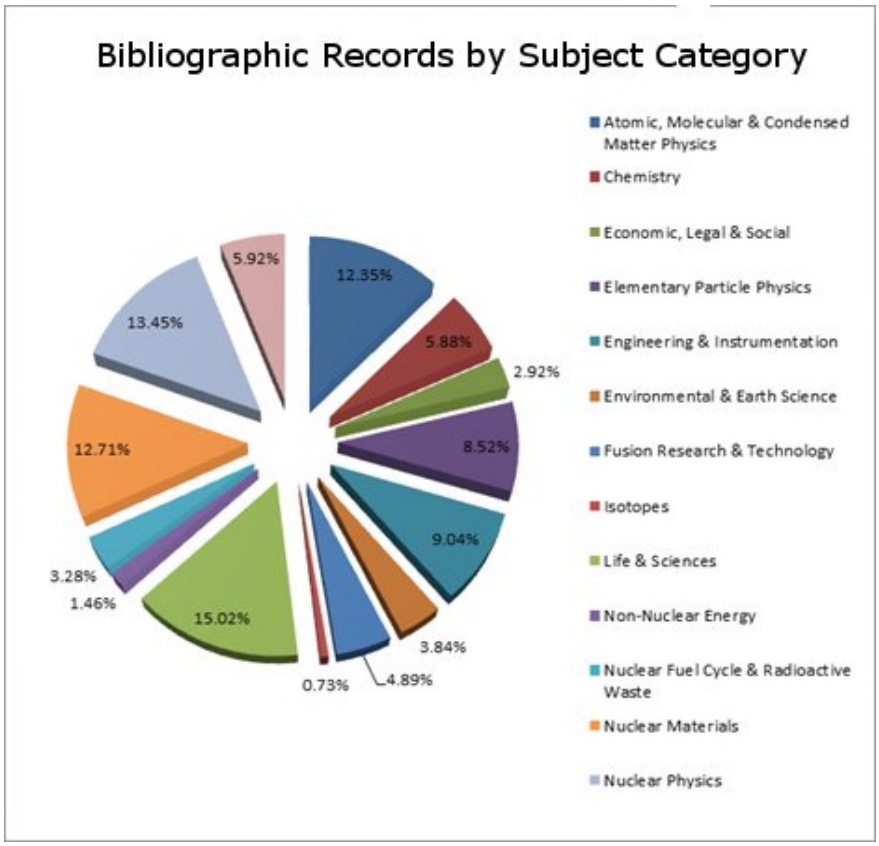


Figure 1: Bibliographic Records by Subject Category

The INIS repository covers all aspects of the peaceful uses of nuclear science and technology, such as nuclear reactors, reactor safety, nuclear fusion, applications of radiation and radioisotopes in medicine, agriculture, industry and pest control, as well as related fields of nuclear chemistry, nuclear physics and materials science. Special emphasis is placed on the environmental, economic and health effects of nuclear energy. Legal and social aspects associated with nuclear energy are also covered. Figure 5 lists a complete set of INIS Subject Categories.

Regarding ***information formats***, the INIS repository consists of seven types of literature – computer media, patents, books, reports, journal articles, miscellaneous and audio-visuals. The main purpose and goal of INIS is to collect non-conventional or grey literature on nuclear science and technology from around the world. This includes various technical reports, project documents, feasibility studies, government reports and any other publications which are not commercially available. However, as shown in the figure below, the INIS repository also contains a number of other forms of conventional literature. Still, it should be noted that while the INIS repository includes links to commercially available journal articles, a great majority of full-texts are, in fact, grey, or non-conventional literature.

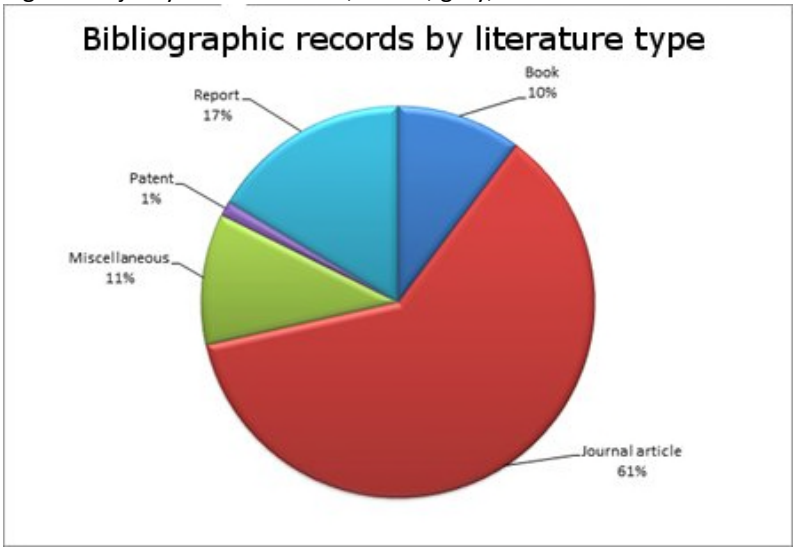


Figure 2: Bibliographic Records by Literature Type

INIS repository

Conventional literature 71%
- Journal articles 61%
- Books 10%

Non-conventional Literature 29%
- Reports 17%
- Miscellaneous 11%
- Patents 1%

Temporal distribution of records in the INIS repository is another aspect of its diversity. As of 31 December 2016, the INIS repository contained 4,000,361 bibliographic metadata records. Of those, 1,093,004 are full-text documents, 761,910 of which are directly available from INIS, with the remaining 331,094 NCLs available from other sources. Only a small portion of the full text documents is restricted and kept for internal use.

During the last 10 years, INIS has added to its repository over 120,000 bibliographic records and 13,000 full text PDF documents annually. The repository is accessible from the INIS website (www.iaea.org/inis).

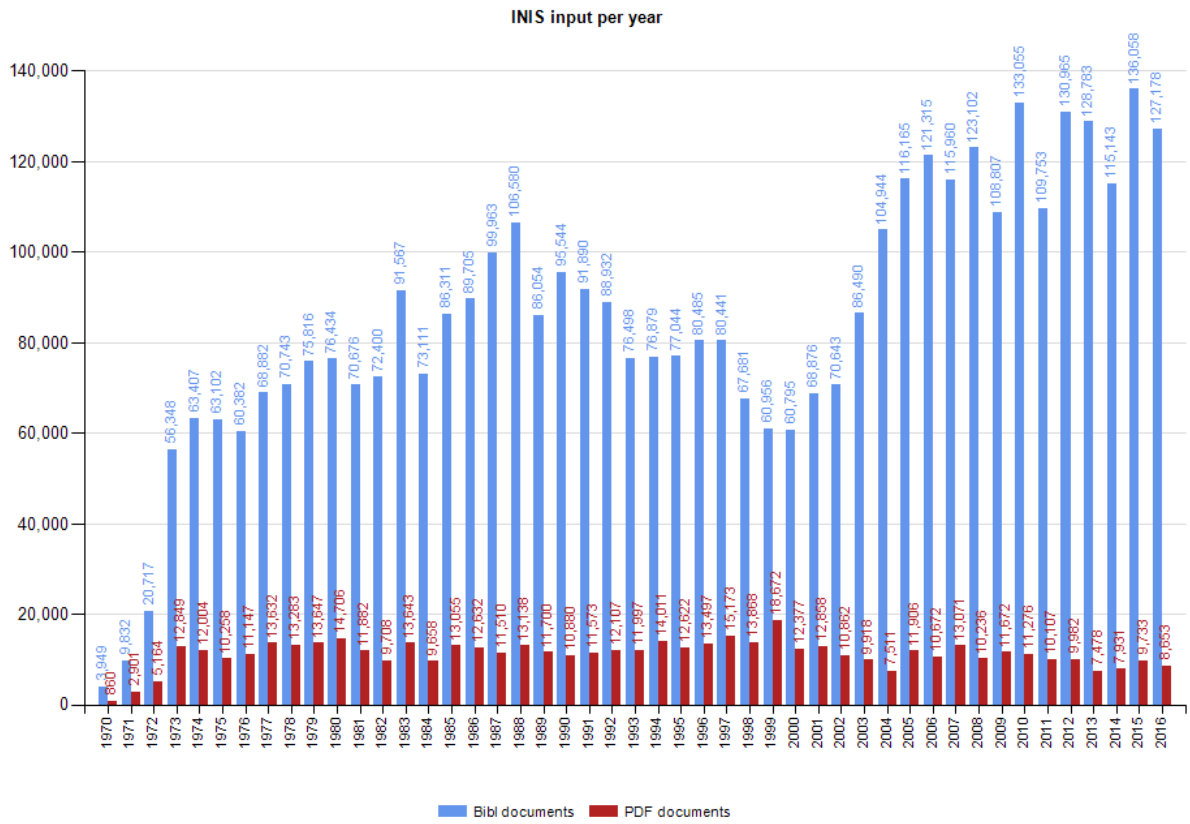


Figure 3: INIS Input per Year

INIS ***geographical coverage*** is almost universal. Currently it includes 150 countries and 24 international organizations. The figure below shows the growth of INIS membership since its creation in 1969.

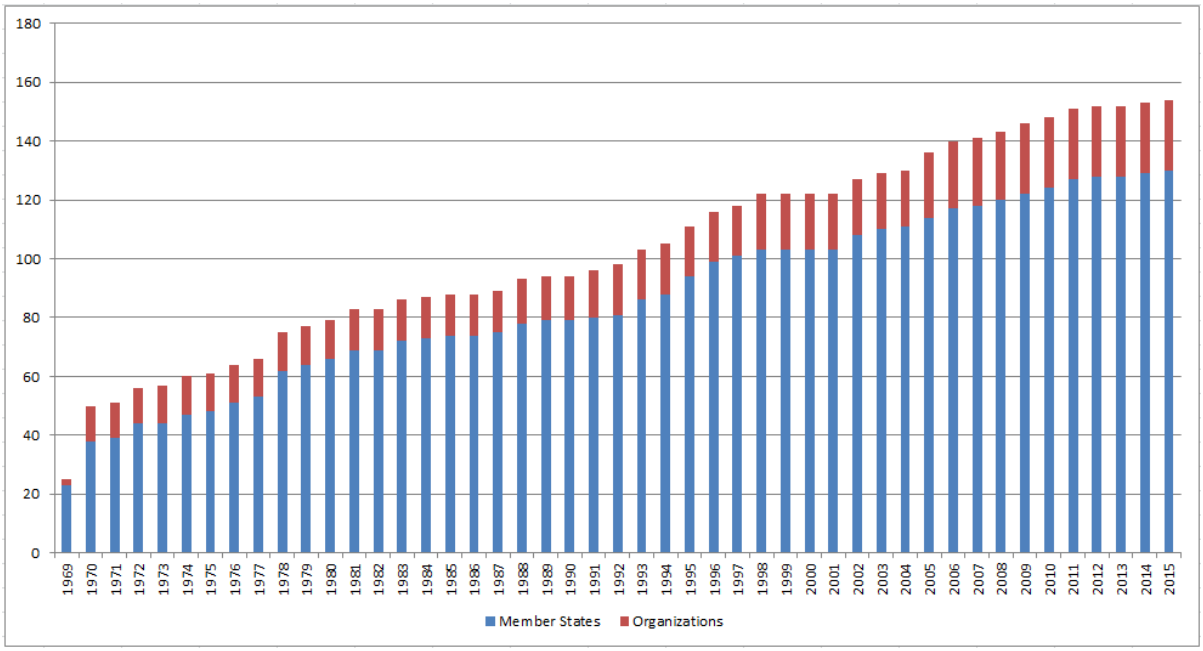


Figure 4: INIS Membership Growth

Besides the geographical coverage of input, that is much diversified as shown above, access and usage of the INIS repository is even more diversified. According to INIS statistics obtained from Google Analytics, users from 229 countries and territories (as defined by Google) accessed the INIS repository in 2016. A list of the 10 countries with the most visits is given below.

	Country	Sessions	% Sessions
1.	United States	233,928	18.06%
2.	India	137,130	10.59%
3.	United Kingdom	68,701	5.30%
4.	Japan	55,635	4.30%
5.	Germany	51,724	3.99%
6.	South Korea	47,543	3.67%
7.	Canada	43,037	3.32%
8.	France	41,748	3.22%
9.	Brazil	35,167	2.72%
10.	Iran	33,513	2.59%

Figure 5: Visits to INIS

Distribution channels for information and documentation available from INIS include the INIS repository search, which is the main channel, as well as Google and Google Scholar, WorldWideScience.org and various widgets developed for use by outside websites.
IAEA ***staff*** represent a world-class, diverse and multi-disciplinary workforce of more than 2,500 employees working in a wide range of nuclear and related disciplines, coming from over 100 countries. The Nuclear Information Section of the IAEA is made up of 30 staff members from 17 countries and 5 continents.

Since its creation in 1969, INIS progress has been marked by a number of very important ***milestones***, which have left a mark on INIS, on the work of the IAEA, and on the increasing importance of grey literature and nuclear information. Some major milestones are listed below.

1970 INIS - the first nuclear database at the IAEA
1979 INIS - the first IAEA database with online access
1991 INIS & AGRIS - the first IAEA databases on CD-ROM

1996 INIS Web Site - the first Web Site at the IAEA
1998 INIS - the first IAEA database available on the Internet
2009 All Internet users given free and open access to the INIS database
2009 INIS database becomes accessible via WorldWideScience.org
2010 INIS offers the first repository search widget
2011 INIS launches new web search interface using Google-based technology
2012 INIS multilingual Thesaurus integrated with the INIS repository search
2012 INIS repository search includes the IAEA Library catalogue
2013 Browse INIS repository by Subject Category
2014 INIS repository becomes searchable through Google and Google Scholar
2015 Launched Open Nuclear Information eXchange System (ONIXS)
First harvesting program based on Open Access Interface-Protocol for Harvesting Metadata (OAI-PMH)

In 2009, INIS announced that access to the INIS database was open to Internet users around the world. Free, unrestricted and ***open access*** was made available from the INIS Homepage. This provided easy access to reliable nuclear information on the peaceful uses of nuclear science and technology, including nonconventional literature, and made nuclear knowledge readily available worldwide.

Managing diversity at INIS

Managing diversity is a complex process with a number of practical requirements and challenges. Not every element can be fully and successfully addressed, so it is important to take a strategic approach. With INIS, the most important element is international collaboration, followed by the successful diversification of formats, services, and ways to collect bibliographic information and the full-text of related nuclear documents. Demographic diversity is already a given since INIS is part of the United Nations (UN). Everything the UN does is based on two fundamental principles: respect for people and continuous improvement.

The impact of diversity on performance was studied by researchers who identified two impacts: the impact on people and the impact on organizations. According to the cognitive resource diversity theory, which advocates that the cognitive resources of each team member contribute to the overall success of the team, a diversity of cognitive resources promotes creativity and decision making capacity (Simons & Rowland, 2011).

Another group of authors (Cox & Blake, 1991) proposed that cultural diversity impacts six direct aspects of organizational effectiveness, including the capability to attract human resources, cost, the issues of innovation, creativity, and problem solving, the marketing advantages of a diverse workforce, and organizational flexibility. My long experience working with the United Nations, and particularly with the IAEA and INIS, tells me that both aspects have had a positive impact on the success of INIS.

Empowering staff members is much easier than rewarding them in institutions, such as international organizations, since flexibility for rewards usually ends with non-monetary awards and praise. Empowering is encountered more often, usually by encouraging staff to demonstrate their interests, propose new ideas, innovate, be creative, and feel in charge or fully responsible for a particular task. Appreciation often shows remarkable results.

Flexible work environments have a real impact on morale, giving employees a feeling of being empowered and trusted to manage their own workloads. Research shows that flexible work hours are related to higher organizational commitment and job satisfaction (Scandura & Lankau, 1997).

Training and mentoring practices in diversity management should continue, to ensure that there is continuous learning and adaptation. Diversified learning that encompasses technical, as well as soft skills, is important for overall success and special attention is given to that, too. It should be mentioned that besides training its own staff members, INIS regularly offers training to its Member States in the form of a training seminar or in the form of ad hoc and e-learning.

Communication is a never ending story. It includes internal, as well as external means and targets. There are many different ways of communicating today, and INIS is using a good variety of them.

Conclusions

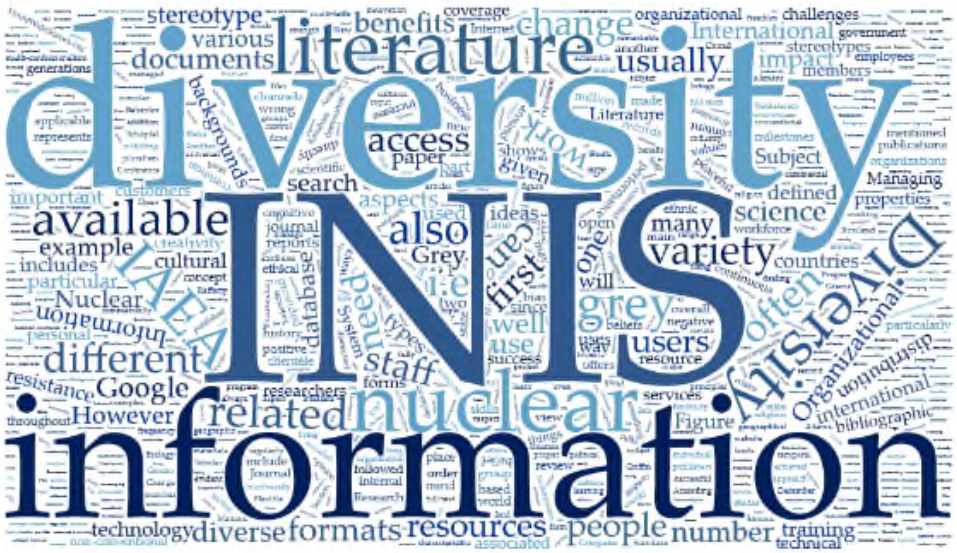
Diversity, defined as the state of having many different forms, types, ideas or properties, is most often associated with cultural diversity or a variety of ethnic backgrounds. However, it is also applicable to information management, usually through the diversity of information resource formats and the variety of information users (i.e. customers or clientèle). INIS represents a very good example of this and its long history of success is proof of that.

Diversity in the work place, and in information management in particular, should be regarded as positive. There is no need to fear diversity. Rather, it should be embraced and used to our benefit. To

obtain maximum benefits from diversity, we need to invest time and effort to get to know our customer's needs and to diversify our information base and services. In other words, we need to leverage all available information resources.

It is also evident that diversity, in itself, does not directly bring about benefits. We need to engage in getting those benefits and we need to promote information services, the use of available information resources, and increase visibility and accessibility, particularly of grey literature, as it is not the type familiar to all users. This could be done through continuous training, education and persistent efforts to increase information literacy.

In summary, proper use of diversity in information management can increase the use of available information resources and therefore, impact creativity and improve overall productivity and staff attitudes.



Bibliography

Cox, T. H., & Blake, S. (1991, August). Managing Cultural Diversity: Implications for Organizational Competitiveness. *The Executive, 5*(3), 45-56.

Grama, B., & Todericiu, R. (2016). Change, resistance to change and organizational cynicism. *Studies in Business and Economics, 3*(11), 47-54.

IAEA. (2010). *Subject Categories and Scope Descriptions.* Vienna: INIS.

Kanahara, S. (2006). A Review of the Definitions of Stereotype and a Proposal for a Progressional Model. *Individual Differences Research, 4*(5), 306-321.

Ng, T. W., & Feldman, D. C. (2012, November). Evaluating Six Common Stereotypes about Older Workers with Meta-Analytical Data. *Personnel Psychology, 65*(4).

Rafferty, A., & Griffin, M. (2008). Organizational Change. In *The Sage Handbook of Organizational Behavior. Vol. 1: Micro Approaches* (pp. 603-620). London: Sage Publications.

Scandura, T. A., & Lankau, M. J. (1997, July). Relationships of Gender, Family Responsibility and Flexible Work Hours to Organizational Commitment and Job Satisfaction. *Journal of Organizational Behavior, 18*(4), 377-391.

Schöpfel, J. (2010). Towards a Prague Definition of Grey Literature. *Twelfth International Conference on Grey Literature, December 6-7, 2010.* Prague: TextRelease.

Seid, G. M., & Venkataram, B. M. (2016). The relationship between workplace ethics and employee attraction, loyalty, and financial performance. *International Journal of Applied Research, 2*(5), 585-596.

Shultz, K. S., & Adams, G. A. (2007). *Aging and Work in the 21st Century.* New Jersey: Lawrence Erlabaum Associates, Publishers.

Simons, S. M., & Rowland, K. N. (2011). Diversity and its Impact on Organizational Performance: The Influence of Diversity Constructions on Expectations and Outcomes. *Journal of Technology, Management and Innovation, 6*(3).

https://doi.org/10.26069/greynet-2025-000.534
TGJ Volume 13, Number 1 2017

Nineteenth International Conference on Grey Literature

Public Awareness and Access to Grey Literature

National Research Council of Italy

Piazzale Aldo Moro 7, Rome, Italy ▪ October 23-24, 2017

Conference Announcement

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

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

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

GL19 seeks to demonstrate how researchers and authors in the last 25 years have made significant inroads in responding to the loss and overload of grey literature. Likewise, this conference will seek to provide new directions in achieving public awareness and access to grey literature on an ever changing information landscape. To this end, information professionals and practitioners in the sectors of government, academics, business and industry are invited to respond to this year's **Call for Papers** reflected in the conference topics.

GreyNet Grey Literature Network Service 1992 2017

Conference Dateline 2017

● April 15	● May 5	● May 12	● May 15	● Sept. 15	● Sept. 25	● Oct. 10	● Oct. 23-24
Close, Call for Papers	Program Committee Meeting	Authors Notified	Open, Call for Posters	Close, Early Conference Registration	Close, Call for Posters	Submission Conference Papers	GL19 Conference

The GL19 Call for Papers remains open until April 15, 2017. After this date, abstracts submitted for presentation in a Plenary Session will be evaluated by the Conference Program Committee set to meet on May 5, 2017 in Pisa, Italy. By mid-May, authors will have been notified of their place on the GL19 Conference Program.

On May 15th a formal Call for Posters will be launched, closing one month prior to the conference or when all available poster slots have been filled. Authors who choose to present a poster instead of a paper are not required to submit a full-text manuscript. However, if they so choose, it will be eligible for publication in the GL19 Conference Proceedings.

Authors elected by the Program Committee to present during one of the conference plenary sessions and who submit a full-text paper qualify to be nominated for the GreyNet Annual Award 2018. Those presenting during the conference poster session will have a chance to win the poster prize 2017.

https://doi.org/10.20069/greynet-2025-000.534

Nineteenth International Conference on Grey Literature

Public Awareness and Access to Grey Literature

National Research Council of Italy

Piazzale Aldo Moro 7, Rome, Italy ▪ October 23-24, 2017

Call for Papers

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

Guidelines for Abstracts

Participants who seek to present a paper at GL19 are invited to submit an English language abstract between 350-400 words. The abstract should deal with the problem/goal, the research method/procedure, an indication of costs related to the project, as well as the anticipated results of the research. The abstract should likewise include the title of the proposed paper, conference topic(s) most suited to the paper, name(s) of the author(s), and full address information. Abstracts are the only tangible source that allows the Program Committee to guarantee the content and balance in the conference program. Every effort should be made to reflect the content of your work in the abstract submitted. Abstracts not in compliance with the guidelines may be returned to the author for revision.

Conference Related Topics

- ☐ **Exposing Grey Literature to Wider Audiences**
- ☐ **Confronting Obstacles in Accessing Grey Literature**
- ☐ **Impact of Emerging Technologies on Grey Literature**
- ☐ **Innovations in Grey Literature powered by Research Data**
- ☐ **Extracting Trusted Content from Social Media**
- ☐ **Digital Preservation, the Lifeline for Grey Resources**

GreyNet Grey Literature Network Service 1992 2017

Due Date and Format for Submission

Abstracts in MS Word must be emailed to conference@textrelease.com on or before **April 15th 2017**. The author will receive verification upon its receipt. By mid-May, shortly after the Program Committee meets, authors will be notified of their place on the conference program. This notice will be accompanied by further guidelines for submission of full text papers, accompanying research data, PowerPoint slides, and required Author Registration.

TextRelease
GL19 Program and Conference Bureau

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

Author Information

Bartolini, Roberto **41**
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 **23, 41**
Stefania Biagioni graduated in Italian Language and Literature at the University of Pisa and specialized in Data Processing and DBMS. She is currently a member of the research staff at the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), an institute of the Italian National Research Council (CNR) located in Pisa. She is head librarian of the Multidisciplinary Library of the CNR Campus in Pisa and member of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She has been the responsible of ERCIM Technical Reference Digital Library (ETRDL) and currently of the PUMA (PUblication MAnagement) & MetaPub, a service oriented and user focused infrastructure for institutional and thematic Open Access repositories looking at the DRIVER/OpenAire vision, http://puma.isti.cnr.it. She has coauthored a number of publications dealing with digital libraries. Her activities include integration of grey literature into library collections and web access to the library's digital resources, including electronic journals and databases. She is a member of GreyNet since 2005. Since 2013 she is involved on the GreyGuide Project.
Email: stefania.biagioni@isti.cnr.it

Cancedda, Flavia **32**
Flavia Cancedda, Technologist, librarian and information specialist. Head of the Italian ISSN Centre (Central Library CNR); Component of “Documentation and Information” Committee of the UNI, National standardization organism (Chairman of the Subcommittee “Presentation, identification and description of documents”; component of the Working Group on National Standard about Professional Requirements for Librarians); Component of the ACNP Libraries Committee (Archivio Collettivo Nazionale Periodici) and of the ACNP Cataloguing Rules Revision Working Group. Since 1991 to 2000 librarian in two different Italian universities and in the National Library of Florence (section: National Bibliography); since 2001 librarian in the Central Library of the Italian National Research Council (CNR). Master’s degree in Libraries Management and Direction. Post-graduate degree as Manuscripts conservator at the Special School for Archivists and Librarians (La Sapienza University in Rome). Author of books and essays about the history of libraries and historical bibliography. Forthcoming publications (2015): ISSN e i suoi fratelli. La lunga marcia dei codici internazionali (part 1 and 2), AIB studi Recent publications/contributes: UNI 11535: il riconoscimento che mancava, AIB studi (Online), 55 (2015), 1, 115-124; UNI 11535:2014, Qualificazione delle professioni per il trattamento di dati e documenti - Figura professionale del bibliotecario - Requisiti di conoscenza, abilità e competenza; La norma UNI sulla professione del bibliotecario, Il Mondo degli archivi (Online), feb. 2014; Authority control, workshop, held in FSR, Faster Smarter and Richer. Reshaping the library catalogue, International meeting, Rome, feb. 2014; UNI ISO 8:2014 Documentazione e informazione - Presentazione dei periodici (translation of: ISO 8:1977 Documentation -- Presentation of periodicals).
Email: flavia.cancedda@cnr.it

De Biagi, Luisa **32**
Luisa De Biagi got her Laurea Degree in Literature and Philosophy at 'La Sapienza' Univ. of Rome (Art history and Cultural heritage). With a Specialization in 'Archivist-Palaeographer' (Vatican School of Palaeo-graphy, Diplomatics and Archivistics at the Vatican Secret Archive) as well as a Specialization Degree in Archivistics, Palaeography and Diplomatics (Archivio di Stato, Rome) and a Degree from the Vatican School of Library Sciences. De Biagi further holds a Master in ‘Business Publishing’ (LUISS Management – Rome). She’s been working for the SIGLE Network (System for Information on Grey Literature in Europe) since 2002, Since 2010 she’s is responsible for the Italian National Referring Centre of Grey Literature at CNR Central Library ‘G. Marconi’ as representative of the European Network and Openarchive OpenGrey. She’s taken part in 3 editions of GreyNet’s GL Conference Series (GL5 Amsterdam, GL13 Washington DC, GL14 Rome and GL15 Bratislava). She’s is also a member of the CNR Working Group for Cedefop-Refernet Project (Consortium for Professional Education and Training coordinated by ISFOL), the Committee for Legal Deposit Acquisition at CNR Central Library, and a member of the European Association of Health Information and Libraries (EAHIL). She’s also responsible for the Library Functional Units ‘Education and Training’ and ‘Cultural Activities Management’, organizing didactics laboratories for students, professional training courses and teaching in professional trainings for librarians, students and users. Email: luisa.debiagi@cnr.it

Author Information CONTINUED

Giannini, Silvia **23, 41**
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 **23, 41**
Sara Goggi is a technologist at the Institute of Computational Linguistics "Antonio Zampolli" of the Italian National Research Council(CNR-ILC) in Pisa. She started working at ILC in 1996 working on the EC project LE-PAROLE for creating the Italian reference corpus; afterwards she began dealing with the management of several European projects and nowadays she is involved with organisational and managerial activities mainly concerning international relationships and dissemination as well as organization of events (e.g. LREC conference series). Currently one of her preminent activities is the editorial work for the international ISI Journal Language Resources and Evaluation, being its Assistant Editor. Since many years (from 2004) she also carries on research on terminology and since 2011 - her first publication at GL13 - she is working on topics related with Grey Literature. Email: sara.goggi@ilc.cnr.it

Pardelli, Gabriella **23, 41**
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

Prost, Hélène **5**
Hélène Prost is information professional at the Institute of Scientific and Technical Information (CNRS) and associate member of the GERiiCO research laboratory (University of Lille 3). She is interested in empirical library and information sciences and statistical data analysis. She participates in research projects on evaluation of collections, document delivery, usage analysis, grey literature and open access, and she is author of several publications.
Email: helene.prost@inist.fr

Savić, Dobrica **55**
Dr. Dobrica Savić 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, including UNV, UNESCO, World Bank, ICAO, and the IAEA. His main interests are creativity, innovation and use of information technology in library and information services.
Email: d.savic@iaea.org

Schöpfel, Joachim **5**
Joachim Schöpfel is senior lecturer at the Department of Information and Library Sciences at the Charles de Gaulle University of Lille 3 and Researcher at the GERiiCO laboratory. He is interested in scientific information, academic publishing, open access, grey literature and eScience. He is a member of GreyNet and euroCRIS. He is also the Director of the National Digitization Centre for PhD Theses (ANRT) in Lille, France. Email: joachim.schopfel@univ-lille3.fr

Smith II, Plato L. **37**
Plato Smith is the Data Management Librarian at the University of Florida with experience in academic research libraries, digital libraries, and data management. He received his doctorate in the field of Information Science from the School of Information within the College of Communication and Information at Florida State University, Florida’s iSchool, Summer 2014. From 2005 to 2012, he was Department Head for the FSU Libraries’ Digital Library where he developed, populated, and managed digital collections in the FSU Libraries’ digital content management system, DigiNole Repository, and electronic theses and dissertations (ETDs) institutional repository. Email: plato.smith@ufl.edu

Notes for Contributors

Non-Exclusive Rights Agreement

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

Submission Requirements

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

REFERENCE GUIDELINES

General

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

Specific

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

Examples

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

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

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

Review Process

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

Journal Publication and Article Deposit

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

https://doi.org/10.26069/greynet-2025-000.534
TGJ Volume 13, Number 1 2017

An International Journal on Grey Literature

'Grey Literature Domain in Broad View'

TGJ Volume 13, Number 1 2017

Altmetrics and Grey Literature: Perspectives and Challenges 5
Joachim Schöpfel and Hélène Prost (France)

Grey Literature citations in the age of Digital Repositories and Open Access 23
Silvia Giannini, Stefania Biagioni, Sara Goggi and Gabriella Pardelli (Italy)

International identification and 'white and grey literature' : Identities, retrieval, reuse and the certainty of knowledge while sharing and connecting information 32
Flavia Cancedda and Luisa De Biagi (Italy)

Developing, linking, and providing access to supplemental genetics dataset vcf files 37
Plato L. Smith II and Lauren McIntyre (United States)

A terminological "journey" in the Grey Literature domain 41
Roberto Bartolini, Gabriella Pardelli, Sara Goggi, Silvia Giannini, and Stefania Biagioni (Italy)

Managing Diversity in the International Nuclear Information System 55
Dobrica Savić (Austria)

Colophon 2

Editor's Note 4

On The Newsfront

GL19 Announcement – Nineteenth International Conference on Grey Literature, October 23-24, 2017 in Rome, Italy 63

GL19 Call for Papers – 'Public Awareness and Access to Grey Literature' Submissions close on April 15, 2017 64

Advertisement

INIS-IAEA, International Nuclear Information System 54

Author Information 65

Notes for Contributors 67