

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

Plato L. Smith II, University of Florida (Libraries), United States  
Lauren McIntyre, University of Florida (Microbiology), 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.

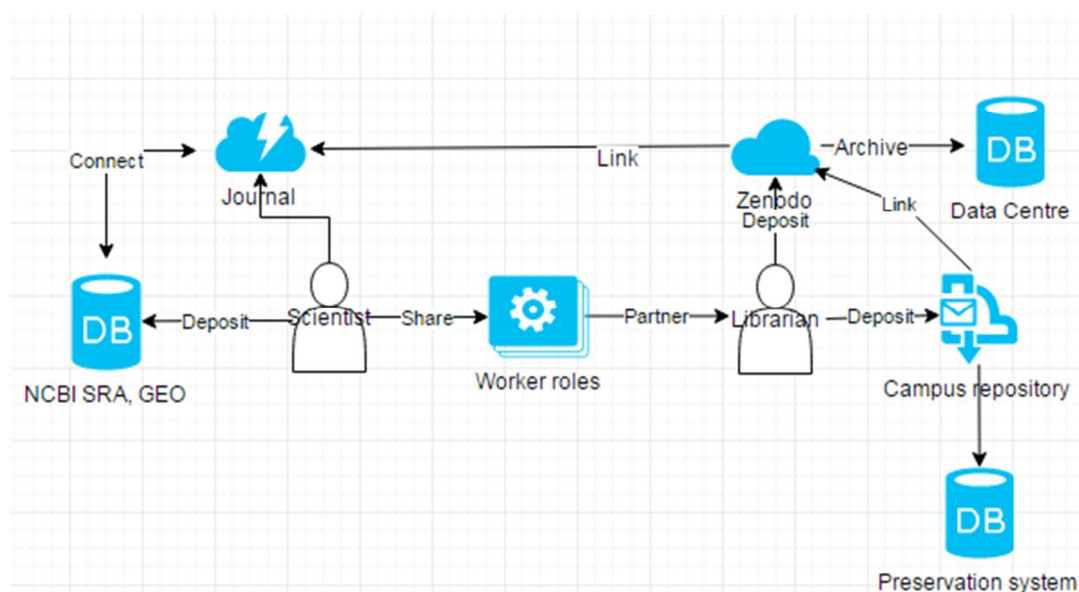
Data generation, raw data files are often text files of large sizes. These raw data- effectively what comes off the machine have been a focus for many in terms of collection and preservation. While there has been a huge effort expended in this area for gene expression data, there has been much less effort 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 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

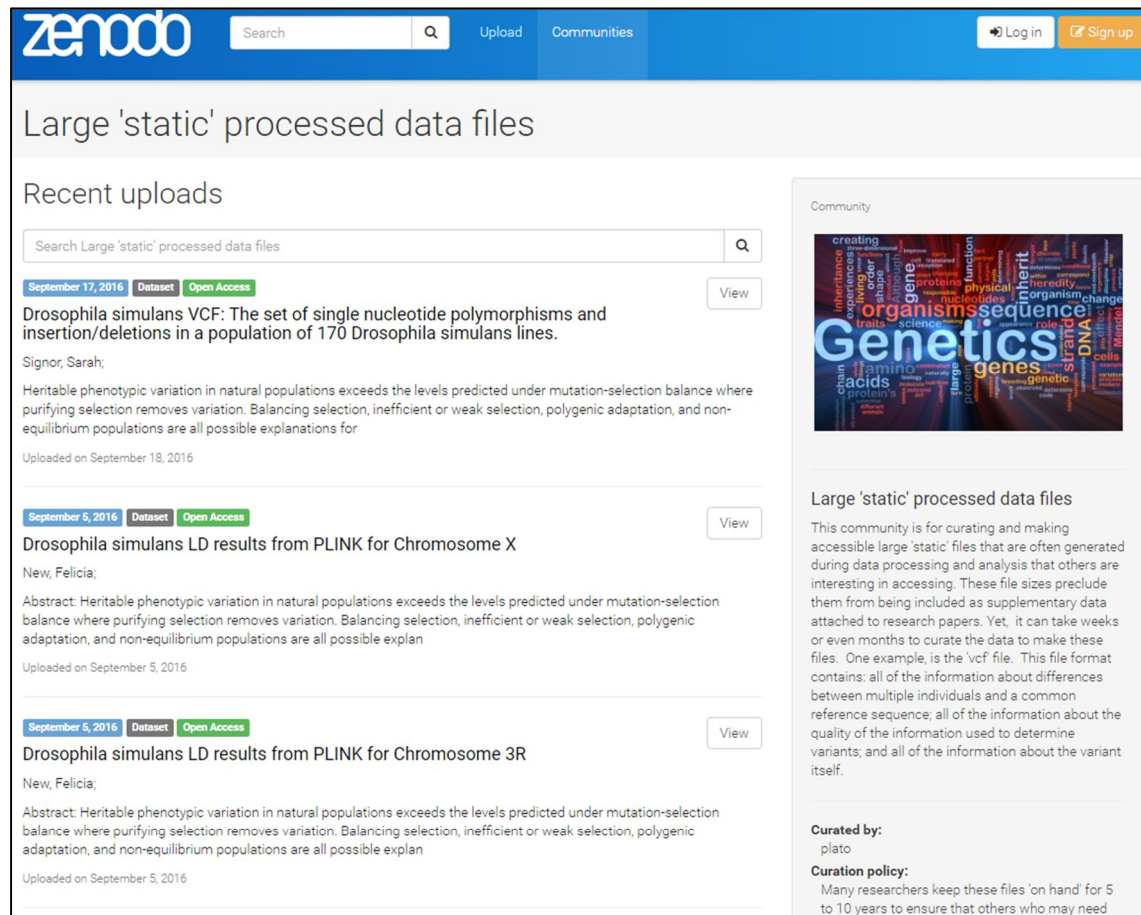
### Using Zenodo as a vcf data repository

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](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 compliant 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 *Drosophila 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>

# National Repository of Grey Literature (NRGL)



**NRGL is**  
digital  
repository  
for grey  
literature

**Free**  
online  
access

## Features

### Provider:

**National Library of Technology  
Prague, Czech Republic**

### Records:

**over 400,000 records**

### Collection provenance:

**Czech Republic**

### Partners:

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

### International Cooperation:

**OpenGrey, OpenAire, ROAR, OpenDOAR, BASE**

## Goals

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

## What else?

**Conference on Grey Literature  
and Repositories**

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

**Informative Web pages**

<http://nrgl.techlib.cz>

**www.nusl.cz**

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

**NU1  
SL**  
národní  
úložiště  
šedé  
literatury



## Policy Development for Grey Literature Resources: An Assessment of the Pisa Declaration

Dobrica Savić, NIS-IAEA, Austria;

Dominic Farace and Jerry Frantzen, GreyNet International, Netherlands;

Stefania Biagioni and Carlo Carlesi, ISTI-CNR, Italy;

Herbert Gruttemeier and Christiane Stock, Inist-CNRS, France

### Abstract

In the spring of 2014, a workshop took place at the Italian National Council of Research in Pisa<sup>1</sup>. The topic of this event dealt with policy development for grey literature resources. Some seventy participants from nine countries took an active part in the workshop – the outcome of which produced what is today known as the Pisa Declaration<sup>2</sup>. This fifteen point document arising from the input of those who attended the workshop sought to provide a roadmap that would help to serve diverse communities involved in research, publication and the management of grey literature both in electronic and print formats.

The Pisa Declaration has been translated and published in some twenty languages. They are all accessible online via the GreyGuide Repository<sup>3</sup> and Portal<sup>4</sup>. Currently, 140 information professionals from renowned organizations worldwide have endorsed this document<sup>5</sup>.

In an effort to assess the impact that the Pisa Declaration has had during the last two years on the policy development for grey literature resources, an online survey among those who endorsed the document was carried out and their responses were analysed. Descriptive statistics and short summaries are used to describe the basic features of the data collected. They are combined with simple graphics that offer easier visual representation of the results achieved. Specific results of the survey analysis indicate those points in the Pisa Declaration that in varying degrees are of relevance and importance to grey literature, as well as points that need further attention and work. Although integral part of library and information management practice grey literature has its own peculiarities and needs that require special attention in order to reach its deserved level of importance in today's research and other activities.

### Introduction

Since its publication in 2014, the Pisa Declaration on Policy Development for Grey Literature Resources has been endorsed by 140 signatories from 74 organizations in 30 countries worldwide. This Declaration has since been translated from the original English text into 20 languages and has come to be termed as the 'roadmap for grey literature in the 21<sup>st</sup> Century'. Now two years on, it is opportune to assess the impact this document has had on library and information practice. It is to this end that an online survey was conducted among its signatories, the results of which are found here recorded.

### Survey sample size and population

Online questionnaire-based Pisa Declaration survey was created and placed on the SurveyMonkey on 25 April 2016. It consisted of 10 multiple choice questions with some of them offering a possibility to leave additional comments. The survey was designed in English language only. First survey replies were received on 30 May and the survey was closed on 18 July 2016. Figure 1 shows the survey response volume and the time distribution.

Requests for completion of the survey were sent to all 133 Pisa Declaration signatories, out of which 60 responded. This marked a 45% response rate.

Generally speaking there are two types of surveys - Surveys distributed internally, such as this one, since it was distributed only to a pre-set group of individuals, and external surveys distributed to wider audience, such as potential customers or general public.

<sup>1</sup> <http://eventi.isti.cnr.it/index.php/en/programme-grey>

<sup>2</sup> [http://www.greynet.org/images/Pisa\\_Declaration,\\_May\\_2014.pdf](http://www.greynet.org/images/Pisa_Declaration,_May_2014.pdf)

<sup>3</sup> <http://goo.gl/72yexP>

<sup>4</sup> <http://greyguide.isti.cnr.it/>

<sup>5</sup> <http://greyguiderep.isti.cnr.it/pisadecla/listaiscritti.php?order=name>

According to SurveyGizmo.com, internal surveys will generally receive a 30-40% response rate on average, compared to an average 10-15% response rate for external surveys.<sup>6</sup> Achieved response rate of 45% with the Pisa Declaration survey is therefore, regarded as above the average.

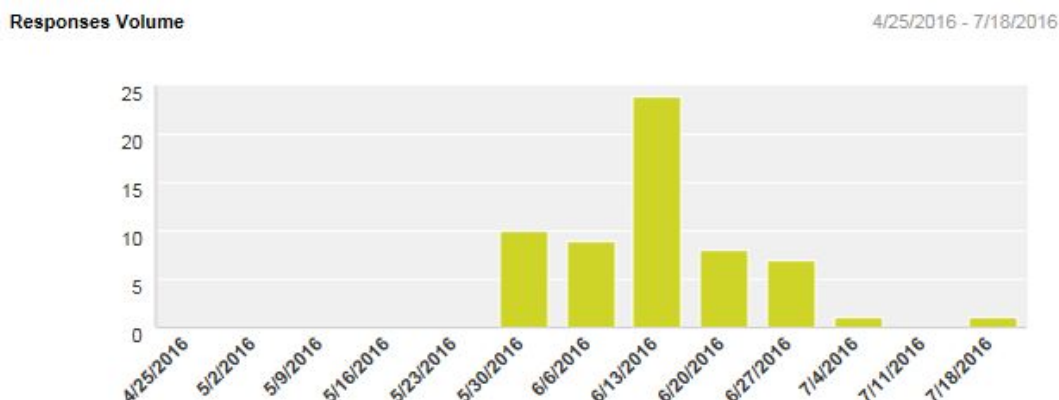


Figure 1: Response Volume

#### Question 1: How did you first come to endorse the Pisa Declaration?

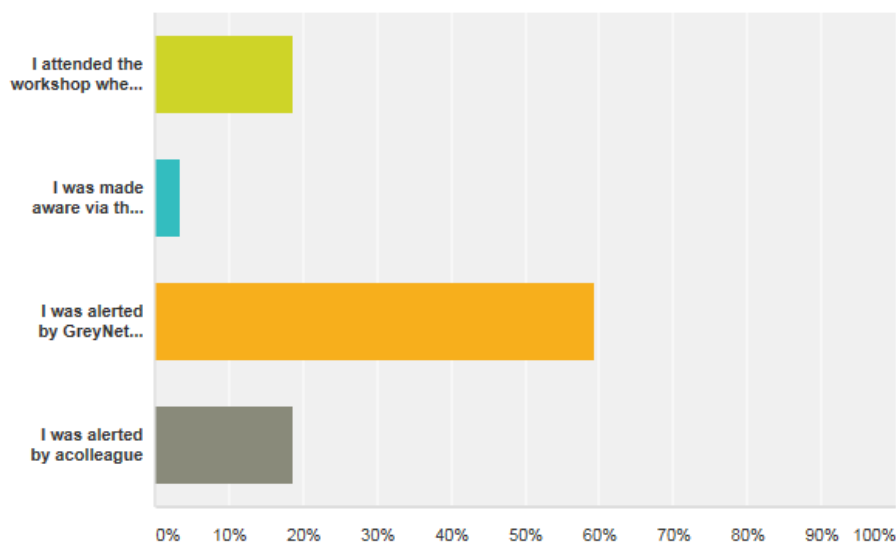
The goal of this question was to find out the way respondent found out about the Pisa Declaration. Out of 60 respondents, 59 answered this question. One respondent skipped this question and left an interesting comment that he/she does not remember how it came out to endorse the Pisa Declaration.

Replies to this question offer some interesting conclusions. First of all, the website is not the best channel for promoting, grabbing people's attention or inciting them to take some action, in this case to endorse the Declaration. Only 3.4% indicated that as the main prompt for endorsing the Declaration. Lots of institutional and organizational resources go into website creation, development and maintenance, but the impact is not always so great. According to the answers received, direct contacts by the GreyNet International (almost 60%), or by a colleague (18.6%) produced the best results which should encourage us to continue maintaining personalized mailing lists and to use multiple opportunities offered by social media, such as Facebook and Twitter.

Interestingly enough, attendance the workshop where the Pisa Declaration was drafted was the reason for only 18.6% respondents to endorse it.

#### How did you first come to endorse the Pisa Declaration?

Answered: 59 Skipped: 1



<sup>6</sup> <https://goo.gl/2l8Zbx>



Answer Choices	Responses
I attended the workshop where the Pisa Declaration was drafted	18.64% 11
I was made aware via the GreyGuide Portal and Repository	3.39% 2
I was alerted by GreyNet International	59.32% 35
I was alerted by a colleague	18.64% 11
Total	59
<a href="#">Comments (5)</a>	

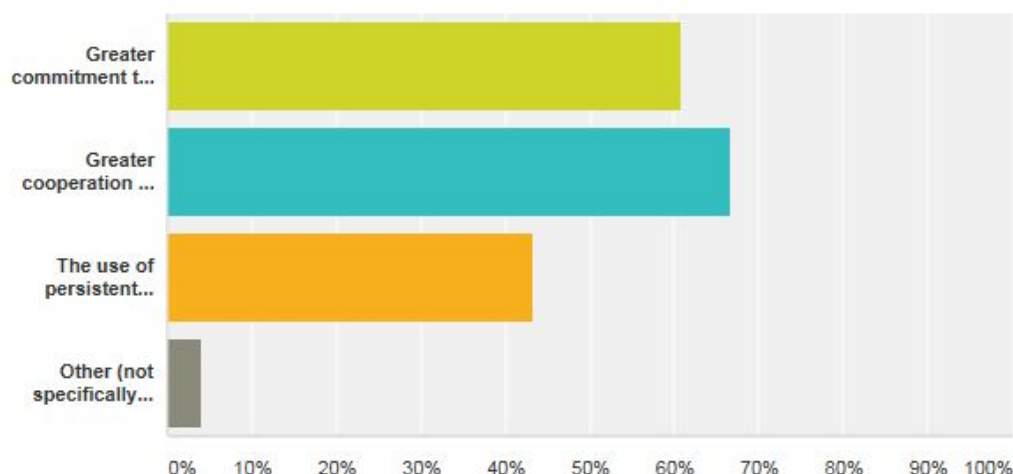
Figure 2: Survey Question No. 1

**Question 2: Indicate the Organizational point(s) stated in the Pisa Declaration that concern your organization most**

Although the question was referring to the Pisa Declaration, it was aimed at finding out more about the main topics of interest or main concerns that organizations have regarding grey literature. Replies indicate that all three indicated areas namely open access, cooperation, and operational standards, currently represent topics of high importance and interest. In a way it is an indication of the current state of grey literature in organizations where much action is required for better processing, dissemination and use of this type of literature.

**Indicate the Organizational point(s) stated in the Pisa Declaration that concern your organization most**

Answered: 51 Skipped: 9



Answer Choices	Responses
Greater commitment to open access by governments and organizations	60.78% 31
Greater cooperation and coordination among organizations engaged in the production, use, collection and management of grey literature	66.67% 34
The use of persistent identifiers and open metadata standards for grey literature	43.14% 22
Other (not specifically mentioned in the Declaration)	3.92% 2
Total Respondents: 51	

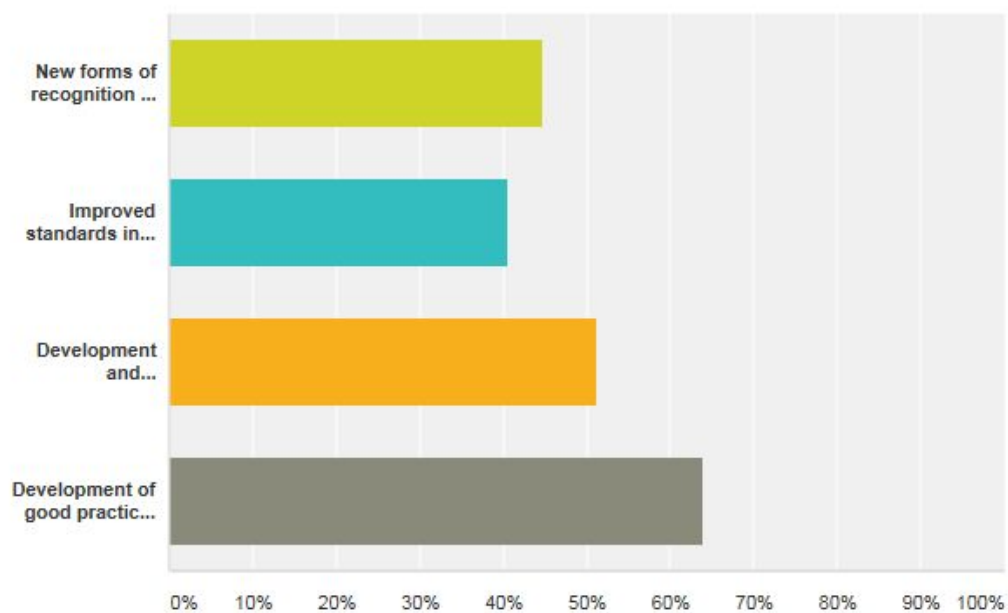
Figure 3: Survey Question No. 2

**Question 3: Indicate the Research and Educational point(s) stated in the Pisa Declaration that concern your organization most**

In a way question number 3 is a continuation of the previous question, since both of them try to find out more about the main topics of interest or main concerns that organizations have regarding grey literature. The difference is that this particular question is concentrated on research and educational points. Similar nature of the question and similar answers received. They also indicate that all four indicated areas – recognition, production and interoperability standards, and good practices, currently represent topics that require attention and further improvement work. Again, it is an indication of the current state of grey literature in general, with many areas and opportunities for improvements.

**Indicate the Research and Educational point(s) stated in the Pisa Declaration that concern your organization most**

Answered: 47 Skipped: 13



Answer Choices	Responses
▼ New forms of recognition and reward for quality grey literature materials by governments, universities and other institutions	44.68% 21
▼ Improved standards in the production and bibliographic control of grey literature	40.43% 19
▼ Development and implementation of interoperable standards in the management of grey literature	51.06% 24
▼ Development of good practice guides for the production, dissemination, and evaluation of grey literature	63.83% 30
Total Respondents: 47	

[Comments \(2\)](#)

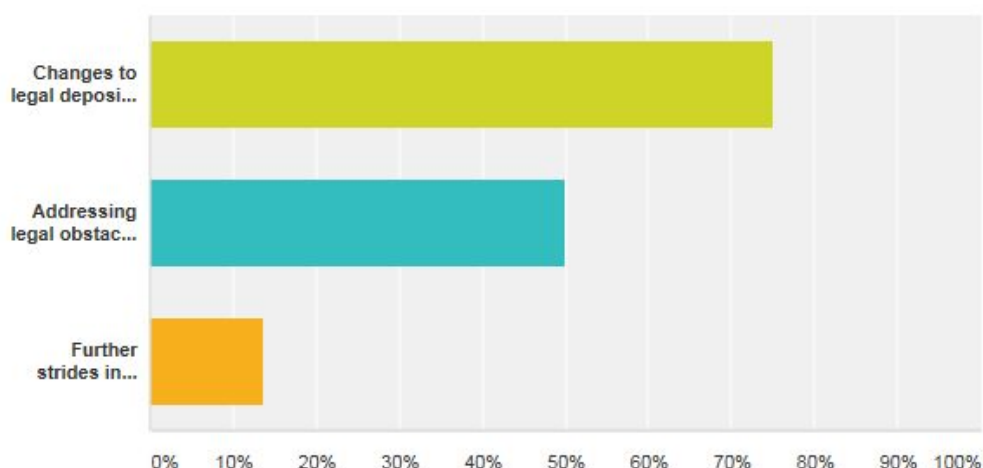
**Figure 4: Survey Question No. 3**

**Question 4: Indicate the Legal point(s) stated in the Pisa Declaration that concern your organization most**

Legal issues and protection of the intellectual property in information management and in management of grey literature is of huge concern for everyone. It is a concern to both – information providers and information users. Challenges are present on both sides, although in different forms. Information providers want to make their documents available, preferably as open source, but still protected as their unique intellectual property. Information users would like to use as much documentation, information and data as possible, but at the same time to be respectful of copyright issues. What the replies to this question on legal concerns indicate is that providers need enhanced copyright regulations that will improve the capabilities of libraries and other collecting services so that they can provide available documentation without much hindrance.

**Indicate the Legal point(s) stated in the Pisa Declaration that concern your organization most**

Answered: 44 Skipped: 16



Answer Choices	Responses
Changes to legal deposit and copyright law that enhance the capacities of libraries, collecting services and educational institutions and programs to collect and provide access to grey literature, particularly non-commercial public interest materials	75.00% 33
Addressing legal obstacles to the dissemination of grey literature	50.00% 22
Further strides in licensing grey content for both commercial and non-commercial purposes	13.64% 6
Total Respondents: 44	
Comments (1)	

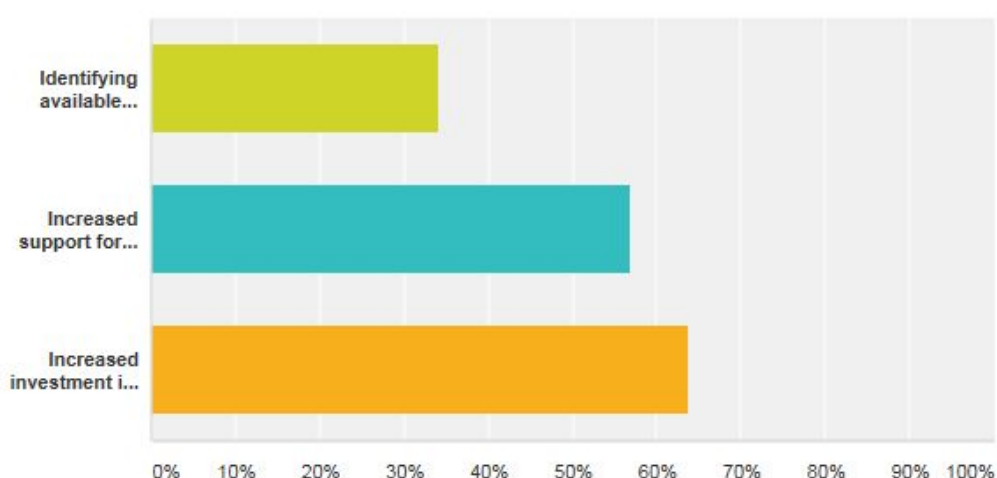
**Figure 5: Survey Question No. 4**

**Question 5: Indicate the Financial and Sustainable point(s) stated in the Pisa Declaration that concern your organization most**

Replies to this question were predictable. Grey literature, all of its sides and activities need more money, better and sustainable financing on a long run. The most urgent seems to be increased investment in infrastructure and new technologies, followed by grey literature long-term preservation. The issue of preservation is particularly vulnerable one since both, paper as well as digital collections, are disappearing quickly. At the same time users demand quick and unrestricted access to full-text documentation. This is a very huge area for further work in making grey literature more relevant and better appreciated.

**Indicate the Financial and Sustainable point(s) stated in the Pisa Declaration that concern your organization most**

Answered: 44 Skipped: 16



Answer Choices	Responses
Identifying available funding for research involving grey literature	34.09% 15
Increased support for collection development and long term preservation of grey literature	56.82% 25
Increased investment in infrastructure and new technologies for accessing and using print and digital grey literature	63.64% 28
Total Respondents: 44	

[Comments \(2\)](#)

**Figure 6: Survey Question No. 5**

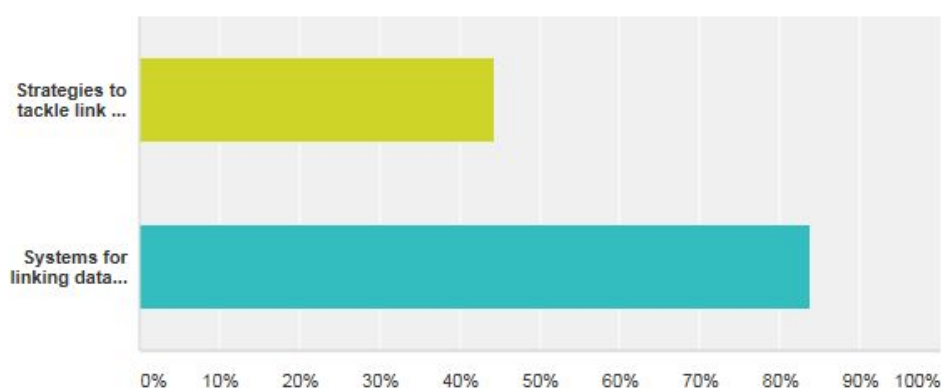
**Question 6: Indicate the Technical point(s) stated in the Pisa Declaration that concern your organization most**

Link rot refers to problems of hyperlinks on individual websites pointing to web pages, servers or other online resources that have become permanently unavailable. This is a problem for Internet in general, not only for grey literature. The link sustainability challenge was indicated as an issue by 44% of all respondents, although we can rightfully conclude that it affects every information provider currently running or using available websites. Question of finding, repairing and preventing broken links would require further study. Use of best practices for preventing link rot, including use of DOI numbers and PURLs requires greater attention among grey literature providers.

A completely new area of linking data and other non-technical content to their grey literature publications together with interoperability standards for sharing grey literature was almost on everyone's mind. 84% of participants indicated that as the greatest technical concern for their organization.

**Indicate the Technical point(s) stated in the Pisa Declaration that concern your organization most**

Answered: 43 Skipped: 17



Answer Choices	Responses
Strategies to tackle link rot and enhance the stability and accessibility of online content	44.19% 19
Systems for linking data and other non-textual content to their grey literature publications togetherwith interoperability standards for sharing grey literature	83.72% 36
Total Respondents: 43	

Comments (1)

**Figure 7: Survey Question No. 6**

**Question 7: Is there a language(s) not listed above in which the Pisa Declaration should be translated and published? If so, please indicate here.**

Pisa Declaration was drafted in English on 16 May, 2014. Due to wide interest and hard work of some of the members of the Grey Literature community, the Declaration was translated into 21 languages. They include: Armenian, Bulgarian, Croatian, Czech, Dutch, French, German, Greek, Hindi, Hungarian, Italian, Japanese, Korean, Macedonian, Russian, Serbian Cyrillic, Serbian Latin, Slovak, Spanish, Tagalog, and Turkish. All of the translated versions are available online<sup>7</sup>.

The intention of the question number 7 was to find out from the respondents if there was a need for translating it into some other language. The following languages were suggested for translation and inclusion: Arabic, Chinese, Portuguese, and Korean.

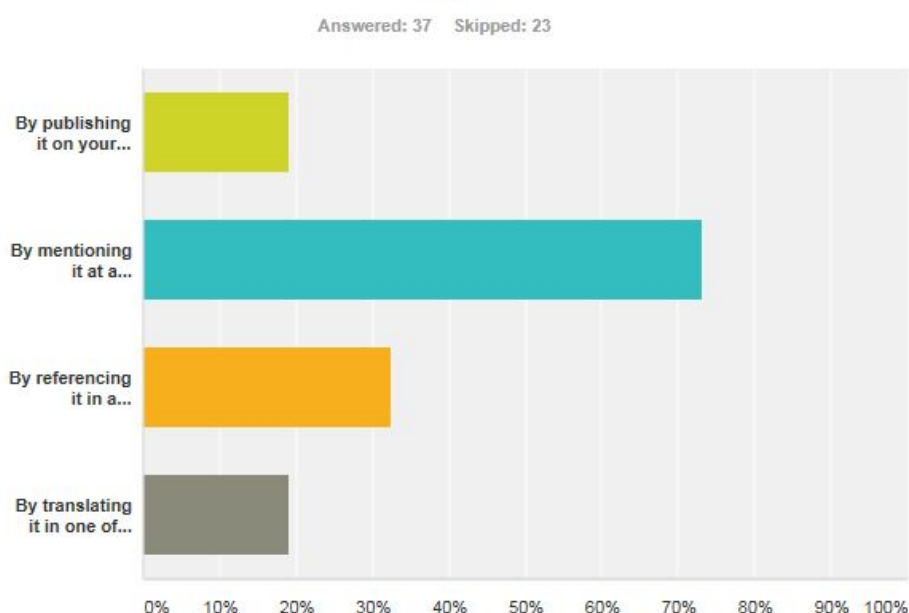
<sup>7</sup> <http://greyguide.isti.cnr.it/>



### Question 8: Have you had the opportunity to promote public awareness to the Pisa Declaration?

Starting assumptions before drafting this survey question was an impression that grey literature in general, as well as the Pisa Declaration, were not being promoted sufficiently. Replies indicate that the assumption was correct. Most of the promotion was done through ad hoc means, such as meetings or conferences. Publishing on the organizational website, as a way for providing a more sustainable presence, was exploited by few participants only. Suggestions received through comments indicated a greater need for promoting the Pisa Declaration through training and social media.

#### Have you had the opportunity to promote public awareness to the Pisa Declaration? If so,



Answer Choices	Responses
By publishing it on your organization's website	18.92% 7
By mentioning it at a conference or meeting	72.97% 27
By referencing it in a publication	32.43% 12
By translating it in one of the above mentioned languages	18.92% 7
Total Respondents: 37	

[Comments \(3\)](#)

Figure 8: Survey Question No. 8

### Question 9: Please take a moment to record any comments, recommendations, amendments, or additions you consider worthwhile for the Pisa Declaration to further benefit policy development for grey literature resources.

Question number 9 asked for comments, recommendations, amendments, or additions considered to be of benefits for further policy development of grey literature. There were 12 comments received. Most of them mentioned training, development of practical manuals and appropriate standards, as well as further studies. Improving awareness among users in developing countries of grey literature benefits was also mentioned.

### Question 10: Your name and email address.

It was curious to notice that out of 60 survey participants 46 identified themselves by leaving their email address. This indicates some kind of devotion to the topic being surveyed and a desire to keep in contact with colleagues and learn more about future progress and developments impacting the world of grey literature.



**Concluding Remarks**

The outcome of the survey leads us to conclude that direct contact via GreyNet and professionals in the grey literature community account for the majority of endorsements to the Pisa Declaration. Its placement on the GreyGuide portal was significant for its formal publication and to facilitate endorsement, however this in itself accounted for the least number of signatories. Across the board, all of the points in the Pisa Declaration are still of concern to the grey literature community.

Legal issues remain the concern not only for content and service providers but also for users of grey literature. Increased investment in new technologies enabling access to the full-text as well as related research and metadata are of equal concern. However, the need to promote public awareness to grey literature is underestimated and would contribute significantly to policy development in this field of library and information science.

## Appendix 1

# Pisa Declaration on Policy Development for Grey Literature Resources

May 16, 2014, Pisa

## Introduction

A wealth of knowledge and information is produced by organizations, governments and industry, covering a wide range of subject areas and professional fields, not controlled by commercial publishing. These publications, data and other materials known as grey literature, are an essential resource in scholarly communication, research, and policy making for business, industry, professional practice, and civil society.

Grey literature is recognized as a key source of evidence, argument, innovation, and understanding in many disciplines including science, engineering, health, social sciences, education, the arts and humanities.

Grey literature document types in print or electronic formats include among others: research and technical reports, briefings and reviews, evaluations, working papers, conference papers, theses, and multimedia content, representing an important and valuable part of research and information.

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

*To achieve the full benefits of grey literature for local, national and global communities we call for and encourage the following:*

### Organizational

1. Greater commitment to open access by governments and organizations.
2. Greater cooperation and coordination among organizations engaged in the production, use, collection and management of grey literature.
3. The use of persistent identifiers and open metadata standards for grey literature.

### Research/Educational

4. New forms of recognition and reward for quality grey literature materials by governments, universities and other institutions.
5. Improved standards in the production and bibliographic control of grey literature.
6. Development and implementation of interoperable standards in the management of grey literature.
7. Development of good practice guides for the production, dissemination, and evaluation of grey literature.

### Legal

8. Changes to legal deposit and copyright law that enhance the capacities of libraries, collecting services and educational institutions and programs to collect and provide access to grey literature, particularly non-commercial public interest materials.
9. Addressing legal obstacles to the dissemination of grey literature.
10. Further strides in licensing grey content for both commercial and non-commercial purposes.

### Financial/Sustainable

11. Identifying available funding for research involving grey literature.
12. Increased support for collection development and long term preservation of grey literature.
13. Increased investment in infrastructure and new technologies for accessing and using print and digital grey literature.

### Technical

14. Strategies to tackle link rot and enhance the stability and accessibility of online content.
15. Systems for linking data and other non-textual content to their grey literature publications together with interoperability standards for sharing grey literature.

## Appendix 2

# Pisa Declaration: An Assessment Study

## Pisa Declaration on Policy Development for Grey Literature Resources

**In the spring of 2014, a workshop took place at the Italian National Council of Research in Pisa. The topic of this event dealt with policy development for grey literature resources. Some seventy participants from nine countries took an active part in the workshop – the outcome of which produced what is today known as the Pisa Declaration.**

### 1. How did you first come to endorse the Pisa Declaration?

- ☐ I attended the workshop where the Pisa Declaration was drafted
- ☐ I was made aware via the GreyGuide Portal and Repository
- ☐ I was alerted by GreyNet International
- ☐ I was alerted by a colleague
- Other (please specify)

### 2. Indicate the Organizational point(s) stated in the Pisa Declaration that concern your organization most

- ☐ Greater commitment to open access by governments and organizations
- ☐ Greater cooperation and coordination among organizations engaged in the production, use, collection and management of grey literature
- ☐ The use of persistent identifiers and open metadata standards for grey literature
- ☐ Other (not specifically mentioned in the Declaration)

### 3. Indicate the Research and Educational point(s) stated in the Pisa Declaration that concern your organization most

- ☐ New forms of recognition and reward for quality grey literature materials by governments, universities and other institutions
- ☐ Improved standards in the production and bibliographic control of grey literature
- ☐ Development and implementation of interoperable standards in the management of grey literature
- ☐ Development of good practice guides for the production, dissemination, and evaluation of grey literature
- Other (not specifically mentioned in the Declaration)

### 4. Indicate the Legal point(s) stated in the Pisa Declaration that concern your organization most

- ☐ Changes to legal deposit and copyright law that enhance the capacities of libraries, collecting services and educational institutions and programs to collect and provide access to grey literature, particularly non-commercial public interest materials
- ☐ Addressing legal obstacles to the dissemination of grey literature
- ☐ Further strides in licensing grey content for both commercial and non-commercial purposes
- Other (not specifically mentioned in the Declaration)

**5. Indicate the Financial and Sustainable point(s) stated in the Pisa Declaration that concern your organization most**

- ☐ Identifying available funding for research involving grey literature
- ☐ Increased support for collection development and long term preservation of grey literature
- ☐ Increased investment in infrastructure and new technologies for accessing and using print and digital grey literature
- Other (not specifically mentioned in the Declaration)

**6. Indicate the Technical point(s) stated in the Pisa Declaration that concern your organization most**

- ☐ Strategies to tackle link rot and enhance the stability and accessibility of online content
- ☐ Systems for linking data and other non-textual content to their grey literature publications together with interoperability standards for sharing grey literature
- Other (not specifically mentioned in the Declaration)

**Two years on, the Pisa Declaration has been translated and published in nearly twenty languages: Armenian, Bulgarian, Croatian, Czech, Dutch, French, German, Greek, Hindi, Hungarian, Japanese, Korean, Macedonian, Russian, Serbian, Spanish, Tagalog, and Turkish – all of which are online accessible via the GreyGuide Repository and Portal.**

**7. Is there a language(s) not listed above in which the Pisa Declaration should be translated and published? If so, please indicate here.**

**8. Have you had the opportunity to promote public awareness to the Pisa Declaration? If so,**

- ☐ By publishing it on your organization's website
- ☐ By mentioning it at a conference or meeting
- ☐ By referencing it in a publication
- ☐ By translating it in one of the above mentioned languages
- Other (please specify)

**Feedback from those who endorsed the Pisa Declaration is thought to provide an up-to-date roadmap serving diverse communities involved in research, publication and the management of grey literature resources**

**9. Please take a moment to record any comments, recommendations, amendments, or additions you consider worthwhile for the Pisa Declaration to further benefit policy development for grey literature resources.**

**10. Your name and email address**

## A Geographical Visualization of GL Communities: A Snapshot

Gabriella Pardelli, Sara Goggi, Roberto Bartolini, Irene Russo, and Monica Monachini,  
Istituto di Linguistica Computazionale "A. Zampolli", CNR Pisa, Italy

### 1. Introduction

***"Today, in the spirit of science, grey literature communities are called to demonstrate their know-how and merit to wider audiences"***

[Farace Dominic J., 2011].

This quotation stresses the important role of the several international organizations in producing and disseminating knowledge in the field of Grey Literature (GL): the paper aims to provide a first snapshot of the geographical distribution of GL organizations and their participation to the annual International Conference on Grey Literature over the time (in the period from 2003 to 2015. See List of Conferences on Table 2 ).

Nowadays a visual representation of data is often associated with the traditional statistical graphs, in particular for representing complex phenomena by means of maps and diagrams, which allow a deeper and more focused analysis of the data. In our case the geographical representation of stakeholders in government, academics, business and industry aims at visualizing the GL community across the globe: it concerns 674 organizations which over the years have contributed to the development of a common vision on the most pressing issues of the field by using new paradigms such as Open Access and the social networks.

Given this scenario the GL Community is visualized by name and country of the organization and by year, as documented by the GL List of Participating Organizations published in the thirteen GL Program Books which can be found on the GreyGuide<sup>1</sup> site. The results are presented in the form of visual graphs, which confirm the international flavor of this field.

### 2. GL Community today

The inter-disciplinary dimension, the specialized themes and the geographical dislocation of its stakeholders are the requisites of attraction of the international Grey Literature community and these elements can represent an advantage for the whole field. Over the years universities, research centers, governmental bodies and industries presented their own research experiences, the technological solutions tested and/or adopted thus facilitating the introduction of new paradigms as well as the giving up of obsolete models.

#### 2.1 GL Community in the world

The most remarkable figure to be reported is the substantial participation of US organizations to the GL conferences: not surprisingly the country stands at the top of the list with 216 organizations<sup>2</sup>. This is the chronological distribution of the American presence to the conference: 2004>32; 2005>10; 2006>25; 2007>10 2008>18; 2009>34; 2010>11; 2011>23; 2012>9; 2013>6; 2014>18; 13; 2015>13.

As shown in graphs 1 and 2, the participation of European institutions of the field is large; but over the years the community has also taken advantage of contributions from institutions coming from somehow countries such as Fiji, Finland, Gambia, Georgia, Iceland, Iran, Latvia, Luxembourg, New Zealand, Romania, Saudi Arabia, Serbia, West Indies, Zimbabwe.

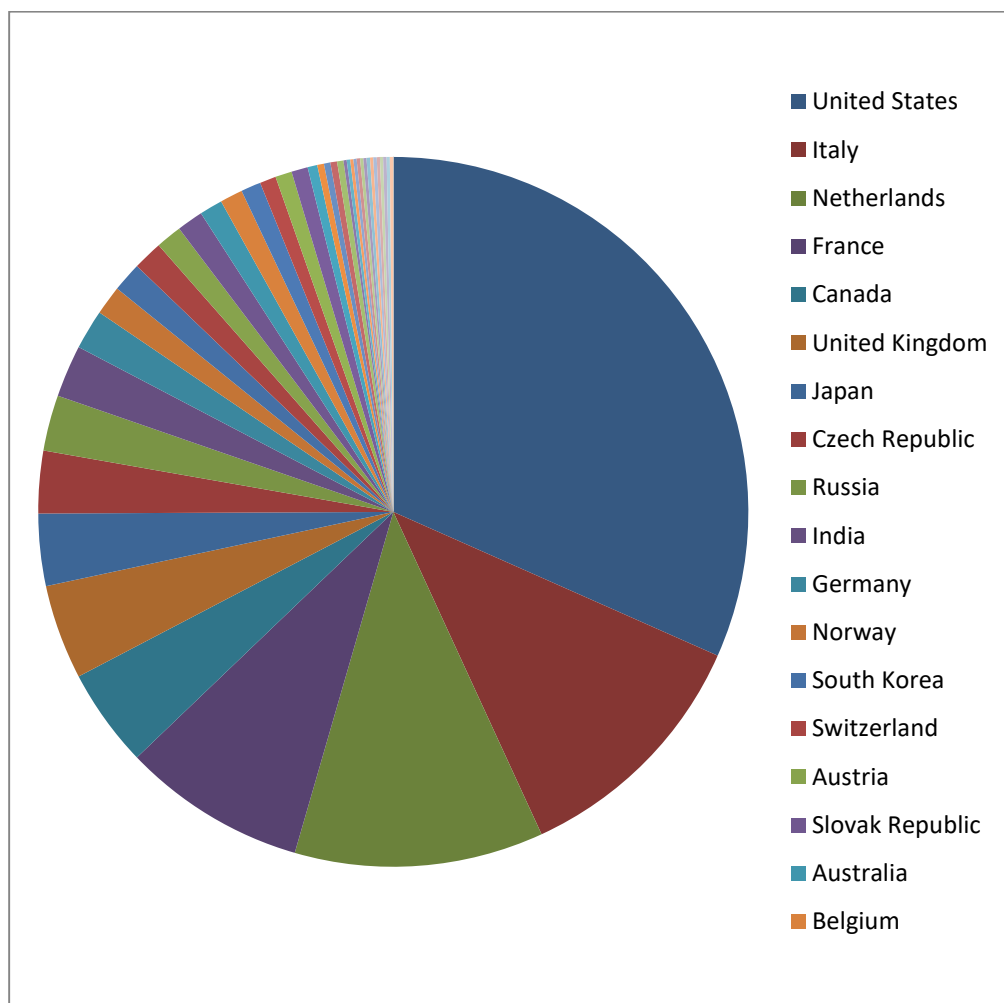
Here below Table 1 shows all the countries which had at least one participating institution from GL5 to GL17.

<sup>1</sup> <http://greyguide.isti.cnr.it/>

<sup>2</sup> The participating organizations have been counted for each single participation (that is, they have been re-counted if attending more than one edition of the conference).



Graph 1: Visualization of GL Community in the world



Graph 2: GL Countries



COUNTRY	NUMBER OF PARTICIPATIONS	COUNTRY	NUMBER OF PARTICIPATIONS
Algeria	2	Latvia	1
		Luxembourg	5
Australia	7	Netherlands	77
Austria	8	New Zealand	1
Belgium	7	Norway	9
Brazil	2	Poland	5
Cameroon	1	Romania	1
Canada	30	Russia	17
Czech Republic	19	Saudi Arabia	1
Denmark	1	Serbia	1
Fiji	1	Slovak Republic	8
Finland	2	Slovenia	6
France	56	South Africa	3
Gambia	1	South Korea	9
Georgia	1	Spain	2
Germany	12	Switzerland	9
Greece	5	Uganda	1
Iceland	1	United Kingdom	29
India	16	United States	212
Iran	1	West Indies	1
Italy	77	Zimbabwe	1
Japan	22		

Table 1: Number of participations by Country.

1.	2003 GL5 Amsterdam, "Grey Matters in the World of Networked Information"
2.	2004 GL6 New York, "Work on Grey in Progress"
3.	2005 GL7 Nancy, France "Open Access to Grey Resources"
4.	2006 GL8 New Orleans, "Harnessing the Power of Grey"
5.	2007 GL9 Antwerp, "Grey Foundations in Information Landscape"
6.	2008 GL10 Amsterdam, "Designing the Grey Grid for Information Society"
7.	2009 GL11 Washington D.C., "The Grey Mosaic: Piecing It All Together"
8.	2010 GL12 Prague, "Transparency in Grey Literature, Grey Tech Approaches to High Tech Issues"
9.	2011 GL13 Washington D.C., "The Grey Circuit, From Social Networking to Wealth Creation", Library of Congress, December 5–6
10.	2012 GL14 Rome, Italy, "Tracking Innovation through Grey Literature", National Research Council, CNR, November 29–30
11.	2013 GL15 Bratislava, Slovak Republic, "The Grey Audit, A Field Assessment in Grey Literature", December 2–3
12.	2014 GL16 Washington D.C. "Grey Literature Lobby, Engines and Requesters for Change", December 8–9
13.	2015 GL17 Amsterdam, "A New Wave of Textual and Non-Textual Grey Literature", December 1–2

Table 2: List of GL conferences.

## 2.2 GL Community and Genre

The information about the entire set of papers presented at the GL conferences in the period 2003-2015 is available on the GreyGuide repository. In addition to the nationality of the authors, we lately decided to extrapolate the information on their gender as well: this type of analysis is usually difficult due to the various ways of writing the names (full name, initials, middle initials). It was therefore needed a cleaning process for being able to divide the authors by gender: for disambiguating the initials of the first names we used portals such as OpenGrey<sup>3</sup>, GreyNet<sup>4</sup>, TextRelease<sup>5</sup> (the section 'Who is in Grey Literature') and GreyGuide. Quite a number of authors have been identified by accessing the repository Google Scholar Citations and social networks such as LinkedIn; publishing houses online

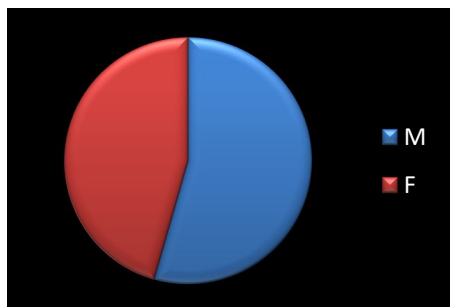
<sup>3</sup> A System for Information on Grey Literature in Europe. <http://www.opengrey.eu/>

<sup>4</sup> <http://www.greynet.org/>

<sup>5</sup> <http://www.textrelease.com/>

libraries and digital archives (i.e., @rchiveSIC<sup>6</sup>) have been consulted as well. For example, the name 'Goggi, S.' has been easily retrieved from various repositories, disambiguated as 'Goggi, Sara' and annotated as "F" (*female*) but with other names the retrieval has been possible only thanks to the pictures available on the web (see the example of the name 'Khan, M.T.M.', identified as "M" (*male*) with the help of his picture on LinkedIn). A few unresolved cases have been annotated with a question mark. Graph 3 shows the participation by gender to GL conferences: of course the names of those who participated to more than one edition – or even presented more than one paper at the same conference – have been counted only once.

The results talk about a sort of gender balance with a slight preponderance of men: female 201, male 240.



Graph 3: GL authors' genders

### 3. Data Extraction and Analysis

Prior to the data analysis, a normalization process of the information provided by the authors was needed: the manual cleaning was mainly carried out on the names of those affiliations which varied both graphically and linguistically over the time. Changes concern also acronyms and abbreviations – often missing – and other types of information which help specifying the organization. Table 1 lists ten examples of variation of names.

N°	Name 1	Name 2	Variant
1	University of Pretoria – UP	University of Pretoria, UP	graphic sign (gs)
2	University of Ljubljana	University of Ljubljana, UNI-LJ	acronym (ac)
3	University of Illinois, UIUC	University of Illinois at Urbana-Champaign	information of a second level (isl)
4	Science & Technology Facilities Council, STFC	Science and Technology Facilities Council, STFC	graphic sign (gs)
5	Biblioteca Centrale "G. Marconi" CNR	Biblioteca Centrale "G. Marconi", CNR Also Biblioteca Centrale "G. Marconi"; CNR Also Consiglio Nazionale delle Ricerche, Biblioteca Centrale	graphic sign (gs)
6	Centre of Information Technologies and Systems of Executive State Authorities	Centre of Information Technologies and Systems, CITIS	information of a second level (isl) + acronym (ac)
7	Data Archiving and Networked Services, DANS	Data Archiving and Networked Services, DANS-KNAW	acronym (ac)
8	Koninklijke Nederlandse Akademie van Wetenschappen – KNAW	Royal Netherlands Academy of Arts and Sciences, KNAW	Langue (la)
9	Institute of Computational Linguistics, ILC-CNR	Istituto di Linguistica Computazionale, ILC	Langue (la)
10	University of Bergen	University of Bergen, Research Documentation Unit – UIB	information of a second level (isl)

Table 3 – Examples of names' variations

<sup>6</sup> Archive Ouverte en Sciences de l'Information et de la Communication. <https://archivesic.ccsd.cnrs.fr/>

Normalization has been therefore essential for the correct identification of the affiliations and the resulting calculation of their presence over the years: the same affiliations are present in different years with different authors and authors of the same institution can even present various activities at the same edition of the conference.

The assessment of the number of affiliations which participated to the GL series over this time span required their ordering in alphabetic tables and then checks on several web sites for disambiguating the graphic and linguistic variants.

Normalization was also performed on languages, for example: Nederlands Instituut voor Wetenschappelijke Informatiediensten – NIWI > Netherlands Institute for Scientific Information Services, NIWI; Consiglio Nazionale delle Ricerche – CNR > National Research Council, CNR; Istituto di Ricerche Popolazione e le Politiche Sociali, IRPPS > Institute of Research on Population and Social Policies, IRPPS.

#### 4 Conclusion and Future Work

With this work we carried out a first introductory mapping of the researchers involved in the international community of Grey Literature, pertaining only their geographical origin. As said, this is still preliminary as other aspects would deserve attention too and thanks to the GreyGuide repository – a sort of international ‘observatory’ for GL – further investigations could be performed with different methodologies in order to achieve new goals.

For example, we plan to more deeply investigate the structure of the research community through the graphs of both collaboration and citation amongst authors, as a sort of social network. This process will help identifying groups of researchers who publish together or usually cite each other.

It is certainly true that the idea of visualizing the tendency of national participation to international conferences – and in our case to the GL series – over the years could be applied to similar research in grey literature thus stimulating further visual surveys from scholars in the field.

#### Essential References

DEL GRATTA R., FRONTINI F., MONACHINI M., PARDELLI G., RUSSO I., BARTOLINI R., KHAN F., SORIA C., CALZOLARI N. (2016). LREC as a Graph: People and Resources in a Network. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odiijk, Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris. Pages 23-28.  
<<http://www.lrec-conf.org/proceedings/lrec2016/index.html>>

DEL GRATTA R., FRONTINI F., MONACHINI M., PARDELLI G., RUSSO I., BARTOLINI R., GOGGI S., KHAN F., QUOCHI V., SORIA C., CALZOLARI N., (2015). Visualising Italian Language Resources: a Snapshot. In Cristina Bosco, Sara Tonelli, Fabio Massimo Zanzotto (eds.), *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*. Accademia University Press, Trento. Pages 100-103.

FARACE DOMINIC J. (2011). Foreword. The Grey Circuit. From Social Networking to Wealth Creation. In *GL13 Conference Program and Abstracts: Thirteenth International Conference on Grey Literature*. GreyNet, Grey Literature Network Service. – Amsterdam : TextRelease, 2011. Page 3.

ZEQIAN S., OGAWA M., SOON TEE TEOH, KWAN-LIU MA (2006). BiblioViz: a system for visualizing bibliography information. In *Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation - Volume 60 (APVis '06)*, Kazuo Misue, Kozo Sugiyama, and Jiro Tanaka (Eds.), Vol. 60. Australian Computer Society, Inc., Darlinghurst, Australia, Australia. Pages 93-102.

## Teaching and Learning about Grey Literature

### *Results from a Poster Presented at the 18th Grey Literature Conference*

Lynne M. Rudasill,  
Center for Global Studies,  
University of Illinois at Urbana-Champaign, United States

Formal opportunities that provide teaching and learning related to the creation, dissemination and conservation of grey literature are few outside of workshops provided by the GreyNet organization. Research by Rabina revealed that library school students in the U.S. received most of their training in the use of grey literature on the job rather than as a part of a designated course. Further, she posits that a cross-curricular approach would most closely reflect the broad scope of grey literature should it be more prominently placed in the curriculum of library and information programs. (Rabina, 2011) With this in mind, the author proposed a poster session on the topic of how we can improve understanding of the creation, access, and preservation of grey literature in I-Schools and in general. The poster included four related questions:

- How do we improve understanding through formal learning?
- How do we expand the general audience and participation?
- What are our learning objectives?
- What methods should we use to disseminate information on grey literature more broadly?

We then provided “sticky notes” to participants and invited them to provide us with their comments. Participation was quite robust with approximately twenty-five individuals participating in discussions about the topic and providing notes on their ideas. The results of the conversations and notes can be broken down into the following topic areas which have been ranked by times mentioned.

#### **Results**

Most of the comments related directly to paths to promote grey literature more broadly. Most frequently mentioned were freely available webinars to provide outreach and training. In a closely related area, promotion of grey literature through video tutorials and other social media was also seen as high priority. A number of individuals were surprised that a robust listserv was not available for the discussion of grey literature and surrounding issues and suggested that one be developed.

A second area of major concern was the preservation and indexing of grey literature. Several individuals strongly supported the concept of better search capabilities for grey literature through development of better metadata for these items. These discussions also included suggestion of the development of a Wikipedia type of open access resource for the storage of otherwise non-indexed articles, reports, etc. Related to this was the promotion of case studies using grey literature and the creation of a bibliography for newcomers.

Finally, there was interest in incorporating more information on grey literature in library and I-School courses and also bringing this information into the undergraduate classroom. We should also be sure that our scholars are aware of this type of literature. Mention was also made of the importance of introducing public library users to grey lit as well as providing information on FOIA laws. A variety of other suggestions were made to improve the “branding” of grey literature, develop alliances with other professional organizations, and perhaps even develop professional degree related to grey literature.

#### **Conclusion**

The GreyNet LIS Education and Training Committee headed by Marcus Vaska will be actively working on teaching and learning for grey literature in the coming year. Please let us know if you have any other suggestions to raise the profile and use of grey literature.

## Reference

Rabina, D. L. (2011). Course and Learning Objective in the Teaching of Grey Literature: The Role of Library and Information Science Education. *Grey Journal (TGJ)*, 7(3), 160-165

## Appendix – Transcript of the results from the conference poster:

There should be webinars for users on the use and value of GL

Quarterly webinars like “Help I’m an accidental government librarian!” on various topics

- Massive indexing effort
- Build researcher/student awareness
- Digitization efforts (ex. See NYPL example) also as institutions create grey lit, they should archive PDF copies

- Hands on training
- Concentrate on
  - Collection development
  - Users/customers
- Offer professional degree/status
- Accreditation? Why not
- Incorporate gl repositories in bigger search engines

Depositing GreyLit in a repository for discovery and preservation and harvesting to aggregators

What is the difference between GL & oA Users better understand the latter.

We need a listserv!!

Teach grey literature in LIS programs

“Streaming” grey literature

1. Use real life inquiries
2. Reach out to public libraries
3. How do I find grey literature in health and medicine
4. Free workshops in the PL and other LIS mtgs,

Possibly add a listserv? (how can these standards be disseminated?)

Provide a bibliography for newcomers?

Why reinforce GL – instead promote metadata & more comprehensive access.

How to evaluate the authoritativeness of information contained in GL. Does GL have an advantage (e.g. timeliness) over traditional professional publications?

What is the brand?

When talking to different communities, how do we describe what grey lit is?

Improve researcher understanding of locating grey literature.

- Training
- Sessions.

The term GL is confusing to many professionals, the terminology using data vocabulary is more attractive.

Alliance w/ other professional organizations?

Case studies

Presenting grey literature on social media platforms

YouTube

Make a video narrated with simple stick figures and examples

It is not important for end users to know what is GL. For them is important to get document, dataset, etc. without the knowledge that is GL. Let's do the GL more accessible = store it, preserve it and make it searchable.

The best way to expand our audience is to push librarians and faculty at universities to force students to use grey literature. It also be influenced in undergrad library classes so that students can use the searching skills their entire college experience.

- Share use cases
- Promote open use of discussion lists
- Trainings on transparency/FOIA law + the like

Provide more broad search

Capabilities i.e., websites, Google  
Outside normal channels

If institutions and individuals are storing data sets why is there not an open source that users can upload information on topics. Such as a Wikipedia for non-indexed articles.

Library Schools

Need to start with the library school administrators who can expand their program curriculum



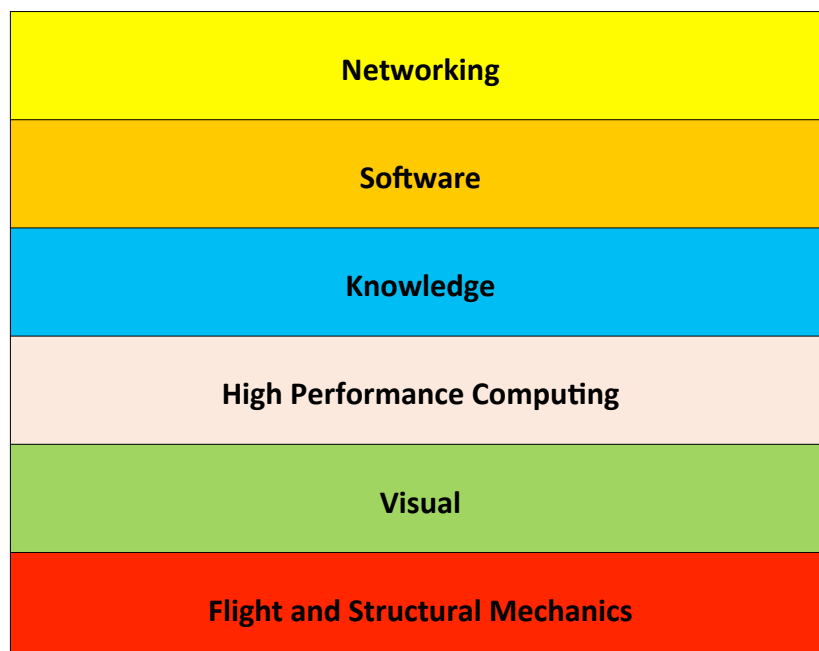
# Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

an Institute of the National Research Council of Italy **CNR**

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

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

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



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

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



## A terminological “journey” in the Grey Literature domain

**Roberto Bartolini, Gabriella Pardelli, Sara Goggi,**

CNR, Istituto di Linguistica Computazionale, “Antonio Zampolli”

**Silvia Giannini and Stefania Biagioni,**

CNR, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, 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." (Cabr , 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 (Pejřov  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)<sup>1</sup>; 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<sup>2</sup> (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

<sup>1</sup> Kindly authorised from Greynet International, <http://www.greynet.org/>.

<sup>2</sup> CNR stands for National Research Council, Italy, <https://www.cnr.it/>.

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.

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<sup>3</sup>.

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.

<sup>3</sup> 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).

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”.

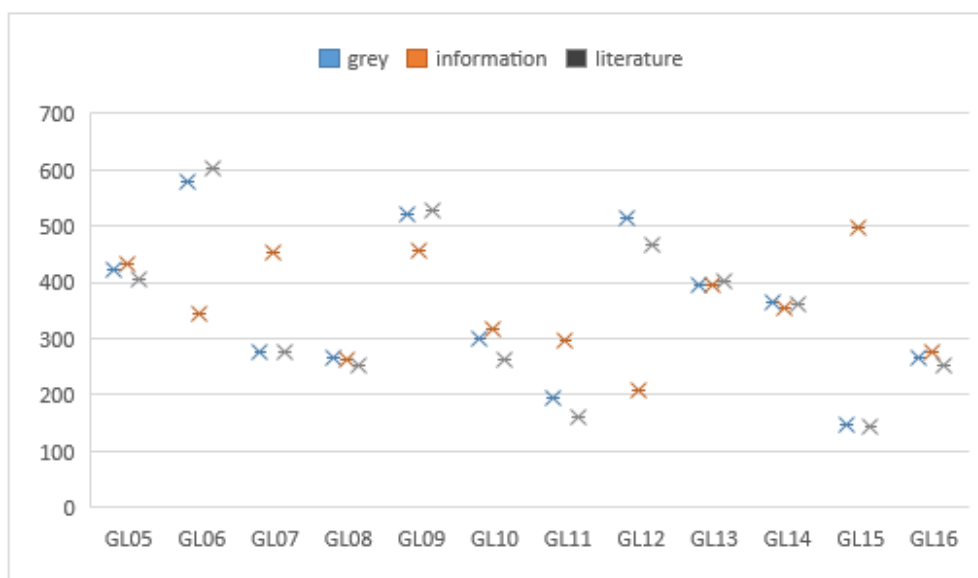
Although “Information” is the monogram with the highest frequency in the entire corpus (4302 occurrences) it occupies the second place in the table representing the high segment terms : the first position belongs to “literature”, one of the more content-related terms, while “grey” shows 3851 occurrences 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”.

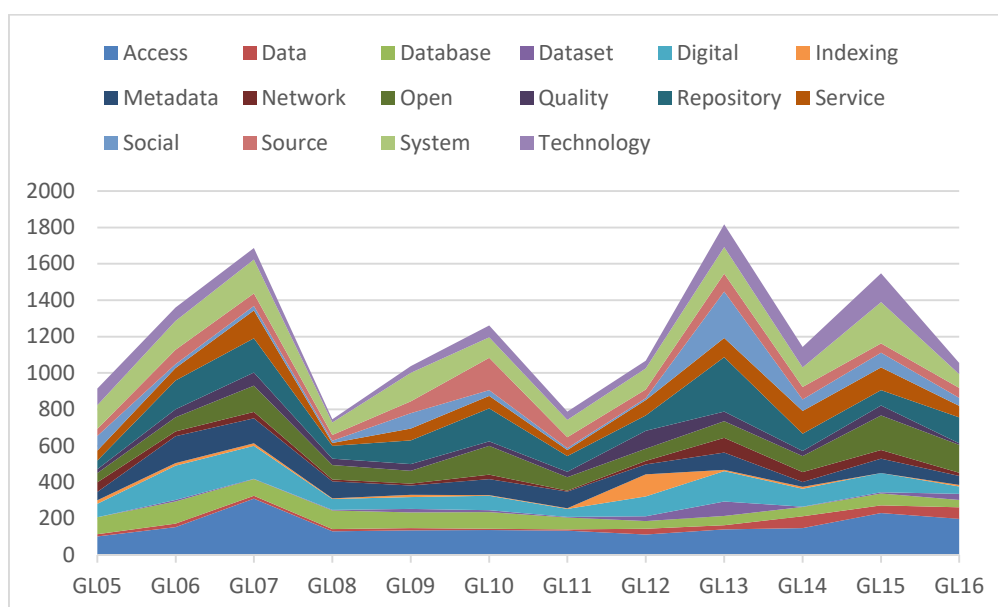
In addition to the pair “grey literature” the term “literature” appears in the following bigrams: “medical literature”, “conventional literature”, “type literature”, “literature repository”, “literature collection”, “repository literature”, “use literature”, “access literature”, “journal literature”, “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

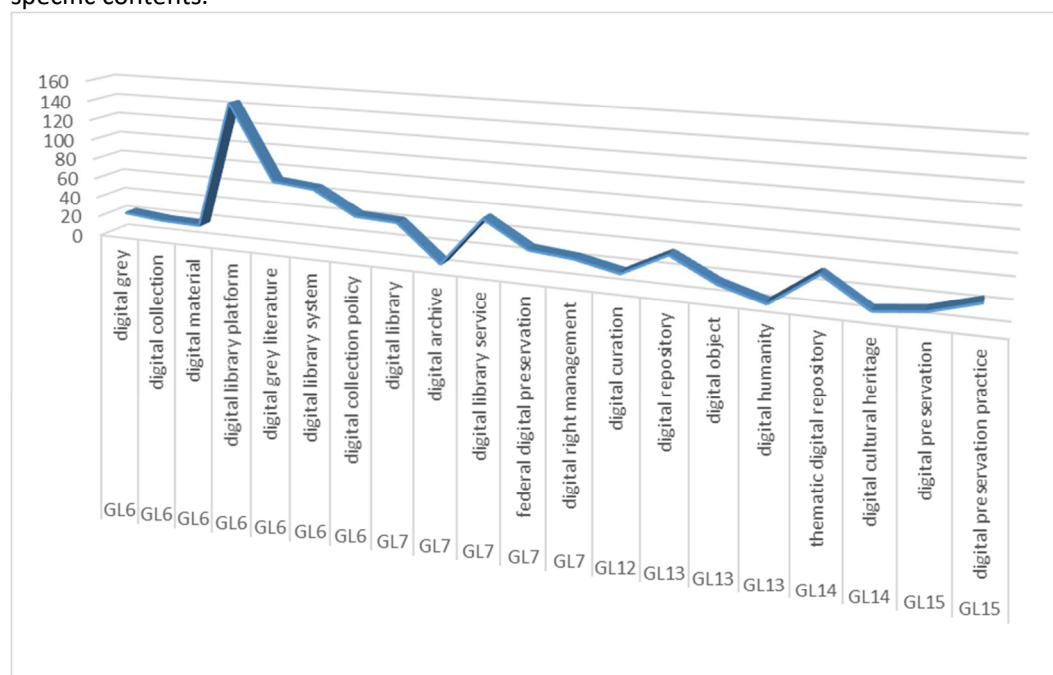
Graph 2 shows the trend of the selected terms over the years: it is clear that - with the exemption of “indexing” and “dataset” – all of them are occurring in each GL glossary. Generally, there are monograms which seem to be constantly used and therefore their trend over the time is stable (e.g. access, 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”.

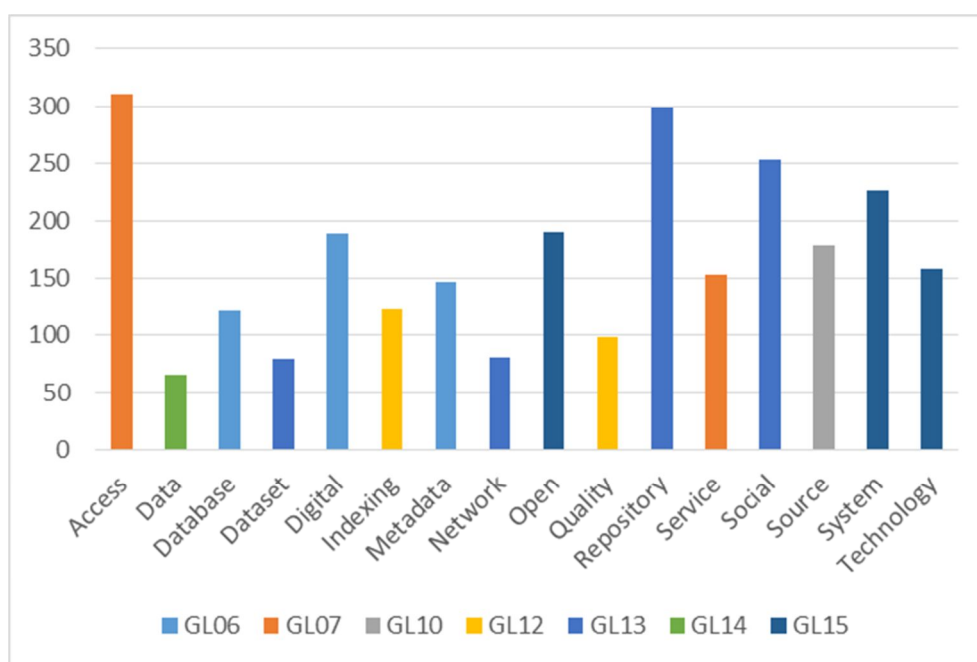
In our context the term “technology” is related to telematics and computer science applications to the documentary field: the single term is paired with “information technology” while the trigram is “technology information system”. Information management is represented by nouns such as “system” and “source”: 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 curious, 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

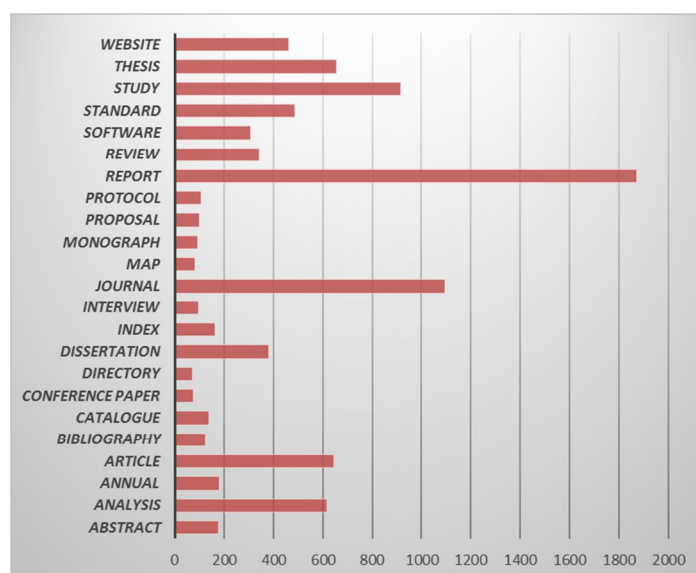
This last chapter is dedicated to the terminology used for describing the types of documents occurring in the corpus.

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, Ministerio 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. Megerdooian 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 Coastal 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 Reviews
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 Cycle
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 terms	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 terms	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 Paper	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 Document	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 Program					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 Material	2						1	1	1				5
Informative Material									1		3	1	5
Normative Document					2	1					2		5
Anthology						1	1		1	1			4
Research Plan									1	1	2		4
Syllabus					3								3
Tertiary Source	1				1			1					3
Corporate Literature											2		2
Habilitation Thesis									1		1		2
Image Material											2		2
Legal Document							1				1		2
Guidebook		1											1
Technical Documentation		1											1

Table 4

## Altmetrics and Grey Literature: Perspectives and Challenges

Joachim Schöpfel, GERiCO Laboratory, University of Lille

Hélène Prost, CNRS 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<sup>1</sup>. 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<sup>2</sup>.

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<sup>3</sup>.

<sup>1</sup> See for instance Ingwersen et al. 2014, also for similar, older studies

<sup>2</sup> The methodological problems to identify theses in bibliographic databases in Larivière et al. (2008) confirm the situation

<sup>3</sup> 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>

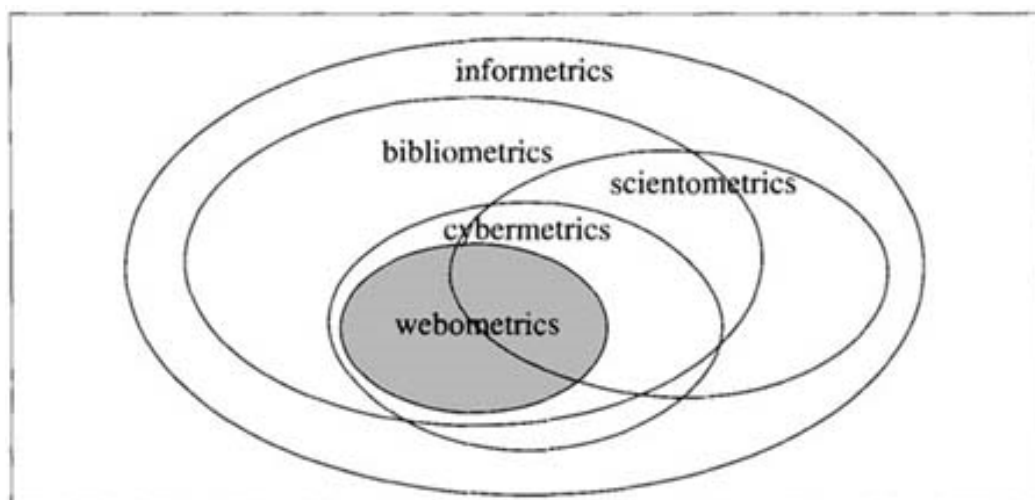


Figure 1: Webometrics in the field of library and information sciences  
(source: Björneborn & Ingwersen 2004)

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<sup>4</sup>. The term was introduced by Jason Priem from Chapel Hill in 2010, in a tweet published on the 29<sup>th</sup> September 2010: “I like the term #articlelevelmetrics, but it fails to imply \*diversity\* of measures. Lately, I'm liking #altmetrics”<sup>5</sup>. 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<sup>6</sup> 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<sup>7</sup>, DORA fostered the awareness for altmetrics and became a reference for the debate, research and development in the field.

<sup>4</sup> See comprehensive reviews by Erdt et al. (2016) and Sugimoto et al. (2016)

<sup>5</sup> <https://twitter.com/jasonpriem/status/25844968813> by @jasonpriem

<sup>6</sup> <http://altmetrics.org/manifesto/>

<sup>7</sup> <http://www.ascb.org/dora/> accessed 7 September 2016

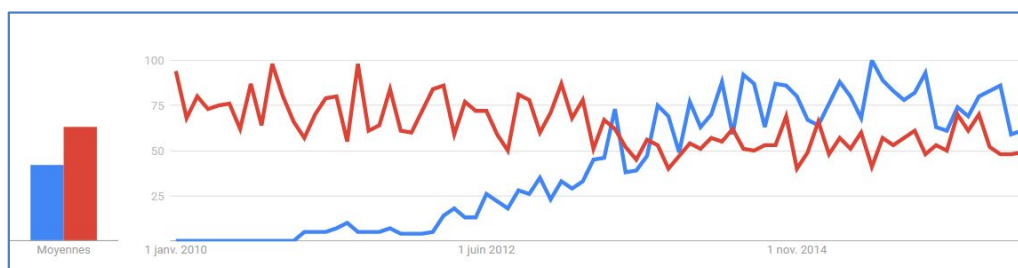


Figure 2: Altmetrics (blue) on Google Trends, compared to scientometrics (red) (2010-2016)<sup>8</sup>

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.

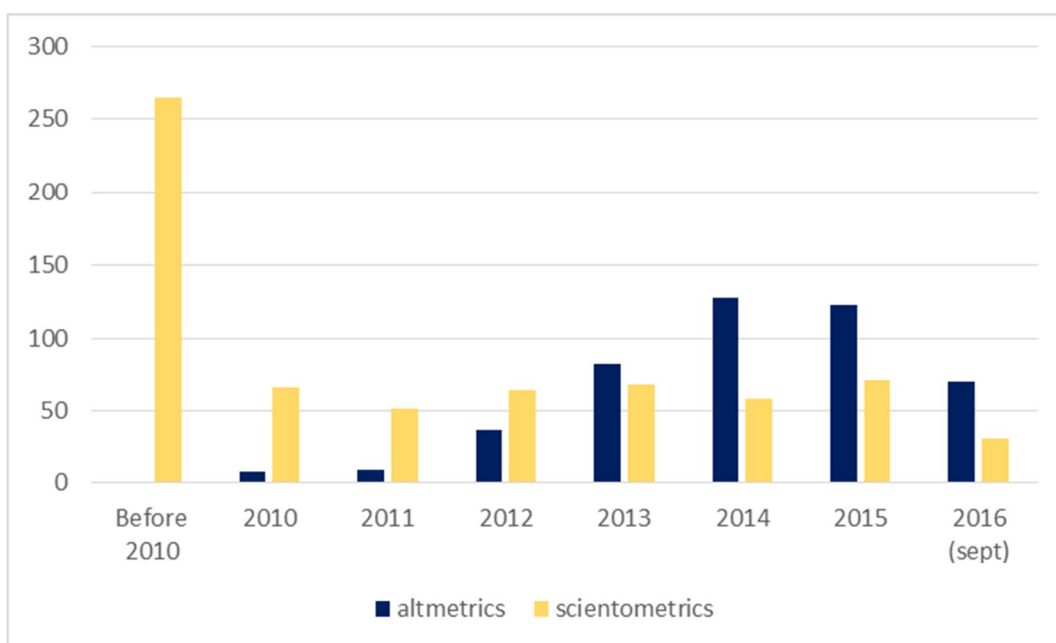


Figure 3: Publications on altmetrics and scientometrics<sup>9</sup>

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<sup>10</sup>, 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).

<sup>8</sup> Data source: Google Trends [www.google.com/trends](http://www.google.com/trends) accessed Sept 3, 2016

<sup>9</sup> Data source: Google Scholar <https://scholar.google.fr> allintitle: altmetrics (or scientometrics), accessed Sept 5, 2016

<sup>10</sup> See <https://www.altmetric.com/about-altmetrics/what-are-altmetrics/>

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”<sup>11</sup>.

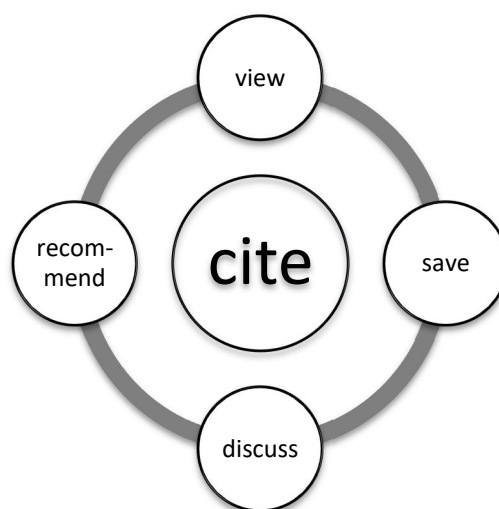


Figure 4: Altmetrics areas of assessment

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”<sup>12</sup>, 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

<sup>11</sup> <https://www.altmetric.com/blog/altmetrics-collection-development/>

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



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.

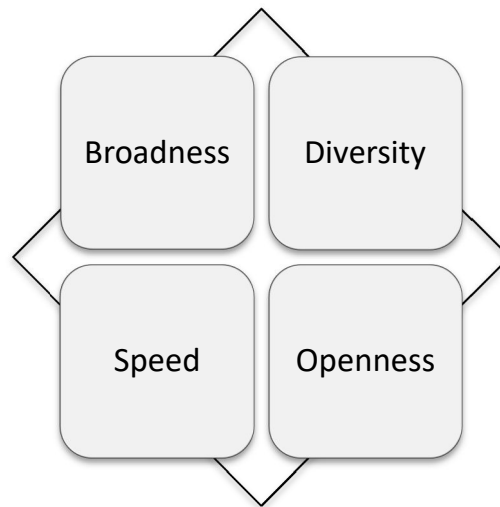


Figure 5: Benefits of altmetrics (source: Bornmann 2014)

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.

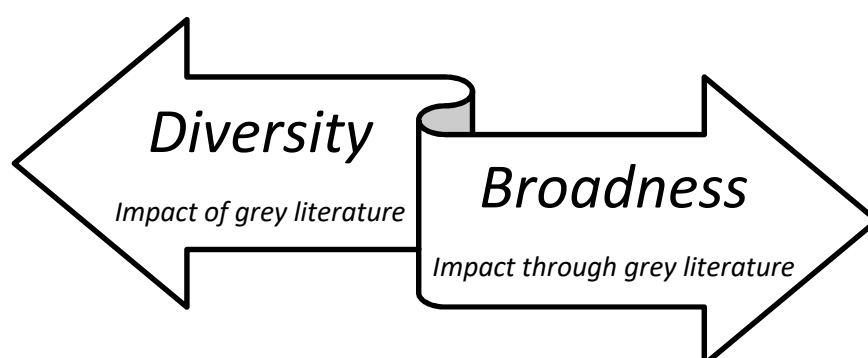


Figure 6: Double potential of altmetrics for grey literature

- **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<sup>13</sup>. 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”<sup>14</sup>.

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.

<sup>13</sup> GS, MAS, ArnetMiner, Mendeley, CiteULike, Zotero, ReaderMeter, ImpactStory, SURE2

<sup>14</sup> See the Directory of Open Access Repositories OpenDOAR, available at <http://www.opendoar.org/>

**HAL<sup>15</sup>**

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<sup>16</sup>. 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 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<sup>17</sup>**

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).

<sup>15</sup> HAL = Hyper articles en ligne <https://hal.archives-ouvertes.fr/>

<sup>16</sup> Accessed 9 September 2016

<sup>17</sup> <https://figshare.com/>

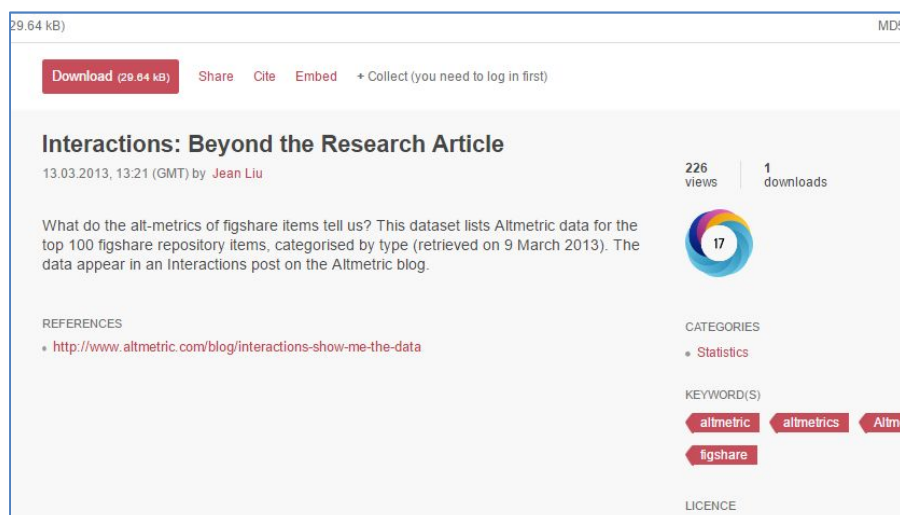


Figure 8: USAGE statistics and altmetrics in figshare

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<sup>18</sup>. 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.

#### IRUS-UK<sup>19</sup>

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<sup>20</sup>

<sup>18</sup> 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)

<sup>19</sup> Institutional Repository Usage Statistics UK <http://www.irus.mimas.ac.uk/>

<sup>20</sup> All IRUS-UK figures and statistics accessed 7 September 2016

**CORE<sup>21</sup>**

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<sup>22</sup> and their impact on scientific communication<sup>23</sup>, 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.

The screenshot shows a ResearchGate profile page for a conference paper. At the top, there are links for 'See all > 0 Citations' and 'See all > 155 Reads', along with 'Download' and 'Request feedback' buttons. The title of the paper is 'Altmetrics Tools: What's out there and what they can do for you?'. Below the title, it is identified as a 'Conference Paper' from 'October 2015' with a DOI of '10.13140/RG.2.1.2961.9686'. The conference is noted as the '17th Conference of Australian Research Management Society, At Singapore'. Two authors are listed: '1st Yew Boon. Chia' with a score of 1.54 and '2nd Joan Wee' with a score of 0.73, both from Nanyang Technological University. An 'Abstract' section follows, starting with 'In the last two years, many publishers have jumped on the altmetrics bandwagon with several high profile acquisitions and the introduction of new services. From these activities, two major altmetrics providers have emerged. Though they have different strengths and weaknesses, particularly in the depth of their coverage of different subject domains, there is also a considerable amount of overlap in their sources. To select the "right" altmetrics provider, it is...'. A small blue icon with a plus sign is at the end of the abstract text.

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

<sup>21</sup> Connecting REpositories <https://core.ac.uk/>

<sup>22</sup> 41m accounts in Academia, 10m accounts in ResearchGate

<sup>23</sup> 14.7m papers in Academia, 100m papers and other items in ResearchGate (24% papers with full-text), mostly STM



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.

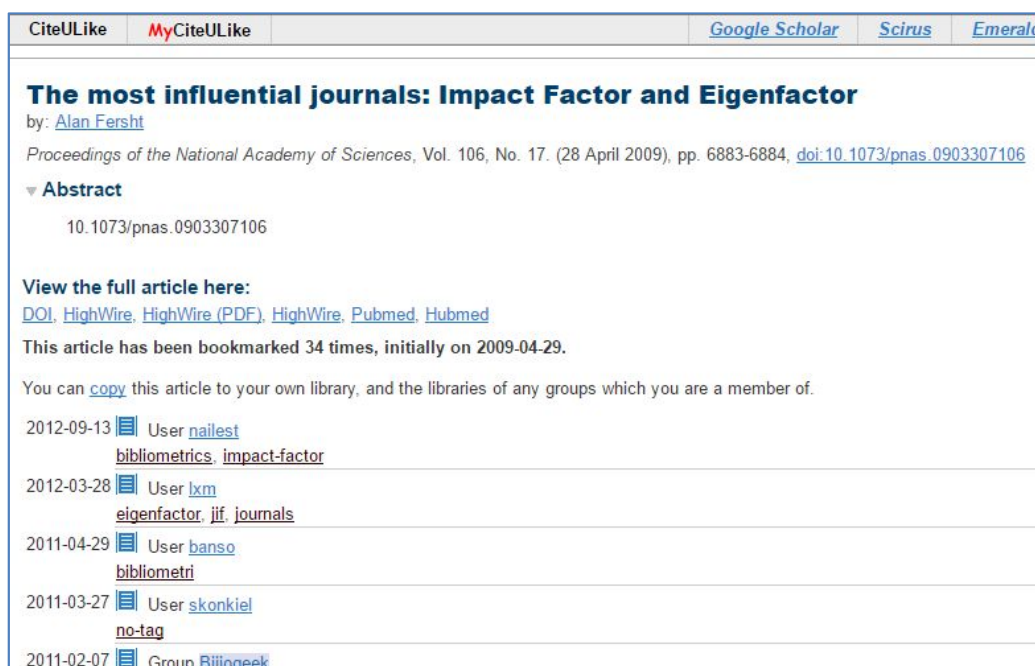


Figure 11: Number of bookmarks in CiteULike

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.

The screenshot shows the Microsoft Academic Search interface. The search bar at the top contains the term 'altmetrics'. Below the search bar, it indicates '1-8 of 452 results for altmetrics (0.3 seconds)'. On the left side, there are filters for 'Date Range' (2007 to 2016), 'Author' (Stefanie Haustein, Jason Priem, Mike Thelwall, Vincent Larivière, Cassidy R. Sugimoto), 'Affiliation' (University of Wolverhampton, École Normale Supérieure, Indiana University Bloomington, University of North Carolina at Chapel Hill, VU University Amsterdam), and 'Field Of Study' (Computer Science, World Wide Web, Data mining, Publishing, Social media). The main results area shows three entries:

- Altmetrics: Value all research products** (2013, *Nature*, volume 493, issue 7431, pp 159-159) by Heather Piwowar. It is cited 168 times\*.
- Do Altmetrics Work? Twitter and Ten Other Social Web Services** (2013, *PLOS ONE*, volume 8, issue 5) by Mike Thelwall (University of Wolverhampton), Stefanie Haustein (École Normale Supérieure), Vincent Larivière (École Normale Supérieure), Cassidy R. Sugimoto (Indiana University Bloomington). It is cited 250 times\*.
- Altmetrics in the wild: Using social media to explore scholarly impact** (2012, Jason Priem, Heather A. Piwowar, Bradley M. Hemminger). It is cited 12 times.

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.

The screenshot shows a Google Scholar search result for a workshop paper. The title is '[doc] How consistent are altmetrics providers? Study of 1000 PLOS ONE publications using the PLOS ALM, Mendeley and Altmetric.com APIs'. The authors listed are Z Zahedi, M Fenner, and R Costas. The paper is from altmetrics 14, 2014, and is available at ndownloader.figshare.com. The abstract states: 'Altmetrics track the impact of scholarly works on the social web. The term was introduced in 2010 (Priem, et al.) as an alternative way of measuring the broader research impact of scholarly outputs using the social web; aimed at enhancing and complementing the more ...'. The citation count is 'Cité 12 fois'. There are links for 'Autres articles', 'Citer', 'Enregistrer', and 'Plus'.

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).

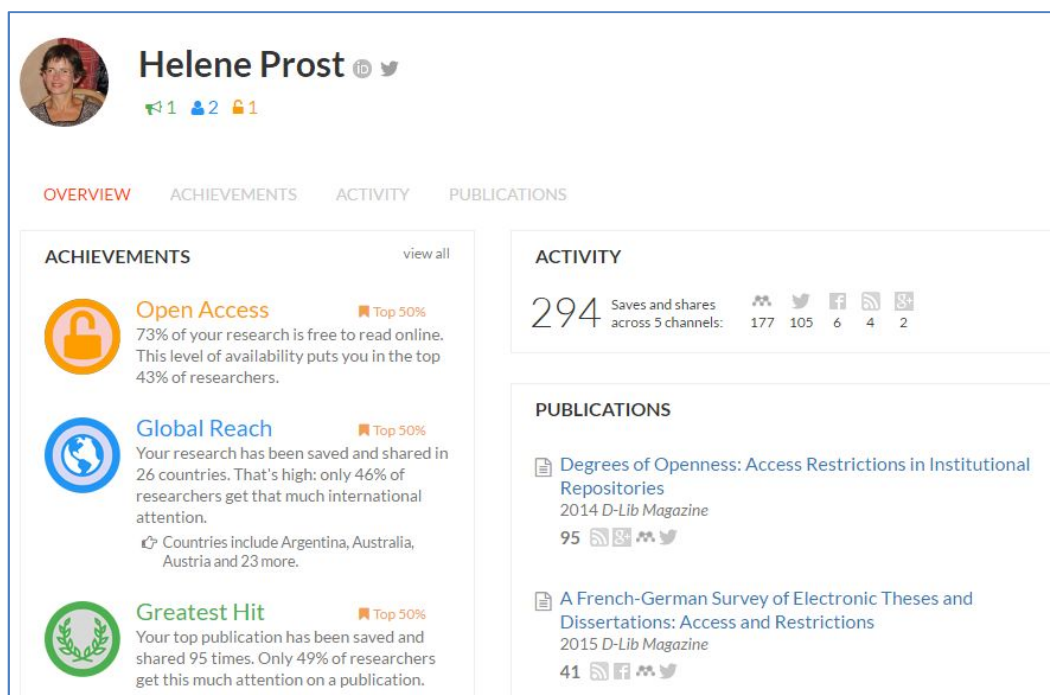


Figure 14: Author's page on ImpactStory

Among these data aggregators one can find Altmeter.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 Altmeter 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. Altmeter 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<sup>24</sup> 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).

<sup>24</sup> <http://www.aminer.org>

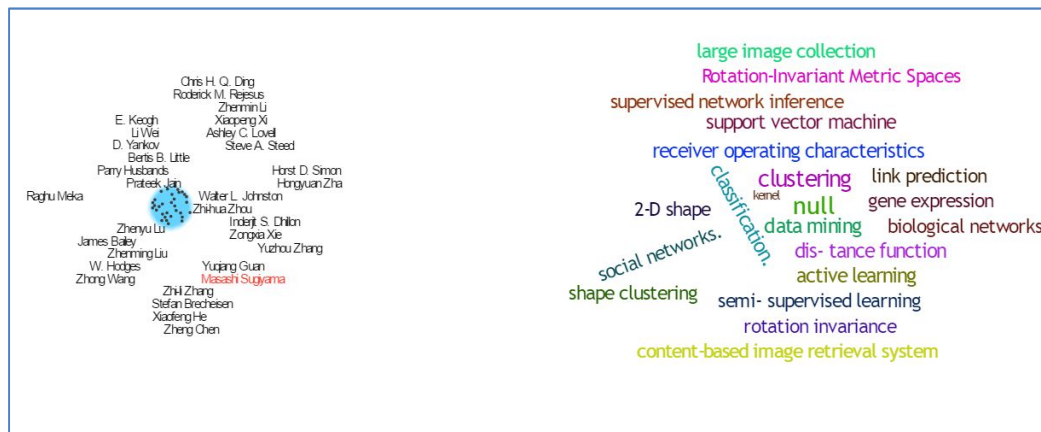


Figure 15: Conference analysis of SIAM International Conference on Data Mining on ArnetMiner

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 Altmeter, grey literature<sup>25</sup> “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, Altmeter has developed a text-mining solution (*Altmeter Policy Miner*) to discover mentions of publications in policy documents on selected websites<sup>26</sup>. 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.

<sup>25</sup> Described as “theses, posters, preprints, patents and policy documents and similar”

<sup>26</sup> 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), Médecins sans Frontières (MSF), NICE Evidence, Oxfam Policy & Practice, UNESCO and World Bank

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<sup>27</sup>. The procedure of 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<sup>28</sup> 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.”

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?

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

<sup>27</sup> In this study NHS Evidence, LG Search, Hein Law Online, EBSCO Business Source Complete, Nexis and Emerald Journals

<sup>28</sup> 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.



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 (...) 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<sup>29</sup> 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<sup>30</sup>, 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<sup>31</sup>, 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<sup>32</sup>. 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

<sup>29</sup> Extraction from the SwePub database of academic publications at Swedish universities <http://swepub.kb.se/>

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

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

<sup>32</sup> Innovations in Scholarly Communication Survey, <http://altmetricsconference.com/who-is-using-altmetrics-tools/>

“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<sup>33</sup>, 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.

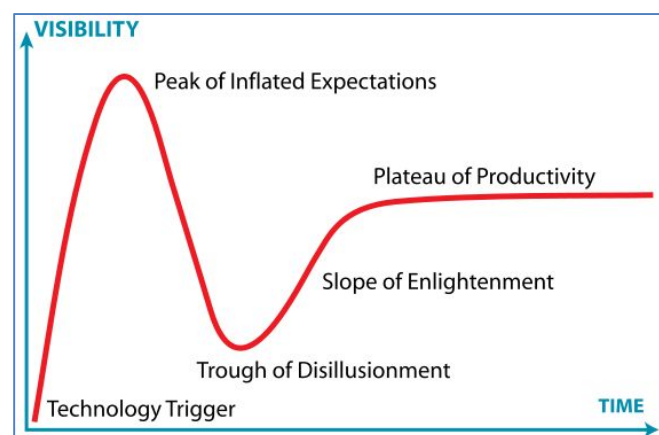


Figure 16: Gartner’s Hype Cycle (source: Wikipedia<sup>34</sup>)

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

<sup>33</sup> <http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp>

<sup>34</sup> [https://en.wikipedia.org/wiki/File:Gartner\\_Hype\\_Cycle.svg](https://en.wikipedia.org/wiki/File:Gartner_Hype_Cycle.svg)



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<sup>35</sup> 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<sup>36</sup>, 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<sup>37</sup>. Perhaps this convergence or happy coincidence may be helpful.

<sup>35</sup> 3:AM Conference, Bucharest 28-29<sup>th</sup> September 2016, <http://altmetricsconference.com/schedule/>

<sup>36</sup> See the Jisc CASRAI-UK pilot on organisational identifiers

<https://jisccasraipilot.jiscinvolve.org/wp/2015/03/06/organisational-identifiers-working-group-outputs-and-update/>

<sup>37</sup> 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., Dias-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 Quarterly* 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.



THE  
NEW YORK  
ACADEMY  
OF MEDICINE

HEALTHY CITIES.  
BETTER LIVES.

# URBAN HEALTH UNCOVERED: The Grey Literature Report

---

**SUBSCRIBE,  
SEARCH,  
DISCOVER AT:**

GreyLit.org or  
greyLitHelp@nyam.org

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

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

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



## The GreyLit Report: Understanding the Challenges of Finding Grey Literature

Danielle Aloia and Robin Naughton,  
The New York Academy of Medicine Library, USA

### Abstract

Searching for and finding grey literature can be difficult. Grey literature, by its nature, is not commercially published and as a result, it requires multiple search strategies to identify and curate quality literature on a subject. Our study into how researchers share grey literature (Aloia and Naughton, 2015) found that researchers speak with colleagues, subscribe to listservs/newsletters, and go to organization websites to find current grey literature. In order to better understand the needs of the health sciences research community, we interviewed GreyLit Report users about their challenges, tools and methods for finding grey literature. The Grey Literature Report (GreyLit Report), developed in 1999 by The New York Academy of Medicine, is a centralized location that makes it easier for health researchers to find grey literature in their field. Speaking directly to librarians and researchers about their needs helped us to better understand how the GreyLit Report website can be enhanced to respond to those needs. Over the course of a week, we conducted online interviews with national and international users of the GreyLit Report. Based on this study, the researchers learned how the GreyLit Report can be enhanced to better serve the grey literature community and add to the growing need for a centralized location to find grey literature. In addition, the paper provides a template for planning and reporting of grey literature searches based on extensive analysis of the research literature.

### Introduction

Commercial databases are abundant, structured and time-honored tools that are the go-to source for researchers and librarians looking for quality resources. This makes it easy for researchers to document and capture the process used to find relevant articles, potentially making the search process reproducible to others. Searching for peer-reviewed articles is standard for the systematic review search process and commercial databases are structured and indexed in a way that facilitates effective searching. But as more research is published through alternative channels, the variability in search strategy has grown. Oftentimes, these materials are published on organization websites or within repositories that are structured and indexed in multiple ways. Organization websites are extremely variable and cannot be counted on to find relevant results. On the other hand, repositories maintained by a university or specific academic discipline can be tailored to meet the needs of their respective users. For these reasons, there is no one place or systematic process to search for grey literature, and as such grey literature searching requires a different set of skills. This research study seeks to understand how researchers and librarians search for grey literature and what resources they use. To do this, the study included a detailed literature review analyzing health science literature that used grey literature for systematic review searching, and semi-structured interviews with researchers and librarians who search for grey literature. The study was guided by three research questions.

### Research Questions

1. What challenges do researchers and librarians face when trying to find grey literature on the health sciences?
2. How are researchers and librarians in the health sciences searching for grey literature?
3. To what extent does the Grey Literature Report help researchers find grey literature in their field?

### Literature Review

Much has been written about the use of grey literature in the research process and its importance in the systematic review but less attention has been given to guidelines concerning the search process and reporting for grey literature. For instance, the *Cochrane*

*Handbook for Systematic Review of Interventions* 2011, provides a small paragraph on grey literature and mentions three or four sources, but no search techniques. The Canadian Agency for Drugs and Technologies in Health (CADTH) publishes *Grey Matters* and *Methods and Guidelines* series to alleviate the myths behind searching for grey literature. Their publications provide excellent guidance on search techniques in health technology assessment that can be used as a model for other topical searches. In *Methods Guide for Medical Test Reviews*, the U.S Agency for Healthcare Research and Quality suggests “Combining highly sensitive searches utilizing textwords with hand searching and acquisition and review of cited references in relevant papers is currently the best way to identify all or most relevant studies for a systematic review.” Currently, there is no one set of search techniques available for conducting a grey literature search. Grey literature searching may be systematic but it is time consuming, hard to replicate or reproduce, and resources vary among disciplines (Adams, 2016). As a result, this literature review explored the tools researchers used to search for grey literature, how search techniques and results were reported, and what types of search strategies were executed.

### Methods

The literature review is based on research papers collected from a PubMed search. Articles and reports that focused on systematic review research were collected and analyzed. Criteria for inclusion included English language, review article/report, mention of search strategies, and published in the past 5 years. The PubMed search (“grey literature” OR “gray literature”) Limits: English, review, and within 5 years) yielded over 1500 results. These results were paired down to 400. Of these 103 were excluded for various reasons, as listed below:

- Fifty-seven focused on grey literature search strategies for systematic reviews,
- Twenty-one didn’t report how or where they searched, so we were unable to determine the search methodology,
- Twelve didn’t report how or where they searched,
- Twelve papers indicated they just searched the traditional databases PubMed, Web of Science to locate grey literature,
- Six we were unable to obtain the full text of the article,
- Five of the articles only reported on the selection bias of not including grey literature in the systematic review, and
- Three papers mentioned explicitly that they excluded including grey literature in their searches.

A total of 297 papers were analyzed for the types of sources reported for searching grey literature. The 57 articles specifically discussing search strategies related to grey literature searching were analyzed for consistency of search methodology and to develop a model to conducting a systematic search.

### Reported Tools for Grey Literature Searching

Systematic review methodologies require the use of grey literature to overcome bias and to be sure that all relevant literature is captured. In the 297 papers analyzed, the most common search tools reported for searching grey literature were websites, handsearch, and Google. Clinical trial databases and registries were the most highly cited (155 mentions) but were not included in this analysis because it is a specific type of grey literature and has its own resources. In Figure 1, the number of mentions means that most authors reported using more than one tool in their search strategy.



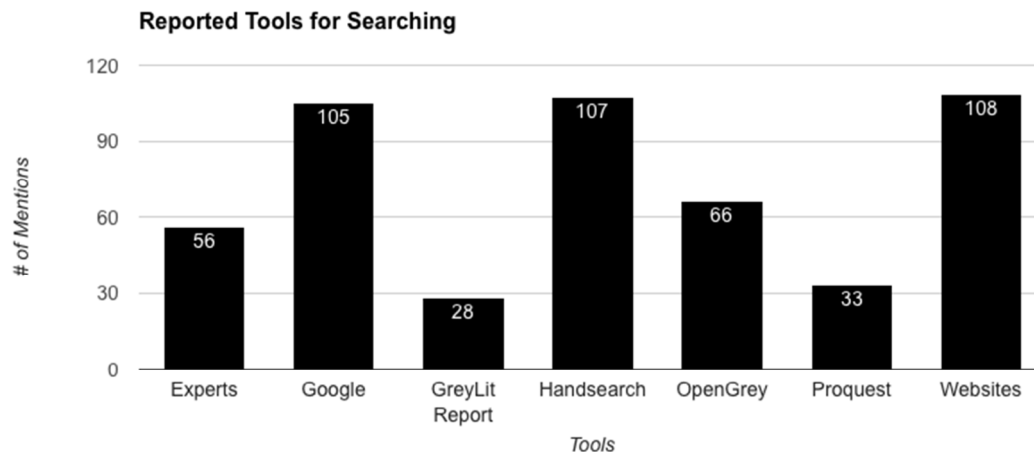


Figure 1: Reported Tools for Searching

OpenGrey, ProQuest, and the GreyLit Report are specialty databases and were cited by about 20% of the papers. Contacting experts to find relevant research was cited highly (56 mentions) as well and is a time honored method within the systematic review process. Handsearching is the “gold standard” for validating peer-review search results and making sure all relevant articles on the research topic are captured. This is almost impossible to do for grey literature searching therefore making it difficult to validate (Adams, 2016). Since Google and websites are the most cited sources for searching for grey literature it is essential to have some standard guidelines for searching. This will help make the search results and reporting transparent and maybe even reproducible.

### Search Strategies

Of the 400 reviews that were analyzed, 57 were specific on how to search for grey literature for systematic reviews. Each article recommended a different set of search strategies for a variety of platforms. A key insight was that a good grey literature search includes these resources: online databases, websites and search engines, repositories, online catalog, and asking experts (Mahood, 2014). This is not surprising because these sources have been highly cited in the literature. Grey literature searching can be time consuming and each grey literature source can take up to 1.5 hrs to search (Saleh, 2014). A good Google search strategy is to use advanced search techniques by searching the title field and checking results past the first 10 pages (Haddaway, 2015). Stansfield, et al. (2016) recommend putting effort into planning the search process early, especially, when it comes to locating resources to search.

### Search Reporting

Of the 400 papers collected 21 did not report any search strategies in their results. A few listed all the resources searched without keywords, some listed resources with keywords used, or included the results as an appendix or supplement. Briscoe (2015), *Web Searching for Systematic Reviews*, analyzed search results of systematic review papers and found that, the majority of papers only reported one search detail and few papers provided links to resources. Briscoe (2015) provides recommendations on reporting results for websites and search engines. Godin, et al. (2015), *Applying Systematic Search Strategies to Grey Literature*, provides guidelines on how to perform and record a grey literature search. Godin et al. provide an amazing amount of detail about their search methods and results in a short and easy to read paper! They created a multi-search plan using four resources: grey literature databases, Google, websites, and experts. They dedicate a paragraph to each search strategy and include an outline of the results in a PRISMA diagram. The authors also provide a timeline for the different methods of their search plan and attach additional files of their search results for each resource.

Below is a checklist of items to include in the search process. For best results in recording and tracking your searches apply these elements to your search strategy:

- Provide name of organization/search engine
- URL
- Dates searched

- Search terms or keywords used
- Some analysis of the results
- How and why results were chosen with links to selected sources
- Time spent

### In-Person Interviews

Interviews were conducted to understand the needs of GreyLit Report users in relation to how they search and their experiences using the GreyLit Report website.

### Methods

Using qualitative research methods with a focus on the semi-structured interview, participants were recruited and interviewed about their experience with the GreyLit Report.

### Recruitment

Participants for the research study were recruited from the GreyLit Report subscribers. There were more than 2000 subscribers to the GreyLit Report who were contacted through email regarding the study. An original email and follow-up emails were sent to subscribers describing the project. Participants were also recruited from an earlier study where they were asked to provide contact information if they would like to be interviewed regarding the GreyLit Report.

### Participants

There were a total of 12 interviews with 14 participants. Two interviews were group interviews because participants invited colleagues to join the scheduled interview who were then included in the discussion. Eleven participants were females and three participants were male. The participants represented a global reach from the United States of America, Australia, Canada, Netherlands, and the United Kingdom. There were librarians, information professionals and researchers. Table 1 shows the breakdown of participants by location and job title.

Table 1: Participants based on location and job title

Location	Job Title	# of Participants
Australia	Liaison Librarian	1
	Librarian	2
	Teaching and Learning Librarian	1
Canada	Information Specialist	1
	Manager, Information Services	1
Netherlands	Director	1
United Kingdom	Information Officer	1
United States of America	Head of Information Services	1
	Graduate Student	1
	Graduate Student/professor	1
	User Services Librarian	1
United States	Informationist	1
	Research Librarian	1
	Senior Research Advisor (Australia)	1
<b>Total</b>		<b>14</b>

Data was collected using semi-structured interview protocol. The interview guide was divided into three parts. The first part of the interview asked participants about their overall experience with grey literature and how they search for grey literature. The second part of the interview focused on the GreyLit Report and asked participants about their experience with the GreyLit website. The third and final part of the interview asked participants demographic information regarding job titles, gender and age range.

## Data Analysis

## Results

Participants were asked how they searched for grey literature and what tools/databases they used for search. Each participant provided multiple responses. Results varied in terms of search strategies and goals, but the majority of participants used Google (10) and organization and government websites (11) to search for grey literature, and OpenGrey (6).



Figure 2: Word Cloud of Participants Search

When asked about specific databases, participants mentioned over 69 databases used, including commercial databases such as PubMed, EBSCO, Web of Science, Proquest Scopus, country-specific databases such as Australian Policy Online and DANS Archive, topic-specific such as Clinical Practice Guidelines, and many more databases particular to the participants' area.

Participants were asked what topical searches would be of interest to them and their domain. They reported over 75 concrete responses in the areas of health information technology, public health, clinical medicine, legal, and physical exercise.

**Other Resources Indexed**

Participants were asked what other types of grey literature they would like to see indexed in the GreyLit Report. Eight participants mentioned Conference Proceedings, three mentioned Datasets, and three mentioned Webinars.

**Experience with the GreyLit Report**

Participants were asked about their experience with the GreyLit Report. In half of the interviews (6), participants used the GreyLit Report as a current awareness tool and would recommend it to others searching for grey literature. In some interviews (5), participants stated that they had a good experience with the GreyLit Report website. Based on the data, a majority of participants did not have as good an experience with the website and suggested that a few aspects of the site could be improved, including clarity on the Academy priority areas, an improved search, and broader scope or coverage.

**Discussion****What challenges do researchers and librarians face when trying to find grey literature on the health sciences?**

The major challenge faced when trying to find grey literature is the variety of sources available among and between disciplines. Respondents to the interviews indicated that sources they used depended upon the research question. Another challenge they faced was that each source has its own search criteria, making it hard to be consistent with search terminology.

**How are researchers and librarians in the health sciences searching for grey literature?**

Librarians and researchers are using a variety of ways to search for grey literature. The most common method, besides the traditional databases, are Google and organizations websites. This was found to be true both in the literature and from in-person interviews.

**To what extent does the Grey Literature Report help researchers find grey literature in their field?**

The Grey Literature Report was helpful to users in that it provided access to publisher names and alerts to new and current resources. Respondents reported difficulty with the search functionality and didn't feel the scope of content was broad enough.

**Conclusion**

Researchers and librarians use a variety of methods to find grey literature. This is supported by both the research literature and the interviews in this study. Google and organization websites were the most cited sources to find grey literature, followed by contacting experts and handsearching the peer-reviewed literature. The GreyLit Report was the fifth most cited in the literature and highly recommended by librarians as a good source for grey literature. The study showed that the major challenge of finding grey literature has to do with the nature of grey literature itself. The same methods used to search commercially published resources cannot be used to find grey literature resources. Grey literature requires a different search strategy and plan.

We recommend developing a search plan for finding grey literature, decide on the resources to use before starting your search, indicate search strategies for each resource using the guidelines above, and make note of results. Be prepared to spend time on each step of the process.

## Bibliography

Adams, Jean, Frances C. Hillier-Brown, Helen J. Moore, Amelia A. Lake, Vera Araujo-Soares, Martin White, and Carolyn Summerbell. "Searching and Synthesising 'Grey Literature' and 'Grey Information' in Public Health: Critical Reflections on Three Case Studies." *Systematic Reviews* 5, no. 1 (2016): 164. doi:10.1186/s13643-016-0337-y.

Aloia, Danielle, and Robin Naughton. "Share #GreyLit: Using Social Media to Communicate Grey Literature." *The Grey Journal (TGJ): An international Journal on Grey Literature* v. 12, no. 2 (2016): <http://hdl.handle.net/10068/1024651>.

Briscoe, Simon. "Web Searching for Systematic Reviews: A Case Study of Reporting Standards in the UK Health Technology Assessment Programme." *BMC Research Notes* 8 (2015): 153. doi:10.1186/s13104-015-1079-y.

CADTH Information Services. "Grey Matters: A Practical Tool for Searching Health-Related Grey Literature." Canadian Agency for Drugs and Technologies in Health, 2015. <https://www.cadth.ca/resources/finding-evidence/grey-matters>.

Godin, Katelyn, Jackie Stapleton, Sharon I. Kirkpatrick, Rhona M. Hanning, and Scott T. Leatherdale. "Applying Systematic Review Search Methods to the Grey Literature: A Case Study Examining Guidelines for School-Based Breakfast Programs in Canada." *Systematic Reviews* 4 (2015): 138. doi:10.1186/s13643-015-0125-0.

Haddaway, Neal Robert, Alexandra Mary Collins, Deborah Coughlin, and Stuart Kirk. "The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching." *PLoS One* 10, no. 9 (2015): e0138237. doi:10.1371/journal.pone.0138237.

Mahood, Quenby, Dwayne Van Eerd, and Emma Irvin. "Searching for Grey Literature for Systematic Reviews: Challenges and Benefits." *Research Synthesis Methods* 5, no. 3 (September 2014): 221–234. doi:10.1002/jrsm.1106.

Saleh, Ahlam A., Melissa A. Ratajeski, and Marnie Bertolet. "Grey Literature Searching for Health Sciences Systematic Reviews: A Prospective Study of Time Spent and Resources Utilized." *Evidence Based Library and Information Practice* 9, no. 3 (2014): 28–50.

Stansfield, Claire, Kelly Dickson, and Mukdarut Bangpan. "Exploring Issues in the Conduct of Website Searching and Other Online Sources for Systematic Reviews: How Can We Be Systematic?" *Systematic Reviews* 5 (November 2016). [http://download.springer.com/static/pdf/34/art%253A10.1186%252Fs13643-016-0371-9.pdf?originUrl=http%3A%2F%2Fsystematicreviewsjournal.biomedcentral.com%2Farticle%2F10.1186%2Fs13643-016-0371-9&token2=exp=1479485891 acl=%2Fstatic%2Fpdf%2F34%2Fart%25253A10.1186%25252Fs13643-016-0371-9.pdf\\* hmac=dc05b476c34a03e4497ac097310af0f9154dbc6efd5135366638c5790f59a6f9](http://download.springer.com/static/pdf/34/art%253A10.1186%252Fs13643-016-0371-9.pdf?originUrl=http%3A%2F%2Fsystematicreviewsjournal.biomedcentral.com%2Farticle%2F10.1186%2Fs13643-016-0371-9&token2=exp=1479485891 acl=%2Fstatic%2Fpdf%2F34%2Fart%25253A10.1186%25252Fs13643-016-0371-9.pdf* hmac=dc05b476c34a03e4497ac097310af0f9154dbc6efd5135366638c5790f59a6f9).

Wildemuth, Barbara M. *Applications of Social Research Methods to Questions in Information and Library Science*. (2009). Westport, CT: Libraries Unlimited.

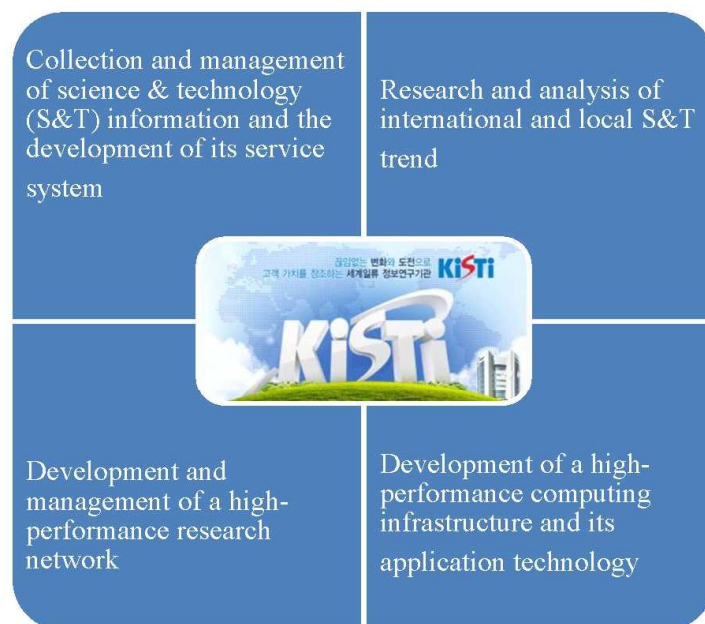
# Korea Institute of Science and Technology Information (KISTI)

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

## \* Vision

World-class information research institute creating values for customers

## \* Main functions



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

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

## \*Contact information

KISTI email address: [hcpark@kisti.re.kr](mailto:hcpark@kisti.re.kr)

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



## Debate about Scientific Popularization in Russian Public Sphere (Based on Grey Literature Material)

Yuliya B. Balashova

Saint Petersburg State University, Russia

### Abstract

The article is devoted to the problematic field associated with the popularization of science, in the reflection of the grey literature. In Russia, the public sphere is arranged in such a way that still many important issues do not discussed in the press, but on various discussions, which materials fixed in the grey literature.

**Keywords:** grey literature, popular science journalism, debate about scientific popularization, Russian public sphere.

### Introduction

At the present time, in Russian increased public interest to scientific problematic. Knowledge itself is a value in all times. Traditionally, science held a high position in Russian society, was vastly included into the public sphere. In Soviet times, the unique model of communication between science and society has been designed, and it didn't duplicate Western experience. Now in Russia a significant number of different activities, related to the science promotion, begin to be carried out. Different conferences, public discussions host in different regions of the country. Media begin to show more interest in all issues of science communication. However, in most cases, evidences of this discussion most accumulated not in the media, but in the grey literature. This is a very interesting phenomenon, peculiar exactly to the Russian public sphere. Traditionally, the most important questions of the public life were not fully reflected in the media. Historically, they informally discussed in friendly circles, clubs, recorded in protocols of the various societies, and in private correspondence (Aronson, M., and Reyser, S., 2000). These documents can be considered as prototype of the "grey literature". Even in the middle of the 19th century in Russia verbal social communication is still in many cases prevailed over writing (Pirozhkova, T. F., 1997). This trend continued in the conditions of Soviet life. In the late Soviet period, the most pressing social issues were discussed mostly in the locker rooms, at the kitchens in the apartments. But in Soviet times, science had a great reputation in the society and seriously supported by the state.

In the modern Russia, science became one of the most important national priorities again. The process of science popularization was broken in the post-Soviet time. As a result, in this area have been accumulated a lot of problems. One of the most significant problems lies in the fact that science in Russia is not sufficiently mediatized. Therefore, actual discussion about the ways of science development is mostly accumulated in grey literature.

### Organisation and methodology of research

The main method of research is the method of participant observation. The author engaged into the process of science popularization in the aspects of education, research, as well as in the public activities. I am myself the creator and the head of the new master's program "Popular science journalism", which started two years ago at Saint Petersburg State University, Russia. The author also is the organizer and participant in a number of scientific conferences, discussions, seminars, devoted to science communications, and science promoting. There materials are mainly reflected in the conference programs and post-releases. Accordingly, they are available to a relatively small circle of interested parties. This is the main reason why this discussion is centered around the same issues forming a germevetic circle.

The debates basically come down to two main issues that have a pronounced methodological nature.

The first of them is the following question: are journalists able to popularize science? A positive answer to this question has long been known. However, here is occurred

polarization of the points of view. Russian scientists believe that science journalists and writers are only make harm (Ivanov, I., 2007). Representatives of the communicative sphere are confident that Russian scientists themselves at the moment are not able to explain to the public the essence of their work (Vaganov, A. G., 2007).

Model of the knowledge in the West is based on the separation of science / art. In such logocentric country, like Russia, this dichotomy is not entirely justified. The very type of national consciousness tends to traditionalism and syncretism. In addition, Russia has not had the historical preconditions for the formation of narrow specialization. One of the most significant cultural reasons is not as consistent as compared with the Western Europe a Russian classic hierarchy development. The idea of the commonwealth of sciences is central to the whole tradition of the national popular science journalism. In the classic Russian "thick" magazines: "Contemporary" ("Sovremennik"), "Notes of the Fatherland" ("Otechestvennyie zapiski") departments of science and literature were mixed. The first issue of the Russian brand of popular science journalism was published in 1890. We are talking about the magazine "Science and life" ("Nauka i zhiz'n"), positioned itself as a "literary, artistic, social, and popular-science magazine". All subject areas were representative in terms of cognition; pre-revolutionary "Nauka i zhiz'n" was opened by mixed department "Science and art". In the popular science magazine science was presented as knowledge of the whole world. Approach to the understanding of scientific knowledge as universal knowledge was very typical to the classical Russian popular science magazine.

Another issue causing debate: is it necessary to popularize humanities? In modern Russian scientific environment has developed evaluative attitude towards humanitarian knowledge as opposed to natural science. In this sense, a logical continuation should be implicit question: "Is humanities research?". Russian grey literature paradoxically convinced that only natural and technical knowledge is a true science. So, here is a characteristic split in modern Russia between humanitarian and natural science. However, exactly humanitarian component has been put into the base of the Russian system of knowledge publicity. Historically has been accumulated considerable experience in the public implementation of humanities: from scientific and popular publications to educational films and lectures (Balashova, Yu. B., 2014).

The main body of the Russian grey literature related to the topic of pseudoscience and the fight against pseudoscience. So, in 1998 on the base of the Russian Academy of Sciences was created the special "Commission Against Pseudoscience and Falsification of Scientific Research". It has a coordination character, also publishes an annual bulletins "In Defense of Science", and conducts a variety of scientific and educational activities. But we have to note that passion for combating pseudoscience appears to be the legacy of the Soviet exposing companies. The latest trend of struggle against pseudoscience, in our view, does not introduce new meanings in a conversation about science popularization, as calls for a system of prohibitive measures, and based on the rhetoric of annihilation.

### Discussion

In those cases, when discussion about science popularization proceeds to the media, it does not become more meaningful because it not bases on the previous experience, reflected in the grey literature. As an example, we give a representative event, which was held on June 28, 2016 in the upper house of the Russian parliament – the Federation Council.

It had potential of a large conference and represented a constructive attempt to combine different participants of the objectively difficult process of scientific popularization. Among them were representatives of the university and academic community, science journalists. Within the framework of this debate it was able to overcome the estimated range of conflicting judgments accompanying discussion around the problem field: science – journalism.

During the speeches sounded the idea that society needs a popular science again. The authority of the Russian science remains the highest in the world, but it lacks the publicity. Therefore, the Russian scientific sphere in particular, needs a mediator between knowledge

and society. In the developed in science promotion countries, and especially the United States, the ability to external communications is competence of the modern scientist, and any academic institution is accompanied any serious research by information campaign.

Despite the constructive nature of the discussions, this event caused a negative response in the press. The author of the negative response in the leading popular science Russian newspaper "Troitsky variant" (under the name of Moscow suburb) was a student (or may be – graduate student) (Troitsky variant, 2016). In accordance with her position, the government is not able to organize this kind of events, as, indeed, to engage science popularization. This position, in principle, contrary to all history and practice of science popularization in Russia. In addition, this view completely ignores the experience of similar events, reflected in grey literature.

From my side, I acted in the same newspaper with an alternative evaluation (Troitsky variant, 2016). And this position received support in the other media (Ecology and life, 2016).

If the discussion on the issues about science popularization takes into account a variety of materials published in grey literature, it would be much more constructive.

### Conclusion

Materials of the different conferences and events, discussion in the blogosphere make up the range of sources about science popularization that require reflection. The scientific studies of the nature, content, specific aspects of the science popularization in modern Russia practically not involved.

Meanwhile, the controversy in the public space towards the popularization of science is centred on those issues that have long known. In the field of public actively discussed the issue could journalist popularize science or not. In Soviet times, the objective of which was to raise the level of the mass audience to scientific, science successfully popularized as the scientists themselves, as well as journalists. A similar situation typical for the Western press too.

Russian public sphere as a whole is arranged in such a way that in order to create a more holistic view of what is happening, it's necessary to appeal to grey literature.

### Acknowledgment

The authors gratefully appreciate the support of the Russian Humanitarian Scientific Fund.

### Funding Information

The article has been prepared with the financial support of the Russian Humanitarian Scientific Fund (RHSF). The project No. 16-03-50128.

### Ethics

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

### References

- Aronson, M., and Reyser, S. (2000), Literary circles and salons. St. Petersburg: Academic project, 400 p.
- Balashova, Yu. B. (2014), General scientific foundation of the new master's program "Popular scientific journalism". – In: Pseudoscience in the modern world: Mediasphere, higher education, school. St. Petersburg: Publisher VVM, pp. 21 – 25.
- Ecology and life, 2016. – <http://ecolife.ru/zhurnal/articles/43490/>
- Ivanov, I. (2007), Anatomy of updates, or how physics actually study the elementary particles. – [http://elementy.ru/nauchno-populyarnaya\\_biblioteka/430431/Anatomiya\\_odnoy\\_novosti\\_ili\\_Kak\\_na\\_samom\\_dele\\_fiziki\\_izuchayut\\_elementarnye\\_chastitsy](http://elementy.ru/nauchno-populyarnaya_biblioteka/430431/Anatomiya_odnoy_novosti_ili_Kak_na_samom_dele_fiziki_izuchayut_elementarnye_chastitsy).
- Pirozhkova, T. F. (1997), Slavophil journalism. Moscow: Moscow State University Press, 221 p.
- Troitsky variant, 2016. – <http://trv-science.ru/2016/07/26/nauchnaya-zhurnalistsika-na-ploshhadke-soveta-federacii/#comments>
- Vaganov, A. G. (2007), Popular science journalism and prestige of science in the public consciousness. – In: Russian Chemical Journal. 51, pp. 86 – 90..

## 'Grey crossroads' in cultural heritage preservation and resource management

Luisa De Biagi, CNR Central Library

Roberto Puccinelli, Telecommunications and Informative Systems Office, Italy

### Abstract

Among the assets that make up the cultural heritage of a country, a special place is assigned to the internal documents produced by organs and entities belonging to the Public Administration. In the public sector, for example, the minutes of meetings of the Boards of Directors are considered historical documents and as such are preserved in the for a long time. Actually from them it is possible to gain insight about the genesis of important decisions which affected the lives of many people. In some countries there is a legal obligation to deposit those documents in long term digital preservation systems, which adhere to ad hoc defined standards. In our opinion, many of those documents can be reckoned as grey literature assets and, beyond "plain and simple" preservation, some additional measures may be deployed in order to extract information and insights from them. In this paper we illustrate a process to collect those assets, cleanse and enrich their metadata and then store them in ad hoc defined data marts, upon which Business Intelligence tools can be used for data navigation and analysis. We finally show some examples of insights that may be acquired from such analysis.

### Introduction

*Grey literature is a field in library and information science that deals with the production, distribution and access to multiple document types produced on all levels of government, academics, business, and organization in electronic and print formats not controlled by commercial publishing, i.e. where publishing is not the primary activity of the producing body* (Greynet, 2011).

Grey literature is produced by entities whose main goal is not publishing. The grey material produced by public administrations and public/ industrial research laboratories, is significant both quantitatively and qualitatively, but has restricted dissemination for internal use. It should be noted that organizations, researchers and academics supplying scientific and technical information in grey literature, trust and support this type of documentation, not only because it contains more detailed outcomes and data, but also because it's produced up to 12 or even 18 months before the official papers are published.

Thus, the two way problem is: accessing grey literature for researchers and reaching identification and acquisition of grey literature for librarians and other information professionals. Moreover, the impact of Grey literature is largely dependent on research field/disciplines and its subjects categories, on methodologies approaches and on sources used. Nowadays newsletters, e-mails, blogs and other social networking sites are community based kinds of GL.

Practice Guidelines are highly important to biomedicine and nursing, working papers are used in Social Sciences (particularly Economics) and patents are important for the so called "hard sciences" (Physics, Geophysics, Chemistry, Biology, Maths) and for applied/technical sciences as Engineering.

Research data are also a kind of grey literature, even considering Social Sciences and Humanities: for example census, geospatial and economic data are used by local governments to formulate policies about preservation, valorization and risk evaluation.

Actually, other branches of cultural heritage are focused on technical aspects, such as archeology and archeometrics, architecture, diagnostics, preservation and restoration, also used for risk evaluation (e.g.: see recent Italian earthquakes and the current discussion about restoration and rebuilding policies and plans).

By consequence, given that some results remain hidden, not published anywhere but internal 'grey'sheets' or data-set archive, there might be more integration between the digital preservation and the various commercial tools and settings managing the long term accessibility of records in database systems useful to decision makers (data warehousing).

Another interesting source of grey literature is the administrative activity of organizations. Many documents are produced during the lives of organizations that in some way trace their history, evolution and end, sometimes providing detailed information about the motivations behind events and decisions. In many countries, there is currently a legal obligation regarding long-term digital preservation of this type of documents. This means not only that the assets will be preserved, but also that their long-term readability will be guaranteed. Moreover, this provides a chance for collecting and controlling useful metadata about the above mentioned assets.

One interesting example are the minutes of the Boards of Directors' meetings. They are considered historical documents and as such must be preserved in the long run. Actually, from them it is possible to gain insight about the genesis of important decisions, which affected the lives of many people.

We think that this long-term accumulation of documents and data lays the basis for the implementation of data warehouses, upon which many types of analysis could be performed. To achieve that goal, a well-designed process must be put in place in order to collect, cleanse and enrich high quality contents and metadata.

In the following we present a short description of models, workflows and architectures for digital preservation. We then describe Business Intelligence framework for metadata and content analysis. We finally illustrate some best practices and European project examples.

#### **Digital Preservation: models, workflows and architectures**

Digital Preservation can be considered a mature field in which a well identified set of standards and best practices has been developed over time and accepted by the reference community.

The **Open Archival Information System (OAIS) model** is currently considered the reference model for Digital Preservation Systems. It has been developed in the context of space agencies, which needed to preserve the huge mass of data coming from satellite observations. First introduced as recommendation in 2002, it has been promoted to ISO standard in 2003 (**ISO 14721**), subsequently updated in 2012 (**ISO 14721:2012**).

It does not describe a technical architecture nor it assumes a particular one. It just defines:

- the **functions** a DP system should implement,
- the **actors** that interact with the system (content producers, repository managers, content consumers),
- the supported **workflows** and
- the **digital objects** that are stored, managed and exchanged with the external world.

According to the OAIS model, a DP system should implement a set of functions that can be associated to the following functional entities:

**Administration:** allows the configuration of the system and the coordination of all the other entities;

**Ingest:** accepts contents and related metadata;

**Archival storage:** stores contents and related metadata and provides access to them;

**Data management:** collects and maintains the information required for the management and access of the contents;

**Preservation Planning:** monitors the contents and guaranties the access and readability of the contents for the designated community, also executing format migration in case of obsolescence;

**Access:** provides controlled access to contents through Authentication and Authorization mechanisms.

The elementary unit for a DP system is the **Information Object (IO)**, which comprises the **Data Object (DO)** - the actual digital content, represented as a bit-stream conforming to a particular format) and the **Representation Information (RI)** needed for the correct fruition and interpretation of the digital content. A DP system accepts, manages and generates **Information Packages (IP)**, which are composed of IOs. The OAIS model identifies the following three types of IP, each one containing one or more digital contents and the related meta-information:

- **Submission Information Package (SIP)**, used by the content providers to submit their objects to the DP system,
- **Archival Information Package (AIP)**, used by the system to archive contents (the contents provided in one SIP could be store in one or model AIPs),
- **Dissemination Information Packages (DIP)**, composed by the system in response to the access requests submitted by consumers.

Let's take the case of AIPs, which are the ones actually stored in a DP system. They are composed of **Content Information** (generally implemented as an IO containing the actual digital content and the related metadata) and the **Preservation Description Information (PDI)**, which is composed of one or more IOs that provide the following types of information:

**Reference Information**, which regards the identifiers assigned to the digital content;

**Context Information**, which documents the creation context of the digital content;

**Provenance Information**, which describe the origin and the history of the object;

**Fixity Information**, which is used to certify provenance and integrity of the object.

In a Digital Preservation system, knowledge can be extracted both from contents and metadata, the latter being the easiest to leverage, being by nature structured and (hopefully) controlled. The main standards used in DP systems for packages and metadata formats are:

- **UNI-SInCRO 11386: 2010**, for IP's structure (with particular reference to AIPs),
- **ISO 15836:2003** Information and documentation - The **Dublin Core** metadata element set, for metadata in general.

### Analysis workflows

In the present section we describe a viable workflow that allows the creation of grey literature Data Warehouse. The reference scenario features a set of repositories (belonging to one or more organizations) that represent the main source of information for the whole system. The workflow can be divided into two main streams: one for metadata the other for actual contents.

#### Metadata

Metadata can be extracted from the system using ETL procedures (Extract-Transform-Load) that fetch metadata from the repositories, perform cleansing, enrich them using of external data sources and reconcile them in a unified data base structure. Additional ETL procedures are then used to populate ad hoc data marts designed to address different types of inquiries and users. Data visualizations and analysis tools are then used to create queries and display results by means of tables and charts.

#### Textual contents

Textual contents are still a challenging asset to manage where it comes to knowledge extraction, due to their unstructured and often inconsistent nature. There are two interesting approaches to extract information from textual unstructured documents. Nevertheless, there is a wealth of text processing techniques and algorithms that can help getting insights from document bases. We identify two interesting approaches in this field: one aimed processing unstructured data to extract knowledge that can be expressed in terms of concepts, relations, topics, categories, etc., which are themselves represented by text; the second is focused on extracting textual and numerical information from



documents, that can be fed into a classic data warehouse and used to produce tables and charts (i.e. provide quantitative insight). The latter approach is closer to classic decision support systems, but the former can provide useful, non-trivial insight. To facilitate the following exposition, we shall call the first approach “semantic” and the second “quantitative”. These terms do not match with any standard or universally accepted taxonomy and will be used just for the sake of clarity.

Examples of the first type of analysis are:

- **Information extraction**

These techniques allow the extraction of structured information from unstructured text. A typical example is the identification of predefined relations in arbitrary text, such as marriages or company mergers. The text “The wealthy John Doe married the gorgeous Jane Doe in the beautiful countryside of Tuscany the 4<sup>th</sup> of July 2016” could be reduced to the following representation

Marriage (John Doe, Jane Doe, 4/7/2016)

if the relation of interest is of the type

Marriage (*husband, bride, date*)

- **Topic Tracking**

The goal of these algorithms is to determine whether a text could be of interest for a particular user, based on criteria such as the user’s profile, his previous readings and selected keywords.

- **Summarization**

A summarization algorithm reduces a lengthy text to few sentences that capture the essence of the whole document.

- **Categorization**

Categorization allows to assign documents to one or more categories belonging to a predefined set.

- **Clustering**

Clustering algorithms identify groups of similar documents (clusters) and assign them labels. In this case the labels are not predefined like the categories of categorization algorithms, but are generated on the fly based on the treated topics.

As regards the “quantitative” approach, an interesting example is described used in text tagging and annotation to analyse documents in order to identify and tag terms that correspond to domain-specific entities (for example, proper noun and numerical expressions) and then feed those terms into a classic star schema that can be queried to provide aggregated and detailed indexes.

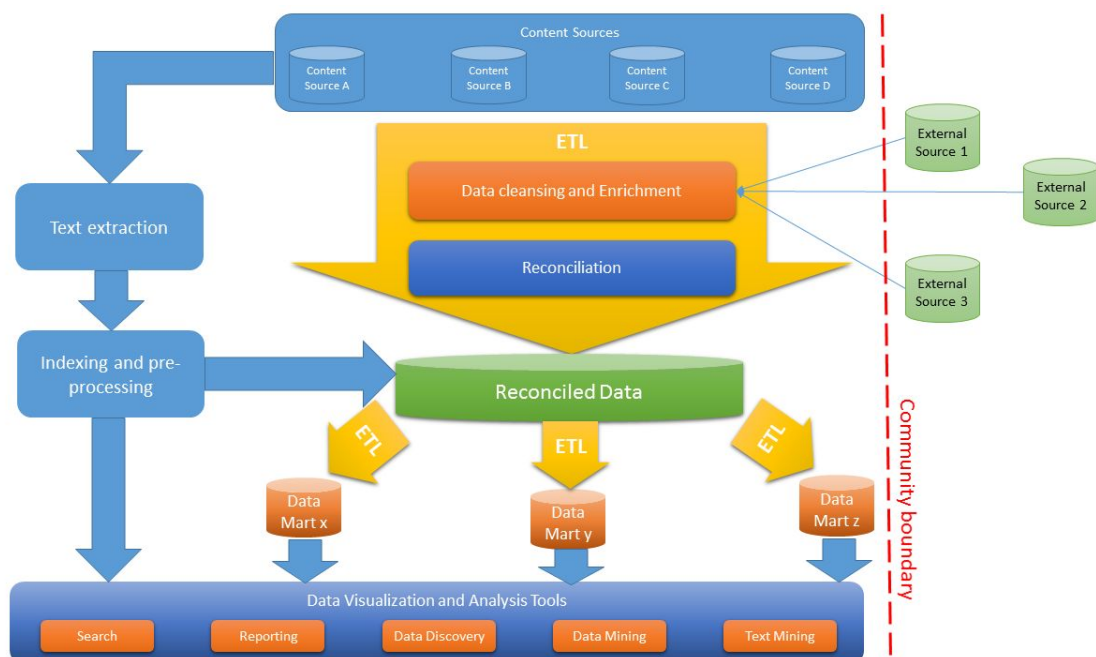


Fig. 1: data flow

In fig 1 we illustrate the data flow. At the top we have the content sources that in our case, as previously stated, are a set of repositories belonging to one or more organizations. On the left we have the flow of textual contents that are indexed and pre-processed in order to extract structured data that can be partly fed into the reconciled data base, which represents the starting point for quantitative analysis. At the centre of the figure we have the ETL procedures that cleanse, transform, integrate and enrich metadata that are then fed into the reconciled data base. Additional ETL procedures populate ad hoc data marts. The external sources on the right are obviously used to enrich metadata. Textual contents and metadata can be searched and analysed by means of the tools that are represented at the bottom of the figure.

### A Business Intelligence framework for metadata analysis

As regards source repositories, we assume that they can be harvested via OAI-PMH. **Fedora**, for example, is a well-established software for Digital Preservation repositories featuring an OAI-PMH provider. Metadata are exported in Dublin core format. Any other process for extracting metadata and contents is acceptable, although open standards are the most reliable choice in the long run.

In our technological framework, the software platform **Talend** is used to implement and execute ETL procedures that perform cleansing, enrichment and reconciliation.

The “Reconciled data” data base and the thematic data marts, are managed via PostgreSQL that, in our opinion, can be considered the most mature enterprise-level open source RDBMS.

As regards textual contents processing we think **ElasticStack** is a reliable integrated solution that features the following components:

- **Elasticsearch**, a distributed, JSON-based search and analytics engine,
- **Beats**, which acquires and sends data to Logstash and Elasticsearch,
- **Logstash**, which allows to compose the data collection pipeline,
- **Kibana**, that enables data presentation and provides a user interface for ElasticStack configuration.

The latter component can be seen as part of the “Data Visualization and Analysis Tools” box in Fig.1, which also includes Jasper (tables and charts) and Knime (Data and Text Mining).

Data discovery is an interesting new trend in data analysis and visualization, that allows data browsing and navigation through non-predefined paths. It leverages in-memory technologies and associative data bases for data storage, thus providing very low response times. To the best of our knowledge, there is currently no open source solution in this field. Commercial platforms are Tableau, Qlik, TIBCO Spotfire, Microsoft Power BI, MicroStrategy, SAP (Lumira), Platfora, Datameer, ClearStory Data, AnswerRocket, and Datawatch.

### Analysis types, possible insights

A key point for any system aiming at extracting value from metadata and digital contents is to ensure that metadata quality in terms of completeness and accuracy. This can be better achieved if the quality control is performed at the content source level and if all the useful information is gathered at ingestion time.

Let’s take the case of Board of Directors minutes. Much insight can be gained from text analysis and text mining, but the task becomes easier if the following information is kept as structured metadata:

- discussion points,
- decisions by type,
- yearly budgets approved,
- budget variation amounts.

In this case it is trivial to identify, for instance, a correlation between BoD decisions and organization’s performance or to evaluate the frequencies of the different discussion topics. Obviously, metadata value should be, as far as possible, numeric or selected from pre-defined lists.

### Best practices and European Project Examples

The launch of UNESCO programme ‘Information for All’<sup>1</sup> provides a platform for discussing actions on information policies and the safeguarding of recorded knowledge, in coordination with the ‘Memory of the World Programme’, which aims to ensure preservation and universal accessibility of the world’s documentary heritage.

To achieve those goals it’s necessary to leverage long-term preservation and big data technologies because they can provide tools to process high volumes of data coming from different sources and represented in different formats.

Other important technological assets for the access, study and protection of cultural heritage products are: new visualization techniques, implementation of 3D models of cultural heritage, search tools for digital libraries, new approaches to digital curation and preservation.

In Europe, beside large-scale aggregators of digital collections like Europeana<sup>2</sup>, outstanding projects about e-Infrastructures for digital cultural heritage are launched and managed by the paneuropean Network DARIAH-EU<sup>3</sup> - Digital Research Infrastructure for the Arts and Humanities - (18 countries, Italy included with CNR as National Coordinator) and CLARIN<sup>4</sup> - European research Infrastructure for Language Resource Technology.

With regard to DARIAH, which takes part in ERIC - European Research Infrastructure Consortium - and ESFRI - European Strategy Forum on Research Infrastructures - , remarkable benefits assured to digital cultural heritage communities and DARIAH-EU affiliated Projects are:

- visibility for National Research in Humanities and possibility of sharing data in a wider community; Technical environment and cooperation (e.g. virtual machines, long-term archiving, collaboration spaces etc.);
- expertise in data modeling and standards for metadata interoperability and virtual research;
- Sustainability: Research data, experiences, outcomes and publications, creating a European legal entity.

In this way DARIAH-EU Project applies for a Network with:

- close interface with Research community, their methods and questions;
- a forum to discuss current work and possible advancements, having a direct feedback from researchers expressing specific needs about digital data management and tools;
- more opportunities for financing of national and international projects.

Other interesting examples of European best practices in Cultural heritage data management and dissemination are DANS (Data Archiving and Networked Services)<sup>5</sup>, and ADS (Archeology Data Service)<sup>6</sup>, funded by the University of York (UK), which is member of Europeana Network and associated member of Digital Preservation Coalition<sup>7</sup>. In ADS the digital archiving of research data is entrusted to the service. ADS uses the Open Archival Information System (OAIS) reference model, which is integrated with various internal policies and procedures aimed to ensure that the data are correctly managed.

ADS main mission is “research, learning and teaching with freely available, high quality and dependable digital resources”. Always keeping a watchful eye on the target audience, ADS promotes updated good practices and technical advices in managing archeological data: a correct way also to enhance ICT research on preservation and exploitation of cultural heritage.

<sup>1</sup> <http://www.unesco.org/new/en/communication-and-information/intergovernmental-programmes/information-for-all-programme-ifap/>

<sup>2</sup> <http://www.europeana.eu>

<sup>3</sup> <http://www.dariah.eu/>

<sup>4</sup> <https://www.clarin.eu/>

<sup>5</sup> <https://easy.dans.knaw.nl/ui/home>

<sup>6</sup> <http://archaeologydataservice.ac.uk/about>

<sup>7</sup> <http://www.dpconline.org/>

Finally, a short note about Vocational Education and Training: it is desirable and suggested, both by academic-scientific community and by professionals, to integrate the “big data” topic in cultural heritage and social sciences curricula studiorum, especially for the sub-fields of information science, librarianship and archival sciences.

### Conclusions

In this paper we have highlighted how some types of Grey Literature assets could be leveraged to gain useful insights regarding the lives of organizations and countries. The legal obligation for long-term digital preservation currently enforced in many countries represent a great chance for building data warehouse infrastructures that collect resources and metadata from DP repositories, cleanse and enrich them and allow different types of analyses performed on ad hoc populated data marts.

### References

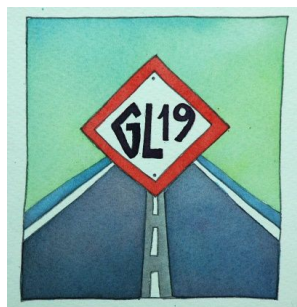
- [1] Consultative Committee for Space Data Systems Secretariat 2012, Reference model for an open archival information system (OAIS): Recommended practice (CCSDS 650.0-M-2: Magenta Book), CCSDS, Washington, DC.  
<http://public.ccsds.org/publications/archive/650x0m2.pdf> (last accessed 18/11/16)
- [2] Salza, S., Il modello OAIS, Work supported by the European Community under the Information Society Technologies (IST) program of the 7th FP for RTDproject APARSEN, ref. 269977
- [3] Lavoie, B., 2014 The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition), DPC Technology Watch Report
- [4] Fan, W., Wallace, L., Rich, S., and Zhang, Z., 2006 Tapping the power of text mining, proceeding of communications of the ACM, pp.78-82.
- [5] Parenteau, J., Sallam, R. L., Howson, C., 2015 The Rise of Data Discovery Has Set the Stage for a Major Strategic Shift in the BI and Analytics Platform Market, Gartner report ID: G00277789
- [6] Prasad, K. S. N., Ramakrishna, S., 2010 Text Analytics to Data Warehousing, International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010, 2201-2207
- [7] Manning, C. D., Raghavan, P., Schütze, H., 2009 An introduction to information retrieval, Cambridge University Press
- [8] Motta, G., Puccinelli, R., Reggiani, L., Saccone, M., 2015 Extracting value from grey literature: processes and technologies for aggregating and analysing the hidden “big data” treasure of organizations, Proceedings Grey Literature Conference 17
- [9] L.Börjesson, Grey literature – grey sources? Nuancing the view on professional documentation: The case of Swedish archaeology, Journal of Documentation, vol 71(2015), nr 6, pp.1158 – 1182
- [10] Guidelines on Cultural Heritage. Technical tools for for heritage conservation and management, September 2012, Conseil de L'Europe, JP - EU/CoE Support to the Promotion of Cultural Diversity (PCDK)
- [11] EU grey literature. Long-term preservation, access, and discovery, Luxembourg, Publications Office of the European Union, 2012 (Cedefop Working papers, nr15), ISBN 978-92-896-1132-9, ISSN 1831-2403
- [12] N. Fajrin Ariyani, U. Laili Yuhana, Generating cultural heritage metadata as linked open data in “2015 International Conference on Information Technology System and Innovation (ICITSI)”, Bandung – Bali, November 16 – 19, 2015, ISBN 978-1-4673-6664-9

## *List of Participating Organizations*

African Studies Centre Leiden, ASC	Netherlands
Alberta Health Services	Canada
American Geosciences Institute	United States
Austrian Academic Library Consortium	Austria
Biblioteca Centrale, G. Marconi; CNR	Italy
Center for Health Innovation; New York Academy of Medicine	United States
Center for Population Health Research; Lankenau Institute for Medical Research	United States
Centre National de Recherche Scientifique, CNRS	France
College of Physicians and Surgeons of Ontario, CPSO	Canada
Court of Appeals for the Second Circuit	United States
Data Archiving and Networked Services, DANS-KNAW	Netherlands
discoverygarden inc.	Canada
EBSCO	United States
East Carolina University; Health Sciences Library	United States
Federal Judicial Center, FJC	United States
Federal Library Information Network, FEDLINK	United States
Georgetown University	United States
GERiiCO laboratory	France
German National Library of Science and Technology, TIB	Germany
Grey Literature Network Service, GreyNet International	Netherlands
Indiana University, School of Informatics and Computing, IUPUI	United States
Institut de l'Information Scientifique et Technique, Inist-CNRS	France
Institute of Computational Linguistics, ILC-CNR	Italy
Institute of Information Science and Technologies, ISTI-CNR	Italy
Irvine Valley College	United States
Japan Atomic Energy Agency, JAEA	Japan
Korea Institute of Science & Technology Information, KISTI	Korea
Library of Congress, LoC	United States
McMaster University, Health Sciences Library	Canada
Metropolitan New York Library Council, METRO	United States
National Library of Technology, NTK	Czech Republic
Network and Information System Office, CNR	Italy
New York Academy of Medicine	United States
New York University; School of Medicine	United States
North Carolina State University Libraries, NCSU	United States
Nuclear Information Section; International Atomic Energy Agency, NIS-IAEA	Austria
Pratt Institute, School of Information	United States
PricewaterhouseCoopers, PwC	Netherlands
Rochester Institute of Technology, RIT	United States
Saint Petersburg State University	Russia
Slovak Centre of Scientific and Technical Information, CVTISR	Slovakia
Smithsonian Libraries	United States
TextRelease, Program and Conference Bureau	Netherlands
United Nations Dag Hammarskjöld Library	United States
University of Arizona, College of Medicine Phoenix	United States
University of Calgary	Canada
University of California, Irvine Libraries, UCI	United States
University of Florida Libraries	United States
University of Illinois at Urbana-Champaign	USA
University of Lille 3	France
University of Toronto, Institute of Health Policy, Management and Evaluation	Canada
University of Wisconsin, Milwaukee	United States
Waseda University	Japan
Wolters Kluwer, Health Learning, Research & Practice	United States
WorldWideScience Alliance	United States

# 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 below.

### 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



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

# TextRelease

GL19 Program and Conference Bureau

Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands  
[www.textrelease.com](http://www.textrelease.com) • [conference@textrelease.com](mailto:conference@textrelease.com)

Tel/Fax +31-20-331.2420



# 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



### Due Date and Format for Submission

1992 2017

Abstracts in MS Word must be emailed to [conference@textrelease.com](mailto:conference@textrelease.com) on or before **April 15<sup>th</sup> 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](http://www.textrelease.com) • [conference@textrelease.com](mailto:conference@textrelease.com)

Tel/Fax +31-20-331.2420

## Author Information

### Aloia, Danielle

151

Danielle Aloia is Special Projects Librarian at The New York Academy of Medicine. She received her MSLS from Catholic University of America, in Washington, DC, in 2005 while working on the AgeLine Database at AARP. She has over 20 years of experience in a variety of library settings, including academic, non-profit and museum. She has been involved with collecting, evaluating, and cataloging grey literature since 2006, first at AARP and then at the United States. Dept. of Transportation. For the past 4 years she has been managing the Grey Literature Report in Public Health, produced by NYAM. Email: [daloia@nyam.org](mailto:daloia@nyam.org)

### Balashova, Yuliya B.

159

Yuliya B. Balashova is a professor of journalism at Saint Petersburg State University, Russia. She is the author of scholarly works on the history of journalism and literature, pedagogy in journalism, science journalism. Taught journalism at Western European universities. As Fulbright scholar currently affiliated with Michigan State University, USA. Email: [ubalash@gmail.com](mailto:ubalash@gmail.com)

### Bartolini, Roberto

109, 117

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](mailto:roberto.bartolini@ilc.cnr.it)

### Biagioni, Stefania

97, 117

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

### Carlesi, Carlo

97

Carlo Carlesi, graduated in Computer Science, worked since 1970 at the IEI (now ISTI) of the Italian National Research Council in Pisa. His interest are focused in many software engineering field such as: Development of data base systems, Software quality and testing, Administration and UNIX system management, Digital library systems, Network security and management. In the year 2000-2010 he was head of the "Information Technology Security Service" of the Institute. He participated in many national and international projects, the last being: Italian Project for Research in the Antarctic - (South Pole project), the aim of the project was the creation of a Multidisciplinary Integrated Information System to manage and query the Set of Antarctic Data Bases; ERCIM Technical Reference Digital Library - (ETRD Project), The Digital Library service allows public access through Internet to the technical reports and other not published document produced by several organizations. It is currently a Research Associate of the Institute ISTI and he is involved in the following projects: PUMA - Publication Management. The Digital Library service allows public access (when permitted) through Internet to the published documents produced by CNR Organizations. GreyGuide - Guide to Good Practices and Resources in Grey Literature. An online forum and repository of good practice and resources in the field of grey literature. Email: [carlo.carlesi@isti.cnr.it](mailto:carlo.carlesi@isti.cnr.it)

### Ćirković, Snježana

81

Snjezana Cirkovic is Director of the Head Office of the Austrian Academic Library Consortium (Kooperation E-Medien Österreich). Her main responsibilities are organizing and coordinating the license agreements for electronic resources, centralized licensing for all consortia members, negotiating with suppliers, representing the consortium at international level. She holds a Master Degree in Library and Information Science from University of Belgrade, Faculty of Philology, with specialization in private libraries in the 18th century. She is currently writing PhD Dissertation with a specialization in European Cultural History during the 18th and 19th century (German-Serbian Cultural Relationships). From 2004 till 2015 she was working as a Senior Chief Librarian at the Institute for German Studies, Faculty of Philology, Belgrade University. She has developed and taught research courses of Information Literacy and Academic Writing for both M.A. and PhD students at the Faculty of Philology. Her main interests are electronic resources, open access, academic writing and scientific research. Email: [snjezana.cirkovic@gmail.com](mailto:snjezana.cirkovic@gmail.com)

### Davis-Castro, Carla

71

Carla Davis-Castro earned her Bachelor of Arts in Dramatic Art in 2008, Master of Public Administration and Master of Science in Library of Science in 2014 from the University of North Carolina at Chapel Hill. She joined the Library of Congress in 2015.

Email: [carlayasmin.daviscastro@gmail.com](mailto:carlayasmin.daviscastro@gmail.com)

**De Biagi, Luisa****162**

Luisa De Biagi got her Laurea Degree in Literature and Philosophy at 'La Sapienza' Univ. of Rome (Art history and Cultural heritage). With a Specialization in 'Archivist-Palaeographer' (Vatican School of Palaeography, Diplomatics and Archivistics at the Vatican Secret Archive) as well as a Specialization Degree in Archivistics, Palaeography and Diplomatics (Archivio di Stato, Rome) and a Degree from the Vatican School of Library Sciences. De Biagi further holds a Master in 'Business Publishing' (LUISS Management – Rome). She's been working for the SIGLE Network (System for Information on Grey Literature in Europe) since 2002. Since 2010 she's responsible for the Italian National Referring Centre of Grey Literature at CNR Central Library 'G. Marconi' as representative of the European Network and Openarchive OpenGrey. 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 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 (EAHL). 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](mailto:luisa.debiagi@cnr.it)

**Farace, Dominic****97**

Dominic Farace is Head of GreyNet International and Director of TextRelease, an independent information bureau specializing in grey literature and networked information. He holds degrees in sociology from Creighton University (BA) and the University of New Orleans (MA). His doctoral dissertation in social sciences is from the University of Utrecht, The Netherlands, where he has lived and worked since 1976. After six years heading the Department of Documentary Information at the Royal Netherlands Academy of Arts and Sciences (SWIDOC/KNAW), Farace founded GreyNet, Grey Literature Network Service in 1992. He has since been responsible for the International Conference Series on Grey Literature (1993-2013). In this capacity, he also serves as Program and Conference Director as well as managing editor of the Conference Proceedings. He is editor of The Grey Journal and provides workshops and training in the field of grey literature. Email: [info@greynet.org](mailto:info@greynet.org)

**Frantzen, Jerry****97**

Jerry Frantzen graduated in 1999 from the Amsterdam University of Applied Sciences/Hogeschool van Amsterdam (AUAS/HvA) in Library and Information Science. Frantzen is the technical editor of The Grey Journal (TGJ). And, since 1996, he is affiliated with GreyNet, Grey Literature Network Service, as a freelance technical consultant. Email: [info@greynet.org](mailto:info@greynet.org)

**Gelfand, Julia****32**

Julia Gelfand is the Applied Sciences, Engineering & Public Health Librarian at the University of California, Irvine Libraries where over the last 35 years she has performed many roles. She is active professionally and currently is a member of the Board of Directors of the Association of College & Research Libraries (ACRL), a division of the American Library Association, a member of the Science & Technology Section of the International Federation of Library Associations (IFLA) and is Secretary of Section T of the American Association for the Advancement of Science (AAAS). She writes and presents frequently on topics related to Scholarly Communication, Collection Management, Digital Scholarship, integration of multimedia in scientific literature, grey literature, social media, library as publisher. A graduate of Goucher College with graduate degrees from Case Western Reserve University, she is the recipient of many awards including the first GreyNet Award presented in 1999 and has been a Fulbright Fellow and a Thomas J. Watson Fellow. Email: [jgelfand@uci.edu](mailto:jgelfand@uci.edu)

**Giannini, Silvia****117**

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](mailto:silvia.giannini@isti.cnr.it)

**Goggi, Sara****109, 117**

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](mailto:sara.goggi@ilc.cnr.it)

**Inagaki, Satomi****24**

Satomi Inagaki is a Librarian at the Central Library of the Japan Atomic Energy Agency (JAEA). Email: [inagaki.satomi@jaea.go.jp](mailto:inagaki.satomi@jaea.go.jp)

**Lin, Anthony****32**

Anthony Lin is the Head of Instruction and Collections at the Irvine Valley College Library. He holds a MSI from the University of Michigan-Ann Arbor, a BA in Spanish from California State University San Marcos, and a BS in Finance from San Diego State University. His interests are emerging technologies, effective bibliographic instruction, and collections management. Email: [alin@ivc.edu](mailto:alin@ivc.edu)

**Lipinski, Tomas****11**

Professor Tomas A. Lipinski completed his Juris Doctor (J.D.) from Marquette University Law School, Milwaukee, Wisconsin, received the Master of Laws (LL.M.) from The John Marshall Law School, Chicago, IL, and the Ph.D. from the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Dr. Lipinski has worked in a variety of legal settings including the private, public and non-profit sectors. He is the author of numerous articles and book chapters; his monographs include, *The Library's Legal Answer Book* co-authored with Mary Minow (2003); *The Copyright Law In The Distance Education Classroom* (2005), *The Complete Copyright Liability Handbook For Librarians And Educators* (2006), and *The Librarian's Legal Companion For Licensing Information Resources And Services* (2012). Recent articles and chapters include, *Click Here to Cloud: End User Issues in Cloud Computing Terms of Service Agreements*, in *Challenges Of Information Management Beyond The Cloud: 4th International Symposium On Information Management In A Changing World*, Imcw 2013 (Revised Selected Papers.), with Kathrine Henderson, *Hate Speech: Legal and Philosophical Aspects*, in *The Handbook Of Intellectual Freedom Concepts* (2014), in 2013 with Andrea Copeland, *Look before you License: The Use of Public Sharing Websites in building Patron Initiated Public Library Repositories*, *Preservation, Digital Technology & Culture* and in 2012, *Law vs. Ethics, Conflict and Contrast in Laws Affecting the Role of Libraries, Schools and other Information Intermediaries*, *Journal Of Information Ethics*. He has been a visiting professor in summers at the University of Pretoria-School of Information Technology (Pretoria, South Africa) and at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Lipinski was the first named member of the Global Law Faculty, Faculty of Law, University of Leuven, Belgium, in 2006 where he continues to lecture annually at its Centers for Intellectual Property Rights and Interdisciplinary Center for Law and ICT. He is active in copyright education and policy-making, chairing the ACRL Copyright Discussion Group, a member of the ALA OITP Committee on Legislation Copyright Subcommittee, a member of the Copyright and Other Legal Matters Committee of IFLA and serves as an IFLA delegate to the World Intellectual Property Organization's Standing Committee on Copyright and Other Rights. In October of 2014 he returned to the University of Wisconsin to serve as Professor and Dean of its i-School, the School of Information Studies. Email: [tlipinsk@uwm.edu](mailto:tlipinsk@uwm.edu)

**Monachini, Monica****109**

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

**Naughton, Robin****151**

Robin Naughton, PhD is the Digital Systems Manager for the New York Academy of Medicine. Prior to joining the Academy, Dr. Naughton was a Digital Consultant with LearningExpress, an EBSCO company, and Oxford University Press, English Language Teaching. She received an Institute of Museum and Library Services (IMLS) fellowship for digital librarianship, and completed her PhD in Information Science from Drexel's College of Computing and Informatics in 2012. Dr. Naughton is a user-centered researcher focused on human-computer interaction (HCI) and mental models, specifically how to design and build interactive systems that are useful, easy to use and enjoyable.

Email: [rnaughton@nyam.org](mailto:rnaughton@nyam.org)

**Pardelli, Gabriella****109, 117**

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](mailto:gabriella.pardelli@ilc.cnr.it)

**Prost, Hélène****131**

Hélène Prost is information professional at the Institute of Scientific and Technical Information (CNRS) and associate member of the GERiCO 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](mailto:helene.prost@inist.fr)



**Puccinelli, Roberto****162**

Roberto Puccinelli is currently head of Section I at CNR's "Information System Office" and he's been working for CNR since 2001. He has previously worked in the private sector as system and network engineer. As adjunct professor, he has held courses for the First University of Rome "La Sapienza" ("Operating Systems II") and for the Third University of Rome ("Programming and Computing Laboratory"). He graduated in Electronic Engineering at the University of Rome "La Sapienza" and holds a master cum laude in Enterprise Engineering from the University of Rome "Tor Vergata". In the past he has worked in several research projects in the field of Grid technologies both at the national and international level (executive manager of Work Package 11 within the DataGrid project – V Framework Programme, et al.). He's currently involved in the design and development of CNR's information system. In particular, he coordinates projects for the development of application systems and is responsible for the design and implementation of CNR's data warehouse. He is also responsible for CNR's Local Registration Authority management. He's currently involved in projects regarding the design and development of research product open archives and persistent identifier registers/resolvers. He is author of several articles in the fields of Grid technologies, Autonomic Computing, Software Engineering, Open Archives and Persistent Identifiers. Email: [roberto.puccinelli@cnr.it](mailto:roberto.puccinelli@cnr.it)

**Rudasill, Lynne****114**

Lynne Rudasill holds the unique post of Global Studies Librarian and Associate Professor, University Library. Professor Rudasill's publications include edited books and book chapters as well as journal articles that focus on dissemination and use of non-governmental organization information and the resulting changes in scholarly communication, as well as user experience with library websites. She is a frequent participant in international conferences in library and information science. Lynne is a member of the Library Executive Committee, former Dean of the Military Education Council, Area Studies Division Chair and Chair of the Professional Committee of the IFLA Governing Board. She was recognized for her service to Association of College and Research Libraries with the Marta Lange/CQ Press Award.

Email: [rudasill@illinois.edu](mailto:rudasill@illinois.edu)

**Savić, Dobrica****49, 97**

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](mailto:d.savic@iaea.org)

**Schöpfel, Joachim****131**

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 GERiCO 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](mailto:joachim.schopfel@univ-lille3.fr)

**Smith, Plato L.****91**

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](mailto:plato.smith@ufl.edu)

**Stock, Christiane****97**

Christiane Stock is the Head of the Monographs and Grey Literature service at INIST, in charge of the repositories LARA (reports), mémSIC (master's theses in information sciences) and OpenGrey. Member of the Technical Committee for the SIGLE database from 1993 to 2005, she also set up the national agency for ISRN (International Standard Report Number). She is member of the AFNOR expert group who prepared the recommended metadata scheme for French electronic theses (TEF). Email: [christiane.stock@inist.fr](mailto:christiane.stock@inist.fr)


**Vaska, Marcus****59**

Marcus Vaska is a librarian with the Knowledge Resource Service (KRS), Alberta Health Services (AHS), responsible for providing research and information support to staff affiliated with an Alberta Cancer Care research facility. A firm believer in embedded librarianship, Marcus engages himself in numerous activities, including instruction and research consultation, with numerous research teams. An advocate of the Open Access Movement, Marcus' current interests focus on showcasing and creating greater awareness of the LibGuides devoted to grey literature in post-secondary institutions across Canada. Email: [mmvaska@ucalgary.ca](mailto:mmvaska@ucalgary.ca)

**Vaska, Rosvita****59**

Rosvita Vaska is a Subject Specialist (retired) with the University of Calgary, responsible for Germanic, Slavic, East Asian, and Arabic Languages and Literatures, as well as Linguistics and Holocaust Studies. Recipient of the 2009 Order of the University of Calgary, Rosvita has been heavily involved in curriculum development, instruction, and research within her subjects of responsibility throughout her career. A firm believer in the importance of grey literature and the Open Access Movement, Rosvita is presently investigating the use of LibGuides as a grey literature document type in instructional pursuits.

Email: [vaska@ucalgary.ca](mailto:vaska@ucalgary.ca)



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

Achieve  
your goals  
with us



## INFORMATION SUPPORT OF SLOVAK SCIENCE

### SCIENTIFIC LIBRARY AND INFORMATION SERVICES

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

### SUPPORT IN MANAGEMENT AND EVALUATION OF SCIENCE

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

### SUPPORT OF TECHNOLOGY TRANSFER

- Technology Transfer Centre at SCSTI
- PATLIB centre

### POPULARISATION OF SCIENCE AND TECHNOLOGY

- National Centre for Popularisation of Science and Technology in Society

### IMPLEMENTATION OF PROJECTS

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

[www.cvtisr.sk](http://www.cvtisr.sk)  
Lamačská cesta 8/A, Bratislava



## ***Index to Authors***

### ***A-B***

Aloia, Danielle	151
Balashova, Yuliya B.	159
Bartolini, Roberto	109, 117
Biagioni, Stefania	97, 117

### ***C-D***

Carlesi, Carlo	97
Ćirković, Snježana	81
Copeland, Andrea	11
Davis-Castro, Carla	71
De Biagi, Luisa	162

### ***E-F***

Ebisawa, Naomi	24
Farace, Dominic	97
Frantzen, Jerry	97

### ***G***

Gelfand, Julia	32
Giannini, Silvia	117
Goggi, Sara	109, 117
Gonda, Mayuki	24
Gruttemeier, Herbert	97

### ***H-I***

Hayakawa, Misa	24
Inagaki, Satomi	24
Itabashi, Keizo	24

### ***J-K-L***

Jones, Kyle	11
Lin, Anthony	32
Lipinski, Tomas	11

### ***M-N***

McIntyre, Lauren	91
Monachini, Monica	109
Naughton, Robin	151
Nozawa, Takashi	24

### ***P***

Pardelli, Gabriella	109, 117
Prost, Hélène	131
Puccinelli, Roberto	162

### ***R***

Rudasill, Lynne	114
Russo, Irene	109

### ***S***

Savić, Dobrica	49, 97
Schöpfel, Joachim	131
Smith, Plato L.	91
Stock, Christiane	97

### ***V***

Vaska, Marcus	59
Vaska, Rosvita	59

Forthcoming  
February 2017

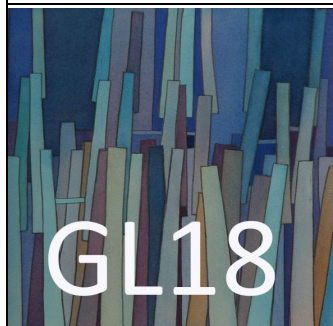
# 'Leveraging Diversity in Grey literature'

The New York Academy of Medicine - November 28-29, 2016

## Publication Order Form

### EIGHTEENTH INTERNATIONAL CONFERENCE ON GREY LITERATURE

Publication(s):	No. of Copies	x	Amount in Euros	Subtotal
GL18 CONFERENCE PROCEEDINGS - Printed Edition ISBN 978-90-77484-30-2 ISSN 1386-2316 <i>Postage and Handling excluded<sup>*)</sup></i>		x	109.00 = €	
GL18 CONFERENCE PROCEEDINGS - PDF Edition ISBN 978-90-77484-30-2 ISSN 1386-2316 <i>Forwarded via email</i>		x	109.00 = €	
GL18 Conference Proceedings - Online Edition ISBN 978-90-77484-30-2 ISSN 2211-7199 <i>Password Protected Access</i>		x	109.00 = €	



POSTAGE AND HANDLING PER PRINTED COPY <sup>\*)</sup>

Holland	<input type="text"/>	x	5.00	€
Other	<input type="text"/>	x	15.00	€

TOTAL EURO = €

Customer Name:	<input type="text"/>
Organisation:	<input type="text"/>
Postal Address:	<input type="text"/>
City/Code/Country:	<input type="text"/>
E-mail Address:	<input type="text"/>

☐ Direct transfer to TextRelease, Rabobank Amsterdam  
BIC: RABONL2U IBAN: NL70 RABO 0313 5853 42, with reference to "GL18 Publication Order"

☐ MasterCard/Eurocard ☐ Visa Card ☐ American Express

Card No.  Expiration Date:

Print the name that appears on the credit card, here

Signature:  CVC II code:  (Last 3 digits on signature side of card)

Place:  Date:

**NOTE:** CREDIT CARD TRANSACTIONS WILL BE AUTHORIZED VIA OGONE/INGENICO DESIGNATED PAYMENT SERVICES

**TextRelease**  
www.textrelease.com

GL18 Program and Conference Bureau  
Javastraat 194-HS, 1095 CP Amsterdam, Netherlands  
T/F +31-(0) 20-331.2420 Email: info@textrelease.com