

An International Journal on Grey Literature



Autumn 2013 – TGJ Volume 9, Number 3

‘IMPROVING GREY LITERATURE THROUGH INNOVATION’

GreyNet

The Grey Journal

An International Journal on Grey Literature

COLOPHON

Journal Editor:

Dr. Dominic J. Farace
Grey Literature Network Service
GreyNet, The Netherlands
journal@grey-net.org

Associate Editors:

Julia Gelfand
University of California, Irvine
UCI, United States

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

Dr. Joachim Schöpfel
Université Charles de Gaulle Lille 3
France

Gretta E. Siegel
Portland State University
PSU, United States

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

Technical Editor:

Jerry Frantzen, TextRelease

CIP

The Grey Journal (TGJ): An international journal on grey literature / D.J. Farace (Journal Editor); J. Frantzen (Technical Editor) ; GreyNet, Grey Literature Network Service. - Amsterdam: TextRelease, Volume 9, Number 3, Autumn 2013. - C.V.T.I.S.R., EBSCO, FEDLINK-Library of Congress, INIST-CNRS, ISTI-CNR, KISTI, NIS-IAEA, NTK, and NYAM are Corporate Authors and Associate Members of GreyNet International. This serial publication appears three times a year - in spring, summer, and autumn. Each issue is thematic and deals with one or more related topics in the field of grey literature. The Grey Journal appears both in print and electronic formats.
ISSN 1574-1796 (Print)
ISSN 1574-180X (E-Print/PDF)
ISSN 1574-9320 (CD-Rom)

Subscription Rates:

€240 institutional, including postage & handling

Contact Address:

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

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

About TGJ

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

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



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

About GreyNet

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

Full-Text License Agreement

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

© 2013 TextRelease

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

Contents

'IMPROVING GREY LITERATURE THROUGH INNOVATION'

Grey literature in the digital culture and practices of the global scholar: the case of molecular biology	122
<i>Chérifa Boukacem-Zeghmouri (France)</i>	
Working for an open e-publishing service to improve grey literature editorial quality	129
<i>Rosa Di Cesare, Marianna Nobile, and Silvia Giannini (Italy)</i>	
Collection of Conference Proceedings and Improving Access to the Full Text of Proceedings	139
<i>Misa Hayakawa, Shun Nagaya, Mayuki Gonda, Takeyasu Fukazawa, M. Yonezawa, and K. Itabashi (Japan)</i>	
Innovation, language, and grey literature	145
<i>Claudia Marzi (Italy)</i>	
Grey in the Innovation Process	152
<i>Keith G. Jeffery (United Kingdom) and Anne Asserson (Norway)</i>	
Data Analytics: The next big thing in information	157
<i>June Crowe and J.R. Candlish (United States)</i>	

Colophon.....	118
Editor's Note.....	121
On the News Front	
Launch GreyGuide Repository	
'A Guide to Proposed and Published Practices in Grey Literature'.....	160
GreyForum Workshop 2.1	
Grey Literature and Information Rights: On Law and Practice.....	161
GL15 Conference Outline and Registration	
'The Grey Audit, a Field Assessment in Grey Literature'.....	162
Advertisements	
INIS, The International Nuclear Information System.....	120
IAEA, Department of Nuclear Energy	138
Author and Title Index 2013.....	164
About the Authors.....	166
Notes for Contributors.....	167

INIS

www.iaea.org/inis

The International Nuclear Information System

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



IAEA

International Atomic Energy Agency

Editor's Note

In December 2004, during GreyNet's First Annual Award Dinner at the New York Academy of Medicine, plans were announced for the launch of The Grey Journal. The following spring, the first issue of the new journal appeared in print and electronic formats entitled "Publish Grey or Perish". This current issue marks the journal's twenty-seventh thematic edition – completing its ninth volume. The range of themes and topics in these volumes cross disciplines, sectors, as well as new frontiers in grey literature. The Grey Journal has truly become that which it set out to be a decade ago, *"the flagship journal for the international grey literature community"*. The journal articles that appear published in this series are due to the work of hundreds of dedicated authors and researchers, journal editors as well as corporate authors whose support sustains its publication. Together these stakeholders are the guarantee of quality and diversity in this field of library and information science. In 2008, The Grey Journal was awarded the coveted Victorinne van Schaick Prize. Over the years, all fifteen GreyNet Award Recipients have published in this series, and the journal continues to benefit through its longstanding license agreement with EBSCO, agreements with abstracting and indexing services, as well as with leading subscription agents. The current issue entitled "Improving Grey Literature through Innovation" focuses on positive change in the field impacted by digital culture, e-publishing, open access, data analytics, and more. On the eve of its 10th Anniversary, this flagship journal takes stock in its past and sets course for yet uncharted waters.

Dominic Farace
journal@grey-net.org

TGJ Volume 1, 2005 1. Publish Grey or Perish 2. Repositories - Home2Grey 3. Grey Areas in Education	TGJ Volume 2, 2006 1. Grey Matters for OAI 2. Collections on a Grey Scale 3. Using Grey to Sustain Innovation	TGJ Volume 3, 2007 1. Grey Standards in Transition and Use 2. Academic and Scholarly Grey 3. Mapping Grey Resources
TGJ Volume 4, 2008 1. Praxis and Theory in Grey Literature 2. Access to Grey in a Web Environment 3. Making Grey more Visible	TGJ Volume 5, 2009 1. Paperless Initiatives for Grey Literature 2. Archaeology and Grey Literature 3. Trusted Grey Sources and Resources	TGJ Volume 6, 2010 1. Government Alliance to Grey Literature 2. Shared Strategies for Grey Literature 3. Research on Grey Literature in Europe
TGJ Volume 7, 2011 1. Transparency in Grey Literature 2. System Approaches to Grey Literature 3. Research and Education in Grey Literature	TGJ Volume 8, 2012 1. Social Networking and Grey Literature 2. Data Frontiers in Grey Literature 3. Managing Change in Grey Literature	TGJ Volume 9, 2013 1. Adapting New Technologies for Grey Literature 2. Tracking Grey Literature Across Disciplines 3. Improving Grey Literature through Innovation

Grey literature in the digital culture and practices of the global scholar: the case of molecular biology

Chérifa Boukacem-Zeghmouri (France)

Abstract

Molecular biology is an innovative field, often studied for its cutting-edge practices. However, very few studies have focused on grey literature. The part played by grey literature in the researcher's process of producing new knowledge remains largely unknown.

By choosing the CRCL, an internationally recognized French laboratory, as the ground for our study, we were able to investigate precisely this question by observing the informational behavior of 13 researchers in their daily work. Our study combines three methodological techniques (semi-structured interviews, ethnographic observation and characterization of the scientific material produced). Our approach consisted in identifying the reality of grey literature in the researchers' informational habits. The results illustrate how a shift has occurred within grey literature, converging towards scientific data, which is becoming the field's new high added value « grey » deposits. Grey literature is more than ever taking a central role in molecular biology in the digital era.

1. Background

On the 2nd of March, 2013, the newspaper “Le Monde” published an article entitled “To whom belongs knowledge?”¹. In the article, a concise diagram retraced the three paths that currently exist to publish research results. These three paths are represented by their colours: white for commercial publications, green for deposits in open archives and gold for free access journals.

The article focused on the end of the publication process, by showing the three main economic models linked to the spread of scientific information. If we were to enhance the diagram by depicting the beginning of the process, as in the development of a publication, we would be tempted to add a new colour: grey.

Grey is not a business model. It represents the documentary material produced through research activity, but which doesn't enter a given commercial circuit: “*Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers i.e., where publishing is not the primary activity of the producing body*”².

However, when surveying the existing literature, it turns out that digital technology hasn't further highlighted the importance of grey in the researcher's informational habits and in the development of scientific knowledge. More research has been centred on the white track. The economic stakes can serve to explain such an interest, but methodological aspects also factor in this phenomenon. In the last few years and for most usage studies, the quantitative approach has mainly been used to analyze the usage statistics of electronic journal platforms. This method has the advantage of giving precise usage behaviour of publishers platforms, but it doesn't take into consideration the sources that researchers access elsewhere. Very few studies have chosen to focus on grey literature³.

When we studied the use of “White” for scientific publications in the field of STM - Sciences, Technology and Medicine⁴ using qualitative methodology, we noticed how important grey literature sources could potentially be for researchers. The semi-structured interviews we conducted at the time allowed us to broaden our observations. Researchers commented on a range of information sources, one of them being grey literature, which they preferred to consult on the Web. Since the interviews were held at the researchers' work places, we were able to notice on site the – sometimes contrasting – manifestations of grey literature, materialized through PhD theses and conference proceedings on office shelves.

A study published by the Research Information Network⁵ revealed that researchers felt dissatisfied about the access to electronic resources and thus made efforts to search for grey literature (PhD theses, reports, conference proceedings). This confirms that grey literature is still relevant for researchers and for the development of their knowledge, despite having increasingly wide access to large publications.

2. Aim of the study

To give more evidence to this point, we chose to examine how much space – or how little – grey literature occupied in a specific field within the wider field of STM, where access to scientific

information is mainly digital. Using Molecular Biology as a filed case study is justified by the fact that researchers in this field tend to fit with the notion of “Global Scholar”. The expression designates researchers who engage with communication and information technologies⁶. The term, coined by Paul Ginsparg⁷ at a conference, articulates the new representation of the values of academia in the digital world: collaboration, sharing, interdisciplinarity, networking... Since then, various people, in many different contexts, have adopted the term; its conceptual scope helps to envision the future of scientific activity in a digital world⁸.

Researchers in molecular biology are at the crossroads of pluri-disciplinary research (chemistry, physics and biology); they use research data, information related to genomics, publications and many other types of knowledge. Therefore, given our research stance, molecular biology is a very relevant field to study to identify the part played by grey literature in the development of new knowledge produced by researchers. The question is: what is the place of grey literature in the researchers’ informational habits and behaviour? Does it play a role in the research life process? These are questions we plan on answering precisely through this study.

Our field study helps us identify the part played by grey literature in the informational habits of researchers in molecular biology. It also allows us to determine which types of grey literature researchers are producing and spreading. Thus, we are ultimately able to determine how grey literature contributes specifically to the research life process.

3. Previous studies

The state of the art for our research question hinges on the information behaviours of research communities, scientific publishing and grey literature. Examining these themes through the lens of molecular biology leads us to investigate topics related to grey literature within a field known for its networked science, multi-scale modelling and scientific data integration⁹.

One of the most extensive studies on practices in biology, conducted by the RIN and the British Library¹⁰, shows that biologists use the Web and search engines intensively to access information. The use of a limited amount of key words is common and concurrent with the “googlisation” phenomenon observed in other fields¹¹. Access to information centres around articles published in peer-reviewed scientific journals, preferably high-impact journals, and research data¹². Biologists are faced with an information overload for articles as much as for data. They use ontologies to favour a semantic search for information that fosters both the production of hypothesis and biological discovery¹³. All these aspects raise the questions of skills levels and training¹⁴.

Collaborative research at national and international levels is essential for projects to advance in biology¹⁵. Researchers must therefore constantly update their computational skills in order to both adapt to the plurality of methodologies used in related fields and to learn using the tools needed for experiments and modelling¹⁶. Biologists share a certain amount of practices with other communities. However, their profile is specific and innovative because they use distributed IT infrastructure and big instruments; they also share research data and favour interdisciplinarity¹⁷.

Research data has been increasingly important to the work of researchers in molecular biology, making it a large scale discipline. Data is so important that it has lodged itself in the transversal structure of the research life process^{18 19}. Molecular biology has joined the Big Science movement, the latter revealing a paradigm shift in science itself²⁰.

The studies which we have briefly outlined don’t, however, place grey literature as the preferred object of study or further investigation. The results and conclusions from these studies allow us to, at best, deduce that grey literature is seldom present or used in molecular biology. But there is no evidence to suggest this. We will therefore attempt to clarify this topic in our study.

4. METHODS

4.1 Sample

Lyon hosts the Cancer Research Centre of Lyon - CRCL²¹, a research team with an international reputation, formed when several research teams from Lyon came together in 2010. The CRCL now hosts 350 French and international researchers. It is organized into 3 departments hosting 17 research teams. Because of the CRCL’s size, its international dimension, the recognized quality of its research and the plurality of themes studied, it became the “natural” ground for our study.

We were able to contact two departments within the centre: « Tumoral Escape » and « Information Flow in the cancer cell », where we met researchers. Thirteen researchers from both departments accepted to participate in our study and opened the door to their work spaces.

4.2 Data collection

Our methodology was organized in three parts. The first phase consisted in identifying the presence of grey literature in the CRCL's production. This wasn't a bibliometric study, but rather a way to characterize production in order to gain an initial understanding of the field. We therefore used SCOPUS²², a bibliometric database, to identify the publications produced by the CRCL. This allowed us to see which types of documents were produced and cited. The second phase was based on a qualitative approach with 13 researchers. We used the semi-structured interview technique, during which each researcher was led to talk about his/her informational needs, habits when searching for information and – possible – use of grey literature. We also used the focus group technique, which was a very important part of our method, since it allowed us to broach the issue of the corpus, allowing people with different approaches and discourses to confront views. Interviews and focus groups lasted between 1 hour and 1 hour 30 min. They were recorded for further study. Verbatim testimonies have been integrated to the study in quotation marks. The third part, led in conjunction with the interviews, consisted in the ethnographic observation of the physical places in which research happened. This part aimed to take into account and to question the organization of work spaces, their content and their attendance. Such observation took place during the individual interviews and the focus groups, which were systematically held on the CRCL premises.

5. RESULTS

5.1 Openness and Collaboration

Openness characterizes the organization of the CRCL teams' workspaces and facilities. Such openness brings the lab table and the computer closer; researchers can go from one to the other with no constraints. There is also physical proximity between researchers since glass partitions separate offices, making spaces visible. Within one place, many different types of activities occur fluidly, favouring exchanges and discussion.

Collaboration is, in fact, the norm within the team, within the lab and at a national level and, even more, at an international level. The importance of international collaboration was confirmed with our census of CRCL scientific production. Results identified 94 articles produced between the end of 2010 (when the CRCL was created) and 2012. These articles were mostly indexed under the fields of Biochemistry, Genetics and Molecular Biology, confirming the CRCL's pluridisciplinary approach within the larger field of biology. Certain articles have up to 100 authors coming from 12 different countries. The United States represents half of the authorship.

5.2 Digital Culture

When observing the workplace, there wasn't much space to store paper amidst the lab tables, computers and cabinets stocking equipment, and there was consequently little paper material to be found. Access to information takes place mainly online. Researchers read documents directly on their computer screens and depend on computers for many simulation and modelling processes. Out of the 94 articles affiliated with the CRCL, we counted 4149 references (almost exclusively articles), all digital. One single article could contain up to 200 references.

The library, in its physical or digital manifestations, isn't part of the researcher's discourse, who considers the library to be the whole of the Web. Because research in molecular biology requires large scale data, researchers need computational skills to carry out studies that are collaborative, shared at a distance and supported by digital networks. The available open access resources reinforce this trend. In order to identify resources, researchers use search engines (Google and Google Scholar) just as much as the traditional databases (PubMed, PubMed Central). Researchers in molecular biology are part of a global digital culture through their navigational, networking and connectivity habits.

5.3 Articles, Everywhere !

Throughout our interviews, researchers stressed the importance of articles. Two corollary notions can explain this, when analysing the interviews. The first notion is that the article certifies research results: authors are recognized by their peers and readers rely on validated results to form new hypotheses. The second notion is linked to the fact that the journal is the exclusive tool used to assess research teams, individual careers and applications for promotions and recruitment. In the same way, the allocation of research subsidies depends on the publication of articles in high impact journals.

5.4 PhD Theses: a collection of articles

Articles have gained such importance that they tend to appear in what is usually defined as grey literature. Interviews revealed that PHD theses defended by doctoral students are in fact collections of articles that have either already been published, on the verge of being published or in preparation. Not only is this practice commonplace, it is even encouraged by thesis supervisors, based on a productivity mind-set. The final document is then a collection of articles, contextualized by an executive summary, transition sections and a conclusion. This practice is justified by the fact that reading times must be optimized to favour more research activity and experimentation. This is why, starting at the beginning of the thesis process, the workload is divided into tasks that are promoted through the publication of articles, which are then integrated to the thesis as a whole.

5.5 Conference proceedings: Articles already published

All the researchers we met enjoyed participating in conferences, finding the encounters and exchanges very beneficial. Conferences are convivial moments and an opportunity to identify colleagues for future collaborations. However, researchers unanimously acknowledged that these scientific events were not meant to spread up-to-date knowledge. The content presented in conferences is largely finalized results that are either in the process of being published or just recently published.

This can be explained by the strong competition between research teams. Although collaboration is key to research activity in biology, competition to publish is also a reality. On-going research remains confidential and conferences are not events showcasing work in progress. They are rather places where the community can present the process or the finalization of “re-heated” research. In the same way, conference proceedings are merely a collection of articles either already published or in the process of publication. Out of the 4149 references associated to the 94 CRCL articles, only 7 references were taken from grey literature (conference proceedings). In the context of the CRCL, grey literature has no impact on citations in molecular biology.

5.6 Data to tell stories

The expressions “tell a story” or “tell a nice story” came back over and over during interviews and focus groups, when researchers were describing conclusive articles. Scientific principles underpin the use of these expressions: research originality, unprecedented methodologies and new results all serve to tell the story. This is a major factor for communities who struggle with information overload. Generating an original and appealing story therefore increases the chances for researchers to see their manuscripts accepted in quality journals, to be read and cited.

The importance of data was revealed through discussions on this particular topic. Until then, data, as the primary outlet of research, was in the realm of the implicit, needing to be articulated at a later stage. Data is part of the development phase in the researcher’s innovation process and is part of his/her daily activity. It sparks a dialogue between what is known and what is observed. Data gives researchers the opportunity to “play it by ear” and to adopt a “short term perspective which drives the research process to action”. Some researchers mentioned the importance of “taking the time to discuss data”.

In order to create the structure of the “nice story”, data appears as soon as the problem is defined. Data is also used when researchers construct their hypotheses. These values serve as tests, to check the obtained results. Finally, data accompanies the published article, taking on various forms. The summary table below, created after analyzing the interviews, gives a concise view of the place occupied by data in the research life process. In order to compare data and articles next to each other, we opted for the table format instead of the circle put forward by Humphrey²³. We therefore notice that data and articles aren’t incompatible, since articles are situated in a “long term perspective with conceptual outcomes”. Therefore, journal articles and datasets are complementary in order to conduct a successful research process.

	Books	Patents	Articles	Theses	Conferences	Data
Reading			X			
Problems			X			X
Hypotheses			X			X
Test						X
Verification						X
Publication			X			X
Promotion			X			

Table 1: Literature in the CRCL research life process

5.7 The « Grey » of data sets

Searching for data is one of the researcher's main occupations, alongside needing to publish and searching for relevant articles amidst the information overload. Data is available online and scattered in two categories of sources: data repositories and isolated data sets.

Data centres and data repositories are key sources for researchers; they are familiar with these centres and consult them on a daily basis. The researchers we interviewed mentioned the following repositories the most: *GenBank – Protein Data Bank*, *DRYAD*, *NCBI*, *PANGEA*, *Gene Expression Omnibus*, *Human Splicing Finder*. These play the same role as core journals. They are taken into account by journal publishers in which researchers publish²⁴, and they host the back office of published and non-published research.

Isolated datasets are scattered among the personal websites of individual researchers who choose to deposit them, or among the websites of research laboratories, where one of the pages may value and make available produced datasets. There are also some databases that are created in the context of specific projects and released online for development and re-use (ex : *Breast Cancer Information Core Database*, *Splice Site Prediction by Neural Network*, *Max EntScan*). Isolated datasets can also appear at the end of journal articles²⁵ under the heading « supplementary data ». Datasets are not always indexed or deposited in data repositories, and are therefore accessible through search engines. This second category of data is crucial in the research process of discovery and innovation. Researchers find this data to be important but hard to discern “deposits”, eluding traditional publishing circuits and accessible through mainstream search engines. For researchers, the added value of these “deposits” is that it allows them to be one step ahead when they identify the data and “stumble upon an original idea which hasn't even occurred to the person producing the data!”

This description allows us to make parallels with the definition of grey literature²⁶. This parallel is reinforced by other elements gleaned during focus groups where researchers expressed their regret concerning the fact that important data pools weren't made available online and couldn't be accessed with search engines. Researchers were also preoccupied about the preservation and sustainability of datasets available on the Web. The issue surrounding the skills necessary to curate the necessary data also arose. The human and financial needs were seen as a significant obstacle, risking the disappearance of such “grey” data.

6. DISCUSSION

These results are confined to a limited panel of researchers in molecular biology and they must be taken into account within this limit. They are representative of the reality we observed within two departments of the CRCL and cannot be generalized to molecular biology researchers as a whole. However, this first study focusing on molecular biology grey literature has the merit of showing certain trends (digital culture, sharing,...), some of which tie in with major trends described in larger scale studies²⁷.

Compared with existing studies, our study's contribution resides in its focus on grey literature within globalized research practices. By taking the stance of being close to the researcher through his information behaviour, we were able to broadly situate the role played by traditional and new grey literature in a context limited to molecular biology.

The analysis demonstrated that a double shift occurred in regards to grey literature in molecular biology. First, “conventional” grey literature has integrated some white literature content – journal articles – in order to align with the practices of the community. Furthermore, research data, which was invisible pre-Internet, now exists digitally in data archives, data bases or engines, allowing it to be used, or even reused. We therefore notice that, although traditional grey literature is replaced by the article, data, which is still largely out of the commercial circuit, tends to be the new type of grey literature. The digital world is therefore not making grey literature disappear, but instead highlighting it and revealing it²⁸. In 2010, an announcement revealed that “the supplemental material associated with a typical Journal article appears to be growing exponentially and is rapidly approaching the size of an article”²⁹. Given these results, it would appear that grey literature's place and role in molecular biology would be on the side of scientific data.

The new emerging questions must take into account issues linked to intellectual property and to quality, all part of the definition of grey literature. These results need to be linked to the results of a 2011 study on data centres³⁰, which highlights the beginnings of the phenomenon. This study stresses the upcoming importance of breaking the isolation and marginalization of these grey literature deposits and of integrating them to the global information and documentary ecosystem (DOI, metadata, discoverability, etc.)³¹.

Another challenge lies in raising awareness among researchers about depositing datasets in repositories. Beyond the idea of sharing, there is the notion of valuing the unknown or unrecognized researcher's work. A rising number of journals are associating datasets to articles and ask researchers to deposit associated data in archives³². We are also witnessing the emergence of journals dedicated to the publication of datasets³³.

Jobs linked to the curation of data have a bright future³⁴. This also represents an opportunity for academic libraries that are searching for missions to serve the academic communities. Through grey literature, the 4th paradigm of scientific communication³⁵ is already fully taking place in molecular biology. The next step would be to integrate data in the academic world's evaluation process.

7. CONCLUSION

The present study, focusing on molecular biology and on the CRCL team, shows that grey literature still appears in the global scholar's research life process but has been taking on a different shape and demands special attention. Grey literature raises questions in terms of quality, Intellectual property, conservation and curation. The research and academic communities must handle these questions in order to uncover the real scientific value of this new kind of grey literature. It has the ability to fulfil the ultimate goal of scientific communication as described by Henri Oldenburg: « *push on to the farthest bounds of things, and strive not to formulate fine and probable opinions, but sure and certain ones, and to penetrate from Nature's antechamber to her inner closet* »³⁶.

In contrast to the theme of this study, it may be interesting in future work to observe and analyse the evolutions surrounding the articulation of research data, as a component of grey literature, to the other colours of scientific publication. Scientific publishing may find a new momentum with grey literature.

Acknowledgments to Sylvie Mazoyer and Didier Auboeuf, (Professors, members of the « Information Flow in the cancer cell » department) and Arnaud Vigneron (lecturer, member of the « Tumoral escape ») from the CRCL, who accepted to meet me and to host this study with their colleagues.

8. REFERENCES

All the sites have been visited in August 20th.

1. http://www.lemonde.fr/sciences/article/2013/02/28/a-qui-appartient-le-savoir_1840797_1650684.html
2. Schöpfel, J., 2011, Towards a Prague Definition of Grey Literature. – In: The grey journal, 7 (1), pp.5-18
3. Rin, 2011, Access to scholarly content: Gaps and barriers. Available in: <http://www.rin.ac.uk/node/1172>
4. Boukacem-Zeghmouri, C., 2012, Online article searching on publisher platforms by STM french scholars: Findings and analysis. – In : The Canadian Journal of Information and Library Science 36 (3/4), pp. 88–105.
5. Rin, 2011, Access to scholarly content: Gaps and barriers. Available in: <http://www.rin.ac.uk/node/1172>
6. Mattelart, A., Mattelart, M., 2004, Histoire des théories de la communication. - La Découverte (Repères), Paris
7. Ginsparg, P., 1996, Electronic publishing in science - winners and losers in the global research village. In: Unesco Conference, Paris, Available in: <http://people.ccmr.cornell.edu/~ginsparg/blurb/pg96unesco.html>
8. Brain circulation: the global movement of research talent, vol.2, n°1, 2012. - In: The Academic executive Brief.
9. Morris, R. W., Bean, C. A., Farber, G. K., Daniel, G., Eric, J., Al, E., 2005, Digital biology: an emerging and promising discipline. – In: Trends in Biotechnology 23 (3), pp. 113–117.
10. Rin, 2009, Patterns of information use and exchange: case studies of researchers in the life science. Available in: http://www.publishingresearch.net/documents/RINPatterns_information_use-REPORT_Nov2009.pdf
11. Boukacem-Zeghmouri, C., Schöpfel, J., 2013, Beyond the Google generation: Towards Community-Specific usage patterns of scientific information? In: Baker, D., Evans, W. (Eds.), Trends, Discovery and People in the Digital Age. Chandos Publishing Online, p. 200
12. Gallezot, G., 2002, Exploration informationnelle et construction des connaissances en génomique. – In: Les Cahiers du numérique, 121–136.
13. Jensen, L. J., Saric, J., Bork, P., 2006, Literature mining for the biologist: from information retrieval to biological discovery. – In: Nature Review Genetics 7 (2), 119–129.
14. Ouzounis, C. A., Apr. 2012. Rise and demise of bioinformatics? Promise and progress.-In: PLoS Comput Biol 8 (4),
15. Gallezot, G., 2002, Exploration informationnelle et construction des connaissances en génomique. – In: Les Cahiers du numérique, 121–136.
16. Morris, R. W., Bean, C. A., Farber, G. K., Daniel, G., Eric, J., Al, E., 2005, Digital biology: an emerging and promising discipline. – In: Trends in Biotechnology 23 (3), pp. 113–117.
17. Collins, F., Morgan, M., Patrinos, A., Apr. 2003, The human genome project: lessons from large scale biology. – In: Science 300, pp. 286–290.
18. Richter, B. G., Sexton, D. P., Jun. 2009. Managing and analyzing Next-Generation sequence data. PLoS Comput Biol 5 (6)
19. De Schutter, E., 2010, Data publishing and scientific journals: the future of the scientific paper in a world of shared data. – In: Neuroinform 8, 151–153.
20. Gray, J. The Fourth Paradigm, Data-Intensive Scientific Discovery, 2009
21. <http://www.crcl.fr>

22. <http://www.scopus.com>
23. Humphrey, C., 2006, E-science and the life cycle of research <http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc>
24. This is the case of Elsevier, in particular, who announced in the summer of 2010 the release of Pangea and the aim to enhance the articles on the ScienceDirect platform with links to datasets "This enrichment functionality offers a blueprint of how Elsevier would like to work with data set repositories all over the world":
25. <http://www.reedelsevier.com/mediacentre/pressreleases/2010/Pages/elsevier-and-pangaea-take-next-step.aspx>.
26. Nucleic Acids Research offers to include supplementary data at the end of each article: <http://nar.oxfordjournals.org> (Visited on August 19th, 2013). This phenomenon is gaining momentum and the number of available corpuses is considerably increasing.
27. Pejsova, P., Vaska, M., 2011, An Analysis of Current Grey Literature Document Typology. – In: The grey journal, 7 (2), pp. 72-80
28. Rin, 2009, Patterns of information use and exchange: case studies of researchers in the life science. Available in: http://www.publishingresearch.net/documents/RINPatterns_information_use-REPORT_Nov2009.pdf
29. Carroll, B. C., Crowe, J., Candlish, J.R., Scientific Data: Increasing Transparency and Reducing the Grey. – In: The grey journal, 7 (1), pp. 33-40
30. Maunsell, J., 2010. Announcement regarding supplemental material. The Journal of Neuroscience 30 (32), 10599–10600.
31. Rin, 2011, Data centres: their use, value and impact. Available in: <http://www.jisc.ac.uk/~media/A922660EE26646E2A08D07FD7738557A.ashx>
32. Richter, B. G., Sexton, D. P., Jun. 2009. Managing and analyzing Next-Generation sequence data. PLoS Comput Biol 5 (6)
33. De Schutter, E., 2010, Data publishing and scientific journals: the future of the scientific paper in a world of shared data. – In: Neuroinform 8, 151–153.
34. Pfeifferberger, H., Wegener, A., Carlson, D., 2011, Earth system science data (ESSD) - a peer reviewed journal for publication of data. D-Lib Magazine 17 (1/2).
35. Wren, J. D., Bateman, A., 2008, Databases, data tombs and dust in the wind. – In: Bioinformatics 24 (19), pp. 2127–2128
36. Gray, J. The Fourth Paradigm, Data-Intensive Scientific Discovery, 2009
37. Letter to Thomas cox, 1656, quoted by Iordan Avramov, An apprenticeship in scientific communication: The early correspondance of Henry Oldenburg (1656-1663).

Working for an open e-publishing service to improve grey literature editorial quality *

Rosa Di Cesare, Marianna Nobile, and Silvia Giannini (Italy)

Abstract

A survey of in-house publishing practices at CNR Institutes is described. Fourth categories are introduced to measure the level of innovation in the management of in-house publications in order to identify the business model used by each CNR Institutes to manage their editorial products, especially digital products. Data used for this descriptive and quality study were obtained from CNR Institute websites.

1. Introduction

The widespread adoption and diffusion of the Internet as well as the increasing application of digital publishing technologies have modified and streamlined functions, processes and products in the scholarly communication chain. The consequences of the application of ICT in context of scientific research and education has been analyzed in many studies. Starting from the first contributions (Roosendaal 1997, 2004; Gierveld 2002) that were focused on the transformation of the linear scientific information chain, through to more recent studies on the changes in scholarly communication focused on e-science collaboration and information and data sharing (Borgman 2007; Tenopir et al. 2012), many analyses are now focused on added—value services embedded in the digital publishing technologies. The latter particularly point out the key role played by academic and research libraries in challenging the traditional business publishing model on in favour of a more sustainable economic model to produce and diffuse scholarly research outputs. Many of these studies refer to e-publishing library services development which *is rapidly becoming a norm for research libraries, particularly journal publishing services* (Mullins et. al. 2011; Iglezakis et al. 2011).

Moreover, the widespread diffusion of electronic publishing technologies is increasing digitization initiatives addressed to traditional libraries' print collections, which involve also Grey literature collections. This means that digital publishing technologies are creating a "second life" for traditional and valuable grey documents and this innovative way of managing grey contents can improve editorial quality, as well as, their diffusion and discovery.

Several studies and reports have demonstrated that moving from print to a digital publishing model reduces production costs through rationalization and automation of editorial procedures (Houghton 2011; Willinsky 2011; Crow et al. 2009). Within this context studies on innovation emphasise the combination of electronic publishing and open access as represent the major drivers of editorial changes.

2. Research hypothesis and aims

Generally an editorial activity is a component of scholarly communication that can be managed as an autonomous activity and/or integrated with the management activities of the scientific production of an academic or research institution.

These activities can be carried out in-house and/or in collaboration with commercial publishers.

Our basic premise is that the editorial activity depends on the organizational-productive context while the number and type of editorial products depends on the disciplinary area. Moreover it could be carried out in an innovative way in terms of process and/or products, to make the editorial process more efficient and effective.

The National Research Council (CNR), is one of the biggest Italian multidisciplinary research institutions and comprises a network of 109 Institutes geographically distributed Institutes, which have scientific and organizational autonomy. One of the institutional missions of CNR is the diffusion of scientific information in Italy and its editorial products reflect this institutional mission. At local level too, CNR institutes have always created their editorial products strictly connected with their studies and research interests, tailored to different target of users (general public and/or their research community), for their production and diffusion they have used and still use conventional and non conventional channels. Of course this activity may vary from Institute to Institute.

The survey analyses in-house production available at CNR Institute web sites to identify: a) type of products and use of bibliographic elements; b) technology used to manage in-house products and c)

* First published in the GL14 Conference Proceedings, February 2013.

degree of innovation concerning the management of the contents; and finally d) access and discovery of products. In other words the analysis concerns the description of the business model used in the editorial production and the policy adopted for the diffusion of its contents.

As the framework of our analysis are new publication models strictly connected with open access movement, the general aim of the survey is to obtain information on current publishing practices at CNR institutes as a means to improve editorial quality of in-house scientific publications and increase visibility of CNR scientific products.

3. Materials and Methods

3.1. Survey design

The object of our analysis is CNR institutes' editorial products published in-house and/or in collaboration with commercial publishers.

The survey was divided into two phases.

In the first phase, we checked each CNR Institute website to gather preliminary information about in-house production, identifying current and ceased editorial products directly produced and managed by CNR Institutes, in a stable standardized way and with continuity. Then we classified Institutes according to a set of criteria that measures the level of innovation in the management of their editorial products (see box).

- I (Innovative)** Institutes that manage in-house-publications applying an editorial control that includes at least an identifiable standardized series title and numbers; and using electronic-publishing systems to manage editorial processes.
- A (Traditional)** Institutes that manage in-house-publications applying an editorial control that includes at least an identifiable standardized series title and numbers.
- Z (No editorial control)**. Institutes that produce GL without applying any editorial control or where we could not find any information on their production on their website
- X (No in-house publications)**. Institutes that do not produce GL at all and/or produce it sporadically - <5 per year-.

In the second phase, we have chosen, among in-house publications, the editorial products, with a minimal set of editorial and bibliographic elements (i.e.: series title and/or number) produced by the Institutes belonging to Letter I and A, and we have analysed their products in terms of:

- Type of products and their publication frequency
- Type of production/diffusion
- Technology used
- Access policy

Data was collected from CNR Institute Websites, between June and September 2012.

For those websites with scarce information about their in-house production, we also checked the CNR central archive that collects both conventional and non conventional literature, and also includes the editorial products produced and managed by CNR institutes.

In addition we decided to conduct informal phone interviews with the manager of the CNR Institute library, to obtain further and more detailed information, in particular on the procedure used to manage the editorial products, critical issues connected with the editorial activity, such as budget and human resources as well as future editorial plans. The phone interview further considered the characteristics of the publishing systems used, focusing in particular on the technical requirements for locally developed systems.

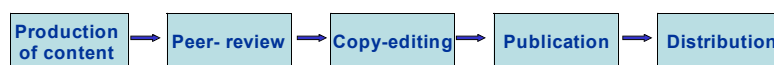
3.2. The business model

An important part of our analysis aimed to identify the business model used to manage in-house publications by CNR Institutes that fall in the category I, namely the Institutes which have introduced changes to the editorial process and/or created a new end product, in other words those Institutes that make the editorial process more efficient. As mentioned before, digital publishing technologies have in addition led to many changes in the core functions of the libraries and publishers, as well as of scholars, affecting the general flow of an editorial processes.

In our view the main steps that are foreseen in an editorial process encompassing different phases.

Fig. 1. Business process model

Analysis of the business process



NO Linear process:

Depending on the business model adopted
on the organisational framework
on the type of products
on the access policies
on the evaluation strategies
on the technology used
.....

It starts with the production of content, that is the acquisition of contents that are going to be published. These contents can be managed as an internal or external activity and can be automated or not. The contents may be subject to peer review or other reviewing systems or not. Then there is the phase of copy-editing where the manuscript is submitted to improve editorial quality and ensure the content's bibliographic and textual style, while the proof reading checking for typos and layout transforms the editorial product for publication and distribution

Most of these activities can be carried out in-house or outsourced to external services providers or also to commercial publishers.

Along with the aforementioned issues the business model also depends on the policies adopted for access to contents, ranging from subscription fees to full OA, but also on the products which can be peer reviewed or not and, of course, on the technology used. In addition the editorial process can be influenced by the organizational framework and human resources available, or by the type of products to be produced. (Journals, monographs or reports are very different in terms of cost. The cost is higher for a monograph compared to for example that of a reports.

To identify the business model we checked for each editorial products the type of business process carried out: a) how the Institute manages products in terms of production & diffusion and/or distribution and which phases were externally managed; b) if the Institute has introduced some changes to the editorial process management (we checked whether the contents of the editorial product were not just a version of the printed one) and, c) if they make use of an online content management and publishing platform or an electronic handling manuscript system (content management system).

4. Results

Table 1 shows the results of the first phase of the analysis that aims to classify the level of innovation of CNR Institutes in the management of in-house production.

Table 1. Number of CNR Institutes by Department according to grouping criterion (see box**)

DEPARTMENTS	Number of Institutes	I	A	Z	X
Earth & Environment	13	2	3	8	0
Agriculture & Food	10	1	2	7	0
Biomedical Sciences	17	0	0	7	10
Chemistry & Materials Techn. Sciences	14	0	0	7	7
Physics Sciences	14	0	2	5	7
Engineering & ICT	21	1	7	12	1
Social Sciences & Humanities	20	6	8	5	1
Total	109	10	22	51	26

The majority Institutes that have a variety of in-house production, managed in different ways (ranging from the most innovative to traditional) are concentrated in the Departments of Earth and Environment, Agriculture and Food, Engineering and Social Sciences and Humanities.

Conversely, institutes belonging to the Department of Biomedical Sciences and Chemistry don't either produce GL or carry out any in-house editorial activity to manage and diffuse their products. They obviously use traditional channels to diffuse their research results.

However, as we can see in the table, there are many Institutes that are classified in category X (26), these are Institutes which produce GL sporadically and in a very limited number ($> 5 \times \text{year}$).

The Institutes classified in category Z (51) produce a lot of GL documents, that are not organized in well-established, standardized series. This is also confirmed in a previous survey (Di Cesare, 2010). These are GL documents produced "ad hoc", such as project deliverables, conference proceedings and so on and this is the case especially for institutes belonging to the Engineering and Information Communication and Technologies Department (ICT).

They were excluded from the analysis because their products lacked the minimal set of editorial and bibliographic elements.

Further to the recent signing of the Berlin Declaration by the CNR, following the development of the CNR Institutional Repository (IR), a working group was established to elaborate specific guidelines for quality and metadata control of the CNR researchers' output, including current and back GL collections. So we hope in the near future to have more suitable procedure in the management of in-house publications, together with consistent editorial policies for all CNR Institutes.

4.1. CNR Institutes editorial products

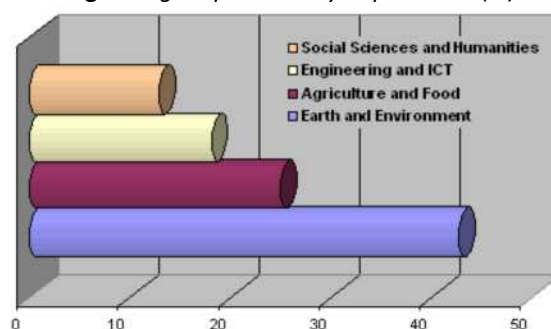
Table 2 shows the number of traditional and digital products broken by CNR Department. We found 106 editorial products with the minimal editorial set (i.e.: series title and/or number, corresponding to inclusion criteria) to be included in the analysis. 106 out of 19 – equal to 18% - are digital products.

Table 2. Number of editorial products by Department

Department	Number of editorial products	Traditional product	Digital product
Earth & Environment	14	8	6
Agriculture & Food	8	6	2
Biomedical Sciences	1	1	0
Chemistry & Materials Techn. Sciences	0	0	0
Physics Sciences	2	2	0
Engineering & ICT	11	9	2
Social Sciences & Humanities	70	61	9
Total	106	87	19

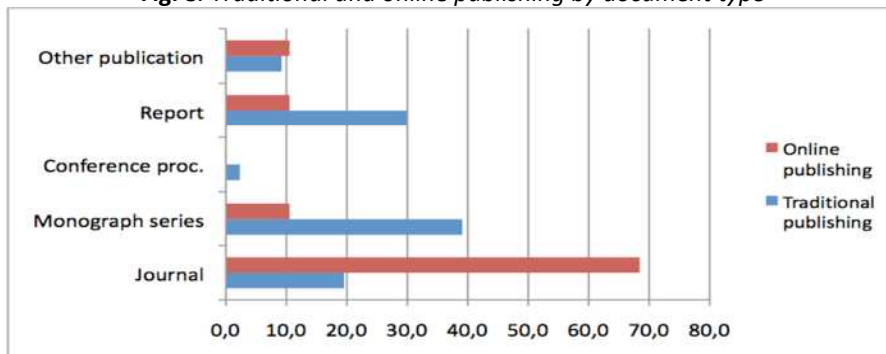
There is an evident concentration of traditional products in the Department of Humanities and Social Sciences, with 70 out of 106 and are predominantly only in print version. The digital products represent a small proportion (fig. 2) of the editorial products (17%). The Department of Earth and Environment accounts for 42% of digital products, a high value considering the total number of products produced and compared to the 12% of the Department of Humanities and Social Sciences.

Fig. 2. Digital products by department (%)



As expected (fig. 3) we found that the journals are the majority of document types that are published online. This is not surprising because usually the Journal product is the first type of document that was supported by electronic manuscript handling system and has now shifted to online publishing. In our sample, journals account for 13 out of 19 – equal to almost 70%, - 2 of which are digital-born and open access. Both are journals produced by the Department of Humanities and Social Sciences.

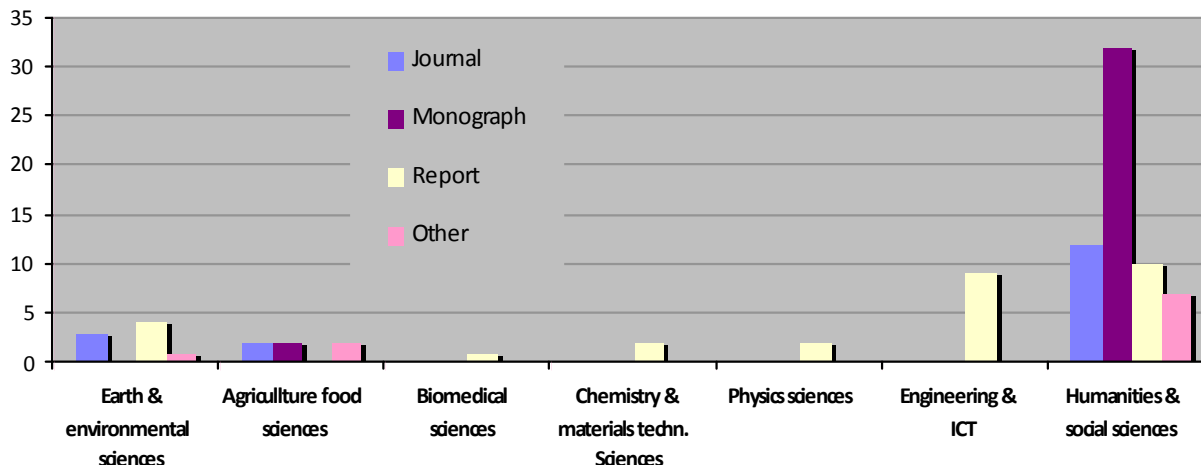
Fig. 3. Traditional and online publishing by document type



4.2 Traditional and online publishing

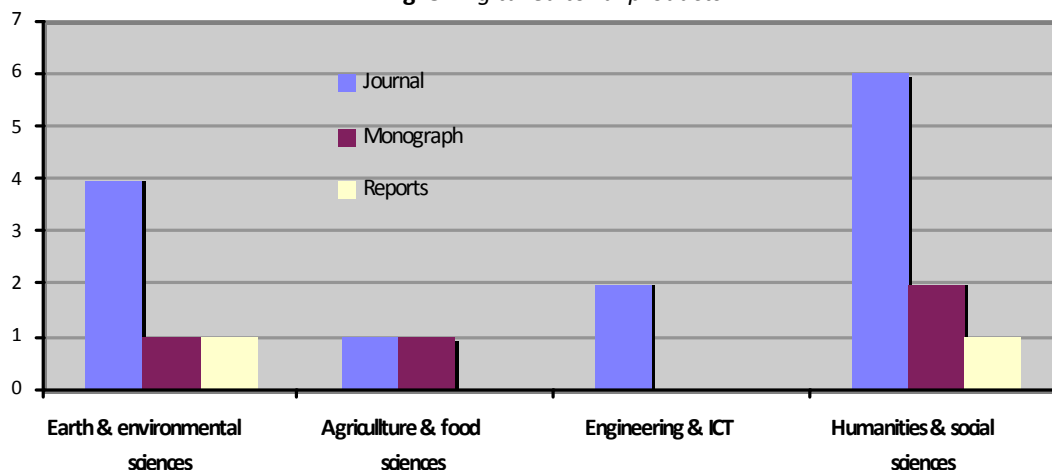
Figure 4 shows traditional editorial products produced by CNR Departments. Many of them are established and valuable collections and have been published without any interruption since the foundation of the Institute. They are products that are typical output of a scientific community, including journals, monograph series and reports. The Department of Humanities & Social Sciences produces the whole range of products, especially Monographs that are well known in the international scientific community for their high-profile. While reports are almost exclusively in-house productions of other Departments and equally quantitatively relevant.

Fig. 4. Traditional editorial products by Department



The figure below shows the digital products managed by electronic publishing systems. They are, as already mentioned, mainly Journals that more frequently shifted from printed to digital formats and adopted proprietary or open source systems to support the whole publishing process taking advantage of flexibility, easy reusability of the content and cost saving. The majority of online publishing journals are concentrated between the Department of Earth & Environment and of the Department of Humanities and Social Sciences, while lower numbers are found respectively for the Departments of Engineering and Information Communication Technology and only a single example for that of Agriculture & Food.

Fig. 5. Digital editorial products



Clear differences in the digital publishing can be noted for the other kind of documents such as reports and monographs. Of course we think that these documents too can benefit from the advantage of electronic publishing systems¹. (Costigan 1999, 2004; Crow 2009).

As regard the reports series, there are not many of them that benefit from electronic publishing systems. The general practice is that “high quality printed report series” have been gradually replaced by digital formats, and sometimes back issues have been digitized and made freely available from the institutes’ website. This conversion from paper to electronic formats has not changed the production process, even if there is a clear advantage for end users. However, the use of electronic publishing systems to manage also the production and distribution of reports could represent a step forward in terms of editorial quality and above it can introduce additional services such as peer review, indexing and abstracting in general and specialized search engines facilitating the web discovery.

In this context, it is well known that the CNR central library collects and manages valuable reports produced over time by CNR research Institutes. They could represent a starting point for a consistent digitations project within CNR. Moreover, it would be interesting to connect this initiative with the print collections stored in OPENSIGLE. This initiative would take into account Lynch’s observation in reported in a recent study (Lynch, 2009; Hahn, 2008),.

In the case of monographs as we have seen previously, (fig. 4), a certain number of them are produced in-house following the traditional publishing process, while only a limited number is currently managed by e-publishing systems (fig. 5), because this would imply the change of the whole editorial process. In fact their shift to an innovative production model is more complex if compared, for instance, with the production of reports. However, considering that monographs, especially in Humanities and Social Sciences, are mainly scientific-based output and often a *niche product*, the use of e-publishing systems could be a valid alternative for cost containment, and we know that monographs in particular suffer budget constraints. Besides they are generally locally oriented, and often publish in their native language since they have greater difficulty finding commercial publishers. Last but not least, they are very expensive to produce and only few commercial publishers accept to publish in narrow and targeted study areas and even if they publish, they do not always guarantee a wide diffusion of contents or indexing and abstracting services.

In this regard many studies and reports stress these critical issues within Monographs. Some of them focus on the insustainability of traditional business models and report examples of successful experiences carried out by e-publishing library services as well as through alliance with more collaborative commercial publishers. (Hahn 2008; Alenius, 2007; Besen 2012, Ferwerda 2010; OAPEN project²)

4.3. Some characteristics of the editorial business process

Table 3 provides an overview of the main characteristics of digital products management that we have analysed. For the majority (68%) of digital products all the activities connected with their production and diffusion are carried out in-house, from content acquisition (including submission and peer review) to the online publications of the content. For a smaller proportion (26%) of digital products, the Institutes externalize part of the editorial process, often the phases of copyediting and publishing. As can be noted paper and electronic diffusion is still the dominant solution and for these reasons some commercial publishers are in charge of providing the printed copy while in other cases print versions are available on demand directly from the producers. Sometimes commercial publishers are responsible for management of membership fees and subscriptions.

Tab. 3 *Profile of digital products (=19)*

Production & Diffusion	n.	%
In-house	13	68,4
Partially in-house	5	26,3
National commercial publisher (for print distribution)	10	52,6
International Commercial publisher	1	5,3
Access policies		
Full OA	15	78,9
Delayed OA	2	10,5
Open access online/Subscription for print	1	5,3
Subscription (online & print)	1	5,3
Technology used		
Content management system	12	63,2
Open source electronic publishing system	6	31,6
Publisher'platform	1	5,3
Copyright & Licensing		
Yes	14	73,7
Not available	6	31,6
Peer review		
Yes	14	73,7
Not available	5	26,3
Scientific committee & editorial board		
Yes	14	73,7
Not available	5	26,3
International standard codes		
Yes	16	84,2

Coming now to the technology used, the table shows that 5 out of 19 digital products are managed using Open Journal System (OJS). OJS is a journal management and publishing system that has been developed by the Public Knowledge Project. It manages every stage of the publishing process, from submissions through to online publication and indexing, including peer review process. It is currently the most suitable and widely used system to manage online publications. OJS it is also OAI-PMH compliant and supports interactive functionalities, such as reading and social network tools (Willinsky 2005; Brian & Willinsky 2010).

However, in our survey we found that the majority of digital products are managed using content management systems. It is interesting to note that in one case the Institute has developed locally an OAI compliant open source publishing system to manage its journal. This is the case of the journal "Archeologia e calcolatori" (Moscati 2009)

With regard to access policies to the content, the majority of products are open access (15 out of 19) and most of them provide information related to copyright and licensing. Peer review and scientific committee and/or an editorial board are also contemplated for the majority of them. Finally, almost all have International standard codes.

Summarizing, in our survey the hybrid business model is the dominant solution: it combines in-house editorial activities with partial externalization. Moreover, it mixes open and toll-access as well as print and electronic format to diffuse the contents. In general we can say that each CNR Institute has its own business model and even within the same Institute there are different management models depending on the type of product.

4.4. Examples

The following examples are representative of different business models adopted by CNR institutes.

1. Journal of limnology - This is an example of the evolution from traditional to innovative publishing. It is a journal directly published by CNR since 1942 in the very specialized field of limnology one. At that time it was a forerunner in environmental studies. Since 1999, it has been an electronic open access journal. The content acquisition and management together with the peer review process is still managed by the CNR institute, while all the activities related to copy-editing and publishing are outsourced to an external e-publishing service that uses OJS platform. If subscriptions for the printed version are required, there is a local commercial publisher that provides them.

2. Archeologia e calcolatori - This is an example of best practice. In 2005 the Institute followed the open access principles and developed an OAI-PMH compliant e-publishing system. The OAI SISTEMA used a simplified solution to manage an OAI-PMH repository. In 2006 Archeologia e Calcolatori was indicated

by Los Alamos National Laboratory study as an example of systems that enabled easy and efficient content discovery.

3. *Geothermics* - The journal *Geothermics* represents a different example compared to the previous ones. It was founded in 1972 by the CNR institute and appeared immediately as an international peer reviewed journal. At the beginning the editor in chief belonged to CNR while at the moment this journal has become one of the Elsevier journals and the business process is completely managed by this international commercial publisher and recently the journal lost its CNR branding.

4. *IRPPS Editoria Elettronica* (e-Publishing service) - This last example is different from the examples described above. The project of introducing an e-publishing system in IRPPS institutes was designed and carried out by the library with the collaboration of internal researchers. We should stress that the introduction of Open Journal System (OJS) in the publication process has varying aims depending on the types of products. IRPPS started with working papers and monograph series, introducing internal peer review for WPS and external peer reviews for monographs. The editorial staff also intend to re-publish old reports that represented a breakthrough in population studies, giving a second life to GL documents. The entire business process is carried out internally.

Conclusions

This preliminary survey focuses on well-established editorial products published by CNR Institutes, with the general aim to better understand to what extent use of new digital publishing technologies have innovated their editorial process and products. Despite a limited number of innovatively managed products, they are in line with scientific scholarly publishing connected with digital publishing technologies and on open access publishing models.

From the results of the survey, it is also clear that the well-established and standardized products, with a solid tradition in print publishing are concentrated in the Department of Humanities and Social Sciences, where the products are predominantly monographs. At the moment innovative products managed by online publishing systems are concentrated in the Department of Earth & Environment.

Concerning the business process, first of all we can say that disciplinary fields do not influence the business model, or the trends in adopting new technologies. What we discovered to be very important is the evolution towards innovative products generally based on products having a long and stable tradition, representing the history of the institute as well as the scientific achievement in a specific field. Of course many back issues aren't accessible online and we consider this the ongoing task in future CNR digitization projects, which we think should be planned and defined at departmental level.

We have seen that the business model adopted is not uniform in all institutes, each one having found its own solution. Sometimes the entire business process is managed in-house, but there is a widespread tendency to contract the publication and distribution of both the electronic and printed format to commercial publishers. From our survey some best practice examples of in-house publication management seem to emerge, especially those using electronic publishing systems with their value-added services. Certainly the use of e-publishing systems increase the quality of editorial products: additional services can increase the visibility; indexing and abstracting of products in search engines make them more easily retrievable, and the peer review process can be quicker.

Moreover, products that represent narrow and targeted study areas with a limited potential market and therefore encounter difficulty finding commercial publishers could benefit greatly from in-house publishing services. Of course, even if costs are reduced by the use of e-publishing systems, it implies the setting up of its organization, training, maintenance and updating. For these reasons a possible sustainable model could be adopted by CNR organizing it at a departmental level in order to achieve economies of scale and to optimize coordination actions.

Taking into account that e-publishing initiatives developed locally by CNR Institutes will grow in the near future, our study was an exploratory pilot study for long-term program publishing activities. In the future we will test the role of CNR libraries in the development of e-publishing services together with the CNR research community, that should be involved in founding the best innovative publishing practices suited to their needs.

In this context it is well known that academic and research libraries that have had a fundamental role in supporting Open access practices in the construction of Institutional repositories, and digitization programs are currently moving toward the development of additional services for their community scholars. This is in line with the onus on libraries to reshape their role in the digital age following changes in scholarly communication models. In this context library publishing services represent a new modality to diffuse scholarly research outputs, improve the quality of in-house published products and decrease costs of publication.

References

- Adema Janneke, Schmidt Birgit (2010). From service providers to content producers: new opportunities for libraries in collaborative open access book publishing. *New review of academic librarianship* 16:S1.
- Alenius Marianne (2007) Open access, University press, and editorial responsibility. Conference paper at the Workshop on Open access, 23-24 April 2007. URL: www.mtp.hum.ku.dk/cgibin/PDFmedopenaccess/Open_Access_Univer_0_0_9788763510868.pdf
- Bareghed A., Rowley J., Sambrioi S. (2009). Towards a multidisciplinary definition of innovation. *Management decision*, 47 (8).
- Besen Stanley M. Kirby Sheila Nataraj (2012). E-books and libraries: an economic perspective. American library association.
- Borgman Christine L. (2007) Scholarship in the Digital Age: Information, infrastructure, and the Internet. Cambridge : MIT Press.
- Brian D. Edgard; Willinsky J. (2010). A survey of the scholarly journal using Open Journal Systems. *Scholarly and research communication*. 1(2). URL: <http://pkp.sfu.ca/files/OJS%20Journal%20Survey.pdf>
- Costigan S. International Affairs Research and the Web (1999). In: Grey Literature GL99. Washington, D.C.
- Cooke M., S. Costigan (2005). Making Your Way Through Grey: Metadata, MARC and User Tools. In: GL6 Conference Proceedings. Amsterdam : TextRelease, 2005.
- Crow R. (2002). The case for institutional repositories: a SPARC position paper. Washington, D.C. : SPARC. URL: http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf
- Crow R. (2009). Campus-based publishing partnership: a guide to critical issues. Washington, D.C. : SPARC. URL: <http://www.arl.org/sparc/partnering/guide/>
- Ferwerda Eelco (2010). Open access monographic publishing in the humanities. *Information Services & Use*, 30 (3). URL: <http://iospress.metapress.com/content/16wg61l0mg6426w8/fulltext.pdf>
- Gierveld H. (2002). A conceptual analysis of functions, processes, and products in the scholarly communication chain. Elpub 2002 proceedings . URL: <http://elpub.scix.net/data/works/att/02-16.content.pdf>
- Hahn, Karla L. (2008). Research library publishing services. New options for university publishing. American Research Library Report 258. URL: <http://www.arl.org/bm~doc/research-library-publishing-services.pdf>
- Houghton John W. (2011). The costs and potential benefits of alternative scholarly publishing models. *Information Research*, 16 (1)
- Iglezakis Ioannis, Synodinou Tatiana-Eleni (2011). E-publishing and digital libraries: legal and organizational issues. Hershey – New York : Information science reference
- Lynch Clifford A. (2009). Special collections at the cusp of the digital age: a credo. Research library issues: a bimonthly report from ARL, CNI and SPARC 267. URL: <http://www.arl.org/bm~doc/rli-267-lynch.pdf>
- Mill Fethy (2000). Trends in publishing academic grey literature: examples from economics. *The International journal on grey literature*, 1(4)
- Moscatti Paola (2009). Archeologia e calcolatori: le ragioni di una scelta. *Archeologia e calcolatori*, 20. URL: http://soi.cnr.it/archcalc/indice/PDF20/12_Moscatti.pdf
- Mullins, James L.; Murray-Rust, Catherine; Ogburn, Joyce; Crow, Raym; Ivins, October; Mower, Allyson; Newton, Mark P.; Nesdill, Daureen; Speer, Julie; and Watkinson, Charles (2012). Library publishing services: strategies for success. Research Report Version 1.0 Libraries Research Publications. Paper 136. URL: http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1166&context=lib_research
- OAPEN project. URL: <http://www.oapen.org/home>
- Ridi R. (2001). Gli incerti confini dell'editoria digitale. URL: <http://www.burioni.it/forum/ridi-confini.htm>
- Roosendaal Hans E. (2004) Driving change in the research and HE information market. *Learned publishing*, 17(1).
- Roosendaal Hans E.(1998). Forces and functions in scientific communication: an analysis of their interplay. CRISP 97 Cooperative Research Information Systems in Physics. URL: <http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html>
- Stempfhuber M. et. Al. (2007). Enhancing visibility: integrating Grey literature in the SOWIPORT cycle. In: 9th International Conference on Grey Literature (GL9). Tools for publishing, archiving and accessing GL. Amsterdam : TextRelease, 2007.
- Tenopir Carol, Birch Ben, Allard Suzie (2012). Academic libraries and research data services. Current practice and plans for the future. An ACRL White paper. www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf
- Willinsky J. (2005). Open Journal Systems; an example of Open source software for journal management and publishing. *Library Hi-Tech*, 23 (4). URL: http://pkp.sfu.ca/files/Library_Hi_Tech_DRAFT.pdf

¹ CIAO (Columbia International Affairs Online) is one of the first digital publishing project to put online scholarly information in the field of international affairs. CIAO includes now full-text online books, journals, seminars and research projects, working papers, reports and conference proceedings. Launched in 1997, CIAO project was presented at a previous GL Conferences (1999, 2005). It is now managed by EPIC (Electronic Publishing Initiative at Columbia University), together with other two e-publishing projects developed at the Columbia University

² OAPEN project stands for Open access publishing in European Networks- OAPEN developed and implemented open access publication model for peer reviewed academic and research monographs in the Humanities and Social Sciences.



IAEA

Department of Nuclear Energy

Any news?

Yes, NE News!

*Download our "NE News" iPad app and get one-stop-access
to IAEA nuclear energy information and publications*



Available on the
App Store



13-32541

www.iaea.org/inis

Collection of Conference Proceedings and Improving Access to the Full Text of Proceedings *

Misa Hayakawa, Shun Nagaya, Mayuki Gonda, Takeyasu Fukazawa,
M. Yonezawa, and K. Itabashi (Japan)

Abstract

Conference Proceedings are “grey literature” due to the fact that they are not made commercially available frequently. While many Proceedings are published on the Internet, there are specific issues that can affect access, such as changes in the URLs. This paper introduces the case of the Japan Atomic Energy Agency (JAEA) library as an example of efforts to improve access of Proceedings using the Internet. JAEA Library uses the Internet to make available presentations by JAEA researchers. The paper notes that the conference secretariats tend to be temporary bodies and the links to conference websites are not permanent. The paper reports our investigation into these problems, and we introduce a new approach to provide access to these Proceedings.

1. Introduction

1.1 Conference Proceedings, a type of “grey literature”

Conference Proceedings are records of the papers and data which researchers submit for a specific theme at a conference or meeting. Proceedings contain information on the latest research trends, therefore, it is important information resources in the fields of science and technology.

These Proceedings papers are characterized by a shorter period between the writing and publication than the journal papers. In recent years, the period between writing and publication has been shortened even further by the Internet [1]. Due to this, the importance of Proceedings is growing.

Many researchers assume that the reliability of Proceedings is lower than journal articles. Nonetheless, many researchers use Proceedings as (1) points of reference in developing research topics, and (2) to investigate the latest research trends [2].

However, Proceedings are often distributed to the participants at a conference, and are often not commercially available. It is difficult for libraries to ensure a complete collection of Proceedings.

Proceedings have traditionally been published as conventional books. However, forms of publication have changed and many Proceedings are published as digital media, such as CD-ROM or Flash Memory, or are published only on the Internet.

Proceedings on the Internet may seem like an easy to collect, but there are in fact many challenges in terms of collection and use. For example, the websites on which Proceedings are usually published are owned and operated independently, and cross-searching these sites is difficult. Additionally, there is a problem of visibility as Proceedings on the Internet cannot be accessed when users do not know if the sites exist [3]. Furthermore, conference secretariats are often temporary bodies and the links to conference websites are not permanent. Changes to URLs happen frequently; a specific challenge relating to access to information on the Internet.

Proceedings are, therefore, still “grey literature” because they are difficult to search and use.

1.2 Purpose of this study

This paper discusses ways to improve access of Proceedings especially on the Internet.

The forms of publishing Proceedings are changing, and number of Proceedings available on the Internet is increasing. They are important as a means of providing the full text of papers to researchers. However, Proceedings published on the Internet cannot be accessed if their existence is not known, and currently it is difficult to search and use Proceedings papers.

We would like to introduce the case of the JAEA library as an example of improving access of full text Proceedings on the Internet.

We also address how the many sources of information about a meeting such as the conference website and papers on institutional repositories can be used to improve the access of Proceedings.

We introduce a new approach in providing conference information and linking to the full text Proceedings on the Internet.

* First published in the GL14 Conference Proceedings, February 2013.

2. Publication forms of Proceedings and library collections

2.1 Survey of the forms of Proceedings publications

We begin with an survey of the current forms of publishing Proceedings.

Fukazawa conducted a survey on the various forms of publishing Proceedings in 1982 and 1992. The survey classified Proceedings by publication forms such as books, journals, reports, and others [4]. 64% of Proceedings were published in journals or books.

We conducted our own survey to ascertain whether there had been any changes in the publication forms. We examined the Proceedings of international conferences in which researchers from the Japan Atomic Energy Agency (JAEA) participated in FY2010 (212 conferences). JAEA researchers mainly participate in conferences in the field of nuclear science.

We classified Proceedings forms as books, journals, reports, and other publications (Figure 1). Then, we classified books by mediums such as book form and CD-ROM (Figure 2).

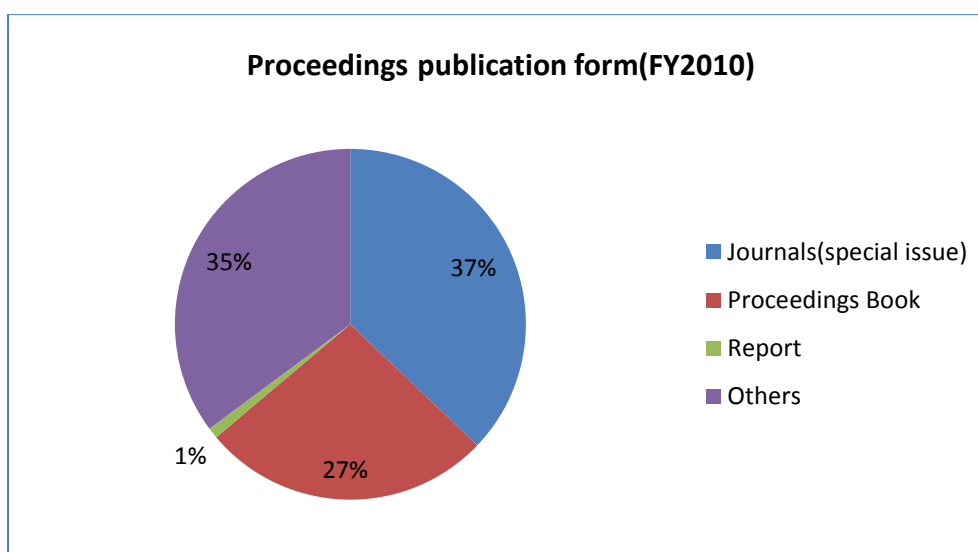


Figure 1: Proceedings publication form

We found that Proceedings are mainly published as journals (in special issues or as conference series). Next, Proceedings were commonly published in books. This is similar to the findings of Fukazawa's survey. Incidentally, the "others" category includes cases where there was no form of publication of Proceedings (where, for example, only an abstract or program book were made available), and where there was an unknown publication form.

Proceedings published in journals are often commercially available, making them easy to access.

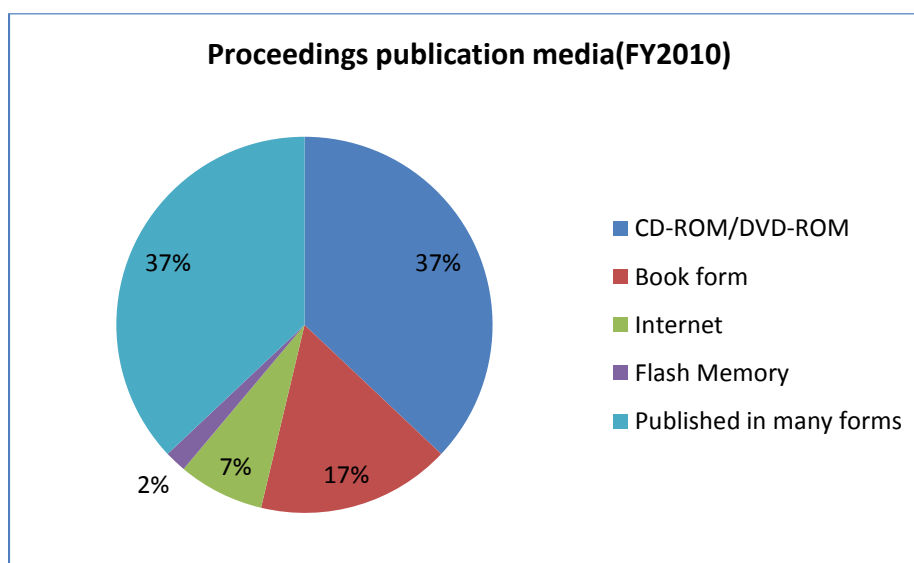


Figure 2: Proceedings publication media

We classified Proceedings by mediums, such as traditional books, CD-ROM/DVD-ROM, Internet sites, and Flash Memory. Proceedings published in two or more forms were classified as “Published in many forms” and duplication was not counted.

In all, 17% of the Proceedings were published as books, much less than the other forms. In comparison, 37% of the Proceedings were published as CD/DVD-ROMs.

Additionally, this survey shows many Proceedings are published more than one publication form. Many of these cases include Proceedings published on the Internet. Proceedings books or CDs were distributed to conference participants, and after the conference they were published on the Internet where they were available to buy. Published in a variety of forms in this way improves the availability of Proceedings.

Only 7% of the Proceedings were published solely on the Internet. In some cases, this publication form was open to the public, while some were only made available to the conference participants. This was, therefore, not necessarily a widely accessible form of providing access to the full text.

A few conference Proceedings were published as flash memory, although we do not dwell on this form of publication in this paper. Flash memory Proceedings were often distributed to conference participants at the conference venue, and pose a particular challenge in terms of collecting Proceedings by libraries. Additionally, it is difficult to preserve flash memory readability because they deteriorate and standards change. Where Proceedings published as flash memory do not conform to a meta-data system, it is also necessary to consider how Proceedings are made available by libraries. JAEA Library currently does provide a flash memory. However, to collect and provide flash memory still remains a challenging issue.

Proceedings published as CD-ROMs present similar problems to those associated with flash memory. For example, a Proceeding’s CD-ROM published in 1996 created with Windows 95. CD-ROM also cannot be read when the operating systems or standards have changed. It is necessary to convert CD-ROM / DVD-ROM to up-to-date forms or to print.

2.2 Collection of Proceedings at the Library

We surveyed the library’s collection of papers presented by JAEA researchers in FY2010 to confirm access of the Proceedings (not including those published as journals or reports) in libraries. We inspected the collections at the JAEA Library, the National Diet Library (NDL) [5], and CiNii Books [6] (to look for information about books and journals held by other university libraries in Japan). Additionally, we also investigated the availability of Proceedings on the Internet (Figure 3).

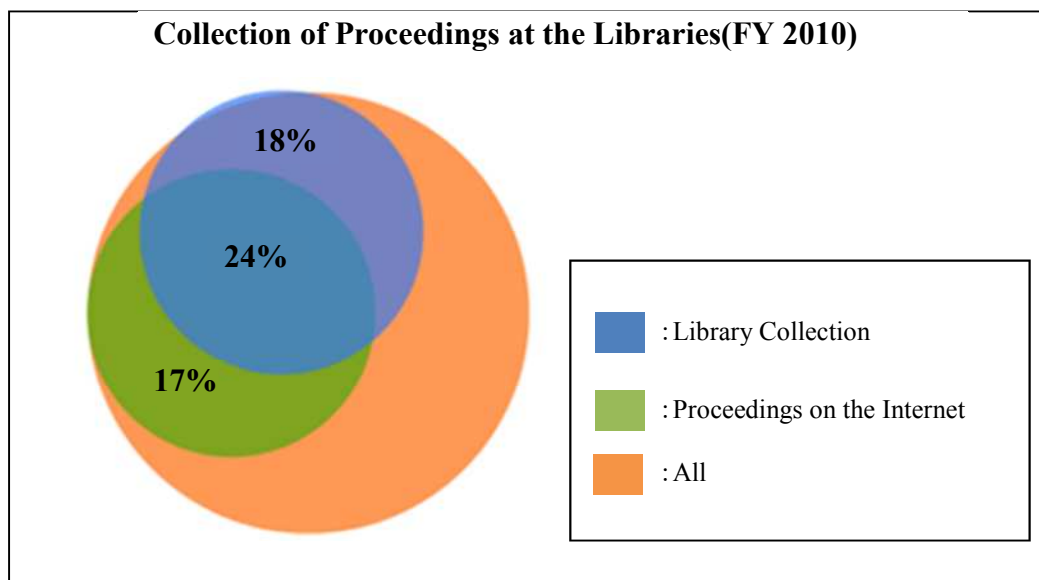


Figure 3: Collection of Proceedings at Libraries

The JAEA Library is one of the largest nuclear information centers in Japan, and we were collecting information on Proceedings specifically related to the field of nuclear science. We collected Proceedings from a variety of sources, including conference websites and from the researchers themselves.

However, we found that the JAEA Library together with the other libraries could provide only 42% of the Proceedings. As already mentioned, 17% of the Proceedings appear only on the Internet. This shows the difficulty of making Proceedings available in library collections. On the other hand, 41% of Proceedings were available as an Internet version. Providing links to the Proceedings on the Internet as well as the library’s collections is important to improve access to Proceedings.

3. Management & providing links at JAEA

3.1 Providing links to Proceedings on the Internet within JAEA

As mentioned above, Proceedings on the Internet is increasing. Providing links to Internet Proceedings will improve access of Proceedings. We also advocated the management and provision of a system of the R&D results in JAEA as well as the links to the full text of the Proceedings.

We manage the publication and presentation of information by JAEA researchers at the JAEA Library. At JAEA, researchers are requested to submit bibliographic data (title, author, journal/meeting name, etc.) via web forms before they undertake presentations or submit articles. Through this procedure, the library staff manage the information and produce an authority file of author names, conference names, and journal/book names.

The ability of this form of unified management by the library is particularly important where conference names have various notations.

Once the information is collated, we disseminate it via the Internet using the JAEA Originated Papers Searching System (JOPSS) [7].

Additionally, since July 2011 we have provided hyperlinks on JOPSS for the full-text versions of articles submitted by JAEA researchers. Providing links via JOPSS allows direct access to the Proceedings on the Internet.

The registration of the DOI/URLs is performed every month. Information is sent in the form of a hyperlink from a JAEA R&D result management system. The search and input of the DOI are then carried out using the Web system (Figure 4).

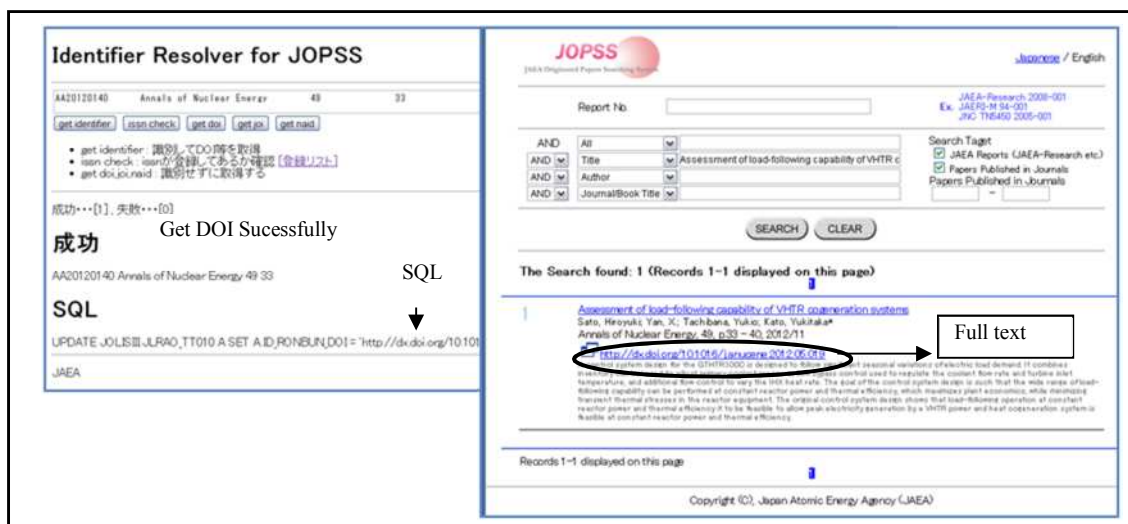


Figure 4: Providing hyperlinks to full text

We currently provide access to 14,459 papers from conference. Among them, 3,137 are hyperlinked to the full-text version or abstract (as of October 2012). Past paper/presentation data will also include DOIs/ URLs. Additionally, conference information (conference name, venue, and date) also provide from JOPSS since September 2012.

However, few Proceedings papers published on the Internet have DOIs (except for those published as journal papers). For example, Proceedings papers published on conference websites almost never have DOIs. In these cases, we provided a URL for the paper (such as a PDF etc.). Additionally, even when full-text versions could not be accessed from the Internet, many conference websites provided abstracts of conference papers. We therefore provided links to the abstracts of conference papers when full-text versions of the Proceedings could not be accessed from the Internet.

3.2 Surveys of Proceedings on Internet/Conference websites

In recent years, many conferences have had dedicated websites which provide conference information (dates, venues, programs, abstracts) to the public and the conference participants. However, this information faces a challenge that is specific to the information on the Internet, i.e., the possibility that the access links to the URLs are either changed or the websites are not permanent.

We also, therefore, investigated access of the conference websites and the contents of websites for the conferences JAEA researchers participated in FY2010.

Access of conference websites with the URL registered before the meeting (Figure 5).

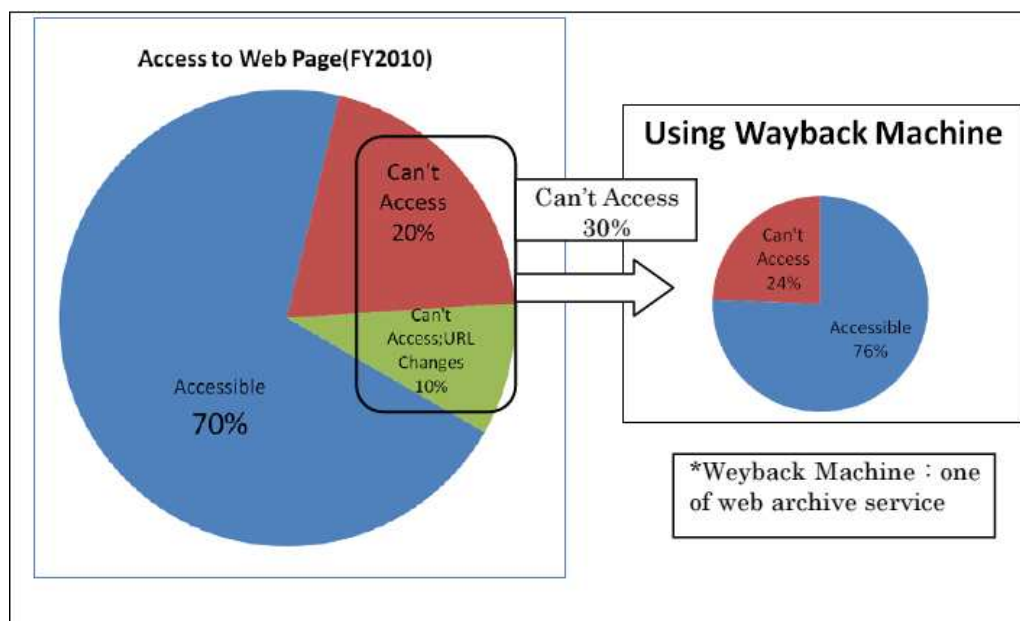


Figure 5: Access to conference webpage

Overall, 30% of the conference websites can no longer be accessed within 3 years. In all, 20% of the websites are no longer accessible because the website no longer exists, while 10% of the websites are not accessible because the URL has been changed. But 76% of these websites can access when use the web archive service.

Our investigation of website content showed that 10% of the websites provided the full text of the Proceedings, while 23% provided the abstracts. Some websites also provided presentation files and movie files, etc.

Some Proceedings available on the Internet were published on conference websites, while some were published on websites dealing with Proceedings in specific fields such as JACoW [8]. Moreover, some Proceedings available on the Internet were published on publisher websites.

Proceedings available on conference websites were published in a variety of formats, such as PDFs of each paper or PDFs of the whole Proceedings combined. Most of them were open to the public for no charge. Some, however, which were accessible only to conference participants. It required the ID and password to access details. These accounted for very few of the Proceedings, but did make it difficult to obtain full-text versions of the papers. Proceedings made available by publishers generally have DOIs, making them more accessible than other forms.

3.3 A new approach to providing Proceedings information

A large amount of conference information and a great number of Proceedings are available to the public on the Internet. The links to Proceedings and conference information on the Internet as well as in the library's collection is useful to improve access to the Proceedings. We, therefore, developed a prototype webpage that provides conference information, library collections of Proceedings, and Proceedings availability on the Internet using Google Calendar. Google Calendar is a free online calendar, and it is useful for information sharing and possible to integrate with other systems (Figure 6).

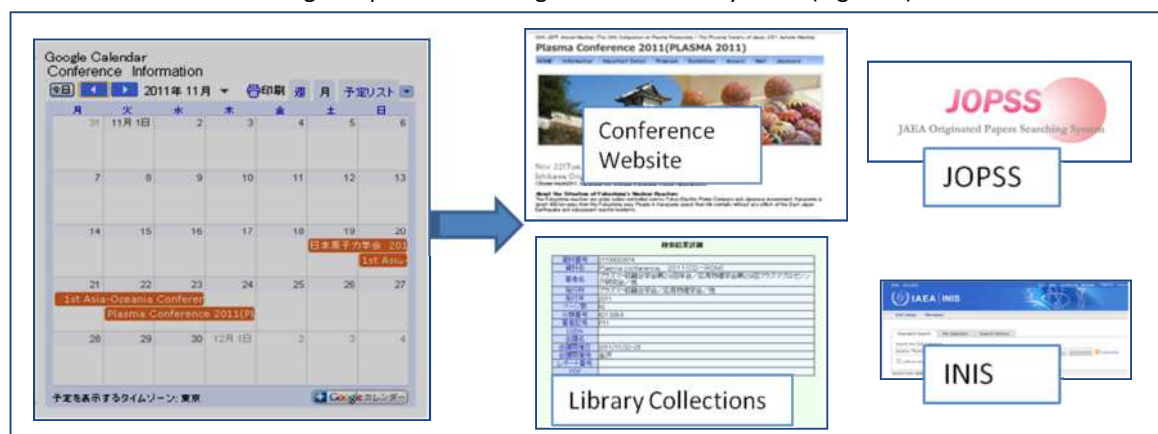


Figure 6: Google Calendar to Internet sources

One of our initial difficulties was in providing conference information where the websites had been deleted within a few years of the conference. We found that it was possible to solve this problem by using the web archives that has developed in recent years. The Web archives collect Internet web pages, archives them, and makes them available to the public. We used the Internet Archive's Wayback Machine [9] to provide conference information from the deleted conference websites. Wayback Machine is a major Internet archive, offering access to Web pages from 1996. As mentioned above, web archives will improve access to deleted conference web pages.

Additionally, we use library collections information from the JAEA Library/NDL/CiNii Books. This covers most of the information on Proceedings in Japan.

Some of the Proceedings papers were included in institutional repositories by the authors. In recent years, institutional repositories have become popular, and some institutional repositories include Proceedings not available anywhere else. For this reason we also provided the hyperlink to the search results of Japanese Institutional Repositories Online(JAIRO) as a way of increasing the availability of conference Proceedings papers. JAIRO includes 11,888 of presentation files (2010/10).

We also provide hyperlinks to the search results of the International Nuclear Information System (INIS), which includes information from conference Proceedings in the fields of nuclear science and technology. INIS hosts the world's largest collection of published information on nuclear science and technology [10], and is appropriate for conference information relating to the participation of JAEA researchers.

We have registered conference information from FY2010.

This conference information on Google calendar is not yet open to the public, but we want to improve the resource and make it publicly available.

4. Summary

There are a variety of forms of publishing Proceedings and the number of Proceedings available on the Internet is increasing. Using links to Proceedings on the Internet as well as library collections can improve the access of the Proceedings. However, websites are not permanent and changes in URLs are frequent. The URLs of Proceedings on the Internet, especially those on conference websites, are particularly susceptible to change. When we provide links to Proceedings, we need to, therefore, undertake periodic checks on the links. Web archive also improve access to web pages when URL changes or deleted pages.

Additionally, some of the Proceedings available on the Internet are not open to the public, making it difficult to access the full text. These accounted for very few of the Proceedings, but we have to consider how to provide these Proceedings.

In recent years, some of the Proceedings papers were included in institutional repositories by the authors. Institutional repositories improve access to full-text Proceedings that are not available anywhere else.

There are a large number of information sources on the Internet. We aim to improve the access of the Proceedings and aim to provide access to this information.

References

- [1] Tsuda, Yoshiomi. Conference Proceedings and materials. The present state of conference Proceedings and materials. *Jyoho no Kagaku to Gijyutu*. 2005, 55(5), p.208-213. (Japanese)
- [2] Toyoda, Yuji. Possibilities for effective use and distribution of preprints. *Jyoho no Kagaku to Gijyutu*. 1998, 48(6), p.336-343. (Japanese)
- [3] Ikeda, Kiyoshi. Grey literature. Overview: rethinking the grey literature's definition. *Jyoho no Kagaku to Gijyutu*. 2012, 62(2), p.50-54. (Japanese)
- [4] Fukazawa, Takeyasu. Analysis of the published Proceedings of international conference-Scientific conference held in Japan. *Jyoho no Kagaku to Gijyutu*. 1993, 43(10), p.913-915. (Japanese)
- [5] NDL OPAC. <https://ndlopac.ndl.go.jp/>
- [6] CiNii Books. <http://ci.nii.ac.jp/books/>
- [7] JOPSS. <http://jolissrch-inter.tokai-sc.jaea.go.jp/search/servlet/interSearch>
- [8] JACoW. <http://www.jacow.org/>
- [9] Wayback Machine. <http://archive.org/web/web.php>
- [10] INIS. <http://www.iaea.org/inis/>

Innovation, language, and grey literature*

Claudia Marzi (Italy)

Abstract

Language and innovation are inseparable. Language conveys ideas which are essential in innovation, establishes the most immediate connections with our conceptualisation of the outside world, and provides the building blocks for communication.

Every linguistic choice is necessarily meaningful, and it involves the parallel construction of form and meaning. From this perspective, language is a dynamic knowledge construction process.

In this article, emphasis will be laid on investigating how words are used to describe innovation, and how innovation topics can influence word usage and collocational behaviour.

Especially in grey materials, and in a context-based approach, the lexical representation of innovative knowledge is closely related to the representation of knowledge itself, and gives the opportunity to reduce the gap between knowledge representation and knowledge understanding.

This will bring into focus the dynamic interplay between lexical creativity and innovative pragmatic contexts, and the necessity for a dynamic semantic shift from context-driven vagueness to domain-driven specialisation.

1. Introduction

Understanding the relationship between language and innovation is connected with understanding that language determines what can and what cannot be talked about, and therefore what can be achieved and what cannot.

Language conveys innovative ideas and gives body to knowledge itself. Language is the common ability of our species that makes us shaping and referring to things, events, and concepts with remarkable precision. This common communicative ability connects people into an information-sharing network, and information allows people to expand their knowledge.

The importance of efficiently deploying knowledge for a complete and successful exchange is easily understandable: through a better understanding of information new ideas can be captured and exploited.

Knowledge transfer and innovation transfer are nowadays ubiquitous processes. The entire system of knowledge refers to knowledge creation and application, by defining a process going from acquisition and sharing to transfer and application. Knowledge extraction involves heterogeneous tasks related to the acquisition, from unstructured textual data in digital format, of structured and classified information relating to research topics. Nonetheless, far from being readily or easily transferred from the originator to the user of a technology, knowledge faces barriers, such as ambiguity, difficulty to be interpreted and absorbed, difficulty to be retrieved.

With particular reference to Grey literature, the spread of Internet has enabled development of better bibliographic scientific databases with significantly improved capacity for storage and retrieval. In recent years, web searching has become the default mode of highly innovative information retrieval, though the main sources of digital information are unstructured or semi-structured documents. Information relating to developments in scientific research is collected in the form of abstracts or full publications, in large and growing bibliographic repositories.

Knowledge ambiguity may also depend on language ambiguity; a shared language is part of a related language, and "for knowledge to be exchanged and combined, there has to be a shared medium of communication" (Hedlund, 1999:11). Language has effects at all stages of knowledge transfer. Language generates on-going impacts beyond a simple knowledge transfer act, and it is simultaneously an active agent in the knowledge transfer process itself.

Our goal is to focus on how words and language structures become vehicle for knowledge generation, and in particular for innovation transfer, and what kind of infrastructural support can enhance innovative knowledge transfer.

* Revised version of the paper published in the GL14 Conference Proceedings, February 2013.

2. Background

2.1. Language of innovation and linguistic innovation

In considering the language of innovation, particular emphasis is laid on overlap and dissonance in terminology and word formation processes associated with innovation. The language of innovation suffers from a problem of lexical overabundance. Not only many words are offered, but different authors define or use these words in different ways; and this because the challenge is the complexity of categorising innovation.

Many terms are used to describe innovation (Linton, 2009, for an overview), and innovation itself offers new collocations and extension of use. Most of the differences in terminology can be accounted for by differences in perspective and domain.

Language change is a fundamental evolutionary phenomenon due to many natural, cultural and historical factors. In particular, we are interested here in shedding light on the phenomenon of terminology innovation and propagation of those novel forms across domains.

Through language cultural novelty can be transmitted vertically (from parents to children), horizontally (from peer to peer) as well as across generational gaps. Although many linguistic innovations fall into the category of what Andersen (1989) has called “fortuitous innovations” (i.e. spontaneous and purposeless innovations as the results of non-functional, non-intentional copying errors), in many cases the very structure of the innovation can be explained with reference to the speakers’ (synchronic) perception and meaningful re-interpretation of linguistic surface forms within the pragmatics of the situation. Meaning itself is a consequence of interaction and context.

The meaning of words can change over time and discourse and, in particular, words can take on new senses when used in novel contexts. Words with emergent novel senses often reflect an extension of use from one domain to another. In this sense, linguistic innovations arise in the context of existing rules which they modify.

2.2. Language ambiguity

Many words have more than one sense, and each of their sense is reflected by their distribution across contexts. Lexical semanticists make a classical distinction between semantically ambiguous words (e.g. *bank*), whose meanings can vary unsystematically, and polysemous words (e.g. *school*), where different senses exhibit a predictable relationship. Both ambiguous words and polysemous words can be disambiguated by defining their context of use. Polysemous words, in particular, can shape their meaning as a function of their context of use. As a consequence of this context-sensitivity, if a polysemous word-type appears more times in the same text (e.g. a single document), it is extremely likely that its different tokens will share the same sense. Although people do not need too much context to perform a disambiguation task, in Natural Language Processing (NLP) and Information Retrieval (IR) larger contexts make the task easier. Methods for computing relational similarities and disambiguating polysemous words, based on large text corpora, can make rough sense distinctions. Although this is far from reaching the sophistication of human judgement, the field is making considerable progress in context-sensitive word sense disambiguation and in the identification of conceptual relations between words.

In the following sections, corpus-based investigations are analysed in the perspective of a lexical representation of innovative knowledge. In the field of corpus linguistics advantages and limits of various corpora are analysed, depending on both linguistic and innovative knowledge research questions.

2.3. Corpus linguistics

Corpus-based linguistics is the study of language on the basis of large text samples – the corpora.

A corpus, as defined by Sinclair (1999: 171), “is a collection of naturally occurring language text, chosen to characterize a state of variety of a language. In modern computational linguistics, a corpus typically contains many millions of words: this is because it is recognized that the creativity of natural language leads to such immense variety of expression that it is difficult to isolate the recurrent patterns that are clues to the lexical structure of the language”. Although it must be considered that an appropriate size of corpus is strongly dependent on the phenomenon to investigate and the purpose itself. Another factor influencing the size of corpora relates to the degree of internal variation in the language or genre under study (Meyer, 2002).

In any case, corpora are incomplete. Rather, the issue is whether they are representative of the inquired phenomena; in other words, a corpus should be large enough to give an adequate representation of the language and more occurrences of the elements under investigation.

The importance of findings, either quantitative or qualitative, depends on the representativeness of the selected corpus for the research question.

Dealing with machine-readable texts offers the basis for purpose-specific research questions.

Occurrence, distribution, and importance, are different issues to be taken into account.

Salient domain-specific concepts and relations are most often conveyed in text through statistically significant terms. Rare words often denote the most salient pieces of content information of a document together with its level of subject-specificity. Recurrent word combinations are defined as COLLOCATIONS.

In corpus linguistics, collocation defines a sequence of words or terms that co-occur more often than would be expected by chance. An example of a phraseological (multi-word expression) collocation is the expression *strong tea*. While the same meaning could be conveyed by the roughly equivalent **powerful tea*, this expression is considered incorrect by English speakers. Conversely, the corresponding expression for computer, *powerful computers* is preferred over **strong computers*. Unlike idioms, collocations have a rather transparent meaning and are easy to decode; yet they are difficult to encode – like idioms – since they are unpredictable for non-native speakers, and moreover they do not preserve the meaning of all their components across languages. Collocations are flexible and they can involve two, three or more words in various ways.

In NLP collocational information derived from corpora is useful in the perspective of text analysis; for instance, in word sense disambiguation collocations are used to discriminate between senses of polysemous words. Frequently occurring collocates give the idea of semantic preference.

Corpus data can be considered as very useful for revealing typically lexico-grammatical patterns and functional aspects of language. Corpus-based studies on word formation show that productivity of derivational suffixes are more pertinent in certain kinds of texts than others. In other words, register variation plays a salient role in word formation. It's, however, essential to state that register distinctions are not defined in linguistic terms, but rest on context, domain, and purpose; and that contextual knowledge allows to support knowledge processes and to better access them.

2.4. The Web as a corpus

The World Wide Web has become a primary meeting place for information and communication, and it provides texts to be mined for lexicographers and linguists. As the web is constantly expanding, it represents an unlimited universe of information and data, and offers ubiquitous accessible information, and large volumes of information are available. Increasingly, corpus linguists have begun using the World Wide Web as a corpus for linguistic analyses.

The Web as a linguistic corpus makes it possible to investigate how words are used to describe innovation, and how innovation topics can influence word usage and collocational behaviour.

As a source of machine readable texts for corpus linguists and researchers in the fields of NLP, IR and Text Mining, the Web offers extraordinary accessibility, quantity, variety and cost-effectiveness. The Web and associated technologies have been both the catalyst for much linguistic creativity and the main vehicle for its dissemination. In contrast, any static corpus is cut off at the moment of its compilation.

However, the web is a particular kind of corpus, as an estimation of its size and especially its composition cannot always be assessed. Moreover, it must be seriously kept into consideration that investigation of selective corpora is better concerned with the description of use and structure of domain-specific language, by inquiring linguistic phenomena, such as co-occurrence distributions, collocational variability, derivational productivity, neologism coinage. While the notion of linguistic corpus as a body of texts rests on some related issues such as finite size, balance, permanence, the very idea of a web of texts brings about notions not only of flexibility but even of non-finiteness and provisionality.

How can data gathered from the web provide new insight into language usage? The Web as a corpus is a rich source of freely available linguistic data covering a lot of topics (Fellbaum 2005), but despite the great advantages in quantity and accessibility, there is no control for example on web posting, and especially concerning English data, a lot of data are posted by non-native speakers. In this sense, the language used on the Web does not represent thoroughly the standard usage. While statistically robust analysis of Web data to discover collocations can give a flavour of what Manning and Schütze (1999) defined as “a conventional way of saying things” by marking the most frequent expressions, high frequent occurrences cannot give a disambiguation of context usage and sense. Moreover, in using the web as a corpus especially when it is accessed through generic search engine, it is virtually impossible to replicate a test on the same data.

In short, the Web offers a huge repository of documents written in a multitude of – more or less standard – languages, of different types or genre, and constantly changing over time, though often

helpless in telling intended purposes and in offering background and contextual knowledge. On the other hand, domain-specific databases or domain-classified document repositories (e. g. *Open Grey*, *Pubmed*, and the like) are strongly associated with the increasing availability and accessibility of GL materials, attesting terminological creativity and lexical innovation by defining in forehand the various disciplinary sub-fields.

In what follows we propose to approach the tight inter-relationship between sense extension, context and innovation, by exploring the usage of words in contexts with NLP technologies. In Computational Linguistics and Computational Lexicology, sense identification and words sense disambiguation are commonly modelled by focusing on the distributional similarity of word usages in context. These techniques provide a key to a deeper understanding of a constructive view of lexical meaning as the by-product of the interaction of a word with its surrounding context (i.e. its collocates), and represent the methodological basis of the ensuing analysis.

3. Methodology and experimental evidence

3.1. Method and materials

The challenge of identifying changes in word sense has only recently been considered in Computational Linguistics.

To investigate the themes discussed in the previous sections genre-oriented and stylistically heterogeneous English texts are analysed, with the support of SKETCH ENGINE (Kilgarrriff et al., 2004), which is a corpus query tool, based on a distributed infrastructure, that generates *word sketches* and *thesauri* which specify similarities and differences between near-synonyms. By selecting a collocate of interest in a *sketched* word, the user is taken to a concordance of the corpus evidence giving rise to that collocate.

Ambiguous and polysemous words have been selected with particular reference to innovative domains, and their collocations are analysed. In particular, we considered the domain of brain sciences and new technologies of brain functional imaging, the domain of knowledge management processes, and the field of information technologies, by mainly focusing on the following test words: IMAGING, RETENTION, STORAGE, CORPUS, NETWORK, GRID.

The selected words present a potentially high degree of semantic ambiguity or polysemy and different degrees of semantic specialisation, which can be analysed objectively by studying their context collocations.

For a terminology exploration, both domain-specific and general-purpose texts materials are selected by using generic search web engine queries (www.google.com by using seed words), domain-specific databases and type coherent multidisciplinary large corpora (e.g. www.opengrey.eu, www.ncbi.nlm.nih.gov/pubmed by selecting the domain). Collocations and concordances are then compared with large balanced corpora (e.g. the British National Corpus, British Academic Written English, New Model Corpus, and the like, whose size ranges between 8 M and 12 G tokens).

3.2. Results

By comparing in different contexts of use collocates and keywords selected from reference corpora – both specific and generic – with simple keywords search in web context, we investigate the ambiguity vs. polysemy gradient, showing how dynamically word meanings are adjusted to novel usage, and how difficult it could be to disambiguate polysemy words without a predefinition of the specific context.

All six terms exhibit distinct senses when they are used in different domains/discourse contexts. However, the extent to which different senses of the same term are mutually related differs considerably from one term to another. The two different usages of *Latin* *CORPUS* as referring to the medical domain, and in particular to brain areas (e.g. *CORPUS CALLOSUM*, *CORPUS STRIATUM*) as opposed to large collection of items in a general sense and to large collections of texts/specimens in the Humanities are related only etymologically, with no systematic sense shift or extension. Moreover, the use of *CORPUS* as ‘collection’ is by far more widely-spread than its (neuro)- anatomical sense. If we are not in a position to constrain automatic word search within particular text domains, the medical usage of *CORPUS* is likely to be severely under-represented, flooded by the vast majority of examples of the more generic sense.

In the case of *IMAGING* on the other hand, the prevalent use of the polysemous term in connection with the medical domain, as referred to specific diagnostic technology, is the result of the application of a general-purpose technology to a specific domain. Technical and scientific bibliographic databases present only this collocate, while by selecting it as a seed word in a web engine search, the more frequent collocation is the one referred to generic visual representations. The selected balanced corpora, on the other hand, exhibit both of them, with a higher frequency of occurrences in the medical domain, in particular in the brain sciences.

NETWORK and GRID represent somewhat extreme cases of such domain-sensitive specialisation, to the point that they appear to be overwhelmingly used in their specialised senses only. NETWORK and GRID are expression of the very popular domain of information technology, and even though related to innovation, they are identified as related to this specific context, exhibiting a very coherent collocation behaviour.

Finally, RETENTION and STORAGE appear to oscillate between their proper and extended senses interchangeably, thus witnessing a paradigmatic case of systematic, context-sensitive polysemy.

RETENTION can select both material and immaterial items, but in the bibliographical references system on Grey Literature presents collocations in the domain of information and knowledge retention, as part of the whole process of acquisition, storage and retrieval, whereas in biomedical scientific database is specialised for fluid retention. STORAGE can make reference to containing units, either physical or computational, to long term memory capacity, related to either computational or cognitive processing. Moreover, the storage process is part of the above mentioned knowledge process. Thus, reference to the storage process by itself does not disambiguate the object to be stored.

In figure 1, logarithmic relative frequency distributions across domains are plotted for the test words. Firstly, for each of the selected words, frequencies – expressed as a relative percentage of occurrences in a balanced corpus – are taken and ranked for the diverse domains. Secondly, the top ranked domains are selected and presented together.

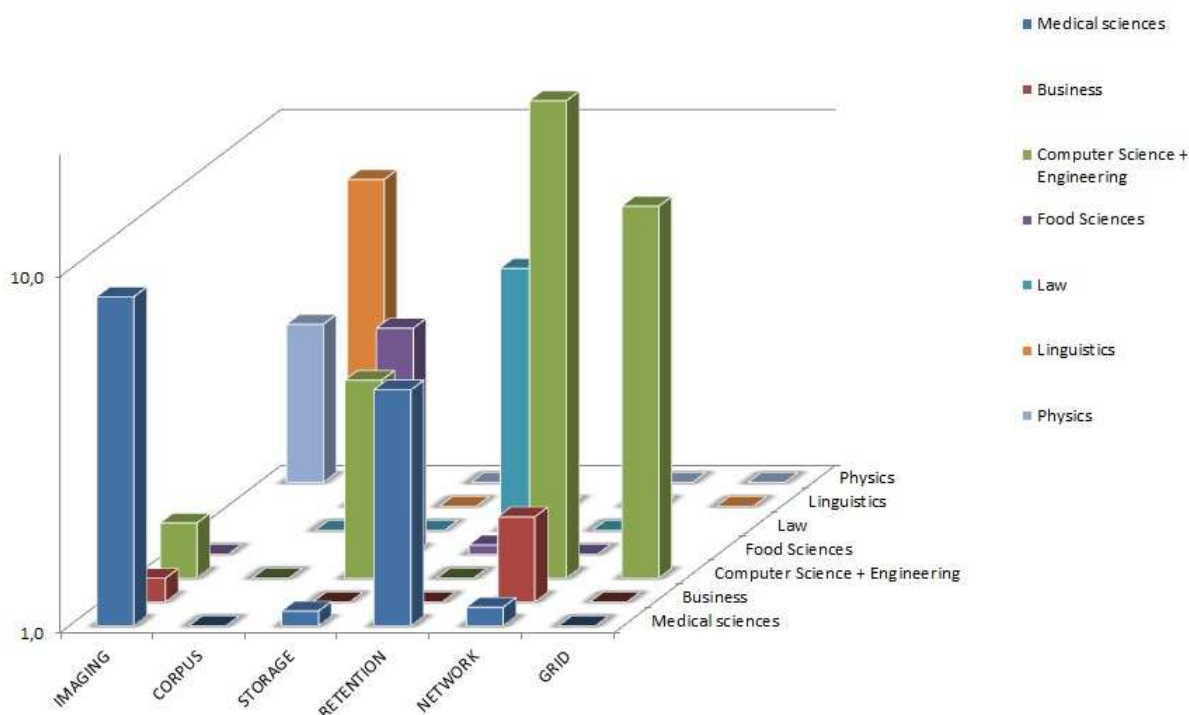


Figure 1- Test words log relative frequency distributions across domains.

Figure 1 provides a snapshot of the usage of our test words across domains, showing that some words are chiefly used in some domains only; however, it does not lay emphasis on the fact that all six terms exhibit multiple senses but relate them differently. The inter-sense relationship can be described in terms of i) AMBIGUITY, when multiple senses show no common schema/relation (as in the case of CORPUS), ii) POLYSEMY, when senses are specialisations/extensions of core meanings (see for example IMAGING); and iii) VAGUENESS, when words can interchangeably be used in both generic and domain-specific contexts with no appreciable sense shift (as with RETENTION and STORAGE). There are words (NETWORK and GRID) which show domain-sensitive specialisation and are chiefly used in their specialised sense.

To further investigate the relationship between different senses, we measured to what extent words which are used both in a specialised context and in a general context tend to co-occur with the same collocates, i.e. tend to be used in similar contexts. This is illustrated in Figure 2, where, for each test word, we counted the number of top-ranked collocates for each of two domains (medical vs. general), and then the number of collocates present in both ranks. As expected, the term CORPUS exhibits the lowest number of common collocates, due to the fact that the two senses of the word share no common meaning core.

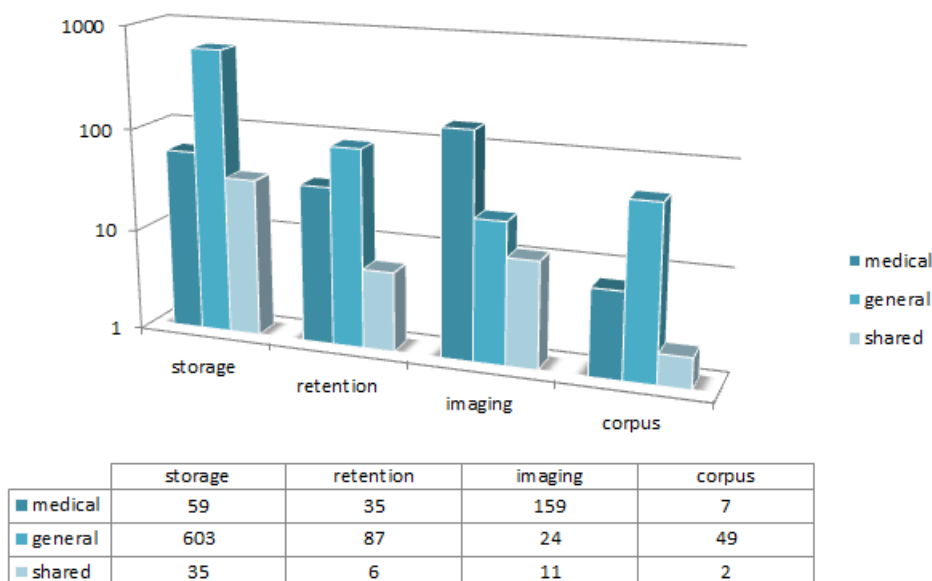


Figure 2 – Medical sciences and general domain shared collocates

It is important to note, however, that words like STORAGE, RETENTION and IMAGING present more overlapping contexts and different degrees of domain specialisation. In particular, IMAGING is by far the most frequent term in its medical usage, even in a non-specialised domain, as shown by the considerable number of collocates that are common to both medical and general domains. Moreover, it is often very difficult to distinguish two different senses of the same term on the basis of the observation of its contextual behaviour only. It turns out that terminological innovation – as expressing conceptual innovation – often implies a gradient extension of a core meaning rather than a radically different usage. We doubt that distributional analyses of contexts can provide a fully automatic basis to tackle sense distinctions at this level of granularity and flexibility. Nonetheless, innovation and domain-driven specialisation can be characterised distributionally in terms of a correlation between typical contexts of usage and knowledge domains.

To sum up, IMAGING, RETENTION, STORAGE, and CORPUS versus NETWORK and GRID are examples of linguistic knowledge in a very specific context-based approach, and therefore examples of phenomena such as ambiguity and polysemy. They are closely related to a representation of innovative knowledge, though still representing an existing gap between the semantic, context and domain dimensions. On the other hand, the relationship between meanings and reality, and lexical semantics are bridged in linguistic data related to concepts of the field of information technology, which is widely known, identified, spread and popular so to enhance a common and coherent understanding.

4. Discussion and concluding remarks

The spread of internet has enhanced the development of better bibliographic scientific databases, such as domain-classified grey material repositories, with a significantly improved capacity for storage, access and retrieval. In recent years, such automated search of grey literature repositories has become the default mode of scientific information retrieval.

As an answer to the need for having better domain-specific databases available for improved storage and retrieval of scientific content, bibliographic domain-specific and type-oriented databases are developed, and emerged over the last two decades to offer materials to a specialised readership, thus providing highly-selected, well-targeted documents.

Such specific multidisciplinary infrastructures define a communication network, where writers and readers – like speakers – are in a sufficient proximity to each other to have very high probability of communication with each other, by sharing the same “language”, intended as the population of utterances in a speech community (Croft, 2000). The informative purposes are defined by the context, in line with Jakobson’s model of communication (1960), where six constituents of the communication act are modelled as functional roles: the addresser or encoder, a message or a signifier, the addressee or decoder, a context or the signified – where by context Jakobson means referent, a code or shared mode of discourse, and a contact or channel.

Analysis of word usages in large multidisciplinary database offers the opportunity to investigate how words and language structures become vehicle for innovative knowledge generation and transfer. Lexical co-occurrences and collocations can be of considerable help in retrieving text materials which are relevant to a specific domain of interest, but they can be of little help in distinguishing one particular

sense of a polysemous word from its other senses. This is especially true of innovative usages of existing terms, which appear to transfer and adapt their original and more general collocates to one or more specialised domains. This is the reason why automated sense disambiguation based on the distributional analysis of words in context proves to be less effective, than disambiguation based on *a priori* subject classifications, especially in distinguishing word usages which are most strongly related to innovation.

Due to ambiguity, polysemy and homography, most terms are multi-referential or multi-contextual (Renouf, 1993) in use. Low-precision results could be improved by restricting the contextual domain. A more supervised approach to the problem, where domains and texts are classified preliminarily by domain experts, rather than being bootstrapped from patterns of word distribution only, promises to be more successful in this task.

We investigated the hypothesis that extension of usage process and polysemous disambiguation correlate significantly with genre- and domain- oriented texts and intended readership, thus providing a convenient way to track down well-targeted, highly technical repositories of openly available text materials.

In requiring a dynamic shift from context-driven vagueness – in term of semantic polymorphism – to domain-driven specialisation – in term of terminological usage, the lexical representation of innovative knowledge in a context-based approach is closely related to the representation of knowledge itself, and represents the opportunity to reduce the gap between knowledge representation and knowledge understanding.

Our main emphasis is laid on the importance of effective information technology repositories and distributed infrastructures for the implementation of knowledge processes, and for an efficiently and widely distributed dissemination of research and innovation results, so to enhance future research.

References

- ANDERSEN H. (1989). Understanding linguistics innovations. In Breivik, L.E., Jahr, E.H. (eds.), *Language Change: Contributions to the Study of its Causes*. Mouton de Gruyter, Berlin. 5-27.
- BAEZA-YATES R., RIBEIRO-NETO B. (1999). *Modern Information Retrieval*. Addison Wesley, ACM Press, New York.
- BIBER D. (1989). A typology of English text. *Linguistics*, 27, 3-43.
- BUITELAAR P., CIMIANO P., MAGNINI B. (2005). *Ontology learning from text*. IOS Press, Amsterdam.
- CHURCH K., HANKS P. (1989). Word Association Norms, Mutual Information and Lexicography. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada.
- CROFT W. (2000). *Explaining Language Change. An Evolutionary Approach*. Longman, London.
- DEUMERT A. (2003). Bringing speakers back in? Epistemological reflections on speaker-oriented explanations of language change. *Language Sciences*, 25, 15-76.
- FELLBAUM C. (2005). Examining the constraint on the benefactive alternation by using the World Wide Web as a corpus. In Kepser S., Reis M. (eds.) *Linguistic evidence: empirical, theoretical, and computational perspectives*. Walter de Gruyter, Berlin. 209-238.
- FLETCHER W.H. (2011). Corpus Analysis of the World Wide Web. In *Encyclopedia of Applied Linguistics*. Wiley-Blackwell.
- FRANTZI K. T., ANANIADOU S., MIMA H. (2000). Automatic Recognition of Multi-Word Terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), 115-130.
- HEDLUND G. (1999). The Multinational Corporation as a Neatly Recomposable System. *Management International Review* 39,1. 5-44.
- JAKOBSON R. (1960). Linguistics and Poetics: Closing Statement. *Style in Language*. MIT Press, Cambridge, MA.
- KILGARRIFF A., RYCHLY P., SMRZ P., TUGWELL D. (2004) The Sketch Engine. *Proceedings EURALEX 2004*, LORIENT, FRANCE. 105-116.
- LINTON J. (2009). De-babelizing the language of innovation. *Technovation*, 29, 729-737.
- LÜDELING A., EVERT S., BARONI M. (2006). Using web data for linguistic purposes. In Hundt M., Nesselhauf N., Biewer C. (eds.), *Language and Computers, Corpus Linguistics and the Web* 18, 7-24.
- MANNING C. D., SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- MARZI C. (2012). Knowledge Communities in Grey. *Proceedings of the GL13 – International conference on Grey Literature: the Grey Circuit*. TextRelease, Amsterdam. 34-40.
- MEYER C.F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge University Press, Cambridge.
- NOOTEBOOM B. (2000). *Learning and innovation in organizations and economics*. Oxford University Press, Oxford.
- PANGARO P. (2008). Innovation, Language, and Organizations. *Continuum Itaú Cultural magazine*. 7.
- SERETAN, V. (2011). *Syntax-Based Collocation Extraction*. Springer, Heidelberg-London-New York.
- SINCLAIR J. (1991). *Corpus, concordance, collocation: Describing English language*. Oxford University Press, Oxford.
- SMADJA, F. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics*, MIT, Cambridge MA, USA, 19(1), 143-177.
- STUBBS, M. (2001). On inference theories and code theories: Corpus evidence for semantic schemas. *Text*, 21 (3), 437-465.
- RENOUF, A. (1993). What the Linguist has to say to the Information Scientist. In Forbes, G. (ed.) *The Journal of Document and Text Management* vol. 1/2, 173-190.
- VETERE G., OLTRAMARI A., CHIARI I., JEZEK E., VIEU L., ZANZOTTO F.M. (2011). Senso Comune, an Open Knowledge Base for Italian. *TAL*, 52 (3), 217-243.
- WELCH D., WELCH L. (2008). The importance of Language in International Knowledge Transfer. *MIR*, 48(3), 339-360.
- THE OPEN GREY <http://www.opengrey.eu/>
- THE SKETCH ENGINE <http://www.sketchengine.co.uk>

Grey in the Innovation Process *

Keith G. Jeffery (United Kingdom) and Anne Asserson (Norway)

Abstract

The research lifecycle has multiple objectives materialised as outputs, outcomes and impacts. Typical outputs are research publications (including grey literature), patents and products such as research datasets and software, many kinds of art or prototype engineering artifacts. Outcomes include patent licence income, value of a company set up to exploit the output or trained research staff. Impacts include employment creation, a commercial product that saves lives or labour or development of a new field of knowledge and research such as genomics since the 1950s.

Commonly research in progress may be documented as grey literature – such as technical reports, laboratory notebooks or instructions for operating new equipment. There is a decision point when grey literature is produced.

One can innovate academically. The output is peer reviewed publications; the outcomes include developing trained researchers; the impact leading to a new field of research. This route provides academic recognition.

Alternatively one can innovate along the wealth-creation route. The output could be a patent; the outcome license income or a new company; the impact employment, dividends to shareholders or a new 'wonder product'. This route provides wealth and possibly improvement in the quality of life.

If research is published this usually precludes following the wealth-creation route since the novel idea is now in the public domain and not protected by patent(s).

Increasingly research funding and research performing organisations wish to demonstrate that the research they fund or do leads to impacts of relevance to society. Tracing of impacts back to the original research is not easy, partly because the eventual impact may not be known for many years. The key is an accurate recording of the research lifecycle including important dates so that the innovation cycle from idea to impact and back to further ideas can be demonstrated.

Recent work – especially in UK in the JISC-funded MICE project – has produced a taxonomy of outputs, outcomes and impacts. In parallel an extension to CERIF (Common European Research Information Format – an EU recommendation to member states) has been developed and approved by euroCRIS. This extension re-uses typical CERIF entities of persons, projects, organisations, publications, patents, products but relates them (with temporal validity and appropriate role) to the production or utilisation of outputs, outcomes and impacts. Naturally grey literature is a key component within this model.

1. Innovation

Innovation may be defined as the development of new customers' value through solutions that meet new needs, inarticulate needs, or old customer and market needs in new ways (Wikipedia). There are essentially three kinds of innovation:

1. Academic innovation: leading to new research techniques and results and providing also trained researchers;
2. Commercial innovation: using the results of research to create wealth via patents or products taken up by industry to provide a commercial product or service that generates the wealth;
3. Societal innovation: using the results of research for improvement in the quality of life in environmental, health, cultural or social aspects.

Innovation is a process with a time dimension and along that time dimension are produced outputs, outcomes and impact.

The grey literature output of research sits within a much broader context of other outputs (such as peer-reviewed publications, patents, products). Nonetheless, as argued in (Jeffery and Asserson 2007) grey literature objects are a very important output since they may lead to outcomes and impact. The outputs are not the innovation: this is achieved by utilising the outputs to produce outcomes and impact.

Outcomes are activities derived directly from the outputs: from research publications or grey literature one might have as outputs trained researchers and new research techniques. From patents (themselves considered grey literature) an outcome could be license income and possibly consultancy work (income) for the researchers to assist exploitation of the patent. From products an outcome could be the setting

* First published in the GL14 Conference Proceedings, February 2013.

up of a spin-out company which employs people and produces products or services. Interestingly, many of the outcomes of research are documented by grey literature since patents, company technical reports, government internal (or external) reports are generated and these are not peer reviewed 'white literature'.

Impact is difficult to define but it is the effect the research has on society. Commonly impact is detected many years (commonly 10-15) after the research is completed. Some examples may illustrate:

1. An impact may be millions of lives saved due to a drug made available after extensive trialling, produced by a pharmaceutical company after further in house development and based on an output of research at a university.
2. Impact may be due to a policy change – possibly enforced by law – based on research. An example is the reduced deaths from lung cancer in Western society because of laws based on policies derived from research on the effect of tobacco on human health.
3. Similar examples exist in the environmental domain where research has led – in time – to policies concerning the provision of clean water supplies.

Of course in some of these cases there is associated wealth creation (e.g. for the pharmaceutical company in the first example) and associated provision of employment and hence further wealth creation.

2. BACKGROUND

2.1 Previous Work

For more than two decades, the authors have worked on research information in the widest sense comprising information not only about grey literature (grey objects) but also all the outputs of research (products, patents, publications) and the context within which the research was done including projects, organizations, funding, persons, facilities, equipment, events. Within the GL community we have highlighted the issues as we see them:

1. the need for formal metadata to allow machine understanding and therefore scalable operations (Jeffery 1999);
2. the enhancement of repositories of grey (and other) e-publications by linking with CRIS (Current Research Information Systems) (Jeffery and Asserson 2004);
3. the use of the research process to collect metadata incrementally reducing the threshold barrier for end-users and improving quality in an ambient GRIDs environment (Jeffery and Asserson 2005);
4. an architectural model for scalable, highly distributed, workflowed repositories of grey literature based on hyperactive 'intelligent' documents (Jeffery and Asserson 2006).
5. A 'from 10,000 metres altitude' view of the grey information landscape 'Greyscape' based on the hypothesis that grey literature is the foundation for the knowledge economy (Jeffery and Asserson 2007).
6. An analysis of interoperation architectures among research information systems 'INTEREST' (Jeffery and Asserson 2008).
7. A proposal that Grey Literature should be seen within the context of e-Science supported by a CERIF-CRIS (Jeffery and Asserson 2009).
8. A proposed architecture 'GLASS' using CERIF metadata to demonstrate transparency in the Grey process (Jeffery and Asserson 2010).

Although this corpus of work demonstrates how CERIF provides the required context for processing grey objects, in particular the 'Greyscape' paper (Jeffery and Asserson 2007) related Grey Literature described by CERIF metadata to the knowledge economy and is thus a relevant piece of previous work related to the current topic: Innovation.

2.2 The Requirement

The requirement is to record the innovation derived from research and in the context of this community research recorded as grey literature (or grey objects) meaning outputs not formally peer-reviewed. The innovation is recorded as outcomes and impact. However, the outputs, outcomes and impact need to be related back to the research project, the persons involved, the organisations involved (e.g. university/faculty / department / research group or maybe a commercial company), the funding and funding organisation (including where appropriate research programme and topic), the facilities and equipment used in the research.

3. The Hypothesis

We assert that a solution – CERIF – exists already which covers these requirements. CERIF has already been in use widely in 42 countries for recording research activity and is an EU Recommendation to Member States. CERIF is maintained, developed and promoted by euroCRIS (www.eurocris.org) at the request of the European Commission. In particular, work done during the MICE (JISC-funded) project in UK extended the CERIF datamodel (which included already outputs) to include indicators and measurements which can record outcomes and impact (and aggregated outputs) (Gartner, Cox, Jeffery 2012). This proposed extension to CERIF was ratified through the euroCRIS process for inclusion in the CERIF datamodel, and provides a way to record unambiguously and in context outcomes and impact derived from research outputs – including grey literature.

4. Proposed Architecture

4.1 Introduction

The CERIF datamodel is already quite well-known in the Grey Literature Community but the overall model is reproduced here (Figure 1) to illustrate the entities that are recorded together with their relationships thus giving the context of the research.

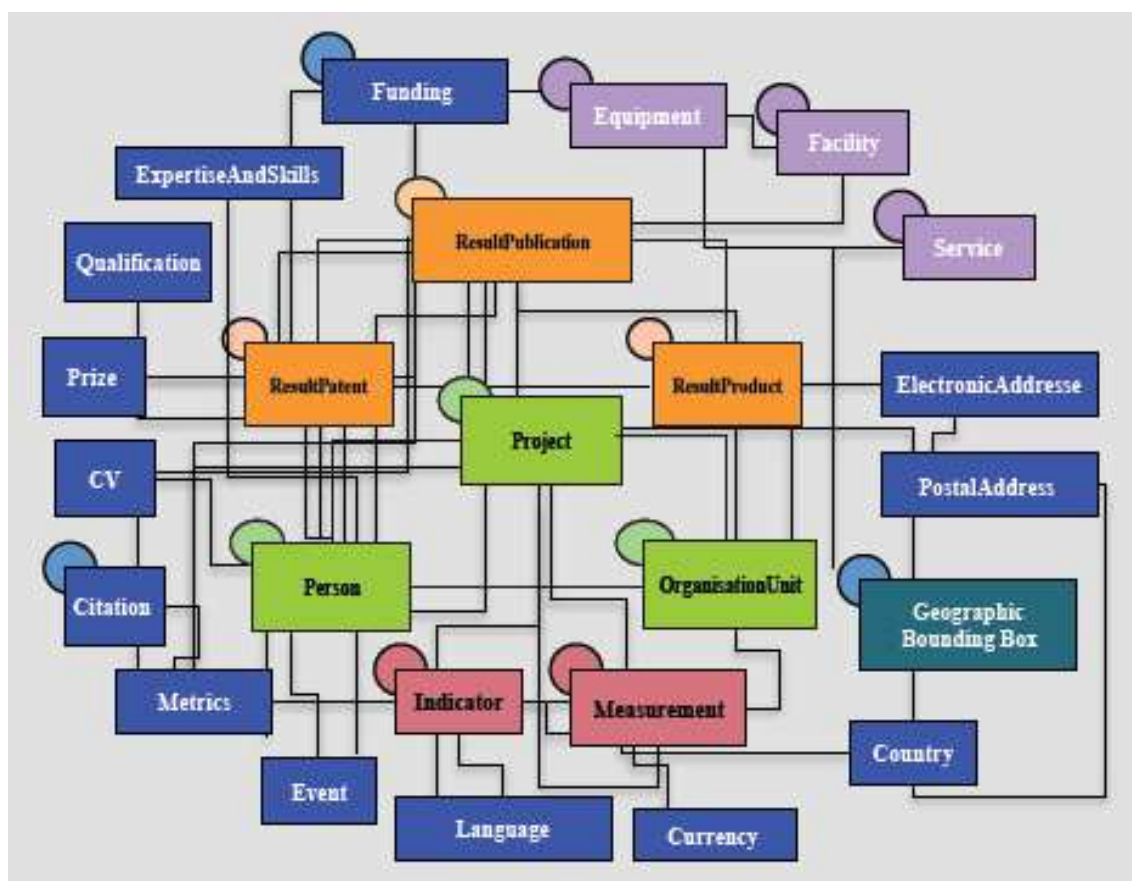


Figure 1: The CERIF Datamodel

The new entities for managing indicators and their measurements for the purposes of innovation (outputs, outcomes and impact) are indicated in the diagram above but dealt with in more detail below.

4.2 Indicators and Measurements for Outcomes and Impacts

The part of the datamodel concerning indicators and impacts is reproduced in detail here (Figure 2) and in particular it should be noted how the entities representing indicators and measurements relate to the base entities of CERIF (such as publication, patent, product and person, organisational unit, project etc) through the semantically rich temporally bound linking relations.

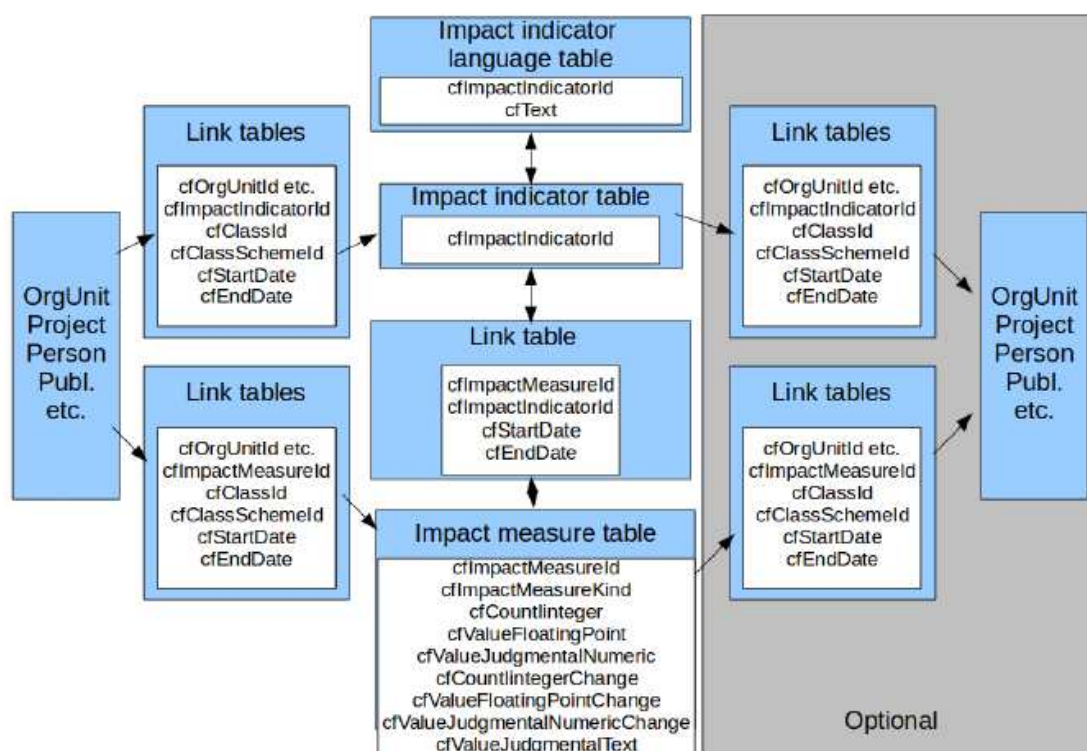


Figure 2: The MICE Datamodel within CERIF

It should be noted that the right side of the diagram (i.e. the instances of entities affected by or benefitting from the outcome or impact) is optional. Furthermore the instances of entities on the right hand side as e.g. beneficiaries of outcomes may appear on the left side as initiators of the transition from outcomes to impacts.

For each indicator there are one or more measurements. Not all measurements are relevant for every indicator; in fact usually only one measurement is appropriate for an indicator. The measurements include an integer count (e.g. how many publications were produced by a person or group in a given time period or how many lives were saved by a new drug); a floating point measurement (e.g. amount of licence income for a patent) and a judgement expressed numerically (e.g. quality on a scale of 1-10). In addition there are the 'delta' measures which record change and compare the value for one period of time with another. Examples would include the increased number of publications, the increased licence income or the improvement in quality. Finally an attribute is made available for a textual statement on judgement to justify the measure for an indicator or to express less precisely the estimated quality.

The MICE project produced a detailed taxonomy of indicators (Gartner, Cox, Jeffery 2012) that could be used, but there are others in the scientometrics and bibliometrics fields. CERIF can, of course, allow the use of any scheme of indicators due to its flexible semantic layer feature.

Another UK project, Snowball, (SnowballProject) has produced a set of indicators for university benchmarking. However, the indicators are dominantly to record performance and less to record outcomes and impact. The 'Snowball Recipe Book' (produced by Elsevier which was a project partner) (SnowballRecipes) was launched at a recent euroCRIS Members' Meeting.

The recently initiated Indicators Task Group of euroCRIS is exploring and researching the available techniques and intends to produce a canonical set of indicators and associated measurements that can be used for benchmarking and comparison across outputs, outcomes and impact. This will provide the basis for measuring innovation and especially innovation in the Grey Process.

5. Conclusion

From the above we may conclude:

1. CERIF provides an appropriate data structure for recording innovation, including within the GREY process;
2. It is being used in significant systems tracking outputs, outcomes, impact related to contextual, temporal, geospatial metadata;
3. euroCRIS has an Indicators Task Group dealing with new scientometrics (including bibliometrics) and new methods for detecting impact (backward chaining).

References

(CERIF) www.eurocris.org/cerif

(Gartner, Cox, Jeffery 2012) Richard Gartner, Mark Cox, Keith Jeffery 'A CERIF-based schema for encoding research impact' The Electronic Library xxx.yyy 2012 Emerald Publishing (in press)

(Jeffery 1999) Jeffery, K G: 'An Architecture for Grey Literature in a R and D Context' Proceedings GL'99 (Grey Literature) Conference Washington DC October 1999 <http://www.konbib.nl/grey-net/frame4.htm>

(Jeffery 2004b) Jeffery, K.G.; 'The New Technologies: can CRISs Benefit' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 77-88 (available under www.eurocris.org)

(Jeffery and Asserson 2004) K G Jeffery, A G S Asserson; Relating Intellectual Property Products to the Corporate Context; Proceedings Grey Literature 6 Conference, New York, December 2004; TextRelease; ISBN 90-77484-03-5

(Jeffery and Asserson 2005) K G Jeffery, A G S Asserson 'Grey in the R&D Process'; Proceedings Grey Literature 7 Conference, Nancy, December 2005; TextRelease; ISBN 90-77484-06-X ISSN 1386-2316

(Jeffery and Asserson 2006c) Keith G Jeffery, Anne Asserson: 'Hyperactive Grey Objects' Proceedings Grey Literature 8 Conference (GL8), New Orleans, December 2006; TextRelease; ISBN 90-77484-08-6. ISSN 1386-2316 ; No. 8-06-X

(Jeffery and Asserson 2007) Keith G Jeffery, Anne Asserson: 'Greyscape' Opening Paper in Proceedings Grey Literature 9 Conference Antwerp (GL9) 10-11 December 2007 pp9-14; Textrelease, Amsterdam; ISSN 1386-2316

(Jeffery and Asserson 2008) Keith G Jeffery, Anne Asserson: 'INTEREST' Proceedings Grey Literature Conference Amsterdam 8-9 December 2008 in Tenth International Conference on Grey Literature : Designing the Grey Grid for Information Society, 8-9 December 2008, Science Park Amsterdam, The Netherlands ed. by Dominic J. Farace and Jerry Frantzen ; GreyNet, Grey Literature Network Service. - Amsterdam : TextRelease, February 2009. GL-Conference series, ISSN 1386-2316; No. 10. - ISBN 978-90-77484-11-1.

(Jeffery and Asserson 2010) Keith G Jeffery, Anne Asserson 'GL Transparency: Through a Glass, Clearly' Proceedings Twelfth International Conference on Grey Literature, Prague December 2010 Ed Dominic Farace and Jerry Frantzen 95-100, TextRelease Amsterdam Series ISSN 1386-2316; no 12; ISBN 978-90-77484-16-6

(SnowballProject) <http://www.snowballmetrics.com/>

(SnowballRecipes) <http://www.snowballmetrics.com/wp-content/uploads/Snowball-Metrics-Recipe-Book.pdf>

(Wikipedia) www.wikipedia.org

Data Analytics: The next big thing in information *

June Crowe and J.R. Candlish (United States)

Abstract

Information is now available in an overabundance, so much so, that distinguishing the noise from the signal has become very problematic. In the past, the collection and storage of information was the primary issue. Currently, there are massive amounts of data both structured and unstructured, that need to be analyzed in an iterative, as well as in a time sensitive manner. In response to this need, data analytical tools and services have emerged as a means to solve this problem.

Grey literature repositories, libraries, and information centers are well positioned to take advantage of these new tools and services. The current trend is to make grey literature more easily discoverable, accessible, and with the new data analytical tools and services, more easily analyzed.

The intent of our survey of the Grey Literature community was to provide a snapshot of the Community's use, planned use, and knowledge of data analytical tools/services for big data as it affects grey literature. The survey summary that follows indicates where the Community currently stands in regards to the use of data analytical tools and services.

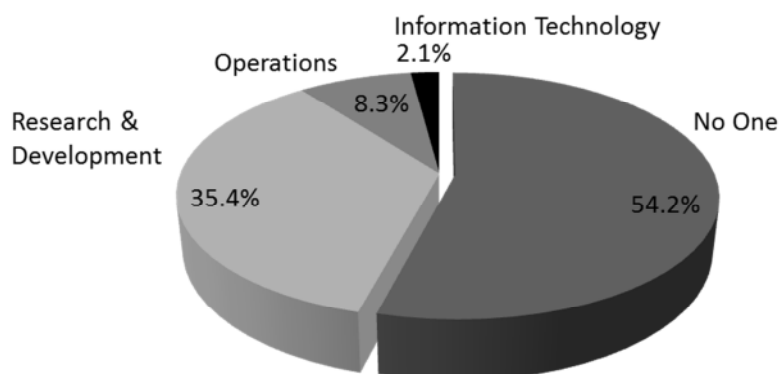
Represented Industries

From September 13 through October 31, 2012, an online survey was conducted and made available through two internet vehicles from: (1) the GreyNet Group on LinkedIn®, and (2) the GreyNet listserv. Forty eight responses scattered across South America (1), New Zealand (1), Africa (2), Asia (4), Australia (8), Europe (16) and North America (16), yielded insight into the Grey Literature Community's knowledge of the Big Data construct. Overall, academia represented 50% of responses, with government and private industry composing the remaining half. Within these industries, nearly 42% of the respondents were at the staff level, indicating that there's great understanding of the Big Data landscape, especially in academia.

The Current Landscape

The current landscape of Big Data products and services revealed several key significant points. First and foremost, a large majority (73%) of respondents indicated that their organization does not currently use Big Data products and services. This situation is partly due to the lack of drivers/champions to adopt them (>54%). However, in the area of Research and Development, there proved to be a significant contributing driver (35%) among the survey population that did denote the current existence and utilization of Big Data discovery and analytical tools (Figure 1).

Figure 1: Survey responses of current drivers for the adoption of Big Data services/products. (N=48)



Since such a large percentage of the Community has not adopted Big Data capabilities, it was not surprising to see that the majority (74%) indicated a novice expertise level. Moreover, 80% of the respondents had not seen any data analytical products/services demonstrated but were planning to use such products for web analytics, predictive analytics, and real-time analytics. For those who were familiar with existing analytical tools, SAP, SAS, and Google BigQuery were among the most popular. The survey question concerning the near future impact of big data analytical platforms, databases, services, and data analytical tools on grey literature (some impact-27%, moderate impact-19%, high impact-33%)

* First published in the GL14 Conference Proceedings, February 2013.

clearly shows that the Grey Literature Community is expecting these services/products to provide some solution to the problem of Big Data.

Importance of Big Data

Big data is important primarily because it is growing at an exponential rate. Over five exabytes is created every two days. The problem with Big Data is not just data analysis, but with discovering, harvesting, curating, storing and its management. Steve Pederson, CEO of BrightPlanet Corporation, stated that 90% of Big Data content lies in the expanding universe of unstructured content; the vast majority of that information is hidden and unknown in the Deep Web segment of the Internet (Pederson, 2012). Not surprisingly, much of grey literature is found in the Deep Web.

The Grey Literature community strongly felt that Big Data will be a huge positive for society just as it will be for science (56%). George Strawn of the Networking and Information Technology Research and Development (NITRD) Program identified four trends in Big Data in science and business:

- (1) bigger data;
- (2) increase in unstructured data;
- (3) increase in distributed data; and
- (4) increase in distributed computing.

These trends will spawn new tools and services for data sharing and collaboration, for data analytics, and for the management of data (Strawn, 2012).

Mobile devices are quickly becoming a primary means of accessing data. However, the survey results indicated that less than half of the respondents responded that it was only somewhat important (36%), on a 1-5 scale, as a method of accessing Big Data results. Figure 2 outlines levels of importance on a Likert scale of 1-5.

Figure 2: Likert scale based on levels of importance.

1	2	3	4	5
Not Important at All	Indifferent	Somewhat Important	Moderately Important	Very Important
19.4%	19.4%	36.1%	19.4%	5.7%

This result indicated that many in the Community do not see the mobile phone as a key tool for accessing data. This was another surprise because Juniper Research is predicting a greater demand in 2013 for data analytics solutions across mobile devices (Juniper Research, 2013).

Barriers and Potential Concerns to Adopting Big Data Products/Services

The two primary barriers to adopting big data services/products are the lack of skilled personnel and/or the lack of sufficient resources. In the Grey Literature community, the lack of sufficient resources proved to be the greatest barrier at 45.7%, with the lack of skilled personnel being a secondary barrier at 33.3%. Currently there is a shortage of personnel skilled in data curation, data integration and re-use, as well as data analysts (Schindel, 2012).

The barriers could conceivably be overcome if there were drivers/champions to lead the cause of adopting these technologies/services, and to clearly identify the return on investments to management. Dennis Gannon, Director of Cloud Research Strategy pointed out that every area of science is now engaged in data-intensive research. Thus, the need for these technologies, products and services are not going away but will continue to increase (Gannon, 2012). As the volume of digital data continues to grow exponentially, there will be a continued need for skilled personnel and a need for adequate financial support. Both will be challenges the Grey Literature Community face in order to reap the benefits of big data analytical products and services.

Biased reporting has proved to be a leading issue to consider as well. Over 90% of respondents were concerned that biased reporting will be a cause of concern across multiple sectors (economic, political, health, scientific, social, etc.). From an historical perspective, an abundant amount of evidence supports this concern. For example, outcome biased reporting within the medical community has been a legitimate public concern when newly developed pharmaceuticals are trial tested then submitted to the Food and Drug Administration (FDA) for approval via New Drug Application (NDA) (Rising et al., 2008).

Big Data Goals

Results from the survey ranked the overall goals of utilizing Big Data products and services. The Grey Literature community favored data discovery (47.2%) with data mining analytics (44.4%) and data

visualization (38.9%). The majority of responses emphasized the importance of these three categories by denoting a five on a 1-5 scale (Table 1).

Table 1: Survey responses on a 1-5 scale (1 being least important, 5 being most important) of which Big Data products/services would be most relevant to your organization's data goals. (N=36)

	1	2	3	4	5
Data discovery	5.6%	8.3%	8.3%	30.6%	47.2%
Data mining analytics	2.8%	2.8%	27.8%	22.2%	44.4%
Data visualization	2.8%	8.3%	19.4%	30.6%	38.9%

Summary of Survey and Future Considerations

Of the 48 survey takers, a trend in the responses revealed that respondents became increasingly impatient/distracted as the survey progressed. In the first third of the survey, responses were nearly 100% participation. Responses decreased by an average of 10 throughout the second third and plummeted by nearly 20 according to the last third of the survey. This indicated that the survey should have been shorter and the options for answers more limited. The survey administrators will take these findings into consideration for future surveys.

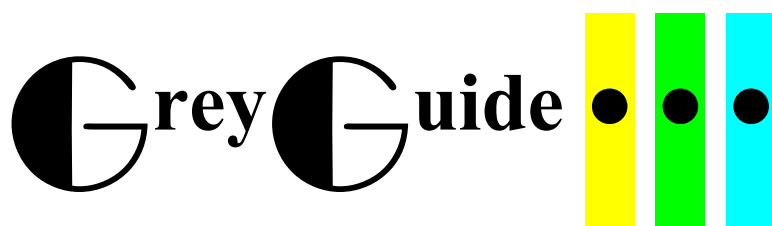
Overall, the Grey Literature respondents are keenly aware of the benefits of using Big Data services and products but have yet to identify people within their organizations as drivers/champions to make it reality. The lack of identified backers who can clearly and consistently make the case to show the immediate benefits and ultimate return on investments in these technologies and services to management has impacted their adoption or hindered their implementation. As indicated from the survey, the Grey Literature community is not using these products in substantial numbers nor have they seen these products/services demonstrated. Yet, the Community sees great value in these products/services for their local economy (>68% of survey takers), and they are planning to use these tools for web analytics, predictive analytics, and real-time analytics. Additionally, if the Community could select big data products/services for common data goals, they would select them first of all for data discovery and then for data mining analytics. Lastly, the lack of adequate financial resources is the greatest barrier to adopting these products/services.

In terms of future considerations, re-distributing the survey in three to five years may yield interesting responses as Big Data initiatives are readily explored. Additionally, as Big Data products and services mature, a better understanding of the developing landscape may reveal insight into trends that cannot yet be foreseen.

References

- Gannon, D. (2012), Science as a service: Data Analytics and Data Mining - The Approaching Tidal Wave. – In: Proceedings of a CENDI/NFAIS/FEDLINK conference held the Library of Congress, Washington DC, Dec. 11, 2012 (http://cendi.gov/presentations/12_11_12_Gannon_Data_Analytics.pdf)
- Juniper Research. 2012. Juniper Research's Top 10 Mobile Trends for 2013. (http://www.juniperresearch.com/shop/download_whitepaper.php?whitepaper=198)
- Pederson, S (2012), Exploiting Big Data from the Deep Web: The new frontier for creating intelligence. BrightPlanet, Sioux Falls, South Dakota. White paper available (<http://www.brightplanet.com/2012/07/creating-intelligence-from-big-data-whitepaper/>)
- Schindel, D.E. (2012), Data Curation: Skill-sets and Workforce Needs. – In: Proceedings of a CENDI/NFAIS/FEDLINK conference held the Library of Congress, Washington DC, Dec. 11, 2012 (http://cendi.gov/presentations/12_11_12_Schindel_Workforce_Needs.pdf)
- Strawn, G.O. (2012), Big Data. – In: Proceedings of a CENDI/NFAIS/FEDLINK conference held the Library of Congress, Washington DC, Dec. 11, 2012 (http://cendi.gov/presentations/12_11_12_Strawn_Big_Data_Overview.pdf)
- Rising K., P. Bacchetti, and L. Bero (2008), Reporting Bias in Drug Trials Submitted to the Food and Drug Administration: Review of Publication and Presentation. PLoS Med 5(11): e217. – doi:10.1371/journal.pmed.0050217

On the News Front



Guide to Good Practice in Grey Literature

[Home](#) [About](#) [Topics](#) [Record Entry](#) [Search](#) [Contact](#) [Disclaimer](#)

Guide to Good Practice Repository

GreyGuide offers information professionals, practitioners, and students common ground for good practice in grey literature

Welcome to the GreyGuide, an online forum and repository of good practice in the field of grey literature. The GreyGuide seeks to capture proposed as well as published practices dealing with the supply and demand sides of grey literature. This is a collaborative project involving GreyNet International and ISTI-CNR Pisa, Italy.

The test-launch of the GreyGuide Repository took place during a summer workshop at the Library of Congress in Washington D.C. on July 25, 2013. The official launch of the GreyGuide Repository will take place on December 2, 2013 during the Fifteenth International Conference on Grey Literature in Bratislava, Slovak Republic. From now until then, the acquisition of both proposed and published good practices is underway. Your content contribution is greatly appreciated.

The GreyGuide is a community driven open resource repository. Your proposed good practice is considered an important contribution to the field of grey literature. In order to facilitate record entry a number of embedded lists and help screens are included throughout the process.

Dr. Stefania Biagioni, Repository Manager/Administrator
Institute of Information Science and Technologies (ISTI-CNR)
Via Giuseppe Moruzzi 1, 56124 Pisa - Italy
Tel. +39 050 3152900 - Fax +39 050 3152810
stefania.biagioni@isti.cnr.it

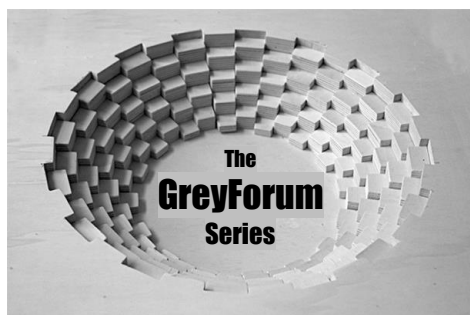
Dr. Dominic Farace, Project Manager
GreyNet International, Grey Literature Network Service
Javastraat 194-HS, 1095 CP Amsterdam - Netherlands
Tel/Fax +31(0)20-331.2420
info@grey.net.org

The **GreyGuide** is an online **Forum and Repository** for both proposed as well as published good practices in the field of grey literature. **Join in the Pre-Launch** of the GreyGuide that is now underway and share the benefits of your experience with others in the grey literature community!

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

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

Grey Literature and Information Rights: On Law and Practice



Artwork by Jeroen Hoogstraten

GreyForum 2.1

Workshop

Grey Literature and Information Rights: On Law and Practice

Date

Thursday, 14 November 2013

Time

10:15 – 15:00 hours

Location

NBC Congrescentrum Nieuwegein
Blokhoef 1
3438 LC Nieuwegein
Netherlands

Workshop Anchors

Prof. Tomas Lipinski, Kent State, USA
Dr. Joachim Schöpfel, U. of Lille, FR
Mr. Rene Dales, EBSCO, BeNeLux
Dr. Dominic Farace, GreyNet, NL

Online Registration

Attendance at the GreyForum Workshop is included in the KNVI Conference Fee. To register simply click, [Online KNVI Registration](http://www.knvi.nl/registration)

Workshop Contact

GreyNet International
Grey Literature Network Service
Javastraat 194-HS, 1095 CP
Amsterdam, Netherlands
info@greynet.org
greynet.org/greyforumseries.html
Phone/Fax +31 (0) 20 331 2420

This workshop is organized together with colleagues from Kent State University (USA), the University of Lille (France), and EBSCO (BeNeLux). It will take place as part of the KNVI (Royal Netherlands Association for Information Professionals) Annual Conference on November 14, 2013

Grey Literature and Information Rights: On Law and Practice

Within grey literature resides a wide range of knowledge based resources often unexploited and underestimated. A clearer understanding of the rights associated with this field of information will enable stakeholders to better understand the relevance and value of grey literature. This workshop will further guide information professionals in their tasks of informing policy and practice in the workplace.



Prof. Tomas Lipinski, Director of the School of Library and Information Sciences at Kent State University (Ohio, USA) will be joined by Dr. Joachim Schöpfel, Senior lecturer in library and information sciences at the University of Lille and Director of ANRT, National Digitization Centre for PhD Theses (France) in presenting on information rights. Rene Dales of EBSCO will demonstrate the benefits of licensing grey literature and Dominic Farace, Director of GreyNet International will engage the workshop participants in the pre-launch of the GreyGuide Repository, an online guide to good practice in grey literature.

Grey Literature and Information Rights is geared to publishers, librarians, archivists, and other information professionals in government, academics, business and industry who are involved in the production and/or management of information resources.

Grounds for Concern • Fair use • Copyright • Intellectual Property • Public Domain • Open Access • Open Data • Licensing • Creative Commons • Digitization • Orphan Works • Preservation • Legislation • Legal Deposit • Accreditation • Plagiarism • Disclaimer • Etcetera

GreyForum Workshop and KNVI Conference Program

09:15 KNVI Registration Desk Open	13:30 Workshop Case Studies
10:15 Workshop Introduction	14:15 Licensing, Open Access
10:30 Basics on Copyright Law	15:00 Close Workshop, Tea
11:15 Intellectual Property	15:30 KNVI Keynotes, etc.
12:00 Lunch	17:00 Conference Reception

More on GreyForum, <http://www.greynet.org/greyforumseries.html>



Fifteenth International Conference on Grey Literature

The Grey Audit: A Field Assessment in Grey Literature

Slovak Centre of Scientific and Technical Information

CVTI SR, Bratislava, Slovak Republic, 2-3 December 2013



Conference Outline and Schedule

DAY 1 Monday	DAY 2 Tuesday
<i>Location: Lamačská cesta 8/A, 811 04 Bratislava</i>	<i>Location: Lamačská cesta 8/A, 811 04 Bratislava</i>
Registration Desk Open 8:15-9:00 Coffee and Tea Service	Registration Desk Open 8:30-9:00 Coffee and Tea Service
Opening Session 9:00-10:30 <i>Welcome Address, Opening Paper, and Keynote Address</i>	9:00-11:00 <i>Poster Session and Sponsor Showcase</i>
Morning Break 10:30-11:00	
Session One 11:00-12:45 <i>"Technology Assessment"</i>	Session Three 11:00-12:45 <i>"Research and Data"</i>
Lunch 12:45-13:45	Lunch 12:45-13:45
Session Two 13:45-15:30 <i>"Sustaining Good Practices"</i>	Session Four 13:45-15:30 <i>"Towards Informed Policies"</i>
Afternoon Break 15:30 -16:00	Pause 15:30-15:45
Introduction to Posters 16:00-17:00 <i>Briefings by Poster Presenters</i>	Closing Session 15:45-16:15 <i>Report by Conference Chair and Moderators, Conference Handoff, and Farewell</i>
Conference Reception 17:00-18:30	Post-Conference Tour CVTI SR 16:15-17:30

Travel Tip! Bratislava, Slovakia and Vienna, Austria are Europe's Twin Capital Cities only 56 km/35 miles apart.

Fifteenth International Conference on Grey Literature

The Grey Audit: A Field Assessment in Grey Literature

Slovak Centre of Scientific and Technical Information

CVTI SR, Bratislava, Slovak Republic, 2-3 December 2013



Conference Registration Form

The Conference Registration Fee includes attendance during the Sessions, a copy of the GL15 Program and Abstract Book, the full-text of the Conference Papers and PowerPoint Slides, as well as the conference pouch and participant badge. Also included are lunches, refreshments during the morning and afternoon breaks, and the Inaugural Reception.

Payment after October 1st 2013 add a €75 surcharge

☐ € 495 Participant Fee ☐ € 395 Author Fee ☐ € 250 Day Pass ☐ € 100 Student Fee



I am employed with **CVTISR** and entitled to a 20% reduction on the conference fee: ☐

I am a *GreyNet* Member and am entitled to a 20% reduction on the conference fee: ☐

Registrant's Name: _____

Organization: _____

Address/P.O. Box: _____

Postal Code/City/Country: _____

Tel/Fax/Email: _____

Please Check One of the Boxes below for the Method of Payment:

☐ Direct bank transfer to TextRelease, Account 3135.85.342 - Rabobank, Amsterdam, Netherlands
BIC: RABONL2U **IBAN:** NL70 RABO 0313 5853 42, with reference to GL15 and Registrant's Name

☐ MasterCard ☐ Visa Card ☐ American Express
Card No. _____ Expiration Date: _____

If the name on the credit card is not that of the registrant, print the name that appears on the card, below

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

Place _____ Date _____

Note: Credit Card transactions can be authorized by Phone, Fax, or Postal services. Email is not authorized.

GL15 Program and Conference Bureau

TextRelease

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

Author and Title Index 2013

No. and Page	A
No. 1, p. 24	Assante, Massimiliano ; Candela, Leonardo; Pagano, Pasquale An Environment Supporting the Production of Live Research Objects
No 3. 152	Asserson, Anne see Jeffery, Keith G.
No. 1, p. 7	Assessing a Subject-based Blog for Current Awareness within a Cancer Care Environment / Yongtao Lin and Marcus Vaska
	B
No. 1, p. 40	Bardi, Alessia ; La Bruzzo, Sandro; Manghi, Paolo Customized OAI-ORE and OAI-PMH Exports of Compound Objects for the Fedora Repository
No 2, 69	Biagioni, Stefania see Luzi, Daniela
No 3. 122	Boukacem-Zeghmouri, Chérifa Grey literature in the digital culture and practices of the global scholar: the case of molecular biology
	C
No. 1, p. 24	Candela, Leonardo see Assante, Massimiliano
No 3. 157	Candlish, J.R. see Crowe, June
No. 1, p. 19	A Centralised National Corpus of Electronic Theses and Dissertations / Július Kravjar and Marta Dušková
No 2, 97	Chalabi, Lydia Characteristics and use of grey literature in scientific journals articles by Algerian researchers: Case study of the University of Science and Technology 'Houari Boumediene' in Physics, Chemistry and Computer Sciences
No. 1, p. 14	Cinini, Alessandra see Sassolini, Eva
No 3. 139	Collection of Conference Proceedings and Improving Access to the Full Text of Proceedings / Misa Hayakawa, Shun Nagaya, Mayuki Gonda, Takeyasu Fukazawa, M. Yonezawa, and K. Itabashi
No 3. 157	Crowe, June ; Candlish, J.R. Data Analytics: The next big thing in information
No. 1, p. 40	Customized OAI-ORE and OAI-PMH Exports of Compound Objects for the Fedora Repository / Alessia Bardi, Sandro La Bruzzo, and Paolo Manghi
	D
No 3. 157	Data Analytics: The next big thing in information / June Crowe and J.R. Candlish
No 2, 69	Data sharing in environmental sciences: A survey of CNR researchers / Daniela Luzi, Roberta Ruggieri, Stefania Biagioni, and Elisabetta Schiano
No 2, 61	De Santo, Elizabeth M. see MacDonald, Bertrum H.
No 3. 129	Di Cesare, Rosa ; Nobile, Marianna; Giannini, Silvia Working for an open e-publishing service to improve grey literature editorial quality
No. 1, p. 19	Dušková, Marta see Kravjar, Július
	E - F
No. 1, p. 24	An Environment Supporting the Production of Live Research Objects / Massimiliano Assante, Leonardo Candela, and Pasquale Pagano
No 3. 139	Fukazawa, Takeyasu see Hayakawa, Misa
No. 1, p. 32	A Funder Repository of Heterogeneous Grey Literature Material with Advanced User Interface and Presentation Features / Ioanna-Ourania Stathopoulou, Nikos Houssos, Panagiotis Stathopoulos, Despina Hardouveli, Alexandra Roubani, Ioanna Sarantopoulou, Alexandros Soumplis, and Chrysostomos Nanakos
	G
No 2, 97	Galatis, Helen see White, Gerald
No 3. 129	Giannini, Silvia see Di Cesare, Rosa
No 3. 139	Gonda, Mayuki see Hayakawa, Misa
No 3. 152	Grey in the Innovation Process / Keith G. Jeffery and Anne Asserson
No 2, 97	Grey literature in Australian education / Gerald White, Julian Thomas, Paul Weldon, Amanda Lawrence, Helen Galatis and Jess Tyndall
No 3. 122	Grey literature in the digital culture and practices of the global scholar: the case of molecular biology / Chérifa Boukacem-Zeghmouri
No 2, 82	Gribling, Armand see Kalentsits, Maria
	H
No. 1, p. 32	Hardouveli, Despina see Stathopoulou, Ioanna-Ourania
No 3. 139	Hayakawa, Misa ; Nagaya, Shun; Gonda, Mayuki; Fukazawa, Takeyasu; Yonezawa, M.; Itabashi, K. Collection of Conference Proceedings and Improving Access to the Full Text of Proceedings
No. 1, p. 32	Houssos, Nikos see Stathopoulou, Ioanna-Ourania
	I - J
No 3. 145	Innovation, language, and grey literature / Claudia Marzi
No 3. 139	Itabashi, K. see Hayakawa, Misa
No 3. 152	Jeffery, Keith G. ; Asserson, Anne Grey in the Innovation Process
	K
No 2, 82	Kalentsits, Maria ; Gribling, Armand What goes up must come down: Publications from developing countries in the Aquatic Common

- No 2, 91 *Kim, Hye-Sun* see **Lee, Seon-Hee**
- No. 1, p. 19 **Kravjar, Július** and *Dužková, Marta*
A Centralised National Corpus of Electronic Theses and Dissertations
L
- No. 1, p. 40 *La Bruzzo, Sandro* see **Bardi, Alessia**
- No 2, 97 *Lawrence, Amanda* see **White, Gerald**
- No 2, 91 **Lee, Seon-Hee**; *Kim, Hye-Sun*
The Research Life Cycle and Innovation through Grey Literature in Nanotechnology in Korea
- No. 1, p. 7 **Lin, Yongtao**; *Vaska, Marcus*
Assessing a Subject-based Blog for Current Awareness within a Cancer Care Environment /
- No 2, 69 **Luzi, Daniela**; *Ruggieri, Roberta*; *Biagioni, Stefania*; *Schiano, Elisabetta*
Data sharing in environmental sciences: A survey of CNR researchers
M
- No 2, 61 **MacDonald, Bertrum H.** ; *De Santo, Elizabeth M.* ; *Quigley, Kevin* ; *Soomai, Suzuette S.* ; *Wells, Peter G.*
Tracking the Influence of Grey Literature in Public Policy Contexts: The Necessity and Benefits of Interdisciplinary Research
- No. 1, p. 40 *Manghi, Paolo* see **Bardi, Alessia**
- No 3. 145 **Marzi, Claudia**
Innovation, language, and grey literature
N
- No 3. 139 *Nagaya, Shun* see **Hayakawa, Misa**
- No. 1, p. 32 *Nanakos, Chrysostomos* see **Stathopoulou, Ioanna-Ourania**
- No 3. 129 *Nobile, Marianna* see **Di Cesare, Rosa**
- P - Q**
- No. 1, p. 24 *Pagano, Pasquale* see **Assante, Massimiliano**
- No. 1, p. 14 *Picchi, Eugenio* see **Sassolini, Eva**
- No 2, 61 *Quigley, Kevin* see **MacDonald, Bertrum H.**
- R**
- No 2, 91 The Research Life Cycle and Innovation through Grey Literature in Nanotechnology in Korea / Seon-Hee Lee and Hye-Sun Kim
- No. 1, p. 32 *Roubani, Alexandra* see **Stathopoulou, Ioanna-Ourania**
- No 2, 69 *Ruggieri, Roberta* see **Luzi, Daniela**
- S**
- No. 1, p. 32 *Sarantopoulou, Ioanna* see **Stathopoulou, Ioanna-Ourania**
- No. 1, p. 14 **Sassolini, Eva**; *Cinini, Alessandra*; *Sbrulli, Stefano*; *Picchi, Eugenio*
Tools And Resources Supporting Cultural Tourism
- No. 1, p. 14 *Sbrulli, Stefano* see **Sassolini, Eva**
- No 2, 69 *Schiano, Elisabetta* see **Luzi, Daniela**
- No 2, 61 *Soomai, Suzuette S.* see **MacDonald, Bertrum H.**
- No. 1, p. 32 *Soumplis, Alexandros* see **Stathopoulou, Ioanna-Ourania**
- No. 1, p. 32 *Stathopoulos, Panagiotis* see **Stathopoulou, Ioanna-Ourania**
- No. 1, p. 32 **Stathopoulou, Ioanna-Ourania**; *Houssos, Nikos*; *Stathopoulos, Panagiotis*; *Hardouveli, Despina*; *Roubani, Alexandra*; *Sarantopoulou, Ioanna*; *Soumplis, Alexandros*; *Nanakos, Chrysostomos*
A Funder Repository of Heterogeneous Grey Literature Material with Advanced User Interface and Presentation Features
T
- No 2, 97 *Thomas, Julian* see **White, Gerald**
- No. 1, p. 14 Tools And Resources Supporting Cultural Tourism / Eva Sassolini, Alessandra Cinini, Stefano Sbrulli, and Eugenio Picchi
- No 2, 61 Tracking the Influence of Grey Literature in Public Policy Contexts: The Necessity and Benefits of Interdisciplinary Research / Bertrum H. MacDonald, Elizabeth M. De Santo, Kevin Quigley, Suzuette S. Soomai, and Peter G. Wells
- No 2, 97 *Tyndall, Jess* see **White, Gerald**
- V – W - Y**
- No. 1, p. 7 *Vaska, Marcus* see **Lin, Yongtao**
- No 2, 97 *Weldon, Paul* see **White, Gerald**
- No 2, 61 *Wells, Peter G.* see **MacDonald, Bertrum H.**
- No 2, 82 What goes up must come down: Publications from developing countries in the Aquatic Common / Maria Kalentsits and Armand Gribling
- No 2, 97 **White, Gerald**; *Thomas, Julian*; *Weldon, Paul*; *Lawrence, Amanda*; *Galatis, Helen*; *Tyndall, Jess*
Grey literature in Australian education
- No 3. 129 Working for an open e-publishing service to improve grey literature editorial quality / Rosa Di Cesare, Marianna Nobile, and Silvia Giannini
- No 3. 139 *Yonezawa, M.* see **Hayakawa, Misa**

Asserson, Anne
152

Anne Asserson holds a master from the University of Bergen, UiB. She has been working with Research Documentation, and has participated in substantial parts of CRIS developmental work, locally and nationally since 1992. Anne Asserson has been part of the establishing and implementing of several CRIS both at the UiB and nationally. For several years she was the chairwoman of the Steering Group of the national CRIS system and project secretary of a National system for academic administration. Anne Asserson is presently representing UiB in the national project group of CRISin. She has also participated in The CORDIS funded European-wide project on " Best Practice" 1996 and was a member of the working group set up 1997 that produced the report CERIF2000 Guidelines (1999) www.cordis.lu/cerif, coordinated by the DGXIII-D4. euroCRIS is now the custodian of the CERIF model www.eurocris.org. Anne Asserson is a member of the euroCRIS board with the responsibility Member Strategy and External Relations. anne.asserson@fa.uib.no

Boukacem-Zeghmouri, Chérifa
122

Chérifa Boukacem graduated with a Master's Degree in library sciences from the University of Algiers (DZ) in 1996. She obtained her Ph.D. from the University of Lyon-2 in 2004 with a comparative study on French academic interlibrary loan and access to digital resources. From 2001 to 2003, she worked at the department of scientific and technical information of the University of Lyon-1, she teaches since 2000 library and information science to students and professionals, is member of the French research group in library sciences, DOCS/GRESI, and co-edits a digital serial, Konex. Her main interests are on e-publishing and multimedia, library software, statistics and Scientometrics and professional counseling. Actually, she is working on the transformation of the digital library and on a longitudinal study of document supply, impact factor and the switch from print to electronic resources. boukacemc@yahoo.fr

Candlish, J.R.
157

Joseph R. Candlish is a Technical Assistant and GIS Analyst at Information International Associates, Inc. (IIa). He received his M.S. in Biosystems Engineering Technology from the University of Tennessee, Knoxville and his B.S. in Environmental Studies: Natural Resources from Sewanee: The University of the South. At IIa he provides secretariat services to CENDI agencies, the federal scientific information manager's group and also fulfills a mapping role for the USGS National Biological Information Infrastructure's Southern Appalachian Information Node (SAIN). jcandlish@iiaweb.com

Crowe, June
157

June Crowe is the Senior Researcher and Group Manager, Open Source Research Division at Information International Associates, Inc. (IIa). She received her AMLS from the University of Michigan, Ann Arbor and her M.Ed. in geographic education from the University of Georgia, Athens. She has extensive experience in the management and operations of library services across government, public, academic, and special libraries. At IIa she manages the open source research division which focuses on medical, socio-cultural, science and technology and business research. Her primary interests are open source information in Grey Literature, digital repositories, and open source intelligence tools. jcrowe@iiaweb.com

Di Cesare, Rosa
129

Rosa Di Cesare is responsible for the library at the Institute for research on populations and social policies of the National Research Council (CNR). She worked previously at the Central library of CNR where she became involved in research activities in the field of Grey literature (GL) as member of the Technical Committee for the SIGLE database. Her studies have focused on the use of GL in scientific publications and recently on the emerging models of scholarly communication (OA and IR). r.dicesare@irpps.cnr.it

Giannini, Silvia
129

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 - infrastructure software for institutional and thematic Open Access repositories of published and grey literature produced by CNR. silvia.giannini@isti.cnr.it

Gonda, Mayuki
139

Mayuki Gonda works as librarian at the Central Library of JAEA (Japan Atomic Energy Agency). He joined JAERI (former JAEA) in 2005, and had been working for management and dissemination of JAEA research results information. Since 2009, he is in charge of selection, classification and indexing for INIS (International Nuclear Information System) Database. He is also a member of editorial committee of the Journal of Information Science and Technology Association (INFOSTA) since 2008. He holds a degree in information science (M.A.) from the Graduate School of Library, Information and Media Studies, University of Tsukuba. gonda.mayuki@jaea.go.jp

Hayakawa, Misa
139

Misa Hayakawa works as librarian at the Central Library of Japan Atomic Energy Agency(JAEA). She Joined JAEA in 2010, and has been working for managing metadata of both papers and oral presentations published by JAEA researchers. In addition, she disseminates such information on the Internet via "JAEA Originated Papers Searching System"(JOPSS). She got the master's degree in Library and Information Science at the Graduate School of Library Information and Media Studies, University of Tsukuba in Japan. hayakawa.misa@jaea.go.jp

Jeffery, Keith G.
152

Keith Jeffery is an information consultant. He is the former Director of International Relations at STFC (Science and Technology Facilities Council) based at Rutherford Appleton Laboratory. Keith previously had strategic and operational responsibility for ICT with 360,000 users, 1100 servers and 140 staff. Keith holds 3 honorary visiting professorships, is a Fellow of the Geological Society of London and the British Computer Society, is a Chartered Engineer and Chartered IT Professional and an Honorary Fellow of the Irish Computer Society. Keith is currently President of ERCIM and President of euroCRIS, and serves on international expert groups, conference boards and assessment panels. He had advised government on security and green computing. He chaired the EC Expert Groups on GRIDs and on cloud computing. keith.jeffery@keithgiefferyconsultants.co.uk

Marzi, Claudia
145

Claudia Marzi is a research fellow at the Institute for Computational Linguistics "A. Zampolli" (ILC), National Research Council (CNR), Pisa. PhD in Acquisitional and Computational Linguistics, University of Pavia, in progress. Laurea degree in Modern Languages (English) at Pisa University in 1998, with the dissertation "The power of words: language creativity in Edgar Allan Poe's narrative". Programme Coordinator of the European Science Foundation Research Networking Programme "The European Network on Word Structure. Cross-disciplinary approaches to understanding word structure in the languages of Europe". Member of board at Institute for Computational Linguistics. Main areas of interest: Computer modelling of the Mental Lexicon; Second language acquisition; Child language; Document and knowledge management. claudia.marzi@ilc.cnr.it

Nobile, Marianna
129

Marianna Nobile graduated in Linguistics, Faculty of Letters and Philosophy at the University of Rome La Sapienza. She had an internship in the library of the Senate of the Republic (Giovanni Spadolini Library), where she worked with the Books Acquisition. Currently she is collaborating with the library of the Institute for Research on Population and Social Science of the Italian National Research Council and she is involved in the activities of indexing electronic resources as well as the development of an e-publishing service. marianna.nobile@irpps.cnr.it

Notes for Contributors

Non-Exclusive Rights Agreement

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

Submission Requirements

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

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

REFERENCE GUIDELINES

General

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

Specific

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

Examples

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

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

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

Review Process

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

Journal Publication and Article Deposit

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

An International Journal on Grey Literature

'IMPROVING GREY LITERATURE THROUGH INNOVATION'

AUTUMN 2013 – TGJ VOLUME 9, NUMBER 3

Grey literature in the digital culture and practices of the global scholar: the case of molecular biology	122
<i>Chérifa Boukacem-Zeghmouri (France)</i>	
Working for an open e-publishing service to improve grey literature editorial quality	129
<i>Rosa Di Cesare, Marianna Nobile, and Silvia Giannini (Italy)</i>	
Collection of Conference Proceedings and Improving Access to the Full Text of Proceedings	139
<i>Misa Hayakawa, Shun Nagaya, Mayuki Gonda, Takeyasu Fukazawa, M. Yonezawa, and K. Itabashi (Japan)</i>	
Innovation, language, and grey literature	145
<i>Claudia Marzi (Italy)</i>	
Grey in the Innovation Process	152
<i>Keith G. Jeffery (United Kingdom) and Anne Asserson (Norway)</i>	
Data Analytics: The next big thing in information	157
<i>June Crowe and J.R. Candlish (United States)</i>	

Colophon.....	118
Editor's Note.....	121
On the News Front	
Launch GreyGuide Repository	
'A Guide to Proposed and Published Practices in Grey Literature'.....	160
GreyForum Workshop 2.1	
Grey Literature and Information Rights: On Law and Practice.....	161
GL15 Conference Outline and Registration	
'The Grey Audit, a Field Assessment in Grey Literature'.....	162
Advertisements	
INIS, The International Nuclear Information System.....	120
IAEA, Department of Nuclear Energy	138
Author and Title Index 2013.....	164
About the Authors.....	166
Notes for Contributors.....	167