

Eighteenth International Conference on Grey Literature

Leveraging Diversity in Grey Literature

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



Program Book

ISSN 1385-2308

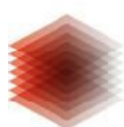
Host and Sponsors



Nuclear Information Section



International Atomic Energy Agency



TIB LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY



Inist



GL18 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



CIP

GL18 Program Book

Eighteenth International Conference on Grey Literature : Leveraging Diversity in Grey Literature – New York, NY USA, on November 28-29, 2016 / compiled by D. Farace and J. Frantzen ; GreyNet International, Grey Literature Network Service. – Amsterdam : TextRelease, November 2016. – 116 p. – Author Index. – (GL Conference Series, ISSN 1385-2308 ; No. 18).

The New York Academy of Medicine (USA), CVTISR (SK), DANS-KNAW (NL), EBSCO (USA), FEDLINK - Library of Congress (USA), Inist-CNRS (FR), ISTI-CNR (IT), KISTI (KR), NIS-IAEA (AT), and NTK (CZ) are Corporate Authors and Associate Members of GreyNet International. This program book contains the schedule for the plenary and panel sessions, as well as the poster session and sponsor showcase. The titles and abstracts of the papers as well as information on the authors are provided. When available, copies of the PowerPoint slides are included in notepad format.

Foreword

LEVERAGING DIVERSITY IN GREY LITERATURE

Scientific information, much of which is published as grey literature, can play a pivotal role in the search for solutions to global problems. Diversity invigorates problem solving and science benefits from a community that approaches problems in a variety of creative ways. Despite their diversity, the hundreds of authors and researchers across the globe involved in grey literature can be seen as part of the same community contributing to the scientific enterprise in valuable ways.

Diversity speaks directly to the effectiveness of information professionals working together as a team and is an essential ingredient for innovation. People from different backgrounds bring with them new information. If you want to build teams, communities, and organizations capable of innovating, you need diversity. It enhances creativity and encourages the search for new information and nuanced perspectives, leading to better decision making and problem solving. Diversity can improve the bottom line of companies as well as organizations, because exposure to it changes the way one thinks. A diverse community of researchers anticipate differences and understand that they will have to work harder to achieve consensus, but their diligence can lead to better outcomes. Authors in the GL-Conference Series come from different societal cultures and geographic regions; however in their research, they are united by the culture of science, which is without borders. This diverse community has over the past two decades applied research methods and offered explanations that have helped this field of information through blind spots, shedding light on what were once seen only as inherent problems. Their evidence based approaches have opened up new areas of research in grey literature. Where in the early '90s the focus was primarily on the demand side of grey literature, equal emphasis today is directed to its supply side. Speed and scale of communication are significant factors that contribute to diversity. The proliferation of technologies has allowed for an exponential growth of knowledge in information science just as in other sciences. However, the diverseness of grey literature resources has become a major challenge to its exploitation. The availability of systems for collecting and aggregating data and its semantic analysis has now become a priority.

GL18 focusses on evidence and seeks to further raise awareness among the wider public to the strength of grey literature based on a shared commitment by a diverse community of authors and researchers responsible for its production, open access, and digital preservation.

Dominic Farace
GREYNET INTERNATIONAL

Amsterdam,
NOVEMBER 2016

A terminological “journey” in the Grey Literature domain

Roberto Bartolini, Gabriella Pardelli, Sara Goggi,

CNR, Istituto di Linguistica Computazionale, “Antonio Zampolli”

Silvia Giannini and Stefania Biagioni,

CNR, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, Italy

“When we read the articles or papers of a particular domain, we can recognize some lexical items in the texts as technical terms. In a domain where new knowledge is generated, new terms are constantly created to fulfil the needs of the domain, while others become obsolete. In addition, existing terms may undergo changes of meaning...” (Kageura K., 1998/1999). According to Kaugera, our aim with this work is to make a “journey” in the Grey Literature (GL) domain in order to offer an overall vision on the terms used and the links between them. Moreover, by performing a terminological comparison over a given period of time it could be possible to trace the presence of obsolete words as well as of neologisms in the most recent research fields.

Within this scenario, the work analyzes a corpus constituted of the entire amount of full research papers published in the GL conference series over a time span of more than one decade (2003-2014) with the aim of creating a terminological map of relevant words. “... corpora used to extract terminological units can be further investigated to find semantic and conceptual information on terms or to represent conceptual relationships between terms. (Bourigault D. et al., 2001). Another interesting inquiry is the terminology used in the GL conferences for describing the types of documents (Pejšová P. et al., 2012).

The work is split up in four sections: creation of the corpus by acquiring the digital papers of GL conference proceedings (GL5 – GL16)¹; data cleaning; data processing; terminological analysis and comparison.

The corpus - made up of 231 research papers (for a total amount of 785.042 tokens) - was processed using a Natural Language Processing (NLP) tool for term extraction developed at the Institute of Computational Linguistics “Antonio Zampolli” of CNR (Goggi et al. 2015; 2016). This tool is what is called a “pipeline” (that is, a sequence of different tools) which extracts lexical knowledge from texts: in short, this is a rule system tool for knowledge extraction and document indexing that combines NLP technologies for term extraction and techniques to measure the associative strength of multi-words. This tool extracts a list of single (monograms) and multi-word terms (bigrams and trigrams) ordered by frequency with respect to the context. The pipeline - used as semantic engine within the MAPS project - has been customized for the extraction of terms from our corpus.

This survey on the results of the information extraction process performed by the described NLP tool has been a sort of linguistic path in the past and present of terminology used in GL proceedings. By means of samplings, it has been possible to obtain the terminological flow in GL domain and to determine if and how the lexicon was evolving over these twelve years and investigate on its dynamic nature.

¹ Kindly provided by Greynet International, <http://www.greynet.org/>.

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

Bionotes

Roberto Bartolini - Expertise on design and development of compilers of finite state grammars for functional analysis (macro-textual and syntactic) of Italian texts. Expertise on design and implementation of compilers of finite state grammars for analysis of natural language texts producing not recursive syntactic constituents (chunking) with specialization for Italian and English languages. Skills on acquiring and extracting domain terminology from unstructured text. Skills on semi-automatic acquisition of ontologies from texts to support advanced document management for the dynamic creation of ontologies starting from the linguistic analysis of documents. Email: roberto.bartolini@ilc.cnr.it



Gabriella Pardelli was born at Pisa, graduated in Arts in 1980 at the Pisa University, submitting a thesis on the History of Science. Since 1984, researcher at the National Research Council, Institute of Computational Linguistics "Antonio Zampolli" ILC, in Pisa. Head of the Library of the ILC Institute since 1990. Her interests and activity range from studies in grey literature and terminology, with particular regard to the Computational Linguistics and its related disciplines, to the creation of documentary resources for digital libraries in the Humanities. She has participated in many national and international projects including the recent projects:- BIBLOS: Historical, Philosophical and Philological Digital Library of the Italian National Research Council, (funded by CNR); - For digital edition of manuscripts of Ferdinand de Saussure (Research Programs of Relevant National Interest, PRIN - funded by the Ministry of Education, University and Research, MIUR). Email: gabriella.pardelli@ilc.cnr.it



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

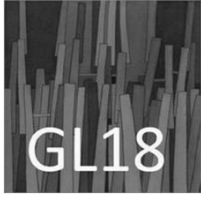


Silvia Giannini graduated and specialized in library sciences. Since 1987 she has been working in Pisa at the Institute for the Science and Technologies of Information "A. Faedo" of the Italian National Council of Research (ISTI-CNR) as a librarian. She is a member of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She is responsible of the library automation software "Libero" in use at the CNR Research Area in Pisa and coordinates the bibliographic and managing activities of the ISTI library team. She cooperates in the design and development of the PUMA (Publication Management) & MetaPub, an infrastructure software for institutional and thematic Open Access repositories of published and grey literature produced by CNR. Email: silvia.giannini@isti.cnr.it



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






A TERMINOLOGICAL "JOURNEY" IN THE GREY LITERATURE DOMAIN

Roberto Bartolini, Gabriella Pardelli, Sara Goggi, CNR-ILC, Pisa Italy
Silvia Giannini, Stefania Biagioni, CNR-ISTI, Pisa Italy

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

SUMMARY

- Scenario & Objectives
- GL Corpus and Method
- Terminological Analysis - GL Topics
- Types of documents - Conclusions



SCENARIO & OBJECTIVES

"When we read the articles or papers of a particular domain, we can recognize some lexical items in the texts as technical terms. In a domain where new knowledge is generated, new terms are constantly created to fulfill the needs of the domain, while others become obsolete. In addition, existing terms may undergo changes of meaning." (Haugen, 1959)

This work analyzes a corpus constituted of the entire amount of full research papers published in the GL conference series over a time span of more than one decade (2003-2014) with the aim of

- making a "journey" in the Grey Literature (GL) domain in order to offer an overall vision on the terms used and the links between them;
- creating a terminological map of relevant words;
- tracing the presence of obsolete words as well as of neologisms in the most recent research fields;
- analyzing the terminology used in the GL conferences for describing the various types of documents.

A TERMINOLOGICAL "JOURNEY" IN THE GREY LITERATURE DOMAIN

GL CORPUS AND METHOD

The work is split up in four sections:

- creation of the corpus by acquiring the digital papers of GL conference proceedings (GL5 – GL16);
- data cleaning;
- data processing using the NLP "pipeline" tool;
- terminological analysis and comparison.

➤ **GL Corpus:**
made of 231 research papers (for a total amount of 785.042 tokens: monograms, bigrams and trigrams);

➤ **Natural Language Processing (NLP):**
data was processed using a tool for term extraction, a sort of "pipeline" (that is, a sequence of different tools) which extracts lexical knowledge from texts. This tool extracts a list of single (monograms) and multi-word terms (bigrams and trigrams) ordered by frequency with respect to the context.

GL CORPUS & METHOD
