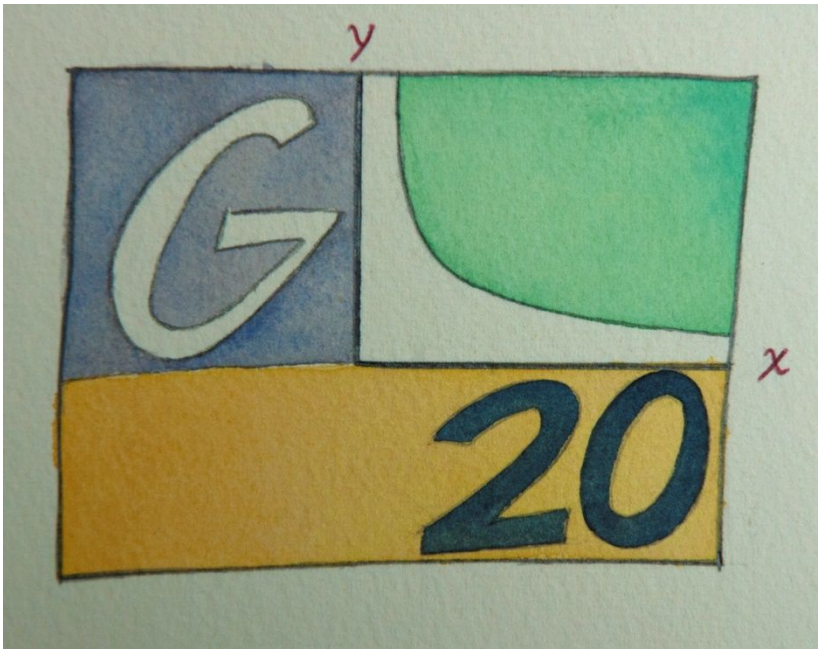
Twentieth International Conference on Grey Literature

Research Data Fuels and Sustains Grey Literature

Loyola University New Orleans, USA • December 3-4, 2018

Program Book

ISSN 1385-2308



Program and Conference Sponsors

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

CENTRUM VEDECKO-TECHNICKÝCH INFORMACIÍ SR

DANS

Nuclear Information Section
NIS - IAEA
International Atomic Energy Agency

TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

KiSTi www.kisti.re.kr

Inist cnrs

EBSCO PUBLISHING

LIBRARY OF CONGRESS FEDLINK

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Technical Library

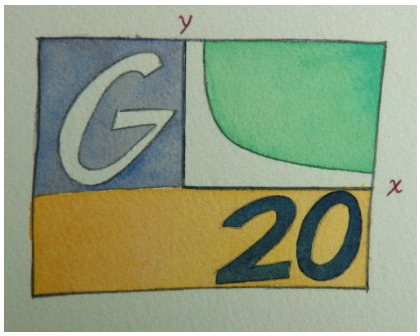
UF George A. Smathers Libraries UNIVERSITY of FLORIDA

UNIVERSITY of WISCONSIN UWMILWAUKEE School of Information Studies

GL20 Program and Conference Bureau

TextRelease

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



CIP

GL20 Program Book
Twentieth International Conference on Grey Literature “Research Data Fuels and Sustains Grey Literature”. - Loyola University New Orleans, USA December 3-4, 2018 / compiled by D. Farace and J. Frantzen ; GreyNet International, Grey Literature Network Service. – Amsterdam : TextRelease, December 2018. – 136 p. – Author Index. – (GL Conference Series, ISSN 1385-2308 ; No. 20).
TIB (DE), DANS-KNAW (NL), FEDLINK-Library of Congress (USA), CVTISR (SK), EBSCO (USA), Inist CNRS (FR), ISTI-CNR (IT), KISTI (KR), NIS-IAEA (UN), NTK (CZ), and University of Florida; George A. Smathers Libraries (USA) 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.

ISBN 978-90-77484-34-0

Foreword

RESEARCH DATA FUELS AND SUSTAINS GREY LITERATURE

The definition of research data is as encompassing as the field of grey literature. What should be included and what should be excluded is and remains an issue of concern. Research data can be defined as factual materials collected by diverse communities of practice required to validate findings. While the majority of research data is created in digital format, research data in other formats cannot be excluded. The formats in which research data appear are multiple and the types of research data are diverse. This also holds for the numerous document types in which grey literature appear published.

Today, while emphasis is placed on big data, the fact that the majority of research projects are small to medium size is overlooked. This is but another characteristic that holds true for grey literature. Nonetheless, one should be aware that research publications are not research data, for they are often managed separately from one another. Just as there are a number of stakeholders involved in the production, access, and preservation of grey publications, so too are there stakeholders tasked with the creation and management of research data. Libraries and data management librarians have the responsibility for the curation of the data they collect and preserve. And, it is important to stress the need to maintain appropriate metadata related to research data in order to facilitate their interpretation and further reuse.

Over the past quarter century, grey literature communities have worked diligently to demonstrate how their documents are produced, published, reviewed, indexed, accessed, and further used, applied, and preserved. Today, these communities are now challenged to demonstrate how research data fuels and sustains their grey literature. These communities of dedicated researchers and authors maintain a strong conviction in the uses and applications of grey literature for science and society. Through the years, they have proved willing to share the results of scholarly work well beyond their own institutions. Hence, one can assume they are aware that innovation forfeits with the loss of data as with the loss of information. This 20th International Conference in the GL-Series seeks to address key issues and topics related to grey literature and its underlying research data.

Dominic Farace — *Amsterdam,*
GREYNET INTERNATIONAL — DECEMBER 2018

GL20 Conference Sponsors

ISTI, Italy

Institute of Information Science and Technologies
National Research Council of Italy, CNR

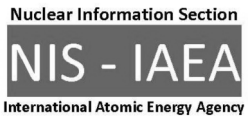
CVTISR, Slovak Republic

Slovak Centre of Scientific and Technical Information

KISTI, Korea

Korea Institute of Science and Technology
Information

EBSCO, USA



NIS-IAEA, Austria

Nuclear Information Section;
International Atomic Energy Agency

TIB, Germany

German National Library of Science and Technology –
Leibniz Information Centre for Science and
Technology University Library

GL20 Conference Sponsors (CONTINUED)

DANS, Netherlands
Data Archiving and Networked Services;
Royal Netherlands Academy of Arts and Sciences

NTK, Czech Republic
National Library of Technology

FEDLINK, USA
Federal Library Information Network;
Library of Congress

Inist-CNRS, France
Institut de l'Information Scientifique et Technique; Centre National de Recherche Scientifique

UW-Milwaukee-SOIS, USA
School of Information Studies (SOIS)
University of Wisconsin, Milwaukee

UF, USA
George A. Smathers Libraries
University of Florida

GL20 Program Committee

Brian Hitson [Chair]
Office of Scientific and Technical Information; U.S. Department of Energy

Robert Bell
Loyola University New Orleans, USA

George Barnum
U.S. Government Publishing Office

Meg Tulloch
U.S. Government Accountability Office

Margret Plank
German National Library of Science and Technology, Germany

Dobrica Savić
Nuclear Information Section, International Atomic Energy Agency, Austria

Christiane Stock
Institut de l'Information Scientifique et Technique CNRS, France

Hana Vyčítalová
National Library of Technology, Czech Republic

Silvia Giannini
Institute of Information Science and Technologies ISTI-CNR, Italy

Ján Turňa
Slovak Centre of Scientific and Technical Information Slovak Republic

Stefania Biagioni
NeMIS Research Laboratory
Italy

Joachim Schöpfel
University of Lille
France

Judith C. Russell
University of Florida Libraries
USA

Plato L. Smith
University of Florida; George A. Smathers Libraries, USA

Henk Harmsen
Data Archiving and Networked Services, Netherlands

Marcus Vaska
Alberta Health Services
Canada

Dominic Farace
GreyNet International
Netherlands

Tomas A. Lipinski
University of Wisconsin Milwaukee, USA

Semantic Query Analysis from the Global Science Gateway

Sara Goggi, Gabriella Pardelli, Roberto Bartolini, and Monica Monachini, ILC-CNR, Italy
Stefania Biagioni and Carlo Carlesi, ISTI-CNR, Italy

Nowadays web portals play an essential role in searching and retrieving information in the several fields of knowledge: they are ever more technologically advanced and designed for supporting the storage of a huge amount of information in natural language originating from the queries launched by users worldwide.

A good example is given by the *WorldWideScience search engine:*

The database is available at <http://worldwidescience.org/>. It is based on a similar gateway, Science.gov, which is the major path to U.S. government science information, as it pulls together Web-based resources from various agencies. The information in the database is intended to be of high quality and authority, as well as the most current available from the participating countries in the Alliance, so users will find that the results will be more refined than those from a general search of Google. It covers the fields of medicine, agriculture, the environment, and energy, as well as basic sciences. Most of the information may be obtained free of charge (the database itself may be used free of charge) and is considered ''open domain.'' As of this writing, there are about 60 countries participating in WorldWideScience.org, providing access to 50+databases and information portals. Not all content is in English. (Bronson, 2009)

Given this scenario, we focused on building a corpus constituted by the query logs registered by the *GreyGuide: Repository and Portal to Good Practices and Resources in Grey Literature*[1] *and received* by the ***WorldWideScience.org***[2] *(The Global Science Gateway) portal*: the aim is to retrieve information related to social media which as of today represent a considerable source of data more and more widely used for research ends.

This project includes eight months of query logs[3] registered between July 2017 and February 2018 for a total of 445,827 queries. The analysis mainly concentrates on the semantics of the queries received from the portal clients: it is a process of information retrieval from a rich digital catalogue whose language is dynamic, is evolving and follows – as well as reflects – the cultural changes of our modern society.

Methods and Tools

In order to analyze the available information a considerable pre-processing on four levels has been carried out:

- at the first level, the set of queries has been cleaned: duplicates, alphanumeric strings, strange graphical forms, IP addresses, etc. have been eliminated;
- at the second level, filters have been added and alphabetical order inserted for having a first picture of the contents of these queries;
- the third step consisted of several trials for choosing the focus;
- lastly, natural language processing (NLP) tools have been applied for processing the information and building the sample.

[1] http://greyguide.isti.cnr.it/ - *GreyGuide is the* online forum and repository of good practices and resources in Grey Literature. It was created - and is now edited - by GreyNet International (content provider) and ISTI-CNR, Pisa Italy (service provider): its launch was in December 2013 and since then *GreyGuide* provides a unique resource in the field of grey literature, which was long awaited and responds to the information needs of a diverse, international grey literature community. *Greynet International is one of the WorldWideScience Associate Members https://worldwidescience.org/alliancemembers.html.*

[2] https://worldwidescience.org/ - It is a global science gateway comprised of national and international scientific databases and portals. ***WorldWideScience.org accelerates*** scientific discovery and progress by providing one-stop searching of databases from around the world. WorldWideScience.org is maintained by the U.S. Department of Energy's Office of Scientific and Technical Information as the Operating Agent for the WorldWideScience Alliance.

[3] The General *Query Log* is the record of each SQL statement received from clients, in addition to their connection and disconnection time.

Since the corpus is made of queries collected in only eight months and the cleaning process reduced them consistently, as a result the final is relatively small. In addition, only the queries in English have been registered while those in other languages have been eliminated (there are a few in French, Spanish, Italian, Portuguese, Polish, Albanese, Galician, Corsican, and so on). As a curiosity, some queries also deal with socio-political or historical events (1915 Mexico Guerra, 1929 crisis unidos, 1960s economy, 1963 Sicily earthquake[4], 1979 Iran revolution, 1979 revolucion irani, 1984 George Orwell, 1986 FBI Shootout).
Coming to the NLP analysis, the software team has decided to follow these two steps:

1. free information extraction: it measured the frequency of all the words contained in the corpus. This preliminary investigation provided us with the whole scenario of the lexical variety of the queries and allowed us to focus on a set of terms from which we built a micro-ontology with meaningful terms relating to the queries launched on the portal;
2. ontology-based extraction: the extraction has been performed again using this micro-ontology which has been essentially used for enriching the domain. In this way, the search engine retrieved each single occurrence of those terms (monograms, bigrams, trigrams) which can be found starting from the ontology.

At the end of this pre-processing phase, we chose to focus on a flow of queries launched on the *WorldWideScience* platform concerning only the bigram *social media.*
Why social media?

- ✓ *nowadays social media are obviously a very effective means of communication but can even vehiculate knowledge as their various types (*eg.: blogs, YouTube*, Facebook, Twitter, etc.) are by now often quoted in bibliographical references amongst the more traditional categories (books, journals and so on);*
- ✓ *the subject involves document types pertaining to Grey Literature.*

An example of this type of extraction is given by some terms which belong to the fields of medicine and psychiatry and are paired with the bigram *social media*:

<***cyberbullying*** *social media*>
<***depression*** *social media*>
<***eating disorder*** *social media*>
<***negative effects*** *social media young adults*>
<***anxiety*** *social media*>
<*social media* **compulsive buying**>
<*social media* ***distraction***>
<***fake news*** *social media*>
<*social media* **millennial**>

At once, these terms show a negative connotation in relation to the use of social media: this phenomenon seems rather relevant in the queries of our corpus and therefore deserves further investigation.

Conclusion

The poster will illustrate the main linguistic features of *the Global Science Gateway by showing*:

- ✓ the lexical map representing the most used/recurrent words (in terms of occurrences) as well as the adoption of neologisms (a very interesting one is “netnography”) and hapaxes (such as “hastag”) in the realm of queries on social media;
- ✓ the comparison among the various typologies of social media on polarity[5], similarity and diversity.

[4] It is actually referring to the Sicily earthquake of 1693, not 1963.
[5] “In linguistics, a **polarity item** is a lexical item that can appear only in environments associated with a particular grammatical polarity – affirmative or negative. A polarity item that appears in affirmative (positive) contexts is called a **positive polarity item** (PPI), and one that appears in negative contexts is a **negative polarity item** (NPI)”, Wikipedia.

Bionotes

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

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 projects. Member of board at Institute for Computational Linguistics. She is author and co-author a number of publications dealing with Computational Linguistics, Computational Terminology and Grey Literature. Email: gabriella.pardelli@ilc.cnr.it

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

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

Stefania Biagioni graduated in Italian Language and Literature at the University of Pisa and specialized in Data Processing and DBMS. She is currently an associate 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 currently involved in the activities of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She has been head librarian of the Multidisciplinary Library of the CNR Campus in Pisa till August 2017. She was the responsible of ERCIM Technical Reference Digital Library (ETRDL) Project and currently is the coordinator 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 and grey literature. Her research interest are focused on digital libraries, knowledge sharing and transfer in scientific area, scholarly communication infrastructures, Open Access and Open Science. She has been dealing with grey literature since 90's. Since 2013 she is involved on the GreyGuide Project. Email: stefania.biagioni@isti.cnr.it
ORCID iD https://orcid.org/0000-0001-9518-0267

Carlo Carlesi, graduated in Computer Science, worked since 1970 at the IEI (now ISTI) of the CNR in Pisa. He is currently a Research Associate of the Institute ISTI and he is involved in the following projects: PUMA - Publication Management. The Digital Library service allows public access (when permitted) through Internet to the published documents produced by CNR Organizations. And GreyGuide, portal and repository of good practice and resources in the field of grey literature. Email: carlo.carlesi@isti.cnr.it
ORCID iD https://orcid.org/0000-0001-9808-6268