Eighteenth International Conference on Grey Literature

Leveraging Diversity in Grey Literature

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

Proceedings

ISSN 1386-2316

Host and Sponsors

THE NEW YORK ACADEMY OF MEDICINE
HEALTHY CITIES. BETTER LIVES.

CENTRUM VEDECKO-TECHNICKÝCH INFORMÁCIÍ 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

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

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

LIBRARY OF CONGRESS
FEDLINK

GreyNet
Grey Literature Network Servic

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 Conference Proceedings
Eighteenth International Conference on Grey Literature : Leveraging Diversity in Grey Literature - New York, NY USA, on 28-29 November 2016 / compiled by D. Farace and J. Frantzen ; GreyNet International, Grey Literature Network Service. - Amsterdam : TextRelease, February 2017. -178 p. - Author Index (GL Conference Series, ISSN 1386-2316 ; 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), NTK (CZ), and TIB (DE) are Corporate Authors and Associate Members of GreyNet International. These proceedings contain the full text of 16 conference papers presented during the two days of plenary and poster sessions. The papers appear in the same order as in the conference program book. Included is an author index with the names of some 40 contributing authors and researchers along with their biographical notes. A list of more than 50 participating organizations as well as sponsored advertisements are likewise included.

ISBN 978-90-77484-30-2

A terminological "journey" in the Grey Literature domain

Roberto Bartolini, Gabriella Pardelli, Sara Goggi,
CNR, Istituto di Linguistica Computazionale, "Antonio Zampolli"
Silvia Giannini and Stefania Biagioni,
CNR, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Italy

1. Introduction

"It is by means of terms that the expert usually transfer their knowledge and again through terms scientific communication reaches the highest effectiveness. Therefore we can assert that terminology - in the sense of a set of representative and domain-specific units - is necessary for representing and connecting specialized fields as well as any attempt to represent and/or transfer scientific knowledge requires, more or less extensively, the use of terminology." (Cabré, 2000). "When we read the articles or papers of a particular domain, we can recognize some lexical items in the texts as technical terms. In a domain where new knowledge is generated, new terms are constantly created to fulfill the needs of the domain, while others become obsolete. In addition, existing terms may undergo changes of meaning…" (Kageura K., 1998/1999).

Specialized lexicons are made up of the terms which are specific to each field of knowledge, «a subset which is distinct but not separated from the common language» (Cassese, 1992): it is usually difficult to extract the relevant domain-specific terminology, meaning to discern terms which belong to a specialized glossary from those belonging to the common dictionary.

The interest in the study of terminology and the "truth" contained in the above definitions has led us to make a "journey" in the Grey Literature (GL) domain in order to offer an overall vision on the terms used and the links between them.

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

2. GL Corpus and method

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

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

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

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

- transformation of the original document, in our case in Word format, in plain utf-8 format text;
- use of some pre-existing software tools for:
  1. sentence splitting: dividing the text into sentences
  2. word tokenization: splitting sentences into words

[1] Kindly authorised from Greynet International, http://www.greynet.org/.
[2] CNR stands for National Research Council, Italy, https://www.cnr.it/.

3. lemmatization and morphological analysis (part of speech tagging)
4. basic syntactic analysis (chunking: dividing the sentence into non recursive constituents)
5. parsing with the **Ideal** dependency parser, a rule-based system whose specific rules were developed for both Italian and English. This tool was developed specifically for the MAPS project, being the most important part of the NLP pipeline.

The output of the chunking phase produces an intermediate annotated document preliminary to terminology extraction performed by the **Ideal** parser, which relies on rules, written in an ad-hoc language, designed to extract all simple and complex noun phrases in the text[3].
Terminological extraction is in turn necessary in order to be able to correctly index the document in the document base to be later semantically searched. The **Ideal** extraction tool takes the chunked text as input, containing all the required morpho-syntactic information.
The output of this terminology extraction pipeline is a set of terms in a standardised *Jason* format.
Within our corpus made of GL articles, this NLP tool – already used as semantic engine within the MAPS project (GL16 and GL17 papers) - extracts a list of single (monograms) and multi-word terms (bigrams and trigrams) ordered by frequency with respect to the context.

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

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

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

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

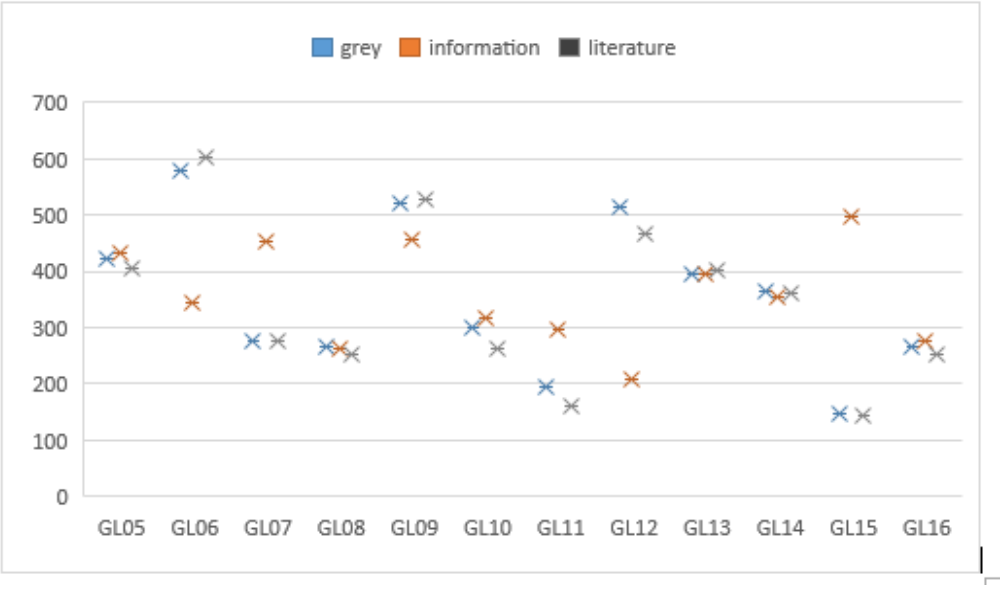
In Appendix 1, we can read the terms occurring most frequently in the high segment: the only two monograms which consistently remain in this segment and in all GL glossaries are “Literature” and “Research”.

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

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

3.2 Mapping

We started the terminological mapping from observing the term that occurs most frequently in the entire corpus: “information” and the two terms more closely related to the context, "grey" and "literature".
Graph 1 shows that the terms “grey” and “literature” have the highest frequency in GL6 (2004) and the lowest in GL15 (2013) while the term “information” has the peak in GL15 (2013) and the bottom in GL12 (2010).



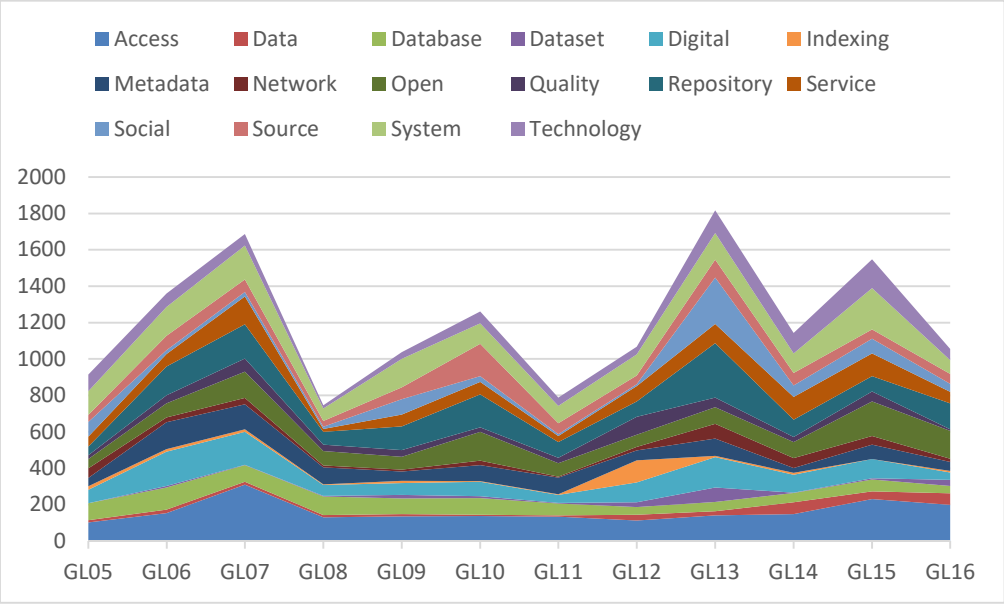
Graph 1 – “Grey”, “Information”, “Literature” – Trend over the years

As expected, the bigram “grey literature” is the most used with 2816 frequencies in the entire corpus while the bigrams “grey material” (66 occurrences) and “grey document” (98 occurrences) are not present in all GL proceedings and their frequencies are much lower. The bigram “grey documentation” only appears in GL5, GL9 and GL16. Among the other bigrams we find: “grey medical”, “grey document”, “digital grey”, “grey publisher”, “grey content”, “grey object”, “grey resource”, “grey collection”. Amongst the trigrams we have: “grey literature collection”, “grey medical literature”, “grey literature community”, “grey literature document”, “grey literature repository”, “grey literature resource”, “grey literature material”, “grey literature typology”, “grey literature report”, “digital grey literature”, “grey literature field”, “grey literature problem”.

In addition to the pair "grey literature" the term "literature" appears in the following bigrams: "medical literature", "conventional literature", "type literature", "literature repository", "literature collection", "repository literature", "use literature", "access literature", "journal literature", "literature collection", "conference literature", "trade literature", "definition literature", "literature document", "topic literature", "literature repository", "literature review". The trigram "non conventional literature" is only used in GL7 and GL14 terminology. Excluding the already mentioned trigrams in which "literature" appears associated with grey, there are: "bibliographic control literature", "new generation literature", "scholar information literature", "digital curation literature", "digital repository literature", "strategic collection literature", "literature network service", "web-based dissemination literature", "digital library literature".
The most common bigrams with the term "information" are in GL15: "Information object" is the top term (39 occurrences) while the bottom one is "Information retrieval" (17 occurrences) in GL14. Amongst the others we find: "information system", "source information", "scientific information", "information interaction", "information system", "internet information", "access information".

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

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

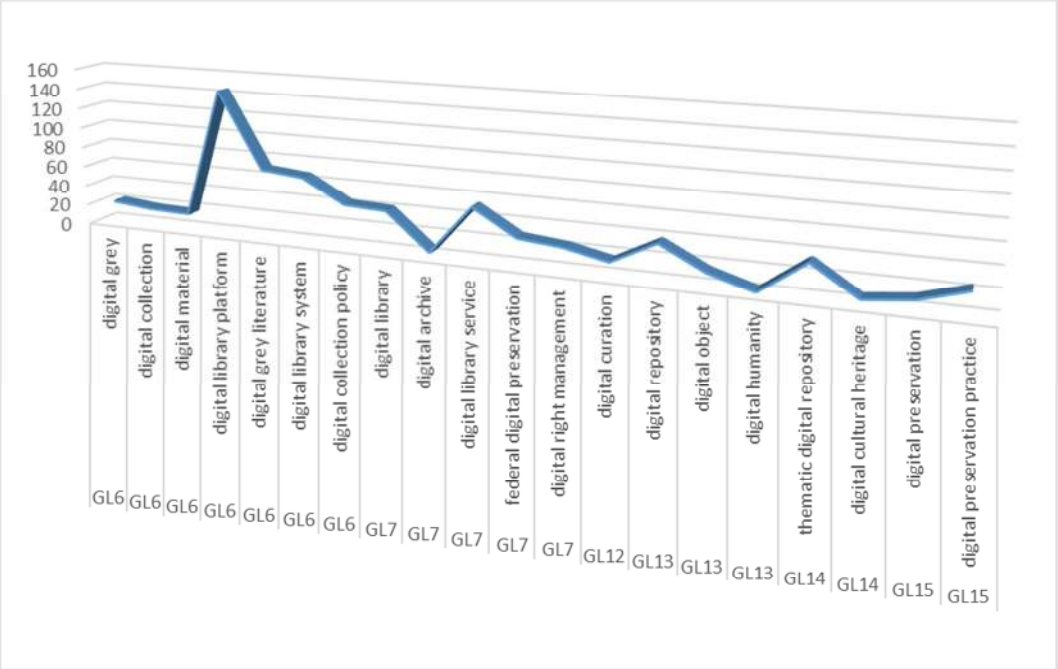


Graph 2. – Selected terms

Graph 2 shows the trend of the selected terms over the years: it is clear that - with the exemption of "indexing" and "dataset" – all of them are occurring in each GL glossary. Generally, there are monograms which seem to be constantly used and therefore their trend over the time is stable (e.g. access, database and digital) while the vast majority of terms alternate high and low frequency peaks.

The monogram “access” has the highest number of occurrences (1928) and “dataset” the lowest (196); amongst the most frequent terms, also “system”, “repository”, “open” and “digital” can be found.

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



Graph 3. “Digital” – bigrams & trigrams

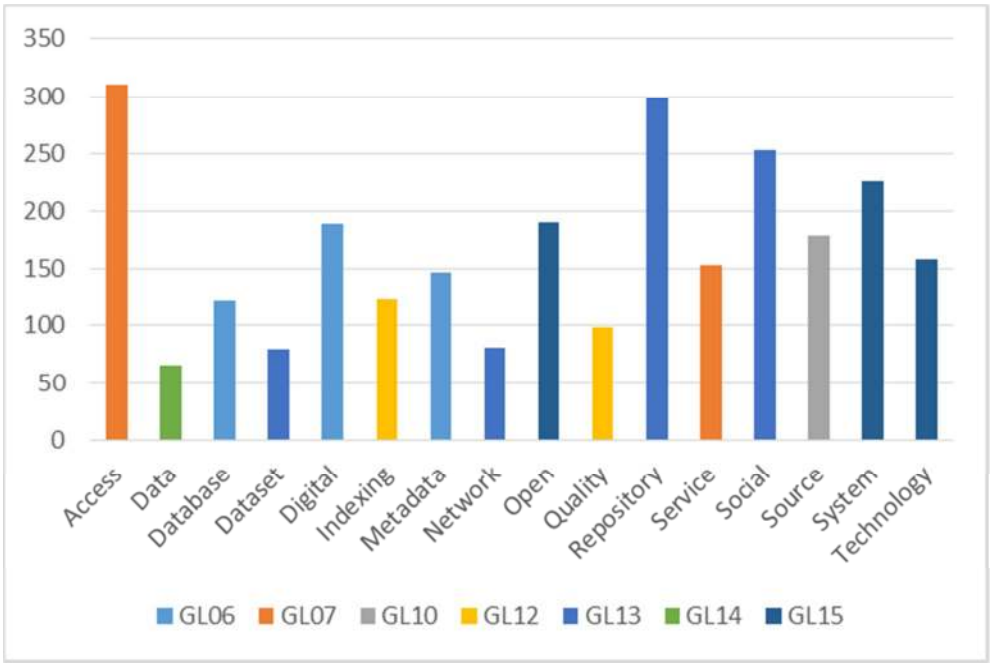
The term “data” shows the highest frequency as a bigram, “big data”: it introduces the set of problems about gathering, managing, representing and accessing huge volumes of data which are dynamically generated from various sources. The bigram first appears in GL14 (2012) and then again the next year thus witnessing the community’s immediate appeal for the subject; as a trigram is mostly in combination with terms like “discovery”, “service” and “product”.
The term “database” cannot be neglected too: it is used and reused in different contexts as a synonym of an archive of structured and connected data and occurs in the entire time-span associated with various semantic values: "citation databases", "technological databases", "grey literature database". “Database” and “metadata” register the highest number of occurrences in the papers of the GL6 (2004) conference, exactly like “digital”.
The exam of the term “metadata” points out its presence in all GLs in the mid-range segment and already in GL6 (2004) and GL7 (2005) in the high frequency one: these are years when there is a considerable discussion over themes as standards, document

management and fast information retrieval systems and the term is often found in association with nouns and adjectives which highlight the importance of properly describing and organizing documentation in the field of digital libraries: “metadata schema”, “navigational metadata”, “descriptive metadata”, “metadata format”, “metadata harvesting” and again “Dublin Core metadata” (the highest percentage), “right management metadata”, “standardized metadata schema” and “metadata quality control”. In GL7 and GL10 “metadata” is often combined with standards and schemes such as Dublin Core and Cerif.
The term “dataset”, in the two variations “dataset” and “data set”, appears in 2005 and remains constantly present in the following editions forming the most frequent bigrams “scientific dataset” and “dataset archive” while is more occasionally associated with “accessibility”, “collection” and “management”.
Already in GL6 (2004) the GL community faces the need to examine the quality of information available on the web: the term “quality” is repeatedly associated with “assessment” or “control”, in particular in the forms “metadata quality control”, “quality assessment metadata”, “quality information”, “quality performance”, “high-quality information” and “metadata quality certification”.
Another interesting term is the adjective “social”. Although we found the topic “Social Networking” only in GL13 (2011), this bigram is in use since GL7 (2005) and the monogram “social” is steadily used in the GL lexicon since GL5 (2003). The adjective “social” is combined with a large number of nouns to form bigrams, trigrams and strings of words with a strong semantic impact. In GL8 the multi-word expression "social network" appears, as a “neologism”, in the GL lexicon. Other linguistic forms emerging from the terminology are linked to the same concept: "virtual social networking”, "social networking tools", "social networking sites", "new social networking technologies". The MWE “social media” was “born” instead in the GL9 conference (2011).

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

3.3 GL Conference topics

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



Graph. 4 – Terms and Topics

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

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

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

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

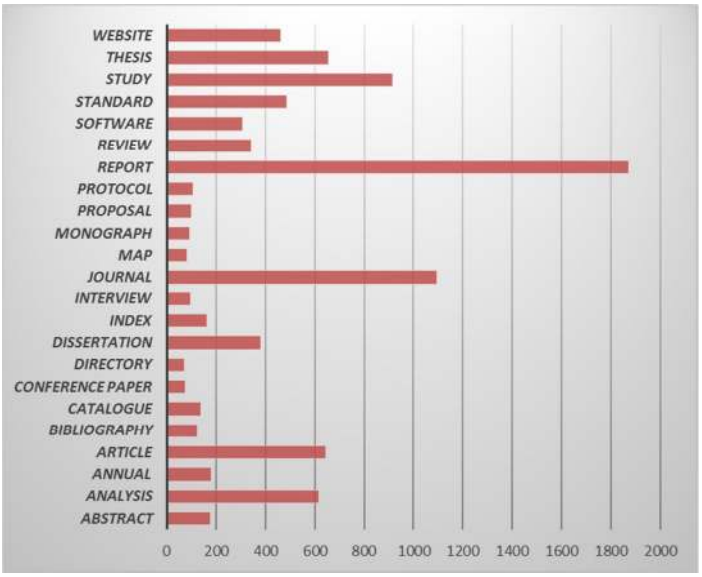
Finally three topics are dedicated to "open access" in GL conferences: "Open Access to Grey Resources", "Open Access and Wealth Creation" and "Open Access to Research Data" (GL16 - 2014).

3.4 Types of documents

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

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

From this perspective, the presence of the *Vocabulary* terminology within our corpus has been verified: the table in Appendix 3 lists the terms appearing in the various GLs and their quantitative consistency. This table is ordered by frequency and the results – in terms of the most occurring terms - are therefore very clear.
In the attempt of making a partition of this list – however arbitrary - we can circumscribe a first area where the frequencies decrease from 1871 to 307 and the terms retrieved are: "report", "journal, "study", "thesis", "article", "analysis", "standard, "website, "dissertation, "review", "software". In the intermediate area where frequencies decrease from 196 to 30 (with a remarkable gap between the last occurrence of the first zone and the first occurrence of the last zone) the terms found are the following: "dataset", "annual", "abstract", "questionnaire", "index", "patent", "catalogue", "bibliography", "annual report", "protocol", "proposal", "interview", "monograph", "map", "conference paper", "preprint", "directory", "newsletter", "manual", "bulletin", "curriculum", "poster".
The terms used with the lowest frequency (from 24 to 1) for describing the types of documents are: "brochure", "proceedings", "government document", "glossary", "memorandum", "handbook", "timeline", "announcement", "conference program", "essay", "press release", "chronicle", "leaflet", "course material", "informative material", "normative document", "anthology", "research plan", "syllabus", "tertiary source", "corporate literature", "habilitation thesis", "image material", "legal document", "guidebook", "technical documentation".



Graph 5. – Types of documents retrieved in all GLs

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

Conclusions

To conclude, this survey on the results of the information extraction process performed by the described NLP tool has been a sort of linguistic path in the past and present of the terminology used in GL proceedings with the goal of drawing a picture of the lexicon used by the GL community and thus contributing to get a deeper knowledge of the GL domain.
Many of the terms encountered cannot have synonyms because they reflect specific concepts devoid of the ambiguities peculiar to the common language. Some expressions

such as “grey resources” and “open access” or nouns as “library” and “repository” refer straight and univocally to the “documentary science”, that is they belong to a specific semantic field.

By adopting a diachronic point of view, a significant terminological stability can be noticed. However some terms have been pointed out as obsolete while others emerged as very up-to-date, the latter are those chosen for assembling studies in the same domain or even for labeling emerging fields of knowledge. This is the case, for example, of the bigram “electronic dataset” retrieved in 2004 and 2007 glossaries and then substituted by the bigram “digital dataset” in 2010 and 2014.

Examples could be endless but the size of the corpus had made necessary to delimit the study to a sample by choosing some of its parts and pertaining taxonomies.

In these last twelve years we have witnessed the establishment of new paradigms of scientific communication, the stunning development of information technology and the creation of new infrastructures for storing, preserving and disseminating scientific information. A fact clearly comes to light from this analysis: the grey literature field has a dynamic and cross nature, its community is sensible to technological innovation and proves to be able of keeping pace with the changes.

The lexicon adopted in the GLs’ scientific papers has confirmed that the “grey” community soon paid specific attention to topics like “open access”, “repository” , “digital objects” and “preservation”, just to cite a few. At the same time the almost stable use of a technical and specialized terminology over the time indicates the interest and the willingness to deepen the knowledge of some themes by reporting updates and novelties.

Lastly, this work must be considered a preliminary analysis of the GL corpus, a linguistic resources to be further investigated with different purposes and different tools.

Essential References

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

Appendix 1 – Frequency

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

Table 1

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

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

Table 2

Appendix 2 – GL Conference topics

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

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

Table 3

Appendix 3 - Types of documents

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

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

Table 4