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

A TERMINOLOGICAL "JOURNEY" IN THE GREY LITERATURE DOMAIN

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…" (Kageura K., 1998/1999).

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.

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.

HIGH, MEDIUM AND LOW FREQUENCY

TERMINOLOGICAL ANALYSIS

1. Identification of the monograms of ***high, medium and low frequency*** within the glossaries provided by the extraction → overview of the single terms used in the papers;

The study of the terms grouped according to their decreasing frequency allowed us to:
a) select some of the most frequently used terms;
b) examine the co-occurrences: bigrams and trigrams;
c) determine the variations between them.

2. Analysis of taxonomic chains in order to shed light on the usage of specific terms within the topics of the various GL conferences.

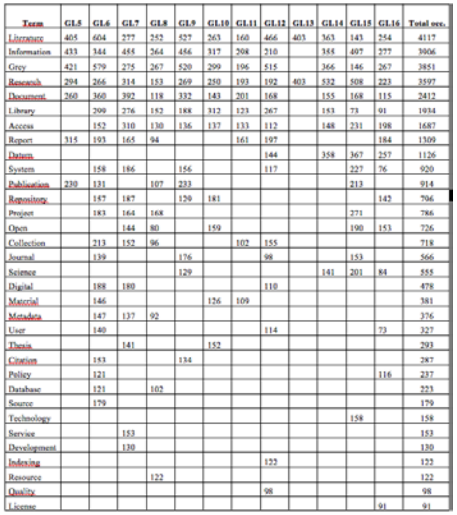
HIGH, MEDIUM and LOW FREQUENCY

In Tables 1 and 2 are grouped – respectively - the terms of the highest and medium segment of each GL Corpus.

For frequency segment of vocabularies we mean the organization of the words for decreasing frequencies, starting from the word with freq$_{max}$ and coming to those with freq$_{min}$, usually with only one occurrence (hapax).

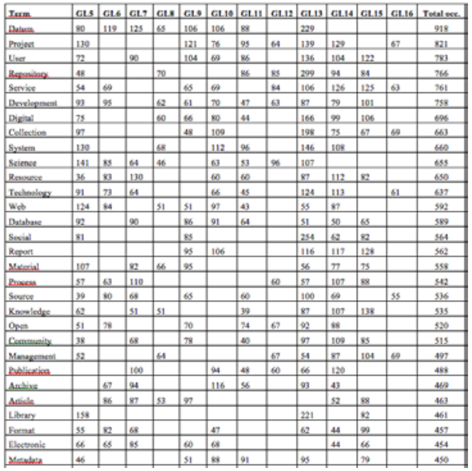
The results is that the highest percentage of terms is to be found in the lowest frequency segment: this applies to all GLs'.

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 GL05 followed by GL14.



Term	GL5	GL6	GL7	GL8	GL9	GL10	GL11	GL12	GL13	GL14	GL15	GL16	Total occ.
Literature	405	604	277	252	527	263	160	466	403	363	143	254	4117
Information	433	344	455	264	456	317	208	210		355	407	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		120	181						142	706
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. High segment



Term	GL5	GL6	GL7	GL8	GL9	GL10	GL11	GL12	GL13	GL14	GL15	GL16	Total occ.
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

....

Table 2. Medium segment

MAPPING

The mapping starts from the observation of the term that occurs most frequently in the entire corpus, which is "information", and from the two terms more closely related to the context, "grey" and "literature".

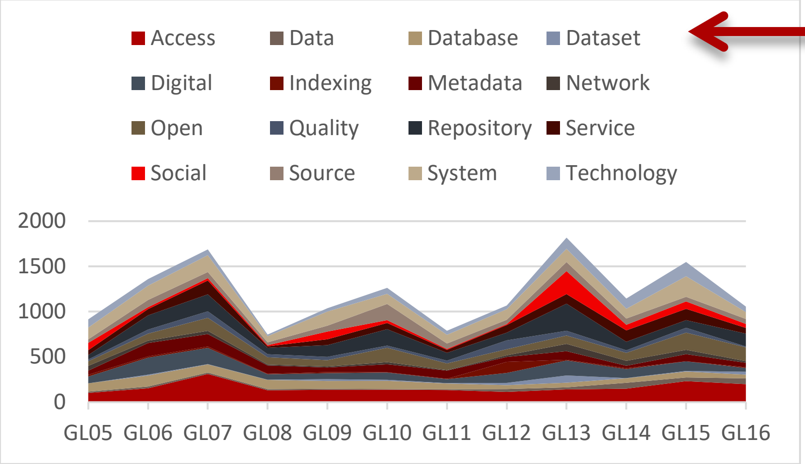
Graph 1 shows that the terms "grey" and "literature" have the highest frequency in GL6 (2004) and the lowest in GL15 (2013), while the term "information" has the highest in GL15 (2013) and the lowest in GL12 (2010).

The bigram "grey literature" is the most used (2816 occurrences in the 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 …

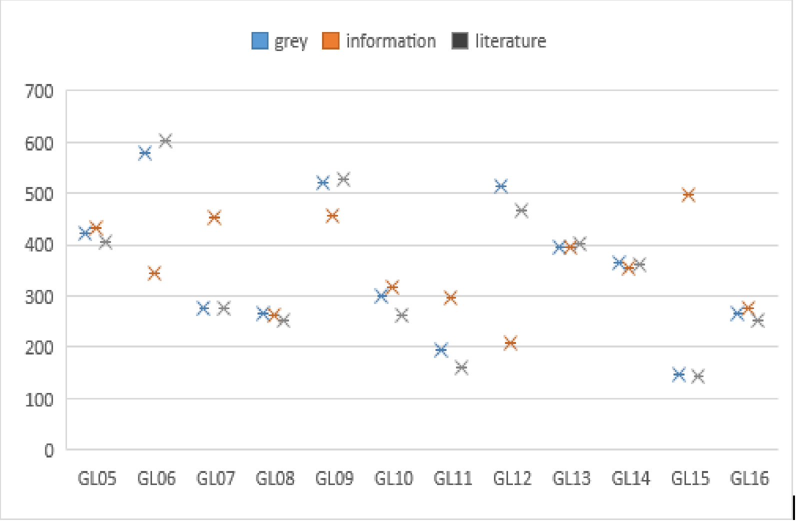
The most common bigrams with the term "information" are in GL15: "Information object" is the top term (39 occurrences) while the bottom is "Information retrieval" (17 occurrences in GL14).

As trigrams: we have "Open Source Information" as top term with 228 occurrences and "Heterogeneous Information Object" as bottom term with 56 occurrences…

Graph 2. – Selected terms - Trend over the years



Graph 1. - grey, information, literature – Trend over the years



Given the size of the corpus and its chronological extension, the terms have been selected according to their technical nature and mainly with respect to a very dynamic and cross field: **ICT (Information and Communication Technology).**

Graph 2 shows the trend of the selected terms over the years: it is clear that – with the exemption of "indexing" and "dataset – all the terms 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. come "access", "database" and "digital") while the vast majority of terms alternate high and low frequency peaks.

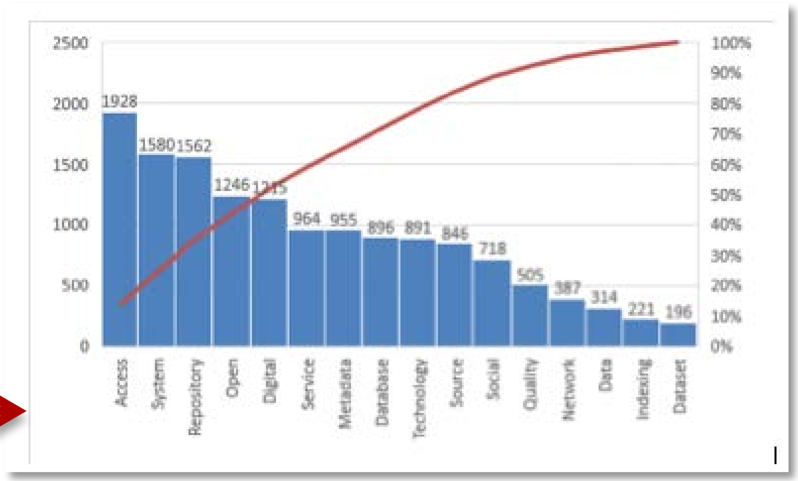
MAPPING

Graph 3. shows the total amount of occurrences for each selected monogram.

Highest number of occurrences: “access” (1928)
Lowest number of occurrences: “dataset”   (196)

Amongst the highest, also “system”, “repository”, “open” and “digital” can be spotted.

Graph 3. – Total occurrences



MAPPING - Examples

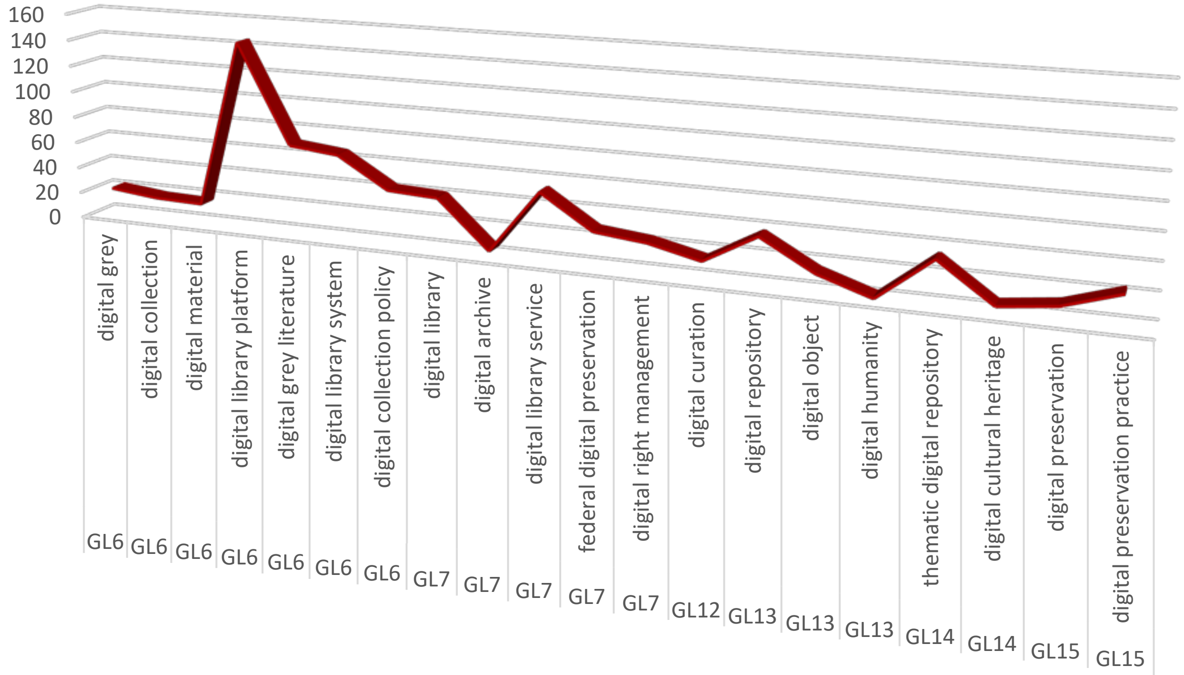
The analysis starts from one of the most versatile adjectives of the corpus: **“DIGITAL”**

The nouns, verbs and multi-word expressions (MWEs) combined with the term “digital” immediately disclose the technological nature of GL community: **infrastructure, platform, system, software, network**.

The occurrences **“digital humanities”** and **“culture heritage”** characterize the fields of knowledge which usually require an expertise crossing between computer science and human and social sciences.

MAPPING

DIGITAL - Graph 4. “digital” – bigrams & trigrams



“DIGITAL”

As for bigrams and trigrams: “*digital library*” and “*digital library platform*” are the most frequent MWEs;

Some lower occurrences:

“infrastructure”, “platform”, “system”, “software”, “network” show the technological nature of GL community;

“digital humanities” and *“culture heritage”* spot activities crossing human/social sciences and computer science.

Among bigrams: “digital library” appears in 2005 (GL7). The community does not ignore themes such as “digital preservation” which appears in 2013 (GL15) and even uses the trigram “digital preservation practice”.

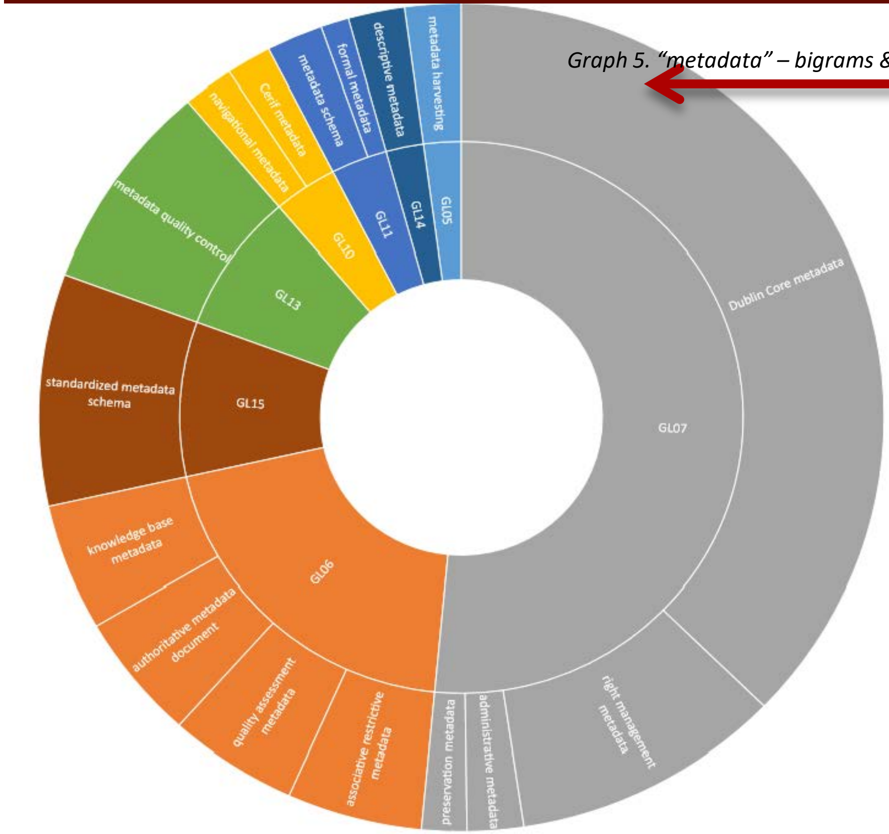
Among trigrams: “digital library platform” has the highest frequency in 2004 (GL6). In 2005 (GL7) “digital library service” can often be found and appears as “ thematic digital repository” in 2012 (GL14).

Within 2004 (GL6) and 2005 (GL7) glossaries, “digital” displays the highest frequencies in the two forms “digital library” and “digital library platform”.

Since GL13 the expression “digital repository” tends to substitute “digital library” though it does not have the overall meaning of handling of a document life cycle which “digital library implies.

METADATA & QUALITY

MAPPING



Graph 5. "metadata" – bigrams & trigrams

"METADATA"

It can be found in all GLs in the medium segment and in the high segment already in GL6 and GL7, when discussion on standards and document management appears;

As for bigrams and trigrams, *"metadata"* comes with nouns and adjectives which highlight the importance in the Digital Library field: *"navigational metadata", "descriptive metadata", "metadata format", "metadata harvesting", "formal metadata", "metadata schema" ….*

In 2005 (GL7) there are topics on "Right management metadata" and "preservation metadata" and administrative metadata".

In GL7 and GL10 the term is associated with specific standards such as Dublin Core and Cerif.

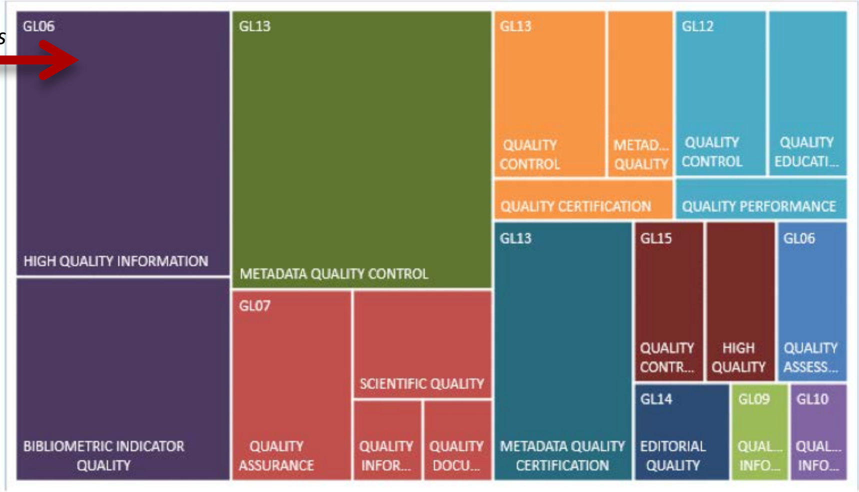
"QUALITY"

Since 2004 (GL6) there is the necessity of testing the quality of information available on the web and the term can be found:

as a bigram:
"quality assessment", "quality control", "quality information", "quality performance".

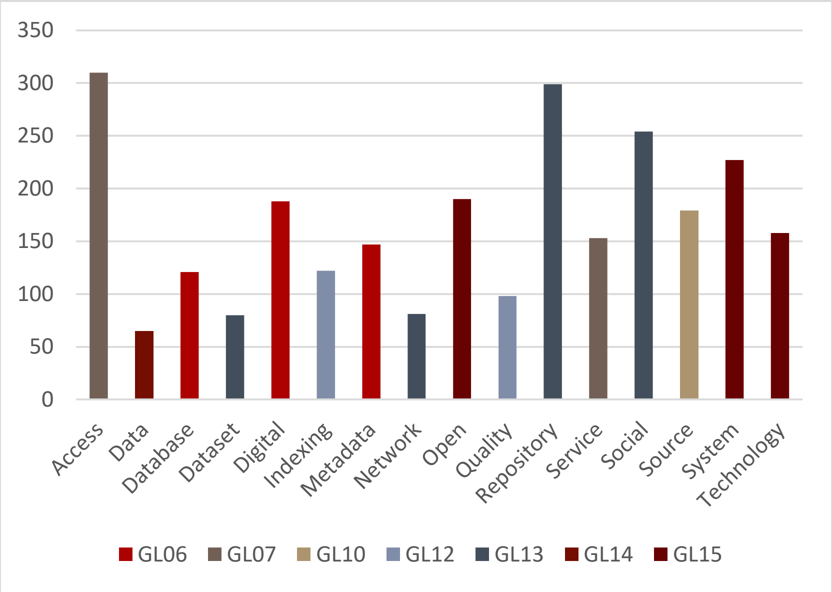
as a trigram:
"metadata quality control", "quality assessment metadata", "high-quality information" and "metadata quality certification".

Graph 6. "quality" – bigrams & trigrams



GL TOPICS

Graph 7 – Terms and topics



GL TOPICS

The term “repository” can never be found amongst the topics of the conference in its singular form.

It is very common in the plural form “repositories” since GL6 and then in GL7, GL8, GL10 and GL11.

“Repository”, together with “dataset”, “network” and “social” are the terms with the highest number of occurrences in the GL13 when the conference topics were:

“Social Networking”, “Special Collections”, “Open Access” and “Wealth Creation”, “Data Frontiers”.

Although we found the topic “Social Networking” only in 2011 (GL13), this bigram is in use since 2005 (GL7) and the monogram “social” is steadily used since 2003 (GL5).

In GL8 the multi-word expression "social network" appears, as a “neologism”, in the GL lexicon.

GL TOPICS

The flow of themes discussed in these years is represented by the topics appearing in the twelve Call for Papers. Therefore some terms have been selected and then analyzed in relation to the topics of all GL conferences following two steps:

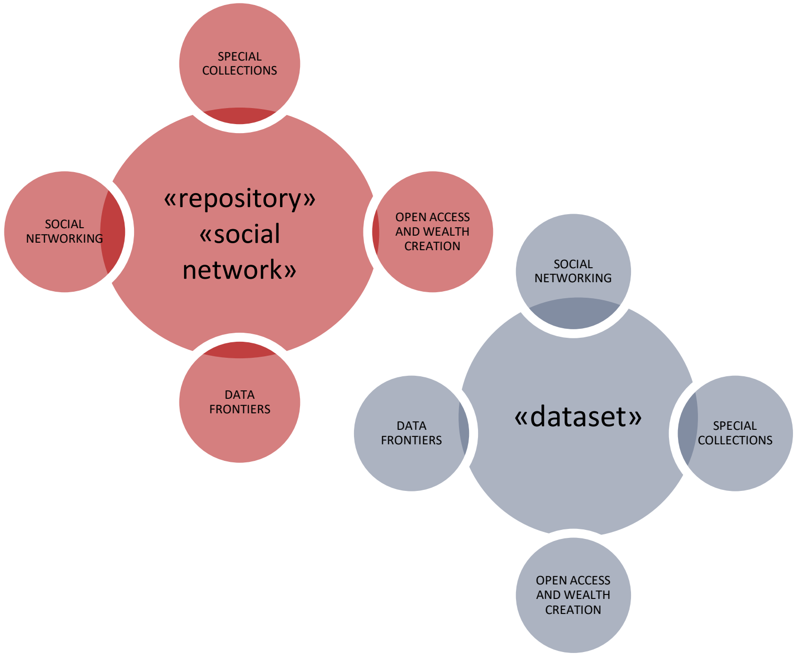
- retrieval of terms with the highest frequency
- their comparison to the conference topics

Graph 7 shows that the frequency peaks are limited to the GL6, GL7, GL10, GL12, Gl13, GL14 and GL15 editions while the other conferences are excluded.

Highest peaks are in:

- GL7 with the term “access” and
- GL13 with “repository” and “social”

REPOSITORY, SOCIAL NETWORK AND DATASET



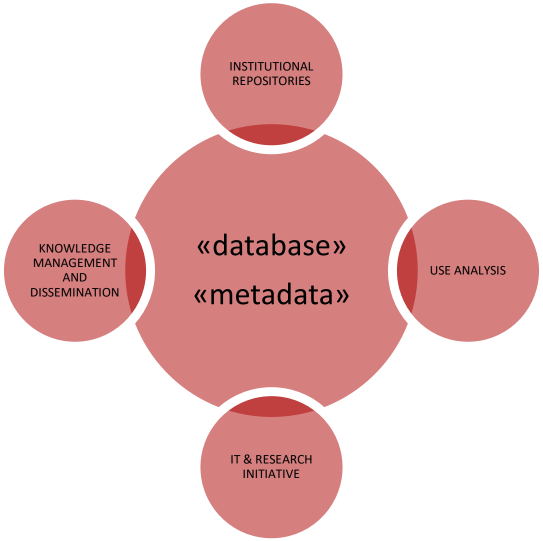
GL TOPICS

The highest number of occurrences of the terms “digital”, “database” and “metadata” is in GL6 (2004) which had the following topics:

“Institutional Repositories”, “Use Analysis”, “IT & Research Initiative”, “Knowledge Management and Dissemination”, “Collection Development and Resource Discovery”.

It is interesting to notice that the monogram “database” never appears among the conference topics and “metadata” is to be found only once, in 2006 (GL8).

DATABASE AND METADATA


INSTITUTIONAL REPOSITORIES
KNOWLEDGE MANAGEMENT AND DISSEMINATION
«database»
«metadata»
USE ANALYSIS
IT & RESEARCH INITIATIVE

GL TOPICS

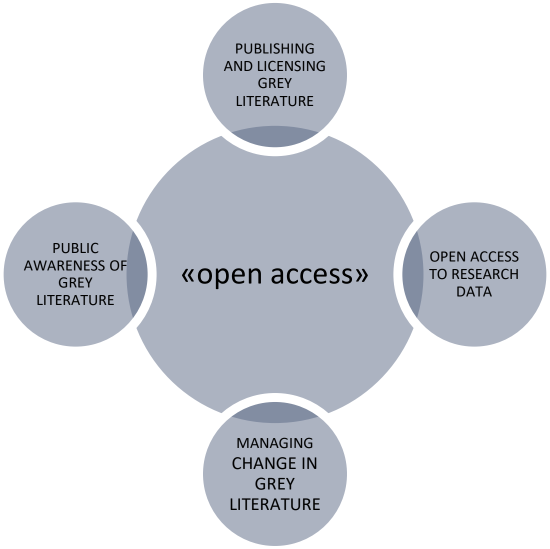
The bigram “open access” is a constant feature in the grey literature lexicon. It is used since the far GL5 (2003) in the two graphic variations “open access” and “open-access” that live together in some GLs’.

We found three topics dedicated to “open access” in GL conferences:

“Open Access to Grey Resources”, “Open Access and Wealth Creation” and “Open Access to Research Data”.

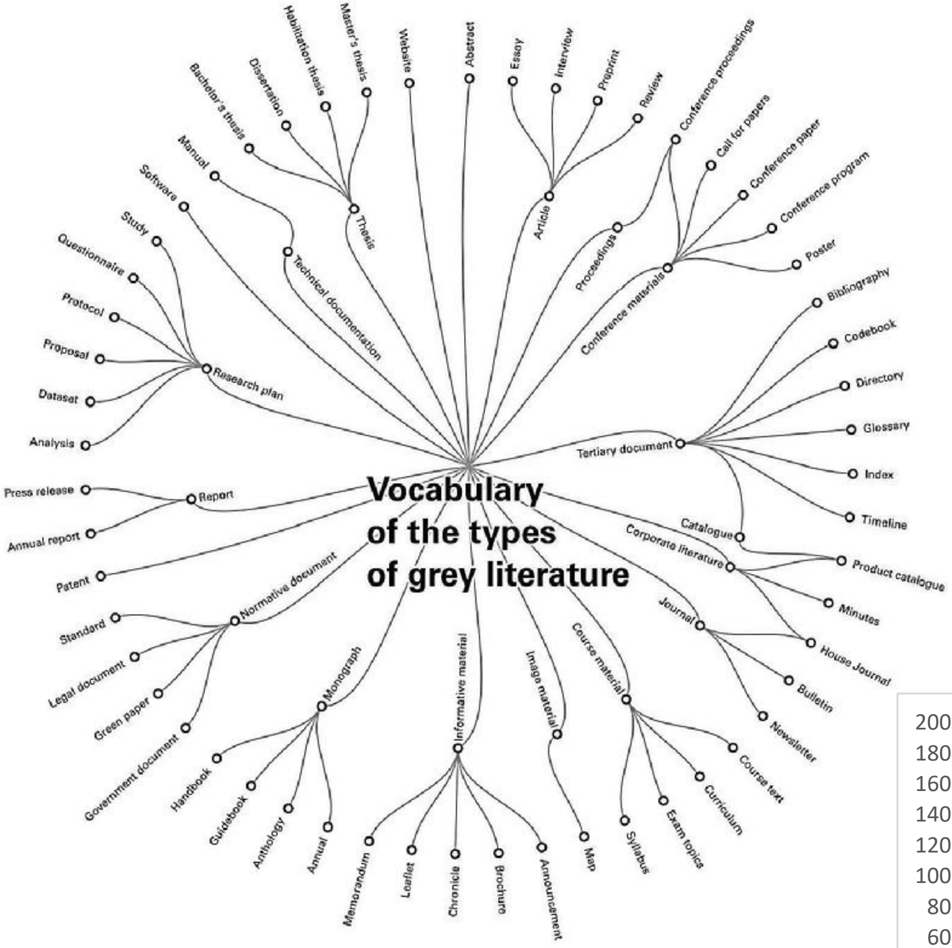
The peak of the highest frequency is reached with *“Open Access to Research Data”* in 2014 (GL16).

OPEN ACCESS


PUBLISHING AND LICENSING GREY LITERATURE
PUBLIC AWARENESS OF GREY LITERATURE
«open access»
OPEN ACCESS TO RESEARCH DATA
MANAGING CHANGE IN GREY LITERATURE

GL TOPICS

TYPE OF DOCUMENTS



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 automatically verify the actual correspondence between the term and its context.

TYPE OF DOCUMENTS: *Graph. 8 – Type of documents*

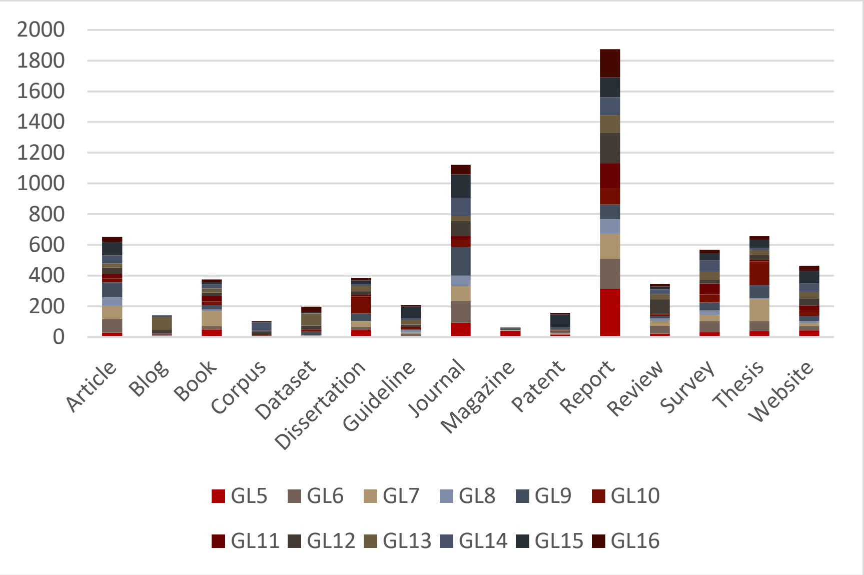
The last step started by comparing the terms extracted from the *Vocabulary of the types of Grey Literature* (2011) with those extracted from our corpus.

A few outstanding examples:

✓ the monogram "journal" is present every year with a total of 1121 occurrences;

✓ the monogram "report", one of the most widely used GL document type, is also present every year with 1893 occurrences:

✓ the monogram "thesis" appears in every edition as well with 659 occurrences, although only a specific topic is devoted to this document type: "Theses and Dissertation" in GL7 (2005).



CONCLUSIONS

This survey 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. Examples could be endless but the size of the corpus makes the analysis of data rather a difficult challenge in terms of defining the subset of the lexicon and the taxonomies to take into account.

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.

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 GL 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 paid soon 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.