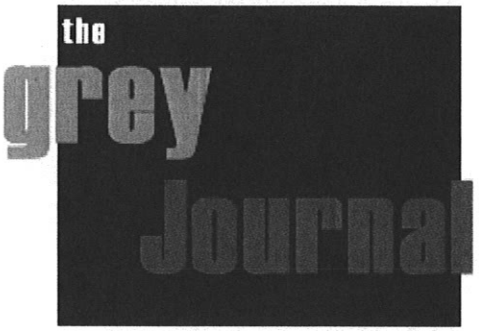
ISSN 1574-1796

An International Journal on Grey Literature



Autumn 2019, Volume 15, Number 3

'ADVANCING OPEN SCIENCE AND GREY LITERATURE'

GreyNet

www.textrelease.com Grey Literature Network Service www.greynet.org

Contents
'Advancing Open Science and Grey Literature'

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

1. Introduction

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.
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.

The following quotation by Bronson gives a good description of 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)

While the World Wide Web keeps on growing, the development of ever more sophisticated search tools within the universe of public and private infrastructures allows to optimize the users' approach to technology: a new generation of web users – such as the so-called "millennials" – is exponentially connected for getting access and share information on social networks.
In 2005 Bettelle states: "Most searchers –and linguists may be no exception – are instead incredibly lazy, generally typing in a few words and expecting the engine to bring back perfect results, ignoring that it is only the act of offering more data in the query that often dramatically improves the results "(Battelle 2005: 23-25). Nowadays it is possible to retrieve knowledge from the web also by means of query logs, linguistic elements which are useful for monitoring a wide range of information.

This Corpus – called GSGCorpus (Global Science Gateway Corpus) – has been processed with Natural Language Processing (NLP) tools: it talks the *web language*, made up of terms originating from the most various domains and styles. 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.

---

* First published in the GL20 Conference Proceedings, February 2019.

[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.

2. Methods and Tools

This project includes eight months of query logs[3] registered between July 2017 and February 2018 for a total of 445,827 queries.

The preliminary phase has essentially dealt with the huge amount of non-relevant information, the so-called *noise* which had to be filtered and eliminated.

Therefore, 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.

Since the corpus is made up 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).

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.

In the following paragraph, it is described the process of information retrieval from a rich digital catalogue of queries.

2.1 NLP Analysis

The free information extraction from the GSGCorpus measures the frequency of the words contained in the corpus; examines the lexical variety of the queries and finally focuses on a set of terms to build a micro-ontology. This extraction was preceded by the cleaning process already described above and of which some examples are listed here:

- Graphical variants:
  - *(micro-fluidic* "micro fluidic*" microfluidic)*
  - *(trypodes basapyrazilique pdf, trypodes Base pyrazilique PDF)*
  - *(Adam Smith, Adams, SM.)*
- Queries in languages other than English:
  - *(alimentos proteicos, Alteracion proteica, Entrainement isometrique, 1. Peste-des-petits-ruminants virus fusion protein F) gene, complete cds, Trypanosoma cruzi: contribution Ã l'identification de substances chimiques et naturelles ayant une activitÃ© trypanocide, alfa1 fetoproteina, hrabanus maurus lehrer, abt und bischof, anabolismo proteina, AND Tanztheater Pina Bausch: Spiegel Gesellschaft, polilì tico, Espanìfa)*
- Duplicates:
  - *(spinal injury) electrical nerve electrical stimulation)*
  - *(spinal injury) electrical nerve electrical stimulation)*
- Alphanumeric strings and IP addresses:
  - *AND colegio MÃ©xico hazaÃ±a*
  - *TÃ©cnicas*
  - *http://www.repositoriodgb.buap.mx:2095/*
  - *http://www.scielo.org.mx/pdf/tca/v6n3/v6n3a2.pdf*

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

- *http://www.sciencedirect.com/science*
- *http://www.sciencedirect.com/science/article/pii/S0019850199001133*
- *http://www.sciencedirect.com/science/article/pii/S0048712002732908*
- *http://www.sciencedirect.com/science/article/pii/S0176161712001848*
- *http://www.sciencedirect.com/science/article/pii/S0211563811001684*
- *http://www.sciencedirect.com/science/article/pii/S0308814614003628*
- *http://www.sciencedirect.com/science/article/pii/S1632347500719722*
- *http://www.xvideos.com/video27283249/jynx_maze_anal_banged_on_bangbros_chongas_in_1080p_ch13211_*
- *https://AsPredicted.org/wsxx7.pdf*
- *https://cirworld.com/index.php/jssr/article/view/3380*
- *https://doaj.org/article/c4a86ed60b7a4961baf52e2b45951d65*

➢ Disambiguation:

- *AND terapeuta errores Jaquelin cortez*
  AND terapeuta errores jaquelin Fortes
- *sandra massoini*
  sandra massoni
- *bio pelicula [spanish]*
  bio peliculua [termine inesistente]
- *social media reslut [termine inesistente]*
- *facebook adverstis [termine inesistente]*
- *social nedia marketing [termine inesistente]*

➢ "Empty" words like articles and prepositions.

As for the most frequent words, the extraction from the GSGCorpus has provided:

I. the decreasing frequency of monograms like nouns, adjectives and adverbs, that is the number of occurrences of each word;
II. the decreasing frequency of bigrams and trigrams and the number of occurrences of each of them in the corpus.

On the basis of their frequency, monograms have been divided in three areas depending on their frequency: high, medium and low. In the highest there are a very few words, while in the lowest there are many but with an irrelevant number of occurrences and the presence of *hapax legomena*[4]. Table 1 and Figure 1 show the twelve most recurrent monograms and how the adjective <social> ranks first with 2412 occurrences.

TERMS	# OCCURRENCES
SOCIAL	2412
MARKETING	1714
MANAGEMENT	1682
EFFECT	1578
MEXICO	1160
SCIENCE	1135
EDUCATION	1122
BUSINESS	1054
CHILD	994
PSYCHOLOGY	916
HEALTH	910
RESEARCH	906

Table 1– Twelve most recurrent monograms

As shown in Figure 1, the most relevant nouns in the corpus are: <business>, <child>, <education>, <effect>, <health>, <management>, <marketing>, <Mexico>, <psychology>, <research>, <science>. The only adjective retrieved is <social>.

[4] In corpus linguistics, a ***hapax legomenon*** (/ˈhæpəks lɪˈgɒmɪnɒn/ also /ˈhæpæks/ or /ˈheɪpæks/;[1][2] pl. *hapax legomena*; sometimes abbreviated to *hapax*) is a word that occurs only once within a context, either in the written record of an entire language, in the works of an author, or in a single text.

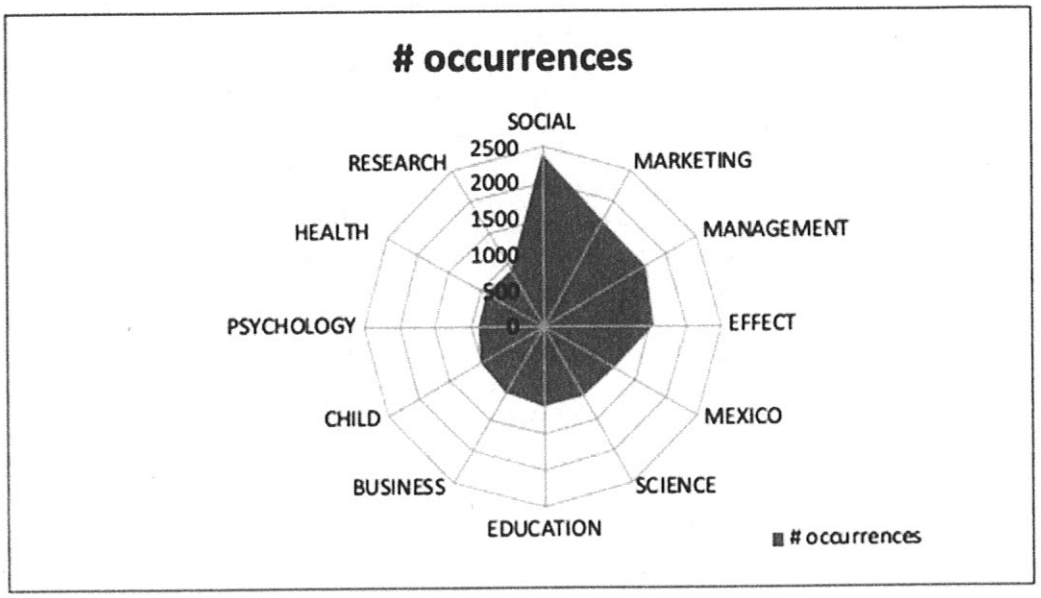


Figure 1 – Most frequent words

Figure 2 shows the bigrams with the higher frequency in the Corpus.

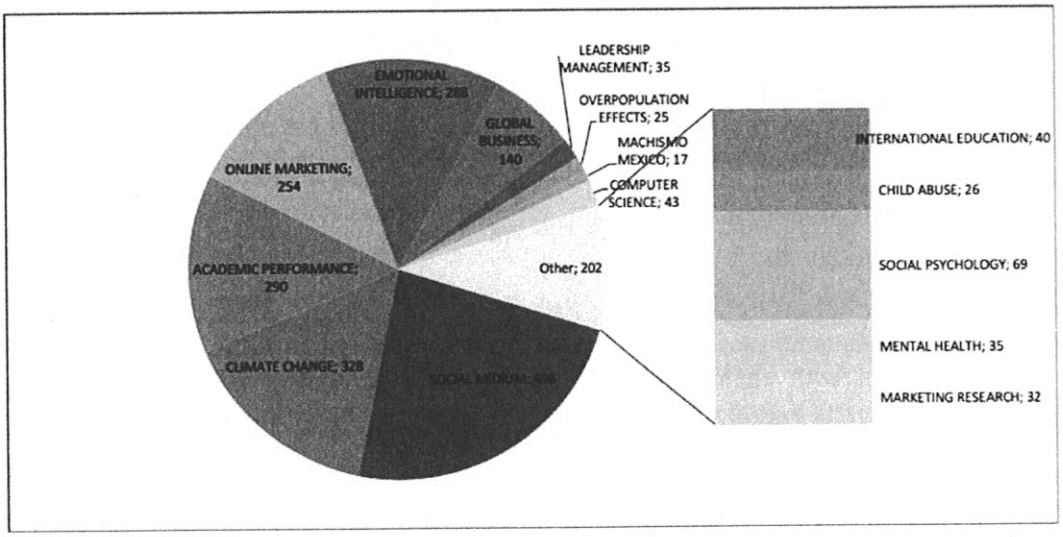


Figure 2 – Most frequent bigrams

3. A case study

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.* A case study has been carried out involving medicine, psychiatry and '*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.*

NLP analysis allowed to browse the corpus through the most and less queried terms: once *social* has been identified as the most frequent one, the analysis was channeled into '*social media*' and the pertinent contexts.

Ontology-based extraction: enrich the domain; retrieve each occurrence of those terms contained in the ontology by using a search engine.

In particular:

✓ Some low-frequency terms (hapax) carry a negative connotation[4], similarity and diversity. in relation to the use of 'social media';
✓ An analysis of negative connotations in connection with child/children, is further investigated.

3.1 Results

The topics concerning social media selected from the corpus of queries (Figure 3):

- The terms of the queries which reveal who are the subjects involved:
  - *The celebrities [The term "celebrities" connote people according to the fame and public attention accorded by the mass media]*
  - *The influencers [The "influencers" can be defined as internet users who have established a relevant number of virtual relationships]*
  - *The Millennials [The Millennials - or generation Y - are those born in full digital revolution]*
  - *The students*
  - *The teenagers*
- The terms of the queries which follow the expression <social media **cause**> with a negative polarity:
  - *anxiety*
  - *depression*
  - *dietary diseases*
  - *disadvantages*
  - *distraction*
  - *insecurity*
  - *sleep deprivation*
  - *teenager becomes antisocial*
  - *teenager neglect real world interaction*
- The terms of the queries which follow the expression <social media **impact on**>:
  - *anxiety*
  - *democracy*
  - *families*
  - *hospitality industry*
  - *maintaining relationships with others*
  - *sales*
  - *society*
  - *teens' lives*
- The terms of the queries which follow the expression <social media Social media **influence**>:
  - *brand loyalty*
  - *consumers purchase decision*
  - *criminality*
  - *fashion trends*
  - *teenagers' body image*
- The terms of the queries which follow the expression <social media **help>** with a positive polarity:
  - *business growth*
  - *maintain relationships with people*
  - *young people stay connected to distant people*
- The terms of the queries involving the words<child/children>: bigrams and trigrams with a negative polarity, see Figure 3:
  - *child abuse*
  - *child labor Indonesia*
  - *Child Marriage*
  - *Child psychiatry*
  - *depression child*

[4] "In linguistics, a **polarity** 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.

- *domestic violence children*
- *domestic violence children behavior*
- *internet safety children*
- *Obesity child*
- *obesity children*
- *Psychopatic assassin children*
- *punishment children*
- *The Evil Child Marriage*
- *violent video games child*

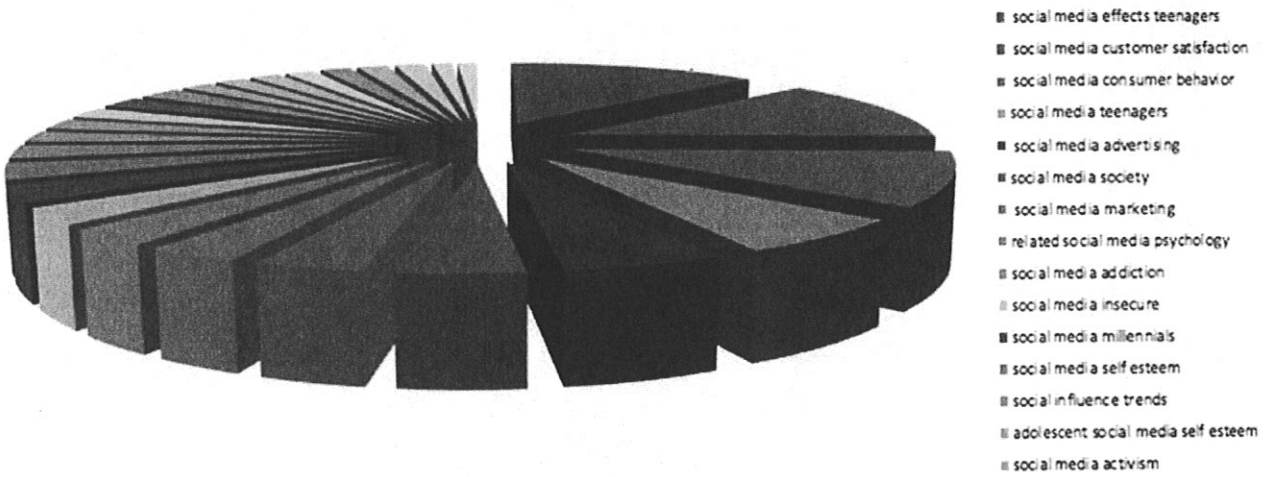


Figure 3 'Social media' occurrences

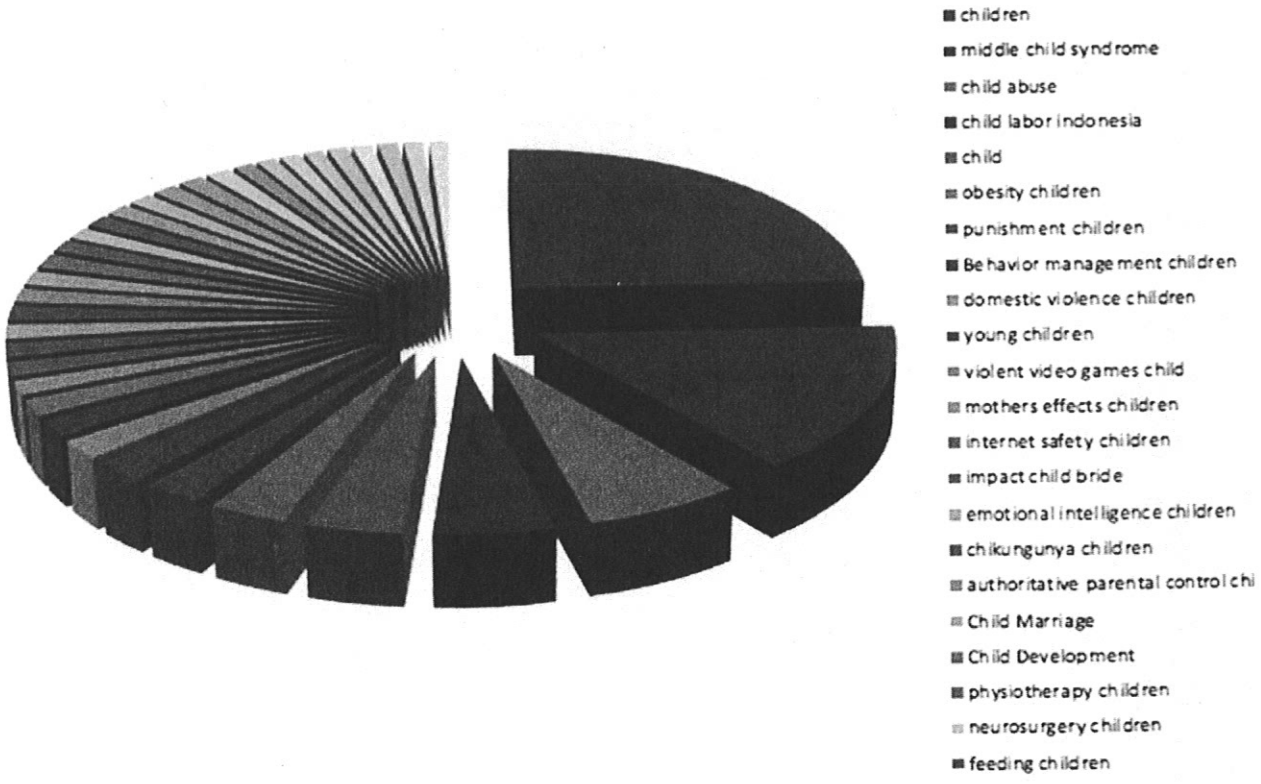


Figure 4 – Child/children context queries

4. Conclusion and final remarks

This work, which does not have any sociological purpose, provides two possible keys for interpreting the sample data of the GSGCorpus.

I. use of terms resulting from the recently exploded digital technology, in particular the contexts of the bigram <social media>[5] with either positive, negative or neutral polarity (see Figures 3, 5).

II. use of terms with a mostly negative polarity which have been retrieved from the contexts of the words <child/children>: examples are given by this string of words which have been extracted from the corpus < prevalence child marriage Burkina Faso>[6] as well as by the search of the book "The Evil Child Marriage" by Radhika Kapu[7].

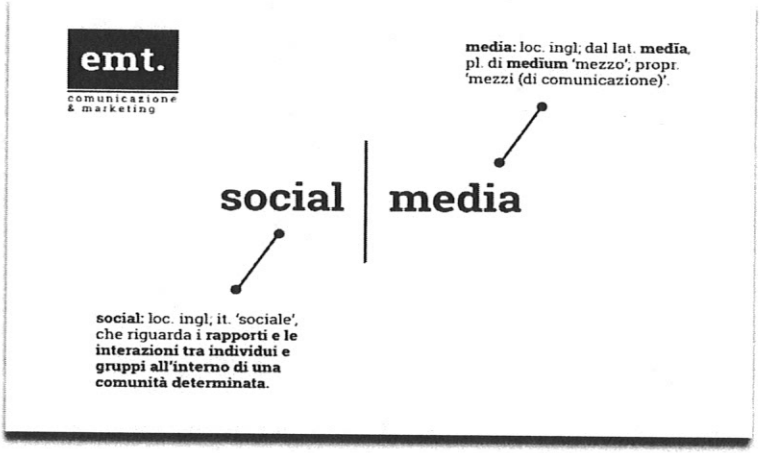


Figure 5. Meaning of the two terms

a) From the sample of queries where the bigram <social media> appears, polar verbs (affirmative and negative) and bipolar verbs (without a clear-cut tendency) have been extracted. In agreement with Manfred Klenner and Stefanos Petrakis, "we observed verbs with a relatively clear positive or negative polarity preference, as well as cases of verbs where positive and negative polarity preference is balanced (we call these bipolar-preference verbs)".

Just to make an example, a verb like 'to cause' has a negative polarity while 'to help' has a positive polarity; 'to impact' and 'to influence' have a neutral polarity (please see par. 3.1 for a detailed description of the terms with a negative polarity used in conjunction with <social media>).

The main results show that the use of social media

- can cause physical harm and widespread diseases;
- can help strengthening virtual relationships;
- can sustain e-commerce and therefore business grown.

b) From the sample of queries associated to the words <child/children>, only a negative polarity of the lexical forms emerges. The use of the following verbs, nouns and adjectives clearly certifies this assumption: 'to abuse', 'to labor', 'marriage', 'psychiatry', 'depression', '(domestic) violence', 'obesity', 'assassin', 'punishment'; 'domestic', 'psychopathic', 'violent'.

Terms extracted from the corpus of queries are largely referring to topics pertaining to the major problems of today's society, eg. *alcoholism, depression, obesity, pornography, drugs, violence, etc.*

Some critical issues can be identified in the following points: a diachronic analysis of the terms was not possible given the short temporal window taken into account; queries in different languages and many spelling/grammatical errors made our task more complicated by weighing the cleaning process down.

[5] https://enricomtomassi.com/cosa-social-media/

[6] Burkina Faso has a child marriage prevalence rate of 52%. On average, almost one out of two girls in Burkina will be married before the age of 18. The rates of child marriage vary from one region to another, and are as high as 86% in the Sahel region and 76% in the East region.
< https://www.girlsnotbrides.org/child-marriage/burkina-faso/>

[7] https://www.researchgate.net/publication/323771530_Child_Marriage_The_Social_Evil

Essential Bibliography

Battelle, J. 2005. The Search: How Google and Its Rivals Rewrote theRules of Business and Transformed Our Culture. Nicholas Brealey Publishing.

Bronson Fitzpatrick R., (2009) WorldWideScience.org: The Global Science Gateway, Medical Reference Services Quarterly, 28:4. {363-370, DOI:10.1080/02763860903256121}.

Chiari, I. 2007-2008. Liste di frequenza. Analisi del testo letterario 1 -, 2007-2008

Hendry David G., Jenkins J. R., McCarthy Joseph F. (2006). Collaborative Bibliography, *Inf. Process. Manage.*,May 2006,volume 42,3. Pergamon Press Inc. Pages 805-825. {http://dx.doi.org/10.1016/j.ipm.2005.05.007}.

Kalgarriff A., Grefenstette G. *Introduction to the Special Issue on the Web as Corpus*. ACL {https://www.mitpressjournals.org/doi/pdf/10.1162/089120103322711569}.

Klenner M., Petrakis S. (2012). Polarity Preference of Verbs: What Could VerbsReveal about the Polarity of Their Objects? In G. Bouma et al. (Eds.): NLDB 2012, LNCS 7337, pp. 35–46, 2012.Springer-Verlag Berlin Heidelberg 2012

Smith A., Anderson M.(2018). *Social Media Ue in 2018*, Report, March 1, 2018. Pew Research Center Internet & Technology {http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/}.

Trevisan M., Dini L. Barbu E., Barsanti I., Lagos Ni., Segond, F., Rhulmann M., Vald, Ed (2012). Query Log Analysis with GALATEAS LangLog, in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, ACL, Stroudsburg. Pages 87-91. {http://dl.acm.org/citation.cfm?id=2380921.2380939}.

Web Search Queries as a Corpus

https://www.celi.it/blog/2016/06/web-corpora/
<http://www.fondazionemilano.eu/blogpress/weaver/2014/03/22/181/>
<https://www.uniba.it/docenti/gatto-maristella/attivita-didattica/materiale-didattico/Gatto_light.pdf>
<https://worldwidescience.org/>
<http://greyguide.isti.cnr.it/>
https://www.researchgate.net/publication/323771530_Child_Marriage_The_Social_Evillarity_item

Appendix

SEMANTIC EXTRACTION SCHEME

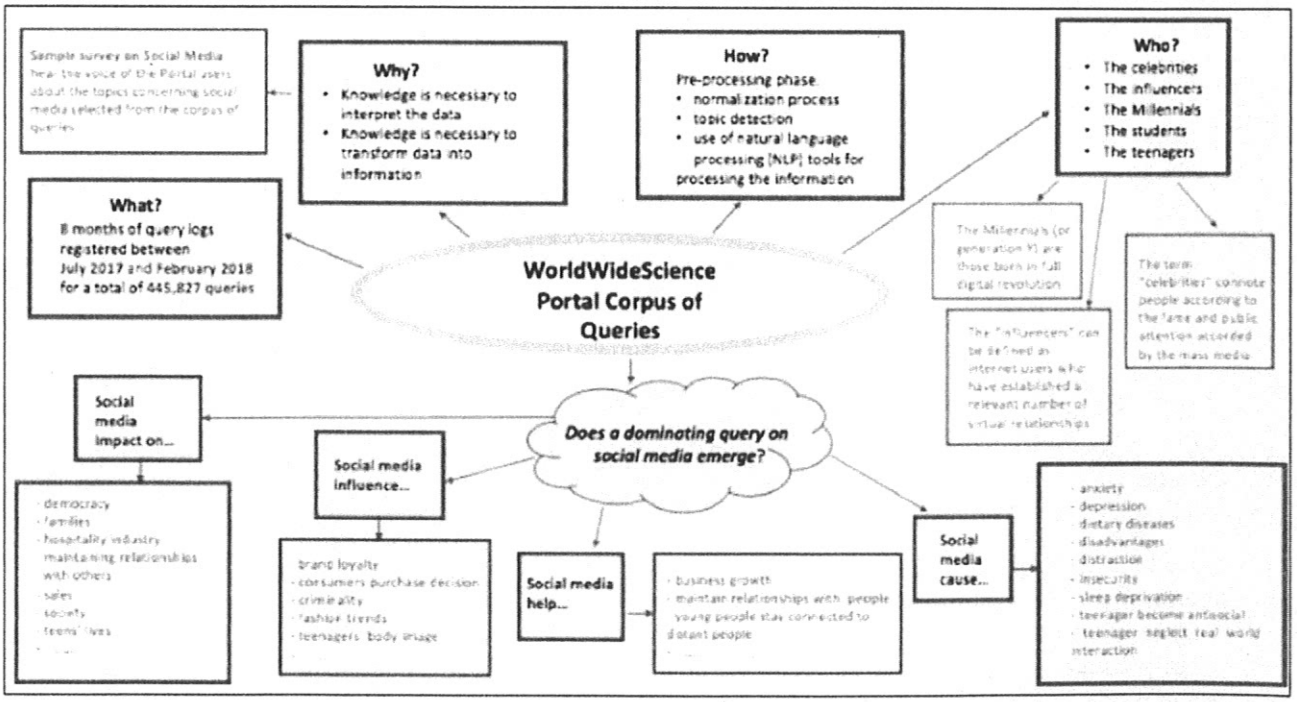


Fig. A1

THE GLOBAL SCIENCE GATEWAY
<https://worldwidescience.org/index.html>

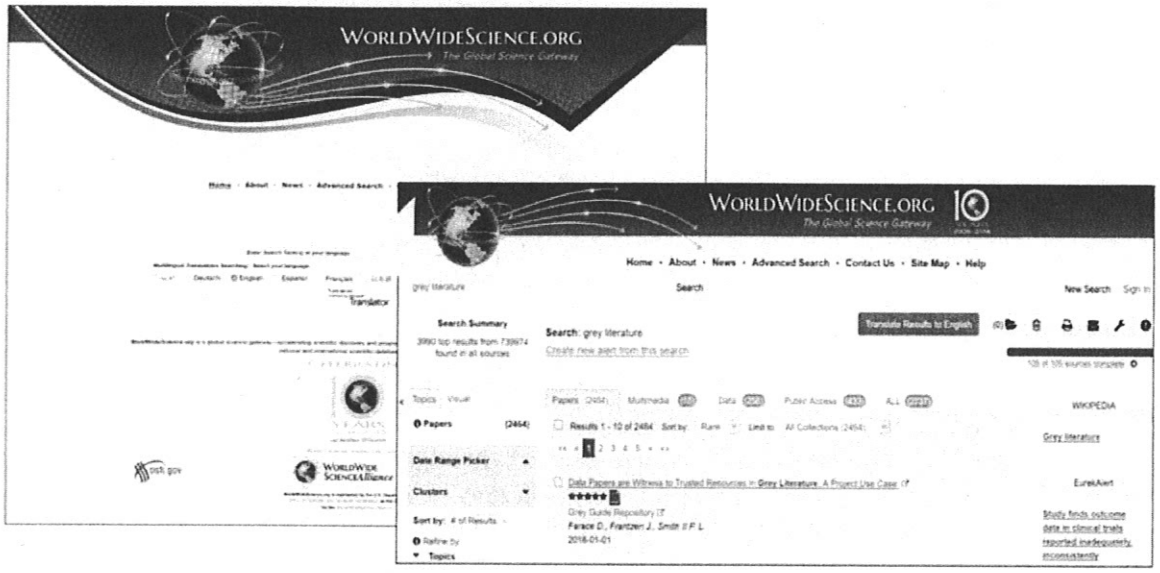


Fig. A2

REPOSITORY AND PORTAL TO GOOD PRACTICES AND RESOURCES IN GREY LITERATURE
<http://greyguide.isti.cnr.it/>

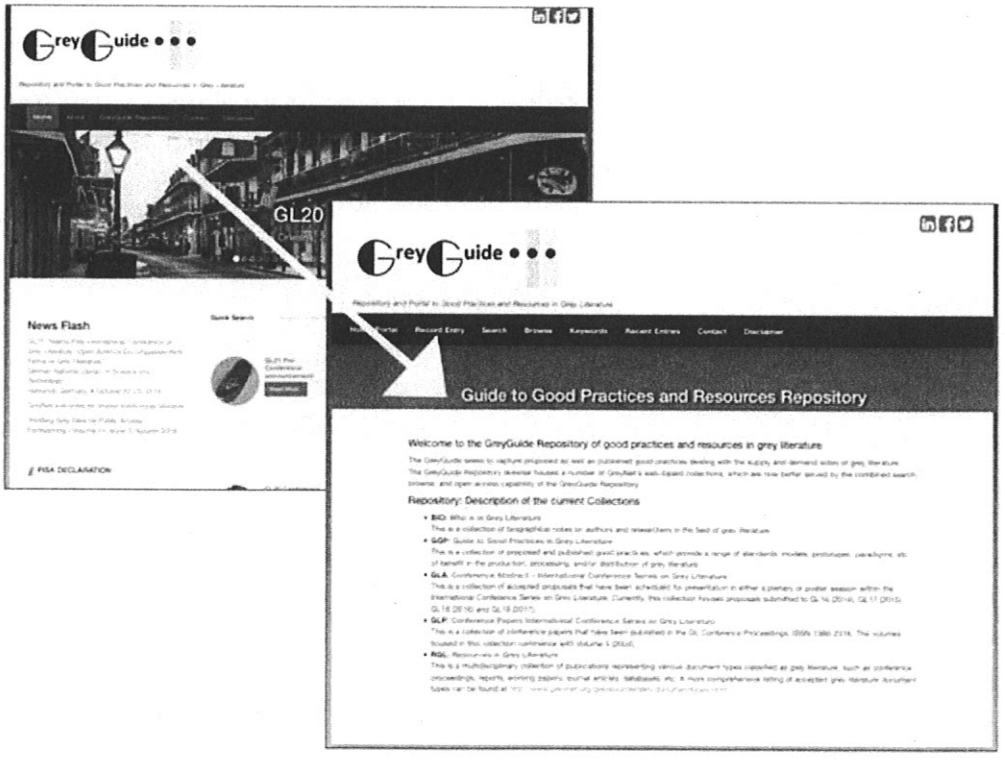


Fig. A3