Twentieth International Conference on Grey Literature. Loyola University, New Orleans, USA
3-4 December 2018

Semantic Query Analysis from the Global Science Gateway

Sara Goggi, Gabriella Pardelli*, Roberto Bartolini*, Monica Monachini*, Stefania Biagioni**, Carlo Carlesi***

*Istituto di Linguistica Computazionale, "A. Zampolli", CNR-ILC, Italy **
*Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR-ISTI, Italy ***

Scenario

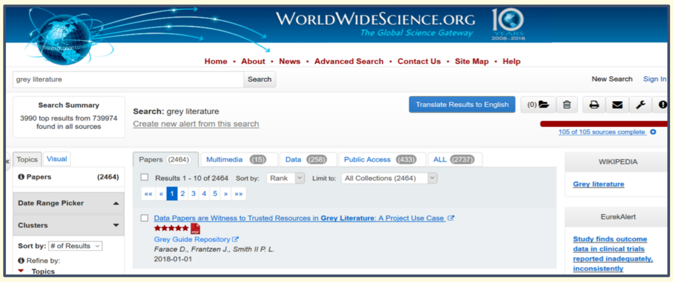
- ❑ Web portals play an essential role in searching and retrieving information in the several fields of knowledge
- ❑ Web portals support the storage of a huge amount of information in NL originating from the queries launched by users worldwide
- ❑ A good example is given by *WorldWideScience.org* (The Global Science Gateway)

Objective

- ❑ The aim is to retrieve information related to *social media* which as of today represent a considerable source of digital data more and more widely used for research ends

Focus

- ❑ The query logs registered by the *GreyGuide: Repository and Portal to Good Practices and Resources in Grey Literature* and received by the *WorldWideScience.org portal*



grey literature

Search Summary
3990 top results from 739974 found in all sources

Search: grey literature
Create new alert from this search

105 of 105 sources complete

Papers (2464)

Results 1 - 10 of 2464 Sort by: Rank Limit to: All Collections (2464)

«« « 1 2 3 4 5 » »»

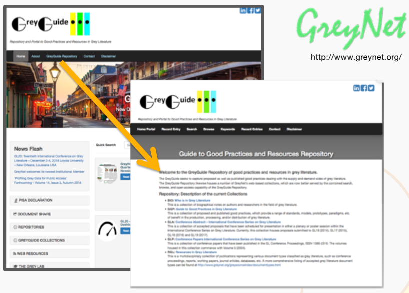
Data Papers are Witness to Trusted Resources in Grey Literature: A Project Use Case
Grey Guide Repository
Farace D., Frantzen J., Smith II P. L.
2018-01-01

Sort by: # of Results

Grey literature

Study finds outcome data in clinical trials reported inadequately, inconsistently

https://worldwidescience.org/



Repository and Portal to Good Practices and Resources in Grey Literature

Guide to Good Practices and Resources Repository

Welcome to the GreyGuide Repository of good practices and resources in grey literature.

Repository: Description of the current Collections

http://www.greynet.org/

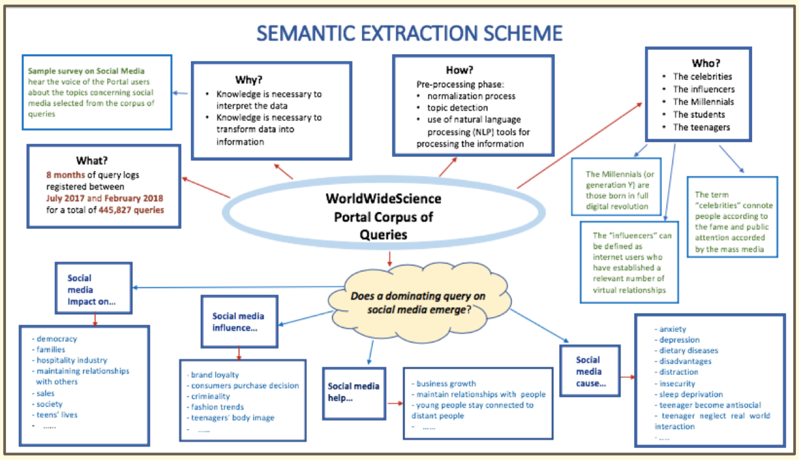
http://greyguide.isti.cnr.it/

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

Methods and Tools

- ❑ A process of information retrieval from a rich digital catalogue of queries
  1) cleaning of the set of queries;
  2) filtering and ordering (alphabetically);
  3) using several trials for choosing the focus;
  4) processing the information and building the sample by NLP tools.

- ❑ NLP analysis
  - a) free information extraction:
    - ➢ measure the frequency of the words contained in the corpus;
    - ➢ examine the lexical variety of the queries ;
    - ➢ focus on a set of terms to build a micro-ontology.
  - b) ontology-based extraction:
    - ➢ enrich the domain;
    - ➢ retrieve each occurrence of those terms contained in the ontology by using a search engine.



SEMANTIC EXTRACTION SCHEME

Sample survey on Social Media
hear the voice of the Portal users about the topics concerning social media selected from the corpus of queries

Why?
- Knowledge is necessary to interpret the data
- Knowledge is necessary to transform data into information

How?
Pre-processing phase:
- normalization process
- topic detection
- use of natural language processing (NLP) tools for processing the information

Who?
- The celebrities
- The influencers
- The Millennials
- The students
- The teenagers

What?
8 months of query logs registered between July 2017 and February 2018 for a total of 445,827 queries

WorldWideScience Portal Corpus of Queries

The Millennials (or generation Y) are those born in full digital revolution

The "influencers" can be defined as internet users who have established a relevant number of virtual relationships

The term "celebrities" connote people according to the fame and public attention accorded by the mass media

Does a dominating query on social media emerge?

Social media Impact on...
- democracy
- families
- hospitality industry
- maintaining relationships with others
- sales
- society
- teens' lives
- ......

Social media influence...
- brand loyalty
- consumers purchase decision
- criminality
- fashion trends
- teenagers' body image
- ......

Social media help...
- business growth
- maintain relationships with people
- young people stay connected to distant people
- ......

Social media cause...
- anxiety
- depression
- dietary diseases
- disadvantages
- distraction
- insecurity
- sleep deprivation
- teenager become antisocial
- teenager neglect real world interaction
- .....

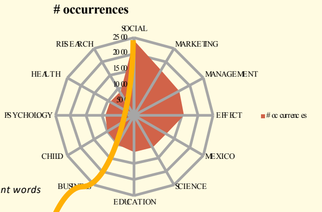


Fig.1 – Most frequent words

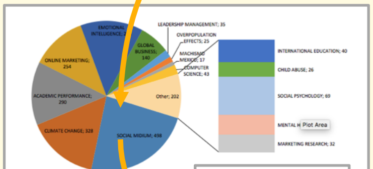


Fig. 2 – Most frequent bigrams

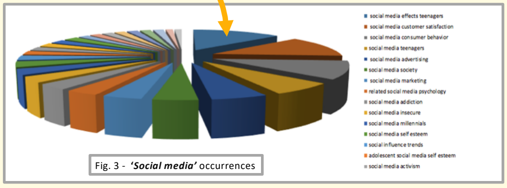


Fig. 3 - *'Social media'* occurrences

WorldWideScience and '*Social Media*'

Why 'social media'

- ❑ Social Media are a very effective means of communication and vehiculate knowledge
- ❑ They are often quoted in bibliographical references amongst the more traditional categories
- ❑ The subject involves document types pertaining to Grey Literature
- ❑ A case study has been carried out involving medicine, psychiatry and '*social media'* Figures 1, 2, 3

- ❑ Some low-frequency terms (hapax) carry a negative connotation in relation to the use of '*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>;

- ❑ An analysis of negative connotations in connection with *child/children*, is further investigated, as shown in Fig. 4

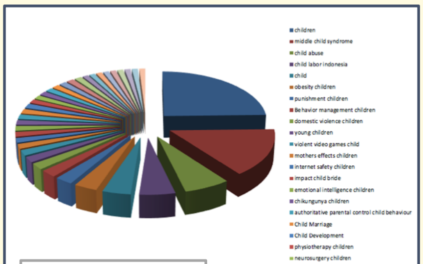


Fig. 4 – Child/children context queries

Final Remarks

- ❑ The case study illustrates the main linguistic features of the *Global Science Gateway* by showing the lexical map which represents the most used/recurrent words
- ❑ Some critical issues: 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
- ❑ 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….*
- ❑ 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.

*Contacts: * sara.goggi@ilc.cnr.it – ** stefania.biagioni@isti.cnr.it*