Geophysical Research Abstracts
Vol. 18, EGU2016-6413, 2016
EGU General Assembly 2016

Marine Planning and Service Platform: specific ontology based semantic search engine serving data management and sustainable development

Giuseppe M.R. Manzella (1), Andrea Bartolini (2), Franco Bustaffa (3), Paolo D'Angelo (1), Maurizio De Mattei (3), Francesca Frontini (2), Maurizio Maltese (3), Daniele Medone (3), Monica Monachini (2), Antonio Novellino (1), and Andrea Spada (4)

(1) ETT SpA, Genova, Italy (giuseppe.manzella@ettsolutions.com), (2) CNR, Istituto di Linguistica Computazionale, Pisa, Italy, (3) Delta Progetti S.L., La Spezia, Italy, (4) Sy.O. SrL., La Spezia, Italy

The MAPS (Marine Planning and Service Platform) project is aiming at building a computer platform supporting a Marine Information and Knowledge System. One of the main objective of the project is to develop a repository that should gather, classify and structure marine scientific literature and data thus guaranteeing their accessibility to researchers and institutions by means of standard protocols. In oceanography the cost related to data collection is very high and the new paradigm is based on the concept to collect once and re-use many times (for re-analysis, marine environment assessment, studies on trends, etc). This concept requires the access to quality controlled data and to information that is provided in reports (grey literature) and/or in relevant scientific literature. Hence, creation of new technology is needed by integrating several disciplines such as data management, information systems, knowledge management.

In one of the most important EC projects on data management, namely SeaDataNet (www.seadatanet.org), an initial example of knowledge management is provided through the Common Data Index, that is providing links to data and (eventually) to papers. There are efforts to develop search engines to find author's contributions to scientific literature or publications. This implies the use of persistent identifiers (such as DOI), as is done in ORCID. However very few efforts are dedicated to link publications to the data cited or used or that can be of importance for the published studies. This is the objective of MAPS. Full-text technologies are often unsuccessful since they assume the presence of specific keywords in the text; in order to fix this problem, the MAPS project suggests to use different semantic technologies for retrieving the text and data and thus getting much more complying results.

The main parts of our design of the search engine are:

- Syntactic parser - This module is responsible for the extraction of "rich words" from the text: the whole document gets parsed to extract the words which are more meaningful for the main argument of the document, and applies the extraction in the form of N-grams (mono-grams, bi-grams, tri-grams).
- MAPS database - This module is a simple database which contains all the N-grams used by MAPS (physical parameters from SeaDataNet vocabularies) to define our marine "ontology".
- Relation identifier - This module performs the most important task of identifying relationships between the N-gram extracted from the text by the parser and the provided oceanographic terminology. It checks N-grams supplied by the Syntactic parser and then matches them with the terms stored in the MAPS database. Found matches are returned back to the parser with flexed form appearing in the source text.
- A "relaxed" extractor - This option can be activated when the search engine is launched. It was introduced to give the user a chance to create new N-grams combining existing mono-grams and bi-grams in the database with rich-words found within the source text.

The innovation of a semantic engine lies in the fact that the process is not just about the retrieval of already known documents by means of a simple term query but rather the retrieval of a population of documents whose existence was unknown. The system answers by showing a screenshot of results ordered according to the following criteria:

- Relevance – of the document with respect to the concept that is searched
- Date - of publication of the paper
- Source – data provider as defined in the SeaDataNet Common Data Index
- Matrix - environmental matrices as defined in the oceanographic field
- Geographic area - area specified in the text
- Clustering – the process of organizing objects into groups whose members are similar

The clustering returns as the output the related documents. For each document the MAPS visualization provides:

- Title, author, source/provider of data, web address
- Tagging of key terms or concepts
- Summary of the document
- Visualization of the whole document

The possibility of inserting the number of citations for each document among the criteria of the advanced search is currently undergoing; in this case the engine should be able to connect to any of the existing bibliographic citation systems (such as Google Scholar, Scopus, etc.).