Toward a knowledge portal in environmental health: taking advantage of the semantic web

Bernard Moulin, Dr. Professor
Computer Science Department and Geomatics Research Centre,
Laval University, Pouliot Building
Ste-Foy, Québec G1K 7P4 , Canada,
Phone:1- (418) 656-5580,
Email: Moulin@ift.ulaval.ca

To reduce and guard against health risks of environmental origin requires easy and rapid access to high quality statistics and information which must be analysed in a useful manner to support decisions and interventions. The problems that environmental and public health officials (EPHO) must confront are extremely diverse. Sources of contaminants may be local, regional, or planetary and may be measured in humans or in several other vectors. The knowledge needed by EPHOs is also very diverse (medecine, environment, chemistry, etc.) and the sources of information are varied (data bases, books, proceedings, journals, radio recordings, video recordings, etc.). When confronting a new case (ex: during the winter season several meningitis cases identified in Quebec city area), EPHOs must act rapidly (ex: decide whether to launch a vaccination campaign or to wait for the detection of other cases, inform parents, etc.). To this end s/he must get the best available information that will help make a decision. The Internet is a very abundant source of information that health professionals use everyday. However, as most professionals trying to search information on the Internet, they are confronted to the problem of getting back from search engines huge lists of URLs, most of them being irrelevant to their requests.

Considering the specialized domain of environmental health, we developed an initial version of a domain ontology with the help of senior EPHOs and domain specialists. We developed a tool (a kind of meta-search engine) that uses this ontology to help users create precise queries in terms of conjunctions of domain expressions. These queries are submitted to search engines such as Google. The search results are filtered based on an analysis of the content of the descriptions of URLs (for instance the small summaries that Google associates to URLs) in order to verify how well they match with the content of the initial requests: do they contain the expressions used in the user's query, in which order? How close are they of each other?

In order to improve the filtering process, it would be very useful to compare the semantic representations of the query and of the URL descriptions. Such semantic representations are not available yet and we work on techniques that exploit lexico-semantic knowledge to analyze the words the user's expressions found in the URL description as well as their relationships. In some ways we try to use lexical knowledge to approximate semantic knowledge. But, this is a temporary solution. In the future we will need more elaborate semantic descriptions of document summaries.

Another problem is related to multimedia documents (audio, video documents). How to index them and how to retrieve them on the basis of their semantic content. The MPEG7 proposal which provides guidelines to index multi-media documents is still uncomplete

and has not yet been largely adopted. It would provide some ways to specify the semantic content of multi-media documents. We need search engines which take advantage of these semantic representations of documents.
These various indexing and search facilities will be integrated in a Knowledge Portal devoted to environmental health. Our plan is to create a repository of URLs relevant to EPHOs organized into categories corresponding to the domain ontology. This repository will be created and continuously updated by a population of agents that will query domain specific search engines, gather the results and filter them according to the URL repository content. Other agents (crawlers) will explore the web seraching for new categories that might be relevant to EPHOs. Such an URL repository will guarantee users precise results, excellent response time, notification of new entries that they have not consulted yet, as well facilities to generate summaries of relevant documents.

When semantic content will be available for documents accessible on the web, it will be interesting to measure the gains of efficiency of searches and indexing over classical methods.