Multi-lingual Search Engine to Access PubMed Monolingual Subsets: A Feasibility Study.

Stéfan Darmoni, Lina F. Soualmia, Nicolas Griffon, Julien Grosjean, Gaétan Kerdelhué, Ivan Kergourlay, Badisse Dahamna

MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-966

Multi-lingual Search Engine to Access PubMed Monolingual Subsets: A Feasibility Study

Stéfan J. Darmoni[a,b], Lina F. Soualmia[a,b], Nicolas Griffon[a], Julien Grosjean[a], Gaétan Kerdelhué[a], Ivan Kergourlay[a], Badisse Dahamna[a]

[a] *CISMeF & TIBS, LITIS EA 4108, Rouen University Hospital, Rouen, France*
[b] *INSERM, Unité Mixte de Recherche en Santé (UMR_S) 872, équipe 20, Paris, France*

Abstract and Objective

PubMed contains many articles in languages other than English but it is difficult to find them using the English version of the Medical Subject Headings (MeSH) Thesaurus. The aim of this work is to propose a tool allowing access to a PubMed subset in one language, and to evaluate its performance.

Translations of MeSH were enriched and gathered in the information system. PubMed subsets in main European languages were also added in our database, using a dedicated parser. The CISMeF generic semantic search engine was evaluated on the response time for simple queries.

MeSH descriptors are currently available in 11 languages in the information system. All the 654,000 PubMed citations in French were integrated into CISMeF database. None of the response times exceed the threshold defined for usability (2 seconds).

It is now possible to freely access biomedical literature in French using a tool in French; health professionals and lay people with a low English language may find it useful. It will be expended to several European languages: German, Spanish, Norwegian and Portuguese.

Keywords:

Databases, bibliographic; French language; Information storage and retrieval; PubMed; User-Computer interface;

Introduction

MEDLINE, created by the U.S. National Library of Medicine (NLM®), is the most used bibliographic database in the world. Currently (October, 5 2012), it contains 19,986,088 citations from 5,631 indexed journals from 81 countries around the world. Each MEDLINE record is indexed with the NLM's controlled vocabulary, the MeSH thesaurus. MEDLINE is the largest component of PubMed (URL: http://pubmed.gov/).

Few tools have already been developed and published to help non-native English speakers to query PubMed/MEDLINE in their native language, in particular BabelMeSH and PICO Linguist. We have also developed a tool to perform PubMed queries in French via a French MeSH browser.

A multilingual search engine to access the PubMed/MEDLINE subset in any one language (e.g. French, German, Spanish or Norwegian) would be of great interest for health professionals and lay people, who are unable to read sufficiently well in English. The goal of this paper is to present such tool, named $MLPubMed_{Ln}$ (Ln stands for language, for French, it will be named $MLPubMed_{Fr}$), and to evaluate it.

Methods

Recently, the CISMeF semantic search engine has improved in two ways, it is now: (1) a generic tool able to describe and index not only web resources but also PubMed citations or Electronic Health Records; (2) multi-lingual by allowing queries in multiple terminologies and several languages.

To integrate NLM data in the information system, a specific java sax parser was developed. Currently, XML parsing and database feeding are separated in two steps. All PubMed citations in French were integrated at the same time. In a production environment, this process will of course be batched with only new data integration. Several MeSH translations were also gathered in the information system.

To perform the feasibility study, response time and number of PubMed citations for 20 MeSH Descriptors were measured. These 20 MeSH Descriptors were chosen from the Top 100 MeSH Descriptors used as Major Topics in the entire PubMed database.

Results

As proof of concept, thousands of PubMed citations were included in $MLPubMed_{Ln}$ in the following languages: French, German, Portuguese, Spanish and Norwegian. The end-user was able to choose one language, then query in the same language. The $MLPubMed_{Ln}$ interface was translated into French, German, Portuguese, Spanish and Norwegian. The same main metadata displayed by default in the PubMed database was also displayed in the $MLPubMed_{Ln}$ search engine, in particular the indexing (MeSH Descriptors and Qualifiers) and the title of the article. The direct link to the full-text article in the same language was also presented to the end-user via Digital Object Identifier (DOI). The objective to create a bibliographic database extracted from PubMed in several languages and completely available in one specific language was then fulfilled as a proof of concept.

To perform a feasibility study of the $MLPubMed_{Fr}$ search engine, all PubMed citations in French (n=654,096) were extracted from PubMed and included in the $MLPubMed_{Fr}$ semantic search engine. The results showed that all response times for all queries were below the limit of two seconds.

Conclusion

The feasibility study to create a multilingual search engine to query monolingual PubMed subsets has been considered as successful for French, and will be extended to the main European languages.