Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 243-244

Subject Access Metadata on the French Web

Ewa Nieszkowska
École nationale des sciences de l'information et des bibliothèques
Lyon, France
niesz@enssib.fr

Summary

The article presents four French projects on subject metadata use: a medical portal (Caducee.net[1]), a subject gateway (Les Signets[2]), a catalogue of patents (INPI, Institut National de la Propriété Industrielle[3]) and a full-text database of the daily newspaper "Libération". The last project is not a public Web application yet but it presents the most innovative approach to subject metadata usage discussed in the article.

These projects, both completed and in progress, as a common characteristic share the use of controlled documentary languages. By this means, they try to increase the efficiency of information retrieval for the remote user.

The article tries to determinate the "remote user" characteristics: he or she is defined as a person searching for information (often for professional purposes) and who often needs exhaustive information in the situation of chronic time shortage. The most popular search engines cannot satisfy such users, who need a more organised Web, and more efficient search. In fact, they might also need a librarian, although they do not know it yet!

However, when they sit alone facing their computer screens, they do not receive assistance from information retrieval specialists (as, for example, librarians). In this situation, it is the role of a resource provider to help remote users in their documentary search and to make this search more user-friendly.

How do the studied projects approach this problem? For Caducee.net and Les Signets, it is done by means of a fairly classical use of indexing languages. In the case of Caducee.net, it is achieved by the use of a standard familiar for the medical public called MeSH[4] (F-MeSH in French). Les Signets face the problem by the planned use of RAMEAU[5], French indexing standard, which has very large number of "used-for references", i. e. non-descriptors that can guide the user to descriptors themselves and, in this way, to relevant resources.

INPI case is more interesting and unusual. Since the indexing language is an alphanumeric one (complex class symbols incomprehensible for a remote user), a linguistic engine is employed to enable search in natural language. Afterwards an index of keywords is generated from existing verbal descriptions of class marks themselves.

All the above-mentioned projects show the importance of natural language tools for remote users. And the fourth project's study seems to indicate that the use of controlled languages in full-text environment can be beneficial for controlled languages themselves: it's the case of the daily "Libération" thesaurus.

This thesaurus appears to prove that full text documentary environment may also be used to create and/or maintain indexing vocabularies and thesauri. The descriptors of the thesaurus are associated (via KALIMA software) with "lexical units" from the full-text articles database. Of all the thesaurus' modules, two seem particularly interesting, as they make an inventive use of an association between "lexical units" and the thesaurus' descriptors. They are called the "Automatic Learning Module" (ALM) and the "Automatic Indexing Module" (AIM):

ALM works by extraction of texts selected for learning, linguistic analysis of their contents, comparison of the contents with their indexing descriptors, finally by saving the results of the comparison in "indexing prototypes". Every time the ALM is used, it generates a new "indexing prototypes". At the same time, the thesaurus' administrator is asked to validate or reject new associations created between the words and expressions coming from the text and the thesaurus' descriptors. The recommendations for validation or rejection of these lexical units are based on their frequency in the given text.

AIM's function is to draw up a list of relevant candidate descriptors of new documents that have been put in the database. This process works by extraction of all the documents' fields (i.e. text title and the body), linguistic analysis of the contents, comparison

with the thesaurus and then final comparison with indexing prototypes of AIM. As the output, the librarian receives predetermined number of the closest descriptors whose relevance has been assessed by the AIM. Afterwards, the librarian's work is to validate, or reject the candidate descriptors and to add those, which have not been generated by the system.

It is important to understand, and underline, that the creation and maintenance of indexing languages is one of the library activities that incur the highest cost. It stems from the fact that for the time being it has been impossible even to part-automate it. The example of the "Libération" thesaurus seems to be opening up a different perspective - not for documents on the Web, but rather for documentary languages …

… Full text contribution is to reduce tedious human workload in the maintenance of indexing standards. It is for the machine to take care of scanning texts and then to compare them with the existing descriptors. This way, the process of assembly, selection and choice of indexing vocabulary, as well as its maintenance, is considerably accelerated. The librarian is made to take final decisions, but whatever can be automated, will be. To put it simply: do not ask what metadata can do for the Web; ask what the Web can do for metadata.

---

[1] http://www.caducee.net
[2] http://www.bnf.fr/pages/liens/index.htm
[3] http://www.inpi.fr
[4] Medical Subject Headings
[5] Répertoire d'autorités-matière encyclopédique alphabétique unifié, http://rameau.bnf.fr