A Semantic Structure to Improve Information Retrieval Using XML

Suela Berisha-Bohé, Béatrice Rumpler, Rocío Abascal

INSA of Lyon – LIRIS
F69621 Villeurbanne cedex, France
{Suela.Bohe | Beatrice.Rumpler | Rocio.Abascal}@insa-lyon.fr

Abstract

Nowadays the information stored in the digital libraries is not completely described, so this information is not really used. The description of information by using metadata seems a good solution to permit the users to find pertinent information. Our proposal is based on the creation and the insertion of new metadata within the document as *«semantic tags»*. These metadata can describe, in our case, the doctoral theses, by taking advantage of XML technology to structure digital documents.

Keywords: digital library, information retrieval, structured document, metadata, XML Schema, knowledge base.

Problem Description

Current State: The scientific library of Doc'INSA set up since 1997 a project named CITHER, which makes possible the diffusion and the access of scientific theses through Internet. Currently, a user can get the contents of *only* one thesis at the same time without being able to select relevant extracts corresponding to a unit of corpus finer than the chapter.

Causes:
1. The use of an inadequate format, such as PDF.
2. The description of the contents only by the keywords added outside the documents
3. The use of the tags proposed by the Dublin Core metadata, which bring only general information

Our recommended solution: Two axes of work
1. Build a document suited to the information retrieval process.
2. Improve the information retrieval tools personalizing the search sessions.

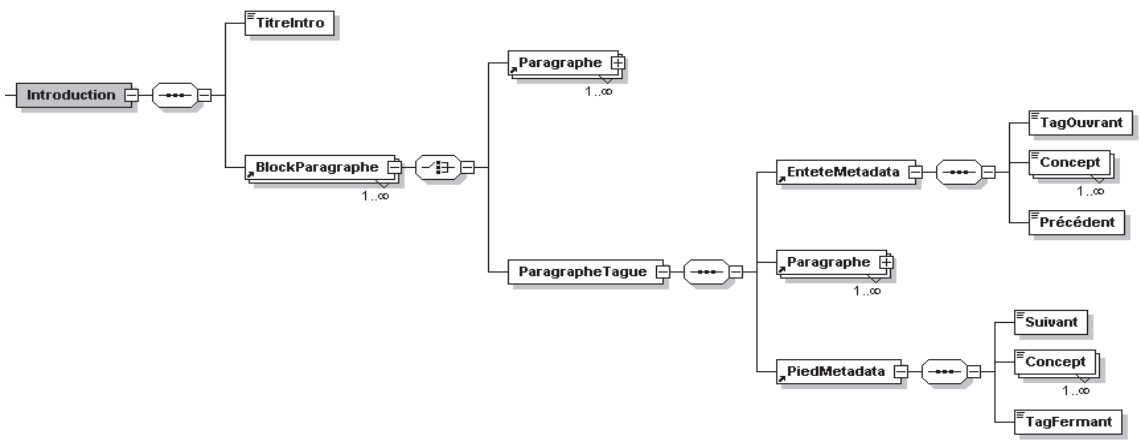

Figure 1. An example of the model of the semantic structure with XML Schema (in XMLSpy)

First axe: Build a Document Suited to the Information Retrieval Process

Tasks:
1. Manual and automatic concept's extraction, study and analysis of the corpus: choice of NLP.
2. Correlation between the use of concepts in the corpus: logical and semantic structure of thesis.
3. Ontology construction by using NLP tools.
4. Building of a knowledge base with the concepts extracted.
5. Creation of a new model of documents (Figure 1) and finally

==> ***Annotation Tool proposed at the author of the thesis*** (Figure 2)

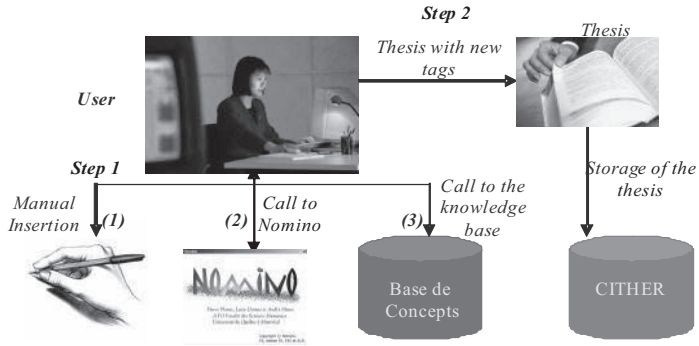

Figure 2. Annotation Tool: Integration of new metadata into the thesis

Second Axe: Improve the Information Retrieval Tools

Tasks: Model the users profile to personalize the search sessions

1. State of art of user modeling tools.
2. Knowledge-based user modeling (Figure 3) and identification of stereotypes.
3. Establishment of CBR technique in the information retrieval (current work).

Tool's Description

1. The PhD student uses this tool at anytime during the writing process of the thesis.
2. The author describes his thesis with concepts inserted manually coming from Nomino's extraction, using the knowledge base, or by personal initiatives.
3. The validated concepts are integrated into the thesis as metadata tags. These new tags will be exploited during a search session.
4. The tool validates document's structure and stores the thesis in the Digital Library

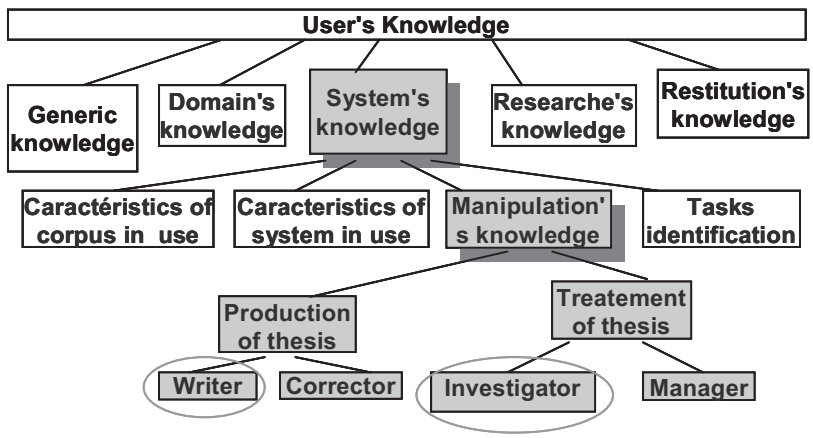

Figure3: Knowledge-based user modeling

Description

- The end user (student or investigator) can search information by means of the engine search of the tool during his activity.
- In our second part of work, Writer is a stereotype of investigator, which is well known by the tool. So, his case is easily identifiable thanks to the concepts used before.

Conclusion

To be able to find pertinent information using information retrieval tools, it is important to define a specific structure of the digital document during its creation. According to this point of view, we have defined a semantic structure of the document by integrating new metadata in judicious parts of the corpus. This makes possible to identify semantic segments in the scientific theses stored in our digital library: CITHER.

In a search session based on keywords, the system will compare them with the semantic metadata (delimiting the semantic segments) and with the keywords describing the thesis. Thanks to this approach the user can get pertinent fragments of one or several theses corresponding to the semantic segment.
We are currently working on the design of an *«advanced»* system based on the ontology built to find more pertinent information. Another axis of our current work is to model the users profile to personalize the search sessions and to implement the technique of CBR for the information retrieval.

Bibliography

Abascal R., Rumpler B., Pinon J.M., (2004) Information Retrieval in Digital Theses Based on Natural Language Processing Tools, J.L. Vicedo et al. (Eds): *España for Natural Language Processing (EsTAL'04), LNAI 3230*, pp. 172-182, Springer-Verlag Berlin Heidelberg, October 2004, Alicante, Spain.

Abascal R., Rumpler B., Pinon J.M., (2003) An Analysis of Tools for an Automatic Extraction of Concept in Documents for a Better Knowledge Management. IRMA International Conference, Philadelphia Pennsylvania, USA. Ed. Mehdi Khosrow-Pour, *IDEA Group Publishing*, ISBN: 1-59140-097-X, pp. 201-204, May 18-21, 2003.

Jolly C., (2000) Rapport sur la diffusion électronique des thèses. Ministère de l'Éducation Nationale—SDBD, 2000.

Thomasson J-J. (2002) Schémas XML, Ed. Eyrolles, ISBN: 2-212-11195-9, November 2002, 466 pp.