From Full Text Storage to Full Contents Representation: Information Science Between Library Science and Informatics

Information science has traditionally been concerned with the acquisition, storage, retrieval, and use of information that is expressed as scientific articles or other documents. However, through developments in several parts of informatics (i.e., computer science in a broad sense) new opportunities are opening up for addressing not only the characterization of documents by descriptors, but the representation of their major contents in themselves in computer analyzable form. These possibilities emerge from developments in the world-wide web and knowledge representation communities, but they are of great importance for information science and suggest a new and wide field of investigation.

In this article I address what are the appropriate software tools and working procedures in this development. There is a need both for schematic methods, such as text mining, that can process large quantities of documents, and for detailed methods that capture the contents of a single document or other information source with great precision. Combined, they contribute to building up domain models characterizing the general structure of specific disciplines.

The presentation will be illustrated with a few concrete examples where these tasks have been addressed. The primary example will be the work of representing the open-access policies of various publishers, as expressed in textual form in the Romeo database, in strictly structured form whereby they can be used in various software services.