15

Meta-Data Extraction from Bibliographic Documents for the Digital Library

A. Belaïd and D. Besagni

15.1 Introduction

The digital library (DL) [19] has become an increasingly common tool for everyone, a trend accentuated by the success of the Web and the easy access to every kind of information. Among the most important DL projects [17], we can mention "Project Gutenberg", the oldest producer of free electronic books [21], the "Million Book Project" [22], and more recently Google announced its intention to digitize the book collections from several famous universities (Michigan, Harvard, Stanford, Oxford) and from the New York Public Library [20].

The DL provides information located in one specific place to anyone, anywhere in the world, as long as the information can be retrieved. Contrary to the Web at large, the DL offers a more organized access to selected information that is often validated, filtered and structured. With this trend, documents not registered in electronic form will risk becoming invisible. It is the Google effect: "if it isn't in Google then it doesn't exist!". This electronic registration is not sufficient enough to define a DL: the document itself must be in electronic form, which does not mean it is machine readable.

A good DL is not only a good document retrieval system but the content must be at the same time accessible as well by the machines than by the users responding to their multiple needs. There are different aspects revealing the DL qualities relative to the:

(a) *Content:* The more structured a document is, the more useful it is. Added to this, the quality of the meta-data accompanying the document is essential.
(b) *Organization:* The more standardized a format is, the more usable and durable the document is.

(c) *Updating and patrimony valorization:* The main problem is not in the feeding of new digital document but in importing digital documents from other DLs, or in adding patrimonial non-electronic documents.
(d) *Use:* The DL function does not stop with the document consultation, but the more options the better.

Since all the documents are not specifically generated to enter a DL, we need cost-effective tools such as OCR, retro-conversion, hypertextualization and meta-data extraction techniques. There again the content defines the approach. Depending on the expected quality, these techniques will need more (or less) adaptation and depth.

At the DL level, independent of the origin of the document (electronic or not), it is obvious that the most important elements in terms of organization and structure are the meta-data and the hyperlinks. Although the hyperlinks, made popular by the Web, are the "icing on the cake" because they improve the navigation functionalities, the meta-data are more basic, even indispensable, because they include for example the catalogue.

Although some problems remain [9], we have now a generation of OCRs capable of extracting the content with a good quality close to 100%, the research tends towards systems dedicated to structure extraction. This structure, generally of an editorial or logical nature, obviously constitutes a first step towards meta-data generation.

As the documents in general are described in a DL by at least their descriptive meta-data, a straightforward use of a DL can be done by correspondence between the terms outlined in bibliographic documents (like bibliographic references, citations, cards, tables of contents, etc.) and the DL meta-data. Depending on the structure finesse of documents in DL and the precision of the outlined terms (which in our case are roughly recognized by OCR) the meta-data recognition can be considered as a "mapping problem" between the real meta-data and the recognized terms. Proper mapping of the bibliographic reference with its actual content within a DL is a challenging task of research [3].

15.2 The Users' Needs

Considering the DL as a very structured document repository that is well organized and continuously updated, we can envisage its use for some important requests similar to which they are done on the Web. But contrary to the Web, the use of bibliographical data may offer more possibilities in the use of common DL services such as:

- *Information retrieval:* This is related to a simple DL consulting. The major need of DL is to retrieve the actual document from DL based on approximated bibliographic terms (roughly recognized by OCR or provided by the user). This corresponds to about 70% of the real DL