Future Trends in Retrospective Document Conversion

A. Belaïd

CRIN-CNRS, Campus Scientifique, B.P. 239 F-54506 Vandœuvre-lès-Nancy Cedex
France

Abstract. This paper describes a framework for retrospective document conversion in the library domain. Drawing on the experience and insight gained from several projects launched over the present decade by the European Commission, it outlines the requirements for solving the problem of retroconversion and traces the main phases of associated processing.

1 Introduction

The success of library automation, resulting in user-friendly on-line catalogues\(^1\) integrated with the web and other circulation-systems facilities, has created an urgent need for retroconversion of the older parts of catalogues [2,30]. As users get used to the new catalogue medium, the documents not registered in machine-readable form become “invisible” and unreadable. This has meant for many libraries the relegation of an important part of their rich stock of documents to a state of accessibility.

Such obvious waste of library collections in addition to the cost difference between manual handling and an equivalent set of automatic routines has made a strong case for the need to convert a library’s entire collection of works to machine-readable records, in the interest of ensuring an efficient use of the investment in the new technology.

This has led to the search for cost-effective tools for the conversion of old catalogues into machine-readable forms. This search has not been limited to the sole problem of conversion but has been extended to embracing other objectives such as ensuring very high rates of distribution and sharing of documents between several libraries.

Drawing on [35], cataloguing methods can be classified into four main categories:

1. Retrocataloguing creating the catalogues from scratch by using new cataloguing tools and inherent standards ensuring the same structure and level of information as the library OPAC\(^2\).

---

\(^1\) A catalogue is a list of bibliographic descriptions of items held by a specific library, set up to give access to the items.

\(^2\) Online Public Access Catalogue.
2. Retroconversion by direct keying from the paper catalogues and other media using automated input tools without adding or changing information in the input data to make it compatible with common standards.

3. Retroconversion by OCR and automatic formatting, creating a structured and a tagged result of the machine-readable record.

4. Retroconversion by the substitution of old records with preexisting machine-readable records of the same records, provided by conversion agency or service production of national bibliographies.

The choice of method is conditioned by the state of funds available and the quality aimed at [22]. If the aim is to create a structure which is consistent with the rest of the OPAC records and respects its standards, then the first method appears as the most natural choice. Cost and time, however, constitute the two serious drawbacks of this choice.

On the other hand, if the aim is to make a maximum number of records available with the possibility of sharing with other sites, the substitution method then seems the most suitable choice. This choice also corresponds to the 1989 European Council Guidelines recommending not only inter-library cooperation in the creation of machine-readable records but also taking advantage of national bibliographical machine-readable records as well as those of important (private) collections [9-11]. In this context, appointed national agencies are to be charged with the responsibility of converting catalogues of national collections and making the results available to other libraries. The attractiveness of this solution, especially with respect to saving local efforts, is however marred by associated delays in making foreign collections available to users, a solution placing libraries in the unacceptable situation of having to wait for the new catalogues before making foreign collections accessible.

Fully aware of the limitations of these methods, most libraries are embarked on a search for optimal solutions. There is growing understanding that a combination of OCR and structure recognition may provide the basis for serious solutions. We can single out three of the several reasons for this. The first is that this approach avoids manual procedures and allows an integrated solution capable of adapting to the specific structure of the catalogue to be converted. Secondly, it allows a greater flexibility in the structuring of records without the necessity of having to reproduce a structure similar to the rest of the OPAC. Finally, the solution can easily be extended to other situations of retroconversion, requiring less rigid format structures, and where there is need to cope with free variations in formats as in scientific references.

The problem is not satisfactorily solved at the moment. Existing systems using OCR based solutions are not tangible enough to allow an assessment of the success of the technique. There is therefore room and need for more investigations to improve on the modest results obtained so far in this domain.