MORE (Marc Optical REcognition) *technical aspects*

Catherine LUPOVICI, Jouve Systmes d'Information

The MORE system is an automation of the separate functional modules to be set up in a traditional recon system to provide :

catalogue digitalisation : to get an image of each entry to be converted by image processing

character recognition : to get the bibliographic data coded in character mode

structure recognition : to get the UNIMARC logical structure of the bibliographic data

coded information : to create all the coded information needed in full UNIMARC format in the coded information block (fields 1XX)

conversion management : quality checking, controls, abnormal records handling and exchange of information with the library, final ISO 2709 formatting

It was decided for the project to develop this automated modules and to integrate them in the prototype according to a general architecture allowing a) to combine them in a flexible way in order to be able to possibly process manually one of the module b) to be able to occasionally bypass manually the automated modules in order to process abnormal records and exceptions.

This architecture will allow for instance to key data content for an hand written catalogue and to process the result for automated structure recognition and coded data creation, or to structure at the full MARC level a machine readable catalogue created with a less detailed format.

All the modules are composed of batch process providing also automated errors or doubts detection to be solved by an operator using an appropriate editor in a Windows environment.

The management module provides statistics, gathered during the production test, in order to assess the feasibility, performance, costs and quality of the method, to be compared with the classical methods.

OCR/ICR of bibliographic data

OCR/ICR software are created for the office environment to process simple textual information. They generally integrate language dictionaries to help the character recognition by spell checking processing. But bibliographic information is far more complex than textual office information. It is composed of :

Textual information :

- Text in a large number of languages and sometimes in romanized non Latin scripts with diacritics

- Proper names : personal names, corporate body names, geographical names

- Specific bibliographic and cataloguing vocabulary often abbreviated

Numeral information including roman numeral to be identify properly

Punctuation marks, parenthesis, square brackets, long and short dashes etc. which can be either part of the bibliographic data or which can represent logical structure of the bibliographic description

In addition, in printed bibliographies, style is used to emphasise the logical structure (for instance bold for the record main entry, italic for bibliographical address and physical description). Bold often creates characters connection and alternate use of italic and standard style affects the spacing recognition.

MORE OCR/ICR automated module

The project is an application to bibliographic information of the Jouve's OCR/ICR know how, already used in production for instance for the patents (1 million pages a year are processed for the European Patent Office).

The principles are to add to existing software Intelligent Artificial tools for decision making. This allows to avoid no recognition response, and to propose alternative responses with a default one pre selected. For this catalogue three OCR where selected and used in combination : Omnipage from CAERE running on PC, Scanworx from Xerox Imaging System and Recore from Ocron both running on UNIX.

They are integrated in the JOUVE's character recognition package named JCR and the whole system has specific parameters defined for the bibliographic information. The system is then tuned for each specific catalogue.

The error and doubt detection routine integrated in the OCR/ICR module provides a file which is then edited. This routine offers several solutions with a default pre-selected one. The operator will have to validate or select another solution like in spell checking routines of word processing software.

MORE structure recognition module

This software is based on the CRIN know how in structure recognition, and on a preliminary study made in 1991-1992 in partnership with JOUVE within the framework of a PhD on Structure Recognition1. This preliminary study was conducted on the French Bibliothque Nationale 1960-1969 printed catalogue.

The knowledge base of the expert system is a detailed data model written with and ODA/SGML2 syntax. There is a model for each part of the catalogue to be converted : a model of bibliographic description, a model of author-title index and a model of subject index. The model of the bibliographic description is rather sophisticated and ha about 40 levels in depth.

Specific dictionaries of keywords are associated to specific parts of the model. They are created by the library from an analysis of the Belgian Bibliography and can be enhanced as the conversion is going on.

The structure analysis software, driven by the model, go through the record character string and build hypotheses on content portions which are tagged logically and checked against left and right context. Weighting technology is used to tune the system.

MORE coded data creation module

The module processes the result file of structure recognition in order to create the following coded data according to UNIMARC :

country code using an atlas of town names with the corresponding code is created from the data tagged in 210 $a sub field

language code using general language dictionaries and weighting techniques. It is applied to the title (200 $a) and also to parallel titles (200 $d and 225 $d)

Both codes are edited through a single editor, developed for the project by checking the proposals of an automated correction routine.

Project results

One of the project objective was to test the interest of introducing OCR/ICR and structure recognition technologies in Retrospective conversion in order to reduce the cost at the same quality level.

The assessment was done by testing the system in real production environment. One year of the bibliography was converted using the prototype developed during the project and tuned for this bibliography, both for batch processing and for editing. Editing was done by the retrospective conversion staff of Jouve, full time employed for classical retrospective conversion, and the test and result were assessed both from the bibliographic point of view and from the production point of view.

Bibliographic quality level

The objective was to reach the quality currently provided by classical double keying, coding and quality checking. This quality is assessed on a random sample, following the methodology defined by the ISO 2859 standard. The application of this standard to retrospective conversion leads to check :

the wrong characters

the structure and coding errors, in relationship with the error importance regarding to the future electronic catalogue search and retrieve usage or processing.

For the production test, quality was assessed following the same methodology and rules, and the final quality is (after automatic processing, editing, batch automatic quality control and correction) :99,974 % characters accuracy

100% fields accuracy, and 99,28% sub fields accuracy

98,4 % of language and country codes accuracy

Main problems were the title sub fields in UNIMARC with a lot of parallel statements relating to the Belgium official multilinguism. For OCR and language code the Dutch dictionary used in the project was not enough powerful.

In conclusion, the quality level reached and the possible level after enhancing the system is equivalent to classical retrospective conversion if the appropriate tools (OCR software, general and specific dictionaries, modelling) can be found or build for the catalogue to be converted.

Productivity

Principles of automatic processing in an industrial production environment is to process regular cases automatically and to keep human intervention for processing only the exceptions, in order to reduce the human time and cost. At the same time computer costs increase, and the human time reduction has to balance this computer increasing.

Assessment of cost comparison between using this methodology and using classical retrospective conversion was made by comparing the real statistics of the production test and an evaluation of time processing for this bibliography, as it is done building a commercial offer for retrospective conversion services, on the basis of the same library specifications.

Human intervention and automated processing figures are :

10 doubts to be checked per record for OCR/ICR editing

47,5 % of records completely automatically processed for structure and code and after quality control processing. More records were processed automatically but some of them were edited after the quality control either for structure of for coded data

5,4 % were sent back to the library as non conforming to the specifications

47,1 % of records edited for structure or/and code

Global human time from the beginning of the process until the final tape formatting is reduced by 1/3. Half reduction can reasonably be expected in full production and the project conclusion is that such methodologies allow costs reduction in some cases. In addition they will be enhanced as the basic software integrated will progress and as they become better known and managed.

[1] Chenevoy, Yannick. Reconnaissance structurelle de documents imprims : tudes et ralisations. Thse Institut National Polytechnique de Lorraine, Dcembre 1992.

[2] Office Document Architecture (ODA) ISO 8613

Standard Generalized Markup Language(SGML) ISO 8879