Retrospective Conversion of Old Bibliographic Catalogues*

A. Belaïd

LORIA UMR, Campus Scientifique, B.P. 239

F-54506 Vandœuvre-lès-Nancy Cedex, France

e-mail : abelaid@loria.fr

Abstract

This paper describes a framework for retrospective document conversion in the library domain. Drawing on the experience and insight gained from the MORE project launched over the present decade by the European Commission, it outlines the requirements for solving the problem of retroconversion of old catalogues in UNIMARC format. Based on OCR technique and automatic structure recognition, the system proposes a direct schema for the conversion of references in machine readable records. Furthermore, as the system is meant for a real production chain, the paper describes the industrial constraints and gives a complete benchmark realised on this chain for 11 volumes and 4568 references. Without any manual intervention, the recognition rate of the system is greater than 75%.

Keywords : Retrospective Conversion, Library Catalogue, Reference Recognition, Structure Analysis, OCR, UNIMARC

*This work was funded by the EEC libraries programme LIB-MORE

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 [1, 6, ?, 11, ?]. 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 inaccessibility.

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 heavily on the experience and insight gained from MORE[2] [2, 3, 10] this paper outlines the main phases of retroconversion for a real production chain and states the relevant requirements of the retroconversion operation in such a chain.

2 Automatic Retroconversion

The use of generic tools to manipulate bibliographical information almost invariably poses the same problems. These are related to the following facts :

- *Heterogeneous Content.* The reference catalogue have usually been produced over a long period of time during which the cataloguing rules have changed. It contains references produced by different cataloguing agencies each applying

[1] A catalogue is a list of bibliographic descriptions of works.

[2] Marc Optical REcognition

their own rules. Many catalogues to be converted contain many different types of references: main entry references with headings representing authors or titles. Added entries by secondary authors, title, subjects, etc. Entries covering more than one reference. The system will have to be able to differentiate between these types and handle the information according to the type.

- *Typographic imperfections*: Bibliographical information is made up of text containing a large number of abbreviated words, not only in the document language but in the cataloguing language as well. It also contains numerical information, sometimes in Roman numerals, and an important quantity of names. To these must be added the multiplicity of languages used and the use of a wide range of stressed characters not in keeping with Latin writing styles. There is higher frequency of punctuation marks than in ordinary text. In addition to their natural role, punctuation marks are used as separations to delimit logical elements of information. The presence of several similar character sets such as hyphens and long dashes, parentheses and square brackets, further increase their frequency. Printed catalogues make use of typography to differentiate between sets of elements belonging to the same logical category. Unlike card catalogues, the layout is more elaborate, including systematic justification of text, variable spacing, and at times word cutting at the end of line. Some of the word cuts belong to the very publication language covered by the catalogue to be converted.

- *Linguistic variabilities*: The recognition of some fields depends on the recognition of some key words in specific lexicons. In these lexicons we can find all the cataloguing vocabulary and all the words that exist in bibliographical work titles and insertions concerning the "authorship responsibility". Punctuation is currently less reliable than that of ISBD [7]. Some words are related to the publication language (title fields, edition, address, collection) and others are related to the cataloguing language (collation and notes). Finally, all the words have to be taken into account in a complete form and also in an abbreviated form, knowing that they were not normalized at the time of the tests.

- *Higher Density of Structure*: The main problem posed by the bibliographical references resides in the density of their logical structure and the multiplicity of choice of information sequences. In fact, several cataloguing entities are optional and repetitive. These information elements are required only for the cataloguer, if the information exists in the catalogued document. Furthermore, these elements can depend on the kind of the document and of course on the kind of references, such as "monograph" or periodical publications, or as in certain catalogues, on "principal" or "secondary" reference. Finally, a practice inherited from printed catalogues is at the root of the current use of punctuation marks as a means of condensed representation of information. The ISBD normalization on the international level further reinforces this.

3 The Belgian Catalogue

3.1 Structural Aspects

The Belgian Bibliography is presented as a series of monthly catalogs on paper. Each catalog is divided into two parts. The first part contains the bibliography body while the second is filled with indexes leading to authors, subjects treated (titles, collections, rubrics in French and in Dutch, etc.).

3.1.1 Layout Reference Structure

The layout structure is very poor; it is partitioned into five areas (cf. figure 1). The first area, composed of the first line, contains on the right hand side, the "CDU" code (Classification Décimale Universelle) which gives some information about the library classification of the reference. The second area contains the reference body. It is composed of a series of fields describing the work referred in the reference such as : "*heading*" (author name or beginning of title), "*title*", "*address*", "*collation*" (material description of the work : location, editor, year, format, etc.). The body is often typed

in many lines. The third area contains the "*Collection*" field (description of the series, volume, etc.). The fourth area contains the "*Note*" field which gives information about, for example, the title (abbreviated, complete, original, etc.). These last two areas are optional and so are not always present in some references. The last area, located on the last line of the reference, contains the "*reference*", on the left hand, and the "*order number*", on the right hand.

	159.962	UDC
Liger-Belair (Gérard). Je suis fakir. ([Par] Gérard Liger-Belair). (*Verviers, Editions Gérard & C°*, 1973), 32° *carré, couv., ill.*, 158 *p.* (30 *fr.*).		Body
(Marabout-flash, 352).		Collection
[Titre introductif : Souvenirs, révélations, conseils].		Note
B.D. 14.814 352	*73-2108*	Ref

Figure 1: Example of a library reference.

3.1.2 Logical Reference Structure

The logical structure is, on the other hand, more dense. A "heading area", representing the first author or the beginning of a title is always located at the begining of the "body". As for the rest, there is an enormous number different possibilities. We can find, for example, depending on the references, "principal authors" or "secondary" (introduced by some characteristic expressions) which can be physical persons or legal entities, some "main titles", "parallel" (printed in different languages), or "partially", "sub-titles", "publishers" with their "addresses" and the "date" of publication, an area "collation" describing the characteristics of the work (number of pages, format, supporting documents, etc.).

4 Automatic Recognition System

Figure 2 shows the main phases of the recognition process of references. In the following, we briefly describe these different components for the Belgian Library.

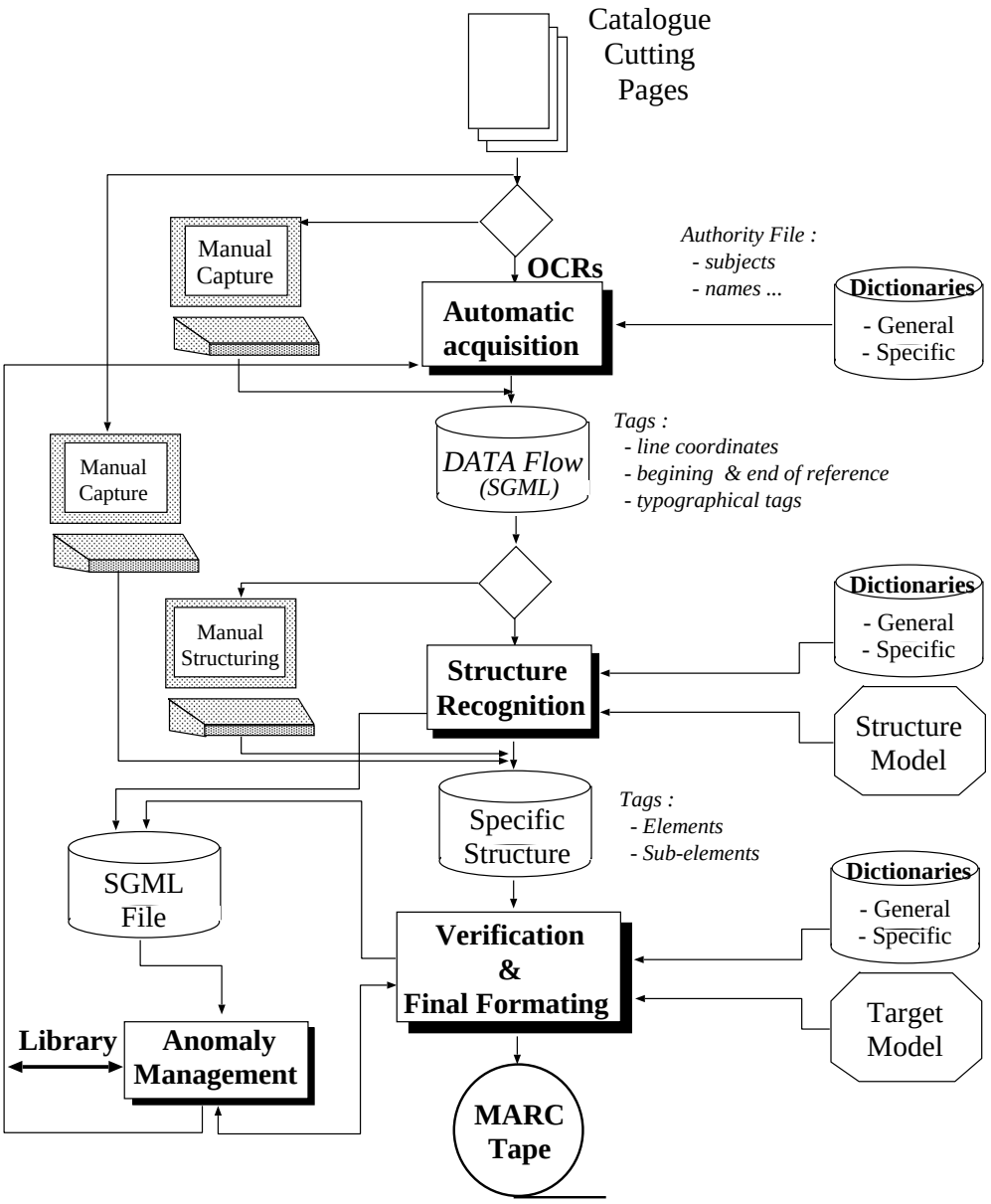

Figure 2: System Overview.

4.1 Data Acquisition

The main problems with handling catalogues are related to the automatic feeding of the pages or cards, the existence of cards printed on both sides, and the variable quality of well used machinewritten cards.

For catalogues in volume, as used in MORE project, pages are separated and feeded separately. The project has identified scanners able to handle a great quantity of pages at acceptable speed. In fact, the speed of the scanning process does not depend entirely on the scanner itself, but also on the controller page as well as the speed of the controlling system. For the resolution and because of the variations in printing quality, many tests were operated in order to determine the average resolution (here equal to 400 dpi) which can be used for all the catalogue pages without changing during the feeding.

Data acquisition also includes data formatting. Being individually pasted into pages, the reference images are altered (skew angle, font changing, cut or connected characters, etc.). Specific algorithms had to be developed in order to take into account these particularities [2]. At the end of this process, each reference is extracted from the page image and given to the recognition system as a list of successive lines.

4.2 Text Conversion

Each reference is passed through a series of commercial OCRs. The results of these OCRs are combined to obtain the best response. The reason for this is that references contain a lot of different symbols (such as punctuations, indices, exponents, and multilingual words typed in different sizes and styles) which are very difficult to recognize using only one OCR. We thought that combining the results from different specialized OCRs will give a maximum of information on the text, its style, its language, and on its separators.

The result of these tasks is a data flow containing the reference text coded in SGML [8]. The tags separate the lines and different information such as style or lexical

class corresponding to each word (token). Figure 3 shows the flow corresponding to the reference of figure 1.

The reference is located in this flow between two successive tags "<NOT" and "</NOT>". Useful tags for the document analysis are "LEX" which gives the lexicon affiliations of words, "I" for italic style, "B" for bold style, and "S" for the number of spaces. The defaults style is standard and as such not tagged. It is possible to have some errors during this first conversion (especially in recognition of style and punctuation). For example, the exponent "o" in c^o is replaced by the character "o". Anther initial recognition error concerns the style of the end of the secondary title which is identified as "standard" instead of "italic".

4.3 Structure Modelling

Knowing that the problem is to find the sub-fields within reference areas, the model specification concentrated on the description of sub-field properties, by the distinction of their typographic styles, the existence of particular words or group of words and their appearance in certain lexicons, and essentially their limits (type of initials and finals such as capital letters, particular words or type of punctuation separating the sub-fields).

The model is given by a context-free grammar written in the EBNF formalism. The format of a production rule is as follows :

Term	::=	Constructor subordinate_Objects[Qualifier] |
		Constant | Terminal
Constructor	::=	seq_td | seq_lr | seq | aggr | cho | import
Separator		Name subordinate_Objects
Attributes		[Name Weight]$^+$

```
<DOC   % image source        1st reference last reference  Directory
TY=N PROV=ENRLEX EG=OK NPN=2085 NDN=2114 IMA=users/brb/juin73/images>

     <PAG % number              bounding box
    NP=1 NOM=0008.ima> <COL XHG=63 YHG=1900 XBD=1027 YBD=2912>

        <NOT % number coordinates
            NON=2108 EN=OK>
         <LIG XHG=870 YHG=2215 XBD=1000YBD=2266 YBSL=2256 ST=t>
         <REDF=85.69>159.962</LIG>
         <LIG XHG=149 YHG=2260XBD=1001 YBD=2313 YBSL=2298 ST=p><B>Liger-Belair</B>
           <I>(Gérard).</I><LEX L=GFR,GNL><REDF=50.00>Je <LEX L=GFR>suis <LEX
           L=GGB,GFR>fakir.<LEX L=GGB,GFR,GNL><RED F=99.99>([Par] <REDF=99.97>Gé-</LIG>
         <LIG XHG=148YHG=2304 XBD=1002 YBD=2356 YBSL=2342 ST=p>rard Liger-Belair).<RED
           F=89.99> <I>(Verviers, <LEX L=GGB> <RED F=100.00>Editions <LEX L=GNL>
           <REDF=99.99>Gérard</I> <I>\&amp;amp; </I></LIG>
         <LIG XHG=151 YHG=2350 XBD=1000 YBD=2403 YBSL=2388 ST=p>C0,<RED F=83.33>1973),
           320<LEX L=GFR,GNL><I>carré, <RED F=66.66>couv.,ill.,</I><RED F=99.97>158
           <RED F=25.00>p.30 <I>fr.</I>).</LIG>
         <LIG XHG=149 YHG=2406 XBD=549 YBD=2457 YBSL=2443 ST=p><LEX L=GGB,GFR,GNL>
           Marabout-flash,352).</LIG> <LIG XHG=148 YHG=2447 XBD=1001 YBD=2499 YBSL=2485
           ST=p> <LEXL= GFR><RED F=99.99>[Titre <LEXL=GFR>introductif:<LEX L=GGB,GFR>
           <RED GFR>F=89.99>Souvenirs,<LEX L=GFR>rivilations, <LEX GFR> L=GGB,GFR,GNL>
           con-</LIG> <LIG XHG=149 YHG=2494 XBD=245 YBD= 2545
           YBSL=2530 ST=p>seils].</LIG>
         <LIG XHG=148 YHG=2546 XBD=1002 YBD=2599 YBSL=2584 ST=t> <RED F=43.75>B.D.
         14.814 <REDF=99.97>352<SN=15><I>73-2108</I> </LIG>
       </NOT>
     </PAG>
</DOC>
```

Figure 3: Flow of the reference given in figure 1.

4.3.1 Constructors and qualifiers

A term, the left hand of a rule, can be either simple (constant or terminal) or composed of subordinate objects. In the last case, a constructor describes the relationship between objects. The constructor precises the order of the appearance of subordinate objects such as SEQUENCE : top-down (*seq_ td*), left-right (*seq_ lr*) or logical (*seq*), AGGREGATE (*aggr*) or CHOICE (*cho*). A special constructor "*import*" is used to inherit for the term some or the total description of another existent and similar term. Furthermore, to express the object occurrence in the term, each object may be accompanied by a qualifier such as OPTIONAL (*opt*), REPETITIVE (*rep*) and OPTIONAL-CONDITIONAL (*optc*) precising the condition under which an object may appear.

4.3.2 Separators

As the structure is not enough sufficient to characterize the fields and to separate them, the limits between consecutive field are introduced to reinforce the field description. Separators can be *specific punctuation marks* as point, comma, bracket, parenthesis, etc. *Mode changing* (Capital letter in the beginning of the field), numeric area, *font style changing*, etc.

4.3.3 Attributes

Because of the weakness of the physical structure and the multitude of choices represented in the model, we add to the previous description some attributes given by the library specification to better precise the description of the reference components.

Several kinds of attributes have been defined, among them, *Type* (string, line, word, char, etc.), *Mode* (capital, numeric, alphabetic, punctuation, etc.), *Style* (bold, italic, standard, etc.), *Position* (beginning of line, inside, end), *Lexicon* affiliation (author index, countries, towns, abbreviations, articles, etc.), *Weight* which specifies the degree of importance of subordinate objects, etc.

4.4 Structure Analysis

The structure analysis is based on the model and on the entry data flow. For the model, the grammar rules are converted by a compilation procedure into a working structure. The input data flow is also reorganized into a working table by a filtering task. This table contains useful tokens extracted from the flow such as style, token, size, etc. and a pointer to a buffer containing the corresponding content. Figure 4 summarizes the principal functioning mode of the structural analysis.

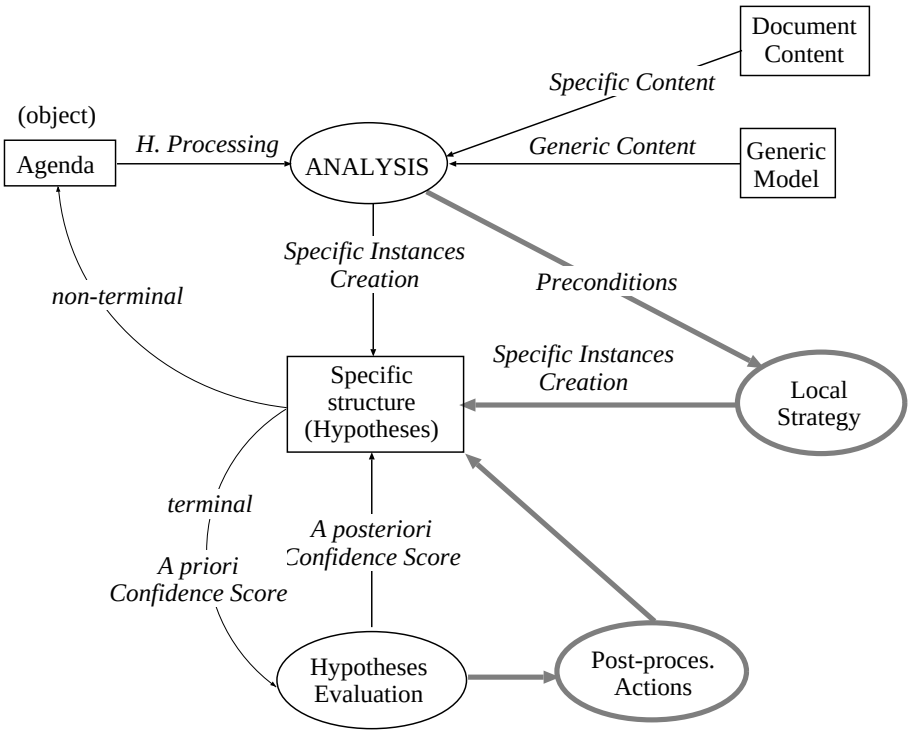

Figure 4: Functioning Scheme of the Structural Analysis.

4.4.1 Model Compilation

This step allows to adapt the analysis process to the application model. It generates working files containing the specific terms, actions and attributes for the application. References as well as indexes (containing authors and subjects) are modelized as three different applications. During the analysis, these files are converted into dynamic tables of terms where the entries correspond to term codes. Each term is given by a

list of characteristics gathered in a characteristic table. This allows the system to read rapidly the characteristics of each analyzed term.

4.4.2 Hypotheses Management

At each step of the analysis, the system proposes for the current object different choices for its decomposition (analysis). These choices which are not already verified are called *hypotheses*. We use a structural tree to store these hypotheses. A confidence score (*a priori* score) is computed for each generated hypothesis. This score allows to choose, in an *agenda*, among all the current hypotheses which one to process first. The score computing is initialized by the weights given in the model for the current object (for its attributes and subordinate objects). This score is successively updated as the hypotheses are verified and becomes a recognition score. At the end of the analysis, each tree path corresponds to a possible structure (for the input reference) weighted by a recognition score. This qualitative reasoning allows to reduce errors and to isolate possible doubtful areas.

The hypotheses are chosen from the agenda according to the importance of their *a priori* scores (apr). Thus, the analyzer is said to function in an opportunistic mode. Terminal terms (tree leaves) are directly verified. On failure or success, the *a priori* score is up-dated and becomes an *a posteriori* score (aps) which is propagated from bottom to top in the corresponding path (see figure 5).

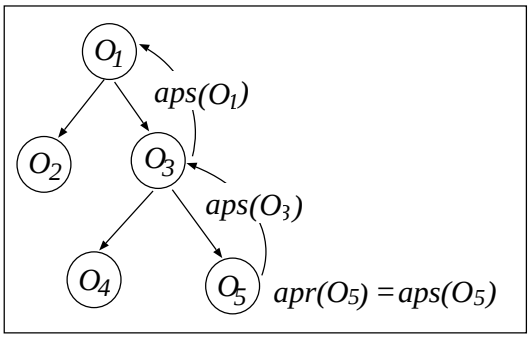

Figure 5: Score Propagation.

The *a priori* score of a current object o depends on the result of the observation

of its attributes (a_t) for each token t_k of o ($C(a_t, t_k)$). It is also function of the tokens length (L) and of the weight W of each attribute.

$$apr(o) = \frac{\sum_{a_t} \sum_{t_k} C(a_t, t_k).W(a_t).L(t_k)}{\sum_{a_t} W(a_t).L(o)}$$

The *a posteriori* score of o is updated from the *a posteriori* scores of its subordinate objects (o_i) by taking into account their corresponding weights (p).

$$aps(o) = \frac{\sum_{i} p(o_i).aps(o_i).L(o_i)}{\sum_{i} p(o_i).L(o_i)}$$

With this method, the different objects and attributes influence the final score according to their importance given in the model (weight) and in the input data string (length).

4.4.3 Local Strategies

We show here some examples of actions executed before the general analysis. Depending on the status returned by these actions, they can play the role of pre-conditions, in such a case, the analysis continues normally, or of local strategies, stopping general strategy. When an action plays the role of a local strategy, it has the control of new hypotheses (possible decomposition of the current object) to submit.

Author searching . In library references studied, it was fitting to identify the secondary authors of the publication. Contrary to principal authors, secondary authors are introduced by a particular expression ("par", "introduit par", "illustration de", etc.). It suffice to recognize this expression and to verify that what follows corresponds to an author. The problem here comes from the fact that authors are not necessarily presented in the same format in indexes and within the references. Furthermore, the list of expressions is not exhaustive. It is fitting to apply a fine syntactical analysis to recognize these secondary authors, as shown by this example:

ZATZME	::=	Seq ZAT ZME?
Sep		Ponct1
Action		+InitAuteurs(Expressions,IndexAuteurs,...)

Parameters `Expressions`, `IndexAuteurs`, etc. correspond to a list of lexicons used by the local strategy `InitAuteurs`

Style Searching . In order to minimize the hypotheses number submitted during the analysis, we have developed some heuristics allowing to cut an area by searching typographic characteristics. The following example shows an action which cuts the current object at the first punctuation preceeding the beginning of the italic area. This pre-cutting allows, in fact, the analysis part by part and makes the economy of several hypotheses which, in all cases, will failed.

ZATX	::=	Seq ZATZME ZIC
Sep		Ponct
Action		+SplitField(italic,Ponct)

Suppression of Useless Hypotheses . Some objects to recognize are easily identifiables (for example, a town found in town dictionary). In this case, it is interesting to delete all hypotheses in the queue which contain, in an another context, the same search area. The action `KillAmbiguities` in the example below is activated if the object `MotEd1` is perfectly recognized. It goes through the specific structure tree and suppresses all waiting hypotheses that contain the same content as `MotEd1` and that do not belong to other instances of `MotEd1`. This action may be used carefully because every new hypothesis on this area, which is not an instance of `MotEd1`, will be forbidden.

MotEd1	::=	Terminal
Alex		Edition //opl. tir. uitg. éd, etc.
Nature		mot
Action		KillAmbiguities() RestituteField()

4.4.4 Output Flow Restitution

When the analysis is finished, it is necessary to go through the structure tree to produce a structured flow corresponding to the result. This running is realised depth first. The structure is represented by a mark up format like SGML. Each tagged field is given by a confidence score.

Figure 6 gives the analysis result of the reference given in figure 1. All the sub-fields were correctly localized. They are coded and tagged in UNIMARC. "QSTR" indicates the evaluation score (maximum 10 000).

```
<675 I=bb QSTR=10000> <$a QSTR=10000>159.962</$a> </675>
<200 I=0b QSTR=9834>  <$f QSTR=9487>Gérard Liger-Belair</$f>
       <$a QSTR=10000>Je suis fakir</$a> </200>
<700 I=b0 QSTR=10000> <$a QSTR=10000>Liger-Belair</$a>
       <$b QSTR=10000>Gérard</$b> </700>
<210 I=bb QSTR=9705>  <$a QSTR=10000>[Verviers]</$a>
       <$c QSTR=9519>[Editions Gérard & Cî]</$c>
       <$d QSTR=10000>[1973]</$d> </210>
<215 I=bb QSTR=9750>  <$d QSTR=7353>32f carré</$d>
       <$c QSTR=8601>couv., ill.</$c>
       <$a QSTR=10000>158 p.</$a> </215>
<010 I=bb QSTR=10000> <$d QSTR=10000>30 BEF</$d> </010>
<225 I=2b QSTR=10000> <$a QSTR=10000>Harabout-flash</$a>
       <$v QSTR=10000>352</$v> </225>
<517 I=0i1 QSTR=10000><$a QSTR=10000>Souvenirs, révélations,
       conseils</$a> </517>
<900 I=bb QSTR=9772>  <$a QSTR=10000>B.D. 14.814 352</$a>
       <$b QSTR=9285>73-2108</$b> </900>
```

Figure 6: Structural Analysis Result of the Given Reference.

In the event of errors, the system generates a fictive UNIMARC code 903 which it uses to demarcate the zone it should have recognized for a field but which does not quite fit the characteristics as specified by the user. This helps in modifying the model to take care of exceptional cases or to really determine that the reference was badly

formed as a result of OCR errors, the printers devil or outright bad transcription of the reference.

When the system finds more than one solution for a given zone, it equally generates a fictive UNIMARC code 902 that it puts around each of the possible solutions which are then presented to an operator who has to make a choice.

4.5 Results and Discussion

The global evaluation of the prototype is made up after the treatment of all the 11 catalogue volumes of the Belgian Library, e.g. 4548 references. The volume of june is discarded because it was used in the first phase for the control quality evaluation. Performances will be discussed in the following points:

- *OCR/ICR.* 6.69 doubts per reference for only the body of the bibliography and 9.87 doubts per reference if include the rest: main and secondary entries.

- *Structure Recognition.* 67% of references have been recognized automatically by the system.

- *Attribution of language and country codes.* 77.7% of references have their codes created automatically by the system.

  However, considering all the operations of correction provided for the automatic structure and code generation, as well as the corrections effected on references with "risk", only 47.5% of references have been entirely recognized automatically without any manual intervention.

- *Speed up.* The speed up of the prototype is about 1'30 per notice. This depends on the complexity of the structure and the correction procedures launched by the system.

  The table 1 give the time spent by the system for the different modules for all the 4548 references.

Automatic Module	Time in hours	% total time
OCR/ICR	16.5	14%
Structure Recognition	99	83.5%
Others	3	2.5%

Table 1: Time spent by the Automatic Processing.

Manual Intervention

The table 2 gives statistics on manual interventions either for OCR correction or for re-treatment of the structure or the codes generation.

Module	Defect Cases	manual interventions
OCR/ICR	44920 doubts examined	9.87 doubts per reference
Structure	1494 references unstructured totally or partially	33% of references
Codes Country Language	1014 references with in less one non-generated code	22.3% of references
Structure + Country Codes + language	2083 references corrected in less one time	52.5% of references
Anomaly after Quality Control	246 references returned to the Library	5.4% of references

Table 2: Statistics on Manual Interventions.

4.5.1 Problems encountred

The main problems encountred in the technical realisation of this project concern the treatment by OCR of the bibliographic information, the structure modelling of the

Library catalogues and the moving to an industrial production.

OCR and Bibliographic Information . The variability of the typography seriously handicaped the straightforward conversion of the bibliography by OCR techniques. Many reasons have been signaled in section **??**. The main deficiencies encountred in the Belgian catalogues are:

- *Typographic aspects*: connected characters for bold data and use of standard numeric characters within textual areas in italic;
- diacritics added by hand,
- use of long dash line for all the parallel areas and for someones of the collection sub-areas;
- intensive use of square brackets.

3.6% of references were returned to the Library because of the presence of non latin charaters to transliterate by the Library.

Bibliographic Catalogue Modelling . This problem is already encountred in a traditional retrospective conversion process in which the Library writes specifications for the conversion of its catalogue. These specifications must be validated on several references and modified in a continuous manner in order to adjust the model in order to take into account the exceptions and new encountred problems. In the MORE project, these specifications were very detailed but with a point of view oriented more for the cataloguing than for the automatic conversion by computer. This needed a more adaptation of the two populations (from Library and Laboratory) to better harmonize their dialogue.

In the other hand, the structure of the bibliographic information is very difficult because of this three main characteristics:

- Catalogues are written before the apparition of the standard ISBD, leading to different structures with particular rules for layout and punctuation;
- The correspondance between the pré-ISBD cataloguing rules is sometimes difficult to establish with the UNIMARC format for the transcription of the titles areas and responsability mentions. This difficulty is lower in USMARC where the main cataloguing elements are grouped into three sub-areas non-repetitive in only one possible sequence. In UNIMARC, the same information can be shared in six sub-areas all of them repetitive and with a high number of possible sequential combinations. The same difficulty is encountred in the modelling of the edition and collection areas.
- Some catalogues has a cataloguing with hierachical levels in the case of the monography in severals volumes, with significant titles for each volume. The model has to take into account some specific considerations for the treatment of these volumes. In the Belgian Library, the cataloguing of volumes belonguing to a monography in many volumes, as well as the treatment of collectif titles was very difficult to model and had created some anomalies returned to the client (the Belgian Library). The presence of many official languages in this bibliography (French ans Dutsh) have also led to a great number of parallel mentions in titles and notes, with a very complex structure of titles and responsability mentions. 78.4% of the 246 references have been returned for manual control because of bad structure, 46.53% for the title structure and 12.25% for the validation of collectivies authors.

5 Conclusion

The aim of this paper was to enhance understanding of the issues involved in the retroconversion process and to show the advances in the field of character recognition and structure interpretation and their usefulness in the development of solutions to the retroconversion problems.

The system presented here gave good results on tested library references. The errors encountered were due to incomplete specification (reference not falling into any of the categories we were provided information on) or OCR errors. The ambiguities encountred were partly due to a combination of incoherence in the specification (which allows different legal segmentations) as well as OCR substitution errors. The evaluation allows the observation of the quality of each reference and each field in the reference and allows the user to intervene or not for manual correction.

References

[1] Beaumont J., Cox J. P.: Retrospective Conversion. A practical Guide for Libraries. Meckler, Westport/London. 1989. 198 p.

[2] Belaïd A., Chenevoy Y., Anigbogu J. C.: Qualitative Analysis of Low-Level Logical Structures. In *Electronic Publishing EP'94*, volume 6, pages 435–446, Darmstadt, Germany, April 1994.

[3] Belaïd A., Chenevoy Y.: Document Analysis for Retrospective Conversion of Library Reference Catalogues, ICDAR'97, ULM, Germany, August 1997.

[4] CEC, DG XIII B: Libraries Programme, Telematics Systems in areas interest 1990-1994: Libraries, Synopses of Projects. http://www2.echo.lu/libraries/en/libraries.html

[5] Council of Europe: Guidelines for Retroconversion Projects prepared by the LIBER Library Automation Group, Council of Europe, Council for Cultural Co-operation, Working Party on Retrospective Cataloguing, 1989.

[6] Crawford R. G., Lee S.: A prototype for fully Automated Entry of Structured Documents. In The Canadian Journal of Information Science, (15)4, pp. 39–50, 1990.

[7] ISBD (G): General International Standard Bibliographic Description: Annotated Text. Prepared by the Working Group on the General International Standard Bibliographic Description set up by the ILFA Committee on Cataloguing. London, 1977. 24 p.

[8] International Standard Organization: Information processing, text and office systems, standard generalized markup language (sgml). Draft International Standard ISO/DIS 8879, International Standard Organization, 1986.

[9] ISO 8859-1 to 7: Information Processing - 8-bit single-byte Coded Graphic Character Sets - Part 1-7: Latin Alphabet No. 1 to 7. International Standards Organisation. 1987.

[10] Lib More: Marc Optical Recognition (MORE), Proposal No. 1047, Directorate General XIII, Action Line IV: Simulation of a European Market in Telematic Products and Services Specific for Libraries, 1992.

[11] Schottlaender B.: Retrospective Conversion: History, Approaches, Considerations. Haworth Press, NY. (1992).

[12] Süle G.: Bibliographic Standards for Retrospective Conversion. In IFLA Journal (16)1, pp. 58–63, 1990.

[13] Valitutto V. and Wille N. E.: A Framework for the Analysis of Catalogue Cards. FACIT Technical Report no 2). Statens Bibliotekstjeneste, Copenhagen. October 1996.