RETROCONVERSION THROUGH OCR/ICR AND STRUCTURE RECOGNITTION: THE MORE (marc Optical REcognition) - project.

Willy VANDERPIJPEN

I. GENERAL PRESENTATION OF THE PROJECT

1. Purpose of the project

The MORE project was selected after the first cal for proposals (1991) of the Libraries Programme of the European Commission. It is classified under Action Line 4, Theme 17 «New bibliographic products and services applying internationally recognised standards». The objective of the project is a feasibility study of OCR/ICR technologies as an approach to the retrospective conversion of library catalogues and the development of specific software for structure recofnition of bibliographic descriptions, through a) the development of a prototype tool, b) the integration of the prototype in a production environment, c) test and assessment of methods under real conditions.

Using existing OCR/ICR-methods, a bibliographic description is entered, of which the structure is recognised and which will be presented in UNIMARC-format.

The source catalogue is the Belgian National Bibliography, year 1973, which is composed of pre-ISBD records for which written cataloguing rules were applied, and which is close to the automated bibliography (starting in 1975) used to create specific dictionaries (a.o. a dutch one) for the project.

2. Partners, duration

- Jouve Sysrtems d'Information, Paris, project coordinaton. Jouve has already been involved for several years in retrospective conversion of library catalogues and in Intelligent Character Recognition, both in a production environment.

- The Royal Library, Brussels, interested in applying new technologies for the retrospective conversion of the printed national bibliography.

- The CRIN (Centre de Recherche en Informatique de Nancy) which is a University research laboratory involved in structure recognition.

The initial project duration was one year, but took a sligtly longer run for technical reasons.

3. The source catalogue: the national Belgian Bibliography (1973).

The Belgian Bibliography has, for the last 50 years before automation, the same physical structure and theoretically applies the same written cataloguing rules. Each year is composed of two parts: the national bibliography itself published in 12 montly issues plus a cumulative annuel index issue, and a special issue devoted to Belgian authors published abroad and to foreign publications on Belgium. The MORE project focuses on the national bibliography itself (monographs and serials published in Belgium). For this part, the languages of publication are Dutch, French, German and English. As Belgium is a bilingual country, the cataloguing languages are Dutch and French, which doubles the cataloguing vocabulary to be treated.

Each issue contains two parts:

- the bibliographic corpus contains the bibliographic descriptions classified in the main UDC classes with chapter fitles and then diled according to the detailed UDC classification number. The bibliographic descriptions are numbered sequentially over the year.

Each issue contains three indexes:

- an alphabetical authors-title index with entries for all the autors, for anonymous titles (more than two authors or collective publications), for serials titles, and cross references....

- an alphabetical French subject headings index

- an alphabetical Dutch subject headings index

All the index entries include sequential numbers of the corresponding bibliographic description.

The page layout is organised into comumns, with a specific first page layout for each part. Records can be split into two columns on the same page on on two pages. Standard, bold and italic styles are used to emphasise the logical structure. The text is theoretically, only in Latin alphabet with diacritics. The

font used was not large, and som diacritics were added by hand before printing (for instance hacek). Justification to the right margin is applied and words are hyphenated at the end ot lines.

4. Conversion: general rules

The conversion is done in full UNIMARC format close to the implementation of the current cataloguing on the VUBIS system of the Royal Library, as it is defined for specific national needs such as the multingual environment.

Retrospective conversion is done without any attempt to create the links with the national authority files which are updated continuously on the current system. As usual in a retrospective conversion process, these links will be created when loading the result into the current system. The authority files are only processed to provide appropriate specific dictionaries.

All the information of the bibliographic descriptions is converted, except the mail adress of the publisher which has te be dropped. When an information is not following the current cataloguing rules, a separate field number is used (for instance, the separate subject heading field for retroconverted data).

As described in the cataloguing rules of 1973 and before, author's names may appear on two different positions in the bibliographic description:

- heading form: if the authors od the first responsibility statement are in the headinh form, this responsibility statement is restituted at the appropriate place according to ISBD-rules (titel and responsibility-field). The heading form is mentionned in the appropriate UNIMARC tag.

- for the secondary authors, a heading form is created, including the relator code, using both specific dictionaries of keywords and the author-titel index where the name is in the heading form.

The serial entry is cheked with the author-title index where the name is in the heading form.

The subject indexes are converted into subject headings and added to the appropriate bibliographic description(s).

5. General results of the project

The structure-recognition software made it possible to recognize entirely the structure of 67% of the bibliographic descriptions. For 27,6% of the descriptions, UNIMARC codes have been added by keyboarders of the company Jouve. 5.4% of the descriptions have been sent as anomalies to the library.

6. Example of a printed and authomatically formated bibliographic description.

Document:

641.5 (083.12)

Courtine (Robert J.). La cuise francaise et internationale. Paris-Brux., Elsevier Sequoia, (1972), 8o, couv,, ill., 357p (Rel. 395 fr.). (Multiquide Elsevier).

B.D. 26.854 4

73-2268

Notice: 732268

Guide: 00647 nam 2200229 n 450

001		732268
010		$d Rel. 395 BEF
100		$a 1993100d1972 u u0???y0103 a
101	1	$a FRE
102		$a fr
200	0	$a La cuisine francaise et internationale $f Robert J. Courtine
210		$a Paris $a Brux. $c Elsevier Sequoia $d (1972)
215		$a 357 p. $c couv., ill. $d 8o
300		$a Multiguide Elsevier
675		$a 641.5 (083.12)
700	0	$a Courtine $d Robert J.
900		$a B.D. 26.854 4 $b 73-2268
901	f	$a Cuisine
901	f	$a Sciences appliquees
901	n	$a Kookkunst
901	n	$a Toegepaste wetenschappen

II. SPECIFIC DICTIONARIES AS TOOLS FOR STRUCTURE RECOGNITION

The automatic recognition of the structure of a bibliographical description is a form of artificial intelligence. To reach that aim the computer has to be hold the relevant elements of the fields and subfields of the bibliographical description to be converted.

ISBD-descriptions have the advantage of a precise punctuation. Unfortunately this is not the case for order descriptions - and these happen to be the first target of retroconversion. In that case there is mostly a need of other recognition marks. The formal outlook of type characters can be of use in a number of cases (italics, bold type etc.). In the case of the «Belgian Bibliography», however, it has been proved that this aspects was not distinctive in all cases, partly while not always unambiguous.

With simple descriptions punctuation and typeface will do. This is the case with the following example (cfr. 73-11).

Document:

Omnibus van de Vlaamse humor. Jos Ghysen,
Louis Verbeeck, Gaston Durnez, jan de Spot.
4de druk. Brussel, Reinaert Uitgaven, (1973),
12o, 576 blz. (Geb. 200 fr.).
(Reinaert-omnibussen).
B.D. 9789d 163 73-2362

Notce: 732362

Guide: 00671nam 2200ss9 n 450

001		732362
010		$d Geb. 200 BEF
100		$a 1993100d1973 u u0???y0103 a
101	1	$a DUT
102		$a bl
200	0	$a Omnibus van de Vlaamse humor $e jos Ghysen, Louis Verbeeck, Gaston Durnez, Jan de Spot
205		$a 4de druk
210		$a Brussel $c Reinaert Uitraven $d (1973)
215		$a nlz. $d 12o
300		$a Reinaert-omnibussen
675		$a 8393 (DE Spot, J., + ...2)
900		$a D.B. 9789d 163 $b 73-2362
901	f	$a Litterature
901	f	$a Litterature neerlandaise
901	n	$a Letterkunde
901	n	$a Nederlandse letterkunde

There is only one author's name as a main entry. It is printed in bold type, so it can easily be recognized and will be converted into the author field following the title, as well as in the field of the entry-from, (tag 700). The title has a simple structure and follows the main entry after a period. The field of the bibliographical address can be recognized as it is printed in italics except the date. The collational formula can also easily be recognized, as well as the Royal Library's at the margin. Most of the other descriptions, however, are not as distinctive as the present one.

Often the next itself is a distinguishinf mark. In certain cases that will be various words or expressions. The functional lists of those words or expressions are called «dictionaries». They will be introduxed in the system. The confrontation of those «dictionaries» and the texts of the bibliographic descriptions read by means of «scanning» and OCR has proved to be a strong tool for automatic recognition and hence also for the encoding of fields and subfields.

For the MORE project the year 1973 of the «Belgian Bibliography» has been chosen as a test set. While describing the specifications of that volume, we built «dictionaries», some very restricted, other ones more extensive. That was necessary when there were no other distinguishing marks. In some cases it was useful as additional marking.

In the following, quite simple, example (73-16), two texts are vital for the identification:

- «Vertaling» (translation) introduces a secundary author, viz the translator. The piece of tect «vertaling» is included in the «dictionary» of tect fragments announcing secondary authors. In that way the «mention of the author» (introcuction text + author's name as printed in the publication) is converted into UNIMARC tag 200 $g and the author's name in its entry form will be converted into tag 702 together with his function-code (730).

«Oorspronkelijke titel» (original title) identifies the annotation (title annotation). This can be recognized through formal elements (square brackets, starting point at margin). The spscific kind of annotation, however, will only be clear through the introuctory formula, and this case «original title».

Morris (Desmond). Intiem dedrag. (Vertaling:
Riet Lyten). Utrecht-Antw., A.W. Bruna &
Zoon, (1972), 8o, omsl., 240 blz. (220 fr).

(Bruna Boeken).

(Oorspr. titel: Intimate behaviour).

B.D. 20.671 128 73-16

Notice: 7316

Guide: 00554nam 2200181 n 450

001		7316
010		$d 220 BEF
100		$a 1994020d1972 u u0???y0103 a
200	0	$a Intiem gedrag $f Desmond Morris $g vertaling: Rift Leyten
210		$a Utrecht $a Antw. $c A.W. Bruna $ Zoon $d (1972)
215		$a 240 blz. $c omsl. $d 8o
225	0	$a Bruna Boeken
312		$a Oorspronkelijke titel: Intimate behaviour
675		$a 159.9 + 30) : 599.9
700	0	$a Morris $b Desmond
702	0	Sa Leyten $b Riet $4 730
900		$a B.D. 20.671 128 $d 73-16.

A list of «articles» was necessary to distinguish between author's names or titles when the main entry is followed by a word/words between brackets: only the first word of a title will be followed by the article between brackets. The brackted element following an author's name is the forename.

Articles are included in a «dictionary».

There was also need for a «dictionary» of «appelations», since these data are to be left out is ISBD descriptions.

Saint's name, however, are included in the entry for author's names.

They too are stocked in another, limited, «dictionary».

The most expented list was formed by the introductory phrases of secondary authors.

Those authors were only recognizable by means of their introductory phrases. Listing them proved to be necessary, since they were not identifiable through unambiguous punctuation sings. As author's hames happen to occur in various forms (with one or more initials, simple or complex structure), the start of a «name» will only be clear when the end of the intriductory phrase will be clear.

One description may also have various sets of secondary authors, each being announced by a proper phrasing.

That «dictionary» has proved to be of paramount importance. It also allowed us to quality the author's names in field 701/702 of UNIMARC with the UNIMARC function code. In the case of the 1973 volume this «dictionary» consists of 210 Dutch phrases, 202 French ones and 57 ones in other languages, totalling 469 different phrases.

The framming of such long lists is very time-consuming, since it implies reading all descriptions in advance. the lists will have to be added to continuously: partly because publishers never used to cate for bibliographical uniformity, partly because compilers of library catalogues always thougt it necessary to copy the exact phrasing of the publication's tile-page. Life would be easier, if ever there had been standard phrases in bibliographical descriptions.

A list of prefixes also helps in identifying surnames consisting of several parts, e.g. when the phrase introducing the author's name consists of several parts, one of which could be part of the author's name.

Other, shorter, «dictionaries» are:

- phrases introducing/indicating «editions, inpressions» and the like;

- annotations on titles;

- annotations on content.

Authority-control and linking with authority-records from UNIMARC fields 700, 701 and 702 will be possible when the bibliographical system to be loaded with converted descriptions possesses those possibilities. This is the case with the Roayal Library's VUBIS/NEWWAVE system. The '700'-tags of the new UNIMARC discriptions will be linked to the authority records of VUBIS-KB. When a name occurring in one of the '700'-tags also occurs in the authority file, the new bibliographical description will be linked to that authority record. The system can also doublecheck in case of homonyms. This will reduce the presence of faults, provided the system explicitly signals the presence of homonyms.

Αναπροσαρμογή δια μέσου OCR/ICR και δομή αναγνώρισης : το MORE (Marc Οπτική Αναγνώριση) - σχέδιο

Willy Vanderpijpen.

I. Γενική παρουσίαση του σχεδίου.

1. Σκοπός του σχεδίου.

Το σχέδιο MORE επιλέχτηκε, το 1991 σύμφωνα με την αρχική πρόταση της Ευρωπαϊκής Κοινότητας για το πρόγραμμα Βιβλιοθηκών, που εντάχθηκε στη Γραμμή Δ, Θέμα 17: «Νέα Βιβλιογραφικά προϊόντα και υπηρεσίες που εφαρμόζουν διεθνώς πρότυπα αναγνώρισης».

Ο αντικειμενικός σκοπός του σχεδίου είναι να διερευνηθεί προσεκτικά από τις Τεχνολογίες του OCR/ICR, η αναδρομική μετατροπή των καταλόγων της Βιβλιοθήκης και η ανάπτυξη ειδικών λογισμικών της δομής αναγνώρισης από βιβλιογραφικές περιγραφές από:

α) την ανάπτυξη ενός πρωτότυπου εργαλείου

β) την ολοκλήρωση μέσω της πρωτοτυποποίησης σε περιβάλλον παραγωγής.

Το More είναι πρόγραμμα που επικεντρώνεται στην Εθνική Βιβλιογραφία (μονογραφίες, περιοδικές εκδόσεις στο Βέλγιο). Για αυτό το μέρος οι γλώσσες της έκδοσης είναι Γερμανικά, Γαλλικά, Δανέζικα και Αγγλικά. Στο Βέλγιο χρησιμοποιούν δύο γλώσσες, γλώσσες καταλογογράφησης είναι Δανέζικα και Γαλλικά, και οι δύο γλώσσες εμφανίζονται στο λεξιλόγιο της καταλογογράφησης.

Κάθε έντυπη έκδοση αποτελείται από δύο μέρη:

- Από το Βιβλιογραφικό σώμα, όπου περιέχονται βιβλιογραφικές αναφορές, ταξινομημένες με UDC και από τίτλους με κεφαλαία γράμματα που συμφωνούν ταξινομικά με τους αριθμούς του UDC.

- Κάθε θέμα περιέχει τρία ευρετήρια

. ένα αλφαβητικό συγγραφέων - τίτλων

. ένα αλφαβητικό Γαλλικό Θεματικό ευρετήριο

. ένα αλφαβητικό Δανέζικο θεματικό ευρετήριο

Ολα τα ευρετήρια καταγράφονται στην Βιβλιογραφική περιγραφή σε διαδοχικούς αριθμούς αναγνώρισης της έντυπης Εθνικής Βιβλιογραφίας.

- Το κέντρο Έρευνας της πληροφορικής στο Nancy που είναι ερευνητικό πανεπιστήμιο, περιλαμβάνεται στη δομή αναγνώρισης.

Το αρχικό στάδιο είχε διάρκεια ένα χρόνο, αλλά για τεχνικούς λόγους παρατάθηκε.

3. Πηγή καταλόγου: Εθνική Βελγική Βιβλιογραφία (1973).

Η Βελγική βιβλιογραφία είχε τα τελευταία 56 χρόνια, πριν την αυτοματοποίηση την ίδια φυσική δομή, και τους ίδιους έντυπους κανόνες καταλογογράφησης.

Κάθε έτος χωρίζεται σε δύο μέρη:

α) την Εθνική Βιβλιογραφία που εκδίδεται σε 12 μηνιαία τεύχη και το πρόσθετο ετήσιο θεματικό ευρετήριο.

β) ένα θεματικό - ειδικό ευρετήριο των Βέλγων Συγγραφέων που έχουν εκδόσει βιβλία σε άλλες χώρες και των ξένων συγγραφέων που έχουν εκδοθεί βιβλία τους στο Βέλγιο.

γ) την εκτίμηση και εξέταση των μεθόδων υπό πραγματικές συνθήκες.

Μια βιβλιογραφική περιγραφή, σύμφωνα με τις υπάρχουσες μεθόδους του OCR/ICR καταχωρείται, αρκεί η δομή της να είναι αναγνωρίσιμη ή να μπορεί να παρουσιαστεί σε μορφή UNIMARC.

Η πηγή του καταλόγου είναι η Βελγική Εθνική Βιβλιογραφία του 1973, που είχε συνταχθεί, πριν το ISBO, με αναφορές σε οποιοδήποτε γραπτό κανόνα καταλογογράφησης και η οποία εμπεριέχεται σε αυτόματη Βιβλιογραφία (αρχή 1975) που χρησιμοποιείται για την δημιουργία ειδικών λεξικών για το σχέδιο (π.χ. ένα Γερμανικό).

2. Μέλη - Διάρκεια

- Jouve συστήματα πληροφόρησης, Παρίσι. Έργο συντονιστής. Jouve ήδη έχει κάνει αναγνώριση αναδρομικής Βιβλιογραφίας σε ICR.

- Η Βασιλική Βιβλιοθήκη στις Βρυξέλλες ενδιαφέρθηκε για την εφαρμογή Νέων Τεχνολογιών στην αναδρομική καταλογράφηση.

- η σελιδοποίηση γίνεται σε στήλες, με σημείο αναφοράς την πρώτη σελίδα,

- οι αναγραφές χωρίζονται σε δύο στήλες.

- Δίνεται έμφαση στη λογική δομή, σταθερά και έντονα γράμματα.

- Το κείμενο είναι μόνο στο Λατινικό αλφάβητο. Τα στοιχεία της εκτύπωσης δεν είναι μεγάλα και κάποια είχαν προστεθεί με το χέρι.

3. Μετατροπές γενικοί κανόνες

Η μετατροπή που γίνεται είναι στην πλήρη μορφή UNIMARC για την ολοκλήρωση της τρέχουσας καταλογογράφησης του VUBIS, συστήματος της Βασιλικής Βιβλιοθήκης, που είναι προσαρμοσμένη σύμφωνα με τις εθνικές ανάγκες.

Η αναδρομική μετατροπή γίνεται χωρίς να δημιουργείται σύνδεση με τα Εθνικά Αρχεία που βρίσκονται στο τρέχον σύστημα.

Συνήθως κατά την μετατροπή αυτή η σύνδεση - ενσωμάτωση μπορεί να δημιουργηθεί αναδρομικά.

Τα αρχεία επεξεργάζονται με τη χρήση ειδικών λέξεων. Στις Βιβλιογραφικές περιγραφές περιλαμβάνονται όλες οι πληροφορίες (εκτός από τις ταχυδρομικές διευθύνσεις των εκδοτών).

Εάν κάποια πληροφορία δεν εμφανίζεται στους κανόνες καταγραφής δημιουργείται αντίστοιχο πεδίο με το σχετικό πεδίο (προστίθεται αντίστοιχος θεματικός αριθμός με την αναδρομική ημερομηνία). Ονόματα συγγραφέων μπορούν να εμφανίζονται σε δύο διαφορετικές θέσεις στην Βιβλιογραφική περιγραφή, ανεξάρτητα εάν οι κανόνες περιγραφικής καταλογογράφησης είναι του 1973.

- Επικεφαλίδα φόρμας: συγγραφείς κύριας υπευθυνότητας εμφανίζεται σε πεδίο (η επικεφαλίδα αναφέρεται στο αντίστοιχο πεδίο Unimarc). Για τον συγγραφέα δευτερεύουσας υπευθυνότητας υπάρχει αντίστοιχο πεδίο.

Η διαδοχική εισαγωγή επαληθεύεται στο ευρετήριο Συγγραφέα - Τίτλου, όπου το όνομα βρίσκεται στην αρχική μορφή.

Τα θεματικά ευρετήρια μετατρέπονται σε θεματικές αναγραφές και προστίθενται στην κατάλληλη Βιβλιογραφική περιγραφή.

4. Γενικά συμπεράσματα του προγράμματος.

Η δομή αναγνώρισης του λογισμικού γίνεται δυνατή ώστε να αναγνωρίζει το 67% των βιβλιογραφικών περιγραφών. Για το 27,6% των περιγραφών, UNIMARC κωδικοί με προσθήκες λέξεων κλειδιών από την ομάδα Jouve, 5.4% από τις περιγραφές που είναι ιδιαιρετότητες (προβλήματα) της βιβλιοθήκης.

II. Ειδικά λεξικά και εργαλεία για τη δομή αναγνώρισης.

Η αυτόματη αναγνώριση μιας βιβλιογραφικής περιγραφής είναι μια μορφή αναζήτησης, αρκεί ο Η/Υ να κατέχει τα πεδία και υποπεδία της βιβλιογραφικής περιγραφής για να κάνει τις κατάλληλες μετατροπές. Οι περιγραφές ISBD έχουν το πλεονέκτημα της συγκεκριμένης στίξης αρχίζοντας από το πρώτο πεδίο αναγνώρισης, κάτι που δεν ισχύει για παλαιές αναγραφές. Για τις παλαιές αναγραφές δημιουργείται η ανάγκη χρήσης άλλων σημείων αναγνώρισης.

Η συμμετρική πρόβλεψη του τύπου των χαρακτήρων χρησιμοποιείται μόνο σε ορισμένες περιπτώσεις. Στην περίπτωση της Βελγικής Βιβλιογραφίας αν και ήταν προφανές ότι ο προσανατολισμός δεν ήταν χαρακτηριστικός για όλες τις περιπτώσεις, αυτό όμως δεν ήταν πάντα σαφές. Αυτό μπορεί να γίνει με απλές περιγραφές στίξης και έντονα τοπογραφικά στοιχεία.

Στην κύρια εγγραφή είναι το όνομα μόνο ενός συγγραφέα. Είναι τυπωμένο με έντονα γράμματα, ώστε να μπορεί να αναγνωριστεί πιο εύκολα και να μετατραπεί στο αρχείο του συγγραφέα, σε συνέχεια στο αρχείο του τίτλου όπως υπάρχει στην φόρμα της εγγραφής (πεδίο 700). Το πεδίο της Βιβλιογραφικής διεύθυνσης μπορεί να αναζητηθεί εφόσον βρίσκεται σε ιταλικούς χαρακτήρες εκτός από την ημερομηνία. Μ' αυτόν τον τρόπο η προβαλλόμενη διατύπωση μπορεί εύκολα να αναγνωριστεί όπως γίνεται στη Βασιλική Βιβλιοθήκη που χρησιμοποιεί δικά της στοιχεία στο περιθώριο.

- Συχνά το ίδιο κείμενο αποτελεί χαρακτηριστικό πεδίο, μπορεί δε σε μερικές περιπτώσεις να ποικίλει σε λέξεις.

- Η λίστα εργασίας των λέξεων όπως παρουσιάζεται εισάγεται στο σύστημα και αποτελεί ένα λεξικό. Η σύγκριση των λεξικών και των κειμένων των βιβλιογραφικών εγγραφών γίνεται μέσα από «Scanning» OCR είναι ένα καλό εργαλείο αναγνώρισης για αποκωδικοποίηση των πεδίων και υποπεδίων.

Η Βελγική Βιβλιογραφία για το MORE το 1973 προτίμησε έναν έλεγχο στο σύνολο.

Περιγράφοντας τις ιδιαιτερότητες ενός τόμου δημιούργησε λεξικά πολύ συγκεκριμένα. Αυτό ήταν αναγκαίο εφόσον δεν υπήρχαν άλλες περιπτώσεις και ήταν χρήσιμο η ύπαρξη πρόσθετης σημείωσης.

Παρουσίαση στα ελληνικά:
Αννα Αντωνιάδου - Τουργέλη
(Βιβλιοθήκη ΕΚΚΕ).