Specific aspects of retroconversion in view of automatic processing

Catherine LUPOVICI, Jouve Systmes d'Information

Retrospective conversion implies, in any case and whatever is the technical solution, that the library defines the rules for converting the data from the source catalogue into the target one. The rules are written in what is called the Technical specifications for the Retrospective Conversion.

The Technical specifications, starting from the detailed target format, for instance UNIMARC, describe all the source data element and all the rules for conversion, and provide general description of the source catalogue and general rules to be applied.

As there is no standard catalogue specificity, there is no standard specifications available and there is no standard automatic system for catalogue conversion.

Automatic processing implies to tell the automatic system what to do and re enforce the need to have enough knowledge of the catalogue specificity in order to set up the system with the appropriate parameters.

Technical specifications are not specific to automated processing but they are more crucial and need special care and detailed information for what is important in automated processing for :

Character recognition

Structure recognition

Coded data creation

Technical specifications for OCR/ICR

Scanning : OCR/ICR processing is recognising a character image and converting the image into the character code. OCR software are working on images and the first thing is the catalogue scanning. Technical specifications will have to provide details on the physical medium and on the sheets size, paper characteristics, colour, binding ...

Segmentation : once the physical units (sheets or cards) scanned, it is important to describe the layout in order to be able to segment the document image into homogeneous regions to be processed as a whole. It can be done at the page level for a printed catalogue, but also at the record level. For instance the classification code region, the body record region, the shelf mark region ... The spatial organisation of the different regions will be described.

Character fonts and styles description with the possible effects on characters connection. Description of the general rules for using the typography in relationship with logical structure and the need to identify the style when processing the OCR.

Scripts : as OCR software are originally written for specific scripts, it is important to have statistical information on the different scripts represented in the catalogue, as well as their possible combination in a single record (occasional other script, complete fields or sub fields). In a single script it is important to know the extension of the basic characters which is needed to handle for the conversion, for instance the diacritical marks classified by languages or by script transliteration, the usage of superscript and subscript.

Punctuation used with description of the different usage specifically the relationship with structure. If applicable, the hyphenation rules at the end of lines will have to be described as well as the rules for recognition and pasting of the cut words.

Cataloguing vocabulary and rules of abbreviation, if possible with the corresponding cataloguing rules.

Language : statistical information on the languages used in the catalogue are very important to select the dictionaries that can improve the character recognition.

Accuracy requested in the conversion process is very important to know before selecting a methodology, and can help to decide if the catalogue can be processed or not by OCR/ICR

Statistical analysis can be made by random sampling and some practical methodologies are well described in the American librarianship literature of the 70s1.

Technical specifications for automatic structure recognition

Automatic structure recognition implies to have a model of the catalogue structure that can be exploited by the system. The model needs to be generic enough to cover all the regular records in the catalogue in term

of the global full structure. At the same time it has to be very detailed in term of data element description to provide all the indices that can be exploited by the system to recognise unambiguously the content portions. All the needed information is taken in the Technical specifications.

Data elements : Structure recognition of a bibliographic record is finding the fields and sub fields data content inside each homogeneous region already segmented. This micro structure is explored by searching information like typographic styles, existence of particular words or group of words which pertains to certain lexicons, their limits (type of initials and finals such as capital letters, particular words, punctuation ...). The sequential order of data elements as well as optional element is also very important to know, to verify that candidate data elements are coherent with their left and right context. The description includes logical and physical details. What is difficult for the librarians in charge of the technical specifications is to provide only the details that are generally applicable to all the records and not to keep details that are not covering enough records. Data elements description must be written from the analysis of a significant random sample of the catalogue as a synthesis of the rules and not accumulation of different rules with some of them possibly exceptional.

Global structure. The global architecture of the data elements will also be described in order to provide a global hierarchical tree description of all the possible sequences of data elements with their mandatory/optional qualifiers.

Accuracy needed must be defined, on the structure basis for the fields and sub fields which are very important for the service to be provided with the converted catalogue, with a clear definition of what is a structure error.

Technical specifications for automatic coded data creation

Automatic coded data creation is analysing the content of specific data elements with dictionaries tools to find out the language or the country to be translated in an international code. For some languages additional technologies can be used, for instance digram in Dutch.

In order to choose the tools to be integrated in the automated system, it is very important to have statistical information on :

The languages which have to be find out in the same catalogue. The language code can only be found from the record itself if the title is significant, and it is important to say when this creation is authorised or not (minimum number of words in the title, proper names, geographical names which can be checked with authority files or with another part of the record for instance classification number ...)

The places of publication and the corresponding languages in which they are written

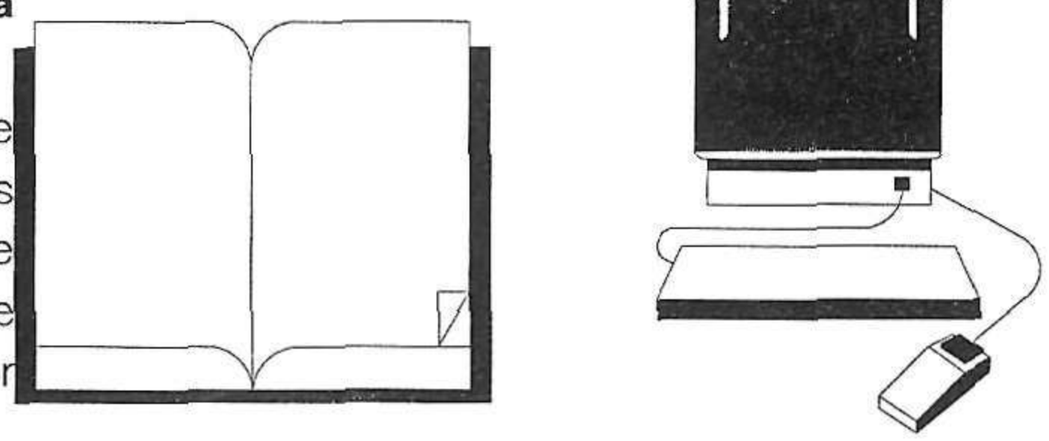
Accuracy requested has also to be defined for each type of code, in order to decide if it can be processed automatically or not.

Conclusion

It is important to remember that there is no obvious information for an "intelligent" automatic system.

In each part of the automatic processing dictionaries are very important. The generic ones can be found in general tools. But some of them are bibliographic specific and can be time consuming for the library to build from scratch. It is important to remember that some tools are already available in machine readable form in the library community, such as the authority files. Otherscan be build from cataloguing rules or standards like ISO 832 providing rules for abbreviation in bibliographic information.

1 The conversion of manual catalogs to collection data bases. In : Library technology reports, March-April 1978, p. 109-206



MORE (Οπτική αναγνώριση Marc) τεχνικές πλευρές ή Τεχνικές πλευρές του MORE

Catherine LUPOVICI, Jouve Systmes d'Information

Το σύστημα MORE είναι μια αυτοματοποίηση των επιμέρους λειτουργικών συστημάτων που μπορεί να εγκατασταθεί σ' ένα παραδοσιακό σύστημα αναγνώρισης ώστε να παρέχει:

. ψηφιακή απόδοση του καταλόγου

. αναγνώριση χαρακτήρων

. αναγνώριση δομής

. κωδικοποίηση πληροφορίας

. διαχείριση μετατροπής

Είχε αποφασιστεί, το πρόγραμμα να αναπτύξει αυτά τα αυτοματοποιημένα συστήματα και να τα ενσωματώσει στο πρωτότυπο σύμφωνα με μία γενική αρχιτεκτονική που επιτρέπει α) τον ευέλικτο συνδυασμό τους ώστε να υπάρχει δυνατότητα χειροκίνητης επεξεργασίας ενός από τα συστήματα και β) τη δυνατότητα περιπτωσιακής παράκαμψης των αυτοματοποιημένων συστημάτων με χειροκίνητο τρόπο προκειμένου να γίνει επεξεργασία αντικανονικών εγγραφών και εξαιρέσεων.

OCR/ICR βιβλιογραφικών δεδομένων

Το λογισμικό OCR/ICR δημιουργήθηκε για το περιβάλλον γραφείου για την επεξεργασία απλής πληροφόρησης υπό μορφή κειμένου. Γενικά, έχουν ενσωματωμένα λεξικά για να βοηθούν στην αναγνώριση των χαρακτήρων με τη διαδικασία του ελέγχου του συλλαβισμού. Η βιβλιογραφική όμως πληροφορία είναι πολύ πιο περίπλοκη από την πληροφορία απλού κειμένου. Αποτελείται από:

. πληροφορία σε μορφή κειμένου

. αριθμητική πληροφορία

. Σημεία στίξης

. Στοιχεία μορφής (πλάγιοι, έντονοι χαρακτήρες) που τονίζουν τη λογική δομή.

Το MORE OCR/ICR αυτόματο σύστημα.

Αυτό το πρόγραμμα είναι μια εφαρμογή της τεχνογνωσίας του OCR/ICR της Jouves στη βιβλιογραφία που χρησιμοποιείται ήδη στην παραγωγή, για παράδειγμα, στις πατέντες: ένα εκατομμύριο σελίδες το χρόνο επεξεργάζεται το Ευρωπαϊκό Γραφείο για τις πατέντες.

Το σύστημα αναγνώρισης δομής MORE

Αυτό το λογισμικό βασίζεται στην τεχνογνωσία του CRIN ως προς την αναγνώριση δομής και σε μία προκαταρκτική μελέτη που έγινε το 1991 - 92 σε συνεργασία με την JOUVE στο πλαίσιο ενός διδακτορικού πάνω στη Δομική Αναγνώριση. Αυτή η προκαταρκτική μελέτη έγινε πάνω στον έντυπο κατάλογο της Εθνικής Βιβλιοθήκης της Γαλλίας (1960 - 1969).

Το σύστημα δημιουργίας κωδικοποιημένων δεδομένων MORE.

Αυτό το σύστημα επεξεργάζεται το αρχείο των αποτελεσμάτων της δομικής αναγνώρισης προκειμένου να δημιουργηθούν κωδικοποιημένα δεδομένα σύμφωνα με το UNIMARC, δηλ. κωδικοποίηση χώρας και κωδικοποίηση γλώσσας.

Και οι δύο κωδικοί διορθώνονται από ένα και μόνο διορθωτή που έχει αναπτυχθεί για το πρόγραμμα με έλεγχο των προτάσεων μιας συνηθισμένης αυτόματης μετάφρασης.

Αποτελέσματα προγράμματος

Ένας από τους σκοπούς του προγράμματος ήταν να δοκιμαστεί το ενδιαφέρον για την εισαγωγή του OCR/ICR και την τεχνολογία της δομικής αναγνώρισης στην Αναδρομική Μετατροπή (των καταλόγων) προκειμένου να μειωθεί το κόστος χωρίς να αλλάξει το ποιοτικό επίπεδο.

Η εκτίμηση έγινε με τη δοκιμή του συστήματος σε πραγματικό περιβάλλον παραγωγής. Ένας χρόνος βιβλιογραφίας μετετράπη με τη χρήση του πρωτοτύπου που αναπτύχθηκε κατά τη διάρκεια του προγράμματος και προσαρμόστηκε σ' αυτή τη βιβλιογραφία και για την επεξεργασία κατά δέσμες και για τη διόρθωση.

Η διόρθωση έγινε από το προσωπικό της Jouve που είχε αναλάβει την αναδρομική μετατροπή, προσωπικό πλήρους απασχόλησης για την κλασσική αναδρομική μετατροπή και η δοκιμή καθώς και το αποτέλεσμα κρίθηκαν και από άποψη βιβλιογραφική και από άποψη παραγωγής.

Επίπεδο βιβλιογραφικής ποιότητας

Στόχος ήταν να φτάσουμε στην ποιότητα που σήμερα επιτυγχάνεται με την κλασσική διπλή πληκτρολόγηση, κωδικοποίηση και τον έλεγχο ποιότητας. Αυτή η ποιότητα κρίνεται από ένα τυχαίο δείγμα σύμφωνα με τη μεθοδολογία που ορίζει το πρότυπο ISO 2859. Η εφαρμογή αυτού του προτύπου στην αναδρομική μετατροπή οδηγεί στον έλεγχο α) των λανθασμένων χαρακτήρων, β) των

λαθών δομής και κωδικοποίησης σε σχέση με τη σημασία των λαθών που αφορά στη μελλοντική έρευνα μέσω ηλεκτρονικού καταλόγου και στη χρήση ανάκτησης ή επεξεργασίας.

Για τον έλεγχο της παραγωγής η ποιότητα κρίθηκε με την ίδια μεθοδολογία και τους κανόνες και η τελική ποιότητα (μετά την αυτόματη επεξεργασία, τη διόρθωση και τον αυτόματο έλεγχο κατά δέσμες):

99,974% ακρίβεια χαρακτήρων

100% ακρίβεια πεδίων και 99,28% ακρίβεια υποπεδίων

98,4% ακρίβεια γλώσσας και κωδικών χώρας.

Το κύριο πρόβλημα ήταν οι τίτλοι των υποπεδίων στο UNIMARC με πολλές παράλληλες αναφορές που σχετίζονται με την επίσημη πολυγλωσσία του Βελγίου. Το Ολλανδικό λεξικό που χρησιμοποιήθηκε στο πρόγραμμα δεν ήταν αρκετά ικανοποιητικό για το OCR και τον κωδικό γλώσσας.

Συμπέρασμα: το ποιοτικό επίπεδο που επιτεύχθηκε μετά την επέκταση του συστήματος είναι ισοδύναμο με την κλασσική αναδρομική μετατροπή αν βρεθούν ή φτιαχτούν για τον υπό μετατροπή κατάλογο τα κατάλληλα εργαλεία (λογισμικό OCR, γενικά και ειδικά λεξικά, πρωτυποποίηση)

Παραγωγικότητα

Αρχή της αυτόματης επεξεργασίας σ' ένα βιομηχανικό περιβάλλον παραγωγής είναι να γίνεται αυτόματα η επεξεργασία των κανονικών περιπτώσεων και να παρεμβαίνει ο ανθρώπινος παράγοντας μόνο στις εξαιρέσεις ώστε να μειώνεται ο χρόνος και το κόστος. Ταυτόχρονα αυξάνεται το κόστος του υπολογιστή και η μείωση του ανθρώπινου χρόνου πρέπει να εξισορροπεί αυτή την αύξηση.

Η εκτίμηση της σύγκρισης του κόστους μεταξύ αυτής της μεθοδολογίας και της κλασσικής αναδρομικής μετατροπής των καταλόγων έγινε μετά από σύγκριση του ελέγχου παραγωγής και αξιολόγηση του χρόνου επεξεργασίας αυτής της βιβλιογραφίας, όπως γίνεται όταν σχεδιάζεται μια εμπορική προσφορά για υπηρεσίες αναδρομικής μετατροπής με βάση τις προδιαγραφές της ίδιας της βιβλιοθήκης.

Ειδικές πλευρές της αναδρομικής μετατροπής ενόψη της αυτόματης επεξεργασίας C. Lupovici

Η αναδρομική μετατροπή συνεπάγεται οπωσδήποτε και με οποιαδήποτε τεχνική λύση, ότι η βιβλιοθήκη ορίζει τους κανόνες για τη μετατροπή των δεδομένων από τον κατάλογο-πηγή στον κατάλογο-στόχο. Οι κανόνες είναι γραμμένοι σ' αυτό που λέγεται «Τεχνικές Προδιαγραφές για την αναδρομική καταλογογράφηση».

Οι Τεχνικές Προδιαγραφές, αρχίζοντας από τη λεπτομερή καρτέλα εγγραφής - στόχο, για παράδειγμα το UNIMARC, περιγράφουν όλα τα στοιχεία των δεδομένων της πηγής και όλους τους κανόνες για μετατροπή και παρέχουν γενική περιγραφή του καταλόγου-πηγής και γενικούς κανόνες που πρέπει να εφαρμοτούν.

Τεχνικές Προδιαγραφές για το OCR/ICR.

Σάρωση: Η επεξεργασία OCR/ICR αναγνωρίζει την εικόνα ενός χαρακτήρα και μετατρέπει την εικόνα σε κωδικό χαρακτήρα.

Διαχωρισμός: Αφού σαρωθούν οι φυσικές ενότητες (φύλλα ή δελτία) είναι σημαντικό να γίνει περιγραφή της διάταξης προκειμένου να διαχωριστεί η εικόνα του εντύπου σε ομοιογενείς περιοχές για να γίνει συνολική επεξεργασία

Περιγραφή τυπογραφικών στοιχείων και μορφών με πιθανά αποτελέσματα στον συνδυασμό των χαρακτήρων.

Γραφές: Επειδή το λογισμικό OCR αρχικά έγινε για συγκεκριμένες γραφές, είναι σημαντικό να υπάρχει στατιστική πληροφόρηση πάνω στις διαφορετικές γραφές που αντιπροσωπεύονται στον κατάλογο καθώς και ο πιθανός συνδυασμός τους σε μία μόνο εγγραφή.

Στίξη που χρησιμοποιείται με την περιγραφή της διαφορετικής χρήσης ειδικά σε σχέση με τη δομή.

Λεξιλόγιο καταλογογράφησης και κανόνες συντομογραφίας, εαν είναι δυνατόν, με τους αντίστοιχους κανόνες καταλογογράφησης.

Γλώσσα: Στατιστική πληροφορία σχετικά με τις γλώσσες που χρησιμοποιούνται στον κατάλογο είναι πολύ σημαντική για την επιλογή των λεξικών που μπορούν να βελτιώσουν την αναγνώριση χαρακτήρων.

Η Ακρίβεια που απαιτείται στη διαδικασία μετατροπής είναι σημαντικό να γνωρίζεται πριν από την επιλογή μεθοδολογίας και μπορεί να βοηθήσει στην απόφαση αν πρέπει ή δεν πρέπει να γίνει η επεξεργασία του καταλόγου με το OCR/ICR.

Τεχνικές προδιαγραφές για αυτόματη αναγνώριση δομής.

Αυτόματη αναγνώριση σημαίνει να υπάρχει ένα μοντέλο της δομής του καταλόγου που να μπορεί να εξερευ-

νηθεί από το σύστημα. Το μοντέλο πρέπει να είναι αρκετά γενικό ώστε να καλύπτει όλες τις κανονικές εγγραφές στον κατάλογο.

Παράλληλα πρέπει να είναι πολύ λεπτομερειακό ως προς την περιγραφή των στοιχείων των δεδομένων ώστε να παρέχει όλα τα ευρετήρια που μπορούν να διερευνηθούν από το σύστημα για να *αναγνωρίζονται χωρίς αμφιβολία οι ποσότητες του περ*ιεχομένου. Όλη η απαιτούμενη πληροφόρηση διαλαμβάνεται στις Τεχνικές Προδιαγραφές.

Στοιχεία δεδομένων: Αναγνώριση δομής μιας βιβλιογραφικής εγγραφής είναι η εύρεση του περιεχομένου των πεδίων και των υποπεδίων των δεδομένων μέσα σε κάθε ομοιογενή περιοχή που έχει ήδη διαχωριστεί.

Σφαιρική δομή. Περιγράφεται προκειμένου να δοθεί σφαιρική ιεραρχική περιγραφή όλων των πιθανών συνεπειών των στοιχείων των δεδομένων με τους υποχρεωτικούς/προαιρετικούς προσδιορισμούς.

Ακρίβεια πρέπει να οριστεί βάσει της δομής με σχέση ορισμού του τι είναι λάθος δομής.

Τεχνικές προδιαγραφές για δημιουργία αυτόματων κωδικοποιημένων δεδομένων.

Η δημιουργία αυτόματων κωδικοποιημένων δεδομένων είναι η ανάλυση του περιεχομένου ειδικών στοιχείων των δεδομένων με λεξικά-εργαλεία για να βρεθεί η γλώσσα ή η χώρα που θα μεταφραστεί σε διεθνή κώδικα.

Για την επιλογή των εργαλείων αυτών που θα ενσωματωθούν στο αυτόματο σύστημα είναι πολύ σημαντικό να υπάρχει στατιστική πληροφορία σχετικά με:

- τις γλώσσες που πρέπει να βρεθούν στον ίδιο κατάλογο

- τους τόπους έκδοσης και τις αντίστοιχες γλώσσες.

Ακρίβεια: Την απαιτούμενη ακρίβεια για κάθε τύπο κωδικού προκειμένου αν αποφασιστεί να είναι δυνατόν να γίνει αυτόματη επεξεργασία.

ΣΥΜΠΕΡΑΣΜΑ

Πρέπει να θυμόμαστε ότι δεν υπάρχει προφανής πληροφορία για ένα «έξυπνο» αυτόματο σύστημα. Τα λεξικά έχουν μεγάλη σημασία για κάθε τμήμα της αυτόματης επεξεργασίας. Τα γενικά μπορούν να βρεθούν σε γενικά εργαλεία. Μερικά όμως είναι εξειδικευμένα βιβλιογραφικά και μπορεί να είναι χρονοβόρο για τη βιβλιοθήκη να ξεκινήσει από την αρχή. Μερικά εργαλεία είναι ήδη διαθέσιμα σε μηχαναγνώσιμη μορφή στους βιβλιοθηκάριους, όπως οι καθιερωμένες εγγραφές. Άλλα μπορούν να σχεδιαστούν από κανόνες καταλογογράφησης και πρότυπα (π.χ. ISO 150.832) που παρέχουν κανόνες συντομογραφίας για βιβλιογραφικές πληροφορίες.

Απόδοση - παρουσίαση: **Μαρία Βακαλοπούλου.**