FACIT : Στόχοι και προοπτικές :
Η ελληνική εμπειρία

Ιωάννα Τσούτσου-Δημοπούλου
Εθνική Βιβλιοθήκη της Ελλάδος

Το πρόγραμμα FACIT, ακρωνύμιο της αναλυτικής ονομασίας Fast Automated Conversion with Integrated Tools, είναι ένα κοινοτικό πρόγραμμα στο οποίο συμμετέχουν η National Library Authority (Statens Bibliotekstjeneste) της Δανίας, η State and University Library (Statsbibliotekt) του Aarus της Δανίας, η Biblioteca Nazionale Centrale της Φλωρεντίας, η Biblioteca Nazionale V.E.III της Νάπολης και η Εθνική Βιβλιοθήκη της Ελλάδος. Στόχος του προγράμματος είναι να μετατραπούν σε μηχαναναγνώσιμη μορφή δακτυλογραφημένα ή τυπωμένα δελτία, που υπάρχουν ήδη στους καταλόγους των βιβλιοθηκών.

Τα τελευταία χρόνια, οι ευρωπαϊκές βιβλιοθήκες αυτοματοποιήθηκαν, το όλο σύστημα διαχείρισης και διακίνησης γίνεται μέσω των ηλεκτρονικών υπολογιστών, το καινούργιο υλικό αναζητείται μέσω των online καταλόγων, με αποτέλεσμα, οι χρήστες να μην έχουν άμεση πρόσβαση στο παλαιότερο υλικό. Εκτός από την, κατά κάποιον τρόπο, "απώλεια" υλικού της βιβλιοθήκης, αυξάνεται και το κόστος με τη διατήρηση αυτών των χειροκίνητων διεργασιών στην όλη λειτουργία της βιβλιοθήκης. Είναι, λοιπόν, προφανής η ανάγκη μετατροπής των υπαρχόντων δελτίων σε μηχαναναγνώσιμη μορφή, ώστε να μπορούν να ενταχθούν στο αυτοματοποιημένο σύστημα. Στα πλαίσια αυτής της ανάγκης, το FACIT αποβλέπει στην γρήγορη και σχετικά φθηνή μαζική μετατροπή δακτυλογραφημένων ή εντυπων δελτίων σε μηχαναγνώσιμη μορφή, με τη χρήση OCR, συμβάλλοντας έτσι στην προώθηση της αυτοματοποίησης των ευρωπαϊκών βιβλιοθηκών.

Στόχος του προγράμματος είναι η δημιουργία μιας "εφαρμογής" για την οπτική αναγνώριση του κειμένου των δελτίων, τον αυτόματο εντοπισμό και τη διόρθωση λαθών από την οπτική αναγνώριση και για την αυτόματη μετατροπή των βιβλιογραφικών εγγραφών σε μορφή UNIMARC. Για την "εφαρμογή" αυτή χρησιμοποιείται η ήδη υπάρχουσα τεχνολογία σε scanning και OCR, για τη μετατροπή των δελτίων σε μορφή ASCII.

Τα δελτία προέρχονται από διαφορετικές βιβλιοθήκες και έγιναν σε διαφορετικές χρονικές στιγμές, επομένως δεν έχουν ομοιομορφία.

Η διαδικασία , για την ολοκλήρωση του προγράμματος, είναι η εξής :

1. Ανάλυση των χαρακτηριστικών των δελτιοκαταλόγων της κάθε βιβλιοθήκης. Ας μην ξεχνάμε ότι κάθε βιβλιοθήκη έχει ακολουθήσει διαφορετικούς κανόνες καταλογογράφησης και επομένως οι πληροφορίες που υπάρχουν σε κάθε δελτίο διαφέρουν. Επίσης, είναι σίγουρο ότι η ίδια βιβλιοθήκη ακολουθεί διαφορετικές πρακτικές καταλογογράφησης στην πορεία των ετών. Στο στάδιο αυτό, η κάθε μία από της συμμετέχουσες στο πρόγραμμα βιβλιοθήκες, έκανε μία λεπτομερή ανάλυση του δελτιοκαταλόγου της, παράλληλα με τις άλλες και βασισμένη σε ένα συγκεκριμένο πλαίσιο, ώστε να επιτευχθεί ομοιομορφία. Δηλαδή, περιγράφηκαν τα είδη δελτίων που υπάρχουν, οι κανόνες που χρησιμοποιήθηκαν για καταλογογράφηση, οι ενδεχόμενες αποκλίσεις από τους κανόνες, η χρησιμοποίηση ή μη χειρόγραφων επεμβάσεων κλπ. Η ανάλυση αυτή ήταν απαραίτητη για να δημιουργηθούν πολύ συγκεκριμένες "προδιαγραφές" και μέθοδοι για την αυτόματη διαμόρφωση των δελτίων σε μηχαναναγνώσιμη μορφή.

2. Scanning δείγματος 2.500 δελτίων από κάθε βιβλιοθήκη για να μπορέσει να γίνει **ανάλυση των λαθών** που προέκυψαν από αυτή τη διαδικασία οπτικής αναγνώρισης του κειμένου των δελτίων, ώστε να δημιουργηθούν συγκεκριμένες "προδιαγραφές" για τον αυτόματο εντοπισμό και διόρθωση των λάθών.

3. Μετατροπή των δελτίων, που είναι ήδη σε μορφή ASCII, σε βιβλιογραφική μορφή UNIMARC. Στο σημείο, δηλαδή, αυτό οι πληροφορίες που υπάρχουν σε κάθε δελτίο τοποθετούνται κάτω από τα κατάλληλα πεδία και

υποπεδία, ώστε να είναι δυνατόν να εντα-
χθούν σε οποιονδήποτε αυτοματοποιημένο
κατάλογο βιβλιοθήκης.

Τελικά, η όλη "εφαρμογή" θα ελεγχθεί με ένα
δείγμα 10.000 δελτίων από κάθε βιβλιοθήκη.

4. Ολοκλήρωση της "εφαρμογής" και παραγω-
γή εγχειριδίου για τη χρήση της.

Το πρόγραμμα αποβλέπει στη διαχείριση πολύγλωσσων καταλόγων, μια και αρκετές ξενόγλωσσες συλλογές βρίσκονται σε εθνικές και μεγάλες ακαδημαϊκές βιβλιοθήκες. Επομένως, ήταν απαραίτητη η ανάπτυξη εργαλείων και μεθόδων για την μετατροπή τόσο πολύγλωσσων δελτίων καταλόγων, όσο και δελτίων που δημιουργήθηκαν πριν από την εφαρμογή των κανόνων ISBD. *Τέτοιου είδους εργαλεία και μέθοδοι* είναι απαραίτητοι για τη μετατροπή, σε ευρεία κλίμακα, των καταλόγων εθνικών βιβλιοθηκών ή άλλων σημαντικών συλλογών σε διεθνες επίπεδο.

Το επίπεδο της βιβλιογραφικής περιγραφής εξαρτάται απόλυτα από τις πληροφορίες που υπάρχουν στην πηγή. Βιβλιογραφικές πληροφορίες που δεν υπάρχουν στην πηγή, δεν θα προστίθενται στη διαδικασία της μετατροπής. Αυτό σημαίνει ότι, σε πολλές περιπτώσεις, το επίπεδο καταλογογράφησης θα βρίσκεται κάτω από τα minimun standards που απαιτούνται σήμερα.

Από την επιτυχία του προγράμματος, θα προκύψουν διάφορα οφέλη. Παράλληλα με τη δημιουργία του λογισμικού της "εφαρμογής", θα έχουν αναπτυχθεί και *διάφορες μέθοδοι ανάλυσης δελτίων, μέθοδοι* ανάλυσης λαθών κατά την μετατροπή, μέθοδοι αξιολόγησης ποιότητας, ταχύτητας και κόστους κατά τη μετατροπή με χρήση OCR, κλπ.

Στόχοι του προγράμματος είναι 1) να διατηρηθούν, κατά τη μετατροπή, οι διαφοροποιήσεις που υπάρχουν στους χαρακτήρες στην πηγή, και 2) οι χαρακτήρες να είναι εύκολα αναγνωρίσιμοι από τον άνθρωπο που θα κάνει τον έλεγχο για τον εντοπισμό των λαθών, είτε στην οθόνη είτε σε χαρτί. Για το σύνολο των χαρακτήρων χρησιμοποιείται το ISO standard UNICODE 10646 για τους λατινικούς χαρακτήρες. Οσον αφορά στους ελληνικούς χαρακτήρες, χρησιμοποιείται το σύνολο των χαρακτήρων που χρησιμοποιεί η ΕΒΕ για τους ελληνικούς χαρακτήρες, το οποίο καλύπτει το πολυτονικό σύστημα.

Και τώρα ας έρθουμε στην ελληνική εμπειρία από το πρόγραμμα FACIT. Θα ήθελα να περιγράψω, εν συντομία, την εικόνα των καταλόγων της ΕΒΕ, ώστε να έχουμε όλοι μια εικόνα για την κατάσταση που περιγράφεται. Η Εθνική Βιβλιοθήκη της Ελλάδος έχει δύο δημόσιους δελτιοκαταλόγους, τον παληό και τον καινούργιο. Στον παληό κατάλογο είναι ταξιθετημένα τα δελτία βιβλίων που εισήχθησαν στη βιβλιοθήκη μέχρι το 1977. Τα δελτία αυτά, διαστάσεων 16,5 x 10 εκ. είναι χειρόγραφα και επομένως δεν εμπίπτουν στους στόχους του προγράμματος FACIT. Στον καινούργιο κατάλογο βρίσκονται τα δελτία που έχουν εισαχθεί από το 1978 και εξής. Ο κατάλογος αυτός είναι χωρισμένος σε ελληνικό και λατινικό τμήμα. Στο ελληνικό τμήμα είναι ταξιθετημένα τα δελτία των οποίων η κύρια αναγραφή είναι στα ελληνικά, στον ξένο κατάλογο είναι ταξιθετημένα τα δελτία των οποίων η κύρια αναγραφή είναι στο λατινικό αλφάβητο. Για παράδειγμα, ένας έλληνας συγγραφέας που γράφει αγγλικά, θα ταξιθετηθεί στο ελληνικό τμήμα του καταλόγου, εφόσον η κύρια αναγραφή είναι στα ελληνικά, ενώ το υπόλοιπο δελτίο είναι γραμμένο αγγλικά. Το δελτίο "τίτλου" του συγκεκριμένου έργου θα ταξιθετηθεί στο λατινικό τμήμα του καταλόγου. Και τα δύο τμήματα του καταλόγου είναι χωρισμένα κατά συγγραφέα και σειρά, κατά τίτλο και κατά θέμα. Ο θεματικός κατάλογος είναι στο ελληνικό τμήμα, ενώ, στο λατινικό τμήμα, υπάρχει ένας πολύ μικρός θεματικός κατάλογος, που αποτελείται από ξένα ονόματα, που χρησιμοποιούνται σαν θεματικές επικεφαλίδες. Και στα δύο τμήματα του καταλόγου, το κείμενο των δελτίων είναι συνδυασμός και των δύο αλφαβήτων.

Τα δελτία της ΕΒΕ παρουσιάζουν ένα μεγάλο πλεονέκτημα έναντι των δελτίων των άλλων βιβλιοθηκών και αυτό είναι η ομοιομορφία. Πρόκειται για, περίπου, 400.000 δελτία, διεθνούς σχήματος (7,5 x 12,5 εκ). Η δημιουργία του καταλόγου άρχισε το 1978, επόμενως όλα τα δελτία έχουν μορφή ISBD: Η καταλογογράφηση *έχει γίνει σύμφωνα με τους Anglo American* Cataloging Rules. Τα δελτία μπορούν να χωριστούν σε δύο μεγάλους κατηγορίες, δηλαδή σε δελτία που έγιναν με γραφομηχανή και σε δελτία που έγιναν με τη χρήση Η/Υ. Τα δελτία που έγιναν με γραφομηχανή καλύπτουν τη χρονική περίοδο 1978-1987. Από το

1988 άρχισε να χρησιμοποιείται Η/Υ για την παραγωγή των δελτίων.

Το 60%, περίπου, των δελτίων είναι γραμμένα στη γραφομηχανή απ'ευθείας ή έχει χρησιμοποιηθεί μεμβράνη. Εχει χρησιμοποιηθεί γραφομηχανή IBM, με γραμματοσειρές courrier 10 και scribe 12. Και οι δύο τύποι γραμματοσειρών χρησιμοποιούνται στο ίδιο δελτίο. Συγκεκριμένα, η γραμματοσειρά courrier 10 χρησιμοποιήθηκε για την κύρια αναγραφή, την περιοχή τίτλου και κύριας υπευθυνότητας, την περιοχή έκδοσης, την περιοχή δημοσίευσης, κλπ., την περιοχή φυσικής περιγραφής, την περιοχή σειράς καθώς επίσης και για όλες τις πρόσθετες αναγραφές. Η γραμματοσειρά scribe 12 χρησιμοποιήθηκε για την περιοχή σημειώσεων, το ίχνευμα, τον ταξινομικό αριθμό, τον αριθμό εισαγωγής και τα αντίτυπα της βιβλιοθήκης. Για το υπόλοιπα 40% των δελτίων έχει χρησιμοποιηθεί Η/Υ. Γενικά, τα δελτία είναι σε καλή κατάσταση, καθαρά, χωρίς χειρόγραφες προσθήκες, κλπ.

Η εικόνα αυτή των δελτίων σημαίνει ότι η μεθοδολογική προσέγγιση τόσο στην ανίχνευση/διόρθωση των λαθών, όσο και στη μετατροπή τους σε μορφή UNIMARC μπορεί να βασιστεί στη στίξη ISBD, σε αντίθεση με ότι συμβαίνει με τις άλλες βιβλιοθήκες.

Τα προβλήματα παρουσιάστηκαν στη φάση εκείνη του προγράμματος που έπρεπε τα δελτία να αναγνωρισθούν από τον OCR για να μετατραπούν σε μορφή ASCII. Συγκεκριμένα, τον Ιούνιο 1993, στην Κοπεγχάνη έγινε το scanning 2.561 αντιπροσωπευτικών δελτίων της ΕΒΕ, που θα χρησιμοποιούντο σαν δείγμα, για να προχωρήσουμε στην ανάλυση των λαθών. Χρησιμοποιήθηκε ο scanner Fujitsu 3096E. Τα δελτία, σε μορφή εικόνας, κάλυψαν 31.8 MB (περίπου 12.4 KB ανά δελτίο). *Για την αναγνώριση χρησιμοποιήθηκε το λο*γισμικό OCR νέο GigaRead για Windows. Το συγκεκριμένο λογισμικό είχε μεγάλο πρόβλημα με την αναγνώριση των ελληνικών χαρακτήρων. Παρόλο που επί πέντε μέρες προσπαθήσαμε να "εκπαιδεύσουμε" τον OCR στην ανάγνωση των ελληνικών χαρακτήρων, το αποτέλεσμα ήταν απογοητευτικό. Στη συνέχεια, δοκιμάστηκε και η παλαιά DOS version του GigaRead, με τα ίδια όμως απογοητευτικά αποτελέσματα.

Εδώ, στην Ελλάδα, χρησιμοποιήσαμε για την οπτική αναγνώριση των χαρακτήρων το λογισμικό Recognita Plus 2.0 (Windows version) της ουγγρικής εταιρείας SLZI, το οποίο θεωρείται ότι έχει μεγαλύτερες δυνατότητες *αναγνώρισης των ελληνικών χαρακτήρων*. Και εδώ, τα αποτελέσματα δεν ήταν ικανοποιητικά.

Τέλος, εξετάστηκε η δυνατότητα χρησιμοποίησης του *αμερικάνικου πακέτου OCR Perceive της εταιρίας* Ocron. Στο πακέτο αυτό, οι θέσεις 128-255 μπορούν να χρησιμοποιηθούν από χαρακτήρες που δίνονται από τον χρήστη. Δυστυχώς, η επαφή με την εταιρεία δεν έφερε κανένα αποτέλεσμα.

Προ μηνός, δοκιμάσαμε ένα ελληνικό OCR, τον "Αναγνώστη" της εταιρίας Ideatech. Το αποτέλεσμα είναι πολύ καλύτερο από ό,τι προηγούμενο δοκιμάσαμε, αλλά οι πολυτονικοί χαρακτήρες παρουσιάζονται από τα μονοτονικά ισοδύναμά τους.

Η εταιρεία αυτή μας διαβεβαίωσε ότι έχει τη δυνατότητα ανάπτυξης ειδικού λογισμικού OCR που να καλύπτει όλους τους ελληνικούς χαρακτήρες, μονοτονικούς και πολυτονικούς, με ελάχιστο ποσοστό λάθους.

Από τα παραπάνω, φαίνεται καθαρά ποιός είναι ο λόγος που η Εθνική Βιβλιοθήκη της Ελλάδος δεν προχώρησε στο στάδιο της ανάλυσης των λαθών. Τα μέχρι τώρα αποτελέσματα από την οπτική αναγνώριση είναι γεμάτα λάθη, επομένως δεν υπάρχει κανένας λόγος ανάλυσής τους. Η Εθνική Βιβλιοθήκη της Ελλάδος έχει ήδη αρχίσει τη διαδικασία για την ανάπτυξη και προμήθεια ειδικού λογισμικού οπτικής αναγνώρισης χαρακτήρων, *που θα καλύπτει τους ελληνικούς* χαρακτήρες. Μόλις γίνει δυνατή η οπτική αναγνώριση του κειμένου των δελτίων, με αποδεκτό ποσοστό λαθών, αμέσως θα προχωρήσουμε στη διαδικασία ανάλυσης των λαθών, με τελικό στόχο τη μαζική οπτική αναγνώριση του κειμένου όλων των δελτίων του καινούργιου καταλόγου και μετατροπή τους σε μηχαναγνώσιμη μορφή. Με δεδομένη της ποιότητα των δελτίων της ΕΒΕ, θεωρούμε ότι το τελικό αποτέλεσμα θα είναι πάρα πολύ θετικό.