ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΑΡΧΕΙΟΝΟΜΙΑΣ ΚΑΙ ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ
Πρόγραμμα Μεταπτυχιακών Σπουδών στην Επιστήμη της πληροφορίας «Διοίκηση και οργάνωση Βιβλιοθηκών με έμφαση στις Νέες Τεχνολογίες της Πληροφορίας»

Ψηφιοποίηση, Κωδικοποίηση και Οπτική Αναγνώριση Χαρακτήρων σε κείμενα πολλών γραφών

Η περίπτωση του Letopis 'Zhurnal 'nykh Statei

Εργασία της:
Κωνσταντίνας Δήμου
Για το μάθημα:
«Ψηφιακές Βιβλιοθήκες»

Διδάσκων καθηγητής: Σαράντος Καπιδάκης

Αθήνα, 2004

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Εισαγωγή

Ένας από τους θεμελιώδεις στόχους των βιβλιοθηκών σήμερα, είναι να σωθεί το υλικό που διαθέτουν σαν μακροπρόθεσμη μνήμη για το αύριο. Οι μεγάλες βιβλιοθήκες έχουν θαυμάσιες συλλογές, οι οποίες αποτελούν την πρώτη ύλη της ιστορίας και γενικότερα της ανθρωπότητας. Αυτές οι συλλογές αποτελούνται κυρίως από έντυπο υλικό, και άλλα συμβατικά τεκμήρια. Είναι λοιπόν ευθύνη του προσωπικού που εργάζεται για αυτούς τους Οργανισμούς, να το συντηρήσει, να το προστατεύσει, να το διαχειριστεί και να το αξιοποιήσει με τον καλύτερο δυνατό τρόπο, καθώς και να εξασφαλίσει τη συνεχή μακροπρόθεσμη πρόσβαση σε αυτό.
Η ανάπτυξη των ψηφιακών βιβλιοθηκών έχει δημιουργήσει ένα μεγάλο ενθουσιασμό σχετικά με την ψηφιοποίηση μερικών από αυτών των συλλογών. Τα τελευταία χρόνια έχουν υπάρξει πολυάριθμα προγράμματα ψηφιακών βιβλιοθηκών ευρείας κλίμακας, τα οποία έχουν οργανωθεί από ποικίλους Οργανισμούς σε όλο τον κόσμο για διάφορους λόγους:

- Ένας λόγος, είναι ότι τα συμβατικά τεκμήρια είναι συχνά εύθραυστα στην αρχική φυσική τους κατάσταση. Η ψηφιοποίηση και η πρόσβαση σε αυτά λύνει τη βασική σύγκρουση μεταξύ διατήρησης και πρόσβασης . Είναι γνωστό ότι η αύξηση των δυνατοτήτων πρόσβασης στα συμβατικά τεκμήρια αυξάνει την φθορά τους και περιορίζει έτσι το χρόνο ζωής τους. Από την άλλη, η φροντίδα για τη διατήρησή τους περιορίζει την πρόσβαση σε αυτά. Η ψηφιοποίηση έρχεται λοιπόν να λύσει το παραπάνω πρόβλημα. Συντηρεί το περιεχόμενο του, και το καθιστά διαθέσιμο σε ολόκληρο τον κόσμο.
- Ένα άλλος λόγος είναι ότι το συμβατικό τεκμήριο, μέσω της επεξεργασίας του και της μετατροπής του σε ψηφιακό, μπορεί να γίνει προσιτό και εκμεταλλεύσιμο με ποικίλους τρόπους (απεριόριστος αριθμός χρηστών, απεριόριστη πρόσβαση κλπ).

Εντούτοις, υπάρχουν σημαντικές προτεραιότητες που πρέπει να εξεταστούν πριν από την έναρξη ενός προγράμματος ψηφιοποίησης. [2,3,7,13]

1. Ψηφιοποίηση

Η ψηφιοποίηση σε ένα αρχειακό περιβάλλον, αναφέρεται συνήθως στη λήψη ενός φυσικού αντικειμένου - όπως ένα βιβλίο, μια εικόνα, ένας χάρτης - και στη διαδικασία μετατροπής του σε ηλεκτρονική μορφή. Συνήθως αφορά μια συλλογή που είναι σπάνια, μοναδική, και συχνά εξαιρετικά εύθραυστη. Η ηλεκτρονική μετατροπή ολοκληρώνεται συνήθως μέσω της σάρωσης, μιας διαδικασίας με την οποία ένα έγγραφο ανιχνεύεται από ένα μηχάνημα , και στη συνέχεια αντιπροσωπεύεται στον υπολογιστή υπό μορφή δυαδικών στοιχείων.
Ωστόσο η διαδικασία της ψηφιοποίησης είναι αρκετά περίπλοκη και απαιτεί μια σειρά από ενέργειες:

1.1. Επιλογή του υλικού για ψηφιοποίηση

Πολλοί Οργανισμοί μπορούν να μπουν στην διαδικασία να ψηφιοποιήσουν όλο το υλικό τους. Όμως, επειδή η ψηφιοποίηση είναι μια αρκετά ακριβή διαδικασία, απαιτείται προγραμματισμός και κατάλληλη επιλογή του υλικού.

Η επιλογή περιλαμβάνει καταρχήν καλό σχεδιασμό, χρησιμοποιώντας τα κατάλληλα κριτήρια κρίσης. Οι καλές τεχνικές επιλογής, εξασφαλίζουν το γεγονός ότι οι οικονομικοί πόροι, επενδύονται στην ψηφιοποίηση των σημαντικότερων και πιο χρήσιμων συλλογών, με το χαμηλότερο δυνατό κόστος, και χωρίς έκθεση του ιδρύματος σε νομικό ή κοινωνικό κίνδυνο.
Η επιλογή είναι μια διαδικασία, που αφορά τους υπεύθυνους της διαχειριζόμενης συλλογής, και οι οποίοι θα πρέπει:

- Να αξιολογήσουν το υλικό που διαθέτουν, να υποδείξουν αυτό που πρόκειται να ψηφιοποιηθεί, και να αιτιολογήσουν τους λόγους επιλογής του.
- Να καθορίσουν προτεραιότητες για τη διάσωση ή τη συντήρηση αυτού του υλικού, με βάση την πνευματική του αξία και τον κίνδυνο φθοράς, έτσι ώστε να ψηφιοποιηθεί κατά σειρά με βάση την αξία του.

Κατά τη διάρκεια των παραπάνω σταδίων, το προσωπικό λαμβάνει αποφάσεις που έχουν σημαντικές επιπτώσεις στη ζωή και τη δυνατότητα πρόσβασης στο περιεχόμενο της συλλογής. Συνεπώς κατά την εξέταση του υλικού για την αξιολόγηση του, την προτεραιότητα συντήρησης, και στην συνέχεια την μετατροπή του σε ψηφιακό, το προσωπικό θα πρέπει να εξετάσει και κάποιους άλλους παράγοντες όπως:

- Δυνατότητα επίτευξης συμφωνιών με κοινοπραξίες και άλλες εθνικές πρωτοβουλίες.
- Αξιολόγηση του από τους χρήστες σε σύγκριση με άλλα υλικά που κατέχει ο οργανισμός.
- Εξασφάλιση της διατήρησης της καινούργιας του μορφής.
- Συμβατότητα του υλικού με τα μέσα τεχνολογίας που διαθέτει.
- Περιορισμούς στην πρόσβαση λόγω του νομικού καθεστώτος.
- Διαθεσιμότητα για χρήση.
- Εξασφάλιση των κατάλληλων μεταδεδομένων για τον προσδιορισμό των εγγράφων, και την πλοήγηση μέσα σε αυτά.
- Το κόστος της όλης διαδικασίας

Έτσι προσδιορίζουν και αφαιρούν το προβληματικό υλικό, και επιλέγουν και δίνουν προτεραιότητα στο κατάλληλο για την ψηφιακή εργασία, εξασφαλίζοντας με αυτόν τον τρόπο μια ομαλή ροή της δουλειάς. [11,2,4,7]

1.2. Μετατροπή του υλικού

Η μετατροπή του έντυπου υλικού σε ψηφιακή μορφή, επεξηγεί τη διαφορά μεταξύ μικρής και μεγάλης κλίμακας προσπαθειών. Ποιος είναι ο καλύτερος τρόπος για να μετατραπούν οι τεράστιες συλλογές σε ψηφιακή μορφή; Ποια είναι η σχέση μεταξύ κόστους και ποιότητας; Ποια είναι η πιθανότητα, οι σημερινές προσπάθειες να αποβούν χρήσιμες μακροπρόθεσμα; Σε ένα μικρό πρόγραμμα το οποίο απαιτεί μόνο μερικές χιλιάδες στοιχεία, το υλικό θα περαστεί μέσω ενός ψηφιακού ανιχνευτή, θα ελεγχθούν τα αποτελέσματα για προφανή λάθη, και θα δημιουργηθούν τα κατάλληλα

μεταδεδομένα που απαιτούνται για την περιγραφή τους. Τι γίνεται όμως με τις βιβλιοθήκες που διαθέτουν τεράστιες συλλογές;

Μερικοί οργανισμοί έχουν αναπτύξει αποτελεσματικές διαδικασίες για την μετατροπή του μεγάλου όγκου του υλικού τους. (Συχνά μέρος της εργασίας στέλνεται σε χώρες όπου οι δαπάνες εργασίας είναι χαμηλές). Εντούτοις, κάθε ένας από αυτούς τους οργανισμούς έχει την δική του μέθοδό. Υπάρχει πληθώρα εργαλείων που χρησιμοποιούνται, αλλά ελάχιστη ανταλλάξιμη εμπειρία. Για τη μετατροπή του κειμένου, η οπτική αναγνώριση χαρακτήρων, η οποία χρησιμοποιεί έναν υπολογιστή για να προσδιορίσει τους χαρακτήρες και τις λέξεις σε μια σελίδα, έχει φθάσει σε ένα επίπεδο αρκετά καλό. Διάφορες ομάδες έχουν αναπτύξει κάποια αξιόλογη εμπειρία, αλλά λίγη από αυτή την πείρα είναι συστηματική ή μπορεί να γίνει κοινή σε όλους.[2]

Οι έννοιες και οι τεχνολογίες που συνδέονται με την ψηφιοποίηση είναι σύνθετες. Υπάρχει μια βασική διαδικασία που περιλαμβάνει διαφορετικά είδη υλικού και λογισμικού που χρησιμοποιούνται σε κάθε βήμα. Ο καθορισμός της κατάλληλης τεχνολογίας συνδέεται άμεσα με την προσδοκώμενη χρήση και το σκοπό της ψηφιοποίησης του υλικού.

Υπάρχουν διάφοροι τρόποι μετατροπής του υλικού σε ψηφιακή μορφή:

- Ο πιο απλός τρόπος μετατροπής, και ευρέως χρησιμοποιούμενος στην πράξη, είναι να δακτυλογραφηθεί εκ νέου το έγγραφο από την αρχή και να προστεθούν οι ετικέτες σήμανσης με το χέρι. Αυτή η μέθοδος έχει το πλεονέκτημα της μεγαλύτερης ακρίβειας για μερικούς τύπους στοιχείων (κατάλογοι, αριθμητικά σύνολα δεδομένων) μη υποκείμενων στα αυτοματοποιημένα μέσα ψηφιοποίησης, και είναι συχνά φτηνότερη από έναν συνδυασμό αυτόματης και ανθρώπινης επεξεργασίας. Ωστόσο η χειρωνακτική εισαγωγή δεδομένων είναι χρονοβόρα - απαιτεί αρκετό εργατικό δυναμικό- και είναι πολύ ακριβή. Δεδομένου ότι η εργασία είναι αρκετά εντατική, πραγματοποιείται συνήθως σε χώρες όπου οι δαπάνες εργασίας είναι χαμηλές. Η Βιβλιοθήκη του Κογκρέσου, η οποία πραγματοποίησε ένα από τα μεγαλύτερα προγράμματα ψηφιοποίησης, γνωστό ως «Πρόγραμμα Μνήμης», έκανε διαγωνισμό και έδωσε την δουλειά σε εξωτερικούς αναδόχους, οι οποίοι κατέληξαν στο συμπέρασμα ότι ο καλύτερος τρόπος ήταν η από την αρχή δακτυλογράφηση.

- Η διαδικασία σάρωσης (scanning), η οποία χρησιμοποιεί υλικό παρόμοιο με τα φωτοτυπικά μηχανήματα (ανιχνευτές), για να πάρει τις ψηφιακές εικόνες των αντικειμένων. Οι ανιχνευτές μπορεί να είναι απλές μηχανές υπολογιστών γραφείου ή πολύ μεγάλα και σύνθετα συστήματα που επεξεργάζονται χιλιάδες έγγραφα. Η φυσική μορφή του αντικειμένου μπορεί να ασκήσει μεγάλη επίδραση στον τύπο εξοπλισμού ανίχνευσης που μπορεί να χρησιμοποιηθεί. Πολλά από τα τρέχοντα συστήματα ανίχνευσης έχουν σχεδιαστεί για επιχειρησιακές εφαρμογές, όπου τα έγγραφα είναι συχνά ενιαία φύλλα ή μέσα σε μια μικρή σειρά μεγεθών, που τα καθιστά ευέλικτα για την αυτόματη ανίχνευση. Το εύθραυστο, τα περίεργα μεγέθη, και οι συνδεδεμένοι όγκοι μερικών υλικών βιβλιοθηκών, παρουσιάζουν μεγαλύτερες δυσκολίες στην ανίχνευση. Στην προκειμένη περίπτωση κάθε έγγραφο σαρώνεται δειγματίζοντας την εικόνα μέσα σε ένα πλέγμα σημείων. Κάθε σημείο αντιπροσωπεύεται από έναν κώδικα φωτεινότητας. Στην απλούστερη μορφή, μόνο το μαύρο και το λευκό διακρίνεται. Με μια ανάλυση 300 dots ανά ίντσα (οριζόντια και κάθετα), μπορούν να παραχθούν καλές εικόνες στις

περισσότερες τυπωμένες σελίδες. Εάν η ανάλυση αυξάνεται στα 600 dots ανά ίντσα, ή εάν οκτώ επίπεδα του γκρίζου κωδικοποιηθούν, μπορούμε να έχουμε άριστη σαφήνεια στην εικόνα. Μία υψηλής ποιότητας αναπαράσταση, απαιτεί τουλάχιστον 24 bits ανά σημείο για να αντιπροσωπεύσει τους κατάλληλους συνδυασμούς χρωμάτων. Αυτό δημιουργεί πολύ μεγάλα αρχεία. Τα αρχεία αυτά, συμπιέζονται για ευκολία στην αποθήκευση και την επεξεργασία, αλλά ακόμη και τα απλά ασπρόμαυρα αρχεία κειμένων χρειάζονται τουλάχιστον 50.000 bytes για να αποθηκεύσουν μια ενιαία σελίδα. Μια σελίδα που έχει ανιχνευτεί αναπαράγει την εμφάνιση της τυπωμένης σελίδας αλλά αντιπροσωπεύει το κείμενο απλά ως εικόνα. Σε πολλές εφαρμογές, αυτό είναι ένα φτωχό υποκατάστατο κειμένου σήμανσης.

- Μια άλλη διαδικασία, είναι αυτή της ανίχνευσης των τυπωμένων σελίδων για να δημιουργηθεί μια ψηφιακή βάση δεδομένων του κειμένου. Αυτή η διαδικασία χρησιμοποιεί το λογισμικό OCR **(**οπτική αναγνώριση χαρακτήρων**)** το οποίο μετατρέπει τους ανιχνευμένους χαρακτήρες του κείμενου σε ισοδύναμους ψηφιακούς χαρακτήρες βάσει κωδικών υπολογιστών. Το λογισμικό πρώτα αναλύει το σχεδιάγραμμα του κειμένου της σελίδας, και μετά διαιρεί το κείμενο σε ζώνες που αντιστοιχούν περίπου στις παραγράφους. Έπειτα καθορίζει την διάταξη των παραγράφων και αρχίζει την ανάλυση του χαρακτήρα. Παρά τις δεκαετίες έρευνας, η οπτική αναγνώριση χαρακτήρα παραμένει μια ανακριβής διαδικασία. Το ποσοστό λάθους ποικίλλει, ανάλογα με το πόσο ευανάγνωστο είναι το αρχικού κείμενο. Εάν το αρχικό έγγραφο είναι σαφές και ευανάγνωστο, το ποσοστό λάθους είναι λιγότερο από 1 τοις εκατό. Όταν όμως έχουμε χαμηλής ποιότητας υλικό, το ποσοστό λάθους μπορεί να είναι πολύ υψηλότερο. Για πολλούς λόγους, ένα ποσοστό λάθους ακόμη και σε μια αναλογία του ενός τοις εκατό είναι πάρα πολύ υψηλό. Αντιστοιχεί σε πολλούς ανακριβείς χαρακτήρες ανά κάθε σελίδα. Διάφορες διαδικασίες έχουν επινοηθεί για να μετριάσουν αυτά τα λάθη. Μια τεχνική είναι να χρησιμοποιηθούν διαφορετικά προγράμματα αναγνώρισης χαρακτήρα για τα ίδια υλικά, με την ελπίδα ότι οι χαρακτήρες που προκαλούν δυσκολία στο ένα πρόγραμμα να μπορούν να επιλυθούν από τα άλλα. Μια άλλη προσέγγιση είναι να χρησιμοποιηθεί ένα λεξικό για να ελέγχει τα αποτελέσματα. Παρόλα αυτά, για να έχουμε υψηλής ποιότητας μετατροπή απαιτείται η ανθρώπινη επέμβαση για την διόρθωση των λαθών που προκύπτουν. Σε μερικά συστήματα, ένα πρόγραμμα υπολογιστή, επιδεικνύει το μετατρεπόμενο κείμενο στην οθόνη και δίνει έμφαση στις λέξεις που αμφισβητούνται, προβάλλοντας μαζί και τις δικές του προτάσεις, έτσι ώστε ο συντάκτης αν θέλει μπορεί να τις δεχτεί ή να τις διορθώσει. Όταν οι μεμονωμένες λέξεις αναγνωριστούν, το επόμενο βήμα είναι να προσδιοριστεί η δομή του εγγράφου και να μπουν οι τίτλοι και άλλα στοιχεία που προσδιορίζουν τη δομή του. Παρά τη σταθερή πρόοδο από παρουσιάζεται τα τελευταία χρόνια, ωστόσο και αυτό απαιτεί επίσης την ανθρώπινη επέμβαση για διόρθωση των λαθών. Όταν οι μεμονωμένες λέξεις αναγνωριστούν, το επόμενο βήμα είναι να προσδιοριστεί η δομή του κειμένου, και να κολληθούν ετικέτες οι επικεφαλίδες καθώς και τα άλλα δομικά στοιχεία. [5,11,2]

Επειδή, όπως αναφέρθηκε και παραπάνω, το OCR παρουσιάζει διάφορα προβλήματα ως προς την ακρίβεια, καλό είναι κατά την εξέταση για την επιλογή του, να έχουμε υπόψη μας τα εξής κριτήρια:

- Αυστηρός καθορισμός του επίπεδου ακρίβειας που θέλουμε, για να ανταποκρίνεται στους ιδιαίτερους στόχους μας.. Οι αποφάσεις για την ακρίβεια πρέπει να λάβουν υπόψη τα χαρακτηριστικά του υλικού της πηγής. Κείμενα που δεν είναι στην αγγλική γλώσσα, μαθηματικά ή χημικά σύμβολα, και άλλοι ειδικοί χαρακτήρες δεν μεταφράζονται επιτυχώς από τις εφαρμογές OCR, και η παρουσία τους πρέπει να ληφθεί υπόψη για την απόφασή μας.

- Μέγεθος του υλικού. Η κατάλληλη προσέγγιση για την παραγωγή των αρχείων κειμένου επηρεάζεται εντυπωσιακά καθώς κινούμαστε από ένα πρόγραμμα 20.000 σελίδων προς ένα πρόγραμμα 200.000 σελίδων, ακόμα κι αν οι στόχοι του προγράμματος είναι οι ίδιοι.

- Ταχύτητα αναγνώρισης

- Κόστος

- Το γεγονός ότι στο μέλλον θα υπάρξουν γρήγορες αλλαγές. Οι ικανότητες λογισμικού OCR έχουν αναπτυχθεί σημαντικά κατά την τελευταία δεκαετία, και οι βελτιώσεις συνεχίζουν να γίνονται. Η δυναμική φύση αυτής της τεχνολογίας σημαίνει, ότι προχωρούμε με γρήγορους ρυθμούς, και τα προγράμματα λογισμικού βελτιώνονται συνεχώς. Συνεπώς θα πρέπει να αξιολογούνται τα νέα προϊόντα που διατίθενται για να καθορίσουμε την καλύτερη δυνατότητα απόδοσης. [5,11]

2. Ψηφιοποίηση κειμένων σε πολλές γραφές

Τα τελευταία χρόνια αρκετοί οργανισμοί σε ολόκληρο τον κόσμο διεξάγουν προγράμματα ψηφιακών βιβλιοθηκών. Αρκετά από αυτά τα προγράμματα περιλαμβάνουν ψηφιοποίηση πληροφοριακών πόρων σε πολλές γραφές. Ένα τέτοιο πρόγραμμα, είναι αυτό του Πανεπιστήμιου της Ιντιάνας το οποίο ανέλαβε την ψηφιοποίηση του Letopis, ενός Ρωσοσοβιετικού εθνικού ευρετηρίου περιοδικών, διάρκειας 20 ετών (1956-1975). Το πρόγραμμα αυτό το οποίο ξεκίνησε στα τέλη του 1999 έχει παρουσιάσει ειδικές τεχνικές προκλήσεις, που οφείλονται εν μέρει στο γεγονός ότι, το Letopis περιέχει υλικό όχι μόνο στα ρώσικα υποσύνολα της κυριλλικής γραφής, αλλά επίσης στα ελληνικά, στη βάση του λατινικού αλφαβήτου, καθώς και τους σύνθετους χαρακτήρες του λατινικού αλφαβήτου (χαρακτήρες που τροποποιούνται από διακριτικά) που χρησιμοποιούνται σε πολλές Δυτικές και Ανατολικές ευρωπαϊκές γλώσσες.

2.1. Επιλογή μιας σειράς κωδικοποιημένων χαρακτήρων, για κείμενα πολλαπλών γραφών

Πολλά ηλεκτρονικά κείμενα που παράγονται από προγράμματα ψηφιακών βιβλιοθηκών, περιλαμβάνουν έγγραφα σε μόνο μία ή δύο γραφές. Τα τελευταία χρόνια πολυάριθμοι οργανισμοί σε πολλές χώρες, έχουν αναπτύξει μια μεγάλη ποικιλία κωδικών σελίδων για ηλεκτρονικά κείμενα. Αυτοί οι κώδικες έχουν προκύψει από τους αρχικούς κώδικες όπως είναι οι CCITT και BCDIC, από τους κώδικες των 7 bit της δεκαετίας το 60 όπως το γερμανικό DIN 66003-1967 και το αμερικάνικο στρατιωτικό FIELDATA, και έχουν βασιστεί στους πρώιμους κώδικες σελίδων των 8 bit όπως είναι οι EBCDIC και ASCII (Αμερικάνικος κώδικας προτύπου για ανταλλαγή πληροφοριών). Κατά τη διάρκεια της δεκαετίας του 80 και 90, οι κώδικες σελίδων των 8 bit χρησιμοποιήθηκαν πάρα πολύ. Στην αρχή της πρώτης δεκαετίας του 21 αιώνα αυτό άρχισε σιγά-σιγά να αλλάζει, με την ανάπτυξη κωδικοποιήσεων πολλαπλών byte. Ωστόσο όμως οι κωδικοποιήσεις των 8 bit εξακολουθούν να παραμένουν σε ευρεία χρήση.

Τα σχήματα κωδικοποίησης των 8 bit, συνήθως αρκούν για να αναπαραστήσουν τους αναγκαίους χαρακτήρες για κείμενα με μια η δύο γραφές, γιατί αυτά τα σχήματα χαρακτήρων των 8 bit, μπορούν να αναπαραστήσουν 256 χαρακτήρες. Γραφές που προέρχονται άμεσα ή έμμεσα από το φοινικικό αλφάβητο (Ελληνικό, Λατινικό, Κυριλλικό κλπ.), γενικά, έχουν λιγότερους από 100 διακεκριμένους χαρακτήρες (υπολογίζοντας τις ανώτερες και χαμηλότερες περιπτώσεις χωριστά, όπως αυτοί βρίσκονται στους κώδικες των σελίδων). Σαν αποτέλεσμα οι 256 δυνατοί χαρακτήρες, συνήθως επιτρέπουν την αναπαράσταση ταυτόχρονα, των ανώτερων και χαμηλότερων περιπτώσεων χαρακτήρων, σε περισσότερες από μία γραφές (επιπλέον και των σημείων στίξης, και των χαρακτήρων ελέγχου του υπολογιστή) μέσα σε ένα δεδομένο κώδικα σελίδας. Τέτοιοι των 8 bit κώδικες σελίδων, συνήθως κατασκευάζονται με το βασικό λατινικό αλφάβητο στην χαμηλή κλίμακα, και με ποικίλους συνδυασμούς γραφών ή σύνθετων χαρακτήρων, στην ανώτερη κλίμακα. Ως εκ τούτου ένας μόνο κώδικας σελίδας, όπως ο «KOH-8» μπορεί να αναπαραστήσει το κείμενο, και στη γλώσσα που χρησιμοποιεί τη βάση του λατινικού αλφαβήτου με σύνθετους χαρακτήρες όπως τα αγγλικά, και στη βάση του ρωσικού αλφάβητου με τους χαρακτήρες που χρειάζονται για τις γλώσσες που στηρίζονται στο λατινικό αλφάβητο στη χαμηλότερη αριθμητική κλίμακα, και επίσης στο ρώσικο υποσύνολο των κυριλλικών χαρακτήρων στην ανώτερη κλίμακα. Αυτή η κωδικοποίηση των 8 bit οδηγεί, σε αυτό το οποίο φέρει τον όρο «περιορισμένη» γλωσσική υποστήριξη πολλαπλών γραφών. Παρόλα αυτά, οι 256 χαρακτήρες είναι σαφώς ανεπαρκείς για να αναπαραστήσουν αρκετές γραφές που βρίσκονται μέσα στα ίδια έγγραφα.

2.2. Το Letopis σαν ένα παράδειγμα κειμένου πολλαπλών γραφών

Λόγω της φύσης του υλικού που έχει ευρετηριαστεί στο Letopis, το πρόγραμμα αυτό ήρθε αντιμέτωπο με την παρουσία πολλών γραφών μέσα σε ένα μοναδικό ηλεκτρονικό έγγραφο. Κάθε εβδομαδιαίο τεύχος του Letopis ευρετηριάζει όλα τα πεδία γνώσης (κοινωνικές, ανθρωπιστικές και θετικές επιστήμες, ιατρική, τεχνολογία, βιομηχανία κλπ.). Ενώ ο κύριος όγκος του Letopis υπάγεται στο ρώσικο υποσύνολο της κυριλλικής γραφής, σε μερικά από τα τμήματα των θετικών επιστημών του ευρετηρίου, υπάρχουν χαρακτήρες του ελληνικού αλφαβήτου, όπως και μοναδικού χαρακτήρα λέξεις και φράσεις του λατινικού αλφαβήτου, καθώς και επιστημονικές/ μαθηματικές σημειώσεις και φραστικοί τύποι. Στις λέξεις και φράσεις των λατινικών γραφών που εμφανίζονται διάσπαρτες στο κείμενο του Letopis, υπάρχουν επίσης και σύνθετοι χαρακτήρες, τόσο από τους δυτικούς ευρωπαϊκούς (Latin 1) κώδικες σελίδων (CP-1252, ISO-9959-1), όσο και από τους ανατολικούς ευρωπαϊκούς (Latin 2) κώδικες σελίδων (CP-1250, ISO-8859-2). Εκτός όμως από τα πρωτότυπα άρθρα, το Letopis ευρετηριάζει και μεταφράσεις άρθρων σε σοβιετικά περιοδικά που έχουν δημοσιευθεί σε άλλες χώρες. Τέτοιες αναφορές συνήθως πρέπει να περιλαμβάνουν μια έκθεση γεγονότων στην πρωτότυπη τοπική γραφή, καθώς και το όνομα του περιοδικού στο οποίο δημοσιεύθηκε το πρωτότυπο άρθρο. Πολλά από αυτά τα άρθρα προέρχονται από περιοδικά που δημοσιεύθηκαν στις τέως χώρες της COMECON, και αλλά προέρχονται από δημοσιεύεις σε δυτικές χώρες. Υπάρχει λοιπόν η ανάγκη να αναπαρασταθούν όλοι οι χαρακτήρες που χρησιμοποιούνται σε γλώσσες, τόσο της Ανατολικής, όσο και της Δυτικής Ευρώπης.

Λόγω του ότι το Letopis είναι ένα πρόγραμμα πολλαπλών γραφών, η χρήση ενός από τα σύνολα των κυριλλικών χαρακτήρων που πιο πολύ χρησιμοποιούνται, όπως το CP-1251, KOH-8 ή ISO-8859-5 μόνο, σαφώς δεν είναι επαρκής. Για να αναπαρασταθεί πλήρως το περιεχόμενο στο Letopis, κάποιος πρέπει να έχει πρόσβαση όχι μόνο στην κυριλλική και βασική γραφή μαζί με τους σύνθετους

χαρακτήρες της Δυτικής Ευρώπης, αλλά επίσης και στους σύνθετους χαρακτήρες της Ανατολικής Ευρώπης και στην ελληνική γραφή. Έτσι αν χρειαζόταν κάποιος να χρησιμοποιήσει για παράδειγμα τα σύνολα χαρακτήρων της Microsoft, θα χρειαζόταν τουλάχιστον όχι μόνο χαρακτήρες CP-1251, αλλά επίσης CP-1252, CP-1250, και CP-1253. Ή αν κάποιος έπρεπε να χρησιμοποιήσει το σύνολο των χαρακτήρων που αντιστοιχούν στο ISO, θα χρειαζόταν χαρακτήρες από το ISO-8859-5,-1,-2 και 7.

2.3. Υπερβαίνοντας το όριο των 256 χαρακτήρων : WGL-4 και Unicode

Στην περίπτωση που ένας πληροφορικός πόρος, περιλαμβάνει μόνο λίγες περιπτώσεις χαρακτήρων που βρίσκονται εκτός του επιλεχθέντος συνόλου βασικών χαρακτήρων, χρησιμοποιούνται πολύ συχνά ποικίλα **workarounds,** όπως ενσωματωμένες εικόνες σε ατομικούς χαρακτήρες, ή λέξεις ή ολότητες αναφορών σε χαρακτήρες. Παρόλα αυτά, τέτοια μέσα μπορούν να κάνουν προβληματική την αναζήτηση κάποιων κειμένων στον υπολογιστή. Υπάρχει ένα ευρύτερο σύνολο χαρακτήρων που εκφράζει μια προσπάθεια υπέρβασης του ορίου των 256 χαρακτήρων, το WGL-4 (Windows Glyph List - 4), το οποίο είναι ουσιαστικά ένα υπέρ σύνολο από κώδικες σελίδων Windows, όπως CP-1250, 1251, 1252,1253 και 1254. Όμως με την υιοθέτηση του Unicode για Windows NT/2002, το WGL-4 δεν έχει ευρέως εφαρμοστεί.
Στην τρέχουσα πρακτική η καλλίτερη επιλογή που απομένει για την κωδικοποίηση κειμένων σε πολλές γραφές, είναι η χρήση ενός κώδικα χαρακτήρων πολλαπλών byte, όπως είναι το Unicode, το οποίο αναπτύχθηκε εν μέρει ως λύση στην διάδοση πολλαπλών, ασύμβατων, και ανεπαρκών συνόλων χαρακτήρων. Το Unicode έχει αναγνωριστεί από τον Διεθνή Οργανισμό Τυποποίησης (ISO) από το 1993 ως παγκόσμιο πρότυπο. Παρέχει τη δυνατότητα κωδικοποίησης όλων των χαρακτήρων που χρησιμοποιούνται από ένα μεγάλο αριθμό γλωσσών του κόσμου, και έτσι ξεπέρασε τον κώδικα ASCII (που καλύπτει μόνο το λατινικό αλφάβητο), και στον οποίο κυρίως βασίστηκε.
Για την κωδικοποίηση του μεγάλου πλήθους των διαφορετικών χαρακτήρων που χρησιμοποιούνται στα αλφάβητα των διαφόρων γλωσσών ο κώδικας Unicode χρησιμοποιεί 16 bits. Τα 16 bits παρέχουν τη δυνατότητα αξιοποίησης 65.536 διαφορετικών συνδυασμών που υπερκαλύπτουν το σύνολο των χαρακτήρων όλων των γνωστών γλωσσών του πλανήτη μας. Έτσι ο κώδικας Unicode, με το πλήθος των συνδυασμών του, επιτρέπει την αναπαράσταση του λατινικού, του ελληνικού, του αρμενικού, του εβραϊκού, του αραβικού, αλλά και πολλών άλλων χαρακτήρων λιγότερο διαδεδομένων γλωσσών. Επίσης καλύπτει, και το ενοποιημένο σύνολο των Κινέζικων , Ιαπωνικών και Κορεάτικων ιδεογραμμάτων. Συμπεριλαμβάνει τα σημεία στίξης, διάφορα διακριτικά, μαθηματικά και τεχνητά σύμβολα, βέλη, τυπογραφικά σημεία κλπ. Με τον τρόπο αυτό διευκολύνονται οι συναλλαγές και η ανταλλαγή αρχείων κειμένου ανάμεσα στις χώρες με διαφορετικές γλώσσες. [9,10]

Η απόφαση να χρησιμοποιηθεί το Unicode για αρχεία κειμένων του Letopis, τα οποία έπρεπε να εγγραφούν σε XML, ανάγκασε το πρόγραμμα να έρθει αντιμέτωπο και με άλλες σχετικές αποφάσεις, όπως ποιος εκδότης κειμένου θα έπρεπε να επιλεχθεί, και ποιες μηχανές αναζήτησης θα έπρεπε να χρησιμοποιηθούν για την πρόσβαση στα δεδομένα. Η XML εξειδίκευση απαιτεί όλα τα XML parsers να είναι ικανά να διαβάσουν κείμενα σε UTF-8 και UTF-16 έκδοση του Unicode. Παρόλα αυτά ο συνδυασμός του Unicode με XML έχει εφαρμοστεί με πολύ πιο αργό ρυθμό τόσο από το εκδότη XML καθώς και τις μηχανές αναζήτησης XML. Για παράδειγμα ένας από τους πιο εμπορικούς XML εκδότες ευρείας χρήσης ο Xmetal δεν είχε μια ευέλικτη εκδοχή του Unicode μέχρι τον Απρίλιο του 2001, πολύ μετά την στιγμή που το πρόγραμμα Letopis είχε αρχίσει να τρέχει. Επιπλέον μερικά πακέτα λογισμικού

χρησιμοποιούν το UTF-8 σαν διορθωτικό του Unicode. Ενώ άλλα χρησιμοποιούν ένα ή περισσότερα, τύπου του UTF-16.
Εξετάζοντας ποικίλους Unicode ευέλικτους απλούς εκδότες κειμένου, όπως και Unicode ευέλικτους XML εκδότες βρέθηκε ότι, το Unicode κείμενο που δημιουργήθηκε ή εκδόθηκε σε έναν εκδότη δεν μπορούσε κατ'ανάγκη να διαβαστεί από άλλον Unicode e-miror. Η εξέταση μερικών Unicode εκδοτών που περιλαμβάνουν Yudit, Linux open source Unicode εκδότη, Unipad, Windows, Microsoft Word 2000 και WordPerfect 9,έδειξε ότι κάποιοι από αυτούς τους εκδότες έμοιαζαν να έχουν ελαφρώς διαφορετικές εφαρμογές από το στερεότυπο του Unicode ή να έχουν παράξενες υποκαταστάσεις των Unicode χαρακτήρων. Για παράδειγμα βρέθηκε ένας εκδότης που μπορούσε να μετατρέψει αυτόματα όλα τα ρωσικού τύπου σημεία που δηλώνουν αγκύλες (Unicode U+00AB και U+00BB) σε λατινικού τύπου σημεία (Unicode U+0022), αλλά κάθε φορά έπρεπε να σώζονται τα αρχεία, να τα κλείνουν και να τα ξανανοίγουν. Έτσι το κωδικοποιημένο κείμενο Unicode για το Letopis πρόγραμμα, από κάποιες απόψεις, δημιούργησε τόσα προβλήματα όσα και έλυσε.
Παρόλα αυτά για τον τελικό χρήστη, η κατάσταση του λογισμικού είναι σχετικά καλή. Η τρέχουσα γενεά των browsers web, όπως ο Netscape 4.x και 6, και ο Internet Explorer 4 και 5 υποστηρίζουν το Unicode Hplay με μια ελάχιστη σχηματοποίηση. Το μέγεθος της αναγκαίας σχηματοποίησης βασίζεται κυρίως στο σύστημα λειτουργίας που χρησιμοποιείται στον υπολογιστή. Για το Microsoft Windows NT 4.0, Windows 2000 ή Windows 98/ME συνήθως δεν είναι αναγκαία κάποια σχηματοποίηση. Οι υπολογιστές που τρέχουν προγράμματα Windows 95 ίσως έχουν ή δεν έχουν εγκατεστημένες Unicode γραμματοσειρές. Ίδια και η περίπτωση των μηχανών Linux που μπορεί να χρειάζεται να έχουν εγκατεστημένες κατάλληλες Unicode γραμματοσειρές. Για τον Macintosh, η υποστήριξη Unicode προστέθηκε με OS 8.5. Στον Macintosh OS 9, μπορεί να χρειαστεί να εγκατασταθεί η γλώσσα kit για Unicode.

3. Οπτική αναγνώριση χαρακτήρων σε Unicode περιβάλλον πολλαπλών γραφών.

Ένας από τους αναφερόμενους σκοπούς του Unicode Consortium's, είναι να αποφεύγεται ο πολλαπλασιασμός της κωδικοποίησης των χαρακτήρων, στο πλαίσιο των γραφών εντός των γλωσσών. Χαρακτηριστικά που είναι ισότιμα στο σχήμα λαμβάνουν ένα μοναδικό κωδικό. Το κριτικό μέρος αυτής της αναφοράς είναι η έκφραση στο πλαίσιο των γραφών.

3.1. Χαρακτήρες που απεικονίζουν γλώσσα (glyphs) έναντι απλών χαρακτήρων

Οι χαρακτήρες που είναι ισοδύναμοι σε μορφή, αλλά που κατηγοριοποιούνται σε διαφορετικές γραφές, αντιμετωπίζονται σαν τελείως διαφορετικές ολότητες με ξεχωριστές Unicode αξίες. Έτσι οι Unicode χαρακτήρες που εμφανίζονται να έχουν όμοια ή ταυτόσημα σχήματα σε διαφορετικές γραφές, αντιμετωπίζονται ως τελείως ξεχωριστοί Unicode χαρακτήρες. Για παράδειγμα το λατινικό γράμμα X έχει την Unicode αξία του U+0058, το ελληνικό Χ έχει την αξία U+03A7 και το κυριλλικό Χ έχει την αξία U+0425. Οι των 8 bit κωδικοί σελίδων, έχουν επίσης το ίδιο πρόβλημα. Για παράδειγμα ο KOX 8, έχει τόσο το λατινικό A και το κυριλλικό A. Όμως το πιο ευρύ σύνολο χαρακτήρων του Unicode πολλαπλασιάζει τον αριθμό των ισοδύναμων σε μορφή χαρακτήρων που περιλαμβάνονται σε ένα μόνο κωδικό σελίδας, και οξύνει το πρόβλημα της οπτικής αναγνώρισης των χαρακτήρων.
Αν και ο διαχωρισμός ομοίων ή ακόμα και ταυτόσημων χαρακτήρων σε ξεχωριστές γραφές, θα μπορούσε σε ορισμένες καταστάσεις να έχει πλεονεκτήματα και μια

ορισμένη λογική σε αφηρημένο επίπεδο (Λατινικό Η και κυριλλικό Η φέρουν πολύ διαφορετικές αξίες για παράδειγμα), σε πρακτικό επίπεδο αυτό μπορεί εμφανώς να αποβεί σε βάρος της ακρίβειας του λογισμικού OCR, πράγμα το οποίο με τη σειρά του έχει συνέπειες στην ικανότητα των μηχανών αναζήτησης να εντοπίζουν μια σειρά χαρακτήρων μέσα σε ψηφιακά κείμενα που έχουν δημιουργηθεί από μια διαδικασία OCR. Σε μικτά κείμενα γραφής όπως είναι το Letopis, εάν το λογισμικό OCR διαμορφωθεί ώστε να αναγνωρίζει πολλαπλές γραφές, υπάρχει ο μεγάλος κίνδυνος αυτό το οποίο σε ένα χρήστη δείχνει σαν κατάλληλη εγγραφή να μην είναι σωστά αναγνωρίσιμο. Για παράδειγμα ένα λατινικό γράμμα Η και ένα ελληνικό Η μπορούν να αντικατασταθούν από ένα κυριλλικό γράμμα Η. Ανάλογα με την επιλεγείσα όψη και το μέγεθος της οικογένειας στοιχείων, αυτό θα μπορούσε να είναι δυσδιάκριτο στο πρόσωπο που διαβάζει το κείμενο, αλλά μια μηχανή αναζήτησης του υπολογιστή σαφώς και θα αποτύγχανε να βρει το κείμενο, γιατί θα έψαχνε για τον χαρακτήρα U+041D, ενώ το κείμενο θα περιείχε χαρακτήρες U+0048 ή U+0397.

Το πρόβλημα είναι ιδιαίτερα εμφανές στα κεφαλαία γράμματα, όπου για παράδειγμα δεν υπάρχει (και πάλι εξαρτάται από τις ειδικές οικογένειες στοιχείων που χρησιμοποιούνται) διαφορά στην εμφάνιση μεταξύ του ελληνικού Α του κυριλλικού Α και του λατινικού Α. Τουλάχιστον στα σχήματα χαμηλότερων περιπτώσεων οι ελληνικοί χαρακτήρες είναι πιο διακριτοί σε σχήμα, από ότι οι λατινικοί και οι κυριλλικοί. Το μέγεθος αυτού του προβλήματος είναι εμφανές, όταν κάποιος εξετάσει τον αριθμό των ομοίων χαρακτήρων στο λατινικό, ελληνικό και κυριλλικό αλφάβητο, δεδομένου ότι και οι 3 γραφές προέρχονται από κοινές ρίζες. Μόνο στην περίπτωση των κεφαλαίων γραμμάτων υπάρχουν τουλάχιστον πάνω από πενήντα πιθανότητες συνδυασμού λάθους, όταν και τα τρία αλφάβητα είναι παρόντα στο κείμενο. Ακόμα και σε μια απλούστερη περίπτωση όπου το κυριλλικό και το λατινικό είναι παρόντα στο κείμενο παραμένουν πολυάριθμές οι δυνατές περιπτώσεις συνδυασμού λαθών. Το παράρτημα παρουσιάζει μερικά παραδείγματα που προέρχονται από το σύνολο της βασικής γλώσσας των ελληνικών χαρακτήρων, το αγγλικό υποσύνολο του λατινικού συνόλου χαρακτήρων, και το ρώσικο υποσύνολο του κυριλλικού συνόλου χαρακτήρων, έτσι ώστε να υπάρξει απεικόνιση του μεγέθους του προβλήματος. Η χρήση του πλήρους κυριλλικού, ελληνικού και βασικού λατινικού, θα μπορούσε να οδηγήσει σε ακόμη περισσότερες περιπτώσεις ισοδύναμων χαρακτήρων.

3.2. Παράγοντες που συμβάλουν σε λανθασμένη αναγνώριση

Από την εφαρμογή του λογισμικού OCR, βρέθηκε ότι οι πιο συχνές περιπτώσεις λάθους, γίνονται στο χαρακτήρα μιας γραφής σε σχέση με το χαρακτήρα μιας άλλης γραφής, όταν ο χαρακτήρας αυτός βρίσκετε σε σχετική απομόνωση. Για παράδειγμα όταν έχουμε το αρχικό ενός κύριου ονόματος ή μεμονωμένους χαρακτήρες όπως οι λατινικοί αριθμοί. Επειδή το λογισμικό OCR που χρησιμοποιήθηκε έλεγχε πλήρες λέξεις σε σχέση με εσωτερικά λεξικά, ήταν λιγότερο πιθανό, να τοποθετήσει ένα γράμμα από μια γραφή στο μέσο μιας πλήρης λέξης από μια άλλη γραφή, η οποία είχε επαληθευτεί σε σχέση με ένα από αυτά τα λεξικά. Δυο παράγοντες συνετέλεσαν στο να είναι τα γράμματα χαμηλότερης περίπτωσης λιγότερο αξιόπιστα ως προς την ακρίβεια του OCR. Αυτά τα γράμματα συνήθως βρίσκονταν μέσα σε ολόκληρες λέξεις, και οι λέξεις αυτές συνήθως ελέγχονταν σε σχέση με το λεξικό που είναι κατασκευασμένο στο εσωτερικό του λογισμικού, και υπήρχαν λιγότερες περιπτώσεις ομοίων χαρακτήρων μεταξύ των γραμμάτων χαμηλής περίπτωσης και στις τρεις γραφές.
Δυστυχώς η δομή των αναφορών που φτιάχνει το πρόγραμμα Letopis ακολουθεί την πρότυπη ρωσική βιβλιογραφική πρακτική, έτσι ώστε όλα τα προσωπικά ονόματα να παρουσιάζονται ως επίθετα, με επιπρόσθετα μόνο τα αρχικά του μικρού και του πατρονομικού ονόματος. Αυτή είναι ακριβώς η κατάσταση, όπου είναι πολύ πιθανή η

σύγχυση στο OCR: μεμονωμένα κεφαλαία γράμματα χωρίς πλαίσιο, με βάση το οποίο, το OCR μπορεί να βεβαιώσει από πια γραφή θα επιλέξει τον πιο κοντινό ισοδύναμο χαρακτήρα που ταιριάζει. Το γεγονός ότι το αρχικό ακολουθείται από μια τελεία, δεν μπορεί να βοηθήσει στο να διακρίνουμε μεταξύ των γραφών, εφόσον η κοινή στίξη είναι ενιαία σε όλες τις γραφές μέσα στο Unicode. Μια τελεία(.) είναι ο χαρακτήρας U+002E, ανεξάρτητα από το γεγονός ότι το περιβάλλον κείμενο μπορεί να είναι ρώσικο, ελληνικό ή να προέρχεται από μια γλώσσα που έχει ως βάση το λατινικό αλφάβητο. Έτσι η παρουσία της τελείας δεν προσθέτει πληροφορίες ως προς την γραφή η οποία αποτυπώνετε σε αυτό.
Μια επιπρόσθετη πολύ γνωστή πηγή λαθών σε όλους τους τύπους του OCR, ήταν η ερμηνεία των μη αναγνωρίσιμων στοιχείων και σημείων πάνω στο χαρτί όπως τα σημεία στίξης ή τα διακριτικά. Αυτό το πρόβλημα μέσω του Letopis παρουσιάστηκε ιδιαίτερα με τους ελληνικούς χαρακτήρες, όπου υπάρχουν μέσα στο σύνολο των ελληνικών χαρακτήρων του Unicode, ορισμένα γράμματα που έχουν επιπρόσθετες σύνθετες μορφές με πρόσθετους τόνους (U+038A) ή διαλυτικά (U+0308). Εμπειρικά βρέθηκε ότι ένα κακοτυπωμένο ή μη διακριτό γράμμα I (U+0049) ερμηνεύτηκε από το λογισμικό OCR σαν κεφαλαίο με διαλυτικά (U+03AA) ή σαν κεφαλαίο ελληνικό με τόνους (U+038A).
Έτσι εκτός από τους OCR λανθασμένους χαρακτήρες λογισμικού με άλλους ισοδύναμους χαρακτήρες, υπήρχε και το πρόβλημα των σύνθετων χαρακτήρων το οποίο οδήγησε σε μείωση της ακρίβειας. Το κεφαλαίο Ι με τις συνδεδεμένες σύνθετες μορφές φάνηκε να είναι μάλλον προβληματικό, γιατί το κεφαλαίο λατινικό I (U+0049) το οποίο συνήθως αναγνωρίζεται λάθος, είναι πολύ πιθανό να εμφανίζεται σε μια σχετική απομόνωση ειδικότερα ως μέρος των λατινικών αριθμών. Άλλα ελληνικά γράμματα που έχουν σύνθετες τέτοιες μορφές, όπως το Ε με τόνο (U=0388) το Η με τόνο (U+0389) και τα λατινικά και τα κυριλλικά γράμματα με τα οποία γίνονται λάθη , έχουν την τάση να εμφανίζονται μέσα σε πλήρες λέξεις, έτσι ώστε η επαλήθευση σε σχέση με τα λεξικά του λογισμικού του OCR να τείνει να ξεκαθαρίσει πολλά από αυτά τα λάθη. Παρόλα αυτά το πρόβλημα δεν έχει περιοριστεί στην αναγνώριση των ελληνικών σύνθετων χαρακτήρων, εφόσον σύνθετοι χαρακτήρες υπάρχουν επίσης σε πολλές παραλλαγές της λατινικής γραφής που χρησιμοποιείται, όπως στα γαλλικά, στα τσέχικα, στα πολωνικά και σε πολλές άλλες.

3.3. Αυξάνοντας την συνολική ακρίβεια μέσω του περιορισμού του αριθμού των γλωσσών που αναγνωρίζονται.

Στη συγκεκριμένη περίπτωση του Letopis, όπου το κύριο σώμα του κειμένου είναι τα κυριλλικά, οι πιθανές προσεγγίσεις του προβλήματος των ομοίων χαρακτήρων που εξετάστηκαν ήταν:

- Να αχρηστευθεί τελείως η αναγνώριση της ελληνικής γλώσσας, και μετά να διορθωθεί το κείμενο, όπου τα ελληνικά φαίνονται κατά την διάρκεια της χειροκίνητης διαδικασίας αναγνώρισης χαρακτήρων.
- Να αχρηστευθεί τόσο η αναγνώριση της λατινικής όσο και της ελληνικής γλώσσας και να γίνει η διόρθωση και στις δυο με το χέρι.
- Να επιτραπεί η πλήρη αναγνώριση και των τριών γραφών και να γίνει η διόρθωση όλων των λαθών που απορρέουν, κατά την διάρκεια της χειροκίνητης διαδικασίας ανάγνωσης των χαρακτήρων.

Διαπιστώθηκε σε σχέση με το πρόγραμμα Letopis ότι, οι εμφανίσεις σύγχρονων ελληνικών γραμμάτων στο κείμενο ήταν πολύ λιγότερες από τον αριθμό των λανθασμένων ελληνικών χαρακτήρων που το λογισμικό OCR είχε εξαγάγει στο κείμενο. Έτσι βρέθηκε ότι στη συγκεκριμένη περίπτωση θα μπορούσε να αυξηθεί η συνολική ακρίβεια των ανεπεξέργαστων εξαγόμενων κειμένων του OCR με την ολοκληρωτική αχρήστευση της αναγνώρισης των ελληνικών. Προς την κατεύθυνση

αυτή, το πρόγραμμα Letopis ήταν τυχερό, γιατί η σχετική παρουσία της ελληνικής γραφής ήταν τόσο χαμηλή ώστε να είναι στην πράξη εφικτό να αχρηστευτεί η αναγνώριση της ελληνικής γραφής, και να προστεθούν οι τυχαίοι ελληνικοί χαρακτήρες μέσα στο κείμενο με το χέρι.
Επίσης βρέθηκε ότι ο αριθμός των σύνθετων λατινικών χαρακτήρων που αναγνωρίστηκαν ως λανθασμένοι, ξεπερνούσε πολύ τον αριθμό των εμφανιζόμενων ως τέτοιων χαρακτήρων, και έτσι αποφασίστηκε να περιοριστεί η αναγνώριση των λατινικών χαρακτήρων μόνο στο βασικό σύνολο χωρίς σύνθετους χαρακτήρες, και όπως στην περίπτωση των ελληνικών χαρακτήρων, να προστεθούν μερικοί σύνθετοι λατινικοί χαρακτήρες πάνω στο κείμενο, κατά τη χειροκίνητη διαδικασία ανάγνωσης. Εάν το κείμενο αποτελείτο από ένα περισσότερο ισορροπημένο μείγμα γραφών, αυτό δεν θα ήταν μια αποδεκτή λύση.
Το Letopis ήταν κατά κάποιο τρόπο τυχερό από το γεγονός ότι, δυνητικά το σύνολο του κυριλλικού κειμένου ήταν σε ρωσική γλώσσα, και έτσι υπήρχε η δυνατότητα να σχηματοποιηθεί το λογισμικό OCR έτσι ώστε να αναγνωρίζει μόνο τους κυριλλικούς χαρακτήρες που χρησιμοποιούνται στη ρωσική γλώσσα. Αυτό είχε το πλεονέκτημα να αποκλειστούν ακόμα μερικοί χαρακτήρες που είχαν χαρακτήρες_ισοδύναμους με τους λατινικούς χαρακτήρες, όπως είναι το J (U+0408) και το S (U+0405) που χρησιμοποιούνται στο σέρβικο υποσύνολο του κυριλλικού.

Συμπέρασμα

Για άλλα δυνητικά προγράμματα ψηφιοποίησης με σύνολα χαρακτήρων σε πολλές γραφές, θα είναι σημαντικό κατά το στάδιο του σχεδιασμού, να αξιολογηθεί ο αριθμός των γραφών που είναι αναγκαίο να συμπεριληφθούν στο στάδιο OCR του προγράμματος. Όσο λιγότερες είναι οι γραφές που πρέπει να αναγνωριστούν, τόσο μικρότερος είναι ο αριθμός των δυνητικά ομοίων ή ισοδύναμων_χαρακτήρων που το λογισμικό OCR μπορεί να αναγνωρίσει λάθος. Από την εφαρμογή προέκυψε ότι, η παρουσία μέσα στο Letopis γραφών πέραν των ρώσικων υποσύνολων του κυριλλικού αλφάβητου, και η απορρέουσα μείωση της ακρίβειας μέσα στο OCR, έχει τριπλασιάσει την ποσότητα του χρόνου που απαιτείται για την διαδικασία ανάγνωσης των εξερχόμενων κειμένων του OCR, σε σχέση με την ανάγνωση κειμένων καθαρά σε ρώσικη γλώσσα.
Καθώς τα προγράμματα ψηφιακών βιβλιοθηκών υπερβαίνουν τα έγγραφα με κείμενα μόνο σε μια ή δυο γλώσσες, το Unicode έχει πιθανότητα να χρησιμοποιηθεί πιο συχνά για την κωδικοποίηση τέτοιων πολύγλωσσων ηλεκτρονικών εγγράφων. Βραχυπρόθεσμα, η υιοθέτηση του Unicode προτύπου, θα μπορούσε να συνεχίσει να βελτιώνετε μέσα σε εκδότες απλών κειμένων, σε XML εκδότες και XML μηχανές αναγνώρισης. Ως αποτέλεσμα, η επιλογή πακέτων λογισμικού προς χρήση με κείμενα Unicode, θα απαιτήσει λιγότερο εντατικό πειραματισμό.
Παρόλα αυτά, το πρόβλημα ισοδύναμων χαρακτήρων εντός των γραφών, φαίνεται να παρουσιάζει ιδιαίτερα προβλήματα για την τρέχουσα γενεά λογισμικών OCR, τα οποία προβλήματα, δεν είναι τόσο φανερά όταν χρησιμοποιούνται περιστασιακές κωδικοποιήσεις των 8 bit, και κατά αυτό τον τρόπο περισσότερο αυστηρά περιορισμένα σύνολο χαρακτήρων. Έτσι η χρήση του Unicode στην κατάσταση την οποία είναι πιο ενδεδειγμένη (κείμενα πολλαπλών γραφών), είναι επίσης η περίπτωση όπου το Unicode μπορεί να οδηγήσει σε δυνητικά προβλήματα για τη διαχείριση της ακρίβειας του OCR. Αυτή η δυνατότητα υποκατάστασης χαρακτήρων μεταξύ όμοιων σχηματοποιημένων χαρακτήρων από διαφορετικές γραφές μπορεί να έχει δυνητικά σοβαρές επιπτώσεις για την ακρίβεια του OCR. Οι δυνητικές ανακρίβειες στα παράγωγα κείμενα, μπορούν να οδηγήσουν σε μείωση της ακρίβειας και της πληρότητας στο σύνολο των αποτελεσμάτων, που προκύπτουν από τις μηχανές αναζήτησης. Η χρήση του Unicode για κάποιους πληροφοριακούς πόρους, είναι αναγκαία για ορισμένους τύπους πολύγλωσσων εγγράφων. Παρόλα

αυτά η πρόσθετη πολυπλοκότητα του χρησιμοποιούμενου Unicode, που συνοδεύεται από τους διάφορους τύπους παγίδων που σκιαγραφήθηκαν σε αυτό το κείμενο, πρέπει να ληφθούν υπόψη κατά το σχεδιασμό των σταδίων κάθε προγράμματος, που μπορεί δυνητικά να χρησιμοποιήσει το σύνολο χαρακτήρων του Unicode.

ΠΑΡΑΡΤΗΜΑ

Μερικά παραδείγματα παρόμοιων χαρακτήρων με κεφαλαία γράμματα στο Ελληνικό, Λατινικό, και Ρωσικό Κυριλλικό αλφάβητο, και η τιμή τους σε Unicode.

Greek script		**Latin script**		**Cyrillic script**	
Α	U+0391	A	U+0041	А	U+0410
Β	U+0392	B	U+0042	В	U+0412
Γ	U+0393			Г	U+0413
Ε	U+0395	E	U+0045	Е	U+0415
Ζ	U+0396	Z	U+005A		
Η	U+0397	H	U+0048	Н	U+041D
Θ	U+0398			Ѳ*	U+0472
Ι	U+0399	I	U+0049	І*	U+0406
Κ	U+039A	K	U+004B	К	U+041A
Μ	U+039C	M	U+004D	М	U+041C
Ν	U+039D	N	U+004E		
Ο	U+039F	O	U+004F	О	U+041E
Π	U+03A0			П	U+041F
Ρ	U+03A1	P	U+0050	Р	U+0420
Τ	U+03A4	T	U+0054	Т	U+0422
Υ	U+03A5	Y	U+0059	у	U+0423
Φ	U+03A6			Ф	U+0424
Χ	U+03A7	X	U+0058	Х	U+0425
		C	U+0043	С	U+0421

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Andrew Spencer, George: Digitization, Coded Character Sets, and Optical Character Recognition for Multi-script Information Resources: The Case of the Letopis 'Zhurnal 'nykh Statei, *LNCS*, 2163, 2001.

[2] Arms W.: *Digital Libraries*, Cambridge Massachusetts, The MIT Press, 2001.

[3] Drenth Pieter: Preservation and access: two sides of the same coin, *LNCS,* 1513, 1998, pp. 743-752.

[4] Hazen Dan, Horrell Jeffrey, Merrill-Oldham Jan: *Selecting Research Collections for Digitization*, Council on Library and Information Resources, 1998.

[5] Haigh Susan: Optical Character Recognition (OCR) as a Digitization Technology, *Network Notes*, 37, 1996.

[6] Kuny Terry: An Introduction to Digitization Technologies and Issues, *Network Notes*, N.14, 1995.

[7] McKay Sally: Digitization in an Archival Environment, *Electronic Journal of Academic and Special Librarianship*, 2003.

[8] Muhleberger Gunter and Stehno Birgit: The Meta Project – Automated Digitization of books and Journals, ECDL 2002, *LNCS* 2458, pp.660, 2002.

[9] Needleman Mark: the Unicode Standard, *Serials Review*, Vol.26, No 2, August 2000, pp. 51-54.

[10] http://www.unicode.org

[11] Sitts Maxine: *Handbook for digital projects: A management tool for preservation and access*, Massachusetts, Northeast Document Conservation Center, 2000.

[12] Vogt – O'Connor Diana: Digitization and Archival Information

[13] Weber Hartmut: Digitization as a Means of Preservation? European Commission on Preservation and Access, Amsterdam, October 1997. **http://www.clir.org/pubs/reports/digpres/digpres.html**