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: 31.1. 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: - ∆υνατότητα 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. - ∆ιαθεσιµότητα 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 4µεταδεδοµένα 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. ∆ιάφορες 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. ∆εδοµένου 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 5περισσότερες 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. ∆ιάφορες 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: 6- 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 ∆υτικές 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. 7Τα 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 ∆υτικής 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 8χαρακτήρες της ∆υτικής 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 ∆ιεθνή 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 9χρησιµοποιούν 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 Χ 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 ΚΟΧ 8, έχει τόσο το λατινικό Α και το κυριλλικό Α. 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 10ορισµένη 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. ∆υο 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. ∆υστυχώς 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 11σύγχυση 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 Ι (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 Ι (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. ∆ιαπιστώθηκε 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 12αυτή, 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 13αυτά 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 A U+0410 Β U+0392 B U+0042 Β U+0412 Γ U+0393 Γ U+0413 Ε U+0395 E U+0045 Ε U+0415 Ζ U+0396 Z U+005A Η U+0397 H U+0048 Η U+041D Θ U+0398 Θ\* U+0472 Ι U+0399 I U+0049 Ι\* U+0406 Κ U+039A K U+004B Κ U+041A Μ U+039C M U+004D Μ U+041C Ν U+039D N U+004E Ο U+039F O U+004F Ο U+041E Π U+03A0 Π U+041F Ρ U+03A1 P U+0050 Ρ U+0420 Τ U+03A4 T U+0054 Τ U+0422 Υ U+03A5 Y U+0059 y U+0423 Φ U+03A6 Φ U+0424 Χ U+03A7 X U+0058 Χ U+0425 C U+0043 C U+0421