Πανεπιστήμιο Αιγαίου
Σχολή Κοινωνικών Επιστημών

ΠΜΣ «Πολιτισμική Πληροφορική και Επικοινωνία»
Ειδίκευση: Μουσειολογίας

«Νέες τεχνολογίες ψηφιακής τεκμηρίωσης χειρόγραφων αρχείων: ψηφιοποίηση και τεκμηρίωση ελληνικού χειρόγραφου αρχείου του 19ου αιώνα με τη χρήση του Transkribus"

Μυτιλήνη 2023

Πανεπιστήμιο Αιγαίου
Σχολή Κοινωνικών Επιστημών

ΠΜΣ «Πολιτισμική Πληροφορική και Επικοινωνία» Ειδίκευση: Μουσειολογίας

Διπλωματική εργασία

«Νέες τεχνολογίες ψηφιακής τεκμηρίωσης χειρόγραφων αρχείων: ψηφιοποίηση και τεκμηρίωση ελληνικού χειρόγραφου αρχείου του 19ου αιώνα με τη χρήση του Transkribus"

των

Βασιλικής Βλάχου

Δήμητρας Πηνελόπης Κασσελούρη

Επιβλέπουσα καθηγήτρια
Ευαγγελία Καβακλή

Εξεταστική επιτροπή
Κωνσταντίνος Κώτης
Ευαγγελία Σαμπανίκου

Μυτιλήνη 2023

«Είμαι συγγραφέας αυτής της Μεταπτυχιακής Διπλωματικής Εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων ή ιδεών, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά για τη συγκεκριμένη μεταπτυχιακή διπλωματική εργασία».

Βασιλική Βλάχου Δήμητρα Πηνελόπη Κασσελούρη

Ευχαριστίες

Θα θέλαμε να ευχαριστήσουμε θερμά την επιβλέπουσα καθηγήτριά μας κα Καβακλή Ευαγγελία για την εμπιστοσύνη που μας έδειξε καθώς, και τις οικογένειές μας για την υπομονή τους και την αμέριστη υποστήριξή τους για την εκπλήρωση αυτού του πονήματος.

Περίληψη

Στον τομέα της ιστορικής έρευνας νέες τεχνολογίες έρχονται να εμπλουτίσουν τις τεχνικές έρευνας σε χειρόγραφες πηγές. Στις μέρες μας η επεξεργασία της αναγνώρισης και μεταγραφής χειρόγραφων κειμένων απασχολεί ένα ποικίλο ερευνητικό πεδίο. Στην παρούσα εργασία παρουσιάζονται οι πρόσφατες εξελίξεις στην ψηφιακή τεκμηρίωση χειρόγραφων αρχείων. Μετά από ανασκόπηση των τεχνολογιών που εξυπηρετούν την διαδικασία αυτόματης μεταγραφής των χειρογράφων, επιλέχθηκε και χρησιμοποιήθηκε ως μελέτη περίπτωσης η πλατφόρμα Transkribus για την αξιοποίηση και εφαρμογή της τεχνολογίας HTR σε ελληνικά χειρόγραφα αρχεία. Στη συνέχεια, για τη διάθεση των πληροφοριών στο κοινό, παρουσιάστηκε το λογισμικό δημιουργίας ψηφιακών βιβλιοθηκών Greenstone. Από τη διαδικασία εξαγωγής και αποδόμησης της πληροφορίας ενός χειρόγραφου αρχείου του 19ου αιώνα, του συμβολαιογράφου Λαρίσης Α. Φίλιου και τη διάθεσή του στους ερευνητές, προέκυψε ότι η ολοένα και μεγαλύτερη χρήση αυτών των τεχνολογιών ανατρέπει βασικές παραδοχές και διαμορφώνει ταυτόχρονα νέες προϋποθέσεις κόστους και χρόνου, εξασφαλίζοντας τη δυνατότητα αυξημένης αλληλεπίδρασης μεταξύ των φορέων της γνώσης.

Abstract

In the field of historical research new technologies are coming to enrich the research techniques in manuscript sources. Nowadays, the processing of the identification and transcription of handwritten texts occupies a diverse research field. In this paper, the recent developments in the digital documentation of manuscripts are presented. After a review of the technologies that serve the process of automatic transcription of handwritten texts, the Transkribus platform was selected and used as a study case for the utilization and implementation of HTR technology to Greek manuscripts. Then, to make the information available to the public, Greenstone digital library creation software was introduced. From the process of extracting and deconstructing the information of a 19th century manuscript file, of Larisa's notary A. Filios and making it available to researchers, it emerged that the increasing use of such technologies overturns basic assumptions and at the same time configures new cost and time conditions, ensuring the possibility of increased interaction among knowledge sectors.

Λέξεις κλειδιά

Αναγνώριση χειρόγραφου κειμένου, Τεχνητή Νοημοσύνη, ψηφιοποίηση, Transkribus, Greenstone, ελληνικά χειρόγραφα αρχεία, αρχείο συμβολαιογράφου Φίλιου.

Key words

Handwriting text recognition, Artificial Intelligence, digitization, Transkribus, Greenstone, Greek manuscript, Filios notary's file.

Περιεχόμενα

Πίνακας εικόνων

Συντομογραφίες

ΓΑΚ - Γενικά Αρχεία του Κράτους
ΤΝ - Τεχνητή Νοημοσύνη (Artificial Intelligence - AL)
CER - Character Error Rate (Ποσοστό σφάλματος χαρακτήρων)
DNN - Deep Neural Network - Βαθιά Νευρωνικά Δίκτυα
HMM - Hidden Markov Models
HTR - Handwriting Text Recognition -Αναγνώριση χειρόγραφου κειμένου
IDP - Intelligent Document Processing - Ευφυή Επεξεργασία Εγγράφων
KWS - Keyword Spotting - Εντοπισμός λέξεων κλειδιά
ΜΜ - Machine Learning - Μηχανική Μάθηση
NN - Neural Network - Νευρωνικά Δίκτυα

Εισαγωγή

Η γραφή αποτελεί το μέσο όπου διαχωρίζεται το μήνυμα από τη ζωντανή παρουσία. Η τυπογραφία συνεχίζει τη διαδικασία αυτή αποσπώντας το άμεσο ίχνος που αφήνει η διαδικασία του γράφοντος. Η πληροφορική είναι αυτή που επιταχύνει τη διαδικασία της γραφής, και από τον παραδοσιακό τρόπο γραφής περνάμε σε ποικίλους τελεστές αυτής, σε νέα συστήματα σημείων και σε όλες τις οπτικές και ηχητικές μορφές (ψηφιακό κείμενο), οι οποίες αλληλεπιδρούν μεταξύ τους (Levy, 2001, σ.115). Ωστόσο, αυτή η διεύρυνση του ψηφιακού κειμένου, το οποίο είναι ευκίνητο, τροποποιείται κατά βούληση και έχει την ικανότητα να κινείται σε διεθνή δίκτυα, χωρίς να έχει εξοβελίσει το χειρόγραφο κείμενο. Πέραν τούτου τα χειρόγραφα κείμενα αποτελούν τροχοπέδη της λήθης του παρελθόντος και πάντα θα δεσπόζουν στη συλλογική μνήμη των ανθρώπων. Ακόμη και αν τροποποιηθούν σε ψηφιακή μορφή δεν τίθεται θέμα διάκρισης πρωτοτύπου και αντιγράφου καθώς έχει πάψει πλέον να είναι προσφυής (Levy, 2001, 62-63).

Κάλλιστα θα μπορούσε να ισχυριστεί κανείς, ότι χάρη στην ψηφιοποίηση των χειρογράφων, παρατηρείται μία άνθηση στην ερευνητική διαδικασία καθώς δίνεται η δυνατότητα στον ερευνητή να ανατρέξει σε πηγές που σύμφωνα με την παραδοσιακή διαδικασία πολλές φορές θα ήταν αδύνατο να τις προσεγγίσει. Παραταύτα, ανακύπτει το πρόβλημα της ανάγνωσης των χειρόγραφων κειμένων, καθώς χρειάζεται εξειδικευμένες γνώσεις για να μπορέσει να ξετυλιχτεί και να σημασιοδοτηθεί το περιεχόμενό τους. Η δυνατότητα της μεταγραφής χειρόγραφων κειμένων κρίνεται απαραίτητη καθώς επικουρεί στην ερμηνεία γεγονότων του παρελθόντος. Παραδοσιακά για την ανάκτηση αυτών των πληροφοριών χρειαζόταν γνώσεις παλαιογραφίας.

Το ενδιαφέρον γύρω από το θέμα της μεταγραφής των χειρόγραφων κειμένων αυξάνεται όλο και περισσότερο στην παγκόσμια κοινότητα. Ιδιαίτερα, σε ένα υβριδικό περιβάλλον συνύπαρξης του χειρόγραφου και του εντύπου με το ψηφιακό, οι λειτουργίες και οι παρεχόμενες ψηφιακές υπηρεσίες δημιουργούν νέα δεδομένα αντιμετώπισης του προβλήματος. Από τον 20ο αιώνα η συμβολή ανάπτυξης της τεχνολογίας μέσω των υπολογιστικών συστημάτων εμπλουτίστηκε με νέες μεθόδους. Τον 21ο αιώνα οι μεταβολές στην τεχνολογία είναι ραγδαίες καθώς εισάγονται νέες τεχνολογίες. Η τεχνητή νοημοσύνη με τη χρήση της μηχανικής μάθησης, σε συνδυασμό με τα νευρωνικά δίκτυα προσδίδουν εξαιρετικές δυνατότητες στη μεταγραφή των χειρόγραφων κειμένων. Αναπτύχθηκαν λογισμικά όπως τα OCR και HTR προκειμένου να αντιμετωπιστεί το πρόβλημα της μεταγραφής των χειρογράφων. Με αφετηρία την έρευνα στο πεδίο χρήσης των τεχνολογιών αυτών, το ενδιαφέρον των ερευνητών επικεντρώθηκε στην τεχνολογία HTR και συγκεκριμένα στην πλατφόρμα αυτόματης αναγνώρισης κειμένου, Transkribus.

Το HTR βασίζεται σε τεχνολογίες Al και σε νευρωνικά δίκτυα. Τα νευρωνικά δίκτυα πρέπει να εκπαιδευτούν χρησιμοποιώντας εικόνες και μεταγραφές για κάθε γραμμή του χειρόγραφου κειμένου. Εν ολίγοις, η εκπαίδευση μοντέλων HTR είναι μια περίπτωση εποπτευόμενης μηχανικής εκμάθησης. Χρησιμοποιώντας ένα αρχείο του 18ου - 19ου αιώνα, του συμβολαιογράφου Λαρίσης Φίλιου Αναστάσιου, δημιουργήθηκαν και

εκπαιδεύτηκαν μοντέλα Al προκειμένου να αναγνωρίσουν και να μεταγράψουν ελληνικά χειρόγραφα της συγκεκριμένης χρονικής περιόδου.

Σκοπός της παρούσας εργασίας είναι η διερεύνηση των τεχνολογιών που μπορούν να βοηθήσουν το έργο της αυτόματης μεταγραφής χειρόγραφων αρχείων και η εκτίμηση της αποτελεσματικότητάς τους όταν αυτές εφαρμόζονται σε ελληνικά χειρόγραφα αρχεία. Στο πλαίσιο αυτό γίνεται αξιολόγηση της πλατφόρμας Transkribus ως αυτόματο εργαλείο μεταγραφής ελληνικών χειρογράφων, το οποίο μπορεί να χρησιμοποιηθεί από πλειάδα πολιτιστικών φορέων. Φιλοδοξία αποτελεί να παρουσιαστεί μια τεκμηριωμένη επισκόπηση για τις δυνατότητες του Transkribus, λαμβάνοντας υπόψη τους διαφορετικούς τρόπους γραφής των ελληνικών χειρογράφων. Αναλύονται τα ποσοστά επιτυχίας, μετά την εκπαίδευσή του και εξάγονται συμπεράσματα για το πώς το σύστημα HTR, που ενσωματώνεται στην πλατφόρμα του Transkribus, μπορεί να βοηθήσει φορείς και μελετητές στη μεταγραφή χειρογράφων.

Συγκεκριμένα στο πρώτο κεφάλαιο γίνεται αναφορά στην ψηφιοποίηση των χειρογράφων αναφέροντας τα πλεονεκτήματα και μειονεκτήματά της. Παράλληλα ορίζονται οι νέες καινοτομίες της τεχνολογίας, όπως η τεχνητή νοημοσύνη και η μηχανική μάθηση, που εξυπηρετούν το σκοπό της εργασίας, αναλύοντας περιεκτικά τη χρησιμότητά τους. Επίσης, γίνεται αναφορά του ερευνητικού στόχου και της μεθοδολογικής προσέγγισης της εργασίας.

Το κεφάλαιο δύο αποτελεί μια ανασκόπηση των υπαρχουσών τεχνολογιών για την αναγνώριση χειρογράφου. Αναφέρεται η σημασία ανάπτυξης αυτών για ένα ευρύ φάσμα επιστημονικών κλάδων και επιχειρείται μια σύντομη ανασκόπηση της κοινότητας της πλατφόρμας του Transkribus. Πολλοί οργανισμοί, όπως βιβλιοθήκες και αρχεία ψηφιοποιούν μεγάλης κλίμακας συλλογές με σκοπό να προσφέρουν στον χρήστη μια εύχρηστη περιήγηση και ανεμπόδιστη επαφή με αυτά τα πολύτιμα κείμενα γνώσης. Η ανάπτυξη της τεχνολογίας ανοίγει νέους ορίζοντες στα χειρόγραφα, δίνοντάς τους νέα πνοή, αφυπνίζοντας παράλληλα το παρελθόν. Ως εκ τούτου, επικεντρώνεται στη σημασία διατήρησης και αξιοποίησης της τεχνολογίας για τα ελληνικά χειρόγραφα αναφέροντας παράλληλα τους φορείς που συντελούν για την ανάδειξη και συντήρηση αυτών.

Το τρίτο κεφάλαιο εστιάζει και αποτελεί τον κύριο στόχο της παρούσας εργασίας περιγράφοντας τα βήματα που χρειάστηκαν για την μεταγραφή του αρχείου με τη χρήση της πλατφόρμας Transkribus, προκειμένου ο εκάστοτε ερευνητής να έχει απρόσκοπτη πρόσβαση πολύπλευρα στα χειρόγραφα. Επίσης, για την ολοκλήρωση και προεπισκόπηση της συλλογής γίνεται λεπτομερής αναφορά στη δημιουργία συλλογής του διαδικτυακού περιβάλλοντος του Greenstone.

Τέλος, στο τέταρτο κεφάλαιο γίνεται συζήτηση για την χρηστικότητα της πλατφόρμας Transkribus, τόσο από τα αποτελέσματα που εξήχθησαν από την δημιουργία μοντέλων εκπαίδευσης για το αρχείο του συμβολαιογράφου Φίλιου, όσο και από προσπάθειες άλλων ερευνητών σε διαφορετικές γλώσσες και παρέχονται συνοπτικά τα αποτελέσματα του πονήματος.

1. Θεωρητικό πλαίσιο

1.1 Η Ψηφιοποίηση των χειρογράφων και των εντύπων

Στην ιστορία του πολιτισμού η τεχνολογία της γραφής καταλαμβάνει κεντρική θέση στον κατάλογο των επινοήσεων του ανθρώπινου γένους και αποτελεί ορόσημο στην τεχνολογία της πληροφορίας. Η γραφή αποτελεί ένα εργαλείο αντικειμενοποίησης των σκέψεων, των συναισθημάτων, των γεγονότων και των απόψεων. Αποτέλεσε με λίγα λόγια τη σωρευτική ανάπτυξη της γνώσης. Κατόπιν η ανακάλυψη της τυπογραφίας αποτέλεσε σημαντικό αρωγό για την ανάπτυξη της πληροφορίας, ωστόσο η δυναμική της "εξοβελίστηκε" από την ηλεκτρονική και ψηφιακή επανάσταση (Eriksen, 2005, σ. 65-68).

Η δεκαετία του 1990 υπήρξε αναμφίβολα η δεκαετία της ψηφιοποίησης. Αρχεία, μουσεία, πανεπιστήμια μερίμνησαν για να εξασφαλίσουν χρηματοδοτήσεις προκειμένου να ψηφιοποιήσουν τις συλλογές τους. Κάθε είδους χειρόγραφα, έντυπα βιβλία, διοικητικά έγγραφα, αντικείμενα τέχνης και πληθώρα διαφόρων ειδών ιστορικών τεκμηρίων ψηφιοποιήθηκαν από ποικίλα συστήματα ψηφιακής μεταφοράς. Στις αρχές της επόμενης δεκαετίας μεγάλο μέρος της παγκόσμιας πολιτισμικής κληρονομιάς είχε υποστεί ψηφιακό αναδιπλασιασμό, και εξαιτίας της ανάπτυξης του Παγκόσμιου Ιστού, προσφερόταν για μελέτη, μετασχηματισμό και αξιοποίησή του. Απόρροια αυτής της εκτενούς ψηφιοποίησης αποτέλεσε ο μετασχηματισμός των Κοινωνικών και Ανθρωπιστικών Επιστημών σε ένα νέο κλάδο τις Ψηφιακές Ανθρωπιστικές Σπουδές (Πατηνιώτης, 2020, σ.7-9).

Επίσης, το μέγεθος και η ποικιλία των ψηφιακών αρχείων τόσο αυτών που προκύπτουν από την ψηφιοποίηση, όσο και των γεννημένων ψηφιακά, καθώς και η εμφάνιση νέων τεχνολογιών, δημιούργησαν τεράστιες προκλήσεις στους ερευνητές αλλά και στα ιδρύματα μνήμης, απαιτώντας καινοτόμες προσεγγίσεις για τον καλύτερο τρόπο αποτύπωσής τους και διατήρησής τους.

Η ραγδαία ανάπτυξη της ψηφιακής τεχνολογίας των τελευταίων ετών επέτρεψε στα ποικίλα σώματα των αρχείων να μετεγκατασταθούν μαζικά στο Διαδίκτυο. Η ψηφιοποίηση των αρχείων έχει ως απόρροια την μεγάλη κλίμακα ποικιλομορφίας των συλλογών και τους διαφορετικούς τύπους μέσων, μέσω των οποίων μπορούν να γίνουν προσβάσιμα από το διαδίκτυο. Τα διαδικτυακά βοηθήματα εύρεσης και τα εργαλεία αναζήτησης έχουν αντικαταστήσει τους παλαιότερους καταλόγους καρτών. Οι ψηφιακές φωτογραφίες έχουν αντικαταστήσει τις φωτοτυπίες και οι φορητοί υπολογιστές παρέχουν την δυνατότητα εξερεύνησης διαδικτυακών συλλογών χωρίς να κρίνεται αναγκαία η φυσική παρουσία του ερευνητή.

Ωστόσο ανακύπτει το ερώτημα τί εννοούμε όταν χαρακτηρίζουμε ένα αντικείμενο ψηφιακό. Σύμφωνα με το Εθνικό Κέντρο Τεκμηρίωσης και Ηλεκτρονικού Περιεχομένου (ΕΚΤ) (2020), «*Ψηφιακό αντικείμενο ενός πολιτιστικού αντικειμένου είναι μία ψηφιακή μορφή (εικόνα, κείμενο, ήχος, βίντεο / κινούμενη εικόνα ή τρισδιάστατη αναπαράσταση) η οποία αντιπροσωπεύει, αναπαριστά, αναπαράγει ή - για εξ' αρχής ψηφιακό δημιούργημα - αποτελεί το ίδιο το πολιτιστικό αντικείμενο*». Το ψηφιακό

αντικείμενο σύμφωνα με τον σκοπό για τον οποίο χρησιμοποιείται αποτελείται από ένα ή περισσότερα ψηφιακά παράγωγα με ποικίλα χαρακτηριστικά, όπως το μέγεθος, το μορφότυπο κ.λπ.. Διακρίνονται τρεις βασικοί τύποι παραγόμενων αρχείων: α) αρχείο πολύ υψηλής ποιότητας για ψηφιακή διατήρηση, β) αρχείο υψηλής ποιότητας για διάθεση μέσω διαδικτύου, γ) αρχείο μέτριας ποιότητας για προεπισκόπηση (Εθνικό Κέντρο Τεκμηρίωσης και Ηλεκτρονικού Περιεχομένου [ΕΚΤ], 2020, σ.10).

Παράλληλα ένα ψηφιακό αντικείμενο μπορεί να διακρίνεται σε α) αντικείμενο ψηφιοποίησης (ditigized) το οποίο αποτυπώνεται οπτικά ως ένα συμβατικό πολιτιστικό αντικείμενο και διαμορφώνεται με τη διαδικασία ψηφιοποίησης - η οποία προκύπτει με χρήση ποικίλων μεθόδων - όπως σάρωση- σε ένα ψηφιακό. και β) σε ένα εξ' αρχής ψηφιακό αντικείμενο (born-digital) του οποίου η οντότητα διαμορφώθηκε από την αρχή σε ψηφιακή μορφή και όχι κατόπιν διαδικασίας ψηφιοποίησης (Εθνικό Κέντρο Τεκμηρίωσης και Ηλεκτρονικού Περιεχομένου [ΕΚΤ], 2020, σ.10-11).

Σημαντικό στοιχείο αποτελεί η κατανόηση ότι η φυσική μορφή των εγγράφων ως έντυπα ή χειρόγραφα καθορίζει και την αρχιτεκτονική τους διάταξη και δομή καθώς και τον τρόπο πρόσβασης. Η ψηφιοποίησή τους αλλάζει τη φύση τους, τη σύστασή τους και τη δομή τους καθώς και την προσβασιμότητά τους καθιστώντας την περισσότερο ευεπίτευκτη (Poster, 2004, σ.23-24). Τα ψηφιακά αρχεία αποτελούν νέα προϊόντα πράξης δημοσίευσης, παραγόμενα με εντελώς νέους όρους από τα συμβατικά, τα οποία ακόμη βρίσκονται στο στάδιο διαμόρφωσής τους. Αυτό απορρέει από το γεγονός ότι το περιβάλλον διαμόρφωσής τους βρίσκεται σε μια διαρκή και με ταχείς ρυθμούς εξέλιξη, ώστε καθίσταται δύσκολο να αποκτήσει μια παγιωμένη τυπολογία. Τα ψηφιακά αρχεία υιοθετούν σε μεγάλη κλίμακα τις τυπολογικές μορφές των συμβατικών οι οποίες υλοποιούνται α) με ψηφιοποίηση, σε πανομοιότυπη μορφή με το συμβατικό πρωτότυπο και β) μέσα από την πρωτογενή παραγωγή ψηφιακών εκδοχών συμβατικών πρωτοτύπων (Μπώκος, 2005, σ.58-59).

Εν κατακλείδι, σύμφωνα με τον Manovich (2020, σ.249-250), τα νέα τεχνολογικά μέσα του ψηφιακού κόσμου αποτελούν μία νέα μοντερνιστική αβαγκάρντ των παλαιών μέσων. Οι νέες τεχνολογίες δεν έχουν να κάνουν με το "βλέπειν" και την αναπαράσταση του κόσμου με νέους τρόπους, αλλά τα νέα μέσα χρησιμοποιούν τα παλαιά ως πρωτογενές υλικό.

1.1.1 *Πλεονεκτήματα και μειονεκτήματα της ψηφιοποίησης*

Ο αναγνώστης ενός βιβλίου που είναι τυπωμένο, έχει στα χέρια του ένα φυσικό αντικείμενο, στο οποίο παρουσιάζεται εξ ολοκλήρου μια εκδοχή του κειμένου. Πάνω σε αυτό το φυσικό αντικείμενο έχει τη δυνατότητα να προσθέσει σημειώσεις, να το κόψει, να ανατρέξει στα περιεχόμενα, στις συλλογές εικόνων, πίνακες ονομάτων ή όρων-κλειδιών κ.λπ.. Το ψηφιακό έρεισμα εντούτοις παρουσιάζει μια μικρή διαφορά, η αναζήτηση στους πίνακες και γενικότερα η χρήση των μέσων προσανατολισμού γίνονται με μεγαλύτερη ταχύτητα. Επίσης, η ψηφιοποίηση επιτρέπει να συνδέσουμε ή ακόμη και να αναμείξουμε εικόνες ή και κείμενα πάνω στο ψηφιακό έρεισμα. Αποτελεί δηλαδή ένα δίκτυο ταχείας και άμεσης πλοήγησης (Levy, 2001, σ.51, 57).

Σύμφωνα με τον Levy (2001, σ.58) «*η ψηφιοποίηση εισάγει μια μικρή κοπερνίκεια επανάσταση*». Ο χρήστης δεν καλείται πλέον να ακολουθεί τις οδηγίες της ανάγνωσης

και να μετακινείται σε ένα φυσικό χώρο, βηματίζοντας σε μια βιβλιοθήκη, γυρνώντας σελίδες, ψάχνοντας στα ράφια, αλλά αποτελεί πλέον ένα "*ευκίνητο*" "*καλειδοσκοπικό κείμενο*" το οποίο έχει την ικανότητα να διπλώνεται και να εκδιπλώνεται με προθυμία μπροστά στον αναγνώστη παρουσιάζοντας όλες τις όψεις του.

Τα ψηφιακά αρχεία είναι συνυφασμένα με την πληροφορία και παράλληλα με τις έννοιες "επικοινωνία" και "μετάδοση". Επίσης, χαρακτηρίζονται από πολυχρηστικότητα καθώς δίνουν τη δυνατότητα χρήσης της ίδιας πληροφορίας από περισσότερους του ενός χρήστη την ίδια χρονική στιγμή. Παράλληλα, τα ψηφιακά αρχεία αποδεσμεύονται από χωρικούς και χρονικούς περιορισμούς. Δύναται η δυνατότητα πρόσβασης σε οποιοδήποτε γεωγραφικό χώρο και χρόνο. "Το εδώ και το τώρα" της έντυπης πληροφορίας μετατρέπεται σε "οπουδήποτε και οποτεδήποτε" στο ψηφιακό περιβάλλον. Πλέον αντί ο χρήστης να μετακινείται στην πληροφορία, η πληροφορία μεταφέρεται στο χώρο του. Στην ουσία έχει επέλθει μια αποτοπικοποίηση του κειμένου δίνοντας τη δυνατότητα μια διαρκούς συμμετοχής στο ψηφιακό κοινό γίγνεσθαι. Ωστόσο, η αξιοποίησή τους μπορεί να αποκτήσει διαφορετική χρησιμότητα από τις προθέσεις ή προβλέψεις των δημιουργών τους. Επιπλέον, η πρόσβασή τους προϋποθέτει τη διαμεσολάβηση τεχνολογικών συσκευών και ο χρήστης απαιτείται να διαθέτει δεξιότητες και εξοικείωση με το μέσο αναπαράστασης (Τσιμπόγλου, 2006, σ.35-38). Η ψηφιακή αναπαράσταση δεν αποτελεί αντικείμενο που δημιουργήθηκε από τη φύση αλλά είναι δημιούργημα του ανθρώπου. Η κατασκευή της είναι δομημένη με τέτοιο τρόπο για να τη χειρίζονται τα ψηφιακά μέσα καθιστώντας τη σύνθετη στο χειρισμό της (Καπιδάκης, 2010, σ.32).

Τα ψηφιακά αρχεία έχουν αρκετά κοινά στοιχεία με τα συμβατικά καθώς στην ουσία τα μιμούνται σε μεγάλο βαθμό. Παρ' όλα αυτά έχουν τα δικά τους μορφοποιητικά στοιχεία τα οποία τα διαφοροποιούν. Επιπροσθέτως, παρατηρείται σύγκλιση των μέσων δηλαδή διαφορετικοί τύποι περιεχομένου, όπως χειρόγραφα, βιβλία άρθρα, χάρτες, εικόνες, ηχογραφήσεις κ.α., όταν ψηφιοποιηθούν παύουν να διακρίνονται σε διαφορετικούς τύπους πληροφορίας και αποτελούν ένα ψηφιακό πληροφοριακό περιεχόμενο (Τσιμπόγλου, 2006, σ.38-40). Σύμφωνα με τον Πατηνιώτη (2020, σ.13), η ψηφιοποίηση δεν αποτελεί μια ισχνή εκδοχή «*αυτού που υπάρχει πραγματικά*» αλλά αποτελεί μια μετάβαση σε ένα διαφορετικό καθεστώς πραγματικότητας.

Ένα μεγάλο πρόβλημα που ανακύπτει είναι η τεράστια ποικιλία και ασυμβατότητα των μορφοτύπων. Τα ψηφιακά αρχεία που δημιουργούνται θα πρέπει να είναι αναγνώσιμα και μετά από χρόνια. Επίσης, ανακύπτουν και νομικά προβλήματα, όπως η πιθανή παραβίαση των πνευματικών δικαιωμάτων. Η νομική διάσταση έγκειται στη γενικότερη δυναμική και κωδικοποίηση της κοινωνικής και οικονομικής οργάνωσης που επικρατεί (Καπιδάκης, σ.104). Ένα προτέρημά τους είναι η μη ανταγωνιστικότητα τους, δηλαδή η οικονομική τους διάσταση. Η φυσική ιδιότητα της μη αναλωσιμότητάς τους καθιστά την πληροφορία που παρέχουν μη ανταγωνιστική στη χρήση της καθώς δύναται να χρησιμοποιηθεί παράλληλα από πλειάδα χρηστών χωρίς ωστόσο να αναλωθεί ή και να απολεσθεί η κατοχή τους (Τσιμπόγλου, 2006, σ.36).

Η παρουσίαση και η διάθεση του ψηφιακού αρχείου στους χρήστες μέσω διαδικτυακού περιβάλλοντος για την ευρεσιμότητα και την επαναχρησιμοποίηση των ψηφιακών πόρων δημιούργησε την ανάγκη να οικοδομηθούν και να αναπτυχθούν σημαντικά αποθετήρια

γνώσης. Ακαδημαϊκά ιδρύματα και πολιτιστικοί οργανισμοί έθεσαν ως στόχο τους τη δημιουργία βάσεων δεδομένων και λειτουργικών διεπαφών με τον χρήστη καθώς και προγράμματα πληθοπορισμού και προγράμματα που θα βελτίωναν την διαδραστικότητα των εφαρμογών.

1.2 Τεχνητή νοημοσύνη, βαθιά μηχανική μάθηση

Η κατανόηση των χειρογράφων πέρα από πνευματικό αγαθό αποτελεί τον καμβά και τον καθρέπτη του επιπέδου πολιτισμού ενός λαού και του παρελθόντος του. Αποθετήρια και τόποι φύλαξης αυτών των θησαυρών αποτελούν οι σύγχρονες βιβλιοθήκες και τα αρχεία. Ωστόσο, με την ανάπτυξη της τεχνολογίας περάσαμε στην ψηφιακή φύλαξη αυτών των ανεκτίμητων πνευματικών προϊόντων.

Όσο η τεχνολογία εξελίσσεται, η επικοινωνία ανθρώπου υπολογιστή ακμάζει και σε αυτό αναμφίβολα έχει συμβάλει η τεχνητή νοημοσύνη (ΤΝ). Η ΤΝ έχει συμπληρώσει πάνω από μισό αιώνα ζωής και εύλογα αποτελεί μία από τις πιο επίκαιρες ερευνητικές περιοχές της επιστήμης των υπολογιστών. Η ΤΝ αφορά τον τομέα της επιστήμης των υπολογιστών, που ως στόχο της θέτει τη σχεδίαση ευφυών υπολογιστικών συστημάτων, τα οποία προσομοιάζουν με την ανθρώπινη συμπεριφορά.

Οι ορισμοί που έχουν δοθεί για την ΤΝ θα μπορούσαν να ταξινομηθούν σε τέσσερις μεγάλες κατηγορίες. Οι κατηγορίες αφορούν την ανάπτυξη συστημάτων που σκέφτονται ως άνθρωποι, αυτές που σκέφτονται λογικά, εκείνες που συμπεριφέρονται όπως οι άνθρωποι και που αντιδρούν λογικά. Ένας γενικός ορισμός που προτείνεται από τους Βλαχάβας, Κεφαλάς, Βασιλειάδης, Κόκκορας, και Σακελλαρίου (2020, σ.3-4) είναι ότι «*ΤΝ είναι ο τομέας των Υπολογιστών που ασχολείται με τη σχεδίαση και την υλοποίηση υπολογιστικών συστημάτων τα οποία είναι ικανά να μιμηθούν γνωστικές ικανότητες, εμφανίζοντας έτσι χαρακτηριστικά που αποδίδουμε συνήθως σε ανθρώπινη συμπεριφορά, όπως για παράδειγμα η επίλυση προβλημάτων, η αντίληψη και κατανόηση εικόνων, η μάθηση, η εξαγωγή συμπερασμάτων, η κατανόηση φυσικής γλώσσας, κ.λπ.*».

Η ΤΝ έχει δημιουργήσει και έχει εξελίξει ένα πλήθος εφαρμογών όπως η ρομποτική, η μηχανική όραση, η αυτόματη μετάφραση κ.α.. Έχει στρέψει το βλέμμα της σε καθημερινά προβλήματα και σε χρήσιμες κατευθύνσεις, όπως στην αναγνώριση εικόνων, στην αυτόματη οδήγηση και εν γένει στην επίλυση προβλημάτων που αφορούν εκατομμύρια ανθρώπους (Βλαχάβας κ.ά., 2020, σ.19). Τα πιο εντυπωσιακά επιτεύγματά της αφορούν το πεδίο της μηχανικής μάθησης και τα νευρωνικά δίκτυα (Neural Network ΝΝ).

Η μάθηση αποτελεί τη διαδικασία βελτίωσης ενός συστήματος σε μια ορισμένη εργασία εφόσον υφίσταται η παρατήρηση πολλών παραδειγμάτων. Ο άνθρωπος εκούσια ή ακούσια δύναται να εφαρμόσει διάφορες διαδικασίες μάθησης σε όλη τη διάρκεια της ζωής του. Για παράδειγμα, από την νηπιακή του ηλικία μαθαίνει τη μητρική του γλώσσα από τις ομιλίες των γονιών του. Επίσης, μπορεί να διαβάσει ένα χειρόγραφο κείμενο ακόμη και αν δεν έχει έρθει σε επαφή με τον γραφικό χαρακτήρα του γράφοντα. Η μηχανική μάθηση (ΜΜ) αποτελεί έναν τομέα της ΤΝ που ασχολείται με την ανάπτυξη αλγορίθμων οι οποίοι βελτιώνουν την επίδοση ενός συστήματος σε διάφορα προβλήματα που ανακύπτουν, όπως τα παραπάνω (Διαμαντάρας & Μπότσης, 2019, σ.18-19).

Οι εφαρμογές της μηχανικής μάθησης βρίσκουν ανταπόκριση σε ένα ευρύ φάσμα πεδίων, όπως στην υγεία, την ενέργεια, την επικοινωνία, την αναγνώριση εικόνας κ.α.. Μάλιστα, η αναγνώριση εικόνας αποτελεί μία από τις πρώτες εφαρμογές μοντέλων μηχανικής μάθησης. Διαιρείται σε διάφορες υποκατηγορίες, όπως η αναγνώριση ασθενειών από ιατρικές εικόνες, η αναγνώριση χειρόγραφου ή τυπωμένου κειμένου και πολλές άλλες. Αξίζει να σημειωθεί ότι τα μοντέλα βαθιάς μάθησης (DNN), τα οποία λαμβάνουν ως είσοδο εικόνες που δεν έχουν υποστεί ιδιαίτερη προεπεξεργασία, έχουν εξελιχθεί σε εξαιρετικά εργαλεία αναγνώρισης εικόνας. Εφαρμόζονται δε, σε ένα μεγάλο φάσμα περιπτώσεων και παρουσιάζουν σημαντική πρόοδο (Διαμαντάρας & Μπότσης, 2019, σ.29).

Τα τεχνητά νευρωνικά δίκτυα (NN) είναι εμπνευσμένα από τη δομή και τη λειτουργία του εγκεφάλου. Ο εγκέφαλος αποτελεί ένα εξαιρετικά πολύπλοκο σύστημα με δυνατότητα να οργανώνει τα δομικά του στοιχεία με τους λεγόμενους νευρώνες. Είναι δομημένοι ώστε να εκτελούν συγκεκριμένους υπολογισμούς και μάλιστα με ταχύτητα μεγαλύτερη από τον γρηγορότερο ψηφιακό υπολογιστή που διατίθεται σήμερα. Στην ουσία ένα NN είναι μία μηχανή σχεδιασμένη ώστε να προσομοιάζει με τον τρόπο που εκτελεί μία εργασία ή λειτουργία ο ανθρώπινος εγκέφαλος. Σύμφωνα με τον Haykin (2010, σ.1-2) *«ένα νευρωνικό δίκτυο είναι ένας τεράστιος παράλληλος επεξεργαστής με κατανεμημένη αρχιτεκτονική, ο οποίος αποτελείται από απλές μονάδες επεξεργασίας και έχει από τη φύση του τη δυνατότητα να αποθηκεύει εμπειρική γνώση και να την καθιστά διαθέσιμη για χρήση»*.

Οι πρώτες απόπειρες για την επίλυση αναγνώρισης χειρόγραφου κειμένου έγιναν με τη χρήση μεθόδων Μηχανικής Μάθησης και συγκεκριμένα με τα Hidden Markov Models (HMM). Ωστόσο, οι δυνατότητες τους αποδείχθηκαν αρκετά περιορισμένες καθώς βασίζονταν ως επί το πλείστον στη χειρωνακτική εργασία για να εξάγουν χαρακτηριστικά, με συνέπεια να περιορίζεται η ικανότητα εκμάθησης (Rufenacht, 2020).

Τα αρχεία αναμφίβολα διαδραματίζουν καθοριστικό ρόλο στην οικοδόμηση και την πρόοδο της κοινωνίας. Αποτελούν ένα φορέα συσσώρευσης καταγεγραμμένων πληροφοριών, ενός ανθρώπου, μιας κοινότητας, ενός έθνους. Οι παγκόσμιες προσπάθειες ψηφιοποίησης μεγάλης κλίμακας οδήγησαν στο μετασχηματισμό της αρχειακής πρακτικής και των ροών εργασίας. Όλο και περισσότερες αρχειακές συλλογές ψηφιοποιούνται και νέο υλικό υποβάλλεται στα αρχεία. Ωστόσο, η αξιολόγηση και διασφάλιση της ποιότητάς τους αποτελούν το θεμέλιο λίθο των αρχείων. Οι αρχειοθέτες χρειάζονται την υποστήριξη της τεχνολογίας ώστε να χαλιναγωγήσουν τα τεράστια αρχειακά δεδομένα. Τα αρχεία μεταμορφώνονται σε ένα τεράστιο οργανισμό μεγάλων δεδομένων και η Τεχνητή Νοημοσύνη (AI), κυρίως με τη μέθοδο μηχανικής μάθησης, αποτελεί ένα εργαλείο αντιμετώπισης αυτού του μετασχηματισμού. Οι οκτώ αποδεκτές αρχές τήρησης αρχείων (λογοδοσία, διαφάνεια, ακεραιότητα, προστασία, συμμόρφωση, διαθεσιμότητα, διατήρηση και διάθεση) αποτελούν τον οδηγό για την οργάνωση των νέων μορφών ψηφιακών αρχείων. Οι τεχνολογίες τεχνητής νοημοσύνης διαδραματίζουν πολλαπλούς ρόλους σε αυτή τη διαδικασία. Η ΤΝ βοηθά τα αρχεία να ξεφύγουν από τους περιορισμούς των παραδοσιακών αρχειακών μονάδων, διατηρώντας ταυτόχρονα τα δυνατά τους σημεία, όπως παραδείγματος χάρη όσον αφορά την προέλευση ενός αρχείου. Οι τεχνικές της ΤΝ χρησιμοποιούνται ευρέως για την απόκτηση νέων γνώσεων

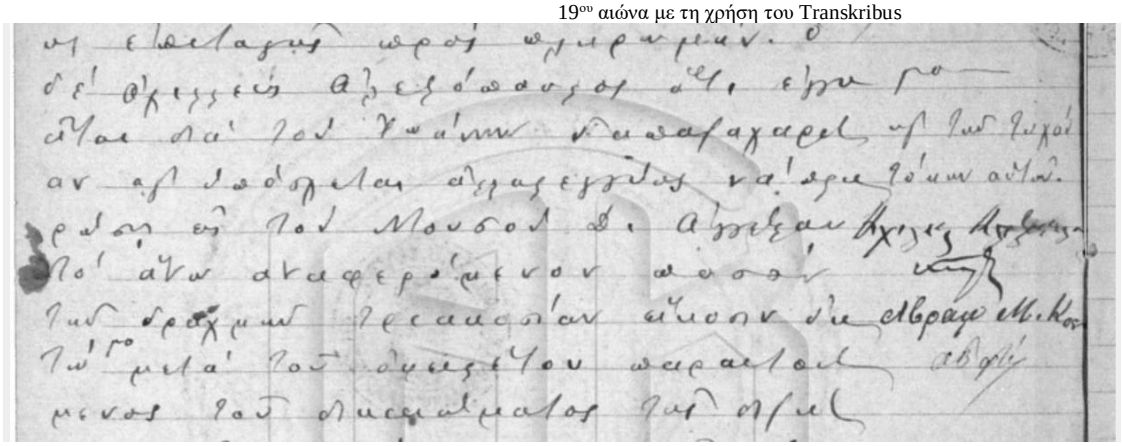
από τις ψηφιακές συλλογές, την πολυφωνική αποδόμηση μέσω καινοτόμων μορφών ανάκτησης, επιτρέποντας παράλληλα στα αρχεία να συνδέονται απευθείας με νέες κοινότητες (Colavizza, Blanke, Jeurgens & Noordegraaf, 2021).

Η μνήμη είναι σύμφυτη με την ανάπτυξη μιας οργανωμένης κοινωνίας και αποτελεί αφετηρία κάθε ανθρώπινης δραστηριότητας. Τα χειρόγραφα αποτελούν ένα μέσο διατήρησης της μνήμης για αυτό και αποτελούν ένα σημαντικό κομμάτι της ανθρώπινης ιστορίας εν γένει. Η ΤΝ με τη χρήση μηχανικής μάθησης (HTR, OCR) για τα χειρόγραφα αποτελεί σημαντικό παράγοντα στην ανάδειξη και στην αξιοποίησή τους. Αρχικά, προφυλάσσει το πρωτότυπο από την φθορά που μπορεί να επέλθει από τη συχνή και ανεξέλεγκτη χρήση του και παράλληλα αυξάνει την ταχύτητα μεταγραφής και την διαθεσιμότητα πρωτογενούς ιστορικού υλικού. Αναδεικνύει το ιστορικό γίγνεσθαι ενός πολιτισμού καθώς δίνει τη δυνατότητα στον εκάστοτε ερευνητή να μπορέσει να κατανοήσει και να αξιοποιήσει την ιστορική αυτή πηγή, ανεξαρτήτως γλώσσας και εθνότητας, δημιουργώντας παράλληλα μια ποικιλόμορφη διεθνή κοινότητα.

1.3. Ερευνητικός στόχος

Τα χειρόγραφα ιστορικά κείμενα αποτελούν αναμφίβολα μία ανεξίτηλη πηγή πληροφοριών για τον ερευνητή. Προκειμένου να διασωθούν και να προστατευτούν αυτά τα ιστορικά μνημεία γνώσης, μεγάλο μέρος τους έχει ψηφιοποιηθεί. Οι ψηφιακές βιβλιοθήκες προσφέρουν πληθώρα ψηφιοποιημένων αρχείων χωρίς ωστόσο να συνυπάρχει και η μεταγραφή τους. Ως απόρροια αυτού, τα ψηφιοποιημένα αρχεία καθίστανται μη προσβάσιμα για μεγάλο μέρος των ερευνητών, αφού έχουν να αντιμετωπίσουν τον μεγάλο σκόπελο της ανάγνωσης και κατανόησής τους. Μπορεί η απρόσκοπτη πρόσβαση στο υλικό να είναι ιδιαίτερα χρήσιμη, ωστόσο η κατανόηση του κειμένου χρήζει γνώσεων παλαιογραφίας, η οποία δεν αποτελεί κτήση των πολλών. Για το λόγο αυτό η μέθοδος αναγνώρισης χειρόγραφων ψηφιοποιημένων αρχείων είναι ένα σημαντικό αίτημα των ερευνητών και της επιστημονικής κοινότητας ευρύτερα.

Η μεταγραφή και η αποκωδικοποίηση (deciphering) των χειρόγραφων κειμένων αποτελεί μία πολύπλοκη και χρονοβόρα διαδικασία. Τα μεθοδολογικά προβλήματα που έχει να αντιμετωπίσει ένας ερευνητής είναι οι ιδιαιτερότητες του γραφέα, κυρίως στην ορθογραφία. Επιπλέον, τα γράμματα δεν είναι στοιχισμένα, είναι ανισομεγέθη, η ένταση του μελανιού ποικίλει και τα διαστήματα ανάμεσα στις λέξεις και στα γράμματα είναι ανισομερή. Παρατηρείται επίσης, από πολλούς γραφείς να γράφουν εκτός των ορίων της σελίδας, ή και να υπάρχουν σημειώσεις σε ακαθόριστα σημεία του εγγράφου (Εικόνα 1).



Εικόνα 1: Κηλίδες μελανιού - Σημειώσεις σε ακαθόριστο σημείο της σελίδας.

(Φίλιος, 1881).

Επίσης, η ύπαρξη του πολυτονισμού, των πνευμάτων, η εναλλαγή μικρογράμματης και μεγαλογράμματης γραφής, οι βραχυγραφίες (abbreviations) οι οποίες στην ελληνική γραφή απαντώνται πολύ συχνά, οι αποκοπές των λέξεων ταλανίζουν τους παλαιογράφους εσαεί (Παπάζογλου, 2009, σ.191,195).

Πέρα όμως από τις δυσκολίες της γραφής και τα χαρακτηριστικά του κάθε γραφέα ανακύπτουν και άλλοι παράγοντες που επηρεάζουν την αποκωδικοποίηση ενός εγγράφου. Τα ιστορικά έγγραφα παρουσιάζουν και υλικές φθορές εξαιτίας της παλαιότητάς τους, όπως κηλίδες από υγρασία και αλλαγή χρώματος. Συχνό φαινόμενο επίσης, αποτελεί η απορρόφηση μελανιού από το χαρτί με αποτέλεσμα να διακρίνεται η πίσω γραμμένη πλευρά της σελίδας. Τέλος, τα σκισίματα και τα τσακίσματα μιας σελίδας δημιουργούν ένα περαιτέρω εμπόδιο για την ευεπίλυτη προσβασιμότητα του.

Νέες τεχνολογίες αναπτύσσονται με σκοπό να δώσουν λύσεις στα παραπάνω προβλήματα και να βοηθήσουν ερευνητές και αρχειονόμους στο δύσκολο έργο τους. Τα λογισμικά ανοιχτού κώδικα που αναπτύσσουν εργαλεία μεταγραφής και αναγνώρισης χειρόγραφου κειμένου (HTR), στοχεύουν στη διάσωση της ιστορικής και πολιτιστικής κληρονομιάς, παρέχοντας αυτόματες μεταγραφές ιστορικών τεκμηρίων σε ένα ευρύτερο κοινό. Κατά πόσο όμως αυτές οι τεχνολογίες μπορούν να εφαρμοστούν σε ελληνικά χειρόγραφα αρχεία δίνοντας αξιόπιστο μεταγραμμένο κείμενο; Ως εκ τούτου οι στόχοι της παρούσας εργασίας είναι:

(α) να διερευνηθούν οι τεχνολογίες που μπορούν να βοηθήσουν το έργο της ψηφιακής τεκμηρίωσης χειρόγραφων αρχείων,

(β) να μελετηθεί η αποτελεσματικότητά τους σε ελληνικά χειρόγραφα αρχεία.

Μέσα από τους στόχους αυτούς ωστόσο, αναφύονται διάφορα ερευνητικά ερωτήματα όπως:

Α)Ποιες είναι οι αυτές οι τεχνολογίες και πως μπορούν να ενταχθούν στο πλαίσιο της ψηφιακής τεκμηρίωσης χειρόγραφων αρχείων;

Β)Κατά πόσο είναι αποτελεσματικές.

Γ)Υπάρχουν εργαλεία που μπορούν να χρησιμοποιηθούν για την ψηφιακή τεκμηρίωση ελληνικών χειρογράφων;

1.4 Μεθοδολογική προσέγγιση

Οι σύγχρονοι οργανισμοί και φορείς, όπως αρχεία και βιβλιοθήκες, έρχονται αντιμέτωποι με τη δυσκολία ανάγνωσης και κατανόησης του υλικού των συλλογών τους που καλούνται να ταξινομήσουν, να ψηφιοποιήσουν και να διαθέσουν στους ερευνητές. Ως εκ τούτου, κύριο μέλημά τους αποτελεί η εξεύρεση μεθόδων και πρακτικών που μπορούν να αποφέρουν ποιοτικές υπηρεσίες στο κοινό τους.

Με την πάροδο των χρόνων, η χρήση μοντέλων αυτόματης μεταγραφής και αναγνώρισης χειρόγραφου κειμένου διαδίδεται όλο και περισσότερο στον ερευνητικό τομέα πλειάδας επιστημών, καθώς αρκετοί είναι οι ερευνητές που για τη διεξαγωγή της έρευνάς τους, έχουν στηριχθεί σε διάφορα λογισμικά περιβάλλοντα. Σκοπός της παρούσας έρευνας αποτελεί η διερεύνηση των τεχνολογιών αυτών που ευελπιστούν να δώσουν λύσεις στα προβλήματα που προαναφέρθηκαν και να διευκολύνουν το έργο των επαγγελματιών της πληροφόρησης. Στο πλαίσιο του πονήματος αυτού, θα πραγματοποιηθεί έρευνα στο διαδίκτυο, σε βιβλιογραφικές πηγές και ηλεκτρονικά περιοδικά, ώστε να καταγραφούν και να προσδιοριστούν κάποιες από τις δυνατότητες των διαθέσιμων ψηφιακών εργαλείων μεταγραφής και αναγνώρισης χειρόγραφου κειμένου. Επίσης, θα μελετηθεί η περίπτωση της πλατφόρμας Transkribus, με τεχνολογία HTR, προκειμένου να προσδιοριστούν οι δυνατότητες και η αποτελεσματικότητά της στην αυτόματη μεταγραφή ελληνικών χειρόγραφων κειμένων.

Για την υλοποίηση του σκοπού της εργασίας, θα δημιουργηθούν μοντέλα Al, τα οποία θα εκπαιδευτούν στην ανάγνωση ελληνικού χειρόγραφου κειμένου με τη μεταγραφή και χειροκίνητη διόρθωση μέρους του αρχείου του συμβολαιογράφου Λαρίσης Αναστασίου Φίλιου. Το αρχείο αυτό περιέχει χειρόγραφο καλλιγραφικό κείμενο του 19ου αιώνα, σημειώσεις, βραχυγραφίες και αριθμούς. Η αποτελεσματικότητά του θα δοκιμαστεί υπό αυτές τις προκλήσεις. Αρχικά, θα μεταγραφούν κείμενα με μη αυτοματοποιημένο τρόπο που θα λειτουργήσουν ως ρετσέτα για την εκπαίδευση των μοντέλων. Στη συνέχεια θα πραγματοποιηθεί η εκπαίδευση και θα αποδοθούν οι αυτόματες μεταγραφές των κειμένων. Έπειτα, θα δοκιμαστούν στα μεταγραμμένα κείμενα κάποια από τα εργαλεία ανάλυσης και αναζήτησης κειμένου. Τέλος, θα εξαχθούν συμπεράσματα για την αξιοπιστία και την αποτελεσματικότητα της συγκεκριμένης πλατφόρμας ως προς την εφαρμογή σε ελληνικό χειρόγραφο κείμενο.

2. Ανασκόπηση τεχνολογιών για αναγνώριση χειρογράφου

Η ανάγκη βιωσιμότητας των ιστορικών εγγράφων, η εξασφάλιση της δυνατότητας ανάκτησης και απρόσκοπτης χρήσης τους, δεν αποτελεί μόνο βασικό ζητούμενο αλλά και μια σύγχρονη πρόκληση. Η εισαγωγή των τεχνολογιών της Πληροφορικής στις Ανθρωπιστικές Σπουδές είχε ως απόρροια την υιοθέτηση νέων μορφών δημιουργίας εργαλείων έρευνας που αποσκοπούν στο σημασιολογικό περιεχόμενο των χειρόγραφων, δηλαδή στην ίδια τη γνώση που εμπεριέχουν οι πηγές αυτές.

2.1 Τεχνολογία Οπτικής Αναγνώρισης Χειρογράφων (OCR)

Η Οπτική Αναγνώριση Χειρογράφων (OCR) αποτελεί ένα από τα εργαλεία εξαγωγής δεδομένων εικόνας για σάρωση παλαιών χειρογράφων και βιβλίων, μιας αίτησης δανείου κ.α.. Ως εκ τούτου κατανοούμε ότι οι εικόνες εγγράφων αποτελούν μια από τις πιο εύχρηστες πηγές δεδομένων. Η εξαγωγή κειμένου από αυτές εξαρτάται από το OCR. Το OCR έχει εφαρμοστεί σε έγγραφα χρησιμοποιώντας Python με σκοπό το υψηλό επίπεδο ακρίβειας. Ωστόσο, όλα αυτά τα μοντέλα έχουν ένα μειονέκτημα, είτε έχουν ένα καλό μοντέλο αναγνώρισης, είτε ένα καλό μοντέλο ανίχνευσης. Δηλαδή ορισμένα μοντέλα είναι γρήγορα στην αναγνώριση κειμένου, αλλά πιο αργά στην ανίχνευση και το αντίθετο (Sridhar, 2022).

Η καινοτομία της τεχνολογίας OCR εντυπωσιάζει καθώς εφαρμόστηκε για πρώτη φορά το 1928. Ο Gustav Tauschek από την Αυστρία δημιουργεί ένα παρόμοιο σύστημα με τη χρήση φωτοκυττάρων. Τα σημερινά ανεπτυγμένα προγράμματα OCR έχουν μεταξύ τους μικρές διαφορές, γενικά όμως έχουν την δυνατότητα να επεξεργάζονται την εικόνα κάθε σελίδας και να αναγνωρίζουν το κείμενο. είτε γραμμή προς γραμμή, είτε χαρακτήρα προς χαρακτήρα, είτε λέξη προς λέξη. Το OCR στηρίζεται στη μηχανική μάθηση δίνοντας τη δυνατότητα στο φυσικό κείμενο να μετατραπεί σε κείμενο μηχανικής εκμάθησης χρησιμοποιώντας ένα σαρωτή για την επεξεργασία του φυσικού εγγράφου. Το λογισμικό προσπαθεί να καθορίσει τη γραμμή βάσης για κάθε γραμμή κειμένου στην εικόνα. Μόλις σκιαγραφηθούν οι δυνητικοί χαρακτήρες από το λογισμικό μπορεί να χρησιμοποιήσει δύο τεχνικές για την εξαγωγή αποτελεσμάτων: την αναγνώριση μοτίβου (κάθε διακριτικό συγκρίνεται σε pixel-to-pixel) και την εξαγωγή χαρακτηριστικών (κάθε διακριτικό συγκρίνεται με διαφορετικούς κανόνες που περιγράφουν το είδος του χαρακτήρα) (Woodford, 2018).

Το λογισμικό OCR έγινε γνωστό από την ABBYY το 1993, παραταύτα η αποτελεσματικότητά του σε σύνθετα σενάρια, όπως το χειρόγραφο ήταν χαμηλή. Σήμερα η ABBYY με τη χρήση της τεχνητής νοημοσύνης έχει δημιουργήσει το ABBYY Blue Prism, το οποίο αποτελεί μία ευφυή επεξεργασία εγγράφων (IDP) , ωστόσο, περιορίζεται στον επαγγελματικό τομέα, όπως επεξεργασία τιμολογίων , αποδείξεων κ.λπ., εξάγοντας βασικά δεδομένα από τα κείμενα αυτά, σε πολλές διαφορετικές γλώσσες (https://www.abbyy.com/).

Η βελτίωση του OCR επήλθε μέσω των μοντέλων HMMs (Hidden Markov Models) τα οποία χρησιμοποιούν εργαλεία μοντελοποίησης που χρησιμοποιήθηκαν αρχικά στην αναγνώριση ομιλίας. Τα τελευταία χρόνια έχει ενσωματώσει τεχνικές μηχανικής

μάθησης βελτιώνοντας τα ποσοστά ακρίβειας (Nockels, Gooding, Ames & Terras, 2022).

Το Tesseract αποτελεί μια δημοφιλή μηχανή οπτικής αναγνώρισης χαρακτήρων. Αναπτύχθηκε αρχικά από την Hewlett Packard Laboratories Bristol UK έως το 2006. Από το 2006 υποστηρίζεται από την Google. Έχει τη δυνατότητα να αναγνωρίσει διάφορες μορφές εικόνας όπως PNG, GPEG και TIFF. Το Tesseract δεν διαθέτει ενσωματωμένο GUI (γραφική διεπαφή χρήστη), βασίζεται σε νευρωνικά δίκτυα και εστιάζει στην αναγνώριση γραμμής. Επίσης, μπορεί να αναγνωρίσει περισσότερες από 100 γλώσσες, και ενώ η προηγούμενη έκδοση παρουσίασε αισθητή βελτίωση με την χρήση νευρωνικών δικτύων, το Tesseract 5 εκπαιδεύεται και αναγνωρίζει ταχύτερα μέσω των floats, ωστόσο για να εξάγει καλύτερα αποτελέσματα, πρέπει να υπάρχει καλή ποιότητα εικόνας (https://github.com/).

Το Kraken είναι ένα πρόγραμμα ανοικτού κώδικα OCR, το οποίο αναπτύχθηκε από την Apache 2.0. Το συγκεκριμένο πρόγραμμα εκτελείται μόνο σε περιβάλλον Linux και Mac OS X και όχι σε Windows. Το Kraken χρησιμοποιείται ως επί το πλείστον για αναγνώριση αραβικών και περσικών χαρακτήρων καθώς μπορεί να δημιουργήσει μοντέλα για δημιουργία μεταγραφών κειμένων γραμμένα από δεξιά προς τα αριστερά και από πάνω προς τα κάτω. Επίσης, υποστηρίζει και μοντέλα για κείμενα γραμμένα σε λατινικές γλώσσες. Ωστόσο, απαιτείται οι εικόνες που χρησιμοποιούνται να είναι υψηλής ευκρίνειας και το κείμενο να εκτίθεται σε μια στήλη (https://digitalorientalist.com/).

Το Calamari είναι πακέτο OCR το οποίο βασίζεται στο ORopy και το Kraken. Το συγκεκριμένο πακέτο ειδικεύεται στη δημιουργία μεταγραφών για πρώιμα έγγραφα, ωστόσο αποδίδει και σε σύγχρονα κείμενα. Δεν εκτελεί ανάλυση διάταξης ή τμηματοποίηση γραμμών και οι εργασίες αυτές θα πρέπει να εκτελούνται χωριστά. Εστιάζει στη μεταγραφή εικόνων γραμμής σε κείμενο. Μπορεί, παρά ταύτα, να ενσωματωθεί με άλλα προγράμματα OCR που βασίζονται σε Python όπως (pyocr, kraken κ.λπ.) για την ολοκλήρωση όλων των σταδίων της διαδικασίας μεταγραφής. OCR (https://pitt.libguides.com/).

Επίσης, το Easy OCR αναπτύχθηκε από τη Jaided Al, μια εταιρεία η οποία ειδικεύεται στην οπτική αναγνώριση χαρακτήρων. Η υλοποίησή του γίνεται με τη χρήση της γλώσσας Python και αναγνωρίζει 42 γλώσσες. Δίνει πιο ακριβή αποτελέσματα σε κείμενα αρχείων pdf με υψηλή ευκρίνεια (Gulati, 2021).

Το Nautilus - OCR είναι μια μηχανή οπτικής αναγνώρισης χαρακτήρων, η οποία αναπτύχθηκε από την Εθνική βιβλιοθήκη του Λουξεμβούργου για τη συλλογή ιστορικών εφημερίδων της. Τα μοντέλα αυτά τα παρήγαγε και τα δημοσίευσε προς δημόσια χρήση. Λειτουργεί με τα σχήματα METS/ALTO, με τη δυνατότητα να λαμβάνει ένα σύνολο δεδομένων αυτών των σχημάτων και να παράγει ένα βελτιωμένο σύνολο δεδομένων METS/ALTO. Το μοντέλο αυτό είναι συμβατό σε περιβάλλον Linux και Mac oS (https://pitt.libguides.com/).

Οι μηχανισμοί OCR έχουν καθιερωθεί σε πολλαπλές εφαρμογές με επιτυχία. Ωστόσο, είναι απαραίτητο να διευκρινιστεί, ότι μόνο στις περιπτώσεις που αφορά έγγραφα υψηλής ευκρίνειας και ποιότητας, η αναγνώριση των χαρακτήρων υπήρξε απρόσκοπτη και επιτυχής. Παραταύτα, σε χειρόγραφα έγγραφα όπου οι λέξεις είναι γραμμένες σε πολυτονικό σύστημα, με βραχυγραφίες, η ποιότητα του χαρτιού είναι χαμηλή, υπάρχουν

ατέλειες στην στοιχειοθεσία και πολλές άλλες αιτίες, αποτελούν όλους εκείνους τους παράγοντες για την χαμηλή απόδοση και την αποτυχημένη αναγνώριση των χαρακτήρων.

2.2 Τεχνολογία Αναγνώρισης χειρόγραφου κειμένου (HTR)

Τα χειρόγραφα κείμενα είναι δύσκολο να προσπελαστούν μέσω των παραδοσιακών προσεγγίσεων του OCR, λόγω του χειρόγραφου στυλ τους ή των ειδικών γραμματοσειρών και πολλών άλλων δυσκολιών. Η μετάβαση από το OCR έγινε στο πιο προηγμένο HTR, το οποίο χρησιμοποιεί προσεγγίσεις μηχανικής μάθησης, όπως βαθιά νευρωνικά δίκτυα με σκοπό την εξαγωγή οπτικών χαρακτηριστικών και την αναγνώριση χαρακτήρων και λέξεων σε μια τμηματοποιημένη γραμμή κειμένου. Το κοινό των δύο τεχνολογιών είναι ότι απαιτείται χειροκίνητη παρέμβαση και εκπαίδευση.

Το HTR (handwriting recognition) αποτελεί μια τεχνολογία η οποία χρησιμοποιείται για να διαβάσει χειρόγραφα κείμενα σε εικόνες. Τα υπολογιστικά συστήματα HTR χρησιμοποιούν προσεγγίσεις μηχανικής μάθησης όπως τα βαθιά νευρωνικά δίκτυα με σκοπό να εξάγουν οπτικά χαρακτηριστικά και να αναγνωρίσουν χαρακτήρες και λέξεις. Το HTR δεν αναγνωρίζει το χειρόγραφο αλλά τη γραφή του χειρογράφου καθιστώντας ικανό να μεταγράφει διαφορετικούς τύπους γραφής. Με λίγα λόγια η τεχνολογία HTR είναι γλωσσικά ανεξάρτητη, η διαδικασία εκπαίδευσης των νευρωνικών δικτύων είναι ίδια, για οποιοδήποτε τύπο αλφαβηταρίου, οποιασδήποτε ημερομηνίας (Muehlberger κ.ά., 2019).

Ο Adam Matthew Digital είναι ο πρώτος και μοναδικός εκδότης που χρησιμοποίησε την τεχνητή νοημοσύνη για αναγνώριση χειρόγραφου κειμένου (HTR) για τις χειρόγραφες συλλογές του, μέσω της πλατφόρμας Quartex με την ανάπτυξη του προγράμματος HTR Transcription. Η συγκεκριμένη πλατφόρμα υποστηρίζει και μεταγραφή οπτικοακουστικού υλικού (https://www.amdigital.co.uk/).

Το Beatrix είναι ένα αυτόνομο σύστημα αυτοεκμάθησης για την αναγνώριση χειρόγραφων κειμένων. Το σύστημα ενσωματώνει τη νευρωνική αναγνώριση με τεχνικές ανάλυσης περιβάλλοντος και το αποτέλεσμα επιτυγχάνεται μέσω μιας διαδικασίας αυτομάθησης που αποτελείται από τέσσερα αλληλεπιδρώντα υποσυστήματα ενισχύοντας κατ' αυτόν τον τρόπο την ικανότητα του συστήματος να αναγνωρίζει το συγκεκριμένο χειρόγραφο, χωρίς ωστόσο να χάνει την ικανότητα του να αναγνωρίζει άλλους τύπους γραφής (Lazzerini, Marcelloni, F., & Reyneri, 1997).

Πολλά προγράμματα έχουν αναπτύξει τις δικές τους εξατομικευμένες λύσεις HTR. Ένα εργαλείο αποτελεί το Monk, το οποίο αναπτύχθηκε από το Πανεπιστήμιο Groningen, από μια ερευνητική ομάδα του ινστιτούτου τεχνητής νοημοσύνης ALICE. Το Monk παρέχει ένα μέσο συνεχούς εκπαίδευσης αναγνώρισης γραφής και δημιουργεί ευρετήρια χειρόγραφων συλλογών. Έχει επεξεργαστεί με επιτυχία διάφορα είδη γραφής, όπως ιερογλυφικά, ιερατική γραφή σε πάπυρο, κινέζικους χαρακτήρες χειρόγραφους και τυπωμένους σε ξύλο και πολλά άλλα (University of Groningen, [χ.χ.]).

Στην Ελλάδα το μDOC.ts είναι μια τεχνολογία που βασίζεται στην Τεχνητή Νοημοσύνη, αποσκοπώντας στην αυτόματη εξαγωγή κειμένου από ψηφιοποιημένα ιστορικά έγγραφα. Αναπτύχθηκε από το Δημοκρίτειο Πανεπιστήμιο Θράκης με τη σύμπραξη του Ερευνητικού Κέντρου Αθηνά και των εταιρειών Omega Technology και Prisma

Electronics ABEE το 2018. Η αυτόματη εξαγωγή μοντέλου επιτυγχάνεται μέσω HTR χρησιμοποιώντας γλωσσικά μοντέλα και τεχνικές εντοπισμού λέξεων (KWS). Μάλιστα, τα αποτελέσματα από την μέχρι τώρα εφαρμογή του ήταν εντυπωσιακά καθώς αγγίζουν μόλις το 5% σε επίπεδο σφάλματος. Επιπροσθέτως, το σύστημα προσαρμόζεται σε οποιαδήποτε μορφή γραφής (Δημοκρίτειο Πανεπιστήμιο Θράκης. Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, ([χ.χ])).

Το Transkribus είναι μια πλατφόρμα για την αναγνώριση χειρόγραφου κειμένου και αναζήτηση ιστορικών εγγράφων με χρήση τεχνητής νοημοσύνης σε οποιαδήποτε γλώσσα. Δημιουργήθηκε από την ΕΕ transcriptorium το 2013 - 2015. Από το 2019 ιδρύθηκε η READ-COOP SCE για τη διατήρηση και περαιτέρω ανάπτυξή της πλατφόρμας. Το Transkribus αναπτύχθηκε στο πλαίσιο του προγράμματος "READ" της ΕΕ Horizon 2020 από μια κοινοπραξία κορυφαίων ερευνητικών ομάδων, από όλη την Ευρώπη, με επικεφαλής το Πανεπιστήμιο του Ίνσμπρουκ (https://readcoop.eu/)

Το Transkribus αποτελεί την πιο δημοφιλή πλατφόρμα για την παραγωγή μεταγραφών ιστορικών κειμένων σε όλους τους κλάδους πολιτισμού. Επιτρέποντας την αυτοματοποιημένη αναγνώριση και μεταγραφή ιστορικών κειμένων καθιστά το υλικό τους ευανάγνωστο, με αποτέλεσμα να διευρύνεται η πρόσβαση σε συλλογές και ταυτόχρονα να επεκτείνονται οι δυνατότητες κατανόησης των κειμένων.

Επίσης, το in Codice Ratio αποτελεί ένα υβριδικό ερευνητικό έργο, σε πειραματικό στάδιο, συνδυάζοντας τις τεχνολογίες OCR και HTR, επιλέγοντας κατά το δοκούν, τα πλεονεκτήματα που προσφέρει η κάθε τεχνολογία βασιζόμενοι στην τμηματοποίηση χαρακτήρων. Στόχος τους είναι να περιορίσουν την εκτεταμένη εκπαίδευση και να προσφέρουν αυτοματοποιημένες λύσεις που θα εξυπηρετούν κάθε ερευνητή (Ammirati κ.ά., 2017).

Η βελτίωση και η αύξηση πρόσβασης στις συλλογές, μέσω της συνεχούς ανάπτυξης της τεχνολογίας HTR, θα ανοίξει το δρόμο στους ερευνητές-χρήστες, παρέχοντας τους την δυνατότητα να εντοπίζουν άμεσα και τελέσφορα τα ζητήματα που τους απασχολούν. Ταυτόχρονα, μέσω της τεχνολογίας αυτής δίνεται μια πρόσθετη παροχή δεδομένων, οι χρήστες μπορούν να απομονώσουν λέξεις, ονόματα, περιοχές, προσθέτοντας με αυτό τον τρόπο, βάθος και εύρος στην έρευνά τους.

Στη συνέχεια παρατίθεται συγκριτικός πίνακας (Πίνακας 1) με τα πλεονεκτήματα και τις δυνατότητες των προαναφερθέντων τεχνολογιών.

Πίνακας 1: Συγκριτικός πίνακας

	Προσαρμογή σε πολλά είδη γραφής	Αναγνώριση πλήθος γλωσσών	Ανοικτός κώδικας	OCR	HTR	NT	Δεν απαιτεί υψηλή ευκρίνεια εικόνας	Μεταγραφή οπτικοακουστικού υλικού	πραγματοποιεί ανάλυση κειμένου	Αυτόματη μεταγραφή	Μη αυτόματη μεταγραφή	Εργαλεία αναζήτησης WEB
ABBYY Blue Prism		✓		✓		✓				✓		✓
Tesseract 5		✓	✓	✓		✓				✓		
Kraken			✓	✓					✓	✓	✓	
Calamari		✓	✓	✓					✓	✓		
Easy OCR		✓		✓						✓		
Nauticus OCR		✓		✓						✓		
Quartex					✓	✓	✓	✓	✓	✓	✓	✓
Beatrix	✓				✓	✓	✓		✓	✓	✓	
Monk	✓				✓	✓	✓			✓	✓	✓
μDOC.ts	✓				✓	✓			✓	✓		✓
Transkribus	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓
Codice Ratio	✓			✓	✓	✓	✓		✓	✓	✓	

Παρατηρώντας τον πίνακα γίνεται κατανοητό ότι οι τεχνολογίες OCR έχουν την δυνατότητα να μετατρέψουν σαρωμένες εικόνες χειρόγραφων ή έντυπων κειμένων, σε κείμενο αναγνώσιμο από τον ηλεκτρονικό υπολογιστή. Ωστόσο, η απόδοσή τους σε χειρόγραφα, εξαιτίας της ποικιλομορφίας της γραφής και όλων των δυσμενών παραγόντων που έχουν αναφερθεί, δεν παράγουν ευανάγνωστα αποτελέσματα και δεν ανταποκρίνονται ικανοποιητικά. Επίσης, έχουν την δυνατότητα αναγνώρισης πλήθους γλωσσών, όμως απαιτείται υψηλή ευκρίνεια εικόνας και δεν παρέχουν εργαλεία αναζήτησης. Οι τεχνολογίες HTR ανταποκρίνονται περισσότερο ικανοποιητικά στη μεταγραφή χειρόγραφων κειμένων. Όλα τα αναφερθέντα λογισμικά HTR χρησιμοποιούν τη ΤΝ με χρήση της μηχανικής μάθησης και των νευρωνικών δικτύων. Παραταύτα, η πλατφόρμα Transkribus ξεχωρίζει καθώς έχει τη δυνατότητα να ανταποκριθεί σε όλα τα ζητούμενα με επιτυχία. Μπορεί να προσαρμοστεί σε πολλά είδη γραφής και να ανταποκριθεί επαρκώς στην αναγνώριση πλήθους γλωσσών. Επίσης, δεν απαιτεί υψηλή ευκρίνεια εικόνας, πραγματοποιεί ανάλυση κειμένου και αυτόματη μεταγραφή του. Διαθέτει εργαλεία αναζήτησης (KWS) και πρόσβαση στο Web. Ωστόσο, και τα υπόλοιπα λογισμικά HTR μπορούν να ανταποκριθούν επαρκώς σε ένα ευρύ πεδίο αυτών

των ζητουμένων, όμως επειδή αποτελούν ιδιωτικές πρωτοβουλίες ή έχουν πειραματιστεί σε συγκεκριμένη μόνο γλώσσα καθώς επίσης, δεν είναι διαθέσιμα για ένα ευρύτερο κοινό αποτελούν εμπόδιο στην περαιτέρω αξιολόγησή τους. Το Transkribus είναι το μοναδικό λογισμικό το οποίο έχει χρησιμοποιηθεί από μία πλειάδα οργανισμών, ιδρυμάτων, φορέων αλλά και ιδιωτών με μεγάλα ποσοστά επιτυχίας η οποία αναφέρεται λεπτομερώς παρακάτω.

2.3 Η σημασία ανάπτυξης τεχνολογιών για τις ανθρωπιστικές επιστήμες

Τα χειρόγραφα τεκμήρια αποτελούν γνήσια στοιχεία της ιστορίας ενός έθνους. Δίνουν την δυνατότητα επικοινωνίας με το παρελθόν, τόσο στο χώρο, όσο και στο χρόνο. Στα χειρόγραφα διαγράφεται η εξέλιξη ενός λαού μέσα από τις ποικίλες δραστηριότητές του. Η αξία τους είναι διηνεκής και παραμένει ακέραιη στο εσωτερικό κάθε πολιτισμού. Αποτελούν εν γένει τα νήματα του κοινωνικού ιστού της ανθρώπινης ιστορίας και πολιτισμού (Langlois, 1989).

Η χρήση των τεχνολογιών της επιστήμης της Πληροφορικής διαμορφώνει ένα καινοτόμο περιβάλλον διατήρησης της ανθρώπινης δραστηριότητας. Η Πληροφορική αποτελεί παράγοντα καθοριστικής σημασίας και αρωγό για τη δημιουργία, αποθήκευση, καταγραφή και πρόσβαση στην πληροφορία. Η υιοθέτηση και η χρήση νέων τεχνολογιών συμβάλλουν σημαντικά στη διαχείριση του περιεχομένου των ιστορικών χειρόγραφων τεκμηρίων, δηλαδή στη διαχείριση και στην κατανόηση της γνώσης που εμπεριέχονται σε αυτά.

Τα τελευταία χρόνια είμαστε μάρτυρες της σφοδρής ανάπτυξης του διαδικτύου, χωρίς ωστόσο να λαμβάνεται υπόψη η διάσταση του χρόνου, καθώς το διαδίκτυο μας προσφέρει το παρόν, συγκεντρωμένο και επίκαιρο. Αυτό σαφέστατα ενέχει μια αρνητική διάσταση, διότι επίκειται να αποτελέσει τα επόμενα χρόνια τη μοναδική πηγή καταγραφής και τεκμηρίωσης πηγών του παρελθόντος άρα είναι πρόδηλο ότι δεν δύναται να αγνοηθεί η χρονική διάσταση (Παπαλεξίου, 2021, σ.350).

Στην προσπάθεια αυτή να επιτευχθεί η χωροχρονική μετατόπιση του διαδικτύου και η διάσωση των ιστορικών πηγών μας, όσον αφορά την Ευρώπη, η ψηφιοποίηση αυτών και η χρήση εξελιγμένων τεχνολογιών εντοπίζεται σε πολλά ερευνητικά κέντρα, οργανισμούς, διαδικτυακά περιβάλλοντα και προγράμματα.

Το Nodegoat αποτελεί ένα διαδικτυακό ερευνητικό περιβάλλον διαμορφωμένο ώστε να εξυπηρετεί τη διάσωση των ιστορικών πηγών μας. Δίνει μάλιστα τη δυνατότητα, τα δεδομένα μεταγραφής του Transkribus, να μπορούν να συνδεθούν με το περιβάλλον του. Το Nodegoat είναι ένα λογισμικό ανοιχτού κώδικα που επιτρέπει στους μελετητές να μοντελοποιούν, να δημιουργούν, να αναλύουν και να οπτικοποιούν σύνολα δεδομένων τα οποία αποδίδονται χωρικά και χρονικά. Στην ουσία αποτελεί μια ιστορική σχεσιακή βάση δεδομένων (https://nodegoat.net/).

Η Europeana είναι μια παν-Ευρωπαϊκή ψηφιακή βιβλιοθήκη με στόχο να τροφοδοτεί, να εμπλουτίζει και να υποστηρίζει την πολιτιστική κληρονομιά στον ψηφιακό κόσμο. Η λειτουργία της ξεκίνησε στις 20 Νοεμβρίου 2008 και από τότε έχει συνεισφέρει και έχει συνεργαστεί με ποικίλα καινοτόμα τεχνολογικά εργαλεία που βελτιώνουν και εμπλουτίζουν τα μεταδεδομένα της. Η Europeana έχει αναπτύξει την πλατφόρμα Transcribathon για μεταγραφές πληθοπορισμού με το εργαλείο αυτόματης μεταγραφής

Transkribus, ενδυναμώνοντας τον τομέα της πολιτιστικής κληρονομιάς στον ψηφιακό μετασχηματισμό της (https://www.europeana.eu/el).

Το Voyant Tools αποτελεί ένα διαδικτυακό περιβάλλον ανάλυσης και ανάγνωσης ψηφιακών κειμένων ανοιχτού κώδικα. Αναπτύχθηκε από μια ομάδα μελετητών των ψηφιακών ανθρωπιστικών επιστημών τους Stefan Sinclair και Geoffrey Rockwell. Απευθύνεται σε πλειάδα χρηστών, όπως φοιτητές, ερευνητές και μελετητές ψηφιακών ανθρωπιστικών επιστημών προκειμένου να διευκολύνει τις αναγνωστικές και ερμηνευτικές πρακτικές. Περιλαμβάνει μία ποικιλία εργαλείων, που διευκολύνουν την ερμηνεία και τη μελέτη κειμένων ακόμη και για ένα αρχάριο χρήστη. Κύριο χαρακτηριστικό των εργαλείων αυτών είναι η διαδραστικότητά τους και η αλληλεπίδρασή τους (https://voyant-tools.org/).

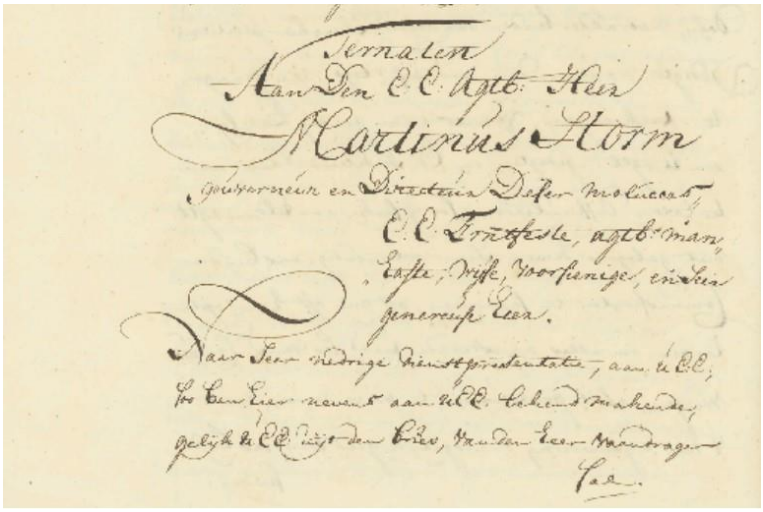
Υπάρχει επίσης μια έκδοση του Voyant (VoyantServer Desktop), χωρίς σύνδεση στο διαδίκτυο, η οποία είναι εύκολη στη λήψη της στην επιφάνεια εργασίας του υπολογιστή και εξασφαλίζει πιο γρήγορη, πιο ιδιωτική και πιο ευέλικτη απόδοση. Οι δημιουργοί μάλιστα της πλατφόρμας ενθαρρύνουν την χρήση της καθώς όπως υποστηρίζουν τα πράγματα μπορεί να γίνουν απρόβλεπτα, στη φιλοξενούμενη έκδοση, αν πατήσουν ταυτόχρονα το ίδιο κουμπί 30 άτομα (https://voyant-tools.org/).

Το Greenstone αποτελεί ένα διαδικτυακό περιβάλλον με σκοπό την ανάπτυξη και τη διανομή συλλογών για τη δημιουργία ψηφιακών βιβλιοθηκών από δημόσιους φορείς και ιδρύματα αλλά και από ιδιωτικές πρωτοβουλίες. Αποτελεί ένα λογισμικό ανοιχτού κώδικα, πολύγλωσσο, το οποίο αναπτύχθηκε από τη New Zealand Digital Library Project του Πανεπιστημίου Waikato σε συνεργασία με την Unesco και τη MKO Human Info. Η διάδοση εκπαιδευτικών και πολιτιστικών πληροφοριών παγκοσμίως αποτελεί το κύριο μέλημα του, δημιουργώντας ένα χρήσιμο εργαλείο για την απρόσκοπτη επαφή με την ψηφιακή γνώση. Το Greenstone χρησιμοποιήθηκε επίσης, ως βοηθητικό εργαλείο του πονήματος, για μια καλύτερη οπτική προσέγγιση του συμβολαιογραφικού αρχείου του οποίου έγινε χρήση (https://www.greenstone.org/).

2.4 Η κοινότητα χρηστών του Transkribus

Από το 2015, όπου έγινε διαθέσιμη η πλατφόρμα Transkribus περισσότερα από 100 ιδρύματα και προγράμματα καθώς και πολλές βιβλιοθήκες και αρχεία έχουν υπογράψει σύμβαση συνεργασίας με την READ-coop. Γίνεται αντιληπτό ότι τα αποτελέσματα της αυτοματοποιημένης μεταγραφής και αναζήτησης επιταχύνουν τις συνδέσεις με το ιστορικό υλικό και παρέχουν τη δυνατότητα να αναδιαμορφώσουν την ερευνητική πρακτική (Muehlberger κ.ά., 2019).

Τα Εθνικά αρχεία της Ολλανδίας ξεκίνησαν πριν 5 χρόνια, μία φιλόδοξη στρατηγική ψηφιοποίησης, με αρχικό σκοπό στα επόμενα 15 χρόνια να ψηφιοποιήσουν το 10% των αρχείων τους χρησιμοποιώντας την τεχνολογία αναγνώρισης χειρόγραφου κειμένου (HTR) για αυτόματη μεταγραφή και μετατροπή σε ψηφιακό αρχείο κειμένου (Εικόνα 2). Δημιούργησαν 6.000 σελίδες δεδομένων εκπαίδευσης με ποσοστό λάθους 7%. Ανέπτυξαν το δικό τους μοντέλο Al το Dutch Handwriting με Transkribus, όπου περιλαμβάνει σαρώσεις από τον 17ο έως τον 19ο αιώνα, το οποίο μπορεί να χρησιμοποιηθεί από οποιονδήποτε χρήστη με παρόμοια έγγραφα (https://readcoop.eu/).



Εικόνα 2: Χειρόγραφο από τα Εθνικά Αρχεία της Ολλανδίας.

(https://readcoop.eu/)

Το ίδρυμα Fondazione Banco di Napoli διατηρεί το μεγαλύτερο αρχείο ιστορικοοικονομικής τεκμηρίωσης στον κόσμο. Το 2018 με το όνομα Pandetta Project ξεκίνησε τη χρήση του Transkribus. Μέχρι στιγμής έχουν ψηφιοποιηθεί και μεταγραφεί εξήντα πανέτες, για τις τράπεζες που δραστηριοποιήθηκαν από το 1586 έως το 1599, με ποσοστό σφαλμάτων μόλις 5%. Από το 2019 ο φορέας έγινε ιδρυτικό μέλος του συνεταιρισμού READ (http://www.fondazionebanconapoli.it/en/).

Το Ινστιτούτο Πολωνικής γλώσσας εργάζεται σε ένα ψηφιακό σώμα πολωνικών κειμένων - διαφορετικών ειδών, στυλ και γλώσσας - του 17ου και 18ου αιώνα, με την ονομασία KorBa. Ανέπτυξαν δύο διαφορετικά μοντέλα τόσο για έντυπα κείμενα, με ποσοστό λάθους 0,29%, όσο και για χειρόγραφα με ποσοστό λάθους 1,8%). Παρόλο που τα χειρόγραφα κείμενα περιείχαν διαφορετικές γλώσσες στην ίδια σελίδα, όπως γερμανικά, λατινικά ή γαλλικά, παραταύτα το Transkribus ανταποκρίθηκε με επιτυχία (https://readcoop.eu/).

Το Γερμανικό Ίδρυμα Ερευνών (DFG) σε συνεργασία με το Freie Universitat Berlin του τμήματος Μεσαιωνικής Ιστορίας, έθεσαν ως κεντρικό στόχο την "Χαρτογράφηση της Μεσαιωνικής Βιέννης" που βασίζεται σε κείμενα του βιεννέζικου κτηματολογίου του 15ου αιώνα. Μεταγράφοντας αρχικά περίπου 30 φύλλα από ένα από τα κτηματολόγια και εκπαιδεύοντας το λογισμικό με νέο υλικό, το ποσοστό σφάλματος χαρακτήρων (CER) είναι κάτω από το όριο του 2% (https://readcoop.eu/).

Το Πανεπιστήμιο της Ζυρίχης του τμήματος Ιδιωτικού Δικαίου με επικεφαλής τον Δρ Walter Boente ανέπτυξαν ένα μοντέλο το Iurisprudentia, κάνοντας ένα ταξίδι στο χρόνο, το οποίο εστιάζει σε ιστορικά έγγραφα του αυστριακού, ελβετικού και γερμανικού νομικού συστήματος (Εικόνα 3). Περισσότερα από 390.000 έγγραφα, χειρόγραφα και τυπωμένα, μεταγράφηκαν με ένα εκπληκτικά χαμηλό ποσοστό σφαλμάτων (https://rwi.app/iurisprudentia/en/iurisprudentia).

Εικόνα 3: Παράδειγμα έκδοσης έργου που γίνεται προσβάσιμο χάρη στο Iurisprudentia.

(https://rwi.app/iurisprudentia/en/iurisprudentia).

Τα Εθνικά αρχεία της Νορβηγίας, το Πανεπιστήμιο του Δουβλίνου Trinity, τα εθνικά Αρχεία της Σουηδίας, η βιβλιοθήκη του Πανεπιστημίου του Μπέργκεν και πολλά άλλα ιδρύματα και φορείς, χρησιμοποιούν το λογισμικό Transkribus με επιτυχία, προκειμένου να αναδείξουν και να κατανοήσουν την πολιτιστική τους κληρονομιά (https://readcoop.eu/).

Ωστόσο, μεμονωμένοι ερευνητές, όπως η υποψήφια διδάκτορας Jessica Cool στο τμήμα αγγλικών στο UCLA, μεταγράφει ολόκληρο το αρχείο της Ada Lovelace - μαθηματικό του 19ου αιώνα και κόρη του ρομαντικού ποιητή Λόρδου Μπάιρον- αποτελούμενο από 14.000 σελίδες. Η Cook δημιούργησε το μοντέλο Lovelace Al στο Transkribus και σημειώνει ότι πολλές φορές το μοντέλο χάνει λέξεις που είναι εύκολα αναγνωρίσιμες από το ανθρώπινο μάτι, αλλά έχει αναγνωρίσει λέξεις, που οι προηγούμενοι ερευνητές όσο και η ίδια είχαν αποκρυπτογραφήσει λανθασμένα (https://readcoop.eu/).
Η γραπτή πολιτιστική κληρονομιά χαρακτηρίζεται από απίστευτη ποικιλία και ιδιαιτερότητες. Μελετώντας τις παραπάνω περιπτώσεις γίνεται κατανοητό ότι το Transkribus αποτελεί μια χρήσιμη τεχνολογία (πλατφόρμα) για την αναγνώριση, μεταγραφή, ευρετηρίαση και τον εμπλουτισμό των χειρόγραφων εγγράφων - όπως και των έντυπων - με μικρά ποσοστά λάθους σε πολλές διαφορετικές γλώσσες και στυλ γραφής. Δίνει την δυνατότητα πρόσβασης σε εκατοντάδες χιλιόμετρα αρχειακών και ιστορικών εγγράφων μεταλαμπαδεύοντας, κατανοώντας και διαδίδοντας κατά αυτόν τον τρόπο την παγκόσμια γραπτή πολιτιστική κληρονομιά.

2.5 Η σημασία αξιοποίησης της τεχνολογίας στα ελληνικά χειρόγραφα

Η εποχή που διανύουμε αναμφισβήτητα αποτελεί την εποχή της τεχνολογικής επανάστασης μεταμορφώνοντας δυναμικά τον κόσμο μας και εξελίσσοντάς τον σε ένα εκκολαπτήριο νέων τεχνολογιών. Η εισχώρηση των νέων τεχνολογιών σε κάθε τομέα της ζωής του ανθρώπου έχει εκσυγχρονίσει και έχει αποτελέσει εφαλτήριο αναζωπύρωσης

των δεξιοτήτων του. Η διείσδυση αυτή έχει επηρεάσει αναντίρρητα και τον χώρο του πολιτισμού, διατηρώντας και αναβιώνοντας κατ' αυτόν τον τρόπο το ιστορικό παρελθόν μας.

Αδιαμφισβήτητο γεγονός αποτελεί, ότι για ένα τεράστιο χρονικό διάστημα, κυρίαρχη θέση στην ελληνική κοινωνία, κατείχε το χειρόγραφο που εξυπηρετούσε τις ανάγκες της εκκλησίας, νομικές υποθέσεις, εμπορικές συναλλαγές, σχολικά εγχειρίδια αλλά και χειρόγραφα βιβλία απλά προς τέρψη των αναγνωστών. Η "αποκρυπτογράφηση" όλων αυτών των χειρόγραφων αποτελεί τη συνέχεια του πολυταξιδεμένου δρόμου της παράδοσης και εν γένει της πολιτιστικής μας κληρονομιάς. Η μεταγραφή αυτών των χειρόγραφων αποτελεί πηγή αληθινής γνώσης αλλά και αληθινών εκπλήξεων για τον εκάστοτε ερευνητή ή φορέα.

θα μπορούσαμε κάλλιστα να ισχυριστούμε ότι η μεταγραφή αυτών αποτελεί ένα κρίκο στην αλυσίδα της γνώσης και για λόγους τόσο δεοντολογίας, όσο και ευγένειας προς τον γράφοντα, πρέπει να είναι ισχυρός αλλά και ακριβής. Σκοπός του μεταγραφέα είναι να ερμηνεύσει το περιεχόμενο του κειμένου και να είναι συνεπής στα λεχθέντα του συγγραφέα. Ένα χειρόγραφο μπορεί να αποτελεί έναν αείροο ποταμό λέξεων χωρίς νόημα για έναν ερευνητή που δεν έχει γνώσεις παλαιογραφίας.

Τα χειρόγραφα επίσης, αποτελούν σημαντική πηγή γνώσεων σε πολλά επιστημονικά πεδία. Αποτελούν αναμφισβήτητα μια εξαιρετική πηγή για την άντληση στοιχείων που συμβάλλουν στην επιστήμη της Παλαιογραφίας. Οι σφραγίδες, το μελάνι, ο τρόπος γραφής κ.λπ. αποτελούν μερικά από τα πολλά στοιχεία που μπορούν να συλλεχθούν και να εμπλουτίσουν περαιτέρω τις γνώσεις των παλαιογράφων. Στο πεδίο της γλωσσολογίας, μας παρέχονται στοιχεία για την εξέλιξη της γλώσσας, διασώζοντας παράλληλα ποικίλους διαλεκτικούς τύπους, ιδιωματισμούς και τοπολαλιές. Βοηθούν στην τοπογραφία, καθώς σε πολλά ιστορικά κείμενα αναφέρονται πλήθος τοπωνυμίων. Παρέχουν επίσης, στοιχεία αρχιτεκτονικής της εποχής και της περιοχής στην οποία αναφέρονται. Αντλούνται πληροφορίες για τη θρησκεία, βοηθούν την αρχαιολογική επιστήμη, τις εθνολογικές, ανθρωπολογικές και νομικές σπουδές κ.α.

Αδήριτο γεγονός αποτελεί ο τεράστιος όγκος αυτών των χειρόγραφων. Η μεταγραφή τους με το χέρι φαντάζει ουτοπική. Οι ώρες και το κόστος εργασίας είναι τεράστιες. Η μελέτη των ανθρώπινων γνωστικών διαδικασιών μέσω των συστημάτων της τεχνητής νοημοσύνης και της βαθιάς μηχανικής μάθησης μπορούν να συμβάλλουν δραστικά και δυναμικά στην "αποκρυπτογράφηση" των ελληνικών χειρογράφων καθώς θα εξυπηρετήσουν τη διάσωση, κατανόηση και διάδοσή τους.

2.6 Φορείς ανάδειξης ελληνικών χειρογράφων

Ο επίσημος κρατικός φορέας στην Ελλάδα που ηγείται για τη διατήρηση, ανάδειξη και διάθεση της πολιτιστικής μας κληρονομιάς είναι τα Γενικά Αρχεία του Κράτους (Γ.Α.Κ.). Ο ογκώδης πλούτος πληροφοριών που παρέχουν είναι ανεκτίμητος και αφορά κάθε πτυχή δραστηριότητας τόσο της δημόσιας, όσο και της ιδιωτικής δραστηριότητας του έθνους. Οι αρχειακές συλλογές ζωντανεύουν μέσα από το σύστημα διαχείρισης αρχειακών δεδομένων "Αρχειομνήμων" με επτά εκατομμύρια και πλέον ψηφιοποιημένες σελίδες. Το Αρχειακό υλικό προέρχεται από την κεντρική και τριανταέξι περιφερειακές υπηρεσίες των Γ.Α.Κ. και αφορά τεκμήρια του 17ου και 20ου αιώνα αναδεικνύοντας την

ποικιλία του χρονικού βάθους που αυτά περιλαμβάνουν. Εξαιρετικά σημαντικό τμήμα των Γ.Α.Κ. είναι το Αρχείο Χαρτογραφικής Κληρονομιάς (ΑΧαΚ), το οποίο διαθέτει στο κοινό εκτός από χαρτώο υλικό και τη ψηφιακή χαρτοθήκη που καλύπτει διάφορες περιόδους και φάσεις της Ιστορίας των χαρτών (http://arxeiomnimon.gak.gr/).

Επίσης τα Γ.Α.Κ. προσφέρουν, εκτός από τις ψηφιακές συλλογές, τη Βάση υδατοσήμων ιστορικών εγγράφων των Γ.Α.Κ. η οποία διατίθεται στη διεθνή πύλη υδατοσήμων Bernstein – The Memory of Paper (http://arxeiomnimon.gak.gr/). Επιπλέον, συμμετέχουν στην Ευρωπαϊκή Πύλη Αρχείων η οποία διαθέτει περισσότερες από 600.000 συλλογές πολλών εκατομμυρίων εγγράφων από 30 χώρες. Στην Ευρωπαϊκή αυτή πύλη συνεισφέρουν με αρχειακές περιγραφές μέχρι στιγμής 7000 αρχειακά ιδρύματα, σε περισσότερες από 20 γλώσσες με 5 διαφορετικά αλφάβητα (https://www.archivesportaleurope.net/about-us/the-portal/).

Το Μορφωτικό Ίδρυμα Εθνικής Τραπέζης (ΜΙΕΤ) συστάθηκε την 1η Απριλίου 1966 με σκοπό την ανάδειξη των επιστημών, των τεχνών και των γραμμάτων. Σημαντικό τμήμα του ιδρύματος αποτελεί το Ιστορικό και Παλαιογραφικό Αρχείο, το οποίο διαθέτει αρχείο μικροφίλμ εννιάμιση χιλιάδων περίπου χειρογράφων καθώς και είκοσι μεγάλων ιστορικών αρχείων. Διαθέτει συλλογές χειρογράφων από τον 11ο ως τον 20ο αιώνα (https://www.miet.gr/palaiografiko-arxio/). Επιπροσθέτως, το Ελληνικό Λογοτεχνικό και Ιστορικό Αρχείο ΕΛΙΑ, που αποτελεί τμήμα του ΜΙΕΤ, διαφυλάσσει αρχεία δύο αιώνων. Ανάμεσά τους ξεχωρίζουν τα αρχεία της οικογένειας Κολοκοτρώνη, του Ελ. Βενιζέλου και του Χ. Τρικούπη. (http://www.elia.org.gr/archives-collections/archives/).

Το Εθνικό Κέντρο Τεκμηρίωσης & Ηλεκτρονικού Περιεχομένου (ΕΚΤ), μη κερδοσκοπικού χαρακτήρα, αποτελεί έναν οργανισμό που προάγει τη γνώση, την καινοτομία και τον ψηφιακό μετασχηματισμό. Η εμβέλειά του είναι εθνική και ο ρόλος του πολύπλευρος. Από τους πιο σημαντικούς, η ψηφιακή διατήρηση της τεχνολογικής και πολιτιστικής πληροφορίας που παράγεται στην Ελλάδα. Διαθέτει σύγχρονη πληροφοριακή υποδομή και εξοπλισμό ψηφιοποίησης αρχειακού υλικού και παράλληλα αποτελεί τον Εθνικό Συσσωρευτή για την ευρωπαϊκή ψηφιακή βιβλιοθήκη Europeana (https://www.ekt.gr/).

Εντυπωσιακή είναι και η ψηφιακή βιβλιοθήκη Theasurus Linguae Graecae (Ο Θησαυρός της Ελληνικής Γλώσσας) η οποία αναπτύχθηκε από το Πανεπιστήμιο της Καλιφόρνια και αποτελεί μία από τις πρώτες προσπάθειες, στις Ανθρωπιστικές Σπουδές, για την συγκέντρωση και την ψηφιοποίηση ελληνικών λογοτεχνικών κειμένων. Η TLG περιέχει ψηφιοποιημένα κείμενα από τον 8ο αιώνα π.Χ. μέχρι και την πτώση του Βυζαντίου 1453 π.Χ. καθώς και μεγάλο αριθμό κειμένων του 20ου αιώνα (http://stephanus.tlg.uci.edu/history.php).

3. Μεταγραφή ελληνικού χειρόγραφου αρχείου 19ου αιώνα

Η ψηφιοποίηση των ιστορικών αρχείων είναι απαραίτητη για τη διάσωση, συγκέντρωση και καταγραφή των πρωτογενών πηγών έρευνας κυρίως για τις ανθρωπιστικές επιστήμες. Σύμφωνα με τη Βίβλο Ψηφιακού Μετασχηματισμού (2021) προγραμματίζονται προς υλοποίηση έργα ψηφιοποίησης όλων των φυσικών αρχείων των Γενικών Αρχείων του Κράτους. Παρόλο όμως που αποτελεί πολύ σημαντικό βήμα για τη διαφύλαξη του πλούσιου αυτού υλικού, δεν είναι αρκετό για την αναζήτηση, μεταγραφή και μελέτη των πηγών που θα οδηγήσει στην εξαγωγή ιστορικών συμπερασμάτων.

Η ανάπτυξη της τεχνολογίας δίνει πλέον τη δυνατότητα πειραματισμών πάνω στο σημαντικό αυτό πρόβλημα. Στη συνέχεια θα παρουσιαστεί η περίπτωση μεταγραφής ενός ελληνικού συμβολαιογραφικού χειρόγραφου αρχείου του 19ου αιώνα με τη βοήθεια των εργαλείων που προσφέρει το λογισμικό transkribus.

3.1. Αρχείο συμβολαιογράφου Αναστασίου Φίλιου

Τα συμβολαιογραφικά αρχεία που βρίσκονται στα Γενικά Αρχεία Λάρισας αποτελούν εθνική παρακαταθήκη. Η ερμηνεία των πηγών αυτών είναι θεμελιώδης για την τεκμηρίωση της τοπικής ιστορίας γιατί, μεταξύ άλλων, αποτυπώνουν τη μετάβαση της κυριότητας γης από τους Οθωμανούς στους Έλληνες.

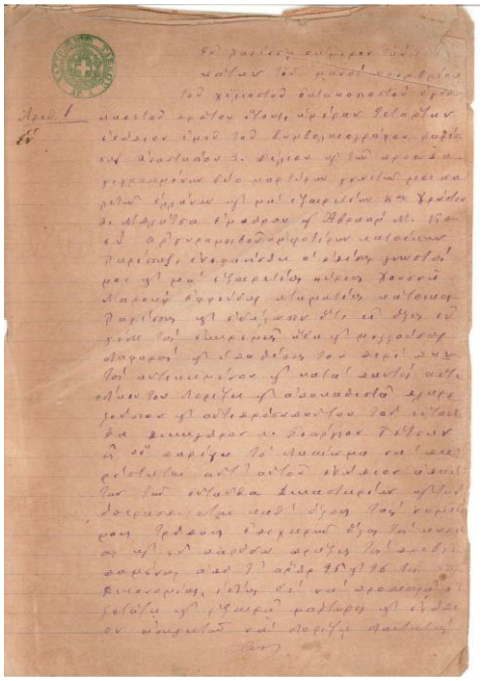
Στην παρούσα έρευνα, ως μελέτη περίπτωσης για τη διερεύνηση τεχνολογιών που μπορούν να προσφέρουν νέες πρακτικές και μεθόδους προσέγγισης ελληνικών χειρόγραφων ιστορικών πηγών, επιλέχθηκε το αρχείο του συμβολαιογράφου Λαρίσης Αναστασίου Φίλιου. Το χρονικό διάστημα που πραγματεύεται είναι από το έτος 1881 έως το 1901 και αριθμεί στο σύνολο 46.450 συμβολαιογραφικές πράξεις. Είναι ένα κλειστό αρχειακό σύνολο που παραδόθηκε στα Γενικά Αρχεία της Λάρισας τον Μάιο του 1998, ψηφιοποιήθηκε μέρος του και αναρτήθηκε στο σύστημα διαχείρισης αρχειακών δεδομένων @ρχειομνήμων των Γενικών Αρχείων του Κράτους (http://gak.lar.sch.gr/).

Η ιστορική του αξία είναι μεγάλη διότι αποτυπώνει από τη μία πλευρά τη μετάβαση της κυριότητας γης από τους Οθωμανούς στους Έλληνες και από την άλλη την κοινωνική και οικονομική κατάσταση των πολιτών όπως διαμορφώθηκε μετά την αλλαγή από υπόδουλο λαό σε ελεύθερους Έλληνες πολίτες.

Επιλέχθηκε επίσης το συγκεκριμένο αρχείο γιατί τα μεθοδολογικά προβλήματα που ανακύπτουν κατά την μεταγραφή του είναι πολυάριθμα. Για παράδειγμα, οι ορθογραφικές ιδιαιτερότητες του συμβολαιογράφου καθιστούν την ανάγνωση πολλών λέξεων δυσνόητη καθώς γίνεται χρήση τόσο της ιστορικής ορθογραφίας όσο και της νέας. Διακρίνονται πολλές βραχυγραφίες και χρησιμοποιείται το πολυτονικό σύστημα. Τα γράμματα αποτυπώνονται σε πολλά σημεία του χαρτιού άναρχα και ανισομερή και η ένταση του μελανιού ποικίλει. Επιπροσθέτως, παρατηρείται συχνά σε πολλά σημεία του εγγράφου να υπάρχουν σημειώσεις εκτός πλαισίου. Το κείμενο είναι γραμμένο στην καθαρεύουσα και σε καλλιγραφία (Εικόνα 4), συνεπώς η ανάγνωση και κατ' επέκταση η κατανόηση και ερμηνεία εμφανίζει δυσκολίες.

Το αρχείο αποτελείται από συμβολαιογραφικές πράξεις διαφόρων κατηγοριών (πωλητήρια γης, οικιών, ενοικιαστήρια, πληρεξούσια, εκθέσεις πλειστηριασμού,

πωλητήρια σίτου, δανειστικά, εξοφλητικά, διανεμητήρια, συμφωνητικά κ.ά.). Η κάθε πράξη δίνει πλήθος πληροφοριών όπως: χρονολογία, τόπο σύνταξης και υπογραφής, ονοματεπώνυμο και πατρώνυμο του συμβολαιογράφου και των συμβαλλομένων, καθώς και τα στοιχεία των μαρτύρων, αντιπροσώπων, διερμηνέων και όλων όσων συμπράττουν στη συμφωνία. Τέλος, δίνονται πληροφορίες για το οικονομικό αντικείμενο, οι οποίες δεν έχουν μόνο ιστορική αξία αλλά και χρηστική. Τα τελευταία χρόνια με τη σύνταξη του Εθνικού Κτηματολογίου πολύ συχνά οι ιδιοκτήτες ακινήτων καταφεύγουν στην έρευνα σε συμβολαιογραφικά αρχεία της περιόδου προσάρτησης της Θεσσαλίας στο Ελληνικό κράτος, για να αποδείξουν σε δικαστήρια τους τίτλους ιδιοκτησίας τους. Συγκεκριμένα, το πρώτο μεταπελευθερωτικό διάστημα συντάχθηκε μεγάλος αριθμός συμβολαίων με πωλητές Οθωμανούς και αγοραστές Έλληνες. Οι πράξεις όμως αυτές θεωρούνταν άκυρες από το ελληνικό δημόσιο γιατί επί των τέως Δημοσίων Οθωμανικών Γαιών οι Τούρκοι πωλητές δεν είχαν δικαίωμα πλήρους κυριότητας, επομένως δεν δικαιούνταν να μεταβιβάσουν ακίνητη περιουσία (Παπαχατζόπουλος, 2016, σελ.105). Απόρροια αυτού είναι πολλές τέτοιες υποθέσεις να καταλήγουν στα ελληνικά δικαστήρια με αποδεικτικά έγγραφα τις συμβολαιογραφικές πράξεις. Κατά συνέπεια, οι εμπλεκόμενοι έρχονται συχνά αντιμέτωποι με τις δυσκολίες ανάγνωσης και μεταγραφής που προαναφέρθηκαν.



Εικόνα 4: Συμβόλαιο αρ. 1.

(Φίλιος, 1881).

3.2. Η πλατφόρμα Transkribus

Το λογισμικό Transkribus δημιουργήθηκε από την κοινοπραξία κορυφαίων ερευνητικών ομάδων της Ευρώπης μέσω του Πανεπιστημίου του Ίνσμπρουκ στην Αυστρία. Είναι μια ολοκληρωμένη πλατφόρμα που μπορεί να χρησιμοποιηθεί από φορείς ή ιδιώτες για την

ψηφιοποίηση αρχείων, την αναγνώριση χειρόγραφου κειμένου, τη μεταγραφή και την αναζήτηση ιστορικών εγγράφων. Λειτουργεί συνεργατικά και μπορεί ο καθένας με μια απλή εγγραφή να εκπαιδεύσει τα δικά του μοντέλα αναγνώρισης χειρόγραφου κειμένου.

Βασίζεται στην τεχνολογία της μηχανικής μάθησης, δηλαδή στην ικανότητα ενός υπολογιστικού συστήματος να δημιουργεί μοντέλα ή πρότυπα από ένα σύνολο δεδομένων (Γεωργούλη, c2015). Συγκεκριμένα, εκχωρούνται ψηφιοποιημένες εικόνες που συνοδεύονται από τη μεταγραφή τους σε μορφή ψηφιακού κειμένου, σε συμφωνία γραμμή προς γραμμή μεταξύ της εικόνας και του κειμένου. Έτσι δημιουργείται ένα μοντέλο HTR το οποίο μπορεί να δώσει αυτόματη μεταγραφή των εικόνων. Το μοντέλο που παράγεται με αυτόν τον τρόπο μπορεί αργότερα να βελτιωθεί με την προσθήκη νέων εικόνων και μεταγραφών επαναλαμβάνοντας τη χειροκίνητη διαδικασία εκπαίδευσης, δίνοντας ένα νέο μοντέλο. Η διαδικασία αυτή μπορεί να επαναληφθεί μέχρι να μειωθεί στο ελάχιστο το ποσοστό λάθους και να δώσει ένα αποτέλεσμα που να ικανοποιεί τους χρήστες.

Οι ελεύθερα διαθέσιμες υπηρεσίες του αφορούν τέσσερις βασικές ομάδες χρηστών, τους αρχειονόμους, τους ακαδημαϊκούς των ανθρωπιστικών επιστημών, τους επιστήμονες των υπολογιστών καθώς και τους ερευνητές που ασχολούνται τόσο με τη μελέτη, όσο και με την εκμετάλλευση των ιστορικών τεκμηρίων. Σε έρευνα που πραγματοποιήθηκε το 2022 σε μηχανές αναζήτησης ακαδημαϊκών δημοσιεύσεων, βρέθηκε ότι από το 2015 έως το 2020 σε 381 μελέτες που δημοσιεύθηκαν με γνωστικό αντικείμενο την πλατφόρμα Transkribus και το HTR, το Transkribus εμφανίζεται κυρίως σε μελέτες που αφορούν στην επιστήμη της αρχειονομίας και βιβλιοθηκονομίας, ενώ μεγάλος είναι και ο αριθμός αυτών που αφορούν στην ιστορία, στην επιστήμη των υπολογιστών, στο δίκαιο και στην εκπαίδευση, αποδεικνύοντας την ευρύτερη δυνατότητα εφαρμογής του εργαλείου. Συγκεκριμένα, το 67% ήταν μελέτες του κλάδου των ανθρωπιστικών επιστημών, 25% μελέτες από ερευνητές που ανήκουν στο πεδίο της τεχνολογίας, 5% από χρήστες γενικών ενδιαφερόντων και 3% ήταν παρουσιάσεις που εξηγούν πώς λειτουργεί η συγκεκριμένη τεχνολογία. Αξίζει να επισημανθεί ότι, εμφανίζονται μελέτες από 31 τομείς επιστημονικού ενδιαφέροντος, όπερ σημαίνει ότι το Transkribus είναι ιδιαίτερα χρήσιμο σε πληθώρα ερευνητών που εργάζονται σε διαφορετικά πεδία ενδιαφέροντος (Nockels, Gooding, Ames, et al.., 2022).

Η ενασχόληση, τα ενδιαφέροντα και η συμβολή αυτών των ομάδων αλληλεπικαλύπτονται και παράλληλα αποτελούν ζωτικό παράγοντα στην δομή και στην οργάνωση του Transkribus. Συγκεκριμένα, τα ιδρύματα μνήμης, οι ερευνητές των ανθρωπιστικών σπουδών και το κοινό, επιτρέπουν την παροχή ψηφιοποιημένων εικόνων και μεταγραφών ως βάση για εξάσκηση στο HTR. Από την άλλη, οι επιστήμονες της πληροφορικής, αποτελούν τους ερευνητές εκείνους που θα υποστηρίξουν αυτήν την τεχνολογία. Με αυτό τον τρόπο κάθε ομάδα χρηστών δράττει τους καρπούς από την συγκεκριμένη πρωτοβουλία.

Προσφέρονται επίσης, ψηφιοποιημένες συλλογές με δυνατότητα αναζήτησης ώστε να διευκολύνεται η έρευνα. Το κοινό μπορεί να ενημερωθεί και να βρει πληροφορίες για τη χρήση της πλατφόρμας και παράλληλα να συμβάλλει δραστικά και ουσιαστικά στην αντιγραφή και στην διόρθωση μεταγραφών ιστορικών τεκμηρίων. Όσον αφορά τους χρήστες που εκπροσωπούν τους επιστήμονες των υπολογιστών μπορούν να

χρησιμοποιήσουν εκ νέου μία ευρεία γκάμα δεδομένων με την μορφή εικόνων του ιστορικού υλικού, σχετικά με την έρευνά τους πάνω στο HTR.

Εκτός από την αυτόματη έκδοση (Transkribus Expert Client) που μπορεί ο κάθε χρήστης να κατεβάσει στον υπολογιστή του, διατίθεται στο κοινό και η έκδοση Transkribus Lite, η οποία διαθέτει φιλικό on line περιβάλλον εργασίας παρέχοντας βελτιωμένη διεπαφή με το χρήστη. Πολλές από τις δυνατότητες του Transkribus Expert Client μπορούν να χρησιμοποιηθούν στο Transkribus lite. Με αυτή την έκδοση δίνεται η δυνατότητα στους χρήστες να χρησιμοποιήσουν το Transkribus στο φυλλομετρητή τους, προσφέροντας εύκολη πρόσβαση σε όλα τα έγγραφα των χρηστών από οποιοδήποτε πρόγραμμα περιήγησης. Επιπλέον, είναι εφικτή η αυτόματη μεταγραφή, η επεξεργασία των εγγράφων με άνεση, καθώς επίσης και η εύκολη συνεργατική εργασία σε ιστορικά έγγραφα.

Μία επιπρόσθετη λειτουργία που προσφέρει, όχι όμως δωρεάν, είναι το Metagrapho api, η οποία επιτρέπει στους φορείς να ενσωματώσουν την πλατφόρμα αναγνώρισης κειμένου στις δικές τους εφαρμογές ή να δημιουργήσουν νέες εφαρμογές αναγνώρισης χειρόγραφου ή τυπωμένου κειμένου. Εκτός από αυτό, υπάρχει αντίστοιχα και ένα δωρεάν εργαλείο αναγνώρισης χειρόγραφου κειμένου μεταφοράς και απόθεσης που έχει δημιουργηθεί με το metagrapho API, το Transkribus ai, με το οποίο μπορούν οι ενδιαφερόμενοι να δοκιμάσουν την πλατφόρμα on line, να επιλέξουν τη γλώσσα του κειμένου και με τη χρήση ενός δημόσιου μοντέλου να μεταγράψουν χειρόγραφο ή τυπωμένο κείμενο. Η εφαρμογή αυτή όμως δεν διαθέτει την επιλογή της ελληνικής γλώσσας.

Εν κατακλείδι, το Transkribus αποτελεί ένα χρήσιμο εργαλείο για αρχειονόμους, βιβλιοθηκονόμους, πληροφορικούς και ερευνητές κυρίως ανθρωπιστικών επιστημών, γιατί:

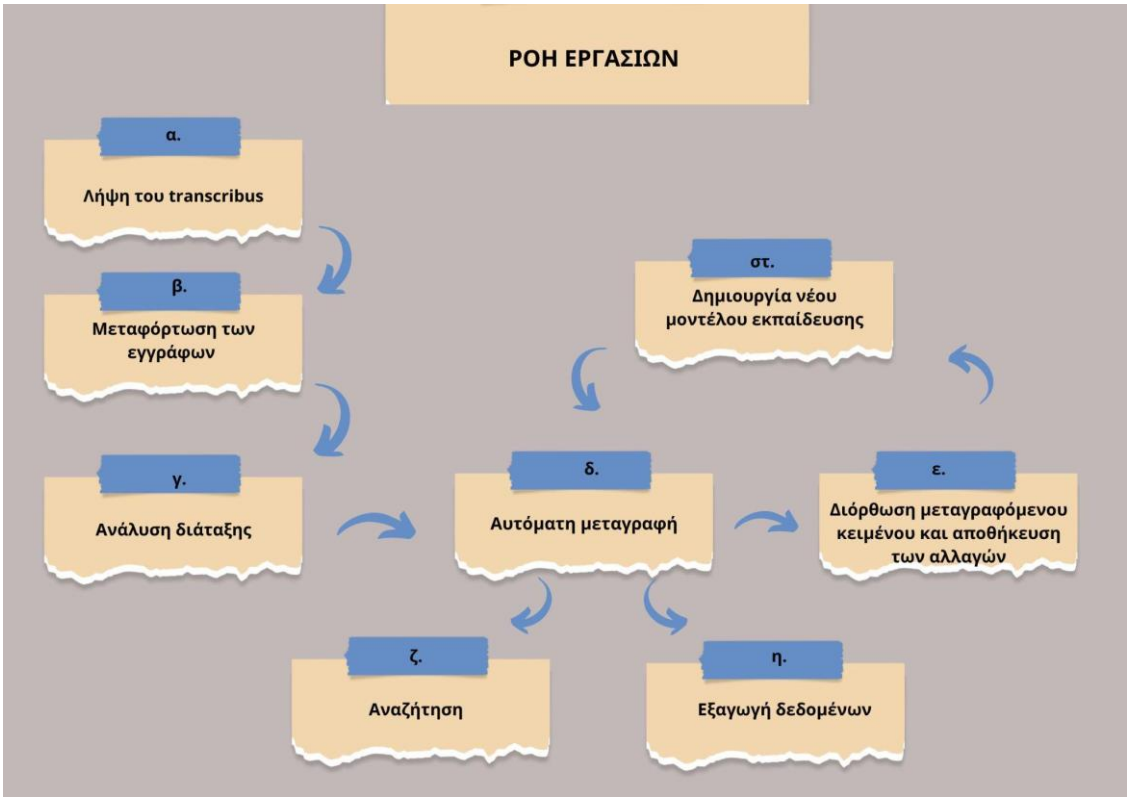
- πραγματοποιεί ανάλυση κειμένου αναγνωρίζοντας με τεχνητή νοημοσύνη τη διάταξη και τη δομή του κειμένου,
- μεταγράφει αυτόματα χειρόγραφο κείμενο με τη χρήση δημοσίων μοντέλων τεχνητής νοημοσύνης, παράλληλα δίνει τη δυνατότητα της μη αυτόματης μεταγραφής ώστε να εκπαιδεύσει ο καθένας τα δικά του μοντέλα,
- δίνει τη δυνατότητα εντοπισμού των εγγράφων με το εργαλείο αναζήτησης πλήρους κειμένου (fulltext), tags και λέξεων κλειδιών (KWS),
- επιτρέπει τη συνεργατική εργασία σε συλλογές εγγράφων,
- κάνει τα ιστορικά έγγραφα προσβάσιμα και αναζητήσιμα στον παγκόσμιο ιστό.

3.3 Εφαρμογή σε ελληνικό χειρόγραφο συμβολαιογραφικό αρχείο

Όπως αναφέρθηκε στο κεφάλαιο 3.1, το αρχείο του συμβολαιογράφου Φίλιου αποτελείται από 46.450 πράξεις, οι οποίες αποτελούν ένα ανομοιοειδές σύνολο, αφού ο τρόπος γραφής σε αρκετές πράξεις διαφέρει ανάλογα με τον γραφέα του συμβολαιογραφικού γραφείου που συντάσσει το κείμενο. Για τις ανάγκες τις παρούσας εργασίας χρησιμοποιήθηκαν 21 πράξεις, που συμπληρώνουν 50 σελίδες χειρόγραφου κειμένου. Αρχικά, επιλέχθηκαν κείμενα γραμμένα με τον ίδιο γραφικό χαρακτήρα για να μειωθεί η πολυπλοκότητα της μελέτης και να υποστηριχθεί η διαδικασία εκπαίδευσης

των μοντέλων HTR. Στη συνέχεια, πριν ξεκινήσει η διαδικασία εκπαίδευσης, έγινε η ανάγνωση και μεταγραφή των κειμένων από τους συγγραφείς της εργασίας ώστε να διευκολυνθεί η διαδικασία διόρθωσης του αυτόματα μεταγραφόμενου κειμένου. Δεν χρειάστηκε σάρωση των εγγράφων, διότι το συγκεκριμένο υλικό είναι ήδη ψηφιοποιημένο στα ΓΑΚ Λάρισας και διατίθεται προς χρήση του κοινού από το Web site της υπηρεσίας (http://gak.lar.sch.gr/).

Για την εφαρμογή της τεχνολογίας του Transkribus, ώστε να αξιολογηθεί η πλατφόρμα ως προς την αποτελεσματικότητα χρήσης σε ελληνικά χειρόγραφα κείμενα, πραγματοποιήθηκε μία σειρά ενεργειών οι οποίες θα παρουσιαστούν αναλυτικά στη συνέχεια (Εικόνα 5).



Εικόνα 5: Διάγραμμα ροής εργασιών.

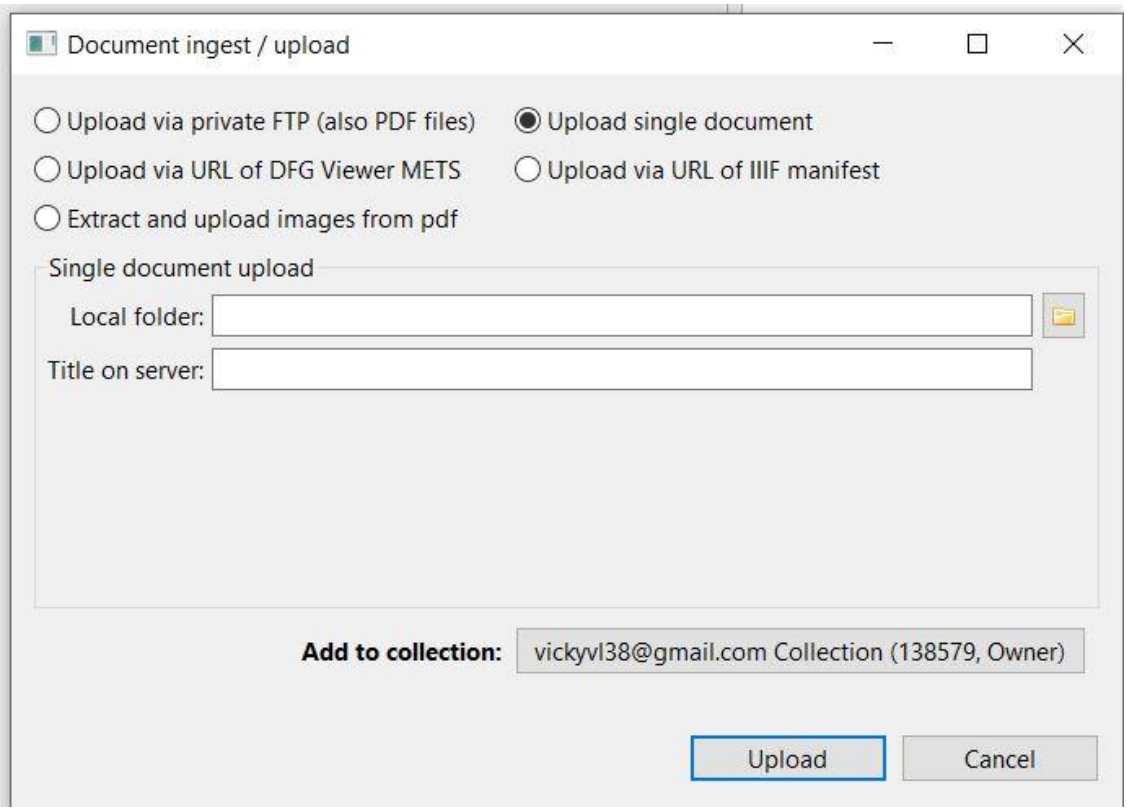
α. Λήψη του Transkribus

Η διαδικασία ξεκίνησε με τη λήψη του Transkribus. Είναι μια απλή διαδικασία κατά την οποία οι ενδιαφερόμενοι εγγράφονται στον ιστότοπο του readcoop.eu δίνοντας το email τους και εγκαθιστούν δωρεάν την πλατφόρμα του Transkribus Expert Client στον υπολογιστή τους. Οι τεχνικές προδιαγραφές δεν απαιτούν εξειδικευμένο εξοπλισμό, αρκεί οι ενδιαφερόμενοι να διαθέτουν Windows 64 bit ή Mac ή Linux λειτουργικό σύστημα και να έχουν εγκατεστημένη μια έκδοση της Java 64-bit. Εναλλακτικά, μπορούν να συνδεθούν διαδικτυακά στο Transcribus Lite κάνοντας σύνδεση με τους κωδικούς της εγγραφής τους.

β. Μεταφόρτωση των εγγράφων

Μετά τη λήψη του Transkribus Expert Client, έγινε η μεταφόρτωση των εγγράφων που προορίζονταν για τη μεταγραφή. Υπάρχουν διαφορετικές επιλογές για τη μεταφόρτωση,

όπως μεταφόρτωση με File Transfer Protocol (FTP), που είναι ιδιαίτερα χρήσιμο για μεγάλα αρχεία εικόνας ή PDF, μεταφόρτωση για μεμονωμένα αρχεία PDF και μεταφόρτωση για μεμονωμένο έγγραφο που απευθύνεται σε αρχεία εικόνας (Εικόνα 6). Επιλέχθηκε η μεταφόρτωση μεμονωμένων εγγράφων επειδή τα σαρωμένα έγγραφα που χρησιμοποιήθηκαν ήταν μόνο αρχεία εικόνας.



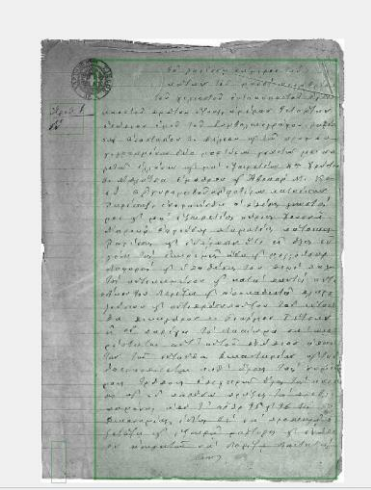
Εικόνα 6: Μεταφόρτωση αρχείων.

γ. Ανάλυση διάταξης

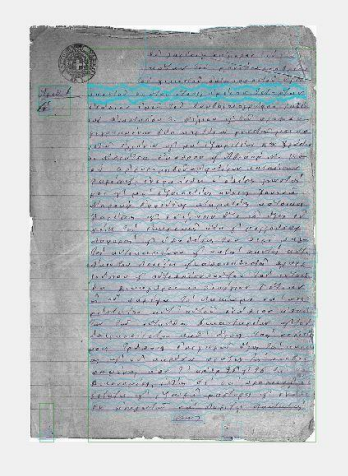
Πολύ σημαντικό βήμα στην όλη διαδικασία αποτελεί η αυτόματη ή χειροκίνητη τμηματοποίηση που προηγείται της αναγνώρισης. Με τη χρήση εργαλείων αυτόματης ανάλυσης διάταξης πραγματοποιήθηκε ο διαχωρισμός των εικόνων σε περιοχές κειμένου και γραμμές.

Αρχικά ζητήθηκε η αυτόματη ανάλυση διάταξης με την τμηματοποίηση του κειμένου σε περιοχές κειμένου, περιοχές γραμμών και γραμμές βάσης (

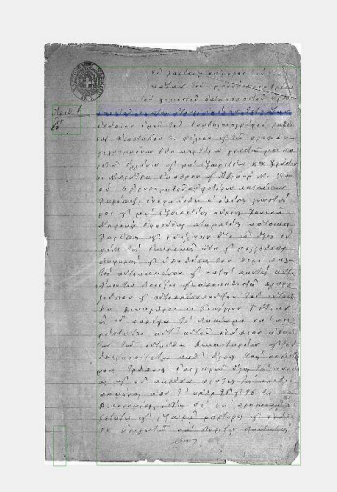
Εικόνα 7, Εικόνα 8, Εικόνα 9). Δίνεται όμως η δυνατότητα της χειροκίνητης επέμβασης και διόρθωσης της τμηματοποίησης. Η λειτουργία αυτή ήταν πολύ βοηθητική, ειδικά σε αυτό το αρχείο, γιατί στο συγκεκριμένο είδος χειρόγραφων τεκμηρίων υπάρχουν περιοχές κειμένου με σφραγίδες, υπογραφές, σημειώσεις, διορθώσεις κ.λπ. οι οποίες κατά τη διαδικασία αναγνώρισης απομονώθηκαν για να μην παρεμποδίζουν τη ροή ανάγνωσης.



Εικόνα 7: Περιοχές κειμένου.



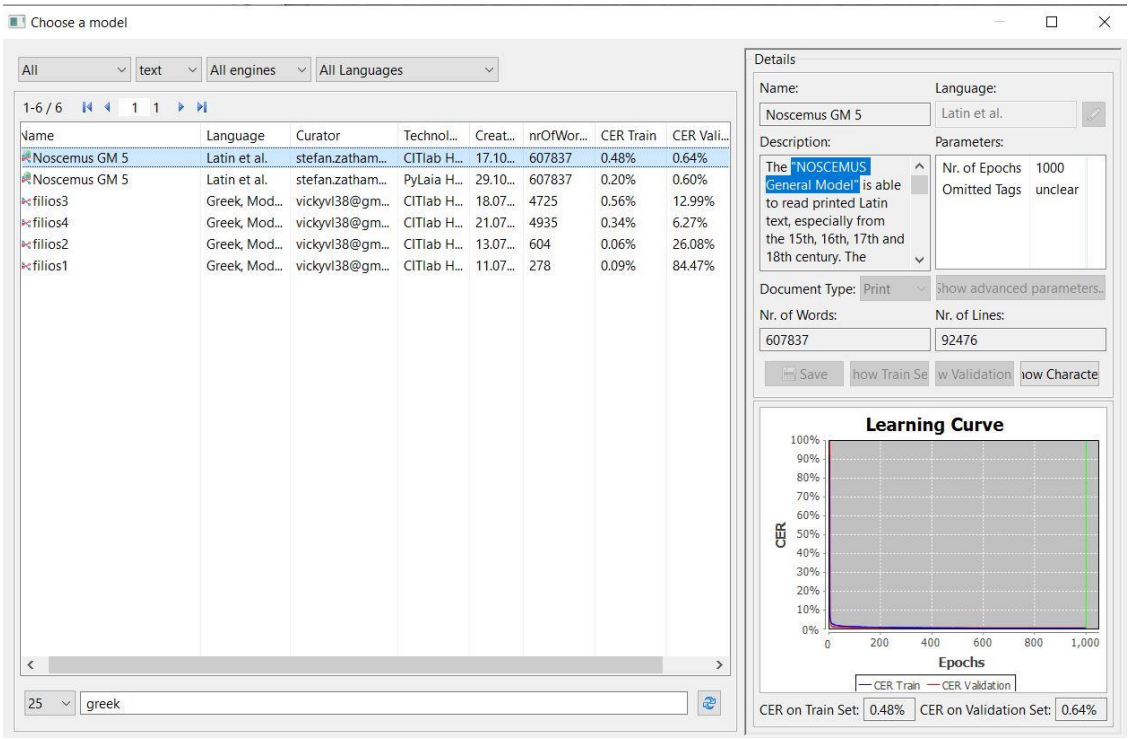
Εικόνα 8: Περιοχές γραμμών.



Εικόνα 9: Γραμμές βάσης.

δ. Αυτόματη μεταγραφή

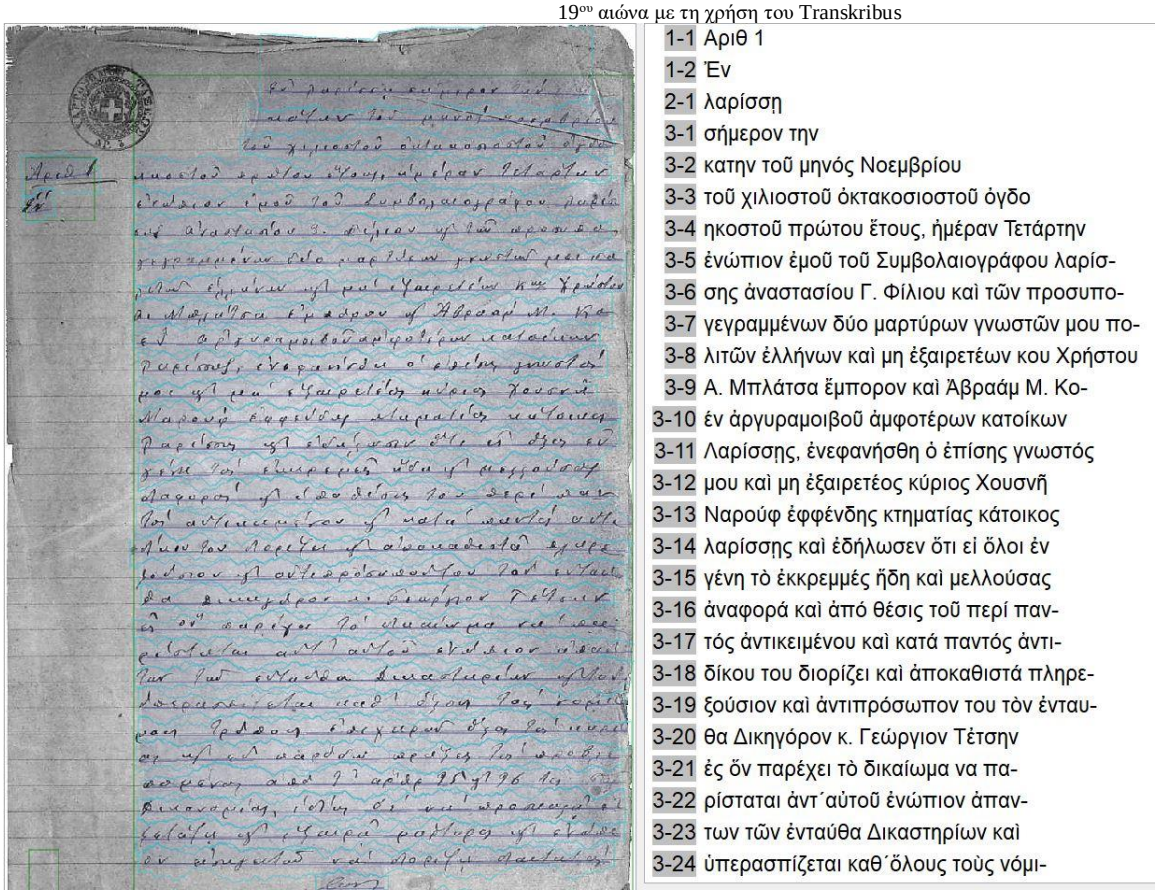
Μετά την μεταφόρτωση των εικόνων και την ανάλυση διάταξης, το σύστημα ήταν έτοιμο για να προχωρήσει στην αυτόματη μεταγραφή των κειμένων. Ως πρώτο βήμα ζητήθηκε το Transkribus να διαβάσει και να μεταγράψει το χειρόγραφο κείμενο, χρησιμοποιώντας ένα ήδη υπάρχον δημόσιο μοντέλο HTR το "NOSCEMUS General Model" (Εικόνα 23). Η διαδικασία ήταν απλή και πραγματοποιήθηκε χρησιμοποιώντας την επιλογή “RUN”, από την ενότητα “Text Recognition” που βρίσκεται στην καρτέλα “Tools”. Μετά από μία σειρά επιβεβαιώσεων που ζητήθηκαν και μία μικρή αναμονή, η διαδικασία ολοκληρώθηκε και το μεταγραμμένο κείμενο εμφανίστηκε στο πρόγραμμα επεξεργασίας κειμένου.



Εικόνα 10: Μοντέλα που προϋπήρχαν και μοντέλα που δημιουργήθηκαν.

ε. Διόρθωση μεταγραφόμενου κειμένου και αποθήκευση των αλλαγών

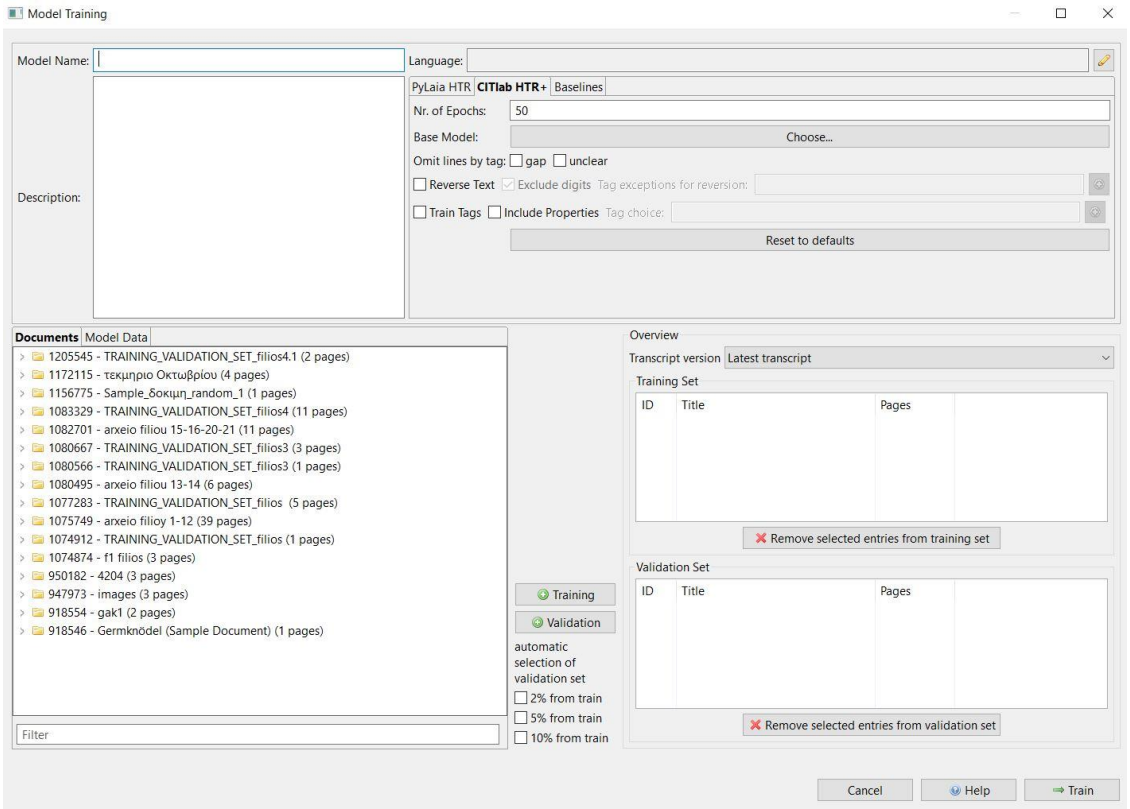
Για κάθε γραμμή βάσης (Εικόνα 9) υπάρχει μία αντίστοιχη γραμμή στον επεξεργαστή κειμένου. Εκεί διορθώθηκε χειροκίνητα η αυτόματη μεταγραφή περνώντας γραμμή-γραμμή τα δεδομένα εκπαίδευσης (Εικόνα 11). Η διαδικασία ολοκληρώθηκε με την αποθήκευση των αλλαγών για τη δημιουργία νέου μοντέλου εκπαίδευσης. Να επισημανθεί σε αυτό το σημείο ότι συνήθως απαιτείται ένας μικρός αριθμός σελίδων (25-75) μη αυτόματης μεταγραφής για να ξεκινήσει η εκπαίδευση του μοντέλου.



Εικόνα 11: Χειροκίνητη διόρθωση μεταγραφής.

στ. Δημιουργία νέου μοντέλου εκπαίδευσης

Οι κύριες επιλογές για την εκπαίδευση ενός μοντέλου βρίσκονται στην καρτέλα "Tools" στην ενότητα "Model Training". Για να πραγματοποιηθεί η εκπαίδευση με την εντολή “Train a new model”, χρειάστηκε πρώτα να προστεθούν μία σειρά επιλογών και πληροφοριών όπως το όνομα του μοντέλου, η γλώσσα, η σύντομη περιγραφή του μοντέλου και του ιστορικού του. Ακολούθησε η επιλογή ενός βασικού μοντέλου και των εγγράφων που συμμετείχαν στο εκπαιδευτικό σετ και στο σετ επικύρωσης. Επιλέχθηκε η μηχανή αναγνώρισης χειρόγραφου κειμένου “CITlab HTR+” (Εικόνα 12) έναντι της PyLaia γιατί το κείμενο προς μεταγραφή ήταν σε καλλιγραφία, επομένως θεωρήθηκε καταλληλότερο αφού η “CITlab HTR+” φέρνει καλύτερα αποτελέσματα σε κείμενο με καμπύλες και περιστρεφόμενες γραμμές (https://readcoop.eu/it/transkribus/howto/how-to-train-pylaia-models-in-transkribus/).

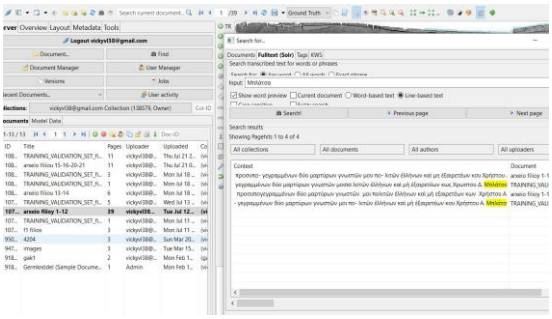


Εικόνα 12: Εκπαίδευση νέου μοντέλου.

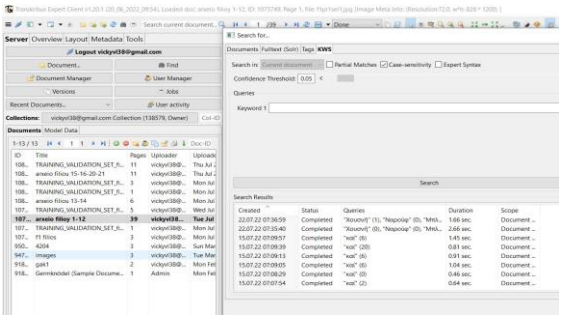
Στη συνέχεια δημιουργήθηκε το μοντέλο filios1 και καθώς προχωρούσε η διαδικασία της χειρόγραφης διόρθωσης και εκπαίδευσης ακολούθησαν τα μοντέλα filios2, filios3 και filios4 (Εικόνα 10).

ζ. Αναζήτηση

Το Transkribus όπως προαναφέρθηκε παρέχει τη δυνατότητα τα μεταγραμμένα έγγραφα να γίνουν προσβάσιμα και αναζητήσιμα στον παγκόσμιο ιστό. Στην έκδοση Transkribus expert client που χρησιμοποιήθηκε στην παρούσα έρευνα, μπορεί να πραγματοποιηθεί εντοπισμός κειμένου σε μεταγραμμένα κείμενα με τη λειτουργία του εργαλείου “Find”, το οποίο καθιστά εφικτή τη δυνατότητα αναζήτησης πλήρους κειμένου ή λέξεων-κλειδιών (Εικόνα 13, Εικόνα 14).



Εικόνα 13: Αναζήτηση πλήρους κειμένου



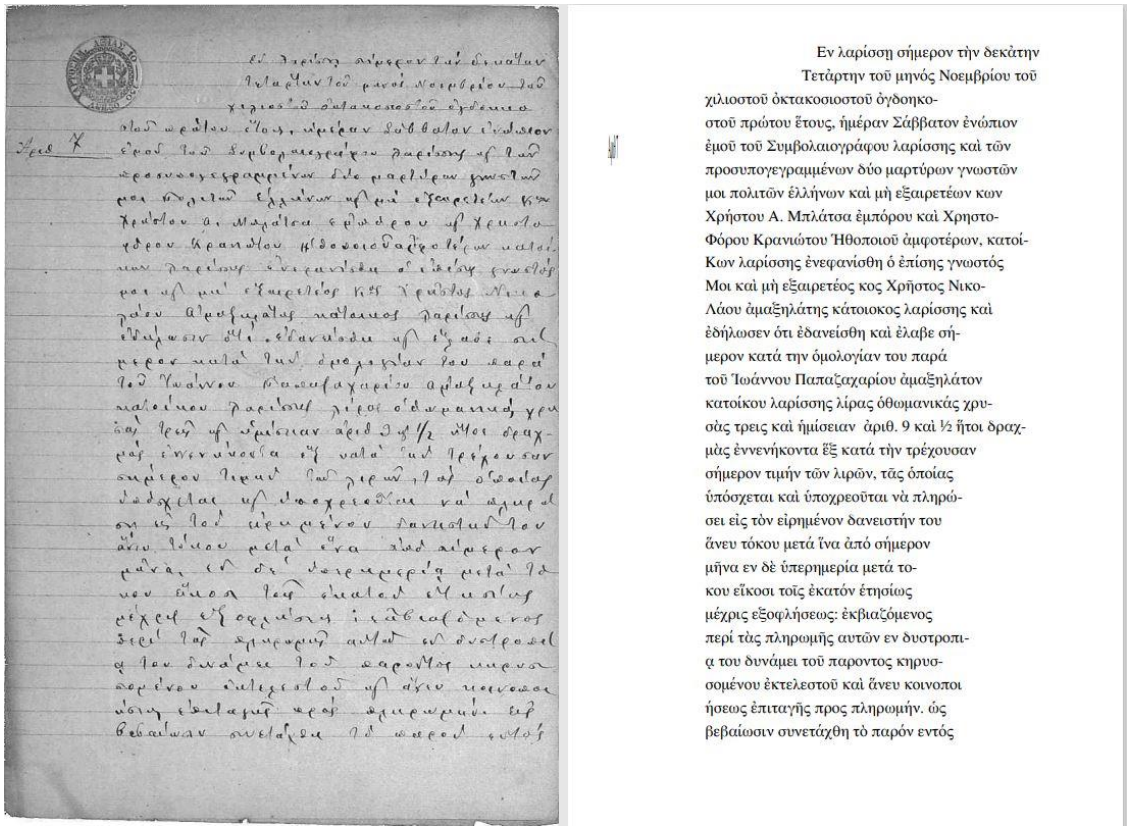
Εικόνα 14: Αναζήτηση με λέξη κλειδί

Συγκεκριμένα, αυτό το εργαλείο πραγματοποιεί αναζήτηση στις τιμές που έχουν αποδοθεί σε χαρακτήρες του κειμένου κατά τη διαδικασία αναγνώρισης και ανακτά όλες τις πιθανές αντιστοιχίσεις για μια δεδομένη λέξη. Αυτή η λειτουργία συνεπικουρεί στην έρευνα των χρηστών και βελτιώνει τις προσφερόμενες υπηρεσίες των επαγγελματιών της πληροφόρησης προς το κοινό.

Ο αθέρας της αναζήτησης θα μπορούσε να είναι η αναζήτηση με tags, η οποία όμως προϋποθέτει την επισήμανση λέξεων στα ήδη μεταγραμμένα έγγραφα. Αυτό σημαίνει ότι δίνεται η δυνατότητα να οριστούν πρόσωπα, μέρη και συντομογραφίες, καθώς επίσης να προστεθούν προσαρμοσμένες κατηγορίες ετικετών και να αναζητηθούν μεμονωμένες ετικέτες στα μεταγραμμένα έγγραφα. Σε αυτή την περίπτωση ο μεταγραφέας περιγράφει το έγγραφο ορίζοντας τα μεταδεδομένα από πεδία που υπάρχουν ήδη (τίτλος, συγγραφέας, μεταφόρτωση, είδος, γλώσσα, τύπο σεναρίου, ημερομηνία γραφής, περιγραφή) ή δημιουργεί τα δικά του πεδία.

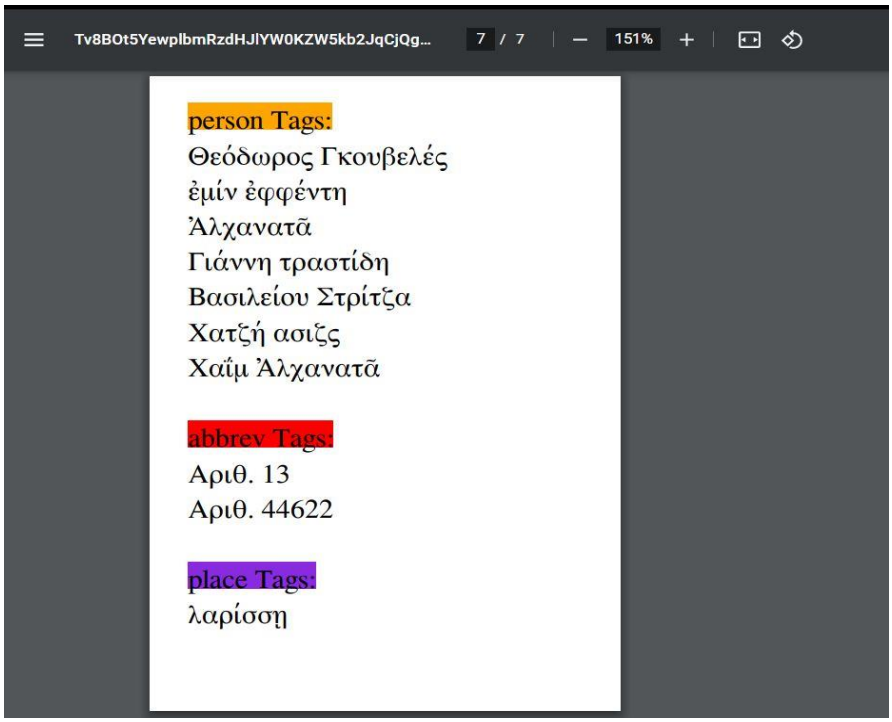
η. Εξαγωγή δεδομένων

Συμπληρωματικά, χρησιμοποιήθηκε το εργαλείο “export document” για την εξαγωγή της μεταγραφής δίπλα από το πρωτότυπο κείμενο. Υπάρχει η δυνατότητα επιλογής διαφόρων μορφών εξαγωγής όπως “Transkribus Document”, “DOCX”, “PDF”, “TXT”, “Excel” κ.ά. Επιλέχθηκε αρχείο PDF και δημιουργήθηκε τόμος με το μεταγραμμένο υλικό (Εικόνα 15). Επιπλέον, αναμφισβήτητα βοηθητική είναι η επιλογή “highlight tags”, που μας δίνει μία λίστα με τα tags που θα επιλεγούν, η οποία μπορεί να έχει θέση έντυπου ευρετηρίου της συλλογής (Εικόνα 16), κάτι που είναι πολύ σημαντικό για ερευνητές που δεν διαθέτουν ψηφιακή εξοικείωση. Τέλος, μία επιπρόσθετη λειτουργία είναι η επιλογή δημιουργίας εξωφύλλου με την οποία μπορούν να προστεθούν πληροφορίες σχετικά με τον τίτλο, τον συγγραφέα, τη γλώσσα και την ημερομηνία του εγγράφου, ή να προστεθεί μια συντακτική δήλωση για να πληροφορηθεί ο χρήστης για το πώς ακριβώς έγινε η μεταγραφή στο συγκεκριμένο έγγραφο.



Εν λαρίσσῃ σήμερον τὴν δεκάτην
Τετάρτην τοῦ μηνός Νοεμβρίου τοῦ
χιλιοστοῦ ὀκτακοσιοστοῦ ὀγδοηκο-
στοῦ πρώτου ἔτους, ἡμέραν Σάββατον ἐνώπιον
ἐμοῦ τοῦ Συμβολαιογράφου λαρίσσης καὶ τῶν
προσυπογεγραμμένων δύο μαρτύρων γνωστῶν
μοι πολιτῶν Ἑλλήνων καὶ μὴ ἐξαιρετέων κων
Χρήστου Α. Μπλάτσα ἐμπόρου καὶ Χρηστο-
Φόρου Κρανιώτου Ἠθοποιοῦ ἀμφοτέρων, κατοί-
Κων λαρίσσης ἐνεφανίσθη ὁ ἐπίσης γνωστός
Μοι καὶ μὴ ἐξαιρετέος κος Χρῆστος Νικο-
Λάου ἀμαξηλάτης κάτοικος λαρίσσης καὶ
ἐδήλωσεν ὅτι ἐδανείσθη καὶ ἔλαβε σή-
μερον κατά την ὁμολογίαν του παρά
τοῦ Ἰωάννου Παπαζαχαρίου ἀμαξηλάτον
κατοίκου λαρίσσης λίρας ὀθωμανικάς χρυ-
σᾶς τρεις καὶ ἡμίσειαν ἀριθ. 9 καὶ ½ ἤτοι δραχ-
μὰς ἐννενήκοντα ἓξ κατά τὴν τρέχουσαν
σήμερον τιμήν τῶν λιρῶν, τᾶς ὁποίας
ὑπόσχεται καὶ ὑποχρεοῦται νὰ πληρώ-
σει εἰς τὸν εἰρημένον δανειστήν του
ἄνευ τόκου μετά ἕνα ἀπό σήμερον
μῆνα εν δὲ ὑπερημερία μετά το-
κου εἴκοσι τοῖς ἑκατόν ἐτησίως
μέχρις ἐξοφλήσεως: ἐκβιαζόμενος
περί τὰς πληρωμῆς αυτῶν εν δυστροπι-
α του δυνάμει τοῦ παροντος κηρυσ-
σομένου ἐκτελεστοῦ καὶ ἄνευ κοινοποι
ήσεως ἐπιταγῆς προς πληρωμήν. ὡς
βεβαίωσιν συνετάχθη τὸ παρόν ἐντός

Εικόνα 15: Εξαγωγή μεταγραμμένου κειμένου



Εικόνα 16: Εξαγωγή λίστας Tags

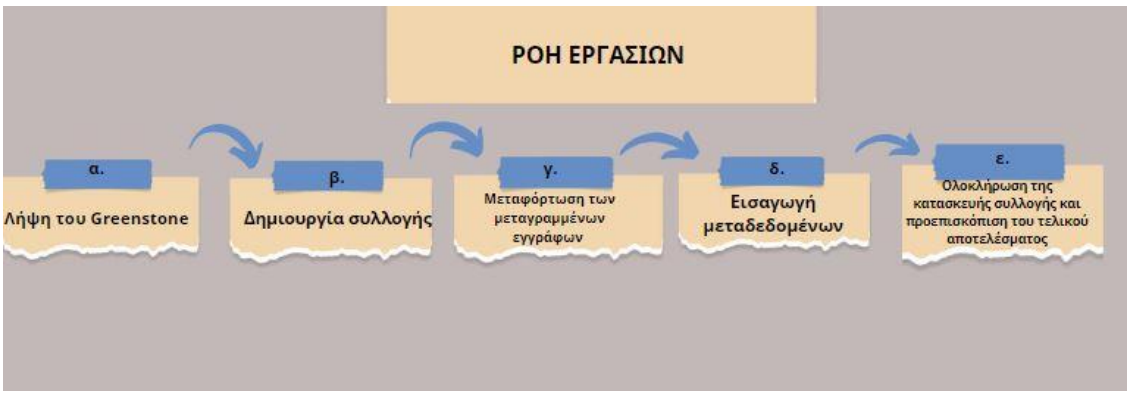
3.3.1 Ανάδειξη και διάθεση του μεταγραμμένου αρχείου προς τους ερευνητές.

Με κύριο στόχο την εξυπηρέτηση των ερευνητών, οι πολιτιστικοί φορείς, δημιουργούν ψηφιακές βιβλιοθήκες για την εύκολη πρόσβαση και τη διάθεση των ιστορικών τεκμηρίων στο κοινό. Για τις ανάγκες τις παρούσας εργασίας δημιουργήθηκε μία συλλογή αποτελούμενη από τα εξαγόμενα αρχεία που είχαν μεταγραφεί κατά την προηγούμενη διαδικασία αναγνώρισης και μεταγραφής χειρόγραφων κειμένων.

Ως εκ τούτου, για τη δημιουργία της ψηφιακής βιβλιοθήκης, επιλέχθηκε ως εργαλείο κατασκευής ψηφιακών βιβλιοθηκών, το λογισμικό Greenstone, διότι είναι ανοιχτού κώδικα, εύκολο στη χρήση και πολύγλωσσο. Επιπλέον, οι προγραμματιστές που το δημιούργησαν βραβεύτηκαν για τη συμβολή τους στην ευαισθητοποίηση των κοινωνικών επιπτώσεων της τεχνολογίας πληροφοριών, και την ανάγκη για μια ολιστική προσέγγιση στη χρήση της πληροφορικής που λαμβάνει υπόψη τις κοινωνικές επιπτώσεις, ένα πολύ σημαντικό βραβείο της Τεχνικής Επιτροπής IFIP για την Ασφάλεια και την Προστασία στα Συστήματα Επεξεργασίας Πληροφοριών (https://www.greenstone.org/factsheet). Αναπτύχθηκε και διανέμεται σε συνεργασία με την UNESCO και τη MKO Human Info στο Βέλγιο.

Το λογισμικό αυτό μπορεί να εκτελεστεί σε δικτυωμένο ή μη δικτυωμένο περιβάλλον. Οργανώνει τις πληροφορίες και τις δημοσιεύει στον παγκόσμιο ιστό με τη μορφή ενός πλήρως αναζητήσιμου ψηφιακού πόρου που βασίζεται σε μεταδεδομένα.

Για τη δημιουργία ψηφιακής βιβλιοθήκης πραγματοποιήθηκαν οι παρακάτω ενέργειες (Εικόνα 17).



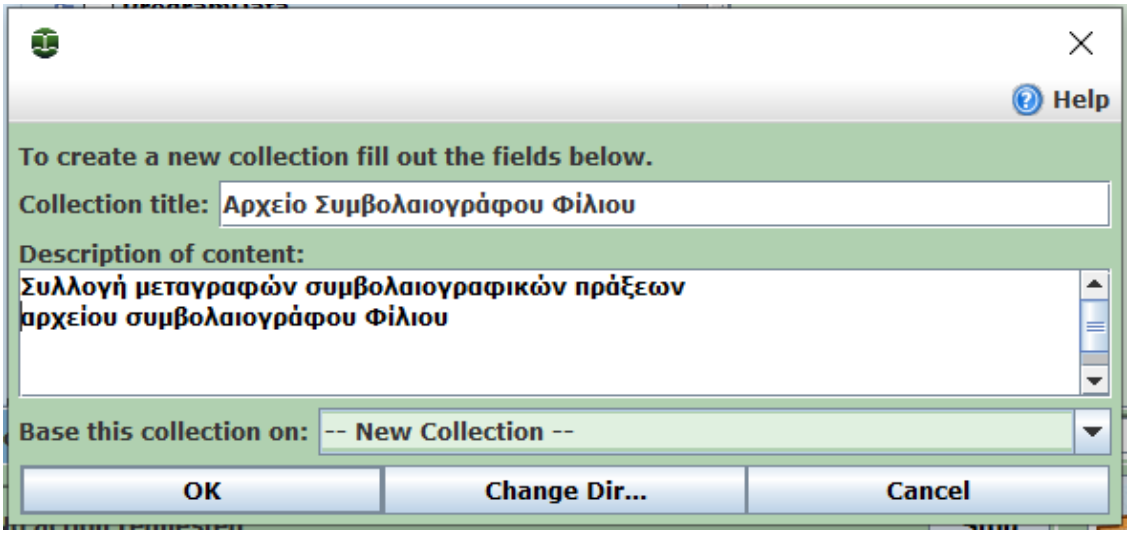
Εικόνα 17: Ροή εργασιών Greenstone

α. Λήψη του Greenstone

Η διαδικασία ξεκίνησε με τη λήψη του λογισμικού από την ιστοσελίδα https://www.greenstone.org/. Χρησιμοποιήθηκε η πρόσφατη έκδοση Greenstone 3.10. Δεν υπάρχουν ιδιαίτερες απαιτήσεις για τη λήψη και εγκατάσταση, ακολουθήθηκαν οι οδηγίες που δίνονται στην ιστοσελίδα.

β. Δημιουργία συλλογής

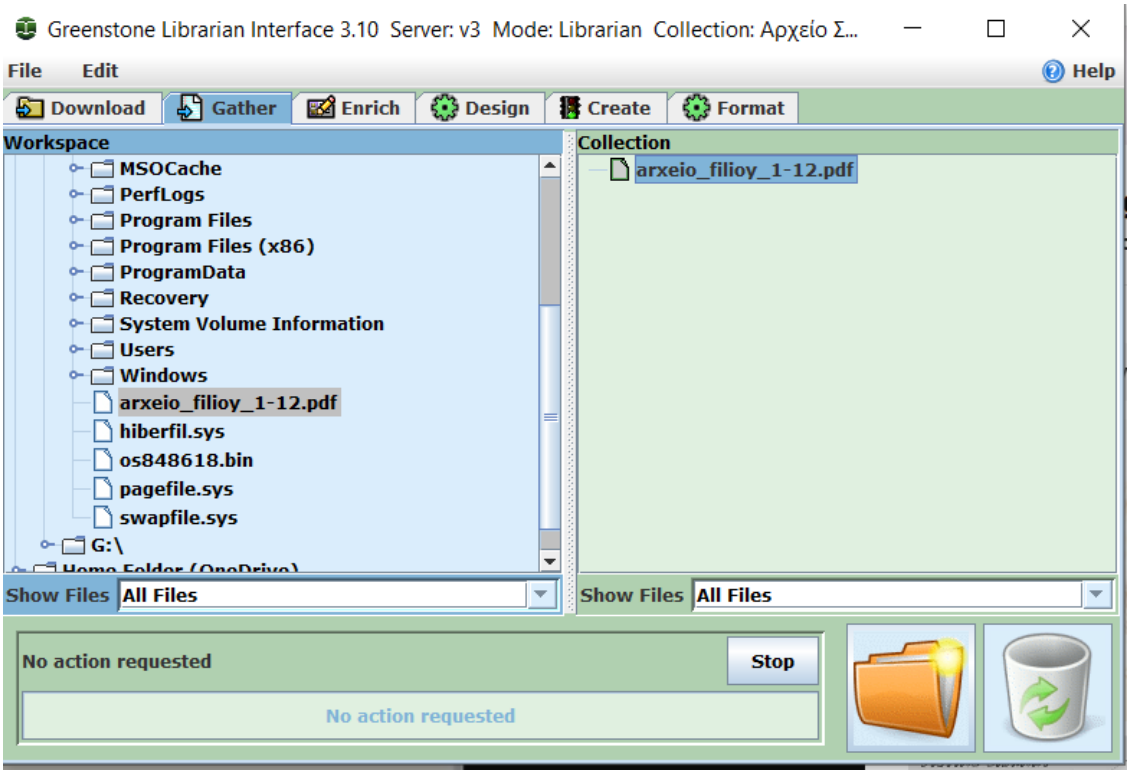
Το επόμενο βήμα ήταν η δημιουργία μιας συλλογής, δίνοντας τίτλο και περιγραφή (Εικόνα 18)



Εικόνα 18: Δημιουργία νέας συλλογής στο περιβάλλον του Greenstone.

γ. Φόρτωση του αρχείου εξαγωγής των μεταγραμμένων εγγράφων από το Transkribus

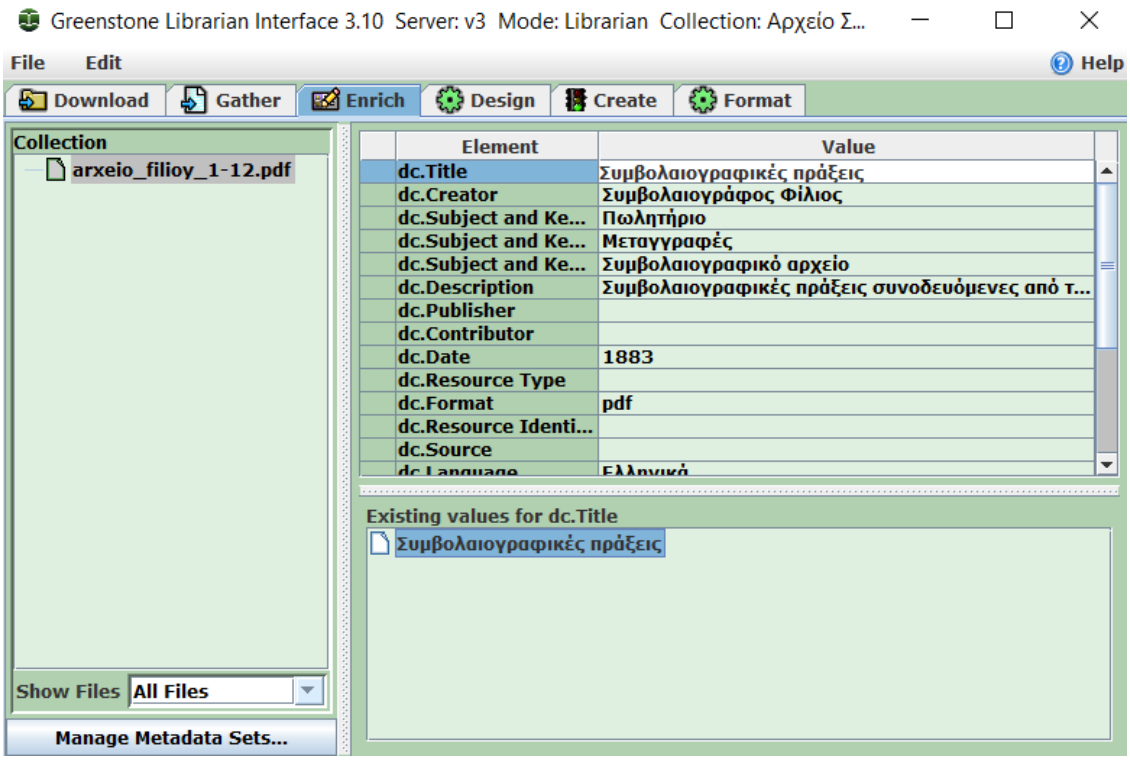
Στη συνέχεια μεταφορτώθηκε από την επιλογή open file το αρχείο με τις μεταγραφές των κειμένων που είχαν εξαχθεί από το Transkribus (Εικόνα 19).



Εικόνα 19: Φόρτωση του αρχείου εξαγωγής των μεταγραμμένων εγγράφων από το Transkribus.

δ. Εισαγωγή μεταδεδομένων

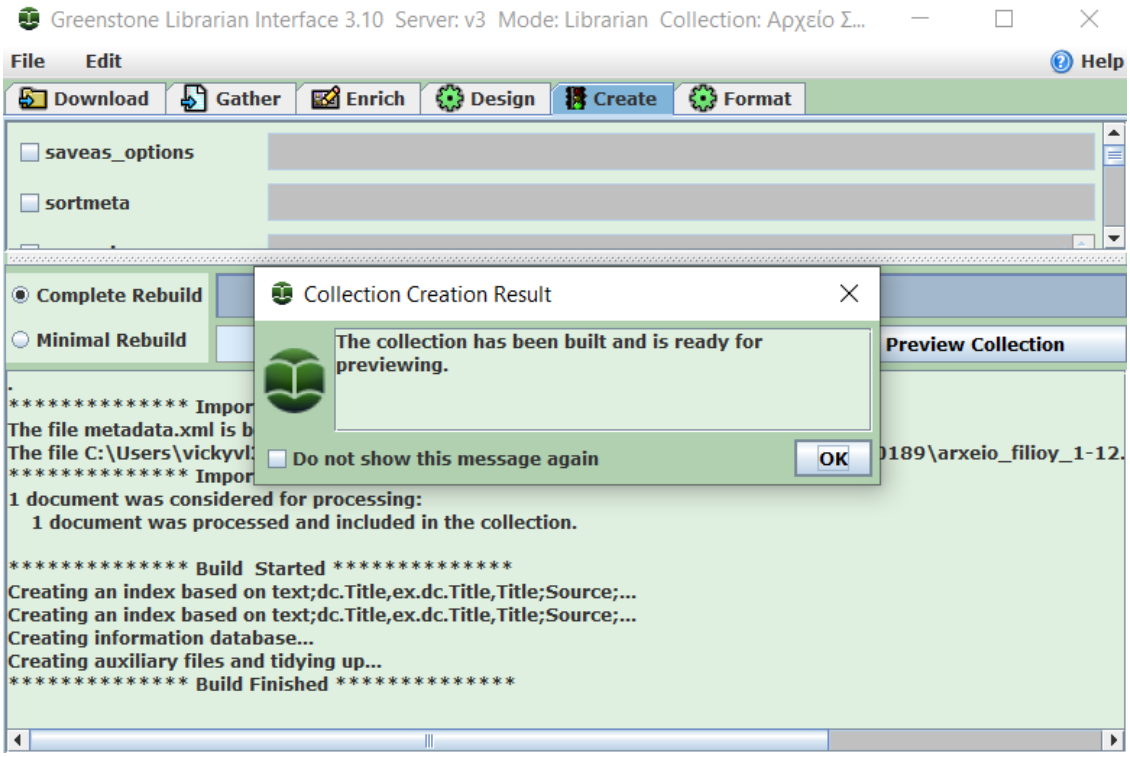
Στο σημείο αυτό περάστηκαν χειροκίνητα κάποια μεταδεδομένα για το μεταφορτωμένο αρχείο επιπλέον από τα μεταδεδομένα που περάστηκαν αυτόματα από το λογισμικό. Συγκεκριμένα ορίστηκαν λέξεις κλειδιά, δημιουργός, τίτλος, περιγραφή, χρονολογία και γλώσσα (εικ.).



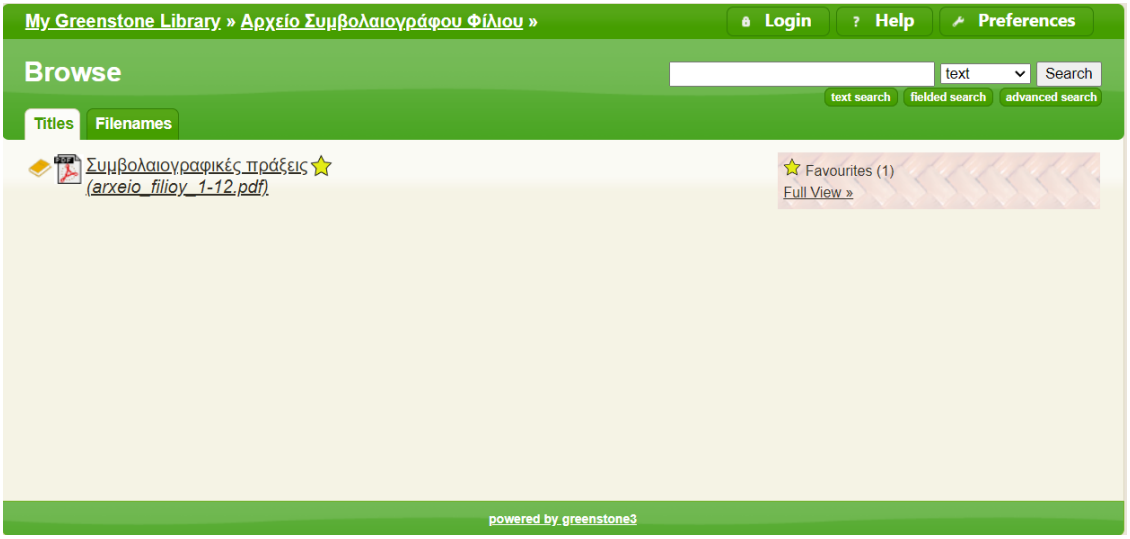
Εικόνα 20: Εισαγωγή μεταδεδομένων.

ε. Ολοκλήρωση της κατασκευής συλλογής και προεπισκόπηση του τελικού αποτελέσματος.

Τέλος, ολοκληρώθηκε η διαδικασία με την επιβεβαίωση της κατασκευής συλλογής και την προεπισκόπιση του τελικού αποτελέσματος (Εικόνα 21, Εικόνα 22).



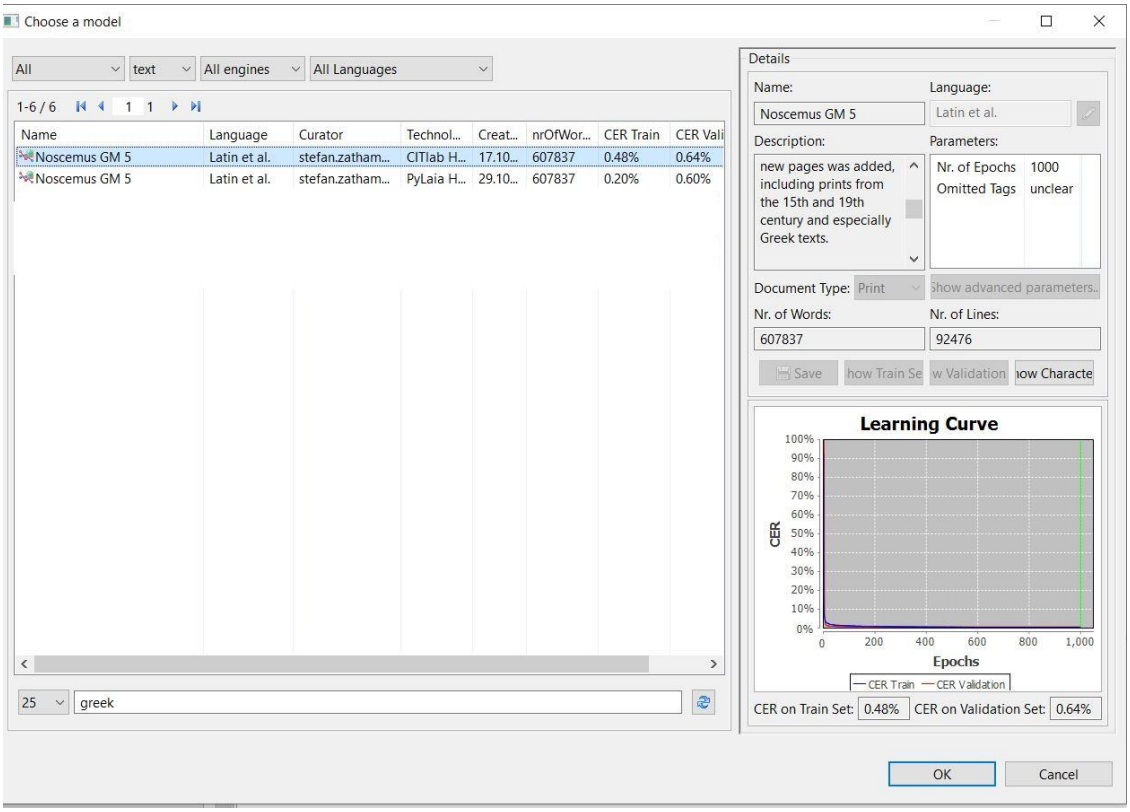
Εικόνα 21: Ολοκλήρωση της κατασκευής συλλογής



Εικόνα 22: Προεπισκόπηση της συλλογής

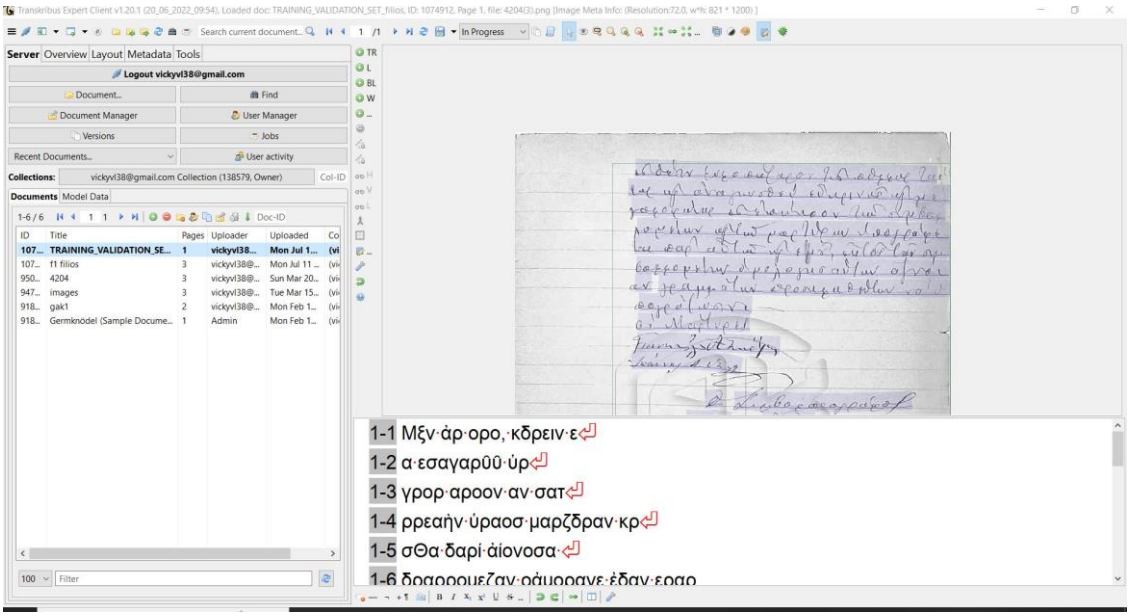
4. Αποτελέσματα

Για την εκπαίδευση ενός νέου μοντέλου αναγνώρισης ελληνικού χειρόγραφου κειμένου χρησιμοποιήθηκαν συνολικά 50 σελίδες χειρόγραφου κειμένου. Αρχικά, ζητήθηκε αυτόματη μεταγραφή δοκιμαστικά με ένα ήδη υπάρχον μοντέλο το "NOSCEMUS General Model" που βρέθηκε με αναζήτηση μοντέλων που αναγνωρίζουν την ελληνική γλώσσα (Εικόνα 23).



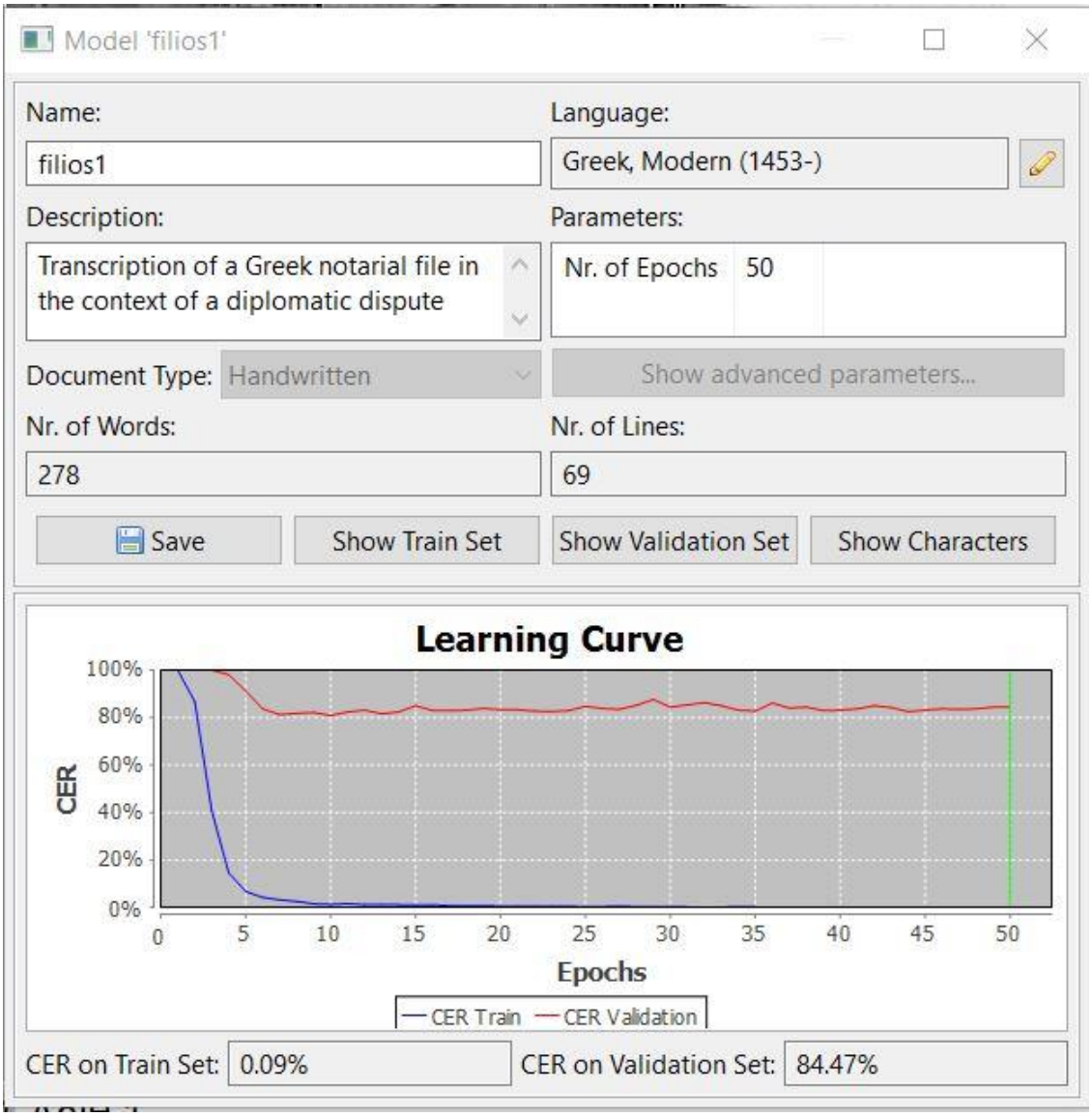
Εικόνα 23: Αναζήτηση για εύρεση μοντέλου που αναγνωρίζει ελληνική γλώσσα

Το αποτέλεσμα που έδωσε όμως δεν ήταν ικανοποιητικό (Εικόνα 24). Όπως φαίνεται και στην εικόνα το μοντέλο αναγνωρίζει κάποια σύμβολα γραμμάτων σωστά, κάποια λάθος και τα υπόλοιπα τα παραλείπει, δίνοντας έτσι κείμενο που ουσιαστικά δεν αναγνωρίζεται.



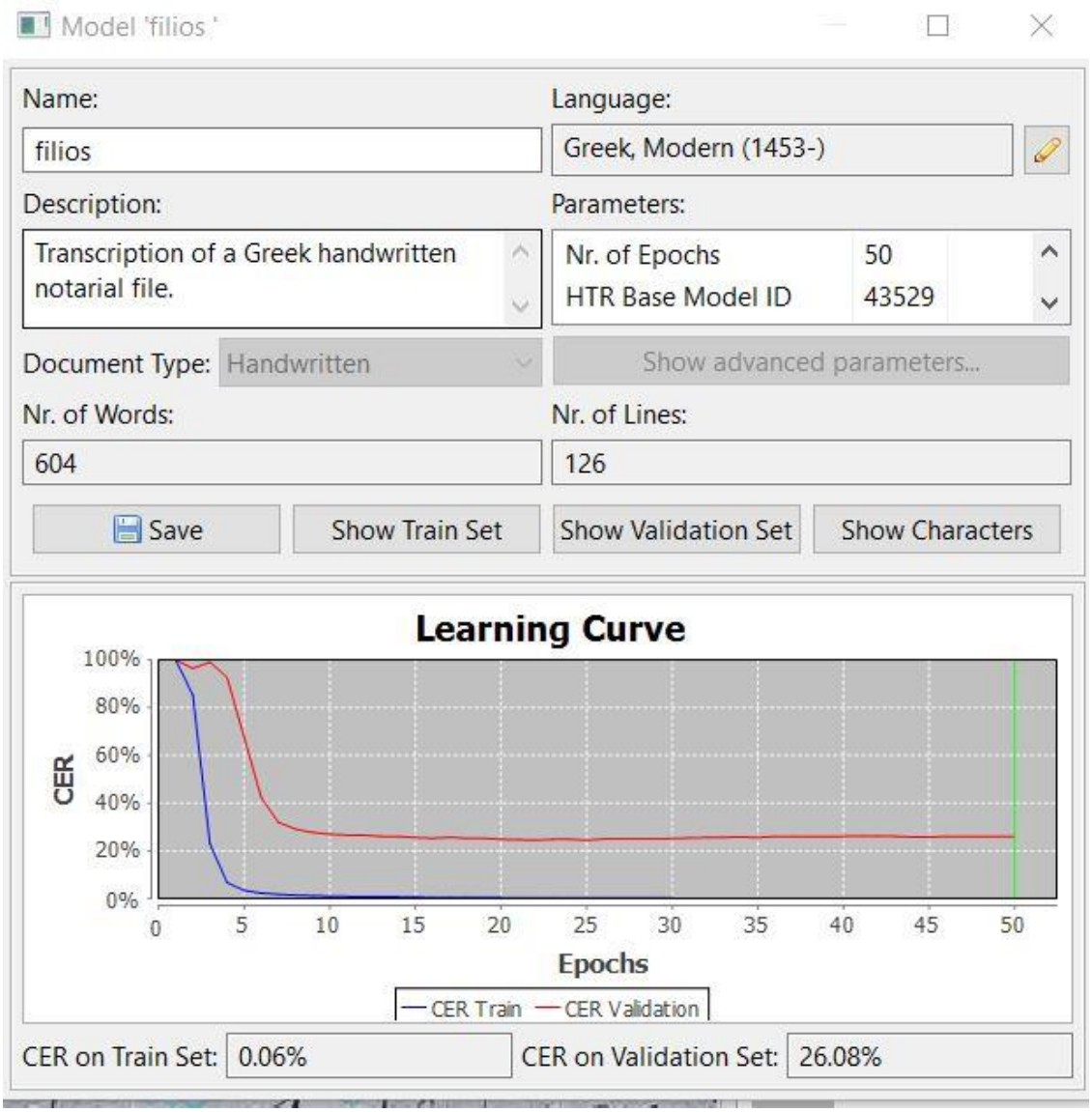
Εικόνα 24: Αποτέλεσμα μεταγραφής με χρήση υπάρχοντος μοντέλου.

Το επόμενο βήμα ήταν να διορθωθεί το κείμενο σειρά σειρά και να αποθηκευτεί ως "Ground Truth". Μετά τη διόρθωση 10 σελίδων χειρόγραφου κειμένου ζητήθηκε η εκπαίδευση ενός νέου μοντέλου με δεδομένα εκπαίδευσης και έτσι δημιουργήθηκε το μοντέλο filios1 (Εικόνα 25).



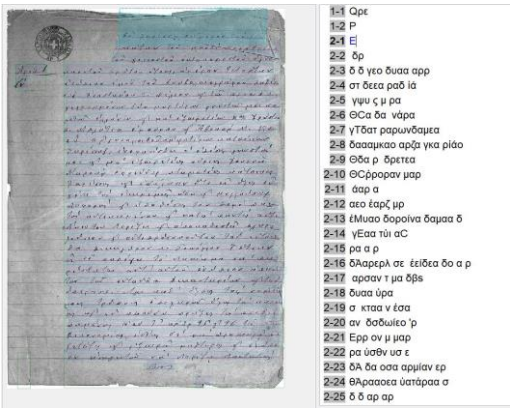
Εικόνα 25: Μοντέλο filios1.

Όπως φαίνεται στην παραπάνω εικόνα η εκπαίδευση έγινε σε 278 λέξεις, 69 σειρές και έδωσε μεταγραφή με ποσοστό σφάλματος χαρακτήρων 84,47% (CER). Η ίδια διαδικασία συνεχίστηκε και η επόμενη δοκιμή έγινε στις 20 σελίδες κειμένου, 604 λέξεις, 126 σειρές. Το νέο μοντέλο ονομάστηκε filios2 και έδωσε ποσοστό λάθους 26,08% (Εικόνα 26).

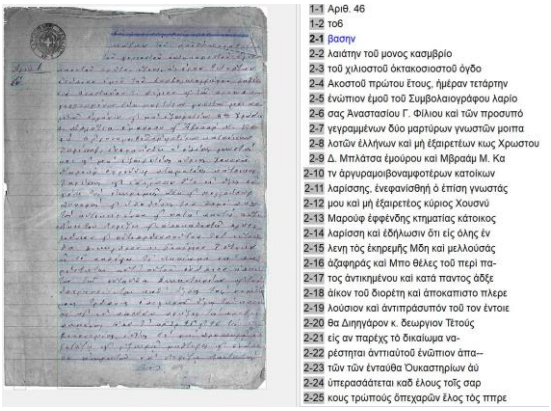


Εικόνα 26: Μοντέλο filios2.

Το αποτέλεσμα στατιστικά ήταν εκπληκτικό για τόσο μικρή αύξηση δεδομένων εκπαίδευσης και η απόδοση στο κείμενο ήταν εμφανής (Εικόνα 27, Εικόνα 28).

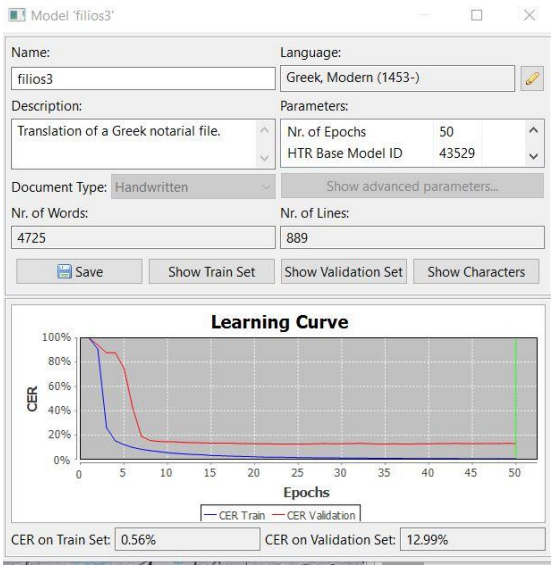


Εικόνα 27: Κείμενο με filios1.

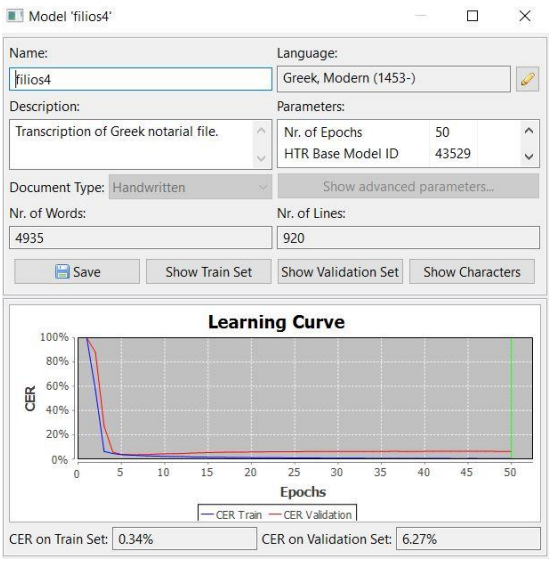


Εικόνα 28: Κείμενο με filios2.

Καθώς προχωρούσε η διαδικασία της χειρόγραφης διόρθωσης και εκπαίδευσης ακολούθησαν τα μοντέλα filios3 και filios4 (Εικόνα 29, Εικόνα 30). Το filios3 δημιουργήθηκε στις 40 σελίδες κειμένου, 4725 λέξεις, 889 σειρές και έδωσε CER 12,99%, ενώ το filios4 στις 50 σελίδες, 4935 λέξεις, 920 σειρές κειμένου και έδωσε CER 6,27%. Η απόδοσή τους στο κείμενο φαίνεται στις Εικόνα 31, Εικόνα 32.



Εικόνα 29: Μοντέλο filios3.

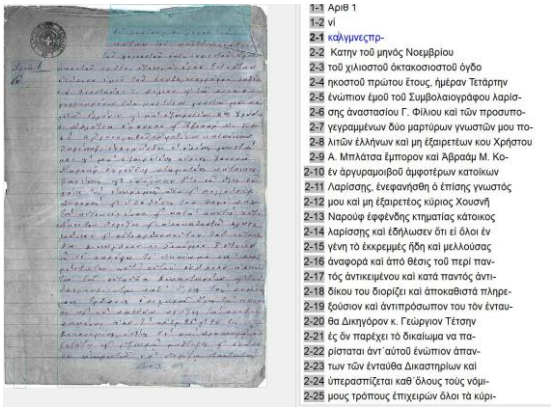


Εικόνα 30: Μοντέλο filios4.

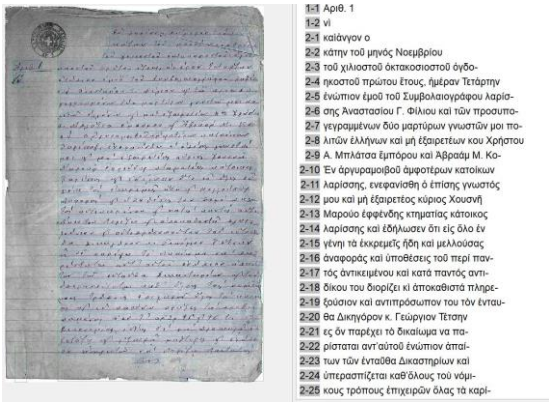
Στον παρακάτω πίνακα φαίνεται η εξέλιξη των μοντέλων σε σχέση με τις διορθωμένες σελίδες κειμένου (Πίνακας 1).

Πίνακας 2: Εξέλιξη των μοντέλων.

	Σελίδες κειμένου	Λέξεις	Σειρές κειμένου	CER
filios1	10	278	69	84,47%
filios2	20	604	126	26,08%
filios3	40	4725	889	12,99%
filios4	50	4935	920	6,27%



Εικόνα 31: Κείμενο με filios3.



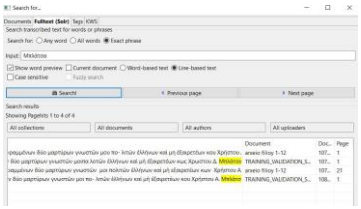
Εικόνα 32: Κείμενο με filios4.

Επίσης, η σύγκριση μεταξύ των εκδόσεων του μοντέλου μπορεί πραγματοποιηθεί και από το εργαλείο "Compare Text Versions". Έτσι, η οπτική αναπαράσταση του τι μετέγραψε σωστά ή λανθασμένα το μοντέλο HTR έδωσε το αποτέλεσμα όπως φαίνεται στην Εικόνα 33. Για την ακρίβεια, όταν έστω και ένας χαρακτήρας είναι λάθος, ολόκληρη η λέξη σημειώνεται με κόκκινο χρώμα ενώ με πράσινο χρώμα, εμφανίζεται η λέξη όπως είναι γραμμένη στη μεταγραφή Ground Truth. Στις προτάσεις χωρίς χρώμα το αναγνωρισμένο κείμενο ταυτίζεται με την Ground Truth επομένως δεν χρειάστηκαν χειροκίνητες διορθώσεις.

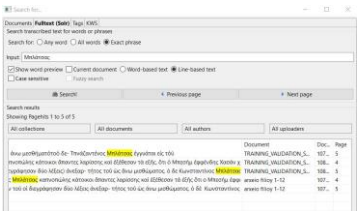
Εικόνα 33: Σύγκριση μεταξύ εκδόσεων μοντέλων.

Οι αυτόματες μεταγραφές που προέκυψαν από την παραπάνω διαδικασία εκπαίδευσης, χρησιμοποιήθηκαν για την εκτέλεση αναζητήσεων "πλήρους κειμένου". Συγκεκριμένα, ερευνήθηκε ως παράδειγμα η αναζήτηση ενός ερευνητή που μελετά την ιστορία ενός καπνοπώλη στην ευρύτερη περιοχή της Λάρισας, ονόματι “Μπλάτσας”, ο οποίος συνέταξε συμβόλαια στον συμβολαιογράφο Φίλιο. Κάνοντας αναζήτηση fulltext με το όνομα του καπνοπώλη δίνονται αποσπάσματα από τα σημεία που βρέθηκε η λέξη. Στο

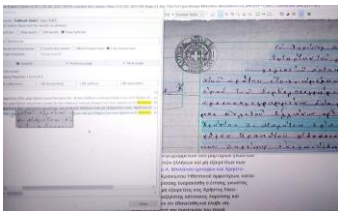
σημείο αυτό πηγαίνοντας τον κέρσορα πάνω στη φράση εμφανίζεται η εικόνα που αντιστοιχεί στο σημείο εμφάνισης της λέξης, ενώ με διπλό κλικ εμφανίζεται η σελίδα μεταγραφής στην οποία ανήκει η λέξη. Οι ίδιες δυνατότητες υπάρχουν και με την αναζήτηση KWS, με τη διαφορά ότι υπάρχει το πλεονέκτημα ανάκτησης όλων των πιθανών αντιστοιχίσεων για μία λέξη, ακόμα και αν ήταν υψηλό το ποσοστό σφάλματος της μεταγραφής. Έτσι στις παρακάτω εικόνες φαίνεται ότι στη συγκεκριμένη αναζήτηση με fulltext βρέθηκε ότι το ζητούμενο "Μπλάτσα" εμφανίζεται σε δύο σελίδες (Εικόνα 34) και "Μπλάτσας" σε άλλες δύο (Εικόνα 35). Επιπλέον, στην Εικόνα 36 βλέπουμε την εμφάνιση του στιγμιότυπου που δίνεται για κάθε φράση.



Εικόνα 34: Αποτέλεσμα αναζήτησης "Μπλάτσα".

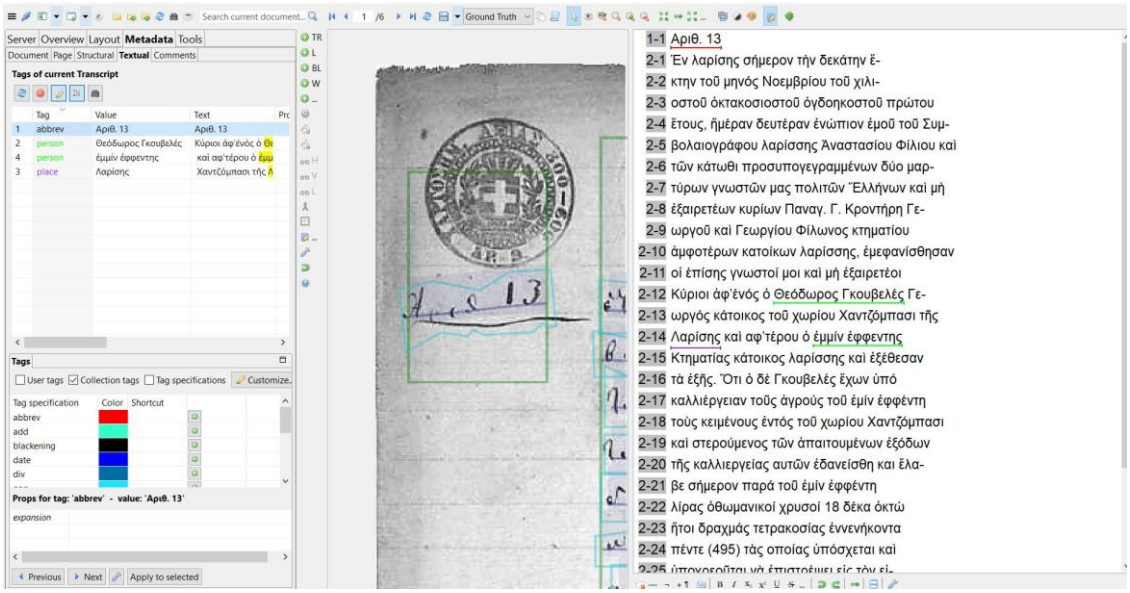


Εικόνα 35: Αποτέλεσμα αναζήτησης "Μπλάτσας".

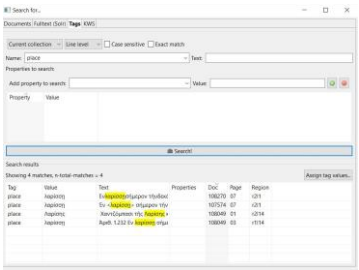


Εικόνα 36: Στιγμιότυπο αποτελέσματος αναζήτησης.

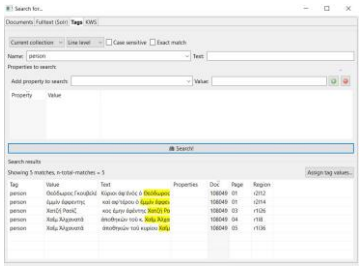
Τέλος, πραγματοποιήθηκε αναζήτηση με ετικέτες που είχαν οριστεί εκ των προτέρων σε λέξεις ενδιαφέροντος (Εικόνα 37) όπως η τοποθεσία, τα ονόματα εμπλεκομένων και ο αριθμός πράξης του συμβολαίου (Εικόνα 38, Εικόνα 39, Εικόνα 40).



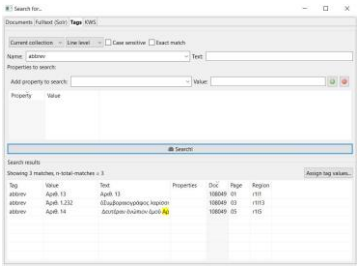
Εικόνα 37: Ορισμός tags.



Εικόνα 38: Αναζήτηση τοποθεσίας.

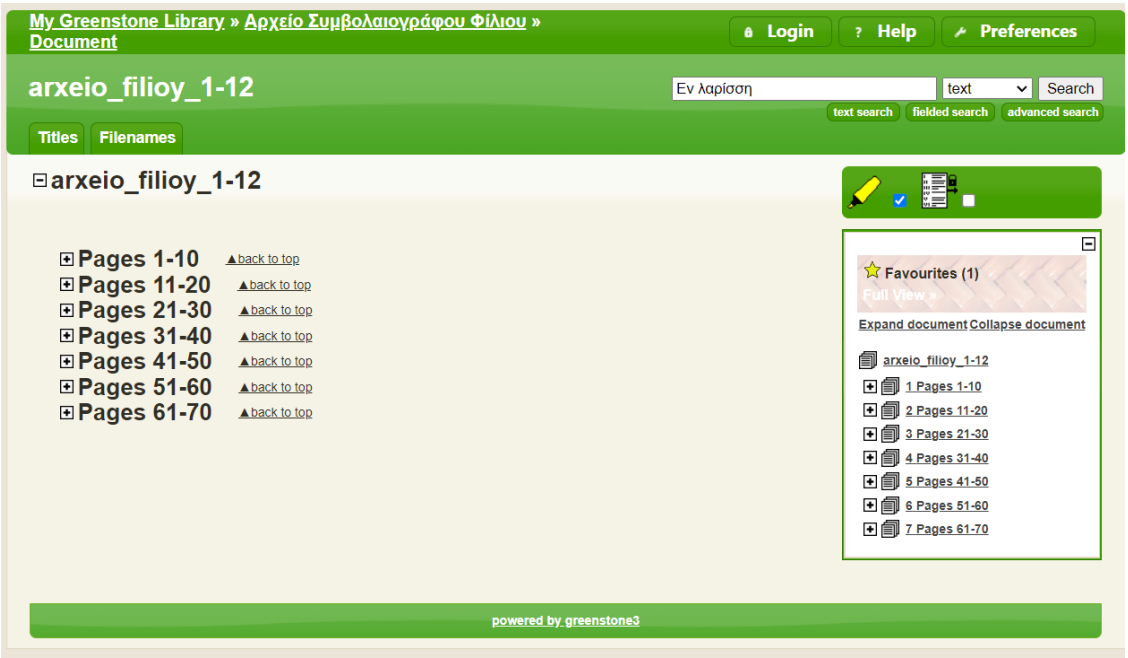


Εικόνα 39: Αναζήτηση ονομάτων εμπλεκομένων.



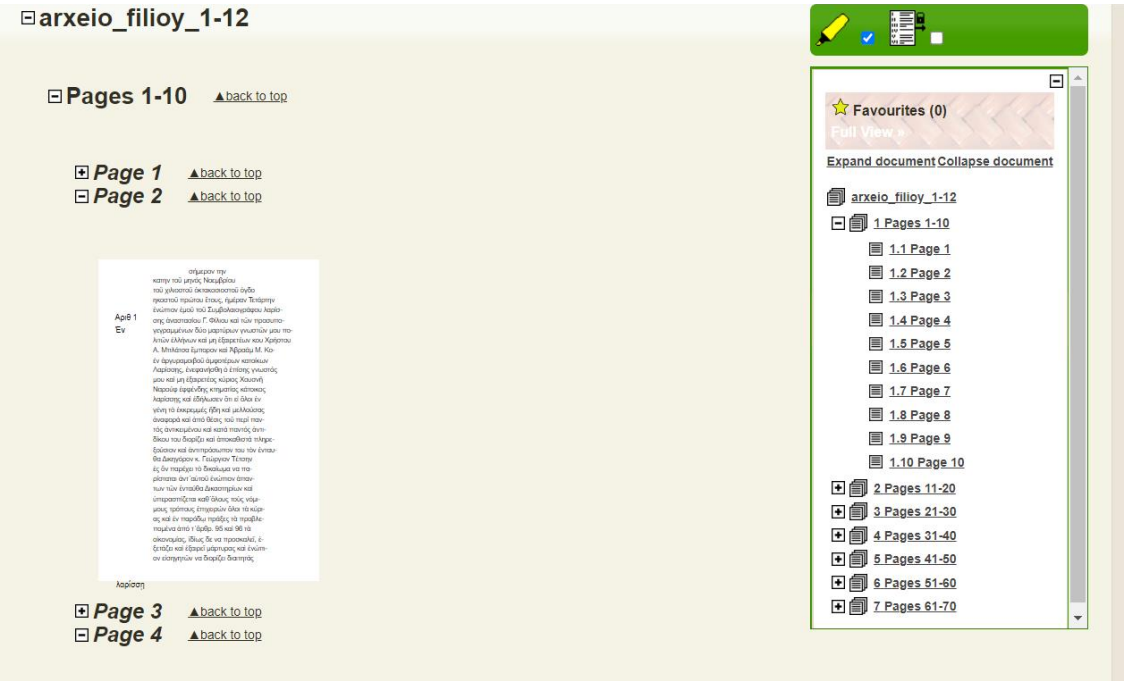
Εικόνα 40: Αναζήτηση αριθμού πράξης.

Για την καλύτερη δυνατή αξιοποίηση των αποτελεσμάτων της παραπάνω διαδικασίας, το εξαγόμενα αρχεία με τις μεταγραφές μεταφορτώθηκαν σε ψηφιακή βιβλιοθήκη που, όπως προαναφέρθηκε, δημιουργήθηκε με το εργαλείο δημιουργίας ψηφιακών βιβλιοθηκών Greenstone 3.10. Σύμφωνα με τις λέξεις κλειδιά που ορίστηκαν κατά τη διαδικασία ορισμού μεταδεδομένων στα αρχεία της συλλογής, έγινε αναζήτηση και εμφανίστηκαν όσα αρχεία είχαν οριστεί εξ αρχής ως “Subject and keywords” π.χ. “Εν Λαρίσση” (Εικόνα 41).



Εικόνα 41: Αποτέλεσμα αναζήτησης λέξεων κλειδιών στο Greenstone .

Με αυτό τον τρόπο οι βιβλιοθηκονόμοι - αρχειονόμοι μπορούν να χρησιμοποιήσουν τη θεματική αποδελτίωση που πιθανώς να είχε προηγηθεί στα χειρόγραφα και να ορίσουν θέματα στα αρχεία για να βοηθήσουν τους ερευνητές να εντοπίσουν έγγραφα πραγματοποιώντας θεματική αναζήτηση. Επιλέγοντας στη συνέχεια σελίδα - σελίδα μπορούν να διαβάσουν το κείμενο (Εικόνα 42).



Εικόνα 42: Μεταγραμμένο κείμενο στη συλλογή του Greenstone προς ανάγνωση.

5. Συζήτηση

Η παρούσα εργασία συμβάλλει στην αξιολόγηση και στην εκπαίδευση της πλατφόρμας Transkribus σε ελληνικά χειρόγραφα του 19ου αιώνα. Για τη δημιουργία αυτόματων και συνάμα επιτυχημένων μεταγραφών απαιτείται η δημιουργία ενός μοντέλου Al, το οποίο βασίζεται σε δεδομένα βασικής αλήθειας (ground truth). Κατόπιν δημιουργούνται υπομοντέλα τα οποία με τη σταδιακή τους ανάπτυξη υπέδειξαν ότι όσο πιο έγκυρα είναι τα δεδομένα βασικής αλήθειας τόσο καλύτερη είναι η απόδοση του μοντέλου. Όσο αυξάνεται ο αριθμός των λέξεων εκπαίδευσης που χρησιμοποιούνται για ένα συγκεκριμένο μοντέλο, τόσο μειώνεται το ποσοστό σφαλμάτων (CER). Ωστόσο, μπορεί να ανακύψει το ερώτημα γιατί στο μοντέλο filios 1 σε σχέση με το μοντέλο filios 2, παρόλο που το κείμενο παρέχει τα ίδια δομικά χαρακτηριστικά και δεν υπάρχει διαφοροποίηση στην ποιότητα των εικόνων το filios 1 είχε CER 84,47% ενώ το filios 2, 26,08%;

Προκύπτει ότι, το τελικό μοντέλο filios 4 περιλαμβάνει τέσσερις χιλιάδες εννιακόσιες τριάντα πέντε λέξεις, έναντι του filios 2 που περιέχει εξακόσιες τέσσερις λέξεις, όσο περισσότερα δεδομένα εκπαίδευσης παρέχονται στο σύστημα, τόσο πιο ακριβής είναι η αυτόματη μεταγραφή με μικρότερα ποσοστά λάθους.

Αναντίρρητα, απαιτείται χρόνος και υπομονή για να εκπαιδεύσει κάποιος ένα μοντέλο. Για την εκπαίδευση του αρχείου του συμβολαιογράφου Φίλιου χρειάστηκαν περίπου 136 ώρες. Μέσα σε αυτό συμπεριλαμβάνονται:

η μεταγραφή των κειμένων εκπαίδευσης, η προετοιμασία των αρχείων pdf, οι τροποποιήσεις στη διάταξη της σελίδας, οι δοκιμές των διαφορετικών μοντέλων και οι χειροκίνητες διορθώσεις (οι οποίες μπορεί να χρειαστούν σημαντικό χρόνο, ειδικά όταν το μέγεθος των αρχείων προς μεταγραφή είναι μεγάλο).

Όπως, αναφέρθηκε παραπάνω, το Transkribus δίνει τη δυνατότητα στους χρήστες, εάν το επιθυμούν, να μοιράζονται τα δεδομένα με άλλους χρήστες. Το λογισμικό διαθέτει πολλά μοντέλα διαφορετικών γλωσσών που προσφέρονται στην κοινή πλατφόρμα. Κατ' αυτόν τον τρόπο θα μπορούσε κανείς να παραλείψει τη διαδικασία κατάρτισης ενός μοντέλου, μειώνοντας το χρόνο εργασίας. Στην περίπτωση της παρούσας εκπαίδευσης χρησιμοποιήθηκε αυτόματη μεταγραφή με το υπάρχον μοντέλο που αναγνωρίζει την ελληνική γραφή το "NOSCEMUS General Model" τα αποτελέσματα όμως ήταν απογοητευτικά καθώς στην ουσία το κείμενο δεν αναγνωρίζεται. Αυτό έγκειται στο γεγονός της ιδιομορφίας της ελληνικής γραφής: οι διαφορετικοί τύποι της ορθογραφίας ανάλογα με την εποχή του χειρογράφου, το πολυτονικό σύστημα, οι βραχυγραφίες και πολλά άλλα που συντελούν στην δύσκολη αποδόμησή της.

Η ανάλυση της πραγματικής απόδοσης των μοντέλων HTR απέδειξε ότι, αν και τα μοντέλα απέχουν πολύ από το να δώσουν αποτελέσματα χωρίς σφάλματα, είναι ιδιαίτερα χρηστικά και συνάμα αποτελούν για τον ερευνητή ένα επωφελές εργαλείο εξοικονόμησης χρόνου και χρήματος. Ωστόσο, αν και το Transkribus άλλαξε σε επιχειρηματικό μοντέλο και δίνει τη δυνατότητα στον χρήστη να κάνει δωρεάν μεταγραφή περίπου τετρακοσίων σελίδων και κατόπιν θα χρεώνεται για κάθε επιπλέον σελίδα, παραταύτα το κόστος και ο χρόνος σε περίπτωση χειροκίνητης μεταγραφής θα είναι ασύγκριτα μεγαλύτερος (https://readcoop.eu/).

Το Transkribus επιπροσθέτως, ανοίγει νέα και άγνωστα ερευνητικά πεδία καθώς, οι μεταγραφές κάθε λέξης χωριστά αποθηκεύονται στο λογισμικό και ακόμη και στην περίπτωση όπου η τελική μεταγραφή είναι εσφαλμένη, ή και αν υπάρχουν λέξεις με ορθογραφική παραλλαγή, οι σωστές μεταγραφές μπορούν να βρεθούν κάνοντας χρήση της λειτουργίας εντοπισμού λέξεων κλειδιών (KWS).

Σύμφωνα με το Read-Coop από το 2021 η κοινότητα του Transkribus έχει αυξηθεί σε 50.000 περίπου εγγεγραμμένους χρήστες από όλο τον κόσμο, με έμφαση κυρίως στις δυτικές λατινικές γλώσσες. Ωστόσο αυξητικός είναι και ο αριθμός των χρηστών και στις κυριλλικές γραφές (https://readcoop.eu/).

Ο Rabus, (2022), όταν χρησιμοποίησε το Transkribus σε γλαγολιτικά χειρόγραφα που ήταν γραμμένα από διαφορετικούς γραφείς, καθιστώντας την πρόκληση ακόμη μεγαλύτερη, έκρινε ότι η συνολική ποιότητα μεταγραφής υπήρξε ικανοποιητική. Το μοντέλο εκπαίδευσης "Handwritten Glagolitic", που δημιουργήθηκε, επέδειξε χαμηλό ποσοστό λάθους (CER) που αναλογεί περίπου ότι στα εκατό γράμματα, έξι περίπου μεταγράφονται λανθασμένα. Επίσης, δημιουργήθηκε και ένα ξεχωριστό μοντέλο για έντυπα κείμενα δίνοντας εξίσου ικανοποιητικά αποτελέσματα. Ο Rabus πιστεύει ότι αν και δεν δίνεται η δυνατότητα να παραχθεί μια μεταγραφή με μηδενικό CER από το λογισμικό, ωστόσο το Transkribus αποτελεί μια χρηστική πλατφόρμα. Εστιάζει περισσότερο στην εξοικονόμηση χρόνου και κόστους και εξάρει το εργαλείο KWS, το οποίο το θεωρεί καινοτόμο και επωφελές για τον ερευνητή. Η αξιολόγηση του Rabus σε μία γραφή, που δεν έχει χρησιμοποιηθεί ευρέως σε μοντέλα εκπαίδευσης του Transkribus, όπως και η ελληνική, έδωσε ικανοποιητικά ποσοστά λάθους. Παρόλο που έπρεπε να εκπαιδευτεί εξαρχής ένα μοντέλο, καθώς τα υπάρχοντα δημόσια μοντέλα της πλατφόρμας απέδιδαν ανεπαρκή αποτελέσματα, η τελική αποτίμηση εξοικονόμησης χρόνου και στις δύο περιπτώσεις ήταν αξιόλογες.

Η Milioni στη διατριβή της σχετικά με το Transkribus ως εργαλείο για βιβλιοθήκες, αρχεία και μελετητές, έκανε μια διττή αξιολόγηση της πλατφόρμας. Διερεύνησε αρχικά μέσω ερωτηματολογίων, εάν γνωρίζουν την πλατφόρμα Transkribus και σε ποιο βαθμό θα τη χρησιμοποιούσαν τα ιδρύματα πολιτιστικής κληρονομιάς, όπως βιβλιοθήκες, αρχεία κ.α. Επίσης, δημιούργησε τα δικά της μοντέλα Al σε λατινικά κείμενα με διαφορετική γραφή. Στο πρώτο μέρος, αναφέρει ότι τα περισσότερα ιδρύματα γνώριζαν τη συγκεκριμένη πλατφόρμα και περίπου τα μισά θα την αξιοποιούσαν αν αποδεικνυόταν αξιόπιστη και αποτελεσματική. Σχετικά, με την αξιολόγηση των μοντέλων της Al συμπέρανε ότι ήταν μια εκτεταμένη εργασία. Εν ολίγοις καταλήγει ότι, η πλατφόρμα έχει αποδειχθεί ευεργετική για μεταγραφή υλικού από τον 18ο αιώνα και μετά, αλλά δεν μπορεί να ανταποκριθεί επαρκώς σε χειρόγραφα του 15ου αιώνα χωρίς τις γνώσεις ενός έμπειρου ερευνητή ή παλαιογράφου. Ωστόσο, εκφράζει την απορία για το ποιες μεταγραφές θεωρούνται επαρκείς και αν τα χειρόγραφα ψηφιοποιημένα έγγραφα είναι σημαντικό να συνοδεύονται από μεταγραφές έστω και μη "επαρκείς" (Milioni, 2020).

Η αξιολόγηση της Milioni όσον αφορά σε κείμενα του 18ου αιώνα και μετά συνάδει με τα αποτελέσματα της έρευνας των χειρογράφων του αρχείου του Φίλιου που χρησιμοποιήθηκε στην παρούσα εργασία. Μπορεί τα ποσοστά λάθους και στις δύο περιπτώσεις να μην ήταν μηδενικά, ωστόσο κρίθηκαν ικανοποιητικά.

Η Milioni (2020, σ.12) αναφέρει επίσης, ότι το HTR μπορεί να επεξεργαστεί ένα τύπο ιστορικών εγγράφων που καλούνται "κρυπτογραφήματα". Αυτά τα ιστορικά κείμενα περιέχουν κρυπτογραφημένες πληροφορίες και εκτός από μεταγραφή, πρέπει πρώτα να αποκρυπτογραφηθούν. Η δυσκολία μεταγραφής τους έγκειται στο γεγονός ότι αποτελούνται από απόκρυφα σύμβολα και επομένως τα λεξικά δεν μπορούν βοηθήσουν. Έχει αναπτυχθεί μία αυτόματη μέθοδος που βασίζεται σε επεξεργασία της εικόνας, έτσι ώστε το σύστημα αναγνωρίζει τα σύμβολα και να τα μεταποιεί σε μορφή αναγνωρίσιμη από τον υπολογιστή.

Ο Schlagdenhauffen μεταγράφοντας το ημερολόγιο του νομικού Eugene Wilhelm είχε να αντιμετωπίσει δύο βασικές προκλήσεις - το χρόνο διαδικασίας γραφής των ημερολογίων (66 έτη), το οποίο οδηγεί σε πολλές παραλλαγές της γραφής και τη χρήση δύο αλφαβήτων του ελληνικού και ρωμαϊκού. Μάλιστα, η διαδικασία μεταγραφής πραγματοποιήθηκε σε συνεργασία με τους μαθητές του. Αξιολογεί λοιπόν το λογισμικό τόσο για την χρηστικότητά του, όσο και στο πλαίσιο διδασκαλίας. Αναφέρει ορισμένους περιορισμούς που παρατηρήθηκαν, όπως ότι η αυτοματοποιημένη μεταγραφή πρέπει να υποβάλλεται εκ των υστέρων χειροκίνητα για διορθώσεις, ωστόσο τα αποτελέσματα είναι θετικά καθώς τονίζει ότι η αυτόματη μεταγραφή είναι ταχύτερη. Εστιάζει στη χρησιμότητα του εργαλείου KWS, το οποίο αποτελεί το κυριότερο χαρακτηριστικό του Transkribus. Όσον αφορά τη διεπαφή με το χρήστη, αυτός και οι μαθητές του δεν αντιμετώπισαν κάποια ιδιαίτερη δυσκολία και την χαρακτηρίζει απλή. Επισημαίνει ότι το Transkribus είναι αποτελεσματικό μόνο για μεσαία ή μεγάλα χειρόγραφα σώματα, καθώς για μικρότερα σώματα το κόστος είναι μεγάλο. Τέλος, υποστηρίζει ότι όσο περισσότερο χρησιμοποιείται το λογισμικό, τόσο πιο εύκολη γίνεται η αναγνώριση ποικιλίας εγγράφων μέσα στο Transkribus (Schlagdenhauffen, 2020). Η αξιολόγηση του Schlagdenhauffen αναφέρεται στις χειροκίνητες διορθώσεις, οι οποίες, όπως παρατηρήθηκε και κατά την επεξεργασία των μοντέλων της παρούσας έρευνας, αναλώνουν σημαντικό χρόνο, μολαταύτα η ανταπόκριση της πλατφόρμας στα ζητούμενα κρίθηκε ικανοποιητική και στις δύο περιπτώσεις με χαμηλά ποσοστά CER, όπως επίσης θετικά αξιολογήθηκε η συμβολή του εργαλείου KWS.

Η έρευνα των Πλατάνου, Παυλόπουλος, Παπαϊωάννου, επικεντρώθηκε στην προσπάθεια αυτόματης μεταγραφής αμιγώς ελληνικών παλαιογραφικών χειρογράφων. Η έρευνα παρουσιάζει τις προκλήσεις που αντιμετώπισε η τεχνολογία HTR με χρήση του λογισμικού Transkribus 1.15.1 σε κείμενα που χρονολογούνται από τον 10ο έως τον 16ο αιώνα. Παρατηρήθηκε μια ανισορροπία στο CER ανά αιώνες. Τα χειρόγραφα που κυμαίνονται από το 10ο έως τον 13ο αιώνα, το CER ήταν μικρότερο σε σχέση με τα χειρόγραφα που κυμαίνονται από τον 14ο έως τον 16ο αιώνα. Μάλιστα, τα χειρόγραφα του 16ου αιώνα παρουσίασαν το μεγαλύτερο ποσοστό λάθους, το οποίο μπορεί να οφείλεται ίσως στην έντονη παρουσία γραμμικού ύφους. Ωστόσο, η έρευνά τους θα συνεχιστεί επεκτείνοντας το σύνολο δεδομένων και θα διευρύνουν την γενίκευση των ευρημάτων τους σε άλλες γλώσσες, ώστε να διακρίνουν αν το υψηλό CER εξαρτάται από τον αιώνα στον οποίο ανήκει το χειρόγραφο (Platanou, Pavlopoulos & Papaioannou, 2022).

Αναμφίβολα μια αυτόματη μεταγραφή αποτελεί μια εύκολη εναλλακτική λύση, ωστόσο κάποιος θα μπορούσε να ισχυριστεί ότι δεν δίνει τη δυνατότητα στον ερευνητή να δει

την πραγματική κατάσταση του κειμένου, τις τροποποιήσεις, τα λάθη που έγιναν από το συγγραφέα κ.α. Παραταύτα, ως αντίλογο σε αυτόν το ισχυρισμό θα μπορούσε κανείς να προβάλλει το γεγονός, ότι πιθανότατα σε πολλές από αυτές τις πηγές, ο ερευνητής, δεν θα μπορούσε να έχει πρόσβαση, είτε λόγω απόστασης, είτε λόγω πρόκλησης φθοράς του χειρογράφου, μα περισσότερο λόγω έλλειψης γνώσεων αποδόμησής του. Το Transkribus εξαλείφει αποστάσεις, προστατεύει τα τεκμήρια και βοηθά, όπως αποδείχθηκε, στην κατανόηση του περιεχομένου του.

Ως πρόσθετο εργαλείο το Transkribus, σύμφωνα με όσα αναφέρθηκαν παραπάνω, είναι χρηστικό καθώς σε ένα μεγάλο όγκο υλικού, βοηθά τον ερευνητή να αναζητήσει - σε ελάχιστο χρόνο - κάτι που τον ενδιαφέρει, ή να αντιγράψει ένα μεγάλο απόσπασμα αντί να το κάνει χειρόγραφα, ή να κάνει μερικές αναλύσεις της συχνότητας σημαντικών λέξεων που θα τον εξυπηρετήσουν στην έρευνά του.

Επιπροσθέτως, το διαδικτυακό περιβάλλον του Greenstone υπήρξε αποτελεσματικό και ιδιαίτερα βοηθητικό, καθώς με την ανάπτυξη της συλλογής των αρχείων του συμβολαιογράφου Φίλιου Α., που δημιουργήθηκε, παρέχει μια ολοκληρωμένη προεπισκόπηση του πονήματος στη διάθεση του κοινού. Συνοπτικά θα μπορούσε να ειπωθεί ότι οι νέες αυτές τεχνικές ξαναζωντανεύουν το χειρόγραφο. Η μεταγραφή τους αποτελεί μια σημαντική προσφορά καθώς συντελούν στην πνευματική ανάταση και μεταφέρουν αξίες και ιδανικά μιας άλλης εποχής.

Συμπεράσματα

Η αξιοποίηση του ποικιλόμορφου αρχειακού υλικού διαφυλάσσεται στα αρχεία, τα πλεονεκτήματα που προσφέρει η ψηφιοποίησή τους, αποτελούν ενισχυτικό παράγοντα για την ανανέωσή τους και φύλαξή τους. Ο ρόλος και η σημασία των αρχείων για τον πολιτισμό και οι πολυπληθείς χρήσεις τους αναδεικνύονται ακόμη περισσότερο με τη χρήση νέων τεχνολογιών. Η ενσωμάτωση των σύγχρονων τεχνολογιών στους αρχειακούς φορείς και ιδρύματα της γνώσης, αποτελεί εφαλτήριο για την επίτευξη μιας βιωματικής προσέγγισης των πρωτογενών πηγών, δίνοντας παράλληλα τη δυνατότητα να αποκωδικοποιηθεί το πολιτικό, ιστορικό και κοινωνικό γίγνεσθαι μιας εποχής, συνεπικουρώντας στην καλλιέργεια της ιστορικής και αρχειακής συνείδησης.

Οι προβληματισμοί σχετικά με την εκπαίδευση του λογισμικού είναι αναπόσπαστα συνδεδεμένοι με την ευχρηστία του. Η διαχείρισή του είναι απλή και αναμφίβολα αποσκοπεί στη "διηνεκή" διαφύλαξη των χειρογράφων και τη νοηματοδότησή τους. Οι στόχοι και τα κίνητρα της έρευνας αναζωογονούν και επανακαθορίζουν την οπτική προς τις νέες τεχνολογίες που αναπτύσσονται. Διευρύνεται ο προβληματισμός για τη χρήση του, εξομαλύνοντας ταυτόχρονα τις αποστάσεις που προκαλεί η ενασχόληση ανθρώπων από διαφορετικά επιστημονικά πεδία έναντι του επιστημονικού πεδίου της πληροφορικής. Παράλληλα, φαίνεται ότι αυτές οι νέες τεχνολογικές προσεγγίσεις προεκτείνουν τις δυνατότητες των ερευνητών, καθώς μεταβάλλουν τις συνθήκες αναζήτησης που έως τώρα καθίστανται χρονοβόρες, δημιουργώντας ένα εύχρηστο και συνάμα βοηθητικό μέσο για την έρευνά τους.

Επιπροσθέτως, θα πρέπει να τονιστεί ότι η χρήση του Transkribus δεν ελέγχει την ορθότητα των αφηγήσεων των χειρογράφων αλλά συνεπικουρεί στην ανάδειξη της πολιτισμική τους αφήγησης και της διαχρονικότητά τους. Θα μπορούσε να λεχθεί ότι λειτουργεί συμπληρωματικά και ανατρεπτικά στην ανάγνωση και αποδόμησή τους, καθώς και στους τρόπους προσέγγισης και επεξεργασίας τους. Επίσης, δίνει τη δυνατότητα σε άτομα ή ομάδες, διαφορετικών εθνοτήτων, με κοινά όμως πολιτισμικά ενδιαφέροντα, να "συναντηθούν" και να προσφέρουν ένα ενιαίο χώρο διαμοιρασμού πολιτισμικών πηγών, εμπλουτίζοντας αέναα τις πηγές πληροφόρησης και γνώσης και παράλληλα την πολυφωνία των πολιτισμών.

Παρουσιάζοντας τέσσερις σύγχρονες αντιπροσωπευτικές έρευνες σχετικά με τη χρήση της πλατφόρμας, σε συνδυασμό με την έρευνα της παρούσας διπλωματικής και μελετώντας όλα τα προτερήματα και μειονεκτήματα που έχουν αναφερθεί παραπάνω, γίνεται κατανοητό ότι τα μοντέλα αναγνώρισης χειρόγραφου κειμένου με τη χρήση του λογισμικού HTR της πλατφόρμας του Transkribus, αποτελούν σημαντικούς αρωγούς σε κάθε είδους έρευνα και στοχεύουν να παρέχουν, όσο είναι δυνατόν, ένα εύκολο περιβάλλον χρήσης, βελτιώνοντας και ανανεώνοντας συνεχώς τα χαρακτηριστικά τους. Συνδυάζοντας τα παραπάνω με τη σταδιακή εξοικείωση των χρηστών στον τρόπο λειτουργίας του, εξαλείφεται σιγά σιγά οποιαδήποτε τυχόν δυσκολία προκύψει κατά τη διάρκεια της χρήσης του. Επίσης, συμπληρωματικά διαδικτυακά περιβάλλοντα, όπως το Greenstone που χρησιμοποιήθηκε στο παρόν πόνημα συμβάλλουν σε μια ολοκληρωμένη συγκεφαλαίωση της έρευνας που πραγματοποιήθηκε.

Ο συνεχής εκσυγχρονισμός της εφαρμογής του Transkribus, διευκολύνει τον εντοπισμό, τη διάσωση, τη διατήρηση και την αξιοποίηση των χειρογράφων, επιτυγχάνοντας με αυτόν τον τρόπο, με γεωμετρική πρόοδο, την απρόσκοπτη προσβασιμότητα, αναπτύσσοντας παράλληλα τη διαλογική εποικοδομητική σχέση μεταξύ των πολιτισμών. Η επιβεβαίωση αυτής της διαπίστωσης προκύπτει από την όλο και συχνότερη συμπόρευση και εμπλουτισμό της κοινότητας του Transkribus ποικίλων φορέων πολιτισμού διεθνώς. Η ενσωμάτωση της παρούσας μελέτης στην διεθνή κοινότητα, μπορεί να συμβάλλει θετικά για την στοχοθέτηση και ανάπτυξη της πλατφόρμας σε σχέση με την ελληνική γλώσσα. Η αντιπαραβολή άλλων μελλοντικών ερευνών, με διαφορετικές συνιστώσες, όπως για παράδειγμα εκπαίδευση με ελληνική γραφή άλλου αιώνα, μπορεί να οδηγήσουν κατά περίπτωση σε ποικίλα συμπεράσματα, αποκαθιστώντας παραταύτα το χειρόγραφο ως πραγματική πηγή πληροφόρησης.

Τέλος, αναντίρρητα, η λειτουργία και η χρήση των τεχνολογιών αυτόματης μεταγραφής χειρογράφων εξελίσσονται. Τα συμπεράσματα σχετικά με την ανταπόκριση της πλατφόρμας στα χειρόγραφα του αρχείου Φίλου Α. ήταν θετικά. Με την διαρκή εξέλιξη της τεχνολογίας, οι δυνατότητές του μπορούν να διευρυνθούν αποτυπώνοντας τις εξαγόμενες μεταγραφές με πληρέστερη ακρίβεια. Παραταύτα, αναμφίβολα η χρήση του μπορεί να συμβάλλει μόνο θετικά σε όλους τους οργανισμούς πολιτιστικής κληρονομιάς.

Βιβλιογραφία

Βλαχάβας, Ι., Κεφαλάς, Π., Βασιλειάδης, Ν., Κόκκορας, Φ. & Σακελλαρίου, Η (2020). *Τεχνητή νοημοσύνη*. Θεσσαλονίκη: Πανεπιστήμιο Μακεδονίας.

Γεωργούλη, Κ. (c2015). *Τεχνητή νοημοσύνη: μια εισαγωγική προσέγγιση*. Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Ανακτήθηκε 18 Σεπτεμβρίου, 2022, από: http://repfiles.kallipos.gr/html_books/93/04a-main.html#_idTextAnchor075.

Δημοκρίτειο Πανεπιστήμιο Θράκης. Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών. ([χ.χ]). *mdoc-ts*. Ανακτήθηκε 11 Σεπτεμβρίου, 2022, από https://mdoc-ts.ee.duth.gr/partners/index.html/.

Διαμαντάρας, Κ. & Μπότσης, Δ. (2019). *Μηχανική μάθηση*. Θεσσαλονίκη: Κλειδάριθμος.

Εθνικό Κέντρο Τεκμηρίωσης και Ηλεκτρονικού Περιεχομένου [ΕΚΤ] (2020). *Καλές Πρακτικές και Προδιαγραφές διαλειτουργικότητας και ποιότητας για τη διαδικτυακή διάθεση ψηφιακού πολιτιστικού περιεχομένου*. Αθήνα: ΕΚΤ.

Ελλάδα. Υπουργείο Ψηφιακής Διακυβέρνησης. (2021). *Βίβλος Ψηφιακού Μετασχηματισμού 2020-2025*. Αθήνα: Ελληνική Δημοκρατία. Υπουργείο Ψηφιακής Διακυβέρνησης.

Καπιδάκης, Σ. (2010). *Εισαγωγή στις ψηφιακές βιβλιοθήκες*. Αθήνα: Γκιούρδας.

Μπώκος, Γ. (2005). Οι ιστότοποι ως δημοσιεύματα: διαχειριστικές απόψεις του έργου της πληροφόρησης. *Τεκμήριον*. 5, 49-87.

Παπάζογλου, Γ. (2009). *Βυζαντινή βιβλιολογία: εισαγωγή στην ελληνική παλαιογραφία και κωδικολογία*. Κομοτηνή: Σταμούλης.

Παπαλεξίου, Ε. (2021), Τεχνολογίες αιχμής και συνεργατικά ερευνητικά εγχειρήματα στην υπηρεσία της ευρωπαϊκής πολιτιστικής κληρονομιάς. Σε Ε. Παπαλεξίου (Επιμ.), *Δημιουργικά αρχεία ως ζωντανά τοπία μνήμης στην ψηφιακή εποχή* (σ. 147-360). Αθήνα: CR.E.ARCH.-FagottoBooks.

Παπαχατζόπουλος, Α. (2016). *Συμβολή στην ιστοριογραφία του Δικηγορικού Συλλόγου Λαρίσης: οι δικηγόροι και ο Δικηγορικός Σύλλογος Λαρίσης κατά την περίοδο 1881-1981*. Λάρισα: Δικηγορικός Σύλλογος Λάρισας.

Πατηνιώτης, Μ. (2020). Στην Κωνσταντινούπολη ψαρεύοντας άθλιο ροζ. Σε Μ. Πατηνιώτης (Επιμ.), *Εισαγωγή στις ψηφιακές σπουδές* (σ. 7-17). Θεσσαλονίκη: Ροπή.

Τσιμπόγλου, Φ. (2006). Μοντέλο συνεργασίας ακαδημαϊκών βιβλιοθηκών: μια συστημική προσέγγιση. *Τεκμήριον*. 6, 29-97.

Ammirati, S., Firmani, D., Maiorino, M., Merialdo, P., Nieddu, E., & Rossi, A. (2017, June). *In Codice Ratio: Scalable Transcription of Historical Handwritten Documents*. Ανακτήθηκε 10 Οκτωβρίου, 2022, από: https://www.academia.edu/36363582/In_Codice_Ratio_Scalable_Transcription_of_Historical_Handwritten_Documents_Extended_Abstract_?from=cover_page.

Colavizza, G., Blanke, T., Jeurgens, C. & Noordegraaf, J. (2021). Archives and AI: An Overview of Current Debates and Future Perspectives. *Journal on Computing and Cultural Heritage*, 15(1), 1-15. doi.org/10.1145/3479010.

Eriksen, T. (2005). *Η τυραννία της στιγμής: γρήγορος και αργός χρόνος στην εποχή της πληροφορίας*. (Α. Σίμογλου, μεταφρ.). Αθήνα: Σαββάλας. (το πρωτότυπο έργο εκδόθηκε [χ.χ.]).

Gulati, A. (2021, Ιούνιος 29). Text detection from images using EasyOCR: Hands-on guide. Ανακτήθηκε 13 Οκτωβρίου, 2022, από: https://www.analyticsvidhya.com/blog/2021/06/text-detection-from-images-using-easyocr-hands-on-guide/.

Haykin, S. (2010). *Νευρωνικά δίκτυα και μηχανική μάθηση*. (Ε. Γκαγκάτσιου, μεταφρ.). Αθήνα: Παπασωτηρίου. (το πρωτότυπο έργο εκδόθηκε c2009).

Langlois, J. (1989). Modern manuscripts: a fragile heritage. *The UNESCO Courier: a window open on the world*. 42, (5), 5-6. Ανακτήθηκε 10 Σεπτεμβρίου, 2022, από: https://unesdoc.unesco.org/ark:/48223/pf0000083206.

Lazzerini, B., Marcelloni, F., & Reyneri, L. M. (1997). Beatrix: A self-learning system for off-line recognition of handwritten texts. *Pattern Recognition Letters*, 18(6), 583-594.https://doi.org/10.1016/S0167-8655(97)00039-1.

Levy, P. (2001). *Δυνητική πραγματικότητα: η φιλοσοφία του πολιτισμού και του κυβερνοχώρου*. (Μ. Καραχάλιος, μεταφρ.). Αθήνα: Κριτική. (το πρωτότυπο έργο εκδόθηκε 1995).

Manovich, L. (2020). Τα Νέα Μέσα από τον Μπόρχες ως την HTML. Σε Μ. Πατηνιώτης (Επιμ.), *Εισαγωγή στις ψηφιακές σπουδές* (σ. 229-316). Θεσσαλονίκη: Ροπή.

Milioni, N. (2020). *Automatic Transcription of Historical Documents: Transkribus as a Tool for Libraries, Archives and Scholars* (μη δημοσιευμένη μεταπτυχιακή εργασία). Uppsala University, Sweden. Ανακτήθηκε 25 Σεπτεμβρίου, 2022, από https://www.diva-portal.org/smash/get/diva2:1437985/FULLTEXT01.pdf.

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E.M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Meunier, J.-L., Michael, J., Mühlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Pérez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Sánchez, J.A., Schofield, P., Sfikas, G., Sieber, C., Stamatopoulos, N., Strauß, T., Terbul, T., Toselli, A.H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H. και Zagoris, K. (2019), "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study", *Journal of Documentation*, 75(5), 954-976. https://doi.org/10.1108/JD-07-2018-0114.

Nockels, J., Gooding, P., Ames, S. & Terras, M. (2022, 17 Ιουνίου). Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. *Archival Science*, 22, 367–392. https://doi.org/10.1007/s10502-022-09397-0.

Platanou, P., Pavlopoulos, J. & Papaioannou, G. (2022, 20 Ιουνίου). Handwritten Paleographic Greek Text Recognition: A Century-Based Approach. *Proceedings of the 13th Conference on Language Resources and Evaluation*, 6585-6589. Ανακτήθηκε 9 Ιανουαρίου, 2023, από https://aclanthology.org/2022.lrec-1.708.pdf.

Poster, M. (2004, 1 Μαΐου). History in the Digital Domain. *Historein*, 4, 17-32. https://doi.org/10.12681/historein.82.

Rabus, A. (2022). HandwrittenTextRecognition for Croatian Glagolitic. *Slovo*, 72, 181–192. Ανακτήθηκε 15 Σεπτεμβρίου, 2022, από https://hrcak.srce.hr/file/391286.

Rufenacht, M. (2020). *Handwritten Text Recognition (HTR) in 2020*. Ανακτήθηκε 9 Σεπτεμβρίου, 2022, από https://parashift.io/en/handwritten-text-recognition-in-2020/.

Schlagdenhauffen, R. (2020). Optical Recognition Assisted Transcription with Transkribus: The Experiment concerning Eugène Wilhelm's Personal Diary (1885-1951). *Journal of Data Mining & Digital Humanities*. Ανακτήθηκε 1 Οκτωβρίου, 2022, από https://jdmdh.episciences.org/6736/pdf.

Sridhar, B. (2022, Απρίλιος 17). *Integrating multiple OCR models to perform detection and recognition separately using Python*. Ανακτήθηκε 15 Οκτωβρίου, 2022, από https://medium.com/quantrium-tech/integrating-multiple-ocr-models-to-perform-detection-and-recognition-separately-using-python-f2c73743e1e0.

University of Groningen. ([χ.χ.]). *Monk*. Ανακτήθηκε 7 Οκτωβρίου, 2022, από https://www.ai.rug.nl/~lambert/Monk-collections-english.html.

Woodford, C. (2018). *OCR (optical character recognition)*. Ανακτήθηκε 2 Νοεμβρίου , 2022, από: https://www.explainthatstuff.com/how-ocr-works.html.

Web sites

https://www.abbyy.com/

https://www.amdigital.co.uk/

https://www.archivesportaleurope.net/about-us/the-portal/

http://arxeiomnimon.gak.gr/

https://digitalorientalist.com/

https://www.ekt.gr/

http://www.elia.org.gr/archives-collections/archives/

https://www.europeana.eu/el

http://www.fondazionebanconapoli.it/en/

http://gak.lar.sch.gr/

https://github.com/

https://www.greenstone.org/

https://rwi.app/iurisprudentia/en/iurisprudentia/

https://www.miet.gr/palaiografiko-arxio/

https://nodegoat.net/

https://pitt.libguides.com/

https://readcoop.eu/

http://stephanus.tlg.uci.edu/history.php/

https://voyant-tools.org/

Πηγές εικόνων

Φίλιος, Α. (1881). *Αριθ. 3*. Αρχείο Συμβολαιογράφου Φίλιου (ΣΥΜΒ.012.01, κυτίο 1, φάκελος 1). Αρχείο Συμβολαιογράφου Φίλιου. Γενικά Αρχεία του Κράτους. Τμήμα ΓΑΚ Λάρισας, Λάρισα. Μεταφορτώθηκε στις 23/05/2016 ως έργο που ανήκει στο Δημόσιο Τομέα (Public Domain)
από http://arxeiomnimon.gak.gr/browse/resource.html?tab=tab02&id=253155.

https://readcoop.eu/model/dutch-handwriting-17th-19th-century/