ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΩΝ

Κατηγοριοποίηση κριτηρίων καταλληλότητας σε ασθενείς με σύνδρομο Sjögren

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Λαμπέτης Νικόλαος

Επιβλέπων: Θεοδώρα Βαρβαρίγου

Καθηγήτρια Ε.Μ.Π

Αθήνα, Νοέμβριος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΩΝ

Κατηγοριοποίηση κριτηρίων καταλληλότητας σε ασθενείς με σύνδρομο Sjögren

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Λαμπέτης Νικόλαος

Επιβλέπων: Θεοδώρα Βαρβαρίγου

Καθηγήτρια Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 4η Νοεμβρίου 2021

Θεοδώρα Βαρβαρίγου	Σίμος Παπαβασιλείου	Μάνος Βαρβαρίγος
Καθηγήτρια Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π

Αθήνα, Νοέμβριος 2021

Λαμπέτης Νικόλαος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π

Περίληψη

Με την διαρκή αύξηση των ιατρικών κειμένων και κατ' επέκταση των κλινικών δοκιμών και σημειωμάτων, αυξάνεται ταυτόχρονα και η ανάγκη για αυτοματοποιημένους τρόπους λήψης πληροφορίας από αυτά.

Σκοπός της παρούσας εργασίας, είναι η ανάπτυξη ενός συστήματος αυτόματης αναγνώρισης οντοτήτων που αφορούν ιατρικές έννοιες σε κριτήρια καταλληλότητας κλινικών δοκιμών, η κατηγοριοποίηση αυτών, και η χρήση του για την εύρεση ασθενών που πληρούν ορισμένα κριτήρια καταλληλότητας σχετικά με το σύνδρομο Sjogren.

Στο σύστημα που αναπτύχθηκε, εφαρμόζεται συνδυασμός λεξιλογικών μεθόδων εύρεσης οντοτήτων με χρήση της οντολογίας HarmonicSS, σε συνδυασμό με την βιβλιοθήκη MeSH του National Library of Medicine για την αναγνώριση ιατρικών εννοιών, μαζί με μεθόδους κανόνων, για την εύρεση αρνήσεων και χρονικών περιορισμών στις κλινικές δοκιμές, ενώ επίσης χρησιμοποιήθηκε το σύνολο σχολιασμένων κριτηρίων Chia καθώς και ένα ειδικά σχεδιασμένο σύνολο σχολιασμένων κριτηρίων για την αξιολόγηση του συστήματος.

Τέλος αναπτύχθηκε μια ακόμη εφαρμογή, η οποία χρησιμοποιεί την αντλημένη πληροφορία και δεδομένα που έχουν καταγραφεί για ασθενείς με σύνδρομο Σιόγκρεν στα πλαίσια του έργου HarmonicSS, για τον εντοπισμό του πλήθους τους που πληρούν τα κριτήρια σε κάθε μία κλινική δοκιμή.

Λέξεις κλειδιά: Σύνδρομο Σιόγκρεν, Κριτήρια Καταλληλότητας, Κλινικές Δοκιμές, Επεξεργασία Φυσικής Γλώσσας, Βιβλιοθήκη MeSH

Abstract

After the constant growth of biomedical knowledge and the increase of clinical trials being conducted and documented, automated solutions for the retrieval of information from them has become a necessity.

The main cause of this diploma thesis is the development of a system capable of recognizing biomedical concepts in clinical trials' eligibility criteria, in order to categorize and look for patients fulfilling certain criteria related to Sjogren Syndrome.

The system developed implements a lexical language recognition approach using the HarmonicSS ontology in conjunction with the National Library of Medicine's MeSH library for the retrieval of biomedical concepts, along with rule-based approaches for the extraction of temporal and word negation information. At the same time the annotated criteria corpus Chia along with custom made annotated criteria were used for the evaluation of the system.

Finally, an application has been developed, which uses data concerning Sjogren Syndrome patients documented in the scope of the HarmonicSS project in order to retrieve the multitude of patients eligible for a given clinical trial.

Keywords: Sjögren Syndrome, Eligibility Criteria, Clinical Trials, Natural Language Processing, MeSH library

Ευχαριστίες

Θα ήθελα αρχικά να ευχαριστήσω την καθηγήτρια Θεοδώρα Βαρβαρίγου, για την ανάθεση και επίβλεψη της διπλωματικής μου εργασίας.

Θέλω επίσης να εκφράσω πολλές ευχαριστίες στον Ευθύμιο Χονδρογιάννη, για την στενή και αποτελεσματική συνεργασία μας καθ' όλη την διάρκεια της συγγραφής της συγκεκριμένης εργασίας.

Θα ήθελα τέλος να ευχαριστήσω την οικογένειά μου, τους συμφοιτητές και τους φίλους μου για τη στήριξή τους όλα αυτά τα χρόνια και ιδιαίτερα τους Δημήτρη Συρράφο, Μαρία Πριμηκύρη, Κωνσταντίνο Λαμπίρη και Βαλένα Δαρσινού.

Λαμπέτης Νικόλαος

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1 ΕΙΣΑΓΩΓΗ

1.1 Οι κλινικές δοκιμές

Οι κλινικές δοκιμές αποτελούν σημαντικό εργαλείο για την απόκτηση γνώσης, χρήσιμης ως προς την ανάπτυξη φαρμάκων, διαγνωστικών μεθόδων και μεθόδων πρόληψης και θεραπείας ασθενειών με την ενεργό συμμετοχή εθελοντών, οι οποίοι πληρούν συγκεκριμένα κριτήρια καταλληλότητας, ορισμένα από τους δημιουργούς της δοκιμής[1].

Ο αριθμός των κλινικών δοκιμών και συνεπώς των κριτηρίων καταλληλότητας προκειμένου να βρεθούν ασθενείς που μπορούν να μετέχουν σε αυτές, έχει αυξηθεί αισθητά[2]. Επίσης, τα κριτήρια καταλληλότητας που ορίζουν τους ασθενείς που μπορούν να μετέχουν στις κλινικές δοκιμές, καταγράφονται πολλές φορές με τρόπους που καθιστούν την ανάγνωση και την επεξεργασία τους από κάποιο υπολογιστικό σύστημα δύσκολη. Τα κριτήρια μπορεί να είναι για παράδειγμα καταγεγραμμένα χειροκίνητα, δίχως να ακολουθούν κάποια ορισμένη δομή.

Ταυτόχρονα, λόγω του γεγονότος πως οι κλινικές δοκιμές προέρχονται από πληθώρα διαφορετικών ανεξάρτητων διοργανωτών και κλάδων, σε αντίθεση με τους οργανισμούς που συμμετέχουν στον σχεδιασμό του τρόπου αναπαράστασης και αποθήκευσης των δεδομένων των ασθενών, δημιουργούνται εμπόδια συμβατότητας λεξιλογίων και ορολογιών, που σημαίνει πως συχνά η διασύνδεση μεταξύ αυτών πρέπει να γίνεται επίσης χειροκίνητα.

Ως αποτέλεσμα η διαδικασία εύρεσης ασθενών από τους διοργανωτές των δοκιμών αποτελεί πολλές φορές αρκετά χρονοβόρα εργασία. Για τον λόγο αυτό ερευνώνται συστήματα αναγνώρισης φυσικής γλώσσας ιατρικών όρων, που θα μπορέσουν να αυτοματοποιήσουν την παραπάνω διαδικασία.

[1] https://clinicaltrials.gov/ct2/about-studies/learn

[2] https://clinicaltrials.gov/ct2/resources/trends

1.2 Στόχος της εργασίας

Στόχος της συγκεκριμένης εργασίας είναι η ανάπτυξη ενός τέτοιου συστήματος, το οποίο επιχειρεί να κατηγοριοποιήσει κριτήρια καταλληλότητας κλινικών δοκιμών που αφορούν ασθενείς με σύνδρομο Sjogren και να πραγματοποιήσει εύρεση των ασθενών που τα πληρούν, χρησιμοποιώντας τεχνικές επεξεργασίας κειμένου και αναγνώρισης οντοτήτων με χρήση κυρίως λεξιλογικών μεθόδων αναζήτησης.

Για την διευκόλυνση της επεξεργασίας των κριτηρίων και της αναγνώρισης οντοτήτων σε αυτά, χρησιμοποιήθηκε το λεξικό MeSH (**Me**dical Subject Headings) της Αμερικανικής Εθνικής Ιατρικής Βιβλιοθήκης, το οποίο περιέχει μεγάλο αριθμό ιατρικών ορολογιών και εννοιών, καθώς επίσης και των διαφορετικών τρόπων γραφής και διατύπωσής τους.

Επίσης, για την εύρεση του πλήθους των ασθενών που τα πληρούν, χρησιμοποιήθηκε ένα σύστημα που αναπτύχθηκε στα πλαίσια του έργου HarmonicSS, το οποίο επιτρέπει σε εγγεγραμμένους στο σύστημα χρήστες να βρουν το πλήθος των ασθενών που πληρούν τα δοσμένα κριτήρια χωρίς να έχουν πρόσβαση σε αυτά, μέσω της έκφρασης ερωτημάτων προς αυτά, βασισμένων στην HarmonicSS οντολογία που αναπτύχθηκε στα πλαίσια του έργου αυτού.

Η αυτοματοποιημένη διαδικασία αναγνώρισης οντοτήτων που αναπτύχθηκε στην εργασία, έχει ως όφελος τον σύντομο και δίχως ανάγκη ανθρώπινου παράγοντα εντοπισμό εννοιών στα κριτήρια και κατ' επέκταση μπορεί να χρησιμοποιηθεί για την κατανόηση του συνόλου των κριτηρίων που ορίζονται σε κλινικές δοκιμές και αφορούν μια συγκεκριμένη ασθένεια, καθώς και τον αυτόματο εντοπισμό του πλήθους των ασθενών που πληρούν τα συγκεκριμένα κάθε φορά κριτήρια.

1.3 Δομή

Στην Ενότητα 2 γίνεται πρώτα παρουσίαση των δυσκολιών και των προκλήσεων που προκύπτουν κατά την επεξεργασία κειμένου και κατά την ανάπτυξη συστημάτων αναγνώρισης όρων/οντοτήτων (Named Entity Recognition) σε κείμενο (ιατρικού περιεχομένου και μη), αναλύονται δημοφιλείς τεχνικές επεξεργασίας κειμένου και αναγνώρισης οντοτήτων και παρουσιάζονται μοντέλα που έχουν χρησιμοποιηθεί σε παρόμοια συστήματα. Έπειτα δίνεται επεξήγηση όσον αφορά τα εργαλεία και τις βιβλιοθήκες/οντολογίες που χρησιμοποιήθηκαν.

Ύστερα στην Ενότητα 3 αναλύεται η αρχιτεκτονική και η μεθοδολογία του συστήματος επεξεργασίας των κριτηρίων καταλληλότητας των κλινικών δοκιμών και κάθε επιμέρους υποσυστήματος που το απαρτίζει. Αναλύεται επίσης ο τρόπος αξιολόγησης του συστήματος και η αρχιτεκτονική του υποσυστήματος που το αξιολογεί.

Στην Ενότητα 4 παρουσιάζονται και σχολιάζονται πρώτα τα αποτελέσματα της αξιολόγησης του συστήματος αναγνώρισης οντοτήτων και ύστερα τα αποτελέσματα της εύρεσης ασθενών και της κατηγοριοποίησης του συνόλου των κριτηρίων καταλληλότητας των κλινικών δοκιμών, χρησιμοποιώντας και ήδη υπάρχουσες επισημασμένες πηγές δεδομένων (annotated dataset) και εκ νέου σχολιασμένες πηγές που δημιουργήθηκαν από εμάς.

Τέλος στην ενότητα 5 συνοψίζονται τα κύρια σημεία της εργασίας ενώ στο παράρτημα αναφέρονται πιο αναλυτικά κάποιες τεχνολογίες και εργαλεία που χρησιμοποιήθηκαν κατά την ανάπτυξη του συστήματος της εργασίας, καθώς και ενδιάμεσα αποτελέσματα που έχουν προκύψει κατά την παραπάνω διαδικασία.

2 STATE OF THE ART - ΕΡΓΑΛΕΙΑ – ΒΙΒΛΙΟΘΗΚΕΣ

2.1 Επεξεργασία κειμένου

2.1.1. Δυσκολίες και προκλήσεις

Η επεξεργασία και αναγνώριση κειμένου αποτελεί τη διαδικασία, κατά την οποία υπολογιστικά μέσα χρησιμοποιούνται για την ανάλυση ή την παραγωγή ελεύθερων κειμένων. Στόχος είναι η κατά το μέγιστο δυνατόν προσέγγιση αναγνώρισης κειμένου όπως αυτή θα γινόταν από άνθρωπο και κατ' επέκταση η άντληση συμπερασμάτων από αυτό. Όταν γίνεται αναφορά σε κείμενο που περιέχει ελεύθερη γλώσσα, είναι προφανές πως αναδύονται δυσκολίες στην επεξεργασία του, οι οποίες δημιουργούνται από το γεγονός πως δεν υπάρχει αυστηρή δομή στον τρόπο που παρουσιάζονται τα δεδομένα του και πρέπει συνεπώς αυτή να αντληθεί. Οι λέξεις, οι προτάσεις και οι σχέσεις μεταξύ αυτών βρίσκονται φαινομενικά τυχαία κατανεμημένες μέσα στο κείμενο και αυτό με την σειρά του δίνεται απλά σαν μια ακολουθία συμβολοχαρακτήρων δίχως κάποια πληροφορία για τα επιμέρους στοιχεία τους. Ταυτόχρονα, τα μέρη του λόγου δύνανται να αποτελούν κομμάτια κάποιας περισσότερο ευρείας έννοιας ή φράσης, που πιθανόν να έχει διαφορετικό νοηματικό περιεχόμενο από τα επιμέρους τμήματα της. Τέλος σε κείμενο γραμμένο σε φυσική γλώσσα δεν αποκλείεται να βρεθούν διαφορές στην γραμματική, την ορθογραφία καθώς ίσως και γραμματικά/ορθογραφικά λάθη.

Γίνεται φανερό λοιπόν, πως σε πρώτη φάση το πρόβλημα που καλείται να αντιμετωπιστεί, είναι η επεξεργασία του κειμένου με κατάλληλο τρόπο, ώστε να γίνει φανερή η δομή του και να μπορούν να διακριθούν σωστά τα μέρη (προτάσεις, λέξεις) που το απαρτίζουν. Σε δεύτερη φάση πρέπει να γίνει προσδιορισμός πιο σύνθετων όρων και εκφράσεων ώστε στο τέλος να εξαχθεί η γνώση που αφορά την πληροφορία που αναζητείται. Για να προσδιοριστούν με περισσότερη ακρίβεια οι προκλήσεις που καλούνται να αντιμετωπιστούν κατά την επεξεργασία κειμένου φυσικής γλώσσας, είναι χρήσιμο να παρουσιαστούν συνοπτικά τα υποπροβλήματα που προκύπτουν και οι τρόποι που αυτά μπορούν να αντιμετωπιστούν για να υπάρξει ουσιώδης ανάλυση ενός κειμένου.

Οι βασικές δυσκολίες λοιπόν που προκύπτουν και καλούνται να αντιμετωπιστούν κατά την διαδικασία επεξεργασίας ενός κειμένου [1], πριν αυτό να καθίσταται έτοιμο για την άντληση νοήματος, απαιτούν μερικά ή όλα από τα ακόλουθα βήματα:

- Η αναγνώριση ορίων των προτάσεων (Sentence boundary detection)
- Ο διαχωρισμός των προτάσεων σε λέξεις-tokens (Tokenization)
- Η προσδιορισμός των λέξεων όσον αφορά την θέση τους σαν μέρος του λόγου (PartOfSpeech tagging)
- Η αναγνώριση λέξεων που δεν προσφέρουν στοιχεία για την άντληση νοήματος (stop words)
- Η αναγνώριση διαφορετικής γραμματικής ή και ορθογραφίας
- Η συλλογή λέξεων-tokens σε ουσιώδη συντακτικά/νοηματικά πακέτα (chunking)

Για την αντιμετώπιση των παραπάνω δυσκολιών όπως προαναφέρθηκαν επιγραμματικά, λαμβάνονται στην παρούσα εφαρμογή κάποια βοηθητικά εργαλεία ή και έτοιμοι πόροι που έπειτα από κατάλληλη επεξεργασία και χρήση συμβάλουν στην εξάλειψη τους. Τέτοια εργαλεία, όπως θα περιγραφούν και παρακάτω, αφορούν είτε έτοιμους αλγορίθμους επεξεργασίας κειμένου όπως για παράδειγμα αλγόριθμοι οριοθέτησης των λέξεων, είτε δομές που αξιοποιούν λίστες από αντικείμενα/λέξεις όπως για παράδειγμα μια λίστα λέξεων που φανερώνουν άρνηση ή μια λίστα πιθανών επεκτάσεων ακρωνύμων.

2.1.2. Βασικά Εργαλεία Επεξεργασίας Κειμένου

2.1.2.1. Εργαλείο parsing

Προκειμένου να επεξεργασθούν οι επιμέρους λέξεις ενός κειμένου και ύστερα να αναγνωριστούν, πρέπει πρώτα να προσδιοριστεί ο τρόπος με τον οποίο αυτές οριοθετούνται μέσα σε αυτό. Πρέπει δηλαδή να βρεθούν οι χαρακτήρες καθώς και οι συνθήκες, κάτω από τις οποίες αυτοί χωρίζουν τις λέξεις και τις προτάσεις. Για παράδειγμα μία τελεία μπορεί να σηματοδοτεί την έναρξη νέας πρότασης αλλά μπορεί και να χρησιμοποιηθεί και σε κάποια συντομογραφία ή σε κάποια έκφραση που περιγράφει δοσολογίες.

Για τον λόγο αυτό γίνεται χρήση ενός parser ο οποίος έχει χρησιμοποιηθεί κατά την επεξεργασία ιατρικών κειμένων και την αναγνώριση συντομεύσεων [2]. Συγκεκριμένα είναι σχεδιασμένος με τέτοιον τρόπο, ώστε να μπορεί να αναγνωρίσει μοτίβα που παρατηρούνται συχνά σε κριτήρια καταλληλότητας, τα οποία σχετίζονται με την χρήση σημείων στίξης κατά την περιγραφή συντομεύσεων και δοσολογιών φαρμάκων. Με τον τρόπο αυτό εντοπίζονται με επιτυχία οι επιμέρους συμβολοακολουθίες που συναντώνται συχνά στα κριτήρια καταλληλότητας, συμπεριλαμβανομένου των συντομεύσεων, των αριθμών και των μονάδων μέτρησης, σε αντίθεση με κάποιον συμβατικό parser, σχεδιασμένο για κείμενα γενικής πληροφορίας, όπως για παράδειγμα ο Stanford-Parser.

2.1.2.2 Το εργαλείο Lexical Tools

Το κύριο λεξιλογικό εργαλείο που χρησιμοποιήθηκε είναι το Lexical Tools[3] της Αμερικανικής Εθνικής Ιατρικής Βιβλιοθήκης, το οποίο παρέχει εργαλεία επεξεργασίας κειμένου. Ταυτόχρονα προσφέρει και εργαλεία Word Indexing και ordering. Το Lexical Tools αποτελείται από μια συλλογή προγραμμάτων η οποία είναι σχεδιασμένη να συνδράμει στην επεξεργασία φυσικής γλώσσας. Είναι ιδιαίτερα χρήσιμο, αφού είναι σχεδιασμένο κυρίως για διαχείριση των λεξιλογίων της Αμερικανικής Εθνικής Ιατρικής Βιβλιοθήκης, όπως σε αυτήν την περίπτωση το MeSH. Περιλαμβάνει κυρίως 3 βασικά λεξιλογικά εργαλεία, το εργαλείο Lexical Variant Generator, το εργαλείο Norm και το εργαλείο WordInd, τα οποία διευκολύνουν πτυχές της επεξεργασίας κειμένου. Παρακάτω περιγράφονται τα δύο πρώτα τα οποία και χρησιμοποιούνται στην παρούσα εργασία.

2.1.2.3 Το πρόγραμμα Lexical Variant Generator (LVG)

Το πρώτο εργαλείο είναι ο παραγωγός παραλλαγών λέξεων (Lexical Variant Generator - LVG). Ο LVG διαθέτει μια σειρά από εντολές οι οποίες επιβάλλουν μετατροπές στο δοθέν κείμενο για να το φέρουν σε ειδική μορφή αντίστοιχη κάθε φορά των εντολών που επιλέγονται. Κάποια από τα βασικότερα βήματα-εντολές που περιέχονται στο πρόγραμμα LVG για τον σκοπό αυτό είναι οι:

[3] https://www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_003.html

Εντολές	Ενέργεια
Remove inflections and conjugations	Απαλείφει στοιχεία τα οποία αφορούν την κλίση και την σύζευξη στο κείμενο.
Order words in multi-word terms	Ταξινομεί τις λέξεις σε κείμενα τα οποία αποτελούνται από περισσότερες από μία.
Remove alphabetic case	Μετατρέπει τα κεφαλαία σε πεζά
Remove punctuation	Αφαιρεί τα σημεία στίξης
Remove possessives	Αφαιρεί κτητικές αντωνυμίες και επιρρήματα

Πίνακας 1: Βασικές εντολές επεξεργασίας κειμένου του εργαλείου LVG

Περισσότερες πληροφορίες σχετικά με το LVG καθώς επίσης παραδείγματα εκτέλεσης και η αναλυτική λίστα των εντολών του βρίσκονται στο παράρτημα.

2.1.2.4 Το πρόγραμμα Norm

Το πρόγραμμα Norm μετατρέπει κατάλληλα κάποια ακολουθία συμβολοσειρών (προτάσεις, φράσεις, λέξεις) εκτελώντας με σειρά εντολές του LVG τέτοιες ώστε να αφαιρούνται από αυτήν, Κεφαλαία γράμματα, Κλίση, Σύζευξη, Διαφορές στην ορθογραφία, Σημεία στίξης, Γένη, Λέξεις χωρίς πληροφορία, Διακριτικά, Συμπλέγματα, Η σειρά των λέξεων και τέλος Αφαίρεση όσων χαρακτήρων δεν είναι ASCII. Το αποτέλεσμα της διαδικασίας είναι η λέξη στην κανονικοποιημένη (normalized) μορφή της. Η κανονικοποιημένη λέξη ή φράση ορίζεται και χρησιμοποιείται στην παρούσα εργασία σύμφωνα με την επεξεργασία που επιδέχεται μια συμβολοακολουθία από τις παραπάνω εντολές

2.1.2.5 Το εργαλείο NegEx

Ακόμα μία πρόκληση που καλείται να αντιμετωπιστεί κατά την επεξεργασία κειμένου και εμφανίζει ιδιαίτερο ενδιαφέρον κατά την επεξεργασία κλινικών σημειωμάτων είναι ο χαρακτηρισμός λέξεων ή προτάσεων ως αρνήσεις. Αρκετά συχνά σε ιατρικά κείμενα και κριτήρια καταλληλότητας γίνεται αναφορά σε όρους οι οποίοι αφορούν την απουσία για παράδειγμα κάποιας ουσίας σε ένα αποτέλεσμα εξέτασης ή ασθένειας σε κάποιον ασθενή, ενώ οι λέξεις οι οποίες ορίζουν την σημασία αυτή πολλές φορές δεν αναγνωρίζονται από συστήματα επεξεργασίας κειμένου σαν πληροφορία [3]. Ένα σύστημα που σχεδιάστηκε για τον σκοπό αυτό και χρησιμοποιείται στην παρούσα εργασία είναι το εργαλείο NegEx, το οποίο με την βοήθεια ενός συνόλου κανόνων κανονικών εκφράσεων (regular expressions) και την χρήση μίας λίστας λέξεων

και φράσεων άρνησης χαρακτηρίζει έναν όρο μέσα σε κάποια πρόταση ως NEGATED ή μη NEGATED.

2.2 Αναγνώριση και Αποσαφήνιση Οντοτήτων

Η αναγνώριση οντοτήτων (Named Entity Recognition - NER) είναι μια διαδικασία επεξεργασίας φυσικού κειμένου κατά την οποία, σε λέξεις ή εκφράσεις, δίνεται κάποιος προσδιορισμός σχετικά με την πληροφορία που προσφέρουν. Κατά την αναγνώριση οντοτήτων χρησιμοποιούνται τεχνικές επεξεργασίας κειμένου όπως αυτές που αναφέρθηκαν παραπάνω σε συνδυασμό με επιπρόσθετα εργαλεία προσαρμοσμένα κάθε φορά κατάλληλα ώστε να επιτευχθεί ο χαρακτηρισμός των στοιχείων του κειμένου ως ονόματα ανθρώπων, τοποθεσιών, οργανισμών κτλ. . Η αναγνώριση αυτών των στοιχειών ως ονοματισμένες οντότητες διευκολύνει στην αποδόμηση του κειμένου και στην άντληση σημαντικών πληροφοριών όσον αφορά το νόημα του.

Για την επίτευξη του παραπάνω σκοπού έχει ερευνηθεί και αναπτυχθεί ένα σύνολο τρόπων προσέγγισης και τεχνικών τέτοιων ώστε κάθε μία να μπορεί να εφαρμοστεί κατάλληλα σε διαφορετικά είδη κειμένων με σκοπό τον αποτελεσματικότερο χαρακτηρισμό και την αποσαφήνιση των οντοτήτων. Παρακάτω περιγράφονται αυτές οι προσεγγίσεις αναλυτικότερα, προσφέρονται παραδείγματα που χρησιμοποιούνται και αναλύονται τα θετικά και αρνητικά τους.

2.2.1 Λεξιλογική προσέγγιση

Στην συγκεκριμένη προσέγγιση, για τον εντοπισμό ευρημάτων μέσα στο κείμενο, γίνεται χρήση ενός συνόλου λέξεων, όρων ή εκφράσεων, το οποίο μπορεί να είναι ένα απλό λεξικό ή κάποιο λεξικό εμπλουτισμένο με πληροφορίες που αφορούν τις σχέσεις μεταξύ των στοιχείων του, όπως για παράδειγμα είναι μια οντολογία ή κάποια βάση γνώσης. Η λεξιλογική προσέγγιση δηλαδή αποτελεί κατά κάποιο τρόπο μία σχεδόν άμεση αντιστοίχιση του συνόλου αυτού και του κειμένου.

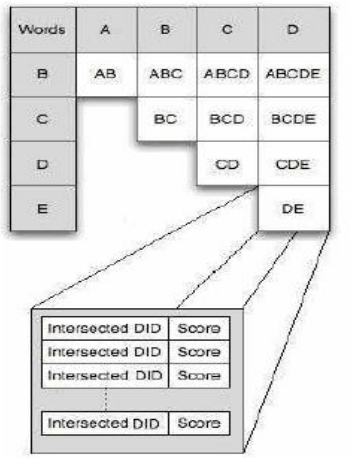
Ωστόσο η διαδικασία αντιστοίχισης λέξεων κάποιου λεξικού με ένα κείμενο περιλαμβάνει και ένα ενδιάμεσο βήμα μορφοποίησης κειμένου, τέτοιας ώστε να είναι δυνατόν να αντιστοιχιστούν λέξεις οι οποίες δύνανται να εμφανίζουν μικροδιαφορές όπως για παράδειγμα στην κλίση και στην ορθογραφία τους. Για τον λόγο αυτό απαιτούνται ειδικά εργαλεία μορφοποίησης κειμένου, ώστε να είναι δυνατόν να ληφθούν υπ’ όψη οι μικρές διαφοροποιήσεις των όρων που πρέπει να εντοπισθούν με τους γνωστούς όρους του λεξικού. Τα εργαλεία αυτά μπορεί να είναι αλγόριθμοι

κανονικοποίησης κειμένου όπως για παράδειγμα εργαλεία για word normalization, stemming, capital stripping και stop word removal.

Για παράδειγμα, στην εργασία [4] χρησιμοποιείται λεξιλογική προσέγγιση αναγνώρισης κειμένου με βάση γνώσης αποτελούμενη από κοινές φράσεις και λέξεις για την αναγνώριση κειμένου προερχόμενου από είσοδο χρήστη, με σκοπό την δημιουργία ενός συστήματος ανταπαντήσεων. Κατά την διαδικασία που ακολουθείται, γίνεται χρήση ενός αλγορίθμου εύρεσης επακολουθιών συμβολοχαρακτήρων αξιοποιώντας έναν πίνακα ομοιότητας (Matching Matrix), ο οποίος αρχικά αναζητεί τις μεγαλύτερες πιθανές συμβολοσειρές [Εικόνα 1] του κειμένου που συμπίπτουν με κάποια λέξη η φράση της βάσης γνώσης και τις προσθέτει σε μια λίστα πιθανών ευρημάτων. Ύστερα γίνεται αξιολόγηση ως προς την ομοιότητα των ευρημάτων με την λέξη ή φράση της βάσης γνώσης, με κριτήριο την διασπορά του ευρήματος, το μήκος του και την διάταξή του ενώ ταυτόχρονα πραγματοποιείται έλεγχος για την απόρριψη υπό-ακολουθιών που υπερκαλύπτονται από ήδη αναγνωρισμένες συμβολοσειρές.

			d	e	l	i	v	e	r	y		d	a	y
		i	0	1	2	3	4	5	6	7	8	9	10	11
	j	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	1	0	0	0	0	0	0	0	0	1	0	0
a	1	0	0	0	0	0	0	0	0	0	0	0	2	0
t	2	0	0	0	0	0	0	0	0	0	0	0	0	0
e	3	0	0	1	0	0	0	1	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	0	1	0	0	0
o	5	0	0	0	0	0	0	0	0	0	0	0	0	0
f	6	0	0	0	0	0	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	0	0	0	1	0	0	0
d	8	0	1	0	0	0	0	0	0	0	0	2	0	0
e	9	0	0	2	0	0	0	1	0	0	0	0	0	0
l	10	0	0	0	3	0	0	0	0	0	0	0	0	0
i	11	0	0	0	0	4	0	0	0	0	0	0	0	0
v	12	0	0	0	0	0	5	0	0	0	0	0	0	0
e	13	0	0	1	0	0	0	6	0	0	0	0	0	0
r	14	0	0	0	0	0	0	0	7	0	0	0	0	0
y	15	0	0	0	0	0	0	0	0	8	0	0	0	1

Εικόνα 1: Διαδικασία εύρεσης μέγιστης κοινής συμβολοσειράς δύο συμβολοσειρών με χρήση πίνακα ομοιότητας



Εικόνα 2: Πίνακας αντιστοίχισης διαδοχικών συμβολοσειρών-λέξεων και αντιστοίχιση με όρους λεξιλογίου

Όμοια τεχνική ακολουθείται και στην προσέγγιση [5], όπου με παρόμοιο τρόπο χρησιμοποιείται ένας πίνακας αντιστοίχισης (Matching Matrix αλγόριθμος) για την εύρεση των μέγιστων γειτονικών ακολουθιών σε κριτήρια καταλληλότητας που μπορούν να αντιστοιχιστούν με την

βάση SNOMED CT. Όπως φαίνεται και στην Εικόνα 2 συλλέγονται με την βοήθεια του πίνακα αντιστοίχισης όλες οι πιθανές γειτονικές ακολουθίες λέξεων, καθώς και όλοι οι όροι του λεξιλογίου που προκύπτουν από κάθε λέξη που περιλαμβάνει η ακολουθία και τέλος γίνεται αξιολόγηση για την επιλογή του καταλληλότερου όρου. Η διαφορά σε αυτήν την περίπτωση είναι πως ο αλγόριθμος δεν εφαρμόζεται σε ακολουθίες συμβολοχαρακτήρων, αλλά σε ακολουθίες λέξεων οι οποίες έχουν αντληθεί από το κείμενο μέσω διαδικασίας τμηματοποίησης του (συγκεκριμένα ένα Maximum entropy boundary detection model) , έχουν έρθει σε κανονική μορφή (αφαίρεση καταλήξεων, κλίσεων, διαφοροποίησης ορθογραφίας) και έχουν χαρακτηριστεί ως μέρη του λόγου (PartOfSpeech-Tags) με χρήση του εργαλείου GENIA-Tagger[4].

Η χρήση τέτοιου είδους προσεγγίσεων είναι αρκετά χρήσιμη, όταν στο κείμενο προς ανάλυση βρίσκονται όροι που κατά κύριο λόγο καλύπτονται από κάποιο γνωστό λεξικό και προσφέρει ικανοποιητικά αποτελέσματα με χαμηλότερο κόστος υπολογισμού από αυτό που άλλες πιθανές προσεγγίσεις μπορεί να απαιτούν για την ίδια διεργασία. Η αδυναμία της λεξιλογικής προσέγγισης είναι πως η αποτελεσματικότητα της εξαρτάται σχεδόν απόλυτα από το μέγεθος και την ποιότητα του λεξικού που διατίθεται. Αν δηλαδή όροι του κειμένου δεν βρίσκονται στην βιβλιοθήκη που χρησιμοποιείται δεν είναι δυνατόν να βρεθούν ενώ ταυτόχρονα για το ίδιο λόγο περιορίζεται αρκετά η δυνατότητα χρήσης του συστήματος για τον εντοπισμό όρων σε κείμενο διαφορετικού περιεχομένου και κατά συνέπεια διαφορετικού λεξιλογίου. Είναι λοιπόν προφανές πως η αποδοτικότητα τέτοιων συστημάτων βασίζεται κυρίως στο πόσο στενά συνδεδεμένο είναι το λεξιλόγιο του κειμένου με το λεξικό ή την βιβλιοθήκη που χρησιμοποιείται. Η μέθοδος αυτή χρησιμοποιείται ευρέως σε κείμενα με ιατρικά δεδομένα, αφού αρκετές βάσεις γνώσης προσφέρουν πολυπληθή δεδομένα με όρους και έννοιες της ιατρικής ορολογίας.

2.2.2 Προσέγγιση με χρήση ενός συνόλου κανόνων

Στην περίπτωση αυτή γίνεται χρήση κανόνων όπως για παράδειγμα η ικανοποίηση κάποιων κανονικών εκφράσεων (regular expressions) καθώς και άλλων μέσων για την αναγνώριση κειμένου. Για παράδειγμα η χρήση κεφαλαίων στην αρχή μιας λέξης που δεν βρίσκεται σε αρχή

[4] http://www.nactem.ac.uk/GENIA/tagger/

πρότασης βοηθάει στην κατηγοριοποίηση του ευρήματος ως όνομα ή τοποθεσία. Επίσης αναγνωρίσιμος εκφράσεων όπως “O is based in L” μπορούν να χρησιμοποιηθούν για να βγει το συμπέρασμα πως η λέξη O είναι οργανισμός και η λέξη L είναι τοποθεσία, και η “P was born in L” ονόματος ανθρώπου και τοποθεσίας αντίστοιχα. Στην περίπτωση αυτή την θέση του λεξικού παίρνει το ίδιο το σύνολο των κανόνων. Η προσέγγιση με χρήση κανόνων δίνει την δυνατότητα συμπερασμού της έννοιας όρων, δίχως αυτές να μπορούν να αντιστοιχιστούν άμεσα με κάποιο λεξικό βγάζοντας συμπεράσματα μόνο από την μορφή τους, τον τρόπο γραφή τους, την θέση τους στην πρόταση και την θέση τους σε σχέση με κάποιες άλλες λέξεις κλειδιά.

Για παράδειγμα στην εργασία [6], χρησιμοποιείται μία τροποποίηση ενός συστήματος αναγνώρισης ονομαζόμενο FASTUS για την εξαγωγή κάποιας πληροφορίας από κείμενο. Το FASTUS αποτελείται από μια σειρά πεπερασμένων αυτομάτων (finite automata) τα οποία ακολουθούν μια συγκεκριμένη σειρά τροφοδότησης του κειμένου μεταξύ τους μέχρι να εξαχθεί από αυτό πληροφορία. Με χρήση λοιπόν συντακτικών κανόνων και κανόνων εύρεσης μοτίβων ακολουθείται μια λογική τεσσάρων τμημάτων [6][7].

Προτού ξεκινήσει η διαδικασία των τεσσάρων τμημάτων του συστήματος οι προτάσεις που δίνονται σαν είσοδος διέρχονται μέσα από έναν parser-tokenizer και ταυτόχρονα ελέγχονται ως προς την δυνατότητα τους να συντάξουν γνωστές κοινές φράσεις οι οποίες αναζητούνται με την βοήθεια μίας βάσης μοτίβων φράσεων. Ύστερα:

1. Σημειώνονται λέξεις οι οποίες υποδεικνύουν σημαντική πληροφορία όπως για παράδειγμα ονόματα ("ABC Corp.", "John Smith") αναγνωρίζοντας μοτίβα κεφαλαιοποίησης χαρακτήρων λέξεων. Για παράδειγμα δύο διαδοχικές λέξεις με τον πρώτο τους χαρακτήρα κεφαλαίο πολύ πιθανόν να συνθέτουν ένα όνομα.

2. Οι προτάσεις διαιρούνται σε συντακτικά γκρουπ/φράσεις και χαρακτηρίζονται ως ομάδες υποκειμένων, ρημάτων, συνδέσμων, ονομάτων και αντωνυμιών με την χρήση μάκρο - κανόνων όπως φαίνεται και στην Εικόνα 3 για το παράδειγμα Υποκείμενο-Ρήμα-Αντικείμενο.

3. Αναγνωρίζονται μοτίβα και φράσεις καθώς και μεταξύ τους σχέσεις που παρουσιάζουν ενδιαφέρον, αντίστοιχα με χρήση συνόλου μάκρο-κανόνων και κανονικών εκφράσεων. Για παράδειγμα στην περίπτωση που αναζητείται πληροφορία σχετικά με την κατηγορία

κάποιου υποκειμένου από κάποια αρχή αναζητείται το μοτίβο <GovtOtficial> accused <PerpOrg> of <Incident>.

```
EVENT-PHRASE --> EVENT-ADJUNCT* (NG[??subj] ({COMPL | COMPL1}))
                    VG[Active=T,Subcat=Basic,??head]
                    (NG[??obj])
                    {P[??prep1] NG[??pobj1] | P[??prep2] NG[??pobj2] |
                     P[??prep3] NG[??pobj3] | EVENT-ADJUNCT}*;
    head = (head 2);
    rule-type = ActiveBase;
    svo-pattern = ??label;
                    ??semantics;;
```

Εικόνα 3: Κανόνας για την αναγνώριση του μοτίβου υποκείμενο-ρήμα-αντικείμενο του συστήματος FASTUS

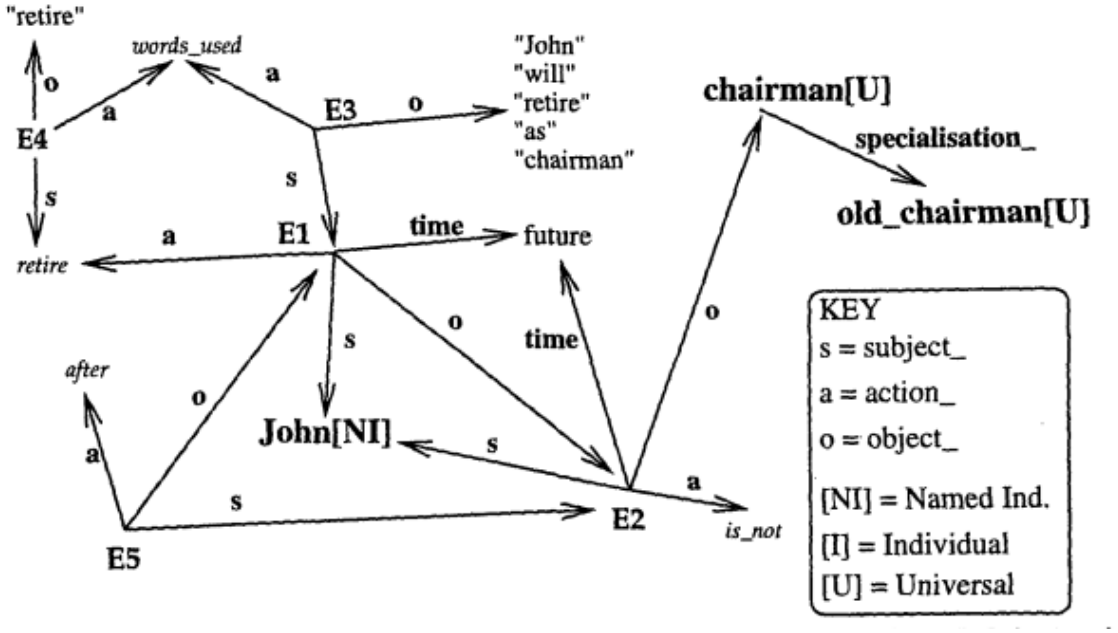
4. Διασταυρώνονται αναγνωρισμένα ευρήματα μεταξύ τους, αν συσχετίζονται. Για παράδειγμα αν ένα όνομα ή μια ενέργεια εμφανίζεται σε δυο αναγνωρισμένα μοτίβα/περιστατικά, αυτά συμπυκνώνονται και δομείται ένα συνολικό περιστατικό το οποίο περιέχει την συνολική πληροφορία όπως φαίνεται στην Εικόνα 4.

Όμοια στην εργασία [8] αναλύεται το σύστημα LOLITA το οποίο εξάγει πληροφορία από κείμενο με χρήση ενός σύνθετου δικτύου/γράφου (SemNet) του οποίου τα επιμέρους στοιχεία αναπαριστούν την εφαρμογή λεξικών και συντακτικών κανόνων για την αναγνώριση εκφράσεων λέξεων και μοτίβων. Πιο συγκεκριμένα το δίκτυο του συστήματος LOLITA αποτελείται από

Incident:	KILLING		
Perpetrator:	"terrorist"		
Confidence:	–		
Human Target:	"Roberto Garcia Alvarado"	Incident:	KILLING
		Perpetrator:	FMLN
and		Confidence:	Suspected or Accused by Authorities
Incident:	INCIDENT	Human Target:	"Roberto Garcia Alvarado"
Perpetrator:	FMLN		
Confidence:	Suspected or Accused by Authorities		
Human Target:	–		

Εικόνα 4: : Δύο μοτίβα αναγνωρισμένα από το σύστημα FASTUS συμπτύσσονται σε ένα γενικότερο

100,000 κόμβους, καθένας των οποίων παριστάνει έννοιες οντοτήτων ή γεγονότων και κατευθυνόμενες ακμές οι οποίες αναπαριστούν τις σχέσεις μεταξύ τους. Κατά την διαδικασία αναγνώρισης μίας πρότασης από το σύστημα, διατρέχονται τα tokens της ύστερα από κατάλληλη προ-επεξεργασία (parsing, normalization, tokenization) και γίνεται ουσιαστικά μια προβολή της στον Γράφο/δίκτυο SemNet. Ύστερα η προβολή αυτή παριστάνει την πρόταση που πρέπει να αναλυθεί προσαρμοσμένη σε γνωστές έννοιες και μοτίβα όπως φαίνεται και στην Εικόνα 5.



Εικόνα 5: Παράδειγμα του συστήματος LOLITA για την αναγνώριση της φράσης "John will retire as chairman.". Συγκεκριμένα ένα τμήμα του δικτύου SemNet το οποίο αναπαριστά τις έννοιες και τις σχέσεις που περιέχονται στην πρόταση

Βασικό πλεονέκτημα των μεθόδων αυτών είναι όπως αναφέρθηκε και προηγουμένως να προσπεραστεί το πρόβλημα της απουσίας κάποιου όρου από γνωστά λεξικά. Ένα αντιπροσωπευτικό παράδειγμα όπως αναφέρθηκε και παραπάνω είναι ο εντοπισμός ονομάτων (ανθρώπων, τόπων ή και οργανισμών), καθώς το πλήθος όλων των υπαρχόντων ονομάτων δεν είναι δυνατόν να καλυφθεί πλήρως από ένα λεξικό ενώ νέα ονόματα προστίθενται στα υπάρχοντα συνεχώς. Για αρκετή πληροφορία δηλαδή που μπορεί να αποκωδικοποιηθεί σε κάποιο στοιχείο, το οποίο ακολουθεί ένα συγκεκριμένο μοτίβο είναι δυνατόν να δημιουργηθεί ένας κανόνας για τον εντοπισμό αυτού του μοτίβου και την κατηγοριοποίηση της λέξης σύμφωνα με αυτό.

Ωστόσο δεν είναι πάντα αποδοτική αυτού του είδους η προσέγγιση, αφού δεν είναι δυνατόν οι πληροφορίες που αναζητούνται να ακολουθούν σε κάθε περίπτωση το σύνολο των κανόνων που έχει οριστεί για την εύρεση τους ενώ παράλληλα, ίδια μοτίβα μπορούν να χρησιμοποιούνται για να περιγράψουν και διαφορετικού είδους όρους από αυτόν που οι κανόνες έχουν σχεδιαστεί να αναγνωρίζουν σε αυτά. Όσον αφορά κείμενα ιατρικού περιεχομένου γίνεται αρκετές φορές χρήση συνόλου κανόνων όταν αντιμετωπίζονται προβλήματα αναγνώρισης μονάδων μέτρησης, ακρωνύμων και ονομασιών που ακολουθούν κάποιο μοτίβο όπως για παράδειγμα οι ονομασίες πρωτεϊνών και κάποιες ονομασίες φαρμάκων και ασθενών.

2.2.3 Με την βοήθεια μηχανικής μάθησης (στατιστική προσέγγιση)

Οι προσεγγίσεις μηχανικής μάθησης χρησιμοποιούν μαθηματικές τεχνικές και μεγάλο όγκο στατιστικών δεδομένων από πολλά διαφορετικά κείμενα για την ανάπτυξη γενικευμένων μοντέλων αναγνώρισης φυσικής γλώσσας, έχοντας με αυτόν τον τρόπο ως βάση πραγματικά παραδείγματα που παρέχονται από τα κείμενα αυτά χωρίς την ανάγκη προσθήκης γλωσσικών κανόνων και σύνθετων λεξιλογίων. Οι τεχνικές προσέγγισης με μηχανική μάθηση μπορούν να χωριστούν σε δύο βασικές υποκατηγορίες: τις επιτηρούμενες από άνθρωπο (supervised) και τις μη επιτηρούμενες (non-supervised). Στην περίπτωση του supervised learning που είναι και η επικρατέστερη μέθοδος ανάπτυξης τεχνικών μηχανικής μάθησης για αναγνώριση κειμένου όπου αναπτύσσεται ένα στατιστικό μοντέλο το οποίο «προπονείται» σύμφωνα με δεδομένα τα οποία έχουν χαρακτηριστεί με ανθρώπινη παρέμβαση.

Βασικές χρήσεις τεχνικών μηχανικής μάθησης στον τομέα της αναγνώρισης φυσικής γλώσσας είναι η αναγνώριση φωνής, η αναγνώριση μερών του λόγου (Part of Speech Tagging), η μάθηση νέου λεξιλογίου, η αναγνώριση σχέσεων μεταξύ λέξεων, σύνθετων γραμματικών κανόνων καθώς επίσης και η μετάφραση κειμένων. Μία κύρια χρήση τεχνικών μηχανικής μάθησης στην αναγνώριση φυσικού κειμένου η οποία χρησιμοποιείται ευρέως είναι όπως αναφέρθηκε η αναγνώριση μερών του λόγου (Part of Speech Tagging) κατά την οποία στόχος είναι η κάθε λέξη σε κάποιο κείμενο να χαρακτηριστεί για παράδειγμα ως υποκείμενο, ρήμα, επίθετο, αντικείμενο, σύνδεσμος κλπ. ενώ μία εξίσου σημαντική εφαρμογή είναι η αναγνώριση οντοτήτων ονομάτων, όπως για παράδειγμα ονόματα τοποθεσιών, ανθρώπων και οργανισμών. Αξίζει να σημειωθεί πως σύγχρονες τεχνικές αναγνώρισης φυσικής γλώσσας βασίζονται αρκετά στην προσέγγιση αυτή αφού έχει αποδειχθεί πως παράγει ικανοποιητικά αποτελέσματα.

Επίσης αξίζει να αναφέρουμε ότι ένα άλλο παράδειγμα εφαρμογής τέτοιου είδους προσέγγισης για την επεξεργασία φυσικής γλώσσας είναι η διαδικασία word embedding [9], κατά την οποία οι λέξεις αναπαρίστανται από κάποιο διάνυσμα το οποίο περιέχει τιμές, η καθεμία από τις οποίες περιγράφει κάποιο «χαρακτηριστικό» της λέξης. Αυτές οι τιμές προσαρμόζονται έπειτα σύμφωνα με κάποιο στατιστικό μοντέλο (πχ. Νευρωνικό Δίκτυο) πάνω σε σώματα κειμένου μέσω της διαδικασίας της μηχανικής μάθησης, βελτιστοποιώντας έτσι τις τιμές των βαρών των κόμβων του νευρωνικού δικτύου. Ως αποτέλεσμα λέξεις με κοντινές τιμές στα διανύσματά τους μπορούν να θεωρούνται και σημασιολογικά ή συντακτικά όμοιες.

Σχετικά με την αναγνώριση των όρων ενός κειμένου, ένα χαρακτηριστικό παράδειγμα χρήσης στατιστικών μοντέλων περιγράφεται στην εργασία [10] όπου γίνεται χρήση τεσσάρων διαφορετικών αλγορίθμων για την κατηγοριοποίηση κειμένου, συγκεκριμένα των: robust linear classifier, maximum entropy, transformation-based learning και Hidden Markov Model (HMM) [11]. Οι αλγόριθμοι που χρησιμοποιούνται εδώ προσδιορίζουν κάθε στοιχείο στο κείμενο επισημαίνοντας το με μια ετικέτα μέρους λόγου, που αντιστοιχεί στη θέση του σε σχέση με μια ονομαζόμενη οντότητα (Named Entity). Είτε δηλαδή αποτελεί την αρχή κάποιας ονομασμένης οντότητας, είτε την επεκτείνει/συνεχίζει, είτε την τερματίζει ή τέλος δεν ανήκει σε καμία οντότητα. Οι robust linear classifier, maximum entropy και transformation-based learning αλγόριθμοι χρησιμοποιούνται αποκλειστικά για την αντιστοίχιση των οντοτήτων με κάποια POS-Tag, ενώ το Hidden Markov Model σε αυτή την περίπτωση βρίσκει χρησιμότητα κυρίως στην άντληση συμπερασμάτων όσον αφορά τις σχέσεις μεταξύ των οντοτήτων του κειμένου. Σημειώνουμε ότι για την διαδικασία εκμάθηση των παραπάνω μοντέλων χρησιμοποιήθηκε ένα σύνολο ονομάτων 50.000 πόλεων, 80.000 ανθρώπων και 3.500 οργανισμών.

Αξιοσημείωτη εφαρμογή στο πλαίσιο της εξέτασης κλινικών σημειωμάτων αποτελεί το μοντέλο που αναπτύχθηκε στην εργασία [12] όπου μαζί με την χρήση συνόλου κανόνων για τον προσδιορισμό όρων ως πιθανοί όροι ιατρικού περιεχομένου και τον συμπερασμό ακρωνύμων χρησιμοποιεί, για την κατηγοριοποίηση κλινικών σημειωμάτων 12000 ασθενών σε εύρος 6 ετών, ένα μοντέλο Conditional Random Fields το οποίο αποδίδει στην δυνατότητα του να αναλύει ακολουθίες όρων με σκοπό τον ονοματισμό τους ως οντότητες. Τα δεδομένα που χρησιμοποιήθηκαν για την ανάπτυξη του μοντέλου προέρχονταν κυρίως από το λεξικό του

UMLS[5] και το λεξικό SNOMED CT (SNOMED International, 2009)[6] όσον αφορά ιατρικές ορολογίες και το MOBY για τους κοινούς όρους της γλώσσας, ενώ διάφοροι πόροι αναλύθηκαν για την εξέταση των ακρωνύμων.

Κατά διαδικασία εκπαίδευσης και αξιολόγησης συστημάτων μηχανικής μάθησης χρησιμοποιείται κάποιο σώμα επισημασμένων δεδομένων (Labeled Data), των οποίων είναι γνωστά τα χαρακτηριστικά τους ως προς τα ζητούμενα του συστήματος. Ένα μέρος των δεδομένων αυτών σε πρώτη φάση συνεισφέρει στην εκπαίδευση του συστήματος (training set) ενώ τα υπολειπόμενα δεδομένα που δεν χρησιμοποιήθηκαν στην εκπαίδευση χρησιμοποιούνται κατά την διαδικασία αξιολόγησης του (testing set). Με τον τρόπο αυτό υπολογίζεται ένας δείκτης αξιολόγησης ο οποίος κρίνει την αποτελεσματικότητα του συστήματος. Συνήθως κατά την αξιολόγηση ενδιαφέρει η δυνατότητα να αναγνωρίζονται όλα τα δεδομένα που παρουσιάζουν ενδιαφέρον (recall) καθώς επίσης και η ορθότητα των αναγνωρισμένων δεδομένων precision.

Ένα από τα πλεονεκτήματα των μεθόδων αυτών είναι η ευελιξία τους εν συγκρίσει με τις δύο προηγούμενες μεθόδους και το γεγονός ότι επιτρέπουν την κατηγοριοποίηση όρων χωρίς να χρειάζεται να γνωρίζουμε αρκετές φορές την ακριβή διαδικασία (λεπτομέρειες του μοντέλου) που το πετυχαίνουν, ενώ επίσης μπορούν με ευκολία να συνδυαστούν με τις προηγούμενες μεθόδους. Προϋποθέτουν συχνά την ύπαρξη επισημασμένων δεδομένων (labeled/training data) τα οποία θα πρέπει να καλύπτουν ικανοποιητικά το πεδίο εφαρμογής προκειμένου το μοντέλο που θα φτιαχτεί να πετυχαίνει ικανοποιητικά αποτελέσματα.

2.2.4 Υβριδικά μοντέλα

Η προσέγγιση αυτή, όπως υποδηλώνει και το όνομά της, συνδυάζει κατάλληλα περισσότερες από μία από τις παραπάνω προσεγγίσεις, προσαρμοσμένες με κατάλληλο τρόπο για την αποτελεσματική αντιμετώπιση του κάθε ξεχωριστού είδους κειμένου. Συγκεκριμένα οι διάφορες προσεγγίσεις μπορούν τροφοδοτώντας η μια την άλλη να μοιραστούν τα προβλήματα της διαδικασίας της επεξεργασίας κειμένου με χρήση διασυνδέσεων τροφοδότησης (pipelines). Τέτοιου τύπου σχεδιασμοί συμβάλουν στην αποτελεσματικότητα συστημάτων αναγνώρισης φυσικής γλώσσας ενώ ταυτόχρονα την αναβάθμιση του συστήματος, αφού αντιμετωπίζοντας κάθε

[5] https://www.nlm.nih.gov/research/umls/index.html

[6] https://www.snomed.org/snomed-ct/five-step-briefing

στοιχείο ενός συστήματος ως διαφορετικά υποσυστήματα, επιτρέπεται με ευκολία να γίνουν μεμονωμένες αλλαγές δίχως να επηρεάζεται η συνολική δομή.

Οι περισσότερες σύγχρονες εφαρμογές αναγνώρισης κειμένου και φυσικής γλώσσας χρησιμοποιούν περισσότερες από μία από τις παραπάνω τεχνικές για να πετύχουν τον σκοπό τους. Πολλά από τα παραπάνω συστήματα που προαναφέρθηκαν και σαν παραδείγματα στις παραπάνω προσεγγίσεις στην πραγματικότητα αποτελούνται από πολυσύνθετα pipelines υποσυστημάτων-modules για να παράγουν τα προτιμότερα αποτελέσματα. Για παράδειγμα στην έρευνα [13] υλοποιούνται ταυτόχρονα όλες οι διαφορετικές προσεγγίσεις κατάλληλα για την υποπερίπτωση της αναγνώρισης φυσικής γλώσσας σε Tweets όπου το κείμενο είναι περισσότερο "ελεύθερο" σε σχέση με αυτό των ειδησεογραφικών άρθρων ή των κλινικών σημειωμάτων. Στην περίπτωση που παρουσιάζεται στο έγγραφο [14] γίνεται συνδυασμός μηχανισμών μηχανικής μάθησης και γλωσσικών κανόνων με σκοπό τον περιορισμό τον λιστών ονομάτων που θα χρειαζόταν κάποιο συμβατικό μοντέλο για την κατηγοριοποίηση τους σε κάποιο κείμενο. Φαίνεται λοιπόν ξεκάθαρα πως η σωστή ρύθμιση και η σωστή χρήση των παραπάνω προσεγγίσεων σε κάποιο σύστημα μπορεί να περιορίσει σημαντικά τους πόρους που χρειάζονται για την ανάπτυξη του εν λόγω συστήματος ενώ ταυτόχρονα να αυξήσει και την αποδοτικότητά του.

2.2.5 Αξιολόγηση Αποτελεσμάτων

Κατά την ανάπτυξη εφαρμογών επεξεργασίας και αναγνώρισης κειμένου είναι σημαντικό να γίνεται αποτίμηση των τελικών αποτελεσμάτων όσον αφορά την ακρίβεια και την ορθότητα τους. Για την υλοποίηση μίας τέτοιας αξιολόγησης καθίσταται αναγκαία η ύπαρξη ενός ήδη σχολιασμένου σώματος κειμένου, σχολιασμένο από τον άνθρωπο, το οποίο θα μπορεί άμεσα να συγκριθεί με την έξοδο της εφαρμογής για την παραγωγή των παραπάνω κριτηρίων αξιολόγησης.

Στα πλαίσια της εργασίας αυτής χρησιμοποιήθηκε η συλλογή σχολιασμένων και κατηγοριοποιημένων κριτηρίων καταλληλότητας Chia [15]. Πρόκειται για μία συλλογή αποτελούμενη από 12,409 σχολιασμένα κριτήρια, στα οποία έχουν βρεθεί και κατηγοριοποιηθεί 41,487 όροι από 14 διαφορετικές κατηγορίες και 25,017 διαφορετικές συντακτικές και νοηματικές σχέσεις από 15 τύπους σχέσεων. Η διαδικασία σύνταξης της Chia προέκυψε «χειροκίνητα» μετά από προσεκτική μελέτη επαγγελματιών στον κλάδο, για την επίτευξη της μέγιστης δυνατής ποιότητας σχολιασμών.

Η μορφή στην οποία προσφέρεται η παραπάνω συλλογή είναι ιδιαίτερα ωφέλιμη, αφού παρουσιάζει τα στοιχεία σε μορφή γράφου μεταξύ όρων και σχέσεων μεταξύ τους και με τον τρόπο αυτό διευκολύνει την δημιουργία SQL ερωτήσεων, όμοιων με αυτά που θέλει να επιτύχει η παρούσα εργασία. Πιο συγκεκριμένα όσον αφορά τον σχολιασμό των όρων/οντοτήτων, αυτές αναπαρίστανται μεταξύ άλλων από τις εξής κατηγορίες *Observation, Condition, Person, Device, Drug, Visit, Procedure* και *Measurement.*

Επίσης υπάρχουν και κάποιες επιπρόσθετες για την περιγραφή όρων άρνησης, επανάληψης και αποτελεσμάτων μετρήσεων. Οι βασικές σχέσεις μεταξύ όρων αναπαρίστανται ως λογικοί τελεστές μεταξύ σχολιασμών και αποτελούνται από τα παρακάτω:

Σχέση	Περιγραφή
AND	Σύζευξη συντακτικά
OR	Διάζευξη συντακτικά
HAS_NEGATION	Σχέση όρου με λέξη που δηλώνει άρνηση
HAS_QUALIFIER	Σχέση όρου με χαρακτηρισμό qualifier
HAS_VALUE	Σχέση όρου με τιμή
HAS_TEMPORAL	Σχέση όρου με εκφράσεις χρονικών περιορισμών

Πίνακας 2: Σχέσεις μεταξύ σχολιασμών όπως εκφράζονται στο Chia Dataset

2.3 Θεματικές επικεφαλίδες ιατρικού περιεχομένου (MeSH)

2.3.1 Εισαγωγή στο MeSH

Το Medical Subject Headings (MeSH) (Θεματικές επικεφαλίδες ιατρικού περιεχομένου) [7] αποτελεί μια εγκυκλοπαίδεια εξειδικευμένου λεξιλογίου, η οποία δημιουργήθηκε από την Αμερικανική Εθνική Ιατρική Βιβλιοθήκη (National Library of Medicine), με σκοπό την καταγραφή, κατηγοριοποίηση και αναζήτηση ιατρικών πληροφοριών και εγγράφων.

Περιλαμβάνει περισυλλογή περιγραφικών δεικτών (Subject Descriptors) θεματικών επικεφαλίδων (Subject Headings) από το MEDLINE/PubMed καθώς και από λοιπές βάσεις δεδομένων της Αμερικανικής Εθνικής Ιατρικής Βιβλιοθήκης ενώ ταυτόχρονα κατατάσσει τους διαφορετικούς

[7] https://www.nlm.nih.gov/mesh/meshhome.html

όρους σύμφωνα με τις έννοιες που περιγράφουν και συνδέει συνώνυμους όρους/ονομασίες και κοντινές έννοιες. Προσφέρει επίσης μια ιεραρχική δομή δέντρου η οποία συμβάλλει στην οργάνωση των εγγραφών από ευρύτερες σε πιο εξειδικευμένες έννοιες και κατηγορίες, δίνοντας με αυτόν τον τρόπο την δυνατότητα πιο αποτελεσματικών μεθόδων αναζήτησης για την εύρεση περιεχομένου.

Πριν αναλυθεί πιο διεξοδικά η μορφή και η ιεραρχία των όρων της βιβλιοθήκης του MeSH είναι σημαντικό να αναφερθούν οι αρχές με τις οποίες συμβαίνει αποτελεσματικά η κατηγοριοποίηση των πολυάριθμων ιατρικών όρων. Το πρωταρχικό πρόβλημα που χρίζεται να αντιμετωπιστεί κατά την διαδικασία αυτή είναι η πολυμορφία και η ποικιλία ονομασιών που χρησιμοποιούνται για τον ορισμό εννοιών από διαφορετικούς ιατρικούς κλάδους ιατρικής επιστήμης. Κάθε διαφορετικός τομέας δύναται να χρησιμοποιεί εναλλακτικές ονομασίες και μορφοποιήσεις ονομασιών, γεγονός που δημιουργεί πρόβλημα στην συλλογή και την κατάταξη τους σε μια ενιαία, οργανωμένη βιβλιοθήκη. Για τον λόγο αυτό καθιερώθηκε μια συγκεκριμένη προτίμηση όσον αφορά την αναπαράσταση και ονομασία τους με την μορφή των Θεματικών Επικεφαλίδων (Subject Headings) οι οποίες ακολουθούν κάποιες βασικές συμβάσεις, όπως θα αναλυθεί παρακάτω.

2.3.2 Τύποι των Θεματικών Επικεφαλίδων

Το MeSH περιλαμβάνει τους παρακάτω τέσσερεις τύπους όρων:

2.3.2.1 Κύριες Επικεφαλίδες (MeSH headings/Main-Headings/Descriptors)

Οι Κύριες Επικεφαλίδες αναπαριστούν έννοιες της βιοϊατρικής ορολογίας. Παίζουν πολύ σημαντικό ρόλο στο λεξιλόγιο του MeSH αφού αποτελούν το βασικό στοιχείο κατάταξης και των ανάκτησης όρων.

Παραδείγματα: *“Body Weight”, “Kidney”, “Dental Cavity Preparation”, “Self-Medication”, “Radioactive Waste”, “Brain Edema”*.

2.3.2.2 Υπό-Επικεφαλίδες (Subheadings/Qualifiers)

Οι Υπό-Επικεφαλίδες είναι συνδεδεμένες με μια κύρια επικεφαλίδα (Heading/Main Heading/Descriptor) και αποσκοπούν στην περιγραφή μιας πτυχής της ευρύτερης έννοιας της κύριας επικεφαλίδας.

Παραδείγματα: *“adverse effects”, “diagnosis”, “metabolism”, “therapy”.*

2.3.2.3 Συμπληρωματικές εγγραφές εννοιών (Supplementary Concept Records)

Οι Συμπληρωματικές εγγραφές εννοιών είναι όροι σε διαφορετική βιβλιοθήκη από αυτή του MeSH. Αφορούν κυρίως ουσίες αλλά συμπεριλαμβάνουν επίσης ορισμένα πρωτόκολλα, ιούς και σπάνιες ασθένειες.

Παραδείγματα: *“cordycepin”, “valspodar”, “tacrolimus binding protein 4”, “MOPP protocol”, “Snyder Robinson syndrome”.*

2.3.2.4 Είδη δημοσιεύσεων (Publication Characteristics or Publication Types)

Τα Είδη δημοσιεύσεων περιγράφουν τον τύπο μια δημοσίευσης όπως για παράδειγμα το είδος της και τα χαρακτηριστικά της έρευνας που αφορά.

Παραδείγματα: *“Letter”, “Review”, “Randomized controlled trial”.*

2.3.3 Δενδρική Δομή (MeSH tree)

Όπως προαναφέρθηκε παραπάνω, οι επικεφαλίδες ακολουθούν μια ιεραρχική μορφή δέντρου. Ειδικότερα, κάθε επικεφαλίδα έχει κάποια θέση στην ιεραρχία. Η ιεραρχία περιλαμβάνει 16 βασικές κατηγορίες/παρακλάδια.

- A. Anatomy
- B. Organisms
- C. Diseases
- D. Chemicals and Drugs
- E. Analytical, Diagnostic and Therapeutic Techniques and Equipment
- F. Psychiatry and Psychology
- G. Phenomena and Processes
- H. Disciplines and Occupations
- I. Anthropology, Education, Sociology and Social Phenomena

- J. Technology, Industry, Agriculture
- K. Humanities
- L. Information Science
- M. Named Groups
- N. Health Care
- V. Publication Characteristics
- Z. Geographicals

Επίσης, κάθε κατηγορία μπορεί να περιλαμβάνει πολλά επίπεδα υποκατηγοριών, όπως στο παράδειγμα που ακολουθεί.

- Anatomy
    - Body Regions
        - Torso
            - **Back**
                - Lumbosacral Region
                - Sacrococcygeal Region

Αρκετοί όροι είναι επίσης δυνατό να εμφανίζονται σε περισσότερες από μια υποκατηγορίες, όπως για παράδειγμα το «Ear» που ανήκει τόσο στο «Head» όσο και στο «Anatomy» όπως φαίνεται πιο κάτω.

- Anatomy
    - Body Regions
    - Head
        - **Ear**
- Anatomy
    - Sense Organs

- **Ear**
    - Ear, External +
    - Ear, Middle +
    - Ear, Inner +

Οι υπό-επικεφαλίδες είναι επίσης οργανωμένες σε λογικές ιεραρχικές ομάδες (οικογένειες). Ακολουθεί ένα παράδειγμα:

- therapeutic use
    - administration & dosage
    - adverse effects
    - poisoning

2.3.4 Περιγραφή Όρων

Παρακάτω φαίνονται τα αποτελέσματα μίας αναζήτησης στον MeSH Browser[8]. Παρουσιάζονται σε τέσσερεις καρτέλες. Η πρώτη περιλαμβάνει βασικές πληροφορίες για την κύρια επικεφαλίδα που αναζητείται, η δεύτερη τις υπό-επικεφαλίδες (Qualifiers), η τρίτη την θέση της επικεφαλίδας στο ιεραρχικό δέντρο του MeSH και η τέταρτη τους προτεινόμενους, συνώνυμους και πιο εξειδικευμένους όρους που της αντιστοιχούν.

Στην πρώτη καρτέλα εμφανίζονται κάποιες πληροφορίες για την ζητούμενη επικεφαλίδα/όρο αναζήτησης που βρέθηκε στην βιβλιοθήκη. Πιο συγκεκριμένα κάποιες από τις περισσότερο χρήσιμες πληροφορίες που αντλούνται:

Όρος	Περιγραφή
MeSH Heading	Το όνομα της όρου αναζήτησης

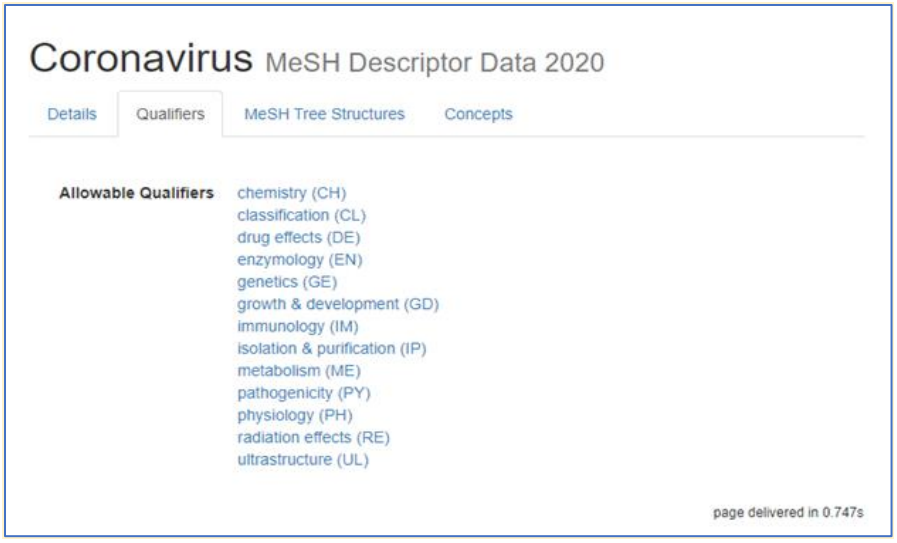
[8] https://meshb.nlm.nih.gov/search

Tree Number(s)	Το παρακλάδι του δέντρου της ιεραρχίας του MeSH στο οποίο βρίσκεται
Unique ID	Το ID του όρου στην βιβλιοθήκη
Annotation	Σύντομη περιγραφή
Entry Term(s)	Μερικοί σχετικοί όροι
Pharm Action	Η φαρμακευτική χρήση του ιατρικού όρου

Πίνακας 3: Βασικές κατηγορίες όρων του MeSH όπως εμφανίζονται στον MeSH Browser

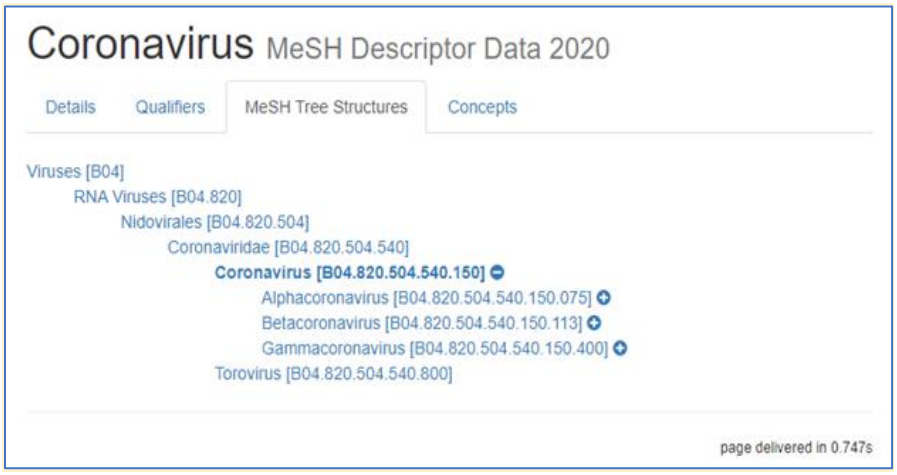
Εικόνα 6: Πρώτη καρτέλα αποτελεσμάτων αναζήτησης του MeSH-Browser με τα βασικά στοιχεία του Main Heading

Στην δεύτερη καρτέλα εμφανίζονται οι επιτρεπόμενες υπό-επικεφαλίδες (Allowable Qualifiers) που μπορούν να αντιστοιχιστούν με τον όρο. Σημειώνεται εδώ, πως κάθε όρος έχει περιορισμένο αριθμό υπό-επικεφαλίδων με τις οποίες υπάρχει νόημα να αντιστοιχισθεί.



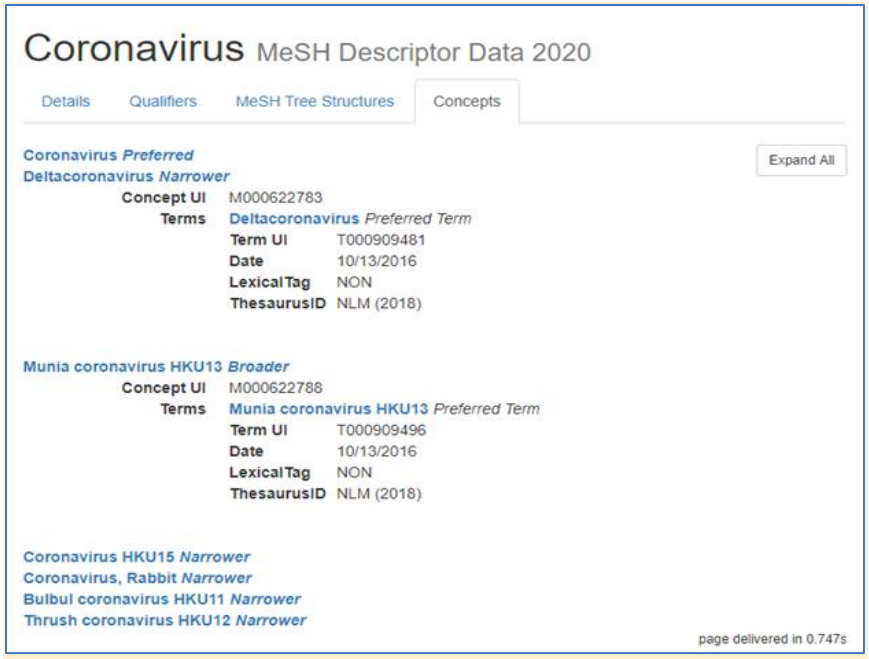
Εικόνα 7: Δεύτερη καρτέλα αναζήτησης αποτελεσμάτων MeSH-Browser με τις επιτρεπόμενες υπό-επικεφαλίδες

Στην τρίτη καρτέλα απεικονίζεται αναλυτικότερα η θέση που ο όρος αναζήτησης κατέχει στο ιεραρχικό δέντρο του MeSH. Σε αυτήν την καρτέλα είναι δυνατόν να εμφανιστούν περισσότερες από μια απεικονίσεις, εφόσον ο όρος μπορεί να συμπεριληφθεί σε πολλαπλές υποκατηγορίες.



Εικόνα 8: Τρίτη καρτέλα αναζήτησης αποτελεσμάτων του MeSH-Browser με την θέση του όρου αναζήτησης στην δενδρική δομή του MeSH

Τέλος στην τέταρτη καρτέλα παρουσιάζονται όλες οι σχετικές με τον όρο που αναζητείται έννοιες (Concepts). Χαρακτηρίζονται ως Preferred, Broader ή Narrower αν είναι προτεινόμενες, πιο ευρείες ή πιο στενές έννοιες αντίστοιχα.



Εικόνα 9: Τέταρτη καρτέλα αναζήτησης αποτελεσμάτων του MeSH-Browser με τις συναφείς έννοιες

Κάθε έννοια στην καρτέλα αναζήτησης περιέχει το όνομα και το ID της έννοιας και τους όρους που σχετίζονται με αυτήν.

Επίσης κάθε όρος σχετιζόμενος με κάποια έννοια περιέχει επίσης πληροφορία για το αν είναι προτεινόμενος ή όχι, το όνομά του και το ID του.

2.3.5 Περισσότερα για τις επικεφαλίδες (headings)

Α. Μορφή των επικεφαλίδων

Όσον αφορά τις ονομασίες των επικεφαλίδων, κατά κύριο λόγο προτιμάται η κανονική μορφή Mitral Valve από την αντεστραμμένη Valve, Mitral. Ωστόσο χρησιμοποιούνται και οι αντεστραμμένες ονομασίες σε περιπτώσεις όπου η επικεφαλίδα ανήκει σε μια υποομάδα όρων για να καθοριστεί με σαφήνεια το εύρος της. Για παράδειγμα: Psychoses, Alcoholic Psychoses, Involutional Psychoses, Senil. Ωστόσο υπάρχουν αρκετές φορές και εξαιρέσεις όπου για χρηστικούς λόγους ή για λόγους νοηματικούς προτιμάται κάποια συγκεκριμένη μορφή της επικεφαλίδας.

Ορθογραφία: Προτιμάται ο αμερικανικός τρόπος γραφής από τον βρετανικό. Για παράδειγμα: Anesthesia αντί για Anaesthesia. Επίσης και στον τρόπο γραφής των ονομάτων υπάρχουν

εξαιρέσεις όπως για παράδειγμα η ονομασία Amoeba (UK) η οποία χρησιμοποιείται αντί της Ameba (US) λόγω προτιμώμενης ονοματολογίας. Σε κάθε περίπτωση όμως όταν γίνεται αναφορά στην αντίστοιχη πάθηση χρησιμοποιείται ο όρος Amebiasis αντί του Amoebiasis.

Β. Τύποι τον επικεφαλίδων

Όπως αναφέρθηκε και στην προηγούμενη ενότητα οι επικεφαλίδες διαχωρίζονται στις εξής κατηγορίες οι οποίες παρακάτω αναλύονται πιο διεξοδικά.

Β.1. Κύριες επικεφαλίδες (Descriptors or main headings): Προσδιορίζουν το εν λόγω θέμα η περιεχόμενο που αναφέρεται.

Οι Descriptors παίζουν τον σημαντικότερο ρόλο στην ορολογία του MeSH, αφού είναι κρίσιμοι για την κατηγοριοποίηση και ανάκτηση δεδομένων από αυτό. Με εξαίρεση τους Descriptors «κλάσης 3» (βλ. παρακάτω) όλες οι υπόλοιπες κατηγορίες τους είναι οργανωμένοι και ιεραρχημένοι στην δομή δέντρου του MeSH. Ανήκουν δηλαδή σε κάποιο βασικό παρακλάδι του ιεραρχικού δέντρου και καθίσταται δυνατή η αναζήτηση τους από περισσότερο σε λιγότερο ευρείες έννοιες.

Οι Descriptors είναι κατανεμημένοι σε τέσσερεις βασικές κλάσεις όπως φαίνεται παρακάτω:

- **Class 1 Descriptors - Main Headings:** Αυτές οι εγγραφές είναι τοπικές επικεφαλίδες οι οποίες χρησιμοποιούνται για να κατατάξουν επικεφαλίδες της βάσης δεδομένων της Αμερικανικής Εθνικής Ιατρικής Βιβλιοθήκης καθώς και άλλων βάσεων αναζητήσιμων μέσω του PubMed. Οι περισσότεροι Class 1 – Descriptors αποτελούν την ανάλυση ενός όρου/έννοιας.

- **Class 2 Descriptors - Publication Characteristics (Publication Types):** Αυτές οι εγγραφές δηλώνουν το είδος του αναφερόμενου όρου. Για παράδειγμα η καταγραφή ενός άρθρου ως Historical Article. Η χρησιμότητα τους αφορά περισσότερο την περιγραφή των χαρακτηριστικών κάποιου αντικειμένου καr όχι την περιγραφή του περιεχομένου του. Στην ιεραρχία του MeSH είναι ονοματισμένα ως "PT" ή <PublicationType> αντί για "MH" ή <MeSHHeading> και βρίσκονται κάτω από την κύρια κατηγορία “V”.

- **Class 3 Descriptors - Check Tags:** Σε αυτή την κατηγορία βρίσκονται οι εξής 2 Descriptor, Check Tag “Male” και Check Tag “Female”

- **Class 4 Descriptors – Geographics:** Πρόκειται για Descriptors οι οποίοι αναφέρονται σε γεωγραφικές πληροφορίες όπως χώρες, ηπείρους, περιοχές και λοιπές υποκατηγορίες τους. Δεν χρησιμοποιούνται για την περιγραφή θεμάτων αλλά για την περιγραφή τοποθεσιών. Στην ιεραρχία του MeSH βρίσκονται κάτω από την βασική κατηγορία “Z”.

Β.2. Υπό-Επικεφαλίδες (Qualifiers/Subheadings): Χρησιμοποιούνται μαζί με τις κύριες επικεφαλίδες και αποσκοπούν στο να συνδέσουν και να ομαδοποιήσουν τα έγγραφα αυτά που αφορούν συγκεκριμένες πτυχές της έννοιας που περιγράφει η κύρια επικεφαλίδα. Για κάθε Descriptor δημιουργείται λίστα από Qualifiers οι οποίοι προσφέρουν περισσότερες δυνατότητες για πιο συγκεκριμένες και εξειδικευμένες αναζητήσεις.

Υπάρχουν 78 τοπικές Υπό-Επικεφαλίδες (Qualifiers) οι οποίες χρησιμοποιούνται για κατηγοριοποίηση σε συνδυασμό με τους περιγραφείς (Descriptors). Επιτυγχάνουν να συλλέγουν κάτω από κάθε τοπικό Qualifier τις εγγραφές αυτές οι οποίες αφορούν μία συγκεκριμένη πτυχή του θέματος ενός Descriptor. Για παράδειγμα ο συνδυασμός Descriptor/Qualifier Liver/drugeffects υποδηλώνει επιδράσεις φαρμάκων στο συκώτι.

Β.3. Συμπληρωματικές εγγραφές (Supplementary Records)

Οι Συμπληρωματικές εγγραφές, αλλιώς Συμπληρωματικές Εγγραφές Χημικών Ουσιών (Supplementary Chemical Records SCR’s) κατηγοριοποιούν χημικά, φάρμακα και λοιπές έννοιες όπως για παράδειγμα σπάνιες ασθένειες οι οποίες δεν βρίσκονται κάτω από κάποια βασική κατηγορία στο MeSH. Αντιθέτως με τους Descriptors οι SCR’s δεν υπάρχουν μέσα σε κάποια ιεραρχική δομή, συνδέονται όμως με έναν οι περισσότερους Descriptors.

- **Class 1 Supplementary Records – Χημικά (Chemicals):** Αφορούν αποκλειστικά χημικά στοιχεία και είναι συνδεδεμένα κατά βάση με κατηγορίας D (“Drug/Chemical”) Descriptors.

- **Class 2 Supplementary Records – Πρωτόκολλα (Protocols):** Αναφέρονται κυρίως σε πρωτόκολλα χημειοθεραπειών. Είναι συνδεδεμένα με την επικεφαλίδα "Antineoplastic

Combined Chemotherapy Protocols" και με χημικά που χρησιμοποιούνται στα εν λόγω πρωτόκολλα από τους Descriptor κατηγορίας D.

- **Class 3 Supplementary Records – Ασθένειες (Diseases):** Εγγραφές που αποκλειστικά αφορούν ασθένειες και είναι συνδεδεμένες με επικεφαλίδες της κατηγορίας C (Diseases) και A (Anatomy).

- **Class 4 Supplementary Records – Οργανισμοί (Organisms):** Εγγραφές οργανισμών (π.χ. ιοί) συνδεδεμένου με Descriptor κατηγορίας B (Organisms)

Γ. Όροι Καταχώρησης (Entry terms): Συνώνυμα ή στενά συνδεδεμένοι όροι οι οποίοι συνδέονται με την κύρια επικεφαλίδα με σκοπό την διευκόλυνση αναζήτησης περισσότερο σύνθετων εννοιών.

Σύνθετοι όροι στο MeSH μπορούν να περιγράφουν με τρεις διαφορετικούς τρόπους. Ο πρώτος από αυτούς είναι ο συνδυασμός δύο η περισσότερων Descriptor. Ο δεύτερος είναι ο Descriptor να συνδυάζεται με κάποιο Qualifier ώστε να δειχθεί με ακρίβεια ποια πτυχή της έννοιας αναζητείται. Τρίτον το MeSH προσφέρει ήδη εξειδικευμένους Descriptor οι οποίοι συσχετίζουν συχνά συνδεδεμένα μεταξύ τους θέματα χωρίς να είναι αναγκαία η χρήση Descriptor με Qualifier.

3 ΜΕΘΟΔΟΛΟΓΙΑ

3.1. Συνοπτική Παρουσίαση του Συστήματος

3.1.1 Περιγραφή Συστήματος

Το σύστημα αναπτύχθηκε κατάλληλα με σκοπό την αναζήτηση ασθενών που πληρούν κριτήρια καταλληλότητας κλινικών δοκιμών που αφορούν το σύνδρομο Sjogren. Αξιοποιεί λοιπόν αυτά τα κριτήρια καταλληλότητας από γνωστές δοκιμές καθώς και το σύστημα διαχείρισης πραγματικών δεδομένων ασθενών κάνοντας χρήση της οντολογίας HarmonicSS και του λεξικού MeSH για να πετύχει τον παραπάνω σκοπό.

3.1.1.1 Είσοδος

Το σύστημα δέχεται σαν είσοδο τα κριτήρια καταλληλότητας κλινικών δοκιμών ασθενών με σύνδρομο Sjögren σε μορφή κειμένου και τα ήδη σχολιασμένα (annotated) κριτήρια για την αξιολόγηση αν πρόκειται αυτή να πραγματοποιηθεί.

3.1.1.2 Έξοδος

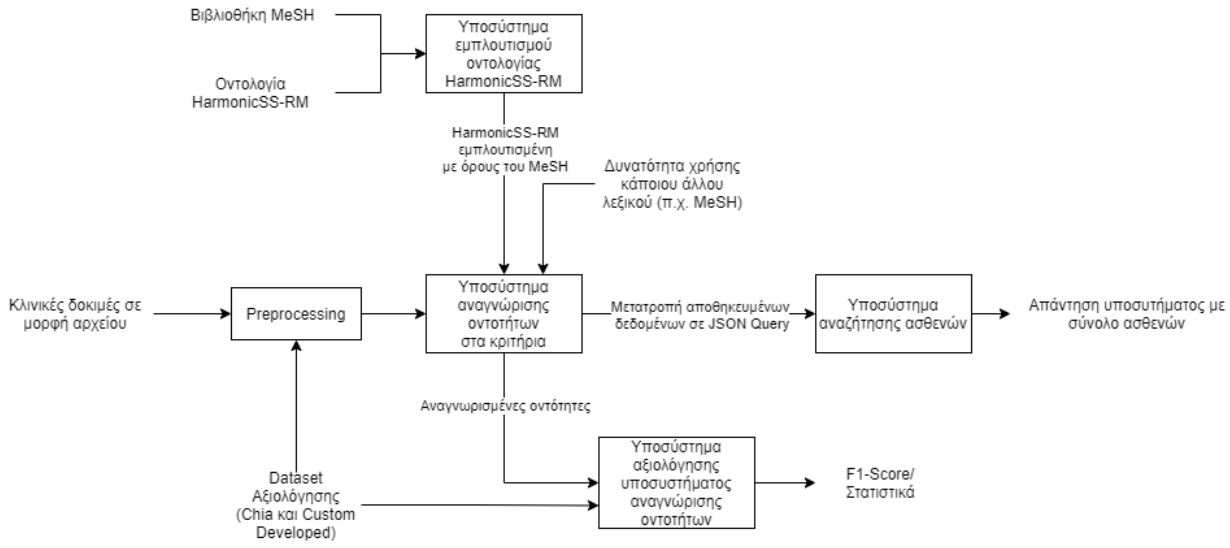
Η έξοδος του συστήματος αποτελείται από τα κριτήρια σχολιασμένα από γνωστούς όρους και από το πλήθος των ασθενών που πληρούν τα εξεταζόμενα κριτήρια καταλληλότητας.

3.1.1.3 Αρχιτεκτονική Συστήματος

Κατά την εύρεση οντοτήτων στα κριτήρια καταλληλότητας αξίζει να σημειωθεί πως γίνεται χρήση του λεξικού του MeSH, σε συνδυασμό με την οντολογία HarmonicSS. Η διαδικασία αυτή απαιτεί το δικό της υποσύστημα στην αρχιτεκτονική του συστήματος, του οποίου σκοπός είναι να αξιοποιηθεί το πλούσιο λεξιλόγιο του MeSH ώστε να προστεθούν συνώνυμες και συναφείς έννοιες στην οντολογία HarmonicSS. Ο λόγος για τον οποίο υλοποιείται το σύστημα αυτό αντί να γίνει χρήση μόνο της βιβλιοθήκης του MeSH ή της οντολογίας ως έχει, αφορά στο γεγονός πως τα δεδομένα στην βάση των ασθενών είναι εκφρασμένα σύμφωνα με μόνο την οντολογία

HarmonicSS. Επομένως η απόκτηση διαφορετικών τρόπων διατύπωσης των όρων της οντολογίας μέσω του εμπλουτισμού της με όρους του MeSH διευκολύνει τον εντοπισμό οντοτήτων, ενώ ταυτόχρονα δεν προκαλεί πρόβλημα συμβατότητας με την βάση των ασθενών.

Το σύστημα λοιπόν, απαρτίζεται α) από το υποσύστημα προεπεξεργασίας(preprocessing) των κριτηρίων, β) από το υποσύστημα εύρεσης όρων της βάσης γνώσης (HarmonicSS Οντολογία, MeSH) σε κάποιο κριτήριο καταλληλότητας, γ) από το υποσύστημα εμπλουτισμού της οντολογίας HarmonicSS με όρους της βιβλιοθήκης MeSH του National Library of Medicine, δ) του υποσυστήματος διαχείρισης και εύρεσης πλήθους ασθενών και ε) του υποσυστήματος αξιολόγησης του συστήματος.



Εικόνα 10: Διάγραμμα αρχιτεκτονικής συστήματος

Όπως φαίνεται και στο παραπάνω σχήμα, ύστερα από την προεπεξεργασία των κριτηρίων, πραγματοποιείται σε αρχικό στάδιο η διαδικασία εμπλουτισμού της οντολογίας προκειμένου να καθίσταται έτοιμη για αναγνώριση όρων στα κριτήρια καταλληλότητας. Έπειτα η εμπλουτισμένη οντολογία και τα κριτήρια τροφοδοτούνται στο υποσύστημα αναγνώρισης όρων των κριτηρίων καταλληλότητας και το υποσύστημα αξιολόγησης του συστήματος αναγνώρισης οντοτήτων αξιολογεί και παράγει αποτελέσματα σχετικά με την αποτελεσματικότητα του. Τέλος το υποσύστημα διαχείρισης ασθενών χρησιμοποιεί τα ευρήματα που εντοπίσθηκαν, τα φέρνει σε κατάλληλη μορφή ερωτήματος και πραγματοποιεί αναζήτηση στο σύστημα διαχείρισης ασθενών.

3.1.2 Συνοπτική περιγραφή των υποσυστημάτων

3.1.2.1 Υποσύστημα εμπλουτισμού της οντολογίας HarmonicSS

Το σύστημα εμπλουτισμού της οντολογίας HarmonicSS είναι το σύστημα το οποίο χρησιμοποιεί την βιβλιοθήκη του MeSH, προκειμένου να εμπλουτίσει την οντολογία με συνώνυμους και συναφείς όρους με σκοπό την διευκόλυνση εντοπισμού οντοτήτων στα κριτήρια και την δυνατότητα εντοπισμού διαφορετικών διατυπώσεων όρων και εννοιών. Το υποσύστημα αυτό, πραγματοποιεί για κάθε ζεύγος όρων του MeSH και του HarmonicSS ελέγχους και αποφασίζει αν θα γίνει προσθήκη κάποιας κατηγορίας όρων του MeSH σε κάποιον όρο τα οντολογίας. Ύστερα το σύστημα συνθέτει την νέα, εμπλουτισμένη οντολογία και προετοιμάζει τα δεδομένα της για να τα δεχτεί σαν είσοδο το σύστημα εντοπισμού οντοτήτων στα κριτήρια καταλληλότητας.

3.1.2.2 Υποσύστημα εντοπισμού οντοτήτων στα κριτήρια καταλληλότητας

Το υποσύστημα εντοπισμού οντοτήτων στα κριτήρια καταλληλότητας κάνει χρήση της οντολογίας HarmonicSS ή της βιβλιοθήκης του MeSH, αξιοποιώντας τεχνικές λεξιλογικής προσέγγισης, για τον εντοπισμό όρων τους στο κείμενο των κριτηρίων. Ταυτόχρονα το υποσύστημα εκτελεί ελέγχους αναγνώρισης αρνήσεων στις εντοπισμένες οντότητες ενώ βρίσκει επίσης και πληροφορία σχετική με χρονικούς περιορισμούς και μονάδες μέτρησης. Τέλος, παράγει ένα αρχείο JSON με τους εντοπισμένους όρους μορφοποιημένους σε μορφή ερωτήματος προορίζοντας τους προς το σύστημα διαχείρισης δεδομένων ασθενών.

3.1.2.3 Υποσύστημα αξιολόγησης του συστήματος

Προκειμένου να αξιολογηθεί η αποδοτικότητα του συστήματος εντοπισμού οντοτήτων στα κριτήρια καταλληλότητας, αναπτύχθηκε σύστημα αξιολόγησης των αποτελεσμάτων του. Το παρόν υποσύστημα, λαμβάνει ως είσοδο ήδη σχολιασμένα (annotated) κριτήρια κα συγκρίνει τους σχολιασμούς αυτούς με την έξοδο του συστήματος εντοπισμού όρων. Αξιολογεί επίσης και τον εντοπισμό των αρνήσεων στους όρους αυτούς, αλλά και τις εντοπισμένες εκφράσεις μονάδων μέτρησης και τιμών. Τέλος υπολογίζει σύμφωνα με τα αποτελέσματα του συστήματος την αποδοτικότητα σε μορφή F1-Score, καθώς επίσης παράγει και πληροφορία σχετική με τα εσφαλμένα ευρήματα.

3.1.2.4 Υποσύστημα διαχείρισης δεδομένων ασθενών

Τέλος, τα δεδομένα που αντλούνται από το υποσύστημα αναγνώρισης οντοτήτων στα κριτήρια καταλληλότητας τροφοδοτούνται στο υποσύστημα διαχείρισης ασθενών, του οποίου σκοπός είναι η μετατροπή τους σε κατάλληλη μορφή ερωτημάτων (Queries), και η αξιοποίηση τους για την αναζήτηση πλήθους ασθενών στην βάση. Ύστερα το υποσύστημα διαχειρίζεται κατάλληλα την απάντηση του ερωτήματος λαμβάνοντας υπ' όψη τα inclusion και exclusion criteria και παρουσιάζει τα αποτελέσματα της αναζήτησης.

3.1.3 Προεπεξεργασία Κειμένου Κριτηρίων καταλληλότητας

Προκειμένου τα κριτήρια καταλληλότητας να έρθουν σε κατάλληλη μορφή, τέτοια ώστε να επιτρέπεται ο εντοπισμός οντοτήτων σε αυτά, πρέπει πρώτα να περάσουν από ορισμένα βήματα προεπεξεργασίας.

Το σύστημα δέχεται αρχικά ως είσοδο τα κριτήρια καταλληλότητας είτε σε μορφή κειμένου ομαδοποιημένα σε κλινικές δοκιμές και σε κατηγορίες inclusion ή exclusion κριτηρίων, είτε σε μορφή κειμένου με την κάθε σειρά του αρχείου να περιέχει ένα κριτήριο καταλληλότητας, πληροφορίες για την κλινική δοκιμή στην οποία ανήκει και το inclusivity του. Στην δεύτερη περίπτωση, τα κριτήρια ομαδοποιούνται με τέτοιο τρόπο ώστε να πάρουν τη μορφή της πρώτης.

Ύστερα κάθε αρχείο διατρέχει το υποσύστημα της *Εικόνας 12*, το οποίο συγκεντρώνει τα κριτήρια σε κατάλληλες δομές αντικειμένων που περιέχουν την απαραίτητη πληροφορία σε μορφή συμβατή με το υπόλοιπο σύστημα.

Η τελική μορφή των κριτηρίων που επιτυγχάνεται μέσα από την διαδικασία του preprocessing είναι μορφή αντικειμένων που περιέχουν πληροφορία για κάθε token κάθε κριτηρίου κάθε κλινικής δοκιμής. Συγκεκριμένα για κάθε λέξη/token αντλείται πληροφορία σχετικά με την θέση του στο κείμενο, με την κανονικοποιημένη του μορφή.

A. Parsing: Στο υποσύστημα αυτό πραγματοποιείται το parsing και tokenization των κριτηρίων. Συγκεκριμένα κάθε κριτήριο διέρχεται από ειδικά διαμορφωμένο parser οποίος διαχωρίζει το κείμενο σε ωφέλιμα κομμάτια/λέξεις (tokens) για να μπορεί το σύστημα στην συνέχεια να τα

επεξεργαστεί ανεξάρτητα. Να σημειωθεί πως για κάθε token συγκρατείται πληροφορία για την συμβολοσειρά που το απαρτίζει και για την θέση του στο κείμενο.

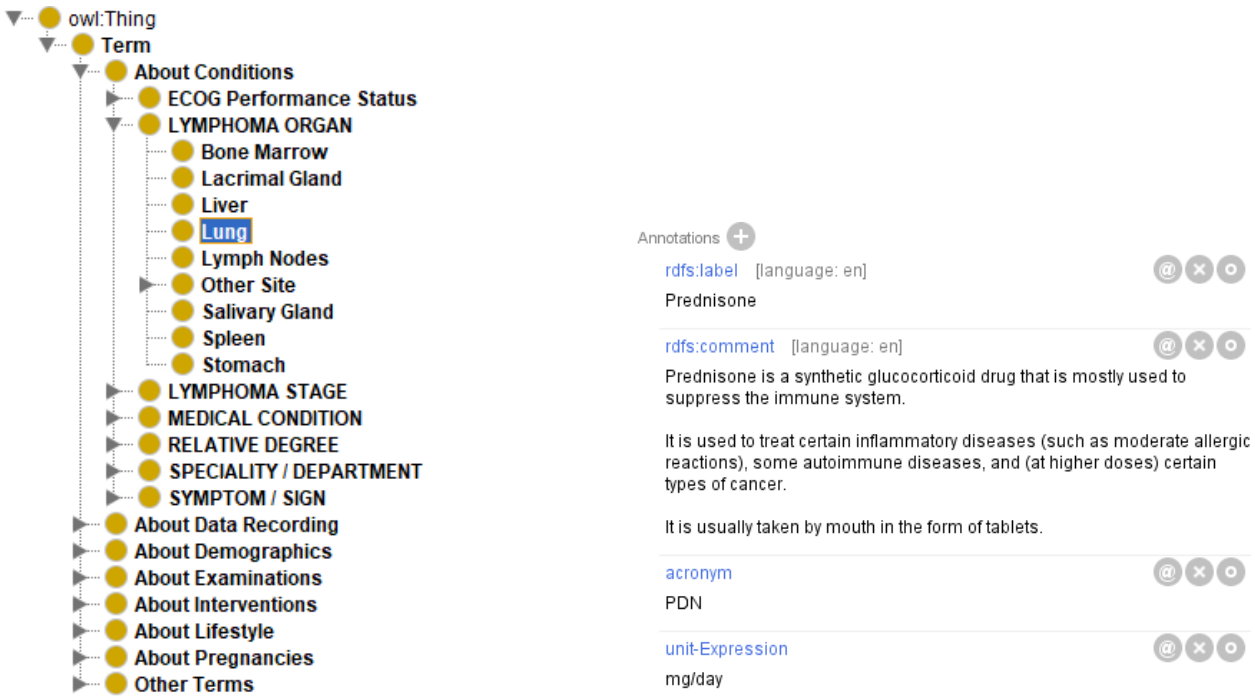
Β. **Normalization:** Ύστερα από την διαδικασία του parsing το σύστημα εφαρμόζει σε κάθε token του κάθε κριτηρίου ειδικούς κανόνες μορφοποίησης με σκοπό οι επιμέρους λέξεις του να έρθουν σε κανονικοποιημένη μορφή. Περισσότερες πληροφορίες για το εργαλείο που χρησιμοποιείται καθώς επίσης και για την διαδικασία της κανονικοποίησης κειμένου βρίσκονται στην [Ενότητα 2.1.3] και στο [Παράρτημα Α]. Η κανονικοποιημένη μορφή που προκύπτει από την παραπάνω διαδικασία αποθηκεύεται σαν επιπρόσθετη πληροφορία σε κάθε στοιχείο token.

3.1.4 Η Οντολογία HarmonicSS

Πριν γίνει μια σύντομη αναφορά στα επιμέρους υποσυστήματα του συστήματος, με σκοπό την ευκολία του αναγνώστη, δίνεται μια συνοπτική περιγραφή της οντολογίας HarmonicSS και των όρων των οποίων περιλαμβάνει.

Η οντολογία HarmonicSS αποτελείται από επιλεγμένους όρους και έννοιες που σχετίζονται με ασθενείς με σύνδρομο Sjogren. Σχεδιάστηκε από το ICCS (Institute of Communications and Computer Systems) του ΕΜΠ και έχει μορφή ιεραρχίας δέντρου αποτυπωμένη σε OWL(Ontology Web language). Για τις έννοιες που περιέχονται στην οντολογία υπάρχει πληροφορία σχετικά με την ονομασία τους και την ιεραρχία τους (π.χ. υπό-κλάση κάποιας ευρύτερης έννοιας) και πιθανόν α) κάποιο συνώνυμο αν υπάρχει, β) κάποια σχετική ακρωνυμία αν υπάρχει και γ) σε περίπτωση που η έννοια αναφέρεται σε κάποια ποσότητα ή κάποιο αποτέλεσμα ιατρικής εξέτασης τις μονάδες μέτρησης και τις εκβάσεις του αποτελέσματος αντίστοιχα.

Παρακάτω φαίνονται παραδείγματα της ιεραρχίας του HarmonicSS καθώς επίσης και παραδείγματα της πληροφορίας που είναι δυνατόν να συναντηθεί μέσα σε κάποιον όρο.



Εικόνα 11: α)Στιγμιότυπο δομής και θέσης του όρου "Lung" στην ιεραρχία της οντολογίας και β)περιεχομένων της οντολογίας HarmonicSS-RM, συγκεκριμένα του όρου "Prednisone"

Αξίζει να σημειωθεί πως η οντολογία όπως φαίνεται και στην Εικόνα 6 απαρτίζεται από τις εξής βασικές κατηγορίες.

Κατηγορία	Επεξήγηση
About Conditions	Έννοιες σχετικές με ασθένειες
About Data Recording	Έννοιες σχετικές με καταγραφή δεδομένων
About Demographics	Έννοιες σχετικές με δημογραφικά
About Examinations	Έννοιες σχετικές με εξετάσεις και διαγνώσεις
About Interventions	Έννοιες σχετικές με θεραπείες και φάρμακα
About Lifestyle	Έννοιες σχετικές με συνήθειες των ασθενών
About Pregnancies	Έννοιες σχετικές με εγκυμοσύνη
Other Terms	Έννοιες που δεν περιέχονται σε κάποια από τις παραπάνω κατηγορίες

Πίνακας 4: Οι βασικές κατηγορίες της οντολογίας HarmonicSS

3.1.5 Σύστημα Διαχείρισης Δεδομένων Ασθενών:

Το σύστημα διαχείρισης δεδομένων ασθενών περιλαμβάνει ένα σύνολο σχεσιακών βάσεων, οι οποίες έχουν σχεδιαστεί με βάση την οντολογία HarmonicSS και περιέχουν τα δεδομένα των ασθενών με σύνδρομο Sjögren. Η πρόσβαση σε αυτά επιτρέπεται στους εγγεγραμμένους χρήστες,

μέσω ενός API που δέχεται τα κριτήρια των ασθενών εκφρασμένα με βάση τους όρους της HarmonicSS οντολογίας, να αναζητήσουν το πλήθος των ασθενών που πληρούν τα κριτήρια συνολικά σε όλες τις βάσεις, με την προϋπόθεση ότι οι διαχειριστές των βάσεων το επιτρέπουν.

Στα πλαίσια αυτής της εργασίας το υποσύστημα αυτό θεωρείται δεδομένο και η επικοινωνία με αυτό βασίζεται σε ανταλλαγή JSON μηνυμάτων που περιέχουν τα κριτήρια εκφρασμένα με βάση τους όρους της HarmonicSS οντολογίας όπως τα παραδείγματα που ακολουθούν:

JSON	Περιγραφή

{
"criterion":"condition_diagnosis",
"condition":"COND-140000",
"date_until_year":"2015"
}
	condition_diagnosis: Το κριτήριο περιγράφει διάγνωση κάποιας πάθησης.
	COND-140000: Αναφέρεται στην εγγραφή της οντολογίας “Cancer”.
	date_until_year (2015): Η διάγνωση έγινε πριν το έτος 2015

{
"criterion":"intervention_medication",
"pharmacological_drug":"CHEM-10000"
}
	intervention_medication: Το κριτήριο περιγράφει λήψη φαρμάκου.
	CHEM-10000: Αναφέρεται στην εγγραφή της οντολογίας “Glucocorticoids”.

{
"criterion":"examination_lab_test",
"test_id":"OCULAR-01",
"outcome_assessment":"ASSESS-20"
}
	examination_lab_test: Το κριτήριο περιγράφει κάποιο αποτέλεσμα εργαστηριακού τεστ.
	OCULAR-01: Αναφέρεται στην εγγραφή της οντολογίας “Schrimer's Test”.
	ASSESS-20: Αναφέρεται στην εγγραφή της οντολογίας “Abnormal” και δηλώνει πως το αποτέλεσμα ήταν μη φυσιολογικό.

Πίνακας 5: Ερωτήματα σε μορφή JSON προς αναζήτηση στην βάση διαχείρισης δεδομένων ασθενών

3.1.6 Τεχνολογίες

Οι τεχνολογίες οι οποίες χρησιμοποιούνται για την ανάπτυξη του συστήματος είναι οι εξής:

Γλώσσα Προγραμματισμού:

- Java
- Javascript

Framework και βιβλιοθήκες:

- Apache JENA[9] – Για την επεξεργασία/ανάγνωση οντολογιών OWL
- Simple JSON[10] – Για την διαχείριση και δημιουργία αρχείων JSON
- SAX XML Parser - Για την διαχείριση και δημιουργία αρχείων XML
- Stanford core-NLP[11]– Βοηθητικό εργαλείο parsing (Stanford parser)
- Lexical Tools[12] – Εργαλείο του NLM για την κανονικοποίηση όρων
- Parser - Semantically-enabled context-aware abbreviations expansion in the clinical domain parser
- NegEx – Εργαλείο κανονικών εκφράσεων για την αναγνώριση αρνήσεων
- JSTree[13] – Για αναπαράσταση δομής δέντρου του HarmonicSS

Όσον αφορά το format των δεδομένων που χρησιμοποιήθηκαν, τόσο για την ανάγνωση όσο και για την παραγωγή τελικών και ενδιάμεσων δεδομένων, χρησιμοποιήθηκε η γλώσσα σήμανσης XML κατά την ανάγνωση και την επεξεργασία της βιβλιοθήκης του MeSH, η γλώσσα σήμανσης OWL για την ανάγνωση και την καταγραφή της οντολογίας HarmonicSS και η γλώσσα σήμανσης JSON για την δημιουργία βοηθητικών αρχείων με πληροφορίες της οντολογίας της βιβλιοθήκης, των κριτηρίων και του ερωτήματος αναζήτησης στο σύστημα των ασθενών.

Βοηθητικά προγράμματα και εργαλεία:

- Protégé για την απεικόνιση των όρων οντολογίας OWL
- MeSH και MeSH TreeView Browser για την αναζήτηση όρων στο MeSH

[9] https://jena.apache.org/

[10] https://github.com/fangyidong/json-simple

[11] https://stanfordnlp.github.io/CoreNLP/

[12] https://www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_003.html

[13] https://www.jstree.com/

3.2. Εμπλουτισμός της Οντολογίας του HarmonicSS με MeSH όρους

3.2.1 Περιγραφή Υποσυστήματος εμπλουτισμού της οντολογίας HarmonicSS

3.2.1.1 Είσοδος

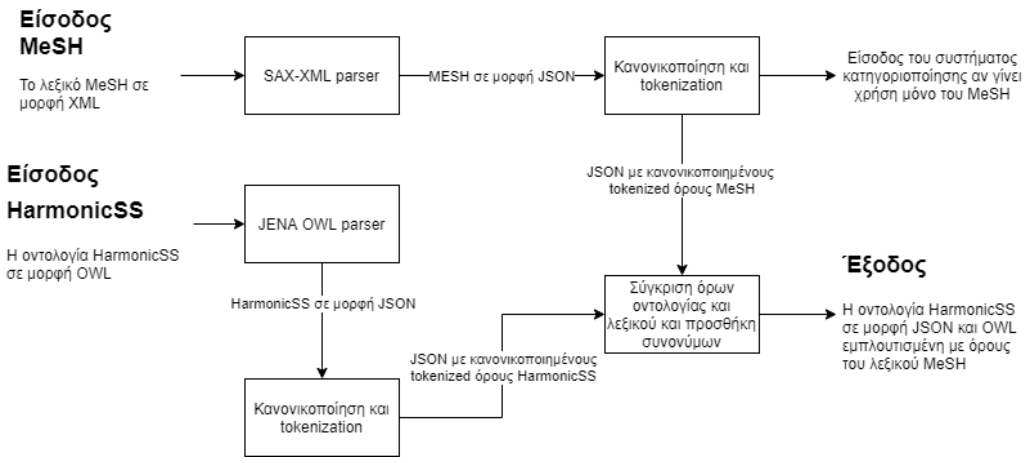
Η είσοδος του υποσυστήματος εμπλουτισμού της οντολογίας HarmonicSS είναι η οντολογία ως έχει και η βιβλιοθήκη του MeSH σε μορφή XML [Παράρτημα Β].

3.2.1.1 Έξοδος

Η έξοδος του συστήματος είναι επίσης μία οντολογία σε μορφή OWL και σε μορφή Json στην οποία όμως, έχουν προστεθεί στους όρους που έχουν αντιστοιχισθεί με την βιβλιοθήκη του MeSH τα αντίστοιχα MeSH-Concepts και MeSH-Descriptor καθώς επίσης και όλοι οι συναφείς ή συνώνυμοι όροι τους (MeSH-Terms), όροι που μπορεί να θεωρούνται λιγότερο ευρείες έννοιες και όροι οι οποίοι μπορούν να θεωρηθούν συναφείς όροι. Ταυτόχρονα από το υποσύστημα προκύπτει και το Mesh σε μορφή Json το οποίο χρησιμοποιείται για την κατηγοριοποίηση κριτηρίων με χρήση μόνο του Mesh. Για όλους του όρους που συγκρατούνται στα τελικά Json αρχεία αποθηκεύεται πληροφορία σχετικά με τις επιμέρους λέξεις/tokens τους στην κανονική και την κανονικοποιημένη τους μορφή.

3.2.1.1 Αρχιτεκτονική Υποσυστήματος

Το σύστημα αποτελείται από 3 βασικά υποσυστήματα. Το πρώτο είναι το υποσύστημα parsing και κανονικοποίησης των όρων της βιβλιοθήκης του MeSH, το δεύτερο το υποσύστημα parsing και κανονικοποίησης όρων του HarmonicSS (τα οποία παράγουν αρχεία σε μορφή JSON), και το τρίτο το υποσύστημα σύγκρισης όρων και ελέγχου ταύτισης και ομοιότητας. Η αρχιτεκτονική του συστήματος φαίνεται και στο παρακάτω διάγραμμα.



Εικόνα 12: Διάγραμμα αρχιτεκτονικής του υποσυστήματος εμπλουτισμού της οντολογίας HarmonicSS με όρους της βιβλιοθήκης του MeSH

3.2.2 Αντιστοίχιση Όρων HarmonicSS Οντολογίας με MeSH

Η κατανόηση και κατ' επέκταση η δυνατότητα της ορθής ανάγνωσης και επεξεργασίας των όρων της βιβλιοθήκης του MeSH είναι ένα σημαντικό εργαλείο που επιτρέπει την αντιστοίχιση συνωνύμων με την οντολογία HarmonicSS-RM. Η διαδικασία αυτή δίνει το πλεονέκτημα του εμπλουτισμού της ήδη υπάρχουσας οντολογίας, με τρόπο που να επιτρέπει μια πληρέστερη ανίχνευση όρων, κάνοντας χρήση συνωνύμων που πιθανόν να μην υπάρχουν στο HarmonicSS-RM. Η αντιστοίχιση της βιβλιοθήκης του MeSH με το HarmonicSS απαιτεί κάποια βήματα τροποποίησης των δεδομένων τους για να βρίσκονται σε θέση να συγκριθούν ορθά.

Παρακάτω περιγράφονται τα υποσυστήματα που αναπτύχθηκαν για την επίτευξη αυτού του σκοπού.

Υποσύστημα Parsing του λεξικού MeSH και της οντολογίας HarmonicSS

Το σύστημα κατηγοριοποίησης κριτηρίων απαιτεί το λεξικό/οντολογία που δέχεται σαν είσοδο, να βρίσκεται σε μορφή JSON. Για τον λόγο αυτό, καθώς επίσης και για να βρίσκονται όλα τα δεδομένα προς επεξεργασία σε μια μορφή, αξιοποιείται αφ' ενός ο SAX-XML parser της Java για την δημιουργία του JSON από το XML αρχείο του MeSH και αφ' εταίρου η βιβλιοθήκη JENA για την μετατροπή της οντολογίας HarmonicSS σε JSON αρχείο. Ταυτόχρονα με τον παρόν υποσύστημα λαμβάνει χώρα και το tokenization και η κανονικοποίηση των όρων όπως περιγράφεται στην παρακάτω ενότητα.

Υποσύστημα Κανονικοποίησης MeSH όρων

Αρχικά λαμβάνονται από την βιβλιοθήκη του MeSH με την μορφή XML οι κύριες επικεφαλίδες και οι συμπληρωματικές εγγραφές (Παράρτημα Β). Για κάθε κύρια επικεφαλίδα (Main Descriptor) και συμπληρωματική εγγραφή (SupplementaryRecord) θα ληφθεί το όνομα της (DescriptorName ή SupplementaryRecordName), το ID (DescriptorID)(SupplementaryRecordID) της, οι επιτρεπόμενοι προσδιορισμοί της (AllowableQualifiers) και οι κατηγορίες που της ανήκουν (TreeNumberList για την περίπτωση των κύριων επικεφαλίδων και την κλάση της συμπληρωματικής εγγραφής για την περίπτωση των συμπληρωματικών εγγραφών). Ύστερα λαμβάνονται όλες οι σχετικές πληροφορίες οι οποίες σχετίζονται με τις έννοιες (Concepts) που με την σειρά τους συμπεριλαμβάνουν όρους (Terms), διατηρώντας πάντα την ιεραρχία τους.

Για κάθε όρο/Termτου κάθε Concept του κάθε Descriptor, ο οποίος λαμβάνεται από το αρχείο XML με την χρήση του SAXParser, της βιβλιοθήκης του MeSH πραγματοποιείται η παρακάτω μετατροπή:

1. Ο όρος διασπάται στις λέξεις/tokens που τον απαρτίζουν με την χρήση του parser που περιγράφεται στην Ενότητα 2.1.2.
2. Το κάθε token που προκύπτει από το βήμα 1 έρχεται στην κανονικοποιημένη του μορφή με χρήση του εργαλείου Norm, όπως αυτό περιγράφεται στις προηγούμενες ενότητες και στο παράρτημα. Για κάθε token επίσης κρατείται και η κανονική του μορφή και η θέση που έχει μέσα στον όρο.

Παραδείγματα όρων και κανονικοποιημένων μορφών τους:

- Cancers, Salivary Gland → cancer salivary gland
- Deficiencies, Steroid Sulfatase → deficiency steroid sulfatase

Η τελική μορφή των τροποποιημένων όρων στο, έχει ιεραρχία όμοια με της βιβλιοθήκης του MeSH, με μοναδική διαφορά πως κρατούνται μόνο οι χρήσιμες πληροφορίες, οι οποίες και μορφοποιούνται για να είναι έτοιμες προς σύγκριση με την οντολογία HarmonicSS και με τα κριτήρια.

Σημειώνεται πως η μορφή JSON που προκύπτει από το παρόν υποσύστημα είναι και η μορφή που θα λάβει σαν είσοδο το σύστημα κατηγοριοποίησης κριτηρίων όταν θα δέχεται σαν είσοδο την βιβλιοθήκη του MeSH, όπως φαίνεται και στο διάγραμμα της αρχιτεκτονικής του συστήματος.

Η μορφής του αρχείου JSON που προκύπτει φαίνεται στην παρακάτω εικόνα που απεικονίζει στιγμιότυπο ενός Concept του Main Descriptor “Calcium Ionophores” του MeSH που αφορά τον όρο “Calcimycin”, για τον οποίο όπως αναφέρθηκε και παραπάνω έχει αποθηκευτεί πληροφορία για το κάθε token που τον απαρτίζει.


```
{
    "Concepts": [
        {
            "ConceptUI": "M0000001",
            "Concept": "Calcimycin",
            "Terms": [
                {
                    "TermNormalized": "calcimycin",
                    "Term": "Calcimycin",
                    "Tokens": [
                        {
                            "TokenRaw": "Calcimycin",
                            "TokenNormalized": "calcimycin",
                            "TokenIndex": 0
                        }
                    ]
                }
            ]
        }
    ],
    "DescriptorID": "D061207",
    "Descriptor": "Calcium Ionophores",
    "TreeNumbers": [
        {
            "TreeNumber": "D03.633.100.221.173"
        }
    ]
}
```


Εικόνα 13: Μορφή του αρχείου JSON με το στιγμιότυπο ενός Concept του Main Descriptor “Calcium Ionophores” του MeSH που αφορά τον όρο “Calcimycin”

Κανονικοποίηση HarmonicSS όρων και σύγκριση με MeSH

Parsing και Κανονικοποίηση

Αντίστοιχα με την διαδικασία της κανονικοποίησης και αποθήκευσης των όρων του MeSH ακολουθούνται τα ίδια βήματα 1. και 2. για κάθε όρο του HarmonicSS με τις παρακάτω διαφορές.

1. Το εργαλείο το οποίο χρησιμοποιείται για την ανάκτηση των όρων από την οντολογία η οποία είναι αποθηκευμένη σε μορφή OWL είναι το Apache Jena
2. Για κάθε όρο του HarmonicSS λαμβάνονται επίσης υπ' όψη όλοι οι συναφείς όροι που είναι αποθηκευμένη με την σήμανση aka (**A**lso **K**nown **A**s).
3. Το τελικό αρχείο Json που προκύπτει δεν διατηρεί την ιεραρχία της οντολογίας.

Σύγκριση και αντιστοίχιση

Η αντιστοίχιση γίνεται, συγκρίνοντας με κάθε κανονικοποιημένο όρο του HarmonicSS κάθε κανονικοποιημένο όρο του MeSH για πλήρη ταύτιση των επιμέρους λέξεων/token τους.

Αν έστω και ένας όρος (Term) κάποιας έννοιας (Concept) κάποιας επικεφαλίδας η συμπληρωματικής εγγραφής του MeSH βρεθεί, τότε η έννοια αυτή, όλοι οι όροι της, η επικεφαλίδα και οι κατηγορίες της, καταγράφονται σαν αντιστοιχισμένες/συνώνυμες με τον όρο του HarmonicSS που εξετάζεται.

Αν κάποιος όρος του HarmonicSS είναι υποόρος κάποιου όρου του MeSH, τότε θεωρείται πως ο όρος αυτός του HarmonicSS στην οντολογία θα έχει υποκλάση τον όρο του MeSH.

Αντίστοιχα αν κάποιος όρος του MeSH είναι υποόρος κάποιου όρου του HarmonicSS, θεωρείται πως οι δύο όροι στην οντολογία θα είναι υποκλάσεις την ίδιας κλάσης της.

Λαμβάνοντας υπ' όψη τα παραπάνω, δημιουργείται ύστερα από την διαδικασία αντιστοίχισης των δύο οντολογιών ένα νέο εμπλουτισμένο οντολογικό μοντέλο, το οποίο τους συμπεριλαμβάνει. Η νέα πληροφορία που εισάγεται στην οντολογία (με νέους σχολιασμούς), είναι η έννοια της βιβλιοθήκης του MeSH που αντιστοιχίστηκε (Concept/ConceptID), οι όροι της (Terms) και η κύρια επικεφαλίδα ή η συμπληρωματική εγγραφή στην οποία ανήκει (Descriptor/DescriptorUI ή SupplementaryRecord/SupplementaryRecordUI). Ταυτόχρονα για κάθε μία από τις παραπάνω περιπτώσεις προστίθεται σχολιασμός σχετικά με το αν ο όρος που έχει αντιστοιχισθεί έχει ρόλο συνωνύμου υποκλάσης ή υποκλάσης της ίδιας κλάσης με τον όρο HarmonicSS με τον οποίο έχει αντιστοιχισθεί.

Η παραπάνω εμπλουτισμένη οντολογία αποθηκεύεται αφ' ενός σε μορφή OWL και αφ' εταίρου σε μορφή JSON η οποία είναι και η τελική μορφή της εμπλουτισμένης οντολογίας που δέχεται σαν είσοδος το σύστημα κατηγοριοποίησης κριτηρίων.

Το παρακάτω δείγμα JSON αποτελεί στιγμιότυπο της τελικής μορφής της εξόδου του υποσυστήματος και της εισόδου του συστήματος κατηγοριοποίησης κριτηρίων. Στην περίπτωση αυτή έχουν προστεθεί σαν match στον όρο “Vaccination” του HarmonicSS όμοιοι όροι της οντολογίας του MeSH. Οι κατηγορίες “SubMatches” και “SimilarMatches” αφορούν τους όρους που δεν χαρακτηρίστηκαν συνώνυμοι, αλλά υποόροι/όμοιοι όροι.


```
{
    "SubMatches": [],
    "HarmonicAKA": [
        {
            "Aka": "Vaccination"
        }
    ],
    "Matches": [
        {
            "ConceptUI": "M0022451",
            "Concept": "Vaccination",
            "Terms": [
                {
                    "TermNormalized": "vaccination",
                    "Term": "Vaccination",
                    "Tokens": [
                        {
                            "TokenRaw": "Vaccination",
                            "TokenNormalized": "vaccination",
                            "TokenIndex": 0
                        }
                    ]
                }
            ]
        }
    ],
    "SimilarMatches": [],
    "MeshTreeNumbers": [
        {
            "TreeNumber": "E02.095.465.425.400.530.890"
        },
        {
            "TreeNumber": "E05.478.550.600.890"
        }
    ],
    "HarmonicsTerm": "Vaccination",
    "HarmonicCategory": "Drug",
    "HarmonicsTermID": "http:\/\/www.semanticweb.org\/ntua\/iccs\/harmonicss\/terminology\/vocabulary#CHEM-60000"
}
```


Εικόνα 14: Στιγμιότυπο του JSON αρχείου με τα στοιχεία της οντολογίας HarmonicSS εμπλουτισμένα με όρους της βιβλιοθήκης του MeSH (Είσοδος του υποσυστήματος κατηγοριοποίησης κριτηρίων)

3.3 Αναγνώριση Οντοτήτων στα Κριτήρια Καταλληλότητας

3.3.1 Περιγραφή Υποσυστήματος αναγνώρισης οντοτήτων

3.3.1.1 Είσοδος

Είσοδος του συστήματος αναγνώρισης οντοτήτων στα κριτήρια καταλληλότητας αποτελούν α) η εμπλουτισμένη οντολογία HarmonicSS σε μορφή JSON ή/και η βιβλιοθήκη του MeSH σε μορφή

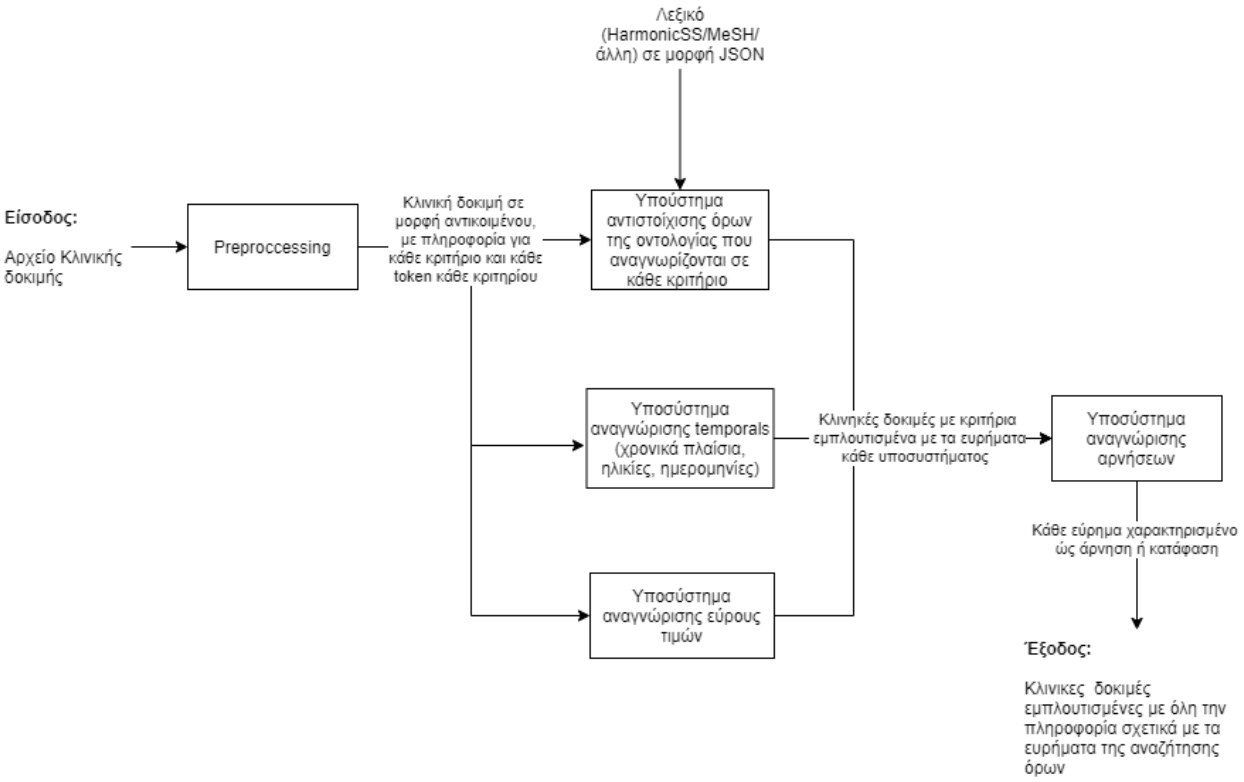
JSON (αναλόγως με το ποιο λεξικό επιθυμεί ο χρήστης να αξιοποιήσει για την αναγνώριση οντοτήτων) και β) τα κριτήρια καταλληλότητας κάθε κλινικής δοκιμής δομημένα όπως περιγράφεται κατά την διαδικασία του preprocessing.

3.3.1.1 Έξοδος

Το σύστημα παράγει ως έξοδο για κάθε κλινική δοκιμή που το διέρχεται, την ίδια κλινική δοκιμή σε μορφή αντικειμένου εμπλουτισμένη για κάθε κριτήριό της με τα ευρήματα όρους που προέκυψαν από την διαδικασία της αναγνώρισης οντοτήτων, καθώς επίσης και για κάθε εντοπισμένη οντότητα πληροφορία σχετικά με το αν κάποια από αυτές τις οντότητες έχει χαρακτηριστεί ως άρνηση (αν αναιρείται δηλαδή η όχι) και το εύρος των τιμών κάποιου χαρακτηριστικού αν αυτό έχει εντοπιστεί στο κριτήριο όπως για παράδειγμα κάποια ημερομηνία ή χρονική περίοδος και κάποια μονάδα μέτρησης.

3.3.1.1 Αρχιτεκτονική Υποσυστήματος

Συγκεκριμένα το σύστημα αναγνώρισης οντοτήτων δέχεται κάθε κλινική δοκιμή σε μορφή αντικειμένου μορφοποιημένη κατάλληλα ώστε να έχει αποθηκευτεί για αυτήν πληροφορία σχετικά με το κάθε κριτήριο που περιλαμβάνει και εν συνεχεία για κάθε κριτήριο συγκρατείται πληροφορία για κάθε



Εικόνα 15: Διάγραμμα αρχιτεκτονικής του υποσυστήματος αναγνώρισης οντοτήτων

3.3.2. Διαδικασία Αυτόματου Εντοπισμού Οντοτήτων

3.3.2.1. Αναγνώριση Όρων Λεξιλογίου

Σε πρώτη φάση, στην διαδικασία εντοπισμού όρων στα κριτήρια και στη συνέχεια στην κατηγοριοποίηση τους, γίνεται χρήση, είτε της εμπλουτισμένης οντολογίας HarmonicSS καθώς επίσης και της κανονικοποιημένης μορφής των όρων της, είτε όποιου άλλου λεξιλογίου (π.χ. MeSH) έχει επιλεχθεί για την κατηγοριοποίηση και έχει μετατραπεί κατά την διαδικασία του Pre-Processing σε κατάλληλη μορφή. Η συνολική πληροφορία σχετικά με τα κριτήρια καταλληλότητας κάθε κλινικής δοκιμής αποτελεί την είσοδο του συστήματος και είναι -όπως προαναφέρθηκε- συγκεντρωμένη σε δομή αντικειμένου το οποίο συνοπτικά περιλαμβάνει:

- Το κάθε κριτήριο της κλινικής δοκιμής σε μορφή συμβολοσειράς
- Το κάθε κριτήριο κλινικής δοκιμής διασπασμένο σε επιμέρους λέξεις/tokens για τα οποία συγκρατείται η αρχική τους μορφή από το κείμενο, η κανονικοποιημένη τους μορφή και η θέση/index τους στο αρχείο της κλινικής δοκιμής

Ταυτόχρονα σε κάθε αντικείμενο κριτηρίου δημιουργείται ένα χαρακτηριστικό, το οποίο θα συμπεριλάβει, ύστερα από την διαδικασία της αναγνώρισης οντοτήτων, πληροφορία σχετικά με τους όρους με τους οποίους αυτό αντιστοιχίσθηκε. Για τον κάθε όρο που θα αντιστοιχισθεί επίσης θα συγκρατούνται:

- Όλα τα χαρακτηριστικά του που αντλούνται από την οντολογία από την οποία προέκυψε όπως για παράδειγμα η κατηγορία του, το ID του και η συμβολοσειρά του στην αρχική και την κανονικοποιημένη του μορφή.
- Οι λέξεις/token του κριτηρίου με τις οποίες αντιστοιχίστηκε
- Η αρχική και η τελική θέση που ο όρος αυτός εντοπίσθηκε στο κείμενο.
- Κάποιο άλλο εύρημα το οποίο συσχετίζεται με το παρόν, όπως για παράδειγμα ένα εύρημα μονάδας μέτρησης ή ηλικίας
- Το γεγονός αν το εύρημα έχει υποστεί άρνηση

3.3.2.2 Δημιουργία της κανονικοποιημένης μορφής των κριτηρίων

Όσον αφορά την κανονικοποιημένη μορφή των κριτηρίων (η οποία λαμβάνει χώρα κατά την διαδικασία του preprocessing Ενότητα 3.1.3) αξίζει να σημειωθεί, πως κάθε κριτήριο διασπάται στις επιμέρους λέξεις/tokens του με την βοήθεια του parser που αναφέρεται στην ενότητα 2.1.2. Ύστερα με την σειρά του κάθε token κανονικοποιείται σύμφωνα με τις εντολές του εργαλείου Norm. Με τον τρόπο αυτό αντλείται από το κάθε κριτήριο η κανονικοποιημένη μορφή κάθε λέξης χωρίς να επηρεάζεται με κάποιον τρόπο η σειρά της μέσα στο κριτήριο. Ταυτόχρονα λέξεις κενού περιεχομένου (stop words) αφαιρούνται. Πλέον δηλαδή το κριτήριο βρίσκεται σε κατάλληλη για επεξεργασία μορφή ώστε κάθε όρος του λεξικού που θα χρησιμοποιηθεί να μπορεί να αντιστοιχισθεί σε αυτό με ευκολία.

Παρακάτω δίνονται παραδείγματα κριτηρίων και των κανονικοποιημένων μορφών τους:

Αρχική μορφή	**Κανονικοποιημένη μορφή**
Patients taking DMARD's, such as hydroxychloroquine, must be on a stable dose.	patient take dmard such as hydroxychloroquine must be a stable dose
History of immunoglobulin E (IgE)-mediated or non-IgE-mediated hypersensitivity	history immunoglobulin e ige mediate non ige mediate hypersensitivity
Patients who have developed dry mouth clearly due to a cause other than Sjögren’s syndrome	patient who have develop dry mouth clearly due a cause other than sjogren syndrome

Πίνακας 6: Αρχική και κανονικοποιημένη μορφή κριτηρίων καταλληλότητας

3.3.2.3 Αναζήτηση Όρων Οντολογίας στα Κριτήρια Καταλληλότητας

Όπως αναφέρεται και στην περιγραφή της αρχιτεκτονικής του συστήματος σε πρώτη φάση γίνεται η αντιστοίχιση στο κριτήριο των όρων, οι οποίο προκύπτουν με χρήση της λεξιλογικής αντιστοίχισης με κάποια οντολογία (HarmonicSS, MeSH). Η εκτέλεση της παραπάνω διαδικασίας έχει ως εξής:

- Για κάθε όρο του λεξικού αναζήτησης ελέγχεται αν όλα τα επιμέρους token που τον απαρτίζουν βρίσκονται μέσα στο κριτήριο καταλληλότητας καθώς επίσης και πόσες φορές αυτά εντοπίζονται μέσα στο κείμενο του κριτηρίου.
- Αφού γίνει η παραπάνω αναζήτηση και πριν προστεθεί το εύρημα/τα ευρήματα στο αντικείμενο κριτηρίου αποθηκευμένο με τον τρόπο που προαναφέρθηκε, ελέγχεται αν ο όρος που βρέθηκε αποτελεί μέρος/υποσύνολο ή περιέχει όρους που έχουν ήδη αναγνωριστεί από την παρούσα διαδικασία σε προηγούμενα βήματα και συγκρατείται πάντα ο όρος που περιέχει τα περισσότερα tokens.

Χαρακτηριστικό ρόλο στο κομμάτι του λεξιλογικού εντοπισμού όρων στα κριτήρια καταλληλότητας παίζει η δυνατότητα του συστήματος να εντοπίζει τα σωστά tokens κάποιου όρου μέσα στο κριτήριο, καθώς επίσης και να εντοπίζει όρους που απαρτίζονται από tokens τα οποία απέχουν μεταξύ τους κάποια απόσταση (η οποία είναι ορισμένη από την είσοδο του προγράμματος). Παρακάτω περιγράφονται τα βήματα που ακολουθούνται για να επιτευχθεί μια περισσότερο ικανοποιητική αντιστοίχιση και να παράγεται μικρότερος όγκος λαθών. Αναλυτικότερα λοιπόν ο αλγόριθμος για κάθε όρο του λεξιλογίου που εξετάζεται:

1. Λαμβάνει όλα τα tokens του σε κανονικοποιημένη μορφή και ελέγχει αν υπάρχουν όλα μέσα στο κανονικοποιημένο κριτήριο ανεξαρτήτως την θέση τους και την απόσταση μεταξύ τους και τα συγκρατεί.

   Στο παρακάτω παράδειγμα ο αλγόριθμος στο παρόν βήμα όταν αναζητήσει τον όρο “Hepatitis B” θα συλλέξει τις λέξεις που είναι σημειωμένες με bold.

   “*Patients that have been diagnosed with **Hepatitis B** or another **hepatitis** virus.”*
2. Υπολογίζει αν ο όρος υπάρχει πολλαπλές φορές μέσα στο κριτήριο και συγκρατεί κάθε αναφορά του σε αυτό με τις αντίστοιχες θέσεις τους στο κείμενο. Για παράδειγμα στο κριτήριο:

"Male *patients that have been diagnosed with **Hepatitis B** or female patients that have been diagnosed with the **hepatitis B** virus."*

Ο αλγόριθμος θα συλλέξει 2 εγγραφές για τον όρο "Hepatitis B".

3. Πραγματοποιεί έλεγχο της απόστασης των επιμέρους tokens που συλλέχθηκαν σύμφωνα με την μέγιστη επιτρεπόμενη απόσταση που δίνεται σαν παράμετρος στο σύστημα. Ταυτόχρονα φροντίζει σε περίπτωση που μπορεί να αναθέσει σε κάποιο εύρημα όρου το ίδιο token από διαφορετικές θέσης της πρότασης, να δίνεται αυτό που είναι κοντινότερο στα υπόλοιπα. Δεδομένου λοιπόν πως η μέγιστη επιτρεπόμενη απόσταση ορίζεται ως 2, είναι δυνατόν να αντιστοιχισθούν ευρήματα τα οποία χωρίζονται από κάποια μικρής σημασίας λέξη/token.
4. Ελέγχει για κάθε στιγμιότυπο του όρου μέσα στο κριτήριο, αν αποτελεί υποσύνολο κάποιας ευρύτερης έννοιας που έχει βρεθεί ή αν κάποιες έννοιες που έχουν ήδη βρεθεί αποτελούν υποσύνολό του.
    a. Στην πρώτη περίπτωση ο όρος απορρίπτεται αφού ευρύτερη έννοια έχει ήδη εντοπιστεί. Για παράδειγμα αν στο παραπάνω παράδειγμα έχει ήδη εντοπιστεί ο όρος "Hepatitis B" και εξετάζεται ο όρος "Hepatitis", το σύστημα δεν θα προσθέσει νέο εύρημα.
    b. Στην δεύτερη περίπτωση αν βρεθεί όρος για τον οποίο στο κείμενο έχουν ήδη αναγνωριστεί σαν ευρήματα υποόροι του στην θέση στην οποία εντοπίσθηκε και αυτός, τότε οι εν λόγω υποόροι διαγράφονται από πιθανές αντιστοιχίες και αντικαθίστανται από τον όρο που εξετάζεται. Για παράδειγμα αν στο παραπάνω κριτήριο είχαν πρώτα εντοπισθεί τα δύο tokens του όρου "hepatitis" και εξετάζεται έπειτα ο όρος "Hepatitis B" τα προηγούμενα ευρήματα θα διαγραφούν και θα αντικατασταθούν από τα στιγμιότυπα του "Hepatitis B".
5. Τέλος τα ευρήματα που ικανοποιούν όλους τους παραπάνω περιορισμούς προστίθεται στο κριτήριο σαν ευρήματα. Για τα ευρήματα αυτά συγκρατείται όλη η πληροφορία σχετικά με την θέση τους στο κείμενο, με τα χαρακτηριστικά και την κατηγορία τους στο λεξικό/οντολογία και σχετικά με την πραγματική και κανονικοποιημένη τους μορφή.

3.3.2.4 Υποσύστημα Αναγνώριση Άρνησης

Η διαδικασία αναγνώρισης άρνησης γίνεται αφού οι όροι εντοπιστούν μέσα σε κάποιο κριτήριο με την βοήθεια του εργαλείου NegEx. Το εν λόγω εργαλείο λαμβάνει σαν είσοδο κάποια έκφραση η οποία έχει αναγνωρισθεί στο κριτήριο καθώς και το κριτήριο αυτό καθ' αυτό και επιστρέφει για την δοθείσα έκφραση, ως αποτέλεσμα, την "αρνητικότητα" της. Αν δηλαδή η λέξη σύμφωνα με το NegEx έχει μη καταφατικό χαρακτήρα επιστρέφει "Negation", αν όχι τότε "Affirmation" αλλιώς αν δεν είναι δυνατόν κάποια φράση να καταταχθεί σε κάποια κατηγορία επιστρέφει "Uncertain" στην οποία περίπτωση η φράση λαμβάνεται ως άρνηση. Η απόφαση του NegEx μαζί με το "Inclusivity" του κριτηρίου συντάσσουν εν τέλη τον τελικό χαρακτήρα του κριτηρίου.

3.3.2.5 Εντοπισμός χρονικών εκφράσεων/περιορισμών/τιμών στα κριτήρια

Για τον εντοπισμό των χρονικών περιορισμών και εκφράσεων καθώς επίσης και εύρους τιμών στα κριτήρια καταλληλότητας, λήφθηκε λίστα εκφράσεων που εμφανίζονται συχνά σε κείμενα κλινικών δοκιμών, για την οποία δημιουργήθηκε κατάλληλος αλγόριθμος εντοπισμού κανονικών εκφράσεων (regular expressions) ώστε να εντοπίζονται στιγμιότυπά τους στα κριτήρια.

Η μορφή της λίστας των παραπάνω εκφράσεων που χρησιμοποιήθηκαν φαίνεται παρακάτω με παράδειγμα μερικές από τις εκφράσεις που δηλώνουν εύρη τιμών:

```
public static final List<String> patternUnitsRangeOperator = new ArrayList<>(
    Arrays.asList(
        "OPERATOR NUMBER UNIT / UNIT x UNL"          ,
        "OPERATOR NUMBER UNIT / UNIT UNL"            ,
        "OPERATOR NUMBER UNIT / UNIT x ULN"          ,
        "OPERATOR NUMBER UNIT / NUMBER UNIT"         ,
        "OPERATOR NUMBER UNIT / UNIT ULN"            ,
        "OPERATOR NUMBER UNIT / UNIT"                ,
        "OPERATOR NUMBER / UNIT"                     ,
        "OPERATOR NUMBER UNIT x UNL"                 ,
        "OPERATOR NUMBER UNIT UNL"                   ,
        "OPERATOR NUMBER UNIT x ULN"                 ,
        "OPERATOR NUMBER UNIT ULN"                   ,
        "OPERATOR NUMBER UNIT"                       ,
        "OPERATOR NUMBER x UNL"                      ,
        "OPERATOR NUMBER UNL"                        ,
        "OPERATOR NUMBER x ULN"                      ,
        "OPERATOR NUMBER ULN"                        ,
        "OPERATOR NUMBER"
));
```

Εικόνα 16: Στιγμιότυπο της λίστας εκφράσεων που δηλώνουν έκφραση εύρους τιμών

Ο αλγόριθμος αναλυτικότερα αναζητά το κριτήριο για κάθε αντιστοίχιση κάποιας συμβολοσειράς με την παραπάνω λίστα και συγκρατεί όλα τα διαφορετικά ευρήματα. Σημειώνεται πως σε

περίπτωση που περισσότερες από μια αντιστοιχίες καταλαμβάνουν ίδιες θέσεις του κειμένου, συγκρατείται πάντα η μεγαλύτερη.

Για παράδειγμα στο κριτήριο **“Doses of >10mg/day”** , το σύστημα θα αναγνωρίσει την έκφραση “**>10mg/day**”.

Παραδείγματα ευρημάτων σε κριτήρια μαζί με την ιεραρχία της οντολογίας HarmonicSS-RM στην οποία ανήκουν:

Παράδειγμα 1:

Κριτήριο Καταλληλότητας:

Females with a diagnosis of Primary Sjogren's Syndrome.

HarmonicSS όροι που εντοπίστηκαν:

- **Sjogren's Syndrome** - [Term, About Conditions, MEDICAL CONDITION, Auto-immune Disease, Sjogren's Syndrome]
- **Females** - [Term, About Demographics, SEX, Female]

Παράδειγμα 2:

Κριτήριο Καταλληλότητας:

Patients taking DMARD's, such as hydroxychloroquine, must be on a stable dose.

HarmonicSS όροι που εντοπίστηκαν:

- **DMARD** - [Term, About Interventions, DRUG / SUBSTANCE, Disease-modifying antirheumatic drugs]
- **Hydroxychloroquine** - [Term, About Interventions, DRUG / SUBSTANCE, Disease-modifying antirheumatic drugs, Conventional Disease-modifying Antirheumatic Drug, Hydroxychloroquine]

3.4 Αξιολόγηση Συστήματος

3.4.1 Περιγραφή Υποσυστήματος Αξιολόγησης

3.4.1.1 Είσοδος

Η διαδικασία αξιολόγησης του συστήματος λαμβάνει σαν είσοδο τα κριτήρια καταλληλότητας, ομαδοποιημένα σε κλινικές δοκιμές και σχολιασμένα με τα αποτελέσματα/ευρήματα του υποσυστήματος κατηγοριοποίησης κριτηρίων, καθώς επίσης και τους ήδη σχολιασμένους όρους του συνόλου αξιολόγησης σε μορφή κειμένου, είτε του Chia-Dataset, είτε προσαρμοσμένους στο εύρος των στοιχείων της οντολογίας HarmonicSS.

3.4.1.2 Έξοδος

Η έξοδος του συστήματος αποτελείται από την βαθμολογία F1-Score που επιτυγχάνει το σύστημα. Παράλληλα παράγονται και στατιστικά για κάθε κλινική δοκιμή και δίνεται δυνατότητα αναλυτικής παρουσίασης των σωστών ευρημάτων και των λαθών του συστήματος κατηγοριοποίησης.

3.4.1.3 Αρχιτεκτονική Υποσυστήματος

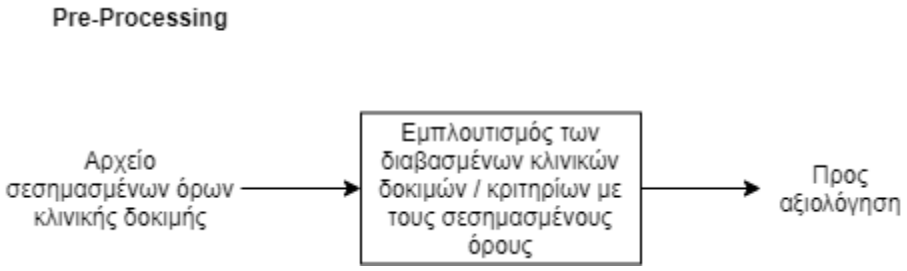
Σε πρώτο στάδιο, αν το σύστημα πρόκειται να αξιολογήσει τα αποτελέσματα του, πραγματοποιείται κατά την διαδικασία ανάγνωσης των κλινικών δοκιμών, ανάγνωση των ήδη σχολιασμένων όρων οι οποίοι προστίθενται σαν χαρακτηριστικά του κάθε κριτηρίου της κάθε κλινικής δοκιμής που αποθηκεύεται.

Ύστερα το υποσύστημα αξιολόγησης διαχειρίζεται τις κλινικές δοκιμές αξιολογώντας κάθε μία εφαρμόζοντας της ελέγχους που εξετάζουν:

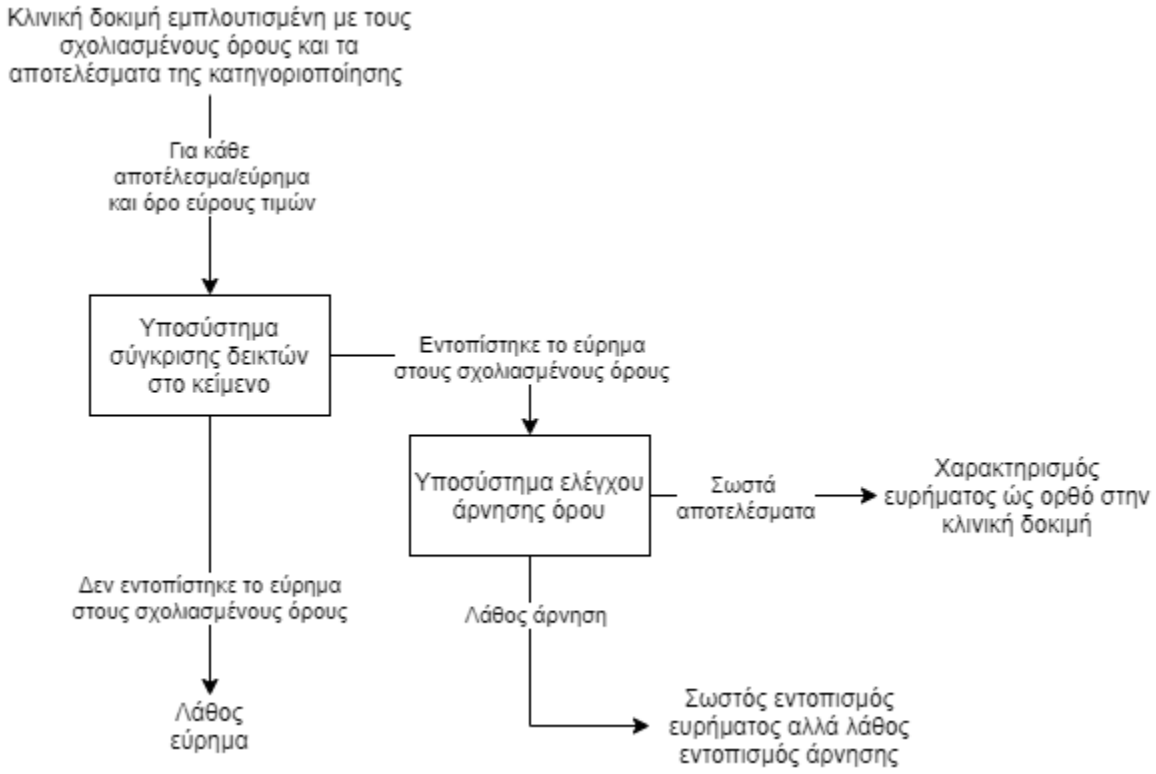
1) Την ύπαρξη του όρου που εντόπισε το σύστημα κατηγοριοποίησης στην ίδια θέση του κειμένου στα σχολιασμένα κριτήρια
2) Αν έχει εντοπιστεί σωστά η άρνηση στους όρους που έχουν εντοπισθεί
3) Αν έχουν εντοπισθεί σωστά οι εκφράσεις ημερομηνιών και τιμών στα κριτήρια

Μετά από κάθε δοκιμή κρατείται το πλήθος των ευρημάτων, σωστών (true positives) και μη (false positives), καθώς επίσης και το πλήθος των σεσημασμένων όρων που θα έπρεπε να βρεθούν (true

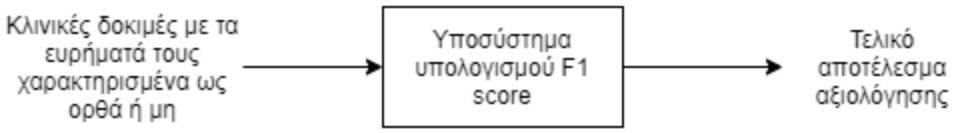
positives + false negatives) και τροφοδοτούνται μετά την προσπέλαση όλων των κλινικών δοκιμών που υπάρχουν στο Dataset σε μία μέθοδο υπολογισμού της Ανάκλησης (recall) και Ακρίβειας (precision) και κατ' επέκταση του F1(*)-Score. Στο παρακάτω διάγραμμα φαίνεται η αρχιτεκτονική του συστήματος αναλυτικότερα.



Εικόνα 17: Η διαδικασία κατά την οποία αν το σύστημα προορίζεται για αξιολόγηση, αποθηκεύονται μαζί με τις κλινικές δοκιμές και οι σεσημασμένοι όροι/οντότητες



Εικόνα 18: Αναλυτικό διάγραμμα του υποσυστήματος που ορίζει κάθε εύρημα του συστήματος αναγνώρισης οντοτήτων ως ορθό ή λανθασμένο



Εικόνα 19: Το υποσύστημα που παράγει το τελικό F1-score και κατά συνέπεια την τελική αξιολόγηση του συστήματος λαμβάνοντας υπ' όψη τα δεδομένα που παράγει το υποσύστημα αξιολόγησης

3.4.2 Η μορφή των Dataset αξιολόγησης (Harmonic-SS, Chia Annotated Criteria)

Όσον αφορά τα Dataset τα οποία χρησιμοποιούνται για την αξιολόγηση του συστήματος, γίνεται χρήση των σχολιασμένων κριτηρίων του συστήματος Chia καθώς επίσης και σχολιασμένων κριτηρίων με κατάλληλα επιλεγμένους σχολιασμούς όρων, ώστε να αντιστοιχούν σε όρους και έννοιες που συμπεριλαμβάνονται στην οντολογία HarmonicSS, όπου τα ευρήματα περιορίζονται σε αυτήν.

Η αξιολόγηση μέσω δύο διαφορετικών συνόλων αξιολόγησης γίνεται για την απόκτηση μιας περισσότερο σφαιρικής εικόνας όσον αφορά την αποτελεσματικότητα του συστήματος. Δίνεται δηλαδή η δυνατότητα αφ' ενός να κριθεί το σύστημα ως προς την αποτελεσματικότητα του και αφ' εταίρου να κριθεί η καταλληλότητα κάποιου λεξικού για τον εντοπισμό ευρημάτων σε ένα ορισμένο σύνολο κλινικών δοκιμών.

Παρακάτω γίνεται ανάλυση της μορφής των Dataset.

3.4.2.1 Η μορφή του Chia Dataset

Το Chia Dataset περιλαμβάνει εκτενή και αναλυτική πληροφορία για τα κριτήρια καταλληλότητας. Διαχωρίζεται σε αρχεία, με το κάθε ένα να αφορά μια κλινική δοκιμή και ένα inclusivity και κατά συνέπεια περιέχει σεσημασμένους όρους από όποιο κριτήριο καταλληλότητας ανήκει σε αυτήν.

Παρακάτω δίνεται ένα παράδειγμα σεσημασμένης κλινικής δοκιμής (μαζί με τα κριτήριά της) πριν αυτή αναλυθεί διεξοδικά:

```
1 Have had one prior platinum-based chemotherapy regimen for the treatment of primary disease.
2 At least 4 weeks since last surgery or radiation therapy.
3 Must have had a treatment-free interval of greater than 6 months following response to platinum.
4 ECOG performance status of 0,1, or 2.
```

Εικόνα 20: Κριτήρια καταλληλότητας της κλινικής δοκιμής NCT00061308_incl

```
R1  AND Arg1:T2 Arg2:T1
R2  Has_temporal Arg1:T1 Arg2:T3
R3  Has_index Arg1:T4 Arg2:T5
R4  Has_index Arg1:T4 Arg2:T6
*   OR T5 T6
R5  Has_index Arg1:T8 Arg2:T9
R6  multi Arg1:T9 Arg2:T10
R7  Has_temporal Arg1:T7 Arg2:T8
*   OR T12 T13 T14
R8  Has_value Arg1:T11 Arg2:T12
T1  Drug 19 54  platinum-based chemotherapy regimen
T2  Condition 76 91 primary disease
T3  Temporal 13 18  prior
T4  Temporal 94 150 At least 4 weeks since last surgery or radiation therapy
T5  Reference_point 117 129 last surgery
T6  Reference_point 133 150 radiation therapy
T7  Condition 167 192   a treatment-free interval
T8  Temporal 196 248    greater than 6 months following response to platinum
T9  Reference_point 228 248 response to platinum
T10 Drug 240 248    platinum
T11 Measurement 251 274 ECOG performance status
T12 Value 278 279   0
T13 Value 278 281   0,1
T14 Value 287 288   .
```

Εικόνα 21: Τα κριτήρια της κλινικής δοκιμής σχολιασμένα στο Chia Dataset

Αξίζει να σημειωθεί πως για κάθε κλινική δοκιμή το Chia Dataset διατηρεί δύο αρχεία κριτηρίων. Ένα για τα κριτήρια που είναι INCLUSION Criteria και ένα για τα EXCLUSION Criteria και για τον λόγο αυτό στο παραπάνω παράδειγμα δεν φαίνεται κάποια σχετική πληροφορία στο περιεχόμενο του αρχείο για το Inclusivity.

Όπως αναγράφεται στο αρχείο της Εικόνας 23 δίνεται πληροφορία για το είδος του ευρήματος νοηματικά και συντακτικά, για την κατηγορία του, την θέση του στην κλινική δοκιμή (ή τους όρους που αφορά, άμα πρόκειται για κάποια σχέση συντακτική η νοηματική) και τέλος τον όρο.

Συγκεκριμένα οι σημάνσεις:

- **T1, T2, T3, ...**: Αφορούν όρους (Terms)
- **R1, R2, R3, …:** Αφορούν σχέσεις μεταξύ όρων και ακολουθούνται από τον τύπο της σχέσης και τις παραμέτρους της (π.χ. όρους T1, T5)
- *** OR:** Αναφέρεται στην διάζευξη

Τέλος αξίζει να σημειωθεί, πως οι μετρήσεις δοσολογιών, οι ημερομηνίες και οι αρνήσεις σημαίνονται ως όροι (T) με ακόλουθο σχόλιο την αντίστοιχη κατηγορία στην οποία αναφέρονται (Value, Measurement, Temporal, Negation).

3.4.2.1 Η μορφή του Dataset σύμφωνα με την οντολογία HarmonicSS

Το Dataset το οποίο χρησιμοποιείται για την αξιολόγηση του συστήματος, μόνο όσον αφορά όρους οι οποίοι συσχετίζονται και εμπεριέχονται στην οντολογία HarmonicSS και προφανώς στους όρους με τους οποίους εμπλουτίστηκε από την βιβλιοθήκη του MeSH, λαμβάνει μορφή όμοια με αυτήν του Chia Dataset, τόσο για λόγους συμβατότητας όσο και για λόγους συνοχής μεταξύ των αξιολογήσεων του συστήματος.

Βασική διαφορά της μορφής μεταξύ του παρόντος αρχείου και αυτού του Chia είναι το γεγονός πως στην στήλη στην οποία αναγράφεται η κατηγορία του ευρήματος βρίσκεται το URI που αντιστοιχεί σε κάποιον όρο της οντολογίας του μοντέλου HarmonicSS-RM. Τα κριτήρια με όμοιο τρόπο ομαδοποιούνται σύμφωνα με το Inclusivity και το ID της κλινικής δοκιμής στην οποία ανήκουν.

3.4.3 Η μορφή του αποτελέσματος

Το αποτέλεσμα της αξιολόγησης αποτελείται από τρία στοιχεία. Το τελικό F1-Score και τις παραμέτρους του Precision και Recall (βλ. Ενότητα 3.4.6).

Ταυτόχρονα για κάθε κλινική δοκιμή δίνονται λεπτομέρειες σχετικά με τους όρους οι οποίοι εντοπίσθηκαν από το σύστημα κατηγοριοποίησης. Δίνονται δηλαδή αναλυτικά οι όροι που βρέθηκαν σωστά οι όροι που βρέθηκαν λανθασμένα, οι όροι που δεν βρέθηκαν ενώ θα έπρεπε να βρεθούν και τα κριτήρια στα οποία δεν εντοπίσθηκε κανένας όρος.

Περισσότερες πληροφορίες για την μορφή των αποτελεσμάτων καθώς και τα αποτελέσματα αυτά καθ’ αυτά βρίσκονται στην Ενότητα 4.

3.4.4 Υποσύστημα ανάγνωσης των Dataset αξιολόγησης (Pre Processing)

Κατά την διάρκεια ανάγνωσης των αρχείων των κλινικών δοκιμών από το σύστημα (βλ. Ενότητα 3.1.3), δίνεται η δυνατότητα να αναγνωσθούν μαζί με αυτές και οι σχολιασμένοι όροι του Dataset αξιολόγησης που έχει επιλεχθεί, σε περίπτωση που το σύστημα προορίζεται για αξιολόγηση. Το σύστημα, εφόσον δοθεί η εντολή να πραγματοποιήσει αξιολόγηση, εξασφαλίζει να αποθηκεύσει μαζί με κάθε κλινική δοκιμή τις πληροφορίες που αντλούνται από αυτό.

3.4.5 Υποσύστημα αξιολόγησης εντοπισμού όρων

3.4.5.1 Παραδοχές κατά την διαδικασία της αξιολόγησης

Ο τρόπος με τον οποίο το Chia Dataset είναι σχολιασμένο, δημιουργεί ορισμένες δυσκολίες όσον αφορά την χρήση του για την αξιολόγηση του συστήματος που έχει αναπτυχθεί. Παρακάτω δίνονται οι λόγοι για τους οποίους προκύπτουν οι εν λόγω δυσκολίες και διατυπώνονται κάποιες παραδοχές οι οποίες λήφθηκαν, προκειμένου να αξιολογηθεί το σύστημα ορθά και σύμφωνα με τις προδιαγραφές του επιθυμητού τρόπου αξιολόγησή.

Τα προβλήματα παρουσιάζονται παρακάτω:

Αφού τα κριτήρια του Chia έχουν σχολιασθεί χειροκίνητα, συμβαίνει, όπως και είναι λογικό, να υπάρχουν λάθη καθώς και ορισμένες παραδοχές στον τρόπο σχολιασμού του. Παρακάτω παραδείγματα λαθών και παραδοχών:

1) Πολλές φορές δεν γίνεται σχολιασμός των όρων “Male”, “Female”, “Man”, “Woman” καθώς επίσης και άλλων δημογραφικών.
2) Όροι μέσα σε παρενθέσεις και αγκύλες πολλές φορές δεν σχολιάζονται.
3) Όροι με πολλαπλές εμφανίσεις σε κάποιο κριτήριο, μπορεί να σχολιάζονται μόνο μία φορά μέσα στην πρόταση.
4) Ορισμένες προτάσεις ή φράσεις (που συμπεριλαμβάνουν πολλές λέξεις) των κριτηρίων μπορεί να είναι σχολιασμένες ολόκληρες, ως “non-representable” ή “non-queryable”

Επίσης πολλοί σχολιασμένοι όροι είναι πιθανό να μην είναι παρόντες στην οντολογία ή το λεξικό που χρησιμοποιείται για την αναγνώριση οντοτήτων στα κριτήρια, γεγονός που σημαίνει ότι κατά την αξιολόγηση προκύπτει αρνητικό αποτέλεσμα ανάκλησης, το οποίο όμως προέρχεται από την έλλειψη όρων στο λεξιλόγιο και όχι από την αδυναμία του προγράμματος να τους εντοπίσει.

Προκειμένου λοιπόν να πραγματοποιηθεί αξιολόγηση η οποία θα παράγει αποτελέσματα αντιπροσωπευτικά της αποτελεσματικότητας του συστήματος γίνονται οι εξής παραδοχές και ακολουθούνται τα παρακάτω:

1) Δεν λαμβάνονται υπ’ όψη λανθασμένα ευρήματα των όρων “Male”, “Female”, “Man”, “Woman” και λοιπών όρων που παράγουν λανθασμένα αρνητική ακρίβεια.
2) Δεν λαμβάνονται υπ’ όψη λανθασμένα ευρήματα μέσα σε παρενθέσεις και αγκύλες.

3) Δεν υπολογίζονται σαν αποτελέσματα σωστά ή λανθασμένα όσα βρίσκονται μέσα στους σχολιασμούς που περιγράφονται παραπάνω στο σημείο 4).

Επίσης, η αξιολόγηση πραγματοποιείται αφ' ενός με όλους τους σχολιασμένους όρους του συνόλου αξιολόγησης και αφ' εταίρου μόνο με τους όρους που βρίσκονται στο λεξικό ή την οντολογία που χρησιμοποιείται για τον εντοπισμό τους.

Το παραπάνω επιτυγχάνεται με τον εξής τρόπο:

Για όλους τους σχολιασμένους όρους του συνόλου αξιολόγησης που χρησιμοποιείται, εφαρμόζεται έλεγχος πριν την αξιολόγηση, αν ο κάθε ένας απ' αυτούς συμπεριλαμβάνεται σε κάποια έννοια του λεξικού ή της οντολογίας. Ο παραπάνω έλεγχος πραγματοποιείται απλά συγκρίνοντας την κανονικοποιημένη μορφή του ενός όρου με του άλλου.

3.4.5.2 Αντιστοίχιση μέσω δεικτών λέξεων

Για κάθε εύρημα του συστήματος σε κάποια κλινική δοκιμή, ελέγχεται αν η κλινική δοκιμή περιέχει σχολιασμένο όρο ο οποίος, όσον αφορά την θέση του μέσα στο κείμενο, περικλείει το εύρημα με την εξής λογική:

1) Για κάθε κλινική δοκιμή συλλέγονται οι εντοπισμένοι από το σύστημα κατηγοριοποίησης όροι, οι οποίοι είναι αποθηκευμένοι μαζί με πληροφορία σχετικά με την θέση τους στο κείμενο και σχετικά με τον αν επιδέχονται άρνηση
2) Συλλέγονται οι σχολιασμένοι όροι των κριτηρίων της κλινικής δοκιμής οι οποίοι επίσης περιέχουν την παραπάνω πληροφορία
3) Για κάθε σχολιασμένο όρο στο dataset αξιολόγησης γίνεται έλεγχος αν έχει εντοπισθεί εύρημα του συστήματος κατηγοριοποίησης τέτοιο ώστε να περικλείεται από τον σχολιασμένο όρο και:
    a. Αν ναι, τότε το εύρημα αυτό θεωρείται σωστό και ο σχολιασμένος όρος θεωρείται πως έχει εντοπισθεί σωστά
    b. Αν όχι, τότε ο σχολιασμένος όρος θεωρείται πως δεν έχει βρεθεί (επηρεάζεται δηλαδή αρνητικά η δυνατότητα ανάκλησης του συστήματος)
4) Παράλληλα αν υπάρχει εντοπισμένος όρος από το σύστημα κατηγοριοποίησης σε σημείο του κριτηρίου στο οποίο δεν υπάρχει κάποιος σχολιασμένος όρος στο dataset

αξιολόγησης, τότε θεωρείται ως λανθασμένο εύρημα. (επηρεάζεται δηλαδή αρνητικά η ακρίβεια του συστήματος)

5) Για τα σωστά ευρήματα του βήματος 4) γίνεται ύστερα έλεγχος, σχετικά με το αν η άρνηση που εντόπισε το σύστημα κατηγοριοποίησης συμπίπτει με την άρνηση που έχει αποδοθεί στον όρο στο dataset αξιολόγησης.

3.4.5.4 Έξοδος υποσυστήματος

Μέσω της παραπάνω διαδικασίας συλλέγονται για κάθε κλινική δοκιμή τα παρακάτω στοιχεία και τροφοδοτούνται στο υποσύστημα υπολογισμού του τελικού F1-score:

- Το πλήθος των στοιχείων που βρέθηκαν από το σύστημα κατηγοριοποίησης σε κάθε δοκιμή είτε αυτά είναι σωστά είτε όχι (True Positives + False Positives ή Retrieved Results).
- Το πλήθος των στοιχείων που βρέθηκαν από το σύστημα κατηγοριοποίησης σε κάθε δοκιμή και ταυτόχρονα υπάρχουν και στο αρχείο αξιολόγησης (True Positives ή (Relevant Results ∩ Retrieved Results))
- Το πλήθος των στοιχείων που υπάρχουν για κάθε δοκιμή στο αρχείο αξιολόγησης (True Positives + False Negatives ή Relevant Results)

3.4.6 Υποσύστημα υπολογισμού F1-score

Η τελική αξιολόγηση λαμβάνει χώρα αφού έχουν συλλεχθεί για κάθε κλινική δοκιμή τα παραπάνω στοιχεία και υπολογίζει το F1-Score το οποίο αποτελεί και την τελική κλίμακα αξιολόγησης του συστήματος. Το F1-Score υπολογίζεται σύμφωνα με τον παρακάτω τύπο:

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Όπου precision είναι η ακρίβεια του συστήματος. Αποτελεί δείκτη ο οποίος επιστρέφει το ποσοστό των αποτελεσμάτων που βρέθηκαν από το σύστημα και είναι ορθά σε σχέση με όλα τα αποτελέσματα που βρέθηκαν και υπολογίζεται όπως φαίνεται παρακάτω σύμφωνα με τα δεδομένα που έχουν συλλεχθεί προηγουμένως:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Και recall είναι η δυνατότητα ανάκλησης του συστήματος. Επιστρέφει το ποσοστό των αποτελεσμάτων που βρέθηκαν και είναι ορθά σε σχέση με όλα τα πραγματικά ορθά αποτελέσματα και υπολογίζεται όπως φαίνεται παρακάτω αντίστοιχα:

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

4 ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1. Αποτελέσματα Αξιολόγησης

Στην παρούσα ενότητα πραγματοποιείται η παρουσίαση και ο σχολιασμός των αποτελεσμάτων του συστήματος, αφ' ενός της διαδικασίας αναγνώρισης οντοτήτων κατά την κατηγοριοποίηση των κριτηρίων με βάση δύο διαφορετικών Dataset αξιολόγησης (Chia Dataset και ειδικά ανεπτυγμένου Dataset σχεδιασμένου στις προδιαγραφές του HarmonicSS) και αφ' εταίρου της διαδικασίας εμπλουτισμού της οντολογίας HarmonicSS με όρους του λεξικού MeSH του National Library of Medicine.

Όσον αφορά την αξιολόγηση της κατηγοριοποίησης των κριτηρίων αναλύεται η δυνατότητα του συστήματος να αναγνωρίζει ονομασμένες οντότητες στα κριτήρια που περιέχονται στο λεξικό στο οποίο χρησιμοποιείται, η δυνατότητα αναγνώρισης άρνησης στο κείμενο και η δυνατότητα αναγνώρισης οντότητες εύρους τιμών.

Κατά την αξιολόγηση του συστήματος εμπλουτισμού της οντολογίας HarmonicSS με όρους του λεξικού MeSH γίνεται παρουσίαση και ανάλυση της δυνατότητας κάλυψης της οντολογίας από όρους του MeSH, κατά πόσο δηλαδή είναι δυνατόν να αντιστοιχισθούν έννοιες και κατηγορίες εννοιών του ενός μοντέλου με το άλλο.

4.1.1 Αποτελέσματα αξιολόγησης συστήματος αναγνώρισης όρων

Κατά την διαδικασία της αξιολόγησης του προγράμματος κάνοντας χρήση του Dataset του Chia, προκειμένου να γίνει άντληση αποτελεσμάτων ενδεικτικών της αποτελεσματικότητας του προγράμματος να αναγνωρίσει όρους ενός λεξιλογίου στα κριτήρια καταλληλότητας, καθίσταται αναγκαίο να περιοριστούν οι σεσημασμένοι όροι του dataset σε αυτούς που υπάρχουν στην οντολογία. Για παράδειγμα αν η οντολογία ή το λεξικό που χρησιμοποιείται για την εύρεση όρων δεν περιλαμβάνει τον όρο «Cancer» τότε το σύστημα δεν είναι δυνατόν να την εντοπίσει σε κάποιο κριτήριο και κατ' επέκταση να αξιολογηθεί σύμφωνα με αυτόν.

Για τον λόγο αυτό δημιουργείται ένα υποσύστημα, του οποίου σκοπός είναι η αφαίρεση από το dataset των παραπάνω μη σχετικών όρων. Το σύστημα αυτό διατρέχει τους σεσημασμένους όρους και για κάθε έναν απ' αυτούς ελέγχει αν όλες οι επιμέρους λέξεις του σε κανονικοποιημένη μορφή συμπεριλαμβάνονται σε κάποιον όρο του λεξιλογίου που χρησιμοποιείται για την αναζήτηση του. Αν κανένας όρος του λεξιλογίου δεν πληροί το παραπάνω κριτήριο τότε ο σεσημασμένος αυτός όρος απορρίπτεται από πιθανό εύρημα.

Εξετάζοντας επίσης τα αποτελέσματα αξιολόγησης του συστήματος και στις δύο περιπτώσεις (πριν και μετά την αφαίρεση των όρων που δεν είναι παρόντες στην οντολογία/λεξικό) μπορούν να προκύψουν και συμπεράσματα σχετικά με την κάλυψη που το καθένα από αυτά προσφέρει σε κάποιο σύνολο κριτηρίων (πόσοι δηλαδή όροι που βρίσκονται μέσα στα κριτήρια υπάρχουν σε αυτό).

Σημειώνεται επίσης πως κατά την διαδικασία της αναγνώρισης όρων στα κριτήρια καταλληλότητας, όταν γίνεται χρήση της βιβλιοθήκης του MeSH χρησιμοποιούνται όροι από τις κατηγορίες Anatomy, Diseases, Organisms και Chemicals and Drugs καθώς επίσης και οι συμπληρωματικές εγγραφές(Supplemental Records), αφού οι υπόλοιπες κατηγορίες του λεξικού περιέχουν πληροφορία η οποία είτε δεν είναι σχετική με το αντικείμενο που θέλουμε να εξετάσουμε, είτε περιλαμβάνει όρους οι οποίοι δεν έχουν χαρακτηριστεί στο Chia dataset.

Λαμβάνοντας υπ' όψη τα παραπάνω, παρουσιάζονται παρακάτω τα αποτελέσματα της αξιολόγησης του συστήματος κατηγοριοποίησης κριτηρίων καταλληλότητας για κάθε συνδυασμό λεξικού/σχολιασμένων κριτηρίων.

4.1.1.1 Αξιολόγηση συστήματος με την βιβλιοθήκη του MeSH και το Chia Dataset:

A. Αποτελέσματα δίχως την απόρριψη όρων που δεν υπάρχουν στο λεξικό

Στα 2000 αρχεία κλινικών δοκιμών που εξετάστηκαν, σύμφωνα με το Dataset του Chia, θα έπρεπε να βρεθούν 36108 όροι από τους οποίους το σύστημα εντόπισε τους 141181 σωστά ενώ εντόπισε 606 όρους οι οποίοι δεν έχουν σχολιαστεί στο Chia Dataset. Παρακάτω παρουσιάζονται αναλυτικά τα ευρήματα κάθε φορά για τις επιτρεπόμενες αποστάσεις μεταξύ των token των όρων (για αποστάσεις μιας, δύο και τριών λέξεων μέσα στο κριτήριο):

Το F1-score και οι τιμές του precision και του recall φαίνονται παρακάτω:

	Distance=1	Distance=2	Distance=3
Recall	0.39273846	0.3897474	0.38224217
Precision	0.95901805	0.94570255	0.9192141
F1-Score	0.5572649	0.5520014	0.53995264
False Positives	606	808	1213
Matched correctly	14181	14073	13802
Correct Negations	13281	13180	12916
Temporals	7032/7582		

Πίνακας 7: Αποτελέσματα αξιολόγησης συστήματος με την βιβλιοθήκη του MeSH και το Chia Dataset

Β. Αποτελέσματα μετά *την* απόρριψη όρων που δεν υπάρχουν στο λεξικό

Στα ίδια 2000 αρχεία κλινικών δοκιμών που εξετάστηκαν και στην προηγούμενη περίπτωση και μετά την αφαίρεση των σχολιασμένων όρων που δεν είναι παρόντες στο λεξικό, σύμφωνα με το Dataset του Chia θα έπρεπε να βρεθούν 14640 όροι από τους οποίους το σύστημα εντόπισε τους 14164 σωστά ενώ εντόπισε 619 όρους οι οποίοι δεν έχουν σχολιαστεί στο Chia Dataset. Παρακάτω παρουσιάζονται αναλυτικά τα ευρήματα κάθε φορά για τις επιτρεπόμενες αποστάσεις μεταξύ των token των όρων (για αποστάσεις μιας, δύο και τριών λέξεων μέσα στο κριτήριο):

Το F1-Score και οι τιμές του precision και του recall φαίνονται παρακάτω:

	Distance=1	Distance=2	Distance=3
Recall	0.9674863	0.95963115	0.94050545
Precision	0.95812756	0.9444706	0.91744405
F1-Score	0.9627842	0.95199054	0.9288316
False Positives	619	826	1239
Matched correctly	14164	14049	13769
Correct Negations	13265	13158	12885
Temporals	7032/7582		

Πίνακας 8: Αποτελέσματα αξιολόγησης συστήματος με την βιβλιοθήκη του MeSH και το Chia Dataset μετά την απόρριψη όρων που δεν υπάρχουν στο λεξικό

4.1.1.2 Αξιολόγηση συστήματος με την εμπλουτισμένη οντολογία HarmonicSS και το Chia Dataset

Α. Αποτελέσματα δίχως την απόρριψη όρων που δεν υπάρχουν στην οντολογία

Στα ίδια 2000 αρχεία κλινικών δοκιμών που εξετάστηκαν, σύμφωνα με το dataset του Chia, θα έπρεπε να βρεθούν 36108 όροι από τους οποίους το σύστημα εντόπισε τους 4136 σωστά ενώ εντόπισε 107 όρους οι οποίοι δεν έχουν σχολιαστεί στο Chia dataset.

Το F1-Score και οι τιμές του precision και του recall φαίνονται παρακάτω:

	Distance=1	Distance=2	Distance=3
Recall	0.114545256	0.114379086	0.11321591
Precision	0.9704364	0.9591268	0.94128484
F1-Score	0.20490463	0.20438461	0.2021211
False Positives	126	152	255
Matched correctly	4136	4130	4088
Correct Negations	3868	3864	3823
Temporals	7032/7582		

Πίνακας 9: Αποτελέσματα αξιολόγησης συστήματος με την οντολογία HarmonicSS και το Chia Dataset

Β. Αποτελέσματα μετά την απόρριψη όρων που δεν υπάρχουν στην οντολογία

Στα ίδια 2000 αρχεία κλινικών δοκιμών και μετά την αφαίρεση των σχολιασμένων όρων που δεν είναι παρόντες στην οντολογία του HarmonicSS-RM, σύμφωνα με το dataset του Chia, θα έπρεπε να βρεθούν 4413 όροι από τους οποίους το σύστημα εντόπισε τους 3617 σωστά ενώ εντόπισε 108 όρους οι οποίοι δεν έχουν σχολιαστεί στο Chia dataset.

Το F1 score και οι τιμές του precision και του recall φαίνονται παρακάτω:

	Distance=1	Distance=2	Distance=3
Recall	0.9365511	0.9331521	0.9225017
Precision	0.9701878	0.95745176	0.9388838
F1-Score	0.9530727	0.9451458	0.9306207
False Positives	127	183	265
Matched correctly	4133	4118	4071
Correct Negations	3865	3852	3807
Temporals	7032/7582		

Πίνακας 10: Αποτελέσματα αξιολόγησης συστήματος με την οντολογία HarmonicSS και το Chia Dataset μετά την απόρριψη όρων που δεν υπάρχουν στην οντολογία

4.1.1.4 Αξιολόγηση συστήματος με την οντολογία HarmonicSS και το Chia δίχως εμπλουτισμό οντολογίας

Τέλος εξετάσθηκε και η αναγνώριση όρων με την οντολογία HarmonicSS δίχως την προσθήκη σε αυτήν όρων του MeSH.

Α. Αποτελέσματα δίχως την απόρριψη όρων που δεν υπάρχουν στην οντολογία

Στα ίδια 2000 αρχεία κλινικών δοκιμών που εξετάστηκαν, σύμφωνα με το dataset του Chia, θα έπρεπε να βρεθούν 36108 όροι από τους οποίους το σύστημα εντόπισε τους 3620 σωστά ενώ εντόπισε 107 όρους οι οποίοι δεν έχουν σχολιαστεί στο Chia dataset.

Το F1 score και οι τιμές του precision και του recall φαίνονται παρακάτω:

	Distance=1	Distance=2	Distance=3
Recall	0.10025479	0.1001994	0.09934086
Precision	0.9712906	0.9596817	0.94270694
F1-Score	0.18174972	0.18145344	0.17974092
False Positives	107	152	218
Matched correctly	3620	3618	3587
Correct Negations	3394	3392	3587
Temporals	7032/7582		

Πίνακας 11: Αποτελέσματα αξιολόγησης συστήματος με την οντολογία HarmonicSS και το Chia Dataset δίχως τον εμπλουτισμό οντολογίας από την βιβλιοθήκη του MeSH

Β. Αποτελέσματα μετά την απόρριψη όρων που δεν υπάρχουν στην οντολογία

Στα ίδια 2000 αρχεία κλινικών δοκιμών που εξετάστηκαν, σύμφωνα με το dataset του Chia, θα έπρεπε να βρεθούν 3997 όροι από τους οποίους το σύστημα εντόπισε τους 3617 σωστά ενώ εντόπισε 108 όρους οι οποίοι δεν έχουν σχολιαστεί στο Chia dataset.

Το F1 score και οι τιμές του precision και του recall φαίνονται παρακάτω:

	Distance=1	Distance=2	Distance=3
Recall	0.9049287	0.902677	0.8944208
Precision	0.9710067	0.9583001	0.940542
F1-Score	0.93680394	0.9296573	0.9169017
False Positives	108	157	226
Matched correctly	3617	3608	3575
Correct Negations	3391	3382	3350
Temporals	7032/7582		

Πίνακας 12: Αποτελέσματα αξιολόγησης συστήματος με την οντολογία HarmonicSS και το Chia Dataset δίχως τον εμπλουτισμό οντολογίας από την βιβλιοθήκη του MeSH και μετά την απόρριψη όρων που δεν υπάρχουν στην οντολογία

4.1.1.3 Αξιολόγηση συστήματος με την οντολογία HarmonicSS και τα Custom Annotated Criteria

Σε αυτήν την περίπτωση εξετάζονται 15 αρχεία κλινικών δοκιμών τα οποία αναφέρονται αποκλειστικά σε ασθενείς με σύνδρομο Sjögren και είναι σχολιασμένα χειροκίνητα σύμφωνα με τις δυνατότητες της οντολογίας HarmonicSS-RM. Θα έπρεπε να βρεθούν 91 όροι από τους οποίους το σύστημα εντόπισε τους 3617 σωστά ενώ εντόπισε 108 όρους οι οποίοι δεν έχουν σχολιαστεί.

	Distance=1	Distance=2	Distance=3
Recall	0.96703297	0.96703297	0.95604396
Precision	1.0	1.0	0.9886364
F1-Score	0.98324025	0.98324025	0.97206706
False Positives	0	0	1
Matched correctly	88	88	87
Correct Negations	81	81	80
Temporals	41/60		

Πίνακας 13: Αποτελέσματα αξιολόγησης συστήματος με την οντολογία HarmonicSS και το custom annotated dataset

Στην παρούσα περίπτωση το σύστημα κατάφερε να πετύχει 100% ακρίβεια στις περιπτώσεις όπου η απόσταση μεταξύ των επιμέρους token όρων είναι ίση με 1 ή 2, ενώ εντόπισε ένα λάθος αποτέλεσμα στην περίπτωση όπου η απόσταση είναι ίση με 3.

Όσον αφορά την τιμή του recall το σύστημα καταφέρνει να εντοπίσει σχεδόν όλους τους όρους που έχουν σχολιαστεί στα κριτήρια, με εξαίρεση κάποιων ορθογραφικών λαθών που συμπεριλήφθηκαν σκόπιμα στους σχολιασμούς.

Όπως είναι αναμενόμενο οι τιμές του F10Score που παράγονται στους ειδικά για την οντολογία σχολιασμένους όρους είναι αρκετά υψηλές.

4.1.2 Σχολιασμός αποτελεσμάτων αξιολόγησης

Γενικά σχόλια

Αρχικά, όσον αφορά τις τιμές του Recall πριν την αφαίρεση σχολιασμένων όρων από τα κριτήρια που δεν είναι παρόντες στην οντολογία HarmonicSS ή το MeSH αντίστοιχα, παρατηρείται πως μετά την αφαίρεση τους προκύπτει σημαντική αύξηση στην τιμή του recall του συστήματος, γεγονός το οποίο οφείλεται στην εκάστοτε αδυναμία του λεξικού ή της οντολογίας να καλύψει τους όρους που έχουν σχολιαστεί.

Παρατηρείται επίσης πως κάνοντας χρήση της βιβλιοθήκης του MeSH το πλήθος των σχολιασμένων όρων που καλύπτεται αυξάνεται αισθητά, γεγονός αναμενόμενο, αφού το MeSH περιλαμβάνει αρκετά μεγαλύτερο εύρος εννοιών από την οντολογία.

Ταυτόχρονα φαίνεται στα αποτελέσματα πως ο εμπλουτισμός της οντολογίας με όρους του MeSH επηρεάζει θετικά την δυνατότητα ανάκλησης όρων συγκριτικά με την χρήση της μη εμπλουτισμένης οντολογίας HarmonicSS.

Τέλος σημειώνεται πως οι τιμές του recall που προκύπτουν κατά την αξιολόγηση του συστήματος ύστερα από την αφαίρεση σχολιασμένων όρων από το dataset αξιολόγησης είναι πλέον αντιπροσωπευτικές των πραγματικών δυνατοτήτων του συστήματος αναγνώρισης όρων στα κριτήρια, αφού πλέον η δυνατότητα ανάκλησης όρων δεν επηρεάζεται από την έλλειψη εννοιών σε κάποιο λεξιλόγιο.

Σχόλια για την δυνατότητα ανάκλησης (Recall)

Θα ήταν αναμενόμενο ύστερα από την αφαίρεση όρων από το dataset αξιολόγησης, που δεν είναι παρόντες στην οντολογία, να προκύπτει recall πολύ κοντά στο 100%. Ωστόσο το recall που παρουσιάζεται στις παραπάνω περιπτώσεις κυμαίνεται κοντά στο 97%. Ο λόγος για τον οποίο αυτό συμβαίνει προκύπτει από το γεγονός πως πολλοί όροι βρίσκονται στα κριτήρια καταλληλότητας εκφρασμένοι, είτε με διάφορες συντομογραφίες, είτε διασπασμένοι σε επιμέρους λέξεις με μεγάλη απόσταση μεταξύ τους, μεγαλύτερη από τις δυνατότητες του συστήματος.

Παρατηρείται επίσης πως συγκρίνοντας τις τιμές του recall για τις διαφορετικές τιμές αποστάσεων εντοπισμένων token κάποιου όρου, αυτές παραμένουν σχεδόν ίδιες με μία ελάχιστη ωστόσο μείωση, γεγονός που οφείλεται στην πιθανή επιλογή λανθασμένων token για την ολοκλήρωση κάποιου όρου μέσα στο κριτήριο. Επομένως το σύστημα αποδίδει καλύτερα στην περίπτωση που οι όροι που εντοπίζονται είναι γειτονικοί.

Σχόλια για την ακρίβεια (Precision)

Όσον αφορά την ακρίβεια που επιτυγχάνει το σύστημα, παρατηρείται πως για κάθε περίπτωση οι τιμές κυμαίνονται περίπου στο 97% για επιλεγμένη απόσταση μεταξύ token όρων ίση με 1. Όταν η απόσταση αυτή αυξάνεται όπως είναι αναμενόμενο είναι δυνατόν να επιλεχθούν λανθασμένα token για να συνταχθεί κάποιος όρος, γεγονός που μειώνει την παραπάνω τιμή precision.

Επίσης, όπως αναφέρεται και στην ενότητα που περιγράφονται οι παραδοχές που έγιναν κατά την διαδικασία αξιολόγησης καθώς επίσης και τα προβλήματα που προκύπτουν από την φύση του συνόλου σχολιασμένων όρων του Chia για την αξιολόγηση του συστήματος, μερικοί παράγοντες που επηρέασαν αρνητικά την ακρίβεια του συστήματος είναι:

1) Το γεγονός πως κάποιοι από τους όρους δεν έχουν σχολιαστεί στο Chia dataset λόγω ανθρωπίνου λάθους ή παραδοχών των συντακτών του(βλ. Ενότητα 3.4.5.1).

2) Το γεγονός πως κάποιοι όροι/κριτήρια έχουν μεν σχολιαστεί αλλά έχουν χαρακτηριστεί με σχολιασμούς που δεν επιτρέπουν την σωστή ταυτοποίηση τους (βλ. Ενότητα 3.4.5.1).

4.2 Αποτελέσματα αντιστοίχισης HarmonicSS οντολογίας με τους όρους του MeSH

Η διαδικασία της αντιστοίχισης των όρων της βιβλιοθήκης του MeSH με αυτούς της οντολογίας HarmonicSS προσφέρει μια ολοκληρωμένη εικόνα όσον αφορά τις κατηγορίες οι οποίες συμπίπτουν και στα δυο μοντέλα. Ταυτόχρονα οι όροι οι οποίοι αντιστοιχίστηκαν εμπλουτίστηκαν με σημαντικό αριθμό συνωνύμων από την βιβλιοθήκη του MeSH οι οποίοι συμβάλλουν στην αποτελεσματικότερη κατηγοριοποίηση των κριτηρίων.

Εκτός από τα άμεσα ευρήματα, τις έννοιες του HarmonicSS-RM δηλαδή που βρέθηκαν στο MeSH και προστέθηκαν στην οντολογία, η διαδικασία αυτή της αντιστοίχισης επιτρέπει να αντληθούν συμπεράσματα και για πιο ευρείς όρους/κατηγορίες του μοντέλου για τους οποίους δεν βρέθηκε κάποιος όρος του MeSH. Έτσι λοιπόν μπορεί να εκφραστεί όλη η ιεραρχική δομή της οντολογίας του HarmonicSS συμφώνα και με άμεσες αντιστοιχίες με το MeSH αλλά και με ποσοστά κάλυψης κατηγοριών.

Πιο αναλυτικά, κατά την διαδικασία εμπλουτισμού της οντολογίας HarmonicSS-RM με όρους της βιβλιοθήκης του MeSH προστέθηκαν στην οντολογία οι παρακάτω κατηγορίες όρων.

- **Mesh Synonyms:** Όροι του Mesh που χαρακτηρίστηκαν ως συνώνυμοι κάποιου όρου του HarmonicSS
- **Mesh Narrower Terms:** Όροι του Mesh που χαρακτηρίστηκαν ως υποόροι κάποιου όρου του HarmonicSS
- **Mesh Wider Terms:** Όροι του Mesh που χαρακτηρίστηκαν ως όμοιοι κάποιου όρου του HarmonicSS (είναι και οι δύο δηλαδή υποκλάσεις της ίδιας κλάσης της οντολογίας)

Παρακάτω δίνεται αναλυτικά η έξοδος του υποσυστήματος που εκτελεί αυτήν την διαδικασία, η οποία αποτελεί ουσιαστικά την κάλυψη της οντολογίας για κάθε μία από τις παραπάνω κατηγορίες.

Για όλες τις κατηγορίες συνολικά από τα 628 στοιχεία της οντολογίας:

	matched	concepts	terms
Synonym	221	248	1576
Wider	215	262	1591

Narrower	201	6303	29734

Πίνακας 14: Αντιστοιχίσεις από όλες τις κατηγορίες συνολικά

Για την κατηγορία Drug από τα συνολικά 63 Drugs της οντολογίας:

	matched	concepts	terms
Synonym	35	51	200
Wider	13	17	197
Narrower	24	211	693

Πίνακας 15: Αντιστοιχίσεις κατηγορίας "Drug"

Για την κατηγορία Condition από τα συνολικά 217 Conditions της οντολογίας:

	matched	concepts	terms
Synonym	104	110	864
Wider	101	113	730
Narrower	82	544	4192

Πίνακας 16: Αντιστοιχίσεις κατηγορίας "Condition"

Για την κατηγορία Examination από τα συνολικά 265 Examination της οντολογίας:

	matched	concepts	terms
Synonym	69	74	477
Wider	96	127	651
Narrower	89	5532	24786

Πίνακας 17: Αντιστοιχίσεις κατηγορίας "Examination"

Για την κατηγορία Demographic από τα συνολικά 21 Demographic της οντολογίας:

	matched	concepts	terms
Synonym	8	8	13
Wider	5	5	13
Narrower	6	16	63

Πίνακας 18: Αντιστοιχίσεις κατηγορίας "Demographic"

Για την κατηγορία Lifestyle από τα συνολικά 4 Lifestyle της οντολογίας:

	matched	concepts	terms
Synonym	1	1	4
Wider	0	0	0
Narrower	0	0	0

Πίνακας 19: Αντιστοιχίσεις κατηγορίας "Lifestyle"

Για την κατηγορία Pregnancy από τα συνολικά 18 Pregnancy της οντολογίας:

	matched	concepts	terms
Synonym	3	3	12
Wider	0	0	0
Narrower	0	0	0

Πίνακας 20: Αντιστοιχίσεις κατηγορίας "Pregnancy"

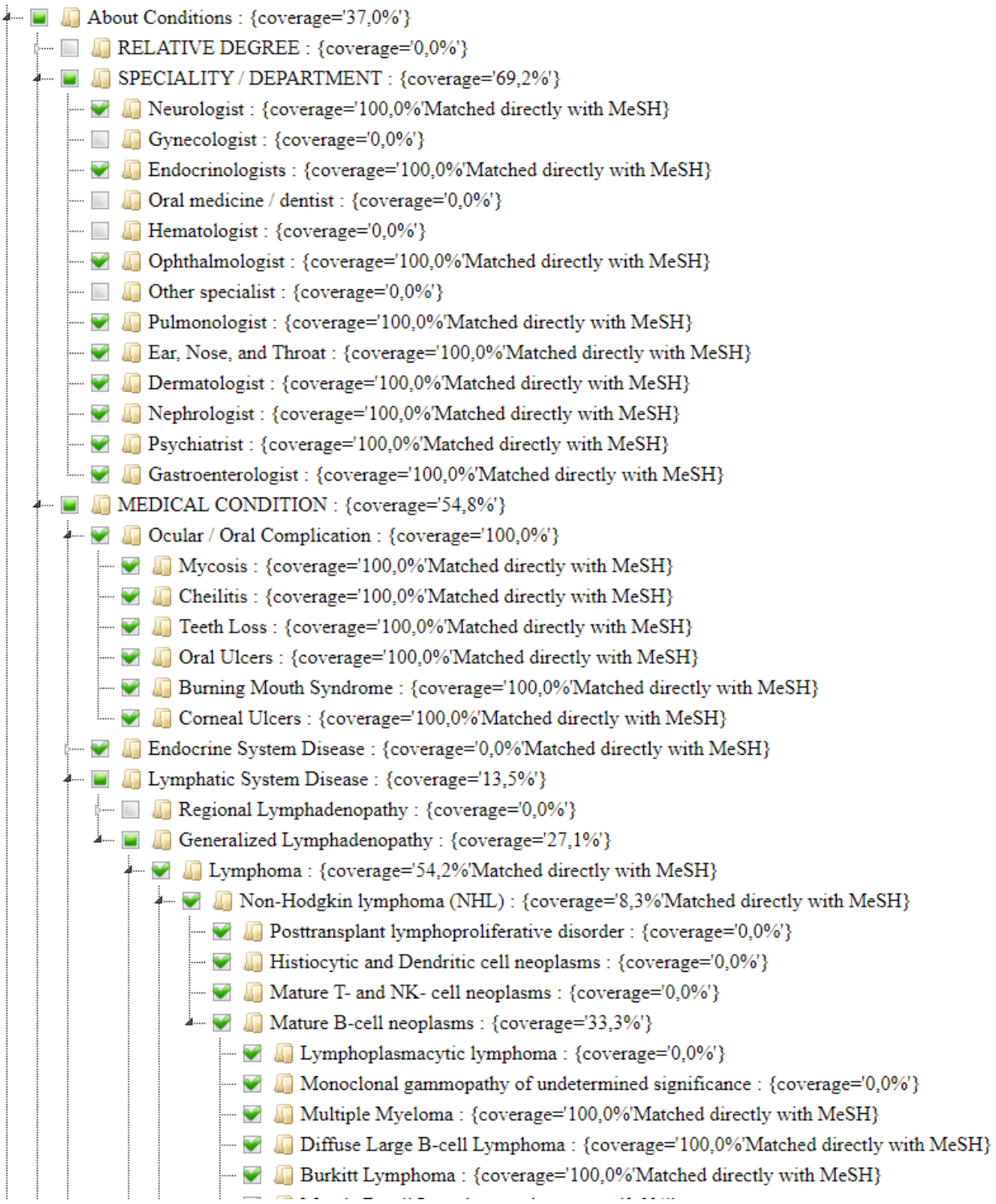
Για την κατηγορία Other Terms από τα συνολικά 40 Other Terms της οντολογίας:

	matched	concepts	terms
Synonym	1	1	6
Wider	0	0	0
Narrower	0	0	0

Πίνακας 21: Αντιστοιχίσεις κατηγορίας "Other Terms"

Παρακάτω φαίνεται μια οπτική αναπαράσταση των όρων του HarmonicSS-RM που αντιστοιχίσθηκαν με όρους του MeSH σε μορφή δέντρου που αναπαριστά την ιεραρχία της οντολογίας. Τα checkbox με το τσεκ δηλώνουν αν κάποιος όρος έχει καλυφθεί πλήρως (έχει βρεθεί συνώνυμο του) από κάποιον όρο του MeSH και τα γεμάτα checkbox δηλώνουν αν όρος έχει καλυφθεί μερικώς από το σύνολο των υποόρων του. Φαίνονται επίσης και τα ποσοστά κάλυψης. Να σημειωθεί επίσης πως τα παρακάτω δεδομένα αφορούν όρους που έχουν χαρακτηρισθεί ως συνώνυμοι του όρου που εξετάζεται.

Εικόνα 22: Οι βασικές κατηγορίες της οντολογίας HarmonicSS και η κάλυψή τους από όρους της βιβλιοθήκης του MeSH



Εικόνα 23:Στιγμιότυπο της ιεραρχίας της οντολογίας HarmonicSS και η κάλυψή των όρων της από όρους της βιβλιοθήκης του MeSH

4.3 Αποτελέσματα εντοπισμού οντοτήτων σε κριτήρια καταλληλότητας που αφορούν το σύνδρομο Σιόγκρεν και κατηγοριοποίηση αυτών

4.3.1 Αποτελέσματα αντιστοίχισης κριτηρίων

Ύστερα από την αξιολόγησή του, γίνεται χρήση του συστήματος για τον εντοπισμό οντοτήτων σε επιλεγμένα κριτήρια καταλληλότητας κλινικών δοκιμών σχετικών με το σύνδρομο Σιόγκρεν. Το σύνολο δεδομένων στο οποίο εφαρμόζεται το σύστημα αποτελείται από 200 διαφορετικές κλινικές δοκιμές, οι οποίες περιέχουν συνολικά 2091 κριτήρια καταλληλότητας. Παρακάτω παρουσιάζονται τα αποτελέσματα εντοπισμού όρων στα κριτήρια κάνοντας χρήση α) του MeSH, β) της οντολογίας HarmonicSS και γ) της οντολογίας HarmonicSS εμπλουτισμένης με σχετικούς όρους του MeSH.

4.3.1.1 MeSH

Κάνοντας χρήση της βιβλιοθήκης του MeSH για τον εντοπισμό όρων στα κριτήρια προκύπτουν τα παρακάτω αποτελέσματα.

- Κριτήρια που αντιστοιχίσθηκαν με τουλάχιστον έναν όρο: 1999 (95.6%)
- Κριτήρια που δεν αντιστοιχίσθηκαν με κανέναν όρο: 92 (4.4%)
- Πλήθος όρων που εντοπίσθηκαν σε όλα τα κριτήρια: 6685
- Πλήθος διαφορετικών μεταξύ τους όρων που εντοπίστηκαν σε όλα τα κριτήρια: 1233
- Κριτήρια με τουλάχιστον έναν περιορισμό τιμών ή χρονικό περιορισμό: 579 (27.7%)
- Πλήθος των περιορισμών τιμών και των χρονικών περιορισμών που εντοπίσθηκαν: 817

Όσον αφορά τις βασικές υποκατηγορίες της ιεραρχίας της βιβλιοθήκης του MeSH παρουσιάζεται παρακάτω η κατανομή των όρων που εντοπίστηκαν σε αυτές.

```
"Anatomy [A]": {
    "A01 Body Regions": 132,
    "A12 Fluids and Secretions": 85,
    "A11 Cells": 66,
    "A03 Digestive System": 53,
    "A07 Cardiovascular System": 20,
    "A09 Sense Organs": 19,
    "A02 Musculoskeletal System": 19,
    "A10 Tissues": 18,
    "A17 Integumentary System": 14,
    "A05 Urogenital System": 14,
    "A08 Nervous System": 13,
    "A04 Respiratory System": 9,
    "A15 Hemic and Immune Systems": 7,
    "A16 Embryonic Structures": 5,
    "A14 Stomatognathic System": 3,
    "A06 Endocrine System": 2,
    "A20 Bacterial Structures": 1
},
"Organisms [B]": {
    "B04 Viruses": 67,
    "B01 Eukaryota": 34
},
"Diseases [C]": {
    "C01 Infections": 279,
    "C05 Musculoskeletal Diseases": 250,
    "C23 Pathological Conditions, Signs and Symptoms": 219,
    "C04 Neoplasms": 121,
    "C20 Immune System Diseases": 95,
    "C11 Eye Diseases": 91,
    "Che null": 82,
    "C14 Cardiovascular Diseases": 65,
    "C17 Skin and Connective Tissue Diseases": 61,
    "C15 Hemic and Lymphatic Diseases": 43,
    "C08 Respiratory Tract Diseases": 41,
    "C10 Nervous System Diseases": 41,
    "C06 Digestive System Diseases": 33,
    "C07 Stomatognathic Diseases": 28,
    "C12 Male Urogenital Diseases": 24,
    "C25 Chemically-Induced Disorders": 22,
    "C18 Nutritional and Metabolic Diseases": 20,
    "C26 Wounds and Injuries": 17,
    "C16 Congenital, Hereditary, and Neonatal Diseases and Abnormalities": 7,
    "C19 Endocrine System Diseases": 7,
    "C13 Female Urogenital Diseases and Pregnancy Complications": 2
},
```

```
"Chemicals and Drugs [D]": {
    "D12 Amino Acids, Peptides, and Proteins": 212,
    "D02 Organic Chemicals": 189,
    "D27 Chemical Actions and Uses": 168,
    "D03 Heterocyclic Compounds": 143,
    "D26 Pharmaceutical Preparations": 124,
    "D04 Polycyclic Compounds": 90,
    "D20 Complex Mixtures": 70,
    "D06 Hormones, Hormone Substitutes, and Hormone Antagonists": 54,
    "D08 Enzymes and Coenzymes": 39,
    "D23 Biological Factors": 33,
    "D01 Inorganic Chemicals": 22,
    "D09 Carbohydrates": 13,
    "Dis null": 10,
    "D05 Macromolecular Substances": 2,
    "D10 Lipids": 1
},

"Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]": {
    "E01 Diagnosis": 426,
    "E02 Therapeutics": 354,
    "E05 Investigative Techniques": 222,
    "E07 Equipment and Supplies": 63,
    "E04 Surgical Procedures, Operative": 16,
    "E03 Anesthesia and Analgesia": 6,
    "E06 Dentistry": 5
},
"Psychiatry and Psychology [F]": {
    "F01 Behavior and Behavior Mechanisms": 207,
    "F02 Psychological Phenomena": 198,
    "F03 Mental Disorders": 19,
    "F04 Behavioral Disciplines and Activities": 1
},
"Phenomena and Processes [G]": {
    "G01 Physical Phenomena": 82,
    "G08 Reproductive and Urinary Physiological Phenomena": 64,
    "G07 Physiological Phenomena": 15,
    "G11 Musculoskeletal and Neural Physiological Phenomena": 12,
    "G09 Circulatory and Respiratory Physiological Phenomena": 3,
    "G10 Digestive System and Oral Physiological Phenomena": 1,
    "G02 Chemical Phenomena": 1,
    "G14 Ocular Physiological Phenomena": 1,
    "G12 Immune System Phenomena": 1,
    "G06 Microbiological Phenomena": 1
},
```

```
"Disciplines and Occupations [H]": {
    "H02 Health Occupations": 101,
    "H01 Natural Science Disciplines": 35
},
"Anthropology, Education, Sociology, and Social Phenomena [I]": {
    "I01 Social Sciences": 93,
    "I03 Human Activities": 11,
    "I02 Education": 5
},
"Technology, Industry, and Agriculture [J]": {
    "J03 Non-Medical Public and Private Facilities": 28,
    "J01 Technology, Industry, and Agriculture": 10
},
"Humanities [K]": {
    "K01 Humanities": 170
},
"Information Science [L]": {
    "L01 Information Science": 116
},
"Named Groups [M]": {
    "M01 Persons": 800
},
"Health Care [N]": {
    "N03 Health Care Economics and Organizations": 39,
    "N04 Health Services Administration": 34,
    "N01 Population Characteristics": 24,
    "N02 Health Care Facilities, Manpower, and Services": 24,
    "N06 Environment and Public Health": 14,
    "N05 Health Care Quality, Access, and Evaluation": 2
},
"Publication Characteristics [V]": {
    "V02 Publication Formats": 38,
    "V03 Study Characteristics": 27,
    "V01 Publication Components": 8
},
"Geographicals [Z]": {
    "Z01 Geographic Locations": 6
}
```


Εικόνα 24α), β), γ), δ), ε) Συχνότητα εμφάνισης εντοπισμένων όρων στα κριτήρια, κατανεμημένους στις βασικές υποκατηγορίες των βασικών κατηγοριών της βιβλιοθήκης του MeSH

Από τους όρους που εντοπίστηκαν στα κριτήρια παρουσιάζονται παρακάτω μερικοί από τους όρους που εμφανίστηκαν με μεγαλύτερη συχνότητα καθώς επίσης και η βασική κατηγορία της βιβλιοθήκης του MeSH στην οποία ανήκουν.

```
{
    "[term=Patients, category=Named Groups]": 365,
    "[term=History, category=Humanities]": 161,
    "[term=Sjogren's Syndrome, category=Diseases]": 154,
    "[term=Screening, category=Analytical, Diagnostic and Therapeutic Techniques and Equipment]": 153,
    "[term=Treatment, category=Analytical, Diagnostic and Therapeutic Techniques and Equipment]": 147,
    "[term=Disease, category=Diseases]": 136,
    "[term=Will, category=Psychiatry and Psychology]": 124,
    "[term=Infections, category=Diseases]": 112,
    "[term=Aged, category=Named Groups]": 104,
    "[term=Diagnosis, category=Analytical, Diagnostic and Therapeutic Techniques and Equipment]": 85,
    "[term=Drug, category=Chemicals and Drugs]": 78,
    "[term=Women, category=Named Groups]": 75,
    "[term=Informed Consent, category=Anthropology, Education, Sociology and Social Phenomena]": 65,
    "[term=Female, category=sex]": 65,
    "[term=Investigators, category=Named Groups]": 63,
    "[term=Diagnoses, category=Analytical, Diagnostic and Therapeutic Techniques and Equipment]": 61,
    "[term=Therapy, category=Analytical, Diagnostic and Therapeutic Techniques and Equipment]": 53,
    "[term=Eye, category=Anatomy]": 50,
    "[term=Randomization, category=Analytical, Diagnostic and Therapeutic Techniques and Equipment]": 50,
    "[term=Pregnancy, category=Phenomena and Processes]": 49,
    "[term=Time, category=Phenomena and Processes]": 45,
    "[term=Oldest Old, category=Named Groups]": 42,
    "[term=Antibodies, category=Chemicals and Drugs]": 42,
    "[term=Classification, category=Information Science]": 42,
    "[term=Surgery, category=Disciplines and Occupations]": 41,
    "[term=Consensus, category=Psychiatry and Psychology]": 40,
    "[term=Opinions, category=Psychiatry and Psychology]": 39,
    "[term=Male, category=sex]": 38,
    "[term=Writing, category=Information Science]": 34,
    "[term=Biopsy, category=Analytical, Diagnostic and Therapeutic Techniques and Equipment]": 33,
    "[term=Corticosteroids, category=Chemicals and Drugs]": 33,
    "[term=HIV, category=Organisms]": 32,
    "[term=Risk, category=Analytical, Diagnostic and Therapeutic Techniques and Equipment]": 30,
    "[term=Malignancy, category=Diseases]": 30,
    "[term=Rituximab, category=Chemicals and Drugs]": 29,
    "[term=Ability, category=Psychiatry and Psychology]": 28,
    "[term=Hydroxychloroquine, category=Chemicals and Drugs]": 28,
    "[term=Serum, category=Anatomy]": 28,
    "[term=Laboratories, category=Technology, Industry, Agriculture]": 27,
    "[term=Methods, category=Analytical, Diagnostic and Therapeutic Techniques and Equipment]": 27,
    "[term=Cyclosporins, category=Chemicals and Drugs]": 26,
    "[term=Prednisone, category=Chemicals and Drugs]": 26,
    "[term=Steroids, category=Chemicals and Drugs]": 25,
    "[term=Dry Eye, category=Diseases]": 25,
    "[term=Cyclophosphamide, category=Chemicals and Drugs]": 25,
    "[term=Hepatitis B, category=Diseases]": 24,
    "[term=Methotrexate, category=Chemicals and Drugs]": 24,
    "[term=Contraception, category=Analytical, Diagnostic and Therapeutic Techniques and Equipment]": 24,
```


Εικόνα 25: Οι συχνότερα εμφανιζόμενοι όροι στα κριτήρια και μαζί με την συχνότητα εμφάνισής τους και την κατηγορία της βιβλιοθήκης του MeSH στην οποία ανήκουν

Παρατηρείται πως κάνοντας χρήση της βιβλιοθήκης του MeSH για τον εντοπισμό οντοτήτων στα κριτήρια καταλληλότητας, επιτυγχάνεται να εντοπιστεί ένα αρκετά μεγάλο πλήθος όρων, το οποίο δεν είναι απαραίτητα σχετικό μόνο με το Σύνδρομο Σιόγκρεν. Συγκριτικά με τα αποτελέσματα τα οποία παράγονται στις παρακάτω ενότητες, κάνοντας χρήση της οντολογίας HarmonicSS, φαίνεται πως στην περίπτωση του MeSH προκύπτουν πολύ περισσότερα αποτελέσματα, όσον αφορά όρους που σχετίζονται με τα γενικότερα στοιχεία και την κατάσταση των ασθενών.

4.3.1.2 Οντολογία HarmonicSS εμπλουτισμένη με MeSH

Κάνοντας χρήση της εμπλουτισμένης οντολογίας HarmonicSS με όρους του MeSH προκύπτουν τα παρακάτω αποτελέσματα.

- Κριτήρια που αντιστοιχίσθηκαν με τουλάχιστον έναν όρο: 1073 (51.3%)
- Κριτήρια που δεν αντιστοιχίσθηκαν με κανέναν όρο: 1018 (48.7%)
- Πλήθος όρων που εντοπίσθηκαν σε όλα τα κριτήρια: 1473
- Πλήθος διαφορετικών μεταξύ τους όρων που εντοπίστηκαν σε όλα τα κριτήρια: 155
- Κριτήρια με τουλάχιστον έναν περιορισμό τιμών ή χρονικό περιορισμό:579 (27.7%)
- Πλήθος των περιορισμών τιμών και των χρονικών περιορισμών που εντοπίσθηκαν: 817

Όσον αφορά τις βασικές κατηγορίες της οντολογίας HarmonicSS παρουσιάζεται παρακάτω η κατανομή των όρων που εντοπίστηκαν σε αυτές.

Κατηγορία	Πλήθος όρων που βρέθηκαν
Condition	568
Examination	408
Drug	379
Demographic	109
Pregnancy	8
Other Terms	1

Πίνακας 22: Συχνότητα εμφάνισης εντοπισμένων όρων στα κριτήρια ανα κατηγορία

Από τους όρους που εντοπίστηκαν στα κριτήρια παρουσιάζονται παρακάτω μερικοί από τους όρους που εμφανίστηκαν με μεγαλύτερη συχνότητα ανά κατηγορία.

Για την κατηγορία Condition:

Όρος	Πλήθος εμφανίσεων
Sjogren's Syndrome	154
Cancer	66
Dry Mouth	36
Dry Eyes	31
Salivary Gland	22
Pulmonary Disease	22
Rheumatoid Arthritis	20

Πίνακας 23: Πλήθος εμφανίσεων των συχνότερων όρων της κατηγορίας Condition

Για την κατηγορία Examination:

Όρος	Πλήθος εμφανίσεων
Renal	35
Biological	35
Pulmonary	25
Hemoglobin	21
Aspartate aminotransferase	16
Anti-Hepatitis C Virus antibody	16
Creatinine	15
Neutrophils	15

Πίνακας 24: Πλήθος εμφανίσεων των συχνότερων όρων της κατηγορίας Examination

Για την κατηγορία Drug:

Όρος	Πλήθος εμφανίσεων
Inhaled Steroids	33
Hydroxychloroquine	29
Rituximab	29
Cyclosporine	28
Prednisone	26
Cyclophosphamide	25
Methotrexate	24

Πίνακας 25: Πλήθος εμφανίσεων των συχνότερων όρων της κατηγορίας Drug

Για την κατηγορία Demographic:

Όρος	Πλήθος εμφανίσεων
Female	66
Male	38

Πίνακας 26: Πλήθος εμφανίσεων των συχνότερων όρων της κατηγορίας Demographic

Για την κατηγορία Pregnancy:

Όρος	Πλήθος εμφανίσεων
Single	5
Twins	2
Abortion	1

Πίνακας 27: Πλήθος εμφανίσεων των συχνότερων όρων της κατηγορίας Pregnancy

Από τους 1473 όρους που εντοπίστηκαν, 1169 (79.4%) από αυτούς έχουν αντιστοιχιστεί με κάποιον συνώνυμο/σχετικό όρο της βιβλιοθήκης του MeSH. Οι πιο συχνά εμφανιζόμενοι από τους

εν λόγω συνωνύμους των όρων που εντοπίστηκαν κατανέμονται στις βασικές κατηγορίες του MeSH όπως φαίνεται παρακάτω.

Υποκατηγορία (Anatomy [A])	Πλήθος εμφανίσεων
A11 Cells	45
A03 Digestive System	39

Πίνακας 28: Συχνότητα εμφάνισης εντοπισμένων όρων στα κριτήρια με αντιστοιχισμένο όρο κάποιο συνώνυμο της βιβλιοθήκης του MeSH ανά κατηγορία, για τις βασικές υποκατηγορίες της κατηγορίας του MeSH Anatomy

Υποκατηγορία (Chemicals and Drugs [D])	Πλήθος εμφανίσεων
D12 Amino Acids, Peptides, and Proteins	148
D03 Heterocyclic Compounds	99
D02 Organic Chemicals	57

Πίνακας 29: Συχνότητα εμφάνισης εντοπισμένων όρων στα κριτήρια με αντιστοιχισμένο όρο κάποιο συνώνυμο της βιβλιοθήκης του MeSH ανά κατηγορία, για τις βασικές υποκατηγορίες της κατηγορίας του MeSH Chemicals And Drugs

Υποκατηγορία (Diseases [C])	Πλήθος εμφανίσεων
C05 Musculoskeletal Diseases	197
C04 Neoplasms	83
C17 Skin and Connective Tissue Diseases	39
C07 Stomatognathic Diseases	37
C11 Eye Diseases	31

Πίνακας 30: Συχνότητα εμφάνισης εντοπισμένων όρων στα κριτήρια με αντιστοιχισμένο όρο κάποιο συνώνυμο της βιβλιοθήκης του MeSH ανά κατηγορία, για τις βασικές υποκατηγορίες της κατηγορίας του MeSH Diseases

Παρατηρείται στην περίπτωση της χρήσης του HarmonicSS για τον εντοπισμό οντοτήτων στα κριτήρια καταλληλότητας, πως τα ευρήματα είναι πιο συγκεκριμένα. Περιορίζονται δηλαδή κυρίως σε ασθένειες, αποτελέσματα εξετάσεων και φάρμακα τα οποία σχετίζονται σε κάποιο βαθμό με το σύνδρομο Σιόγκρεν.

Ταυτόχρονα, από τα παραπάνω αποτελέσματα αντλείται πληροφορία σχετικά με τους συχνότερα χρησιμοποιημένους όρους στις κλινικές δοκιμές που αφορούν ασθενείς με Σύνδρομο Σιόγκρεν.

Φαίνεται λοιπόν, όσον αφορά τα δημογραφικά, πως αναζητούνται κυρίως γυναίκες, ενώ όσον αφορά τις ασθένειες και τα συμπτώματα που αναζητούνται, επικρατέστερα είναι το Σύνδρομο

Σιόγκρεν, καρκίνοι, ξηρότητα στα μάτια και στο στόμα, αναπνευστικές παθήσεις και η ρευματοειδής αρθρίτιδα.

Τέλος, στους πίνακες 23 και 24, δίνεται πληροφορία σχετική με τα φάρμακα και των τύπο εξετάσεων που εμφανίζονται συχνότερα στις αναζητήσεις ασθενών σε αυτού του τύπου τις κλινικές δοκιμές.

4.3.1.3 Οντολογία HarmonicSS χωρίς όρους του MeSH

Κάνοντας χρήση της εμπλουτισμένης οντολογίας HarmonicSS χωρίς όρους του MeSH προκύπτουν τα παρακάτω αποτελέσματα.

- Κριτήρια που αντιστοιχίσθηκαν με τουλάχιστον έναν όρο: 1053 (50.358685%)
- Κριτήρια που δεν αντιστοιχίσθηκαν με κανέναν όρο: 1038 (49.64132%)
- Πλήθος όρων που εντοπίσθηκαν σε όλα τα κριτήρια: 1375
- Πλήθος διαφορετικών μεταξύ τους όρων που εντοπίστηκαν σε όλα τα κριτήρια: 152
- Κριτήρια με τουλάχιστον έναν περιορισμό τιμών ή χρονικό περιορισμό:579 (27.7%)
- Πλήθος των περιορισμών τιμών και των χρονικών περιορισμών που εντοπίσθηκαν: 817

Όσον αφορά τις βασικές κατηγορίες της οντολογίας HarmonicSS παρουσιάζεται παρακάτω η κατανομή των όρων που εντοπίστηκαν σε αυτές.

Κατηγορία	Πλήθος όρων που βρέθηκαν
Condition	521
Examination	379
Drug	358
Demographic	108
Pregnancy	8
Other Terms	1

Πίνακας 31: Συχνότητα εμφάνισης εντοπισμένων όρων στα κριτήρια ανά κατηγορία

Από τους όρους που εντοπίστηκαν στα κριτήρια παρουσιάζονται παρακάτω μερικοί από τους όρους που εμφανίστηκαν με μεγαλύτερη συχνότητα ανά κατηγορία.

Για την κατηγορία Condition:

Όρος	Πλήθος εμφανίσεων
Sjogren's Syndrome	154
Cancer	36
Dry Mouth	35

Dry Eyes	31
Salivary Gland	22
Rheumatoid Arthritis	20

Πίνακας 32: Πλήθος εμφανίσεων των συχνότερων όρων της κατηγορίας Condition

Για την κατηγορία Examination:

Όρος	Πλήθος εμφανίσεων
Renal	35
Pulmonary	25
Hemoglobin	21
Aspartate aminotransferase	15
Anti-Hepatitis C Virus antibody	15
Creatinine	15
Neutrophils	15

Πίνακας 33: Πλήθος εμφανίσεων των συχνότερων όρων της κατηγορίας Examination

Για την κατηγορία Drug:

Όρος	Πλήθος εμφανίσεων
Inhaled Steroids	33
Hydroxychloroquine	29
Rituximab	29
Cyclosporine	28
Prednisone	26
Cyclophosphamide	25
Methotrexate	24

Πίνακας 34: Πλήθος εμφανίσεων των συχνότερων όρων της κατηγορίας Drug

Για την κατηγορία Demographic:

Όρος	Πλήθος εμφανίσεων
Female	66
Male	38

Πίνακας 35: Πλήθος εμφανίσεων των συχνότερων όρων της κατηγορίας Demographic

Για την κατηγορία Pregnancy:

Όρος	Πλήθος εμφανίσεων
Single	5
Twins	2
Abortion	1

Πίνακας 36: Πλήθος εμφανίσεων των συχνότερων όρων της κατηγορίας Pregnancy

Παρατηρείται στο σημείο αυτό, πως τα αποτελέσματα που παρήγαγε η μη εμπλουτισμένη με συνώνυμους όρους της βιβλιοθήκης του MeSH οντολογία, δεν ξεπερνούν πολύ σε πλήθος από τα αποτελέσματα της μη εμπλουτισμένης. Αφ' ενός το MeSH έδωσε την δυνατότητα στο σύστημα να εντοπίσει κάποιους όρους τους οποίους με την οντολογία αυτή καθ' αυτή δεν θα εντόπιζε αλλά αφ' εταίρου οι όροι αυτοί δεν ήταν πολλοί.

Το παραπάνω πιθανόν να οφείλεται στο γεγονός, πως στα πλαίσια του εντοπισμού οντοτήτων για την αναζήτηση στην βάση διαχείρισης ασθενών και την παραγωγή των παρόντων αποτελεσμάτων, προκειμένου να μην προκύψουν σφάλματα συμβατότητας μεταξύ των κατηγοριών των όρων που επιλέγονται για την αναζήτηση των ασθενών, χρησιμοποιήθηκαν μόνο οι αντιστοιχισμένοι όροι της βιβλιοθήκης του MeSH για τους οποίους υπήρχε κάποιος στενά συνδεδεμένος όρος στην οντολογία.

4.3.2 Αποτελέσματα κατηγοριοποίησης κριτηρίων

Στην παρούσα ενότητα παρουσιάζονται αποτελέσματα σχετικά με την κατηγοριοποίηση των παραπάνω 2091 κριτηρίων καταλληλότητας, στις βασικές κατηγορίες της οντολογίας HarmonicSS και της βιβλιοθήκης του MeSH.

4.3.2.1 HarmonicSS

Στον παρακάτω πίνακα παρουσιάζονται οι συχνότητες εμφάνισης των βασικών κατηγοριών της οντολογίας HarmonicSS για τα κριτήρια στα οποία εμφανίστηκε μόνο μία κατηγορία.

Κατηγορία	Αριθμός Κριτηρίων
Condition	326
Demographic	89
Examination	151
Drug	124
Pregnancy	2
OtherTerms	1
Σύνολο	693

Πίνακας 37: Συχνότητες εμφάνισης των βασικών κατηγοριών της οντολογίας HarmonicSS για τα κριτήρια στα οποία εμφανίστηκε μόνο μία κατηγορία

Παρακάτω παρουσιάζονται οι επικρατέστερες κατηγορίες όρων για τα κριτήρια, στα οποία εντοπίστηκαν ευρήματα από περισσότερες από 1 κατηγορίες. Από αυτά τα 94 περιείχαν όρους από 2 διαφορετικές κατηγορίες ενώ 8 περιείχαν παραπάνω από 2.

Κατηγορία	Αριθμός Κριτηρίων
Condition	24
Demographic	4
Examination	16
Drug	16
Pregnancy	0
OtherTerms	0
Περισσότερες από μια κατηγορίες	42
Σύνολο	102

Πίνακας 38: οι επικρατέστερες κατηγορίες όρων για τα κριτήρια, στα οποία εντοπίστηκαν ευρήματα από περισσότερες από 1 κατηγορίες

Παρακάτω παρουσιάζεται η συχνότητα των κριτηρίων που οι επικρατέστερες κατηγορίες τους ήταν κάποιο ζεύγος κατηγοριών.

Κατηγορία	Αριθμός Κριτηρίων
Condition - Examination	15
Drug - Condition	12
Drug - Examination	9
Condition - Demographic	3
Drug - Demographic	1
Pregnancy - Examination	1
Demographic - Examination	1
Σύνολο	42

Πίνακας 39: Συχνότητα των κριτηρίων που οι επικρατέστερες κατηγορίες τους ήταν κάποιο ζεύγος κατηγοριών

Από τα παραπάνω αποτελέσματα παρατηρείται, πως περίπου το 45% των κριτηρίων, στα οποία εντοπίζεται κάποιος όρος αφορούν κάποια ασθένεια, το 18% αφορούν κάποια ασθένεια και το 15% αφορούν κάποιο φάρμακο.

Προκύπτει επίσης το πόρισμα, πως συχνά συνδυάζονται στα κριτήρια όροι ασθενειών με όρους εξετάσεων, όροι φαρμάκων με όρους ασθενειών και όροι φαρμάκων με όρους εξετάσεων.

4.3.2.2 MeSH

Στον παρακάτω πίνακα παρουσιάζονται οι συχνότητες εμφάνισης των βασικών κατηγοριών της βιβλιοθήκης του MeSH για τα κριτήρια στα οποία εντοπίστηκε μόνο μία κατηγορία.

Κατηγορία	Αριθμός Κριτηρίων

Diseases	215
Named Groups	113
Chemicals and Drugs	102
Analytical, Diagnostic and Therapeutic Techniques and Equipment	85
Anatomy	37
Phenomena and Processes	31
Psychiatry and Psychology	30
Anthropology, Education, Sociology and Social Phenomena	23
Information Science	10
Supplementary Record	8
Health Care	8
Disciplines and Occupations	8
Publication Characteristics	6
Organisms	4
Technology, Industry, Agriculture	1
Humanities	1
Demographic	1
Σύνολο	683

Πίνακας 40: Συχνότητες εμφάνισης των βασικών κατηγοριών της βιβλιοθήκης του MeSH για τα κριτήρια στα οποία εντοπίστηκε μόνο μία κατηγορία

Παρακάτω παρουσιάζονται οι επικρατέστερες κατηγορίες όρων για τα κριτήρια, στα οποία εντοπίστηκαν ευρήματα από 2 ή περισσότερες διαφορετικές κατηγορίες.

Κατηγορία	Αριθμός Κριτηρίων
Περισσότερες από μια κατηγορίες	424
Diseases	330
Chemicals and Drugs	130
Analytical, Diagnostic and Therapeutic Techniques and Equipment	96
Anatomy	69
Named Groups	58
Phenomena and Processes	30
Supplementary Record	22
Anthropology, Education, Sociology and Social Phenomena	20
Psychiatry and Psychology	10
Disciplines and Occupations	8
Organisms	6
Information Science	2
Health Care	1
Technology, Industry, Agriculture	1

Σύνολο	1207

Πίνακας 41: Επικρατέστερες κατηγορίες όρων για τα κριτήρια, στα οποία εντοπίστηκαν ευρήματα από 2 ή περισσότερες διαφορετικές κατηγορίες της βιβλιοθήκης του MeSH

Παρακάτω, αναφέρονται τα πιο συχνά εμφανιζόμενα από τα κριτήρια των οποίων οι βασικές κατηγορίες τους αποτελούσαν κάποιο ζεύγος βασικών κατηγοριών.

Κατηγορία	Αριθμός Κριτηρίων
Analytical, Diagnostic and Therapeutic Techniques and Equipment - Diseases	47
Analytical, Diagnostic and Therapeutic Techniques and Equipment - Chemicals and Drugs	34
Named Groups - Diseases	28
Chemicals and Drugs - Diseases	23
Humanities - Diseases	22
Analytical, Diagnostic and Therapeutic Techniques and Equipment - Named Groups	20
Chemicals and Drugs - Named Groups	17

Πίνακας 42: Τα πιο συχνά εμφανιζόμενα από τα κριτήρια των οποίων οι βασικές κατηγορίες τους αποτελούσαν κάποιο ζεύγος βασικών κατηγοριών της βιβλιοθήκης του MeSH

Όμοια με τα αποτελέσματα που αντλήθηκαν από την χρήση της οντολογίας HarmonicSS για τον εντοπισμό οντοτήτων στα κριτήρια καταλληλότητας, κατά την κατηγοριοποίηση των αποτελεσμάτων, παρατηρείται πως οι επικρατέστερες κατηγορίες κριτηρίων αφορούν κυρίως ασθένειες, φάρμακα και εξετάσεις, με την εξαίρεση στην περίπτωση του MeSH, πως εντοπίζονται αισθητά περισσότερα κριτήρια που αφορούν στοιχεία ασθενών.

Αντίστοιχα, και στην περίπτωση χρήσης του MeSH συνδυάζονται στα κριτήρια όροι εξετάσεων με όρους ασθενειών, όροι εξετάσεων με όρους φαρμάκων, όροι φαρμάκων με όρους ασθενειών και στην παρούσα περίπτωση σε αντίθεση με τα προηγούμενα αποτελέσματα και συνδυασμός όρων που αφορούν στοιχεία ασθενών με όρους ασθενειών.

4.4 Αποτελέσματα εύρεσης πλήθους ασθενών

Στην παρούσα ενότητα αναλύονται τα αποτελέσματα της αναζήτησης πλήθους ασθενών στο σύστημα διαχείρισης δεδομένων ασθενών. Ο συνολικός αριθμός των κλινικών δοκιμών που χρησιμοποιήθηκαν για την αναζήτηση είναι 200, ενώ οι συνολικές κλήσεις που

πραγματοποιήθηκαν είναι 4600, αφού κάθε κλινική δοκιμή αναζητήθηκε στις 24 διαφορετικές βάσεις ασθενών του συστήματος.

Στον παρακάτω πίνακα φαίνονται, πόσες από αυτές τις κλήσεις περιείχαν κριτήρια τα οποία χρησιμοποιούνταν από την βάση διαχείρισης ασθενών για την καταχώρηση στοιχείων τους "Criterion - USED", πόσες όχι "Criterion – NOT USED" , και πόσες από αυτές επέστρεψαν μη μηδενικό πλήθος ασθενών.

Αναζήτηση κλινικών δοκιμών σε βάση	
Τουλάχιστον ένα USED κριτήριο	2201
Όλα τα κριτήρια NOT USED	2399
Τουλάχιστον ένα USED κριτήριο και μη μηδενικό πλήθος ασθενών	1082

Πίνακας 43: Πλήθος των κλήσεων στο σύστημα διαχείρισης ασθενών για χρησιμοποιημένα και μη inclusion κριτήρια

Όσον αφορά τα παραπάνω δεδομένα ομαδοποιημένα σε κλινικές δοκιμές προκύπτουν τα παρακάτω:

Από τις συνολικά 200 κλινικές δοκιμές, στις 113 από αυτές βρέθηκε τουλάχιστον μια βάση η οποία χρησιμοποιούσε τα κριτήρια αναζήτησης για την καταχώρηση δεδομένων ασθενών ενώ στις υπόλοιπες 87 δεν βρέθηκε κάποια βάση που να περιέχει τον συνδυασμό των κριτηρίων αναζήτησης της κλινικής δοκιμής.

Αναζήτηση κλινικών δοκιμών σε βάση	
Κλινικές δοκιμές με τουλάχιστον ένα "Used" κριτήριο	113
Κλινικές δοκιμές με κανένα "Used" κριτήριο	87
Κλινικές δοκιμές με τουλάχιστον ένα "Used" κριτήριο και κανένα αποτέλεσμα	24

Πίνακας 44: Πλήθος κλινικών δοκιμών με τουλάχιστον ένα "USED" inclusion κριτήριο, κανένα "USED" inclusion κριτήριο και με τουλάχιστον ένα "USED" inclusion κριτήριο και κανένα αποτέλεσμα

Όσον αφορά τον συνολικό αριθμό του πλήθους των ασθενών που επέστρεψαν τα ερωτήματα στο σύστημα διαχείρισης, λήφθηκαν συνολικά 160449 ασθενείς από όλα τα ερωτήματα που πραγματοποιήθηκαν στην βάση.

Ο μέσος αριθμός των ασθενών που βρέθηκαν σε κλινικές δοκιμές με κριτήρια τα οποία χρησιμοποιούνταν σε τουλάχιστον μια βάση διαχείρισης ήταν 1420 ανά κλινική δοκιμή.

Το ποσοστό ανάκλησης από τον συνολικό αριθμό ασθενών ανά κλινική δοκιμή που έστω και ένα κριτήριο της χρησιμοποιούνταν για αποθήκευση δεδομένων ασθενών, ήταν ίσος με το 45% των ασθενών που περιέχει η βάση.

Παρακάτω δίνονται τα ποσοστά εμφάνισης των κατηγοριών των inclusion κριτηρίων χρησιμοποιημένων “USED” και μη “NOT USED”, καθώς επίσης και των κριτηρίων των οποίων παρά το γεγονός ότι χρησιμοποιούνταν δεν επέστρεψαν κάποιον ασθενή.

Ο παρακάτω πίνακας δείχνει την κατανομή των κατηγοριών τον κριτηρίων που χρησιμοποιούνταν για την καταγραφή ασθενών στην βάση για τα inclusion criteria:

Κριτήριο	Ποσοστό των χρησιμοποιημένων κριτηρίων (%)
condition_symptom	26,5
demographics_gender	21,3
condition_diagnosis	18,2
examination_lab_test	8,3
patient	7,1
intervention_medication	5,5
demographics_ethnicity	5,3
examination_biopsy	5,2
demographics_pregnancy	2,6

Πίνακας 45: Πίνακας κατανομής των “USED” κριτηρίων ανά κατηγορία

Παρακάτω παρουσιάζεται ένας πίνακας με την κατανομή των κατηγοριών των κριτηρίων που δεν χρησιμοποιούνταν για την καταγραφή ασθενών στην βάση για τα inclusion criteria:

Κριτήριο	Ποσοστό των μη χρησιμοποιημένων κριτηρίων (%)
condition_diagnosis	61.5
examination_lab_test	12.5
demographics_pregnancy	10
demographics_ethnicity	4
intervention_medication	4
examination_biopsy	4
condition_symptom	4

Πίνακας 46: Πίνακας κατανομής των “NOT USED” κριτηρίων ανά κατηγορία

Παρακάτω παρουσιάζεται ένας πίνακας με την κατανομή των κατηγοριών των κριτηρίων που χρησιμοποιούνταν για την καταγραφή ασθενών στην βάση για τα inclusion criteria αλλά δεν επέστρεψαν κάποιον ασθενή:

Κριτήριο	Ποσοστό των χρησιμοποιημένων κριτηρίων (%)
patient (birth)	40.4
demographics_ethnicity	18.1
examination_biopsy	11.7
demographics_pregnancy	17.0
condition_diagnosis	9.6
demographics_gender	3.2

Πίνακας 47: Πίνακας κατανομής των "USED" κριτηρίων ανά κατηγορία που επέστρεψαν μηδενικό πλήθος ασθενών

Από τα παραπάνω αποτελέσματα προκύπτει πως το μεγαλύτερο ποσοστό των κριτηρίων των κλινικών δοκιμών που χρησιμοποιούνται από τις βάσεις του συστήματος διαχείρισης ασθενών αφορούν κατά κυρίως συμπτώματα ασθενειών (26,5%) δημογραφικά σχετικά με το φύλο των ασθενών (21,3%) και διαγνώσεις ασθενειών (18,2%).

Από τα κριτήρια που δεν χρησιμοποιούνταν, τα περισσότερα (61,5%) αφορούν κάποια διάγνωση ασθένειας ενώ ακολουθούν κριτήρια που αφορούν κάποιο εργαστηριακό τεστ με 12,5% και δημογραφικά σχετικά με εγκυμοσύνες με 10%.

Για τα κριτήρια τα οποία χρησιμοποιούνταν από την βάση και για τα οποία δεν βρέθηκε κάποιος ασθενής που να τα πληροί, παρατηρείται πως αποτελούνται από κατηγορίες κριτηρίων που αφορούν είτε δημογραφικά ασθενών όπως για παράδειγμα η ηλικία τους η εθνικότητα και κάποια κατάσταση εγκυμοσύνης. Ωστόσο ένα μικρό ποσοστό αυτών αντιστοιχούν και στις κατηγορίες της βιοψίας και των διαγνώσεων. Οι παραπάνω κατηγορίες αναμένονται να βρίσκονται στην παρούσα κατηγορία αφού αποτελούν το σύνολο των κριτηρίων που περιορίζονται συχνά από κάποιον περιορισμό τιμών.

5 ΣΥΝΟΨΗ

Στην παρούσα εργασία, εξετάσθηκαν αρχικά μέθοδοι και εργαλεία επεξεργασίας κειμένου. Ύστερα αναλύθηκαν ήδη υπάρχοντα μοντέλα, τεχνικές και προσεγγίσεις της αυτόματης αναγνώρισης οντοτήτων (Named Entity Recognition) και αξιολογήθηκαν αφενός σύμφωνα με την χρησιμότητα τους σε διαφορετικά είδη κειμένου και αφετέρου σύμφωνα με τον τύπο οντοτήτων που καλούνται να αναγνωρίσουν.

Έπειτα αναπτύχθηκε ένα σύστημα Αναγνώρισης Οντοτήτων σε κριτήρια καταλληλότητας κλινικών δοκιμών, βασισμένο σε μοντέλα λεξιλογικής προσέγγισης, του οποίου βασικός στόχος ήταν ο εντοπισμός εννοιών σχετικών με φάρμακα, διαγνωστικά τεστ, ασθένειες, δημογραφικά ασθενών καθώς επίσης και ο εντοπισμός αρνήσεων και χρονικών περιορισμών αξιοποιώντας τους για την κατηγοριοποίηση των. Εξετάστηκε παράλληλα η δυνατότητα της βιβλιοθήκης του MeSH να συμβάλει στην αποτελεσματικότητα του συστήματος, εμπλουτίζοντας με χρήσιμη πληροφορία την οντολογία HarmonicSS.

Ταυτόχρονα πραγματοποιήθηκε αξιολόγηση του συστήματος κάνοντας χρήση αφ’ ενός του Chia dataset, το οποίο περιέχει επισημασμένα κριτήρια καταλληλότητας γνωστών κλινικών δοκιμών και αφ’ εταίρου ενός συνόλου κριτηρίων καταλληλότητας υπαρχόντων κλινικών δοκιμών σχετικών με το Σύνδρομο Σιόγκρεν που επισημάνθηκαν στα πλαίσια της παρούσας εργασίας προκειμένου να αξιολογηθεί το και η ορθότητα του συστήματος αλλά και η συμβατότητα των λεξικών που χρησιμοποιήθηκαν για τον εντοπισμό οντοτήτων.

Στην συνέχεια πραγματοποιήθηκε χρήση του συστήματος για τον εντοπισμό όρων σε ορισμένες κλινικές δοκιμές που αφορούν αποκλειστικά ασθενείς με Σύνδρομο Σιόγκρεν. Από τα αποτελέσματα που λήφθηκαν κατά την κατηγοριοποίηση των κριτηρίων των κλινικών δοκιμών προέκυψε μια γενική εικόνα για το είδος των όρων που χρησιμοποιούνται ευρέως, όταν πρόκειται να αναζητηθούν ασθενείς με Σύνδρομο Σιόγκρεν. Παρατηρείται όπως είναι αναμενόμενο, πως τα κριτήρια καταλληλότητας συμπεριλαμβάνουν κυρίως γυναίκες και πως αναζητούνται κυρίως ασθένειες και συμπτώματα, με τους περισσότερο συνήθεις όρους σε αυτές τις κατηγορίες την ξηρότητα στα μάτια και στο στόμα, τον καρκίνο και την ρευματοειδή αρθρίτιδα. Αντλείται επίσης

το συμπέρασμα, πως συνδυάζονται στα κριτήρια συχνά όροι που αφορούν κάποια ασθένεια με όρους που αφορούν φάρμακα και εξετάσεις, ενώ τα δημογραφικά των ασθενών που αναζητούνται συνήθως εκφράζονται σε ξεχωριστά κριτήρια.

Τέλος, έγινε χρήση των εντοπισμένων οντοτήτων στα κριτήρια καταλληλότητας, για την εύρεση πλήθους ασθενών που τα πληρούν, κάνοντας αναζήτηση στο σύστημα επεξεργασίας δεδομένων ασθενών. Κατά την παραπάνω διαδικασία προέκυψε από το σύστημα επεξεργασίας κάποιο μη μηδενικό πλήθος ασθενών για περίπου τις μισές από τις κλινικές δοκιμές που εξετάσθηκαν. Τα κριτήρια με τις περισσότερες καταχωρημένες εγγραφές αφορούσαν κυρίως συμπτώματα, ενώ από τις κλινικές δοκιμές με κριτήρια που χρησιμοποιούνταν για την καταχώρηση ασθενών στις βάσεις και επέστρεψαν μηδενικό αριθμό ασθενών, ο μεγαλύτερος περιοριστικός παράγοντας φάνηκε να είναι οι ηλικίες των ασθενών, το οποία ήταν σε κάποιο βαθμό αναμενόμενο αφού οι βάσεις περιείχαν ασθενείς που είχαν διαγνωστεί με Σύνδρομο Σιόγκρεν αρκετά χρόνια πριν την διεξαγωγή των κλινικών δοκιμών.

6 ΠΑΡΑΡΤΗΜΑ

A. Το Εργαλείο Norm

Πιο συγκεκριμένα το εργαλείο Norm χρησιμοποιεί μια σειρά εντολών του εργαλείου LVG (οι κύριες που αναφέρθηκαν προηγουμένως και ο επιπρόσθετες που αναφέρονται παρακάτω). Η ολοκληρωμένη εντολή που εκτελείται με το πρόγραμμα Norm είναι το τρέξιμο του LVG με τις εξής επιλογές:

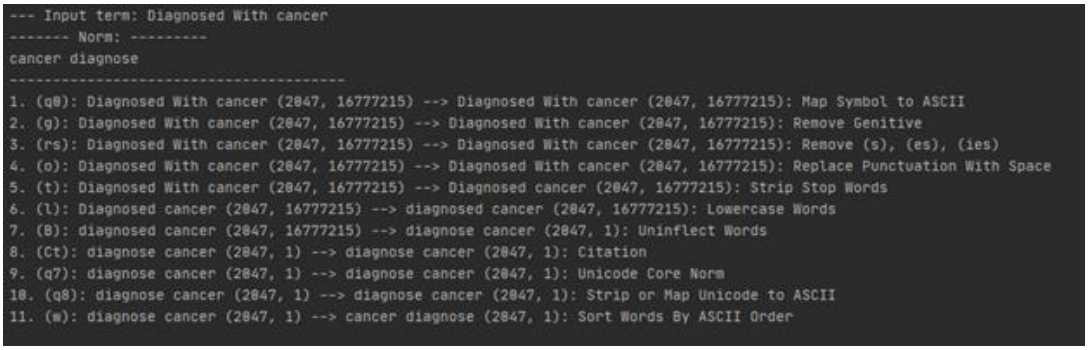
```
-f:q0:g:rs:o:t:l:B:Ct:q7:q8:w
```

Αναλυτικότερα οι εντολές αυτές, όπως ορίζονται από την ιστοσελίδα των Lexical Tools:

Εντολή	Περιγραφή
q0	map Unicode symbols and punctuation to ASCII
g	remove genitives
rs	remove parenthetic plural forms of (s), (es), (ies), (S), (ES), and (IES)
o	replace punctuation with spaces
t	remove stop words
B	lowercase
Ct	uninflect each word
q7	get citation form for each base form
q8	then strip or map non-ASCII Unicode characters
w	sort the words in alphabetical order

Πίνακας 48: Οι εντολές του εργαλείου Norm μαζί με μια σύντομη περιγραφή της λειτουργίας τους

Παρακάτω φαίνεται στο παράδειγμα “Diagnosed With Cancer”, πως εφαρμόζονται οι εντολές αυτές στην πράξη και παράγεται η κανονικοποιημένη έξοδος “cancer diagnose”.



Εικόνα 26: Παράδειγμα τρεξίματος του εργαλείου Norm στην φράση "Diagnosed With Cancer"

Είναι δυνατόν να δημιουργηθούν περισσότερες από μια κανονικοποιημένες μορφές κάποιας ακολουθίας συμβολοσειρών. Επίσης όμοιες κανονικοποιημένες λέξεις μέσω του εργαλείου δύναται να προέρχονται από διαφορετικές ίδιας ρίζας λέξεις.

Η πλήρης λίστα εντολών βρίσκεται στον ιστότοπο των Lexical Tools[14]:

Β. Το MeSH σε μορφή XML

Η Αμερικανική Εθνική Ιατρική Βιβλιοθήκη (NLM) προσφέρει, ανοιχτά για το κοινό στον ισότοπο της, την βιβλιοθήκη του MeSH σε μορφές XML, NTriples-RDF, ASCII και MARC21. Επιλέχθηκε λοιπόν για την συγκεκριμένη εφαρμογή η μορφή XML. Για την ανάκτηση των χρήσιμων πληροφοριών από την βιβλιοθήκη του MeSH σε XML και την μετατροπή τους σε μορφή εύκολα προσβάσιμη και αξιοποιήσιμη είναι αναγκαία η κατανόηση της ιεραρχίας και του συμβολισμού των κατηγοριών και των υποκατηγοριών τους όπως αυτές προσδιορίζονται με τα κατάλληλα XML-Tags.

Β.1 Τα αρχεία XML

Η βιβλιοθήκη του MeSH στην μορφή XML είναι χωρισμένη σε αρχεία, τα οποία διαχωρίζουν τους τύπους δεικτών/επικεφαλίδων της βιβλιοθήκης και είναι τα παρακάτω.

1. **desc2020.xml:** Στο αρχείο αυτό βρίσκονται οι κύριες επικεφαλίδες (Descriptors/Main Headings)

[14] https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2013/docs/designDoc/UDF/flow/index.html

2. **qual2020.xml:** Στο αρχείο αυτό βρίσκονται οι Υπό-Επικεφαλίδες (Subheadings, Qualifiers)
3. **supp2020.xml:** Στο αρχείο αυτό βρίσκονται οι Συμπληρωματικές Εγγραφές (Supplementary Records)

Β.1.1 Δομή του desc2020.xml

Όλες οι κύριες επικεφαλίδες βρίσκονται σε αυτό το αρχείο μέσα στο XML-Tag: **<DescriptorRecordSet LanguageCode = "eng">,** ενώ κάθε επικεφαλίδα περιβάλλεται από το XML-Tag: **<DescriptorRecord DescriptorClass = "X">.** Ο χαρακτηρισμός DescriptorClass αναφέρεται στην κλάση της κάθε επικεφαλίδας όπως ορίστηκε παραπάνω και μπορεί να πάρει τις τιμές 1, 2, 3 και 4.

Μέσα σε κάθε κύρια επικεφαλίδα βρίσκουμε κάποια βασικά XML-Tags:

- <DescriptorUI>
- <DescriptorName>
- <DateCreated>
- <DateRevised>
- <DateEstablished>
- <AllowableQualifiersList>
- <Annotation>
- <HistoryNote>
- <OnlineNote>
- <PublicMeSHNote>
- <PreviousIndexingList>
- <PharmacologicalActionList>
- <TreeNumberList>
- <ConceptList>

Τα παραπάνω XML-Tags περιέχουν τις πληροφορίες για κάθε επικεφαλίδα όπως εμφανίζονται και κατά την διαδικασία αναζήτησης μέσω του MeSH Browser. Παρακάτω περιγράφονται αναλυτικότερα τα XML-Tags <ConceptList> και <TreeNumberList> που συγκρατούν πληροφορία για σχετικούς με την επικεφαλίδα όρους και έννοιες καθώς και για τις κατηγορίες στις οποίες ανήκει.

Το <ConceptList> Tag:

Κάθε έννοια μέσα στο **<ConceptList>** Tag βρίσκεται κάτω από το XML-Tag **<Concept PreferredConceptYN="">**, όπου το PreferredConceptYN="" μπορεί να πάρει τιμές Y, N και δηλώνει αν μια έννοια είναι η προτεινόμενη έννοια μιας επικεφαλίδας.

Μέσα στο **<Concept>** tag βρίσκονται οι εξής βασικές πληροφορίες:

- <ConceptUI>: Το ID της έννοιας
- <ConceptName>: Το όνομα της έννοιας
- <TermList>: Η λίστα των όρων, σχετικών με την έννοια, η οποία με την σειρά της περιέχει του όρους, τον καθέναν μέσα σε ένα XML-Tag <Term> που περιέχει τις παρακάτω πληροφορίες στα αντίστοιχα tags:
    - <TermUI>: Το ID του όρου
    - <String>: Το όνομα του όρου
    - <DateCreated>: Την ημερομηνία δημιουργίας του

Το <TreeNumberList> Tag:

Μέσα στο <TreeNumberList> Tag βρίσκονται όλες οι κατηγορίες του ιεραρχικού δέντρου του MeSH μέσα στις οποίες ανήκει η επικεφαλίδα, η καθεμία μέσα στο αντίστοιχο Tag <TreeNumber>. Αξίζει να αναλυθεί η μορφή στην οποία βρίσκεται η πληροφορία της κατηγορίας της επικεφαλίδας. Για παράδειγμα ο <TreeNumber>D02.705.400.625.800</TreeNumber> φανερώνει αρχικά πως η επικεφαλίδα βρίσκεται στο υπό δέντρο D του ιεραρχικού δέντρου, στην υποκατηγορία D02. Μετά από κάθε τελεία αυξάνεται το βάθος της ιεραρχίας με κάθε αριθμό να αντιπροσωπεύει μια υποκατηγορία του προηγούμενου.

Β.1.2 Δομή του qual2020.xml:

Όλες οι υπό-επικεφαλίδες βρίσκονται σε αυτό το αρχείο μέσα στο XML-Tag: **<QualifierRecordSet LanguageCode = "eng">,** ενώ κάθε υπό-επικεφαλίδα περιβάλλεται από το XML-Tag: **<QualifierRecord>.**

Μέσα σε κάθε κύρια επικεφαλίδα βρίσκουμε τα βασικά XML-Tags τα οποία προαναφέρονται και περιγράφονται στην δομή του desc2020.xml. Η δομή δηλαδή είναι όμοια με την διαφορά ότι το όνομα και το ID της υπό-επικεφαλίδας βρίσκονται μέσα στα XML-Tags: **<QualifierName>** και **<QualifierUI>** αντίστοιχα.

Β.1.3 Δομή του supp2020.xml:

Όλες οι συμπληρωματικές εγγραφές συγκεντρώνονται σε αυτό το XML αρχείο μέσα στο XML-Tag: **<SupplementalRecordSet LanguageCode = "eng">,** ενώ κάθε συμπληρωματική εγγραφή περιβάλλεται από το XML-Tag: **<SupplementalRecord SCRClass = "">** όπου ο χαρακτηρισμός SCRClass αναφέρεται στην κλάση της κάθε συμπληρωματικής εγγραφής όπως ορίστηκε παραπάνω και μπορεί να πάρει τις τιμές 1, 2, 3 και 4.

Μέσα σε κάθε κύρια επικεφαλίδα βρίσκουμε τα βασικά XML-Tags τα οποία προαναφέρονται και περιγράφονται στην δομή του desc2020.xml, με τις εξής παρακάτω διαφορές:

- Το όνομα και το ID της συμπληρωματικής εγγραφής βρίσκονται μέσα στα XML-Tags: **<SupplementalRecordName>** και **<SupplementalRecordUI>** αντίστοιχα.
- Υπάρχει το XML-Tag <HeadingMappedTo> που με την σειρά του περιέχει το <DescriptorRefferedTo> όπου και βρίσκονται οι κύριες επικεφαλίδες με τις οποίες είναι συνδεδεμένη κάποια συμπληρωματική εγγραφή.
- Δεν υπάρχει Tag το οποίο να δηλώνει κάποια κατηγορία, αφού όπως έχει αναφερθεί η κλάση κάποιας συμπληρωματικής εγγραφής είναι και αυτή που κατηγοριοποιεί το περιεχόμενό της.

7 ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

Συντομογραφία	Ερμηνεία
AKA	Also Known As
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
HMM	Hidden Markov Model
JSON	JavaScript Object Notation
LVG	Lexical Variant Generator
MESH	Medical Subject Headings
NER	Named Entity Recognition
NLM	National Library of Medicine
NLP	Natural Language Processing
OWL	Web Ontology Language
POS	Part of Speech
RDF	Resource Description Framework
REGEX	Regular Expression
RM	Reference Model
SQL	Structured Query Language
SS	Sjogren Syndrome
TN-FN	True Negative – False Negative
TP-FP	True Positive – False Positive
UMLS	Unified Medical Language System
XML	Extensible Markup Language

8 ΑΝΑΦΟΡΕΣ ΚΑΙ ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman. Natural language processing: an introduction. 2011

[2] Efthymios Chondrogiannis, Vassiliki Andronikou, Theodora Varvarigou. Semantically-enabled context-aware abbreviations expansion in the clinical domain. 2017

[3] Wendy W.Chapman, WillBridewell, PaulHanburyaGregory, F.CooperabBruce, G.Buchananab. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. 2001

[4] Milan Ojstersek, Marko Ferme. Text analysis with sequence matching. 2011

[5] Jon Patrick, Yefeng Wang, Peter BuddAutomatic Mapping Clinical Notes to Medical Terminologies. 2006

[6] Jerry R. Hobbs, Douglas Appelt, FASTUS: A System for Extracting Information from Text. 1998

[7] Appelt et. al.D. Appelt, and et. al., SRI International FASTUS system MUC-6 test results and analysis, Proceedings of the MUC-6, NIST, Morgan-Kaufmann Publisher, Columbia. 1995

[8] Morgan, R., and et. al., University of durham: Description of the LOLITA system as used for MUC-6 In Proc of the MUC-6, NIST, Morgan-Kaufmann Publishers, Columbia, 1995.

[9] Sahar Ghannay, Benoit Favre, Word Embeddings Evaluation and Combination. 2016

[10] Guillaume Lample, Neural Architectures for Named Entity Recognition. 2016

[11] Radu Florian, Abe Ittycheriah, Hongyan Jing, Tong Zhang, Named Entity Recognition through Classifier Combination. 2003

[12] Yefeng Wang, Annotating and Recognising Named Entities in Clinical Notes. 2009

[13] Alan Ritter, Sam Clark, Mausam, Oren Etzioni, Named Entity Recognition in Tweets: An Experimental Study. 2011

[14] Andrei Mikheev, Marc Moens and Claire Grover, Named Entity Recognition without Gazetteers. 1999

[15] Fabrício Kury, Alex Butler, Chi Yuan, Chia, a large annotated corpus of clinical trial eligibility criteria. 2020