ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάπτυξη Καινοτόμων Σημασιολογικών Μηχανισμών για τη Βελτίωση του Σχεδιασμού μιας Κλινικής Έρευνας

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

Ευθυμίου Κ. Χονδρογιάννη

Αθήνα, Ιούνιος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΟΙΚΙΝΩΝΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑ ΣΥΣΤΗΜΑΤΩΝ
ΠΛΗΡΟΦΟΡΙΑΣ

Ανάπτυξη Καινοτόμων Σημασιολογικών Μηχανισμών για τη Βελτίωση του Σχεδιασμού μιας Κλινικής Έρευνας

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ευθυμίου Κ. Χονδρογιάννη

Συμβουλευτική Επιτροπή : Θεοδώρα Α. Βαρβαρίγου
Μαλαματή Δ. Λούτα
Νεκτάριος Γ. Κοζύρης

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 29η Ιουνίου 2017.

.....	.....	.....
Θεοδώρα Α. Βαρβαρίγου	Μαλαματή Δ. Λούτα	Νεκτάριος Γ. Κοζύρης
Καθηγήτρια Ε.Μ.Π.	Αν. Καθηγήτρια Π.Δ.Μ.	Καθηγητής Ε.Μ.Π.
.....	.....	.....
Εμμανουήλ Βαρβαρίγος	Δημήτριος Ασκούνης	Δημήτριος Κουτσούρης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.
	.....	
	Αναστάσιος Δουλάμης	
	Καθηγητής Ε.Μ.Π.	

Αθήνα, Ιούνιος 2017

.....
Ευθύμιος Κ. Χονδρογιάννης

Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Πρόλογος

Η διδακτορική διατριβή που παρουσιάζεται στις επόμενες σελίδες εκπονήθηκε από τον Απρίλιο του 2010 μέχρι τον Ιούνιο του 2017, στο εργαστήριο Τηλεπικοινωνιών του τομέα Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής, στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου υπό την επίβλεψη της κ. Θεοδώρας Βαρβαρίγου.

Κατά τη διάρκεια της εκπόνησης της διατριβής αυτής, είχα την ευκαιρία να ασχοληθώ με αρκετά ενδιαφέροντα επιστημονικά θέματα που αφορούν κυρίως τους τομείς της κλινικής έρευνας και περίθαλψης, του σημασιολογικού ιστού και της επεξεργασίας κειμένου, και να αποκτήσω πολύτιμη εμπειρία και γνώσεις.

Θα ήθελα να ευχαριστήσω από τα βάθη της καρδιάς μου την καθηγήτριά μου κ. Θεοδώρα Βαρβαρίγου για την υποστήριξη, και την καθοδήγηση που μου παρείχε από την αρχή ως το τέλος της προσπάθειάς μου, καθώς επίσης τους καθηγητές της τριμελούς συμβουλευτικής επιτροπής κ.κ. Λούτα Μαλαματή και Νεκτάριο Κοζύρη.

Επίσης, θα ήθελα να ευχαριστήσω όλους τους συναδέλφους με τους οποίους συνεργάστηκα κατά τη διάρκεια της εκπόνησης της διατριβής μου. Ιδιαίτερες ευχαριστίες ωστόσο θα ήθελα να απευθύνω στους συναδέλφους και φίλους Βασιλική Ανδρονίκου, Αναστάσιο Τάγαρη και Καραναστάση Ευστάθιο, με τους οποίους μοιραστήκαμε πάρα πολλές ώρες ερευνητικής εργασίας και αποδοτικής συνεργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για τη στήριξή τους όλα αυτά τα χρόνια.

Ευθύμιος Κ. Χονδρογιάννης

Ιούνιος 2017

Η σελίδα αυτή είναι σκόπιμα λευκή

Πίνακας Περιεχομένων

Ευρετήριο Σχημάτων

Ευρετήριο Πινάκων

Ευρετήριο Εξισώσεων

Περίληψη

Οι Κλινικές Δοκιμές αποτελούν το μέσο για την εισαγωγή ενός νέου φαρμάκου στην αγορά ή για την ανάδειξη νέων φαρμακευτικών ιδιοτήτων ενός υπάρχοντος φαρμάκου. Στόχος τους είναι να ελέγξουν την ασφάλεια και την αποτελεσματικότητα του φαρμάκου, για την αντιμετώπιση ή θεραπεία μιας συγκεκριμένης ιατρικής κατάστασης. Τα κριτήρια καταλληλότητας (ΚΚ) αποτελούν ένα σημαντικό τμήμα μιας κλινικής δοκιμής, επηρεάζοντας το κόστος της, τη διάρκειά της και γενικότερα τη συνολική της επιτυχία. Η τυπική τους αναπαράσταση σε μια γλώσσα κατανοητή από τον υπολογιστή θα μπορούσε να συμβάλλει σημαντικά στη βελτίωση του σχεδιασμού και της εκτέλεσης της δοκιμής, επιτρέποντας την περαιτέρω επεξεργασία των ΚΚ καθώς και τη διασύνδεσή τους με άλλες πηγές δεδομένων.

Στα πλαίσια της εργασίας αυτής, αρχικά, μελετήθηκαν πρότυπα σχετικά με την αναπαράσταση των ΚΚ καθώς επίσης και υπάρχουσα βιβλιογραφία. Επίσης, εξετάστηκε ένας σημαντικός αριθμός από κλινικές μελέτες (συμπεριλαμβανομένων των ΚΚ) με την ενεργή συμμετοχή ειδικών ερευνητών στον τομέα της κλινικής έρευνας, με απώτερο στόχο τον καθορισμό των παραμέτρων που χρησιμοποιούνται για την έκφραση των ΚΚ. Το αποτέλεσμα της παραπάνω διεργασίας ήταν μια αναπαράσταση των ΚΚ αποτελούμενη από ένα σχήμα συμβατό με τα πρότυπα του CDISC για τον καθορισμό των παραμέτρων ενός κριτηρίου καθώς επίσης και ένα μοντέλο με τα χαρακτηριστικά ενός ασθενούς για την τυπική τους έκφραση σε συνδυασμό με υπάρχοντα συστήματα κατηγοριοποίησης και κωδικοποίησης. Η αξιολόγηση των μοντέλων που αναπτύχθηκαν με βάση 200 τυχαίως επιλεγμένα ΚΚ έδειξε ότι μπορούν να καλύψουν επαρκώς το σκοπό για τον οποίο δημιουργήθηκαν.

Επίσης, αναπτύχθηκαν καινοτόμα συστήματα και μηχανισμοί που διευκολύνουν την πρόσβαση στα δεδομένα των ασθενών (ή γενικότερα των βάσεων δεδομένων), χρησιμοποιώντας τις τεχνολογίες του σημασιολογικού ιστού. Πιο συγκεκριμένα, αρχικά αναπτύχθηκε ένα σύστημα που επιτρέπει την έκφραση περίπλοκων ερωτημάτων, μέσω ενός ιδιαίτερα αλληλεπιδραστικού, φιλικού προς το χρήστη, γραφικού περιβάλλοντος, το οποίο βασίζεται αποκλειστικά και μόνο στην οντολογική περιγραφή του αντίστοιχου πεδίου γνώσης. Ακόμη, αναπτύχθηκαν εργαλεία και μηχανισμοί που επιτρέπουν την πρόσβαση στα δεδομένα μιας βάσης, χρησιμοποιώντας μοντέλα και κωδικοποιήσεις που πιθανώς να διαφέρουν σημαντικά από τα μοντέλα και κωδικοποιήσεις που χρησιμοποιούνται κατά την έκφραση των κριτηρίων.

Τα παραπάνω εργαλεία σε συνδυασμό με τα μοντέλα που αναπτύχθηκαν για την αναπαράσταση και έκφραση των ΚΚ μιας κλινικής δοκιμής μπορούν να συμβάλλουν σημαντικά τόσο στη δομημένη αναπαράσταση των κριτηρίων όσο και στον εντοπισμό των ασθενών που πληρούν τα κριτήρια, μειώνοντας το χρόνο στρατολόγησης των ασθενών και κατά συνέπεια το συνολικό κόστος της κλινικής δοκιμής.

Σημειώνουμε ότι, λόγω της ευρείας χρήσης των συντομεύσεων στην έκφραση των ΚΚ, αυτές μελετήθηκαν εκτενώς στην εργασία αυτή. Ειδικότερα, μελετήθηκαν οι συντομεύσεις που έχουν χρησιμοποιηθεί σε ένα περιορισμένο σύνολο κλινικών μελετών που τυχαία επιλέχθηκαν, καθώς και υπάρχοντες αλγόριθμοι και τεχνικές για τον εντοπισμό της πλήρους μορφής τους. Ακολούθως, αναπτύχθηκε ένα σύστημα για τον αυτόματο εντοπισμό της σημασίας των συντομεύσεων, χρησιμοποιώντας καινοτόμους αλγορίθμους και τεχνικές που λαμβάνουν υπόψη τη διακριτική δυνατότητα των λέξεων που συμμετέχουν στην πλήρη έκφρασή τους. Η αξιολόγηση του συστήματος έδειξε ότι αυτό μπορεί να εντοπίσει με μεγάλη ακρίβεια τη σημασία τους, ενώ η ανάλυση των

δεδομένων που συλλέχθηκαν έδειξε ότι οι πιθανές τους σημασίες είναι πολύ λιγότερες εν συγκρίσει με αυτές των βιοϊατρικών συντομεύσεων.

Η σελίδα αυτή είναι σκόπιμα λευκή

Abstract

Clinical Trials provides the means for bringing new drugs to the market or identifying new pharmacological uses of existing ones. Their purpose is to test the safety and efficacy of a drug towards a specific medical condition. Eligibility Criteria (EC) comprise an important part of a clinical study, being determinant of its cost, duration and overall success. Their formal, computer-processable description can significantly improve clinical trial design and conduction by enabling their intelligent processing, replicability and linkability with other data.

For EC representation purposes, related standards were investigated, along with published literature. Moreover, a considerable number of clinicaltrials.gov studies was analysed in collaboration with clinical experts for the determination and classification of parameters of clinical research importance. The outcome of this process was the EC Representation; a CDISC-compliant schema for organizing criteria along with a patient-centric model for their formal expression, properly linked with international classifications and codifications. Its evaluation against 200 randomly selected EC indicated that it can adequately serve its purpose.

Moreover, innovative systems developed for enabling users' access the data stored in one or more relational databases using semantic web technologies. More precisely, initially, a web application developed that enable users to express complicated queries through the user friendly, highly interactive graphical environment provided that is totally based on the ontological representation of a specific domain of interest. Also, specific tools and mechanism implemented for enabling users access the data stored in a relational database using their own terminology which may pose significant differences in comparison with the elements specified in the relational databases.

The aforementioned tools and mechanisms in comparison with the models developed for EC representation and expression can contribute in the formal expression of the EC of a clinical study as well as the application of EC specified for detecting the eligible subjects for recruitment. Consequently, they can reduce the patients' recruitment period and hence reduce the overall cost of a clinical study.

Since abbreviations are widely used in clinical studies, they have been extensively studied in this work. More precisely, we have studied the abbreviations used in a limited amount of clinical studies we have randomly selected, along with existing algorithms and techniques for abbreviation recognition purposes. Then, we have developed a system for automatically detecting the meaning of abbreviations used in the whole corpus of clinical studies, using innovative algorithms and techniques that take into account the discrimination power of words participating in their long form. The evaluation of the systems developed indicated that it can accurately detect the meaning of abbreviations specified in clinical studies, while the analysis of the collected data indicated that the abbreviations used in clinical studies are much less ambiguous than biomedical abbreviations.

Α. ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ ΤΩΝ ΚΡΙΤΗΡΙΩΝ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΜΙΑΣ ΚΛΙΝΙΚΗΣ ΔΟΚΙΜΗΣ

Στην ενότητα αυτή παρουσιάζουμε μια καινοτόμο, συμβατή με διεθνή πρότυπα, ευέλικτη και εύκολα επεκτάσιμη αναπαράσταση των κριτηρίων καταλληλότητας μιας κλινικής δοκιμής.

Η αναπαράσταση των κριτηρίων καταλληλότητας καθοδηγήθηκε από την ανάλυση ενός σημαντικού αριθμού υπαρχόντων κλινικών δοκιμών και των κριτηρίων καταλληλότητας που έχουν ορισθεί σε αυτές. Επιπρόσθετα, λάβαμε υπόψη τα πρότυπα που έχουν ήδη δημοσιευτεί από διεθνείς οργανισμούς, όπως ο CDISC και HL7, καθώς και υπάρχουσες ονοματολογίες, όπως η Διεθνής ταξινόμηση των ασθενειών (ICD) και η Ανατομική Θεραπευτική Χημική (ATC) κατηγοριοποίηση. Πιο συγκεκριμένα, στην εργασία αυτή προσπαθήσαμε να συνδυάσουμε αποτελεσματικά υπάρχοντα σχήματα, μοντέλα και κωδικοποιήσεις, στην προσπάθειά μας να σχεδιάσουμε μια διά-λειτουργική αναπαράσταση των κριτηρίων καταλληλότητας που θα μπορούσε να υποστηρίξει διαφορετικές ανάγκες.

Στην ανάλυση των κριτηρίων καθώς επίσης και στην αξιολόγηση των μοντέλων που αναπτύχτηκαν συμμετείχαν ειδικοί από τον τομέα της κλινικής έρευνας και περίθαλψης. Όπως φαίνεται στα αποτελέσματα της αξιολόγησης, τα μοντέλα που αναπτύχθηκαν μπορούν να καλύψουν ικανοποιητικά τα κριτήρια καταλληλότητας μιας κλινικής δοκιμής καθώς επίσης και να συνδυαστούν με υπάρχοντα εργαλεία που αναπτύχθηκαν τόσο για τον καθορισμό των κριτηρίων καταλληλότητας όσο και για τη χρησιμοποίησή τους για την εύρεση των κατάλληλων ασθενών.

Η σελίδα αυτή είναι σκόπιμα λευκή

1 *Εισαγωγή*

Οι κλινικές δοκιμές (clinical trials) αποτελούν ένα σημαντικό τμήμα μιας κλινικής έρευνας. Ο πρωταρχικός τους στόχος είναι να εξετάσουν την ασφάλεια και αποτελεσματικότητα μιας νέας θεραπείας για μία διαταραχή ή την ανακάλυψη νέων θεραπευτικών ιδιοτήτων για γνωστά φάρμακα (γνωστό επίσης ως αναπροσανατολισμός φαρμάκου). Το κόστος ανάπτυξης ενός φαρμάκου είναι υπερβολικά μεγάλο και σε αρκετές περιπτώσεις ξεπερνά το 1 δισεκατομμύριο δολάρια με τις κλινικές δοκιμές να αποτελούν παραπάνω από το 1/3 του κόστους αυτού [1][2]. Σημαντικές καθυστερήσεις στη διεξαγωγή των κλινικών δοκιμών εξαιτίας μη ικανοποιητικού σχεδιασμού αλλά και παρατεταμένης περιόδου στρατολόγησης ασθενών (patient recruitment) είναι άρρηκτα συνδεδεμένες με παραπάνω απ' το μισό των εξόδων αυτών, ενώ η χρονική διάρκεια για την ανάπτυξη ενός νέου φαρμάκου συχνά ξεπερνά τα 10 χρόνια [3]. Η αγορά, επίσης, απεικονίζει τις σημαντικές δυσκολίες που αντιμετωπίζουν οι κλινικές δοκιμές στο να μεταφράσουν επιτυχώς τη βασική έρευνα σε αποτελεσματικές θεραπείες για τους ασθενείς. Πιο συγκεκριμένα, μόνο 11% από τα υποψήφια φάρμακα σε "πρώτη-σε-άνθρωπο" κλινικές δοκιμές φτάνουν στην αγορά [4], αλλά ακόμη και αν καταφέρουν να μπουν στην αγορά, η αποτελεσματικότητά τους ή ακόμη και η ασφάλειά τους μπορεί να έχει υπερεκτιμηθεί, όπως δείχνουν νέες μελέτες που γίνονται.

Ανάμεσα στους επικρατέστερους παράγοντες που επηρεάζουν σημαντικά τις κλινικές δοκιμές υπό την έννοια του κόστους, διάρκειας, σχεδιασμού, εσωτερικής και εξωτερικής εγκυρότητας, είναι ο καθορισμός των κριτηρίων καταλληλότητας (eligibility criteria) μιας κλινικής δομικής [5]. Τα κριτήρια καταλληλότητας (ΚΚ – επίσης γνωστά ως κριτήρια εισαγωγής/εξαγωγής – inclusion/exclusion criteria) περιγράφουν τις συνθήκες τις οποίες οι υποψήφιοι ασθενείς θα πρέπει να πληρούν, για να συμμετέχουν σε μια κλινική δοκιμή. Στην προσπάθειά τους να αναδείξουν την αποτελεσματικότητα της υπό-εξερεύνηση θεραπείας, περιορίζοντας πιθανούς κινδύνους για τους ασθενείς, υπό την έννοια της δημιουργίας ενός ιδανικού συνόλου ασθενών, οι ερευνητές, πολύ συχνά, ορίζουν έναν πληθυσμό που απέχει σημαντικά από τον πραγματικό πληθυσμό στον οποίο στοχεύει το φάρμακο. Κατά συνέπεια, οι κλινικές δοκιμές συχνά συμπεριλαμβάνουν πολύ αυστηρά και επιλεκτικά κριτήρια, χωρίς πάντοτε να παρέχεται ικανοποιητική αιτιολόγηση, τα οποία απορρίπτουν ένα μεγάλο τμήμα του πληθυσμού των ασθενών και θέτουν εμπόδια για τη συμμετοχή τους σε κλινικές δοκιμές [6].

Ο κύριος λόγος πίσω από τα προαναφερθέντα προβλήματα πηγάζει από τις δυσκολίες που αντιμετωπίζουν οι ειδικοί στον τομέα της κλινικής έρευνας, στη φιλική προς τον υπολογιστή αναπαράσταση των ΚΚ κατά τη συγγραφή μιας κλινικής δοκιμής και ιδιαίτερα στην εφαρμογή τους στις βάσεις δεδομένων για την εύρεση των κατάλληλων ασθενών. Τα πρότυπα που έχουν δημοσιευτεί από διεθνείς οργανισμούς, όπως ο CDISC (Κοινοπραξία για την δημιουργία Κλινικών Προτύπων ανταλλαγής δεδομένων) [7] και HL7 (Διεθνές Επίπεδο Υγείας 7) [8], έχουν κάνει ένα σημαντικό βήμα για την τυπική αναπαράσταση των παραμέτρων μιας κλινικής δοκιμής καθώς και των δεδομένων των ασθενών, με σκοπό να διασφαλίσουν τη διαλειτουργικότητα ανάμεσα σε πληροφοριακά συστήματα που έχουν ως στόχο τόσο την κλινική έρευνα όσο και την περίθαλψη των ασθενών. Παρόλα αυτά, δεν υπάρχει κάποια ευρέως διαδεδομένη γλώσσα

για την τυπική αναπαράσταση και έκφραση των ΚΚ μιας κλινικής δοκιμής, ενώ οι τομείς της έρευνας και περίθαλψης παραμένουν ασύνδετες ή μερικώς συνδεδεμένες, περιορίζοντας τον αυτόματο εντοπισμό των ασθενών που πληρούν τα κριτήρια. Επομένως, η δημιουργία μιας διαλειτουργικής αναπαράστασης των ΚΚ, η οποία διευκολύνει την εφαρμογή τους στις βάσεις με τα δεδομένα των ασθενών, είναι ένα πολύ σημαντικό και επίκαιρο θέμα στον τομέα της κλινικής έρευνας.

Η ενότητα αυτή είναι οργανωμένη ως εξής. Στο κεφάλαιο 2 περιγράφονται συνοπτικά τα σχετικά πρότυπα και γλώσσες για την αναπαράσταση των κριτηρίων καθώς και οι περιορισμοί τους. Στο κεφάλαιο 3, περιγράφεται αναλυτικά η προσέγγιση που ακολουθήθηκε για την αναπαράσταση των ΚΚ. Στο κεφάλαιο 4 παρουσιάζονται τα μοντέλα που αναπτύξαμε καθώς και ο τρόπος που αυτά χρησιμοποιήθηκαν για την αναπαράσταση των ΚΚ. Στο κεφάλαιο 5 ακολουθεί η αξιολόγηση των μοντέλων που σχεδιάστηκαν. Στο κεφάλαιο 6 συζητείται η εφαρμογή των αποτελεσμάτων για την κάλυψη των αναγκών μιας κλινικής δοκιμής και τέλος στο κεφάλαιο 7 συνοψίζουμε τα βασικά σημεία της ενότητα αυτής.

Η σελίδα αυτή είναι σκόπιμα λευκή

2 *Σχετική Εργασία*

2.1 Διεθνή Πρότυπα και Κωδικοποιήσεις

Για την κατάλληλη αναπαράσταση και ανταλλαγή δεδομένων σχετικών με κλινικές δοκιμές (συμπεριλαμβανομένων των ΚΚ), με απώτερο σκοπό να πάρουμε το μέγιστο δυνατό όφελος από τα καταγραφέντα δεδομένα, οι οργανισμοί CDISC (κυρίως) και HL7 έχουν δημοσιεύσει αρκετά πρότυπα (είτε ανεξάρτητα ο ένας από τον άλλον, είτε μέσω μιας συλλογικής προσπάθειας) συμπεριλαμβανομένων *προτύπων ανταλλαγής πληροφορίας* (exchange protocols) όπως CDISC ODM (Μοντέλο Δεδομένων Λειτουργίας) [9] και HL7, *μοντέλα αναφοράς* (reference models) όπως CDISC SDTM (Μοντέλο/Πίνακας Αναφοράς Δεδομένων Μελέτης) [10] και HL7 RIM (Μοντέλο Αναφοράς Πληροφοριών) [11], *ονοματολογίες και κωδικοποιήσεις* όπως CDISC Controlled Terminology [12] και HL7 v3 Value Sets [13], και *λειτουργικά προφίλ* (functional profile) όπως EHR CR FP (Λειτουργικό Προφίλ Ηλεκτρονικών Αρχείων Υγείας Κλινικών Ερευνών) [14].

Επίσης, ένας σημαντικός αριθμός από ονοματολογίες (δηλαδή, ανοιχτά ή κλειστά σύνολα από όρους ή κωδικούς, κατηγοριοποιήσεις, θησαυρούς και οντολογίες) που σχετίζονται άμεσα με βιοϊατρικούς όρους έχουν δημοσιευτεί από διαφορετικούς διεθνείς οργανισμούς που έχουν ως στόχο την ανάπτυξη προτύπων, όπως η Διεθνής ταξινόμηση των ασθενειών (ICD) [15] που δημοσιεύτηκε από τον Παγκόσμιος Οργανισμός Υγείας (WHO) [16], το Ιατρικό λεξικό για ρυθμιστικές δραστηριότητες (MedDRA) [17] που

αναπτύχθηκε από το Διεθνές Συμβούλιο για την εναρμόνιση των τεχνικών απαιτήσεων για τα φαρμακευτικά προϊόντα για ανθρώπινη χρήση (ICH) [18] και η Βάση με Λογικές Παρατηρήσεις Αναγνωριστικά, Ονόματα, Κώδικες (LOINC) [19] που δημιουργήθηκε και συντηρείται από το Regenstrief ινστιτούτο [20].

Τα πρότυπα που έχουν δημοσιευτεί από το CDISC χρησιμοποιούνται ευρέως στην κλινική έρευνα, καθώς μπορούν να υποστηρίξουν όλες τις φάσεις μιας κλινικής δοκιμής [21], ενώ μπορούν παράλληλα να συνδυαστούν με υπάρχουσες ονοματολογίες και κωδικοποιήσεις για την καταγραφή των κυρίων παραμέτρων μιας κλινικής δοκιμής (π.χ., την ασθένεια ή σύνδρομο που μελετάται) καθώς επίσης και με οποιαδήποτε πληροφορία είναι απαραίτητη κατά το σχεδιασμό, διάρκεια ή ολοκλήρωσή της (π.χ., τις εργαστηριακές μετρήσεις που έλαβαν χώρα). Η εναρμόνιση των στοιχείων που έχουν οριστεί στα CDISC και HL7 Μοντέλα Αναφοράς από την Ομάδα Ολοκληρωμένου Τομέα Βιοϊατρικής Έρευνας (BRIDG) [22] καθώς επίσης και η αντιστοίχιση των όρων που έχουν ορισθεί σε σχετικές ονοματολογίες από το Ενοποιημένο Σύστημα Ιατρικής Γλώσσας (UMLS) [23] διευκολύνει την επικοινωνία ανάμεσα σε πληροφοριακά συστήματα από τον τομέα της έρευνας και περίθαλψης (π.χ., ανταλλαγή δεδομένων για εργαστηριακές μετρήσεις ή τεστ που έγιναν). Σχετικά με την αναπαράσταση των ΚΚ, επειδή αυτά πρακτικά ορίζουν το σύνολο ή το εύρος τιμών στο οποίο τα δεδομένα των ασθενών θα πρέπει να ανήκουν, εισάγουν ένα επιπλέον επίπεδο πολυπλοκότητας στη φιλική προς τον υπολογιστή αναπαράστασή τους, η οποία δυσκολεύει την έκφρασή τους αλλά και την περαιτέρω επεξεργασία των ΚΚ.

Για τον ορισμό των ΚΚ, το CDISC παρέχει μια υψηλού επιπέδου περιγραφή των κριτηρίων εισαγωγής / εξαγωγής ασθενών καθώς και τις παραμέτρους που θα πρέπει να καταγράφονται για το καθένα από αυτά, αφήνοντας τις λεπτομέρειες σχετικά με την τυπική τους αναπαράσταση σε μια γλώσσα κατανοητή από τον υπολογιστή στον εκάστοτε

οργανισμό, όπως φαίνεται στον ορισμό των CDISC προτύπων PRM (Μοντέλο Αναπαράστασης Πρωτοκόλλου) [24], SDM (Μοντέλο Σχεδιασμού Μελέτης) [25] και ODM (Μοντέλο Δεδομένων Λειτουργίας). Οι υπάρχουσες ονοματολογίες μπορούν να καλύψουν ικανοποιητικά τους όρους που χρησιμοποιούνται στα ΚΚ (π.χ., ασθένειες ή τις ευρύτερες κατηγορίες στις οποίες ανήκουν). Απ' την άλλη, οι παράμετροι που έχουν οριστεί στα CDISC και HL7 Μοντέλα Αναφοράς μπορούν να καλύψουν μερικώς τις παραμέτρους που απαιτούνται για την τυπική αναπαράσταση των ΚΚ σε μια γλώσσα κατανοητή από τον υπολογιστή (π.χ., παράμετροι που σχετίζονται με μία διάγνωση), καθώς έχουν σχεδιαστεί για ένα σκοπό διαφορετικό από την αναπαράσταση των ΚΚ.

2.2 Γλώσσες Αναπαράστασης και Έκφρασης Κριτηρίων Καταλληλότητας

Σχετικά με τον ορισμό των ΚΚ, αρκετές γλώσσες αναπαράστασης και έκφρασης έχουν δημοσιευτεί στη βιβλιογραφία, οι οποίες περιγράφονται αναλυτικά από τον Weng κ.ά. [26]. Η Arden Syntax για την περιγραφή Ενοτήτων Ιατρικής Λογικής (MLM) [27] είναι μία γλώσσα για την αναπαράσταση και αναφορά ιατρικής γνώσης που αποτελείται από ανεξάρτητες ενότητες. Κάθε ενότητα περιέχει ικανοποιητική πληροφορία για τη λήψη μιας κλινικής απόφασης και τυπικά αποτελείται από μία ή περισσότερες θυρίδες (slots) που είναι οργανωμένες σε τέσσερα τμήματα (διατήρηση, βιβλιοθήκη, γνώση και πηγές). Η GELLO [28] είναι μία αντικειμενοστραφής γλώσσα, η οποία στοχεύει επίσης στην αναφορά – κοινοποίηση – μοίρασμα ιατρικής γνώσης, με την προϋπόθεση ότι αναφέρεται σε ένα κοινό μοντέλο δεδομένων. Η GELLO βασίζεται στην Αντικειμενοστραφή Γλώσσα ορισμού Περιορισμών (OCL) [29], δηλαδή μία δηλωτική γλώσσα για την έκφραση κανόνων που μπορούν να εφαρμοστούν στη Ενοποιημένη Γλώσσα Μοντελοποίησης (UML) [30] και επομένως παρέχει την απαιτούμενη σημειογραφία για την αναφορά σε κλάσεις καθώς και στις παραμέτρους τους, όπως αυτά ορίζονται στο μοντέλο που

χρησιμοποιείται (π.χ., κάποιο υποσύνολο του HL7 RIM) και επιτρέπει τη δημιουργία περίπλοκων εκφράσεων, χρησιμοποιώντας τους τελεστές που υποστηρίζονται. Η Arden Syntax καθώς και οι GELLO εκφράσεις (υιοθετήθηκαν από το HL7) έχουν σχεδιαστεί κυρίως για την υποστήριξη λήψης αποφάσεων παρά για την αναπαράσταση των ΚΚ μιας κλινικής δοκιμής.

Η Γραμματική και Οντολογία Κανόνων Επιλεξιμότητας (ERGO) [31] είναι μια γλώσσα για την έκφραση ΚΚ. Χρησιμοποιεί τρία διαφορετικά συστατικά (δηλαδή δηλώσεις, εκφράσεις και βοηθητικά στοιχεία) τα οποία επιτρέπουν στους χρήστες να ορίσουν ιδιαίτερα περίπλοκα κριτήρια. Επίσης, διευκολύνει τη διαδικασία μετατροπής των ΚΚ που είναι εκφρασμένα χρησιμοποιώντας τη φυσική γλώσσα σε μια μορφή που είναι κατανοητή από τον υπολογιστή χρησιμοποιώντας ERGO αναφορές [32]. Όμως, η προσέγγιση που ακολουθείται δε διευκολύνει την εφαρμογή των ΚΚ για την εύρεση των ασθενών, καθώς τα δεδομένα τους συχνά βρίσκονται σε μια σχεσιακή βάση δεδομένων και ακολουθούν πιστά κάποιο μοντέλο, ενώ μεγάλο μέρος της πληροφορίας είναι ορισμένο με βάση διεθνή συστήματα κατηγοριοποίησης ή γενικότερα κωδικοποίησης. Κατά συνέπεια, ο καθορισμός της συσχέτισης ανάμεσα στα ERGO στοιχεία και τα στοιχεία της βάσης είναι ιδιαίτερα περίπλοκος (αν όχι αδύνατος), καθώς, για παράδειγμα, μπορεί να μην υπάρχει άμεση αντιστοιχία μεταξύ των ονοματικών φράσεων που χρησιμοποιούνται στα ΚΚ και της ονοματολογίας που χρησιμοποιείται στη βάση με τα δεδομένα των ασθενών.

Στο έργο Συμφωνία σχετικά με τις Τυποποιημένες απαιτήσεις Πρωτοκόλλου για την Επιλεξιμότητα (ASPIRE) υπάρχει σαφής διαχωρισμός ανάμεσα στα κριτήρια που εξαρτώνται από την ασθένεια (π.χ., στάδιο του καρκίνου) και στα υπόλοιπα κριτήρια (π.χ., δημογραφικά, εγκυμοσύνη, κτλ.) που ορίζονται ως ζευγάρια (δηλαδή χαρακτηριστικό - τιμή) [33]. Στο έργο αυτό, η αναπαράσταση των ΚΚ εστιάζει σε

συγκεκριμένου τύπου ασθένειες (διαβήτη και καρκίνο του στήθους) καθώς επίσης είναι στενά συνδεδεμένη με την ονοματολογία που αναπτύχθηκε στο έργο αυτό. Κατά συνέπεια, δεν μπορεί να χρησιμοποιηθεί ως μία γλώσσα αναπαράστασης γενικού σκοπού παρά τις επεκτάσεις που έγιναν.

Στην ανάλυση για την Εξαγωγή και Αναπαράσταση των Κριτηρίων Καταλληλότητας (EliXR) [34], η αναπαράσταση των ΚΚ που ακολουθείται βασίζεται σε κάποια πρότυπα (templates ή frames). Ο σχεδιασμός των προτύπων βασίστηκε σε ένα Σημασιολογικό Δίκτυο που κατασκευάστηκε με βάση την ανάλυση 1000 τυχαίων επιλεγμένων ΚΚ από την τοποθεσία clinicaltrials.gov [35]. Η όλη προσέγγιση που ακολουθείται είναι ιδιαίτερα ενδιαφέρουσα, καθώς η αναπαράσταση των ΚΚ βασίστηκε σε υπάρχοντα κριτήρια μιας κλινικής δοκιμής. Όμως, εξετάστηκε ένας πολύ μικρός αριθμός ΚΚ, τα οποία μπορεί να μην καλύπτουν πλήρως το ευρύ φάσμα των ΚΚ. Επιπρόσθετα, αρκετές δυσκολίες ενδέχεται να προκύψουν κατά τη χρησιμοποίηση των ΚΚ που έχουν ορισθεί με τον παραπάνω τρόπο – εσωτερικά αναπαρίστανται χρησιμοποιώντας ένα συνδυασμό από SQL [36] και Arden Syntax Curly Brackets σημειογραφίας – για την εύρεση των ασθενών, ειδικά όταν τα μοντέλα που χρησιμοποιούνται για την καταγραφή των δεδομένων ενός ασθενούς παρουσιάζουν σημαντικές διαφορές με το Σημασιολογικό Δίκτυο που αναπτύχθηκε.

Ο Luo κ.ά. [37] χρησιμοποίησαν τύπους του UMLS Σημασιολογικού Δικτύου για την κατηγοριοποίηση 3400 τυχαίως επιλεγμένων ΚΚ, χρησιμοποιώντας ιεραρχική κατηγοριοποίηση (hierarchical clustering) [38]. Ακολούθως, συγχώνευσαν τα 41 συγκροτήματα που αυτομάτως ανιχνεύτηκαν σε 27 διαφορετικές σημασιολογικές κατηγορίες (περαιτέρω κατηγοριοποιήθηκαν σε 6 κατηγορίες) μέσω μιας χειροκίνητης – βασισμένης στο χρήστη – διαδικασίας. Σε αυτή την εργασία, επίσης, επιχειρήθηκε διασύνδεση των σημασιολογικών κλάσεων που ανιχνευτήκαν με το BRIDG μοντέλο. Πιο

συγκεκριμένα, 16 από τις 17 σχετικές με ΚΚ παραμέτρους που περιέχονται στο BRIDG μοντέλο (ανιχνεύτηκαν με τη βοήθεια ενός συνόλου από ειδικούς στον τομέα αυτό) μπόρεσαν να συσχετιστούν με τις κλάσεις που ανιχνεύτηκαν. Όμως, υπήρξαν 8 σημασιολογικές κλάσεις (π.χ., κατάσταση οργάνου ή ιστού) οι οποίες δεν μπόρεσαν να διασυνδεθούν με τις παραμέτρους του BRIDG μοντέλου. Παρόλα αυτά, αν και δεν περιγράφεται στην εργασία τους, το γεγονός ότι ένα σημαντικό μέρος από τις κλάσεις που ανιχνεύτηκαν διασυνδέθηκαν με το CDISC μοντέλο διευκολύνει την επεξεργασία των ΚΚ που είναι εκφρασμένα με βάση το παραπάνω μοντέλο από συστήματα που υποστηρίζουν τα CDISC πρότυπα.

Ο Milian κ.ά. [39] επίσης πρότειναν μια αναπαράσταση των ΚΚ βασισμένη σε πρότυπα. Τα πρότυπα καθώς και οι επιμέρους παράμετροί τους ορίσθηκαν χειροκίνητα με βάση την ανάλυση 300 τυχαίων επιλεγμένων ΚΚ σχετικά με τον καρκίνο του στήθους. Οι συγγραφείς, σε συνεργασία με ειδικούς στον τομέα της ιατρικής, όρισαν επίσης τη σχέση ανάμεσα στα πρότυπα που ορίστηκαν, επιτρέποντας στους χρήστες να εντοπίσουν όχι μόνο τα ΚΚ που ανήκαν στο ίδιο πρότυπο αλλά και αυτά με στενότερη ή ευρύτερη σημασία. Η προσέγγιση που ακολουθήθηκε εστίασε στη μελέτη των ΚΚ που ανήκουν σε ένα συγκεκριμένο τύπο κλινικών δοκιμών, ενώ μόνο ένας ειδικός στον τομέα της ιατρικής συμμετείχε στην παραπάνω διαδικασία. Επιπρόσθετα, τα ΚΚ που ορίζονται με βάση τα πρότυπα που αναπτύχθηκαν δεν μπορούν να χρησιμοποιηθούν άμεσα για την εύρεση των ασθενών. Για αυτό το σκοπό απαιτείται συσχέτιση όλων των προτύπων που ορίσθηκαν με ερωτήματα προς τη βάση δεδομένων των ασθενών, μια διαδικασία αρκετά δύσκολη και σε μεγάλο βαθμό κατευθυνόμενη από το χρήστη εξαιτίας των σημαντικών δομικών και σημασιολογικών διαφορών ανάμεσα στο μοντέλο που χρησιμοποιείται εσωτερικά για την αναπαράσταση των προτύπων και στο μοντέλο που χρησιμοποιείται για την καταγραφή των δεδομένων των ασθενών.

Θα πρέπει να σημειωθεί ότι ο όρος «πρότυπο», παρά το γεγονός ότι έχει χρησιμοποιηθεί, για να χαρακτηρίσει την αναπαράσταση των ΚΚ σε διάφορα έργα, δε χρησιμοποιείται πάντα και παντού με την ίδια σημασία. Στην περίπτωση του ERGO [31] αναφέρεται στις «αφηρημένες» προτάσεις, βάση των οποίων γίνεται η έκφραση των ΚΚ χρησιμοποιώντας για παράδειγμα την κατάλληλη ονοματική φράση. Στην EliXR ανάλυση αναφέρεται στο συνδυασμό των στοιχείων του Σημασιολογικού Δικτύου μαζί με συγκριτικούς τελεστές και αριθμητικές τιμές ή ονοματικές φράσεις για την αναπαράσταση των ΚΚ. Για παράδειγμα, η φράση «Ασθενείς που διαγνώστηκαν με ΑΣΘΕΝΕΙΑ τους προηγούμενους ΑΡΙΘΜΌΣ μήνα/μήνες» είναι ένα πρότυπο, όπου οι συμβολοακολουθίες γραμμένες με κεφαλαία γράμματα θα αντικατασταθούν από τα πραγματικά δεδομένα ενός ΚΚ. Εναλλακτικά, σε περίπτωση που μία ασθένεια ή αριθμός εμφανίζεται αρκετά συχνά, μπορεί να είναι μέρος του προτύπου αυτού καθ' αυτού, όπως στην εργασία που δημοσιεύτηκε από τον Bhattacharya και Cantor [40]. Ο όρος «πρότυπο» χρησιμοποιείται με παρόμοια σημασία από τους Milian κ.ά. [39]. Όμως, στη δουλειά τους τα στοιχεία που χρησιμοποιούνται για την έκφραση του συνόλου ή εύρους τιμών, στα οποία τα αντίστοιχα δεδομένα θα πρέπει να ανήκουν, δεν αποτελούν μέρος του προτύπου. Για παράδειγμα, το προαναφερθέν κριτήριο θα γραφόταν ως εξής: «Ασθενείς που διαγνώστηκαν με () μέσα σε ()», όπου το πρώτο στοιχείο θα αντικατασταθεί με την κατάλληλη ασθένεια/ασθένειες και το δεύτερο στοιχείο από την αντίστοιχη χρονική περίοδο.

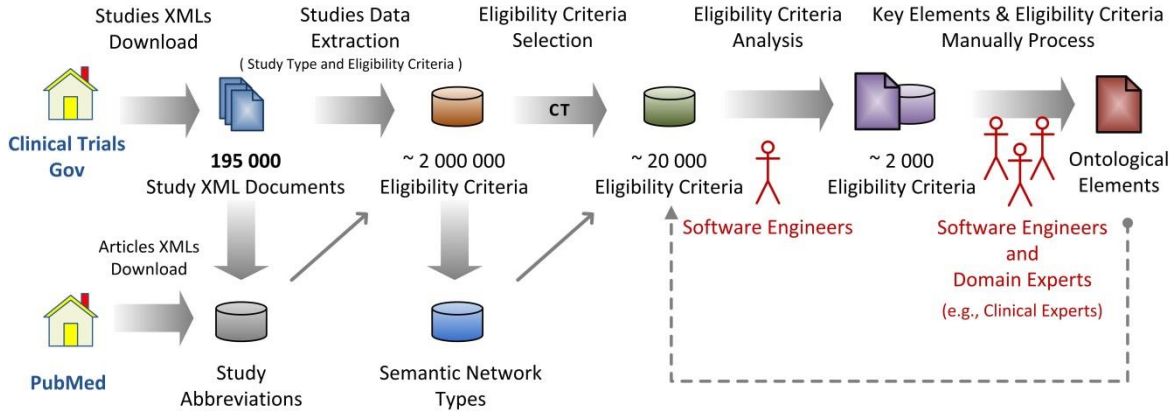
Η αναπαράσταση ΚΚ με βάση κάποιο πρότυπο, εκτός απ' το γεγονός ότι διευκολύνει την αναπαράσταση των κριτηρίων προτείνοντας πιθανές παραμέτρους που εμφανίζονται συνήθως μαζί (π.χ., η ασθένεια σε συνδυασμό με κάποιο χρονικό περιορισμό), περιορίζει το χρήστη, ο ρόλος του οποίου πλέον βρίσκεται στον ορισμό των εσωτερικών παραμέτρων ενός προτύπου, χωρίς γενικά να έχει τη δυνατότητα να

συνδυάσει διαφορετικά στοιχεία για την έκφραση ιδιαίτερα περίπλοκων κριτηρίων. Η ανάλυση της πολυπλοκότητας των ΚΚ από τους Ross κ.ά. [41] επισημαίνει ότι οι χρήστες ορίζουν και απλά αλλά και περίπλοκα κριτήρια (κριτήρια διαφορετικά από το μοντέλο παράμετρος - τιμή), τα οποία αποτελούν πάνω από το 80% των συνολικών κριτηρίων με τους χρονικούς περιορισμούς να αποτελούν περίπου το 40% αυτών [42]. Η ανάγκη για μια ιδιαιτέρως εκφραστική γλώσσα αναπαράστασης ΚΚ που περιλαμβάνει υπάρχουσες, διεθνώς αναγνωρισμένες, ονοματολογίες και κωδικοποιήσεις επίσης υποστηρίζεται από την ανάλυση 24 φαινοτύπων αλγορίθμων σχετικών με τα ηλεκτρονικά δεδομένα υγείας που αναπτύχθηκαν κατά το έργο eMERGE [43], οι οποίοι έδειξαν ότι παρά τις διαφορές ανάμεσα στους αλγορίθμους που χρησιμοποιούνται (πχ. ποικίλλουν στην έκταση και μορφή των δεδομένων) η λογική που βρίσκεται από πίσω είναι αρκετά ομοιογενής και βασίζεται ιδιαίτερα σε εμφωλευμένες λογικές εκφράσεις, περίπλοκους χρονικούς περιορισμούς και ευρέως χρησιμοποιούμενους ICD-9 κωδικούς [44].

3 *Μεθοδολογία*

3.1 Λήψη, Επεξεργασία και Επιλογή Κριτηρίων Καταλληλότητας

Ο σχεδιασμός της αναπαράστασης των Κριτηρίων Καταλληλότητας (ΚΚ) μιας κλινικής δοκιμής καθοδηγήθηκε από την ανάλυση των κριτηρίων [45] που έχουν ορισθεί σε υπάρχουσες κλινικές δοκιμές που δημοσιεύτηκαν στην τοποθεσία ClinicalTrials.gov [35]. Το Σχήμα 1 απεικονίζει τη συνολική προσέγγιση που ακολουθήθηκε, συμπεριλαμβανομένου του μεγέθους των εγγράφων (κλινικών δοκιμών) ή κριτηρίων που χρησιμοποιήθηκαν, καθώς επίσης και τους χρήστες που συμμετείχαν στη διαδικασία αυτή.



Σχήμα 1: Προσέγγιση που Ακολουθήθηκε για την Επεξεργασία των Κριτηρίων Καταλληλότητας.

Αρχικά, τα δεδομένα από όλες τις διαθέσιμες κλινικές μελέτες ελήφθησαν (περίπου 195 χιλιάδες κλινικές μελέτες ήταν διαθέσιμες τον Αύγουστο του 2015) και

κατόπιν εισήχθησαν σε μία σχεσιακή βάση, συμπεριλαμβανομένων των ΚΚ που είχαν ορισθεί. Για την εύρεση των μεμονωμένων κριτηρίων που είχαν ορισθεί, το κείμενο με τα ΚΚ επεξεργάστηκε, λαμβάνοντας υπόψη συγκεκριμένες λέξεις και φράσεις που χρησιμοποιήθηκαν για την εισαγωγή μιας λίστας από κριτήρια εισαγωγής ή εξαγωγής ασθενών, καθώς επίσης και συγκεκριμένα σημεία στίξης ή αριθμούς που χρησιμοποιούνται συνήθως κατά την εισαγωγή ενός νέου κριτηρίου. Η αναλυτική περιγραφή της διαδικασίας που ακολουθήθηκε για την εξαγωγή των κριτηρίων βρίσκεται στο παράρτημα 30.1. Ακολουθώντας την παραπάνω διαδικασία, καταφέραμε να εξάγουμε μια λίστα από 1.9 εκατομμύρια ΚΚ [46].

Για να επιτρέψουμε στους χρήστες καθώς επίσης και σε πιθανά συστήματα να «καταλάβουν» τα ΚΚ, χωρίς να είναι απαραίτητο να λάβουν υπόψη την περιγραφή μιας κλινικής δοκιμής, οι συντομεύσεις (abbreviations aka acronyms, shorthands, initialisms) που χρησιμοποιήθηκαν κατά τον ορισμό των ΚΚ επιλύθηκαν με τη σημασία τους. Αξίζει να σημειωθεί ότι το 22.7% των ΚΚ περιέχει τουλάχιστον μία συντόμευση, χωρίς να έχουμε λάβει υπόψη Λατινικές συντομογραφίες (όπως, π.χ.) και Μονάδες Μέτρησης (όπως, μγ) που συνήθως χρησιμοποιούνται σε συντετμημένη μορφή με τη σημασίας τους να είναι προφανής. Για το σκοπό αυτό, χρησιμοποιήσαμε ένα σύστημα που εμείς αναπτύξαμε [47] για την εύρεση των συντομεύσεων και κυρίως της εκτεταμένης μορφής τους, με την προϋπόθεση ότι αυτές είχαν οριστεί κατά την περιγραφή της κλινικής δοκιμής. Για την εύρεση της σημασίας των υπόλοιπων συντομεύσεων που χρησιμοποιούνται στα ΚΚ, το παραπάνω σύστημα χρησιμοποιήθηκε για τη δημιουργία μιας βάσης δεδομένων με όλες τις πιθανές σημασίες μιας συντόμευσης, λαμβάνοντας υπόψη τα δεδομένα που ελήφθησαν από την τοποθεσία ClinicalTrials.gov καθώς και PubMed [48]. Η χρησιμοποίηση των δύο παραπάνω πηγών δεδομένων ήταν απαραίτητη, καθώς ένας σημαντικός αριθμός των συντομεύσεων που χρησιμοποιήθηκαν στα ΚΚ δεν

είχαν ορισθεί σε καμία άλλη κλινική δοκιμή [49]. Η παραπάνω βάση δεδομένων κατόπιν χρησιμοποιήθηκε για την επιλογή της κατάλληλης σημασίας των συντομεύσεων, λαμβάνοντας υπόψη τα συμφραζόμενα.

Στα πλαίσια της εργασίας αυτής επικεντρωθήκαμε στις 158 χιλιάδες διαθέσιμες παρεμβατικές κλινικές μελέτες και πιο συγκεκριμένα στα 1.7 εκατομμύρια ΚΚ που είχαν οριστεί. Για τον αποτελεσματικό χειρισμό του μεγάλου όγκου των ΚΚ που εξήχθησαν, αναπτύξαμε ένα σύστημα με πρωταρχικό στόχο την οργάνωση των ΚΚ σε κατηγορίες, λαμβάνοντας υπόψη τις έννοιες που χρησιμοποιήθηκαν και κυρίως τις ευρύτερες κατηγορίες στις οποίες ανήκουν (τύπους του UMLS σημασιολογικού δικτύου). Για το σκοπό αυτό χρησιμοποιήθηκε ο Ανοιχτός Βιοϊατρικός Επισημασητής (OBA) [50] καθώς και μια ποικιλία από ονοματολογίες, όπως οι Κλινικοί Όροι Συστηματοποιημένης Ονοματολογίας Ιατρικής (SNOMED-CT) [51] και οι Ιατρικές Θεματικές Ενότητες (MeSH) [52]. Επίσης, ο Mgrep αλγόριθμος [53] χρησιμοποιήθηκε για την ανεύρεση σχετικών όρων οι οποίοι κατόπιν επεξεργάστηκαν για την εύρεση της ευρύτερης κατηγορίας στην οποία ανήκουν, λαμβάνοντας υπόψη προ-υπάρχουσες οντολογίες και συσχετίσεις που υπάρχουν στο UMLS και BioPortal.

Για λόγους ανάλυσης, περισσότερα από 20 χιλιάδες ΚΚ επιλέχθηκαν (περίπου 1% των συνολικών ΚΚ) με μια διαδικασία που είχε ως στόχο να καλύψει το ευρύ φάσμα των ΚΚ. Για το σκοπό αυτό, αρχικά μετρήσαμε πόσες φορές εμφανίζεται καθένας από τους τύπους του UMLS σημασιολογικού δικτύου στο σύνολο των ΚΚ των παρεμβατικών κλινικών δοκιμών, καθώς επίσης και τους όρους που χρησιμοποιούνται σε καθένα από αυτά. Κατόπιν, τα κριτήρια επιλέχθηκαν έτσι, ώστε το ποσοστό των επιλεγμένων κριτηρίων που ανήκουν σε καθεμιά από τις ευρύτερες κατηγορίες να είναι ίδιο με το ποσοστό των συνολικών ΚΚ που ανήκουν στην εκάστοτε κατηγορία, καθώς επίσης και οι

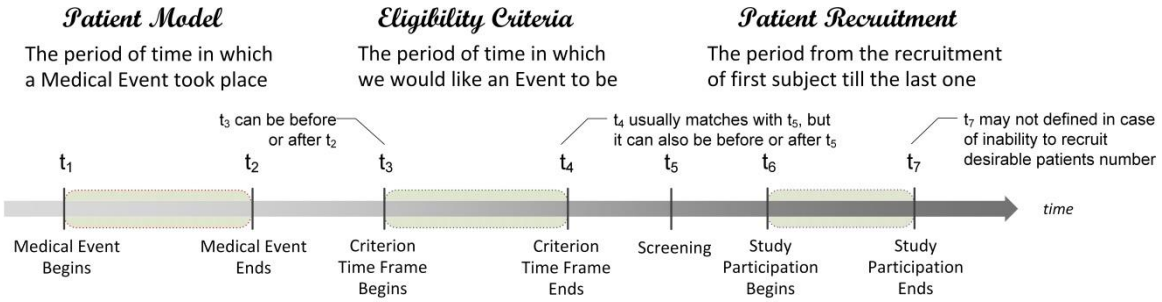
έννοιες που χρησιμοποιούνται στα επιλεγμένα ΚΚ να χρησιμοποιούνται με την ίδια συχνότητα.

3.2 Κατηγοριοποίηση και Ανάλυση Κριτηρίων Καταλληλότητας

Τα τυχαίως επιλεγμένα ΚΚ ακολούθως συνδέθηκαν με τουλάχιστον μία από τις ακόλουθες 7 κατηγορίες: Δημογραφικά (Demographics), Κριτήρια ανά Γένος (Gender-specific Criteria), Τρόπος Διαβίωσης (Lifestyle), Διαταραχές (Disorders), Παρεμβάσεις (Interventions), Διαγνωστικά ή Εργαστηριακά Τεστ (Diagnostic and Lab Tests) και Κριτήρια Συσχετιζόμενα με Κλινικές Μελέτες (Study-related Criteria). Για παράδειγμα, ένα κριτήριο στο οποίο χρησιμοποιήθηκε κάποια Φαρμακολογική Ουσία, διασυνδέθηκε με την κατηγορία της Παρέμβασης. Στην εργασία αυτή, τα κριτήρια που σχετίζονται με μία συγκεκριμένη κατηγορία ασθενών (δηλαδή, με βάση το Γένος) τοποθετήθηκαν σε μία ξεχωριστή κατηγορία. Επίσης, τα κριτήρια που δεν ανήκουν σε κάποια από τις 6 πρώτες κατηγορίες τοποθετήθηκαν στην τελευταία κατηγορία. Ο ορισμός των παραπάνω κατηγοριών βασίστηκε σε μια διεξοδική ανάλυση σχετικής δουλειάς που δημοσιεύτηκε σε συνέδρια και περιοδικά που χαίρουν ιδιαίτερης εκτίμησης [37][34], καθώς επίσης και στη συνεισφορά των 8 κλινικών εμπειρογνωμόνων από διαφορετικούς τομείς (συμπεριλαμβανομένων καρδιολογίας, ενδοκρινολογίας, ψυχιατρικής και φαρμακολογίας) που συμμετείχαν στην ανάλυση των ΚΚ.

Τα κριτήρια που ανήκουν σε καθεμία από τις παραπάνω επτά κατηγορίες εξετάστηκαν λεπτομερώς με κάποιο ημιαυτόματο τρόπο για την εξαγωγή των υποκατηγοριών (π.χ., Χορήγηση Φαρμάκου, Χειρουργική Επέμβαση, κτλ.), καθώς επίσης και των παραμέτρων που σχετίζονται άμεσα με καθεμία από τις παραπάνω κατηγορίες (π.χ., δοσολογία φαρμάκου, τρόπος χορήγησης, μορφή δόσης, κτλ.). Η διαδικασία που ακολουθήθηκε ήταν σε μεγάλο βαθμό χειρωνακτική καθώς υπάρχοντες αλγόριθμοι και τεχνικές για την εύρεση συσχέτισης μεταξύ εννοιών δεν μπορούν να αποδώσουν

ικανοποιητικά αποτελέσματα [54], ειδικότερα όταν αυτές είναι άγνωστες. Παρόλα αυτά, οι φράσεις ή μοτίβα που χρησιμοποιούνται συνήθως για την έκφραση της συσχέτισης μεταξύ των εννοιών καταγράφηκαν και χρησιμοποιήθηκαν, για να φιλτράρουμε τα υπό-επεξεργασία ΚΚ. Με την παραπάνω διαδικασία, συγκεντρώσαμε έναν σχετικά μικρό αριθμό ΚΚ για καθεμία από τις παραμέτρους που εντοπίσαμε, τα οποία κατόπιν χρησιμοποιήσαμε στις συζητήσεις μας με τους ειδικούς στον τομέα αυτό. Συνολικά, περίπου 2 χιλιάδες ΚΚ συγκεντρώθηκαν (δηλαδή, το 10% των κριτηρίων που επιλέχθηκαν), τα οποία εξετάσαμε προσεκτικά με την ενεργή συμμετοχή των ειδικών σε θέματα κλινικών δοκιμών. Η παραπάνω διαδικασία διήρκεσε 3 χρόνια.



Σχήμα 2: Εννοιολογική χρονική τοποθέτηση των κριτηρίων καταλληλότητας σε σύγκριση με χαρακτηριστικά γεγονότα στο χώρο της περίθαλψης και της κλινικής έρευνας.

Τα παραπάνω ΚΚ τα εξετάσαμε υπό την έννοια της σχέσης των χαρακτηριστικών που περιγράφονται με τον ασθενή και το σκοπό που αυτά εξυπηρετούν, καθώς επίσης και της διασποράς τους στον χρόνο. Το Σχήμα 2 απεικονίζει διαφορετικές χρονικές περιόδους στις οποίες τα ΚΚ μπορεί να αναφέρονται (t3:t4) καθώς και τις θέσεις τους, σχετικά με:

α) Γεγονότα σχετικά με την Υγεία του Ασθενούς (t1:t2) τα οποία αναμένεται να καταγράφονται σε έναν οργανισμό φροντίδας υγείας (π.χ., Νοσοκομείο)

β) Χαρακτηριστικά γεγονότα αναφορικά με μία κλινική δοκιμή, όπως έλεγχος ασθενών (t5) καθώς επίσης και των χρονικών σημείων εισαγωγής (t6) και εξαγωγής (t7) από μια κλινική δοκιμή.

Στην πραγματικότητα, η ανάλυση έδειξε ότι, όταν ορίζουμε τον πληθυσμό μιας κλινικής δοκιμής, οι ερευνητές ενδιαφέρονται όχι μόνο για την τρέχουσα κατάσταση του ασθενούς ή το ιστορικό του, αλλά και για τη συμπεριφορά και πρόοδό του στο μέλλον (δηλ, $t4 \geq t5$). Παραδείγματα αποτελούν η ικανότητα ενός ασθενούς να λαμβάνει το υπό-εξερεύνηση φάρμακο, καθώς επίσης και η θέληση του ασθενούς να εκτελέσει μία συγκεκριμένη εργασία κατά τη διάρκεια μιας κλινικής δοκιμής (π.χ., η πραγματοποίηση μιας εξέτασης κάθε 2 εβδομάδες για όλη την περίοδο της κλινικής δοκιμής).

Επίσης, οι ερευνητές μπορεί να μην ορίσουν άμεσα τα χαρακτηριστικά που πρέπει να πληρούν οι ασθενείς που θα λάβουν μέρος στην κλινική δοκιμή, αλλά έμμεσα, έχοντας ως στόχο να περιορίσουν τους παράγοντες που συμβάλλουν στην αύξηση του ρίσκου συμμετοχής ενός ασθενούς και επηρεάζουν αρνητικά την αποτελεσματικότητα της δοκιμής, για παράδειγμα «κρύβοντας» τις επιπτώσεις του φαρμάκου «επισκιάζοντας» την ανάλυση των δεδομένων μελέτης. Τα κριτήρια αυτά συνήθως είναι κριτήρια εξαγωγής ενός ασθενούς από την κλινική δοκιμή και μπορεί να περιλαμβάνουν συγκεκριμένες διαταραχές, μη ικανοποιητική λειτουργία ενός οργάνου, διανοητική δυσλειτουργία που ακολουθείται από συγκεκριμένες θεραπείες, προσδόκιμο ζωής, κτλ.

3.3 Δημιουργία Μοντέλου Περιγραφής Κριτηρίων Καταλληλότητας

Λαμβάνοντας υπόψη το αποτέλεσμα της προηγούμενης διαδικασίας και ειδικότερα τις κλάσεις και παραμέτρους που ανιχνεύτηκαν, κατασκευάστηκε ένα μοντέλο το οποίο εκφράζει τις σημασιολογικές κλάσεις και υποκλάσεις στις οποίες ανήκουν τα ΚΚ, καθώς επίσης και τις παραμέτρους με ιδιαίτερη ερευνητική σημασία που επηρεάζουν την επιλογή των ασθενών. Δεδομένης της πολυπλοκότητας και ποικιλίας των ΚΚ, οι παράμετροι που συμπεριλήφθησαν στο μοντέλο μας ακολούθησαν τη σειρά προτεραιότητας, λαμβάνοντας υπόψη τη συχνότητα εμφάνισής τους καθώς και την

κλινική τους σημασία. Ο σκοπός των κριτηρίων ελήφθη επίσης υπόψη για τον ορισμό των παραμέτρων, καθώς επίσης και τη διασύνδεση του ασθενούς με τις κλάσεις που ορίσθηκαν.

Η ονοματολογία, κατηγοριοποίηση και ορισμός των όρων που συμπεριλήφθησαν στο μοντέλο μας βασίστηκε στην παραπάνω ανάλυση, καθώς επίσης και στην ανάλυση των προτύπων φροντίδας υγείας (healthcare) και κλινικής έρευνας (clinical research) [55][56], με απώτερο σκοπό την ανάπτυξη μιας διαλειτουργικής αναπαράστασης που μπορεί να εξυπηρετήσει διαφορετικές ανάγκες, όπως την «ανταλλαγή» κριτηρίων κλινικών δοκιμών ή τη διασύνδεσή τους με τους οργανισμούς φροντίδας υγείας. Πιο συγκεκριμένα, για κάθε όρο που εντοπίσαμε η διασύνδεσή του με σχετικές έννοιες από τα CDISC και HL7 Μοντέλα Αναφοράς διερευνήθηκε. Το μοντέλο επίσης συνδέθηκε με έννοιες και όρους από τα OpenEHR αρχέτυπα [57], έχοντας ως απώτερο στόχο τη διασύνδεσή του με βάσεις ασθενών. Τα στοιχεία που εντοπίστηκαν στα παραπάνω μοντέλα συμπεριλήφθησαν στον ορισμό του εκάστοτε όρου στο μοντέλο που αναπτύχθηκε.

Οι όροι που χρησιμοποιήθηκαν για τον ορισμό των ΚΚ βασίστηκαν σε διεθνώς αναγνωρισμένες κωδικοποιήσεις και κατηγοριοποιήσεις, όπως το ICD-10, οι Χημικές οντότητες βιολογικού ενδιαφέροντος (ChEBI) [58], το LOINC, κτλ., επιτρέποντας στους ερευνητές να ορίσουν τα κριτήρια με τη λεπτομέρεια που χρειάζεται (π.χ., αποκλείοντας πιθανούς ασθενείς με μη-αλλεργικό άσθμα ή γενικά κάποια χρόνια ασθένεια). Οι επιλεγμένες ονοματολογίες συνδέθηκαν με το μοντέλο που αναπτύχθηκε, προσδιορίζοντας την ονομασία και κυρίως τοποθεσία της εκάστοτε ονοματολογίας στον ιστό. Με αυτό το τρόπο, οι διαθέσιμοι όροι καθώς και η σημασία τους (όπως ορίζονται απ τον οργανισμό που είναι υπεύθυνος για τη δημοσίευσή τους) μπορεί να προσδιοριστούν, ενώ οι όροι αυτοί καθ’ αυτοί εισήχθησαν σε μία ξεχωριστή βάση δεδομένων.

Το Μοντέλο ΚΚ καθώς και οι ονοματολογίες που έχουν προς το παρόν διασυνδεθεί με το μοντέλο εκφράστηκαν με τη μορφή μιας OWL-2 [59] οντολογίας, επονομαζόμενης «Οντολογία Κριτηρίων Καταλληλότητας».

3.4 Αναπαράσταση και Έκφραση Κριτηρίων Καταλληλότητας

Για την οργάνωση και καταγραφή των ΚΚ μιας κλινικής δοκιμής, αναπτύχθηκε ένα Επεκτάσιμης Γλώσσας Σήμανσης (XML) Schema [60] στο οποίο υπάρχει σαφής διαχωρισμός ανάμεσα στα κριτήρια εισαγωγής και εξαγωγής. Επίσης, για κάθε κριτήριο πρέπει να καταγράφεται ένα μοναδικό αναγνωριστικό (ID), μια αναπαράσταση του κριτηρίου, χρησιμοποιώντας τη φυσική γλώσσα καθώς και μηδέν μία ή περισσότερες αναπαραστάσεις – εκφράσεις του κριτηρίου σε γλώσσα του υπολογιστή (τυπική έκφραση κριτηρίου). Το σχήμα που αναπτύχθηκε είναι πολύ πιο απλό από το σχήμα που υπάρχει διαθέσιμο από το CDISC (υπάρχει ένα προς ένα αντιστοίχιση με τα αντίστοιχα ODM και SDM στοιχεία), ενώ υπάρχει ξεκάθαρος διαχωρισμός ανάμεσα στη γλώσσα (π.χ., SPARQL) και στο μοντέλο (π.χ., EC-O) που χρησιμοποιήθηκαν για την έκφραση του κριτηρίου. Στο σχήμα που αναπτύχθηκε, ο χρήστης μπορεί επιπλέον να ορίσει την οντότητα που παρήγαγε την τυπική έκφραση του κριτηρίου (π.χ., από το χρήστη μέσω ενός γραφικού περιβάλλοντος), εάν η έκφραση αυτή εκφράζει με ακρίβεια το κριτήριο που περιέγραψε ο χρήστης, χρησιμοποιώντας τη φυσική γλώσσα καθώς και κάθε σχόλιο που είναι απαραίτητο.

Σχετικά με την τυπική έκφραση ενός κριτηρίου, δύο επιλογές υποστηρίζονται: XML [61] και SPARQL [62]. Και στις δύο αυτές επιλογές η έκφραση του εκάστοτε κριτηρίου βασίζεται στο μοντέλο που αναπτύχθηκε. Τα διαφορετικά χαρακτηριστικά των δύο παραπάνω γλωσσών διευκολύνουν την υλοποίηση διαφορετικών διαδικασιών που είναι άμεσα συνδεδεμένες με τα ΚΚ. Πιο συγκεκριμένα, η XML αναπαράσταση των ΚΚ διευκολύνει τον εντοπισμό και την επεξεργασία των παραμέτρων ενός κριτηρίου και

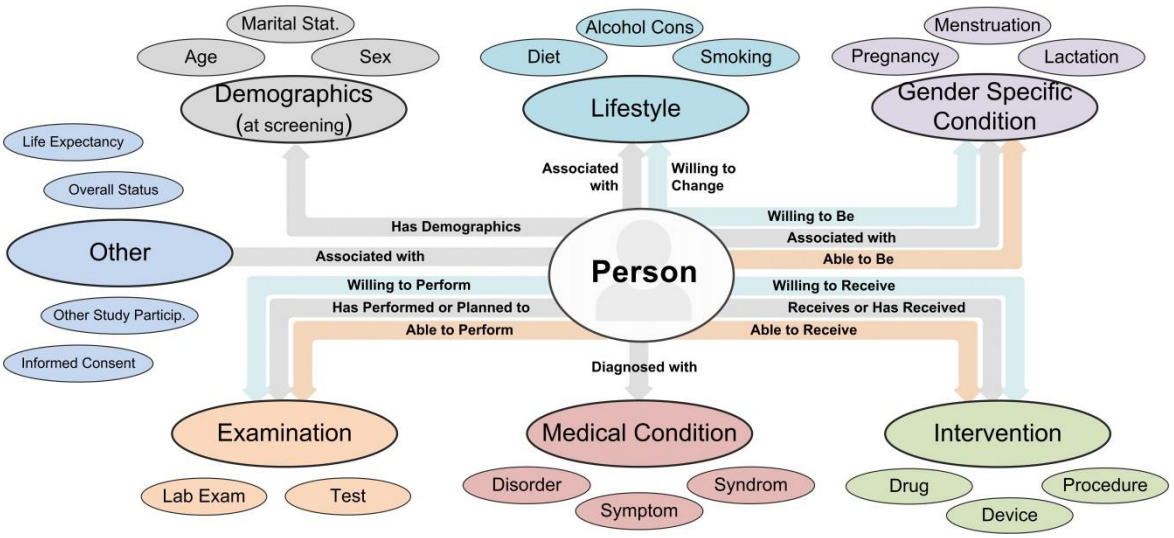
επομένως θα μπορούσε να χρησιμοποιηθεί για λόγους υποστήριξης απόφασης. Απ' την άλλη, η SPARQL αναπαράσταση των ΚΚ διευκολύνει τη μετατροπή των κριτηρίων σε ερωτήματα προς μια εικονική ή πραγματική βάση δεδομένων ασθενών και επομένως μπορεί να συμβάλλει στον αυτόματο εντοπισμό των ασθενών, που πιθανώς να είναι κατάλληλοι να συμμετέχουν στην κλινική δοκιμή (συζητείται στην Ενότητα 6.2). Αξίζει να σημειωθεί ότι τα στοιχεία που ορίσαμε για την XML αναπαράσταση των ΚΚ καθορίστηκαν με βάση τους όρους που υπάρχουν στην οντολογία. Κατά συνέπεια, η μετάβαση του τρόπου αναπαράστασης ενός κριτηρίου από XML σε SPARQL μπορεί να γίνει με κάποιο αυτόματο τρόπο (περιγράφεται συνοπτικά στην Ενότητα 4.2).

Η σελίδα αυτή είναι σκόπιμα λευκή

4 *Μοντέλα Περιγραφής Κριτηρίων*

4.1 Οντολογία Κριτηρίων Καταλληλότητας

Η Οντολογία Κριτηρίων Καταλληλότητας (ΚΚ) που αναπτύχθηκε περιλαμβάνει ένα ευρύ φάσμα από παραμέτρους (πάνω από 500 οντολογικά στοιχεία ορίσθηκαν) για τον ορισμό των ΚΚ μιας κλινικής δοκιμής.



Σχήμα 3: Συνοπτική εικόνα του μοντέλου που αναπτύχθηκε για την έκφραση των Κριτηρίων Καταλληλότητας.

Στο Σχήμα 3, παρουσιάζεται η συνολική εικόνα της οντολογίας. Τα δεδομένα είναι οργανωμένα σε κατηγορίες, επονομαζόμενες Δημογραφικά (συμπεριλαμβανομένων Ηλικίας, Φύλου, Εθνικότητας, Φυλής, Οικογενειακής Κατάστασης), Τρόπος Διαβίωσης (συμπεριλαμβανομένων Δίαιτας, Δραστηριοτήτων και Κατανάλωσης Καπνού),

Παράμετροι σχετικές με το Γένος (συμπεριλαμβανομένων εγκυμοσύνης και θηλασμού), Κατάσταση (συμπεριλαμβανομένων Διαταραχών, Συμπτωμάτων, Συνδρόμων, Δυσλειτουργιών και Βλαβών), Παρέμβαση / Αντιμετώπιση (συμπεριλαμβανομένων Φαρμάκων, Εγχειρήσεων, Ακτινοβολίας κτλ.), καθώς και κάποιες άλλες παραμέτρους που δεν είχαν συμπεριληφθεί στις παραπάνω κατηγορίες (π.χ., συμμετοχή σε άλλη κλινική δοκιμή).

Οι σχέσεις μεταξύ των κατηγοριών των δεδομένων με τον Άνθρωπο/Ασθενή ορίσθηκαν λαμβάνοντας υπόψη την ανάλυση που έλαβε χώρα, η οποία έδειξε ότι τα κριτήρια ορίζουν τα χαρακτηριστικά που θα πρέπει να έχουν ή να πληρούν οι συμμετέχοντες καθώς επίσης και τις δραστηριότητες ή διαδικασίες στις οποίες θα ήταν πρόθυμοι να συμμετέχουν, να εκτελέσουν ή γενικότερα να ακολουθήσουν. Για παράδειγμα, ένας άνθρωπος συσχετίζεται με μία παρέμβαση με διαφορετικούς τρόπους: α) λαμβάνει-ακολουθεί μια παρέμβαση (π.χ. λαμβάνει ένα φάρμακο), β) έλαβε στο παρελθόν ή γ) σκοπεύει να λάβει στο άμεσο μέλλον (π.χ., μια προγραμματισμένη εγχείρηση). Εκτός από αυτές τις προφανείς συσχετίσεις ανάμεσα στις δύο οντότητες, στην οντολογία ορίσαμε επιπλέον την ικανότητα και προθυμία ενός ανθρώπου να λάβει-ακολουθήσει μία παρέμβαση, το οποίο με τη σειρά του πολλαπλασιάζει επί τρία τα κριτήρια που θα μπορούσαμε να ορίσουμε σχετικά με τον τρόπο χειρισμού/παρέμβασης ενός ιατρικού προβλήματος ή γενικότερα διαταραχής. Ο Πίνακας 1 παρουσιάζει τις βασικές παραμέτρους που ορίσθηκαν στην οντολογία για κάποιες από τις προαναφερθείσες κατηγορίες. Για παράδειγμα, στην περίπτωση των διαγνώσεων, το σύνολο των κατάλληλων ασθενών μπορεί πιθανώς να επηρεάζεται όχι μόνο από τη διαταραχή αλλά και από την ημερομηνία που αυτή εκδηλώθηκε, τη σοβαρότητά της, το στάδιο στο οποίο βρίσκεται, τα συμπτώματα που αυτή έχει ή ακόμη το λόγο που την προκάλεσε. Ένα άλλο παράδειγμα έχει να κάνει με τα φάρμακα που έχουν χορηγηθεί σε

έναν ασθενή, για τα οποία είναι σημαντικό να γνωρίσουμε όχι μόνο το φάρμακο ή τη δραστική ουσία αλλά επίσης τη δοσολογία, τη μορφή του φαρμάκου και τον τρόπο χορήγησης, την περίοδο που έχει συνταγογραφηθεί να λάβει το φάρμακο ο ασθενής, τη συχνότητα λήψης φαρμάκου/δοσολογίας, καθώς επίσης και το λόγο που αυτό χορηγήθηκε (π.χ., για λόγους πρόληψης).

Σημασιολογική Κλάση	**Παράμετροι**
Δημογραφικά Χαρακτηριστικά	Ηλικία
	Φύλο ή Γένος
	Εθνικότητα
	Φυλή
	Θρησκεία
	Γλωσσική Ικανότητα
	Επίπεδο Μόρφωσης
	Οικογενειακή Κατάσταση
	Δεδομένα Εργασίας
	Δεδομένα Στέγασης
Διάγνωση (Medical Condition)	Κωδικός Διάγνωσης
	Στάδιο ή Σοβαρότητα
	Ημέρα που Ξέσπασε
	Ημέρα που Κλινικά Αναγνωρίστηκε
	Αιτία/Αιτίες της Ιατρικής Κατάστασης
	Συμπτώματα
Φάρμακα (Intervention)	Κωδικός Φαρμάκου
	Δραστικές Ουσίες
	Ποσότητα Φαρμάκου
	Μορφή Δόσης
	Τρόπος Χορήγησης
	Περίοδος Λήψης Φαρμάκου
	Συχνότητα Λήψης Φαρμάκου
	Αιτία Λήψης Φαρμάκου

Σημασιολογική Κλάση	**Παράμετροι**
Εργαστηριακές Εξετάσεις	Κωδικός Εξέτασης
	Αποτέλεσμα Εξέτασης
	Ημερομηνία Εξέτασης
Κατανάλωση Καπνού (Lifestyle Choice)	Δείκτης Κατανάλωσης
	Συχνότητα Χρήσης
	Ημερομηνία που Ξεκίνησε
	Ποσότητα Καπνίσματος
	Συχνότητα Καπνίσματος
Εγκυμοσύνη, Εμμηνόπαυση (Gender Specific Condition)	Αποτέλεσμα Τεστ Εγκυμοσύνης
	Ημερομηνία Σύλληψης
	Κωδικός Εισαγωγής Εμμηνόπαυσης
	Ημερομηνία Εισαγωγής Εμμηνόπαυσης

Πίνακας 1: Υποσύνολο των παραμέτρων της Οντολογίας Κριτηρίων Καταλληλότητας.

Θα πρέπει να σημειωθεί ότι κατά τον ορισμό των στοιχείων της οντολογίας ελήφθη υπόψη η σημασιολογική απόσταση ανάμεσα στην κλινική έρευνα και περίθαλψη, δίνοντας έμφαση στην ορολογία που χρησιμοποιείται για την αναπαράσταση των ΚΚ, πάρα τους όρους που ορίζονται στον ηλεκτρονικό φάκελο ενός ασθενούς. Για παράδειγμα, σύμφωνα με το μοντέλο μας, ένα όργανο μπορεί να χαρακτηριστεί ως «λειτουργεί ικανοποιητικά», ενώ αυτή η πληροφορία δεν αναμένεται να υπάρχει (τουλάχιστον με αυτή τη μορφή) στη βάση των ασθενών. Επίσης, στην περίπτωση των εργαστηριακών μετρήσεων, συμπεριλήφθηκε η έννοια του Άνω (και Κάτω) Φυσιολογικού Ορίου, το οποίο επιτρέπει στους ειδικούς στον τομέα της κλινικής έρευνας να εκφράσουν κριτήρια σχετικά με τις εργαστηριακές μετρήσεις, ανεξάρτητα απ' το εργαστήριο που αυτές έλαβαν χώρα, καθώς τα παραπάνω όρια ποικίλλουν και παρουσιάζουν σημαντικές διαφορές ανά εργαστήριο και νοσοκομείο. Στο μοντέλο μας ορίστηκαν επίσης παράμετροι που σχετίζονται άμεσα με μία κλινική δοκιμή, όπως η ικανότητα ενός ανθρώπου να δώσει

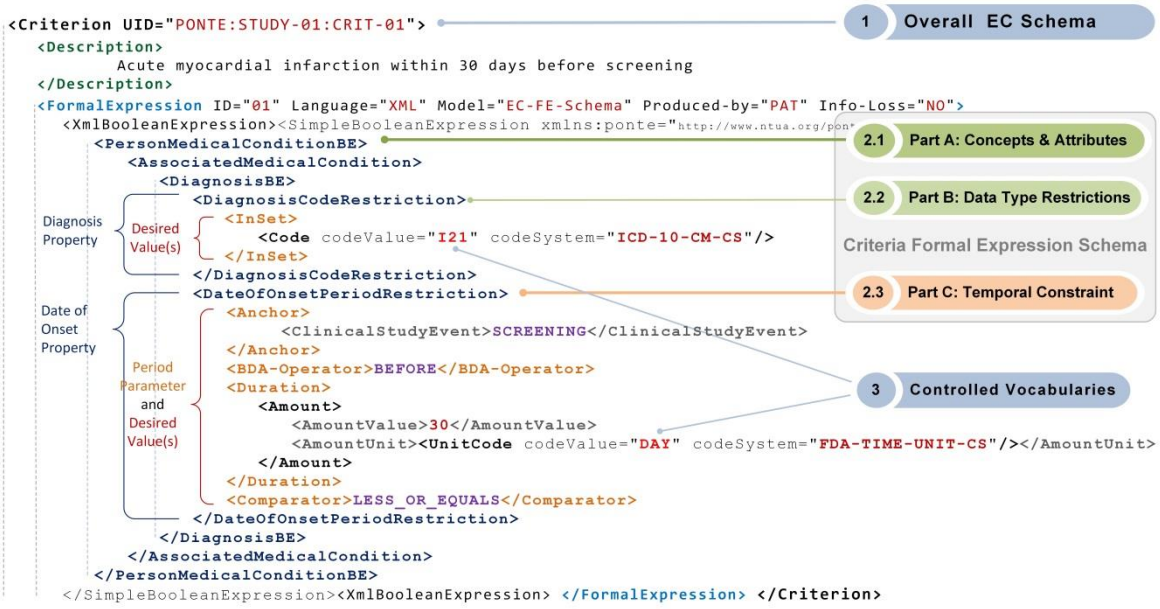
τη συγκατάθεσή του (informed consent) είτε από μόνος του είτε μέσω κάποιου συγγενούς, η συμφωνία του με τις διαδικασίες που ακολουθούνται κατά την κλινική δοκιμή, κτλ.

Οι ονοματολογίες που επιλέχθηκαν ορίσθηκαν στην οντολογία (δηλαδή, αναγνωριστικό, όνομα, περιγραφή, υπεύθυνος οργανισμός και τοποθεσία) και οι όροι/κωδικοί (ως έννοιες) που προέρχονται από την καθεμιά οργανώθηκαν σε ευρύτερες κατηγορίες με βάση το πεδίο που αυτές καλύπτουν. Ακολούθως, στον ορισμό των παραμέτρων, χρησιμοποιήθηκαν οι ευρύτερες κατηγορίες (π.χ., κωδικός διάγνωσης) παρά οι συγκεκριμένες ονοματολογίες (π.χ., ICD-10-CM κωδικός), επιτρέποντας την ανανέωση (π.χ., αντικατάσταση ή προσθήκη) των ονοματολογιών, χωρίς να είναι απαραίτητη οποιαδήποτε αλλαγή στον ορισμό των παραμέτρων (εικόνα 6). Συνεπώς, πολλαπλά συστήματα κατηγοριοποίησης μπορούν να είναι ταυτόχρονα συνδεδεμένα με το μοντέλο μας έτσι, ώστε οι ερευνητές να μπορούν να χρησιμοποιήσουν αυτό που καλύπτει καλύτερα τις ανάγκες μιας κλινικής δοκιμής (π.χ., ονοματολογίες που εστιάζουν σε ένα συγκεκριμένο πεδίο γνώσης, όπως το Διαγνωστικό και Στατιστικό Εγχειρίδιο Ψυχικών Διαταραχών – 5^{η} έκδοση (DSM-5) [63]) ή αυτό με το οποίο είναι περισσότερο εξοικειωμένοι.

4.2 Τυπική Αναπαράσταση Κριτηρίου

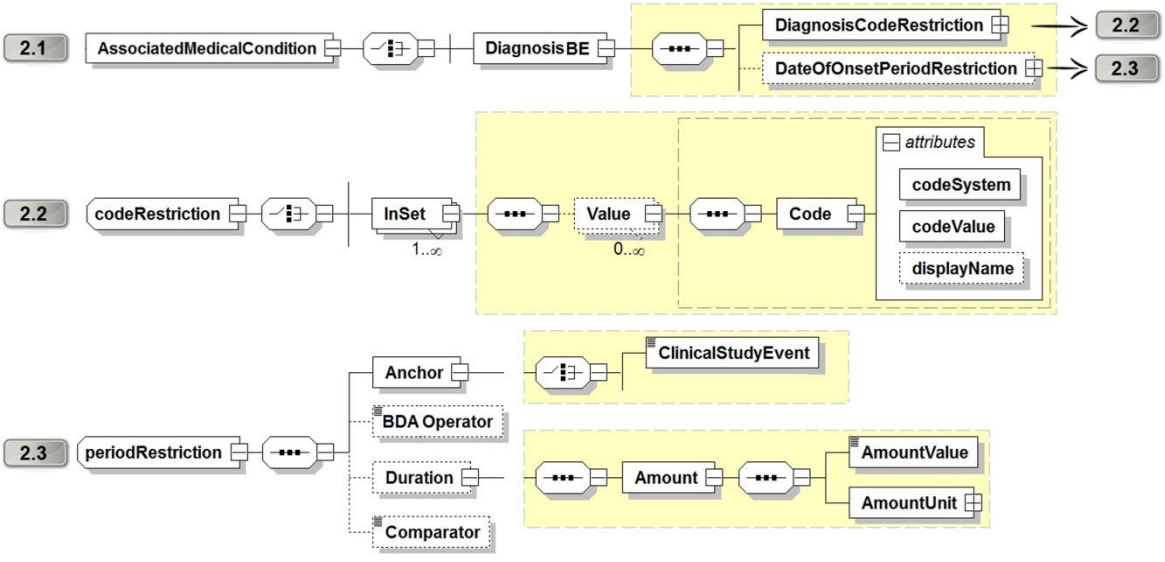
Σχετικά με την τυπική αναπαράσταση ενός κριτηρίου, δύο επιλογές είναι διαθέσιμες, επονομαζόμενες XML και SPARQL, με ένα παράδειγμα για την καθεμιά να υπάρχει στα σχήματα Σχήμα 4 και Σχήμα 6 αντίστοιχα. Η τυπική αναπαράσταση ενός ΚΚ με βάση την XML ορίσθηκε από το χρήστη μέσω κάποιου γραφικού περιβάλλοντος. Από την άλλη, η τυπική αναπαράσταση του κριτηρίου με βάση τη SPARQL παράχθηκε αυτόματα (περιγράφεται συνοπτικά πιο κάτω). Λόγω έλλειψης χώρου, η XML αναπαράσταση του κριτηρίου διαχωρίστηκε από τη SPARQL. Παρόλα αυτά, και οι δύο αναπαραστάσεις μπορούν να συνυπάρχουν στον ορισμό ενός κριτηρίου. Ο τύπος του κριτηρίου (δηλαδή,

κριτήριο εισαγωγής ή εξαγωγής) δηλώνεται μέσω της ευρύτερης κατηγορίας στην οποία αυτό ανήκει.



Σχήμα 4: Κριτήριο σχετικό με τις διαγνώσεις ενός ασθενούς, τυπικά εκφρασμένο σε XML.

Στο Σχήμα 4, παρουσιάζονται οι τρεις παράμετροι του κριτηρίου που ορίσαμε με βάση το ΚΚ σχήμα που αναπτύχθηκε (1).



Σχήμα 5: Τμήμα του Σχήματος Τυπικής Έκφρασης Κριτηρίων Καταλληλότητας για την XML αναπαράσταση ενός κριτηρίου.

Σχετικά με την τυπική περιγραφή του κριτηρίου, τα στοιχεία προέρχονται από το XML σχήμα (Σχήμα 5), ο σχεδιασμός του οποίου καθοδηγήθηκε από τις κλάσεις και παραμέτρους που ορίσαμε στο μοντέλο μας (2.1). Για την αναπαράσταση των συνθηκών που θα πρέπει αυτά να πληρούν, τα αντίστοιχα στοιχεία ορίσθηκαν λαμβάνοντας υπόψη τον τύπο των δεδομένων των εκάστοτε παραμέτρων (2.2). Στην περίπτωση των παραμέτρων, η τιμή των οποίων προέρχεται από κάποια ονοματολογία, ο κωδικός καθώς επίσης και η αναφορά στο εκάστοτε σύστημα κωδικοποίησης καταγράφηκε (3). Οι χρονικοί περιορισμοί (2.3) αποτελούν ειδική περίπτωση και ο καθορισμός τους βασίστηκε στις επόμενες τέσσερις παραμέτρους: Σημείο Αναφοράς, Χρονικός Δείκτης, Διάρκεια και Συχνότητα (Πίνακας 2).

Παράμετρος	**Υ/Π**	**Σύντομη Περιγραφή**
Σημείο Αναφοράς	Υ	Ένα γεγονός σχετικό με την κλινική μελέτη, όπως η διαδικασία επιλογής ασθενών και η λήψη ενός φαρμάκου.
Χρονικός Δείκτης	Υ	Ένας τελεστής που δείχνει εάν το γεγονός έλαβε ή θα λάβει χώρα «πριν», «κατά την διάρκεια» ή «μετά» το σημείο αναφοράς.
Διάρκεια	Π	Η χρονική διάρκεια της περιόδου, συχνά εκφρασμένη με έναν αριθμό ακολουθούμενο από μία μονάδα μέτρησης χρόνου.
Συχνότητα	Π	Ο αριθμός των επαναλήψεων που έλαβε χώρα ένα γεγονός στην χρονική περίοδο που έχει προσδιοριστεί.

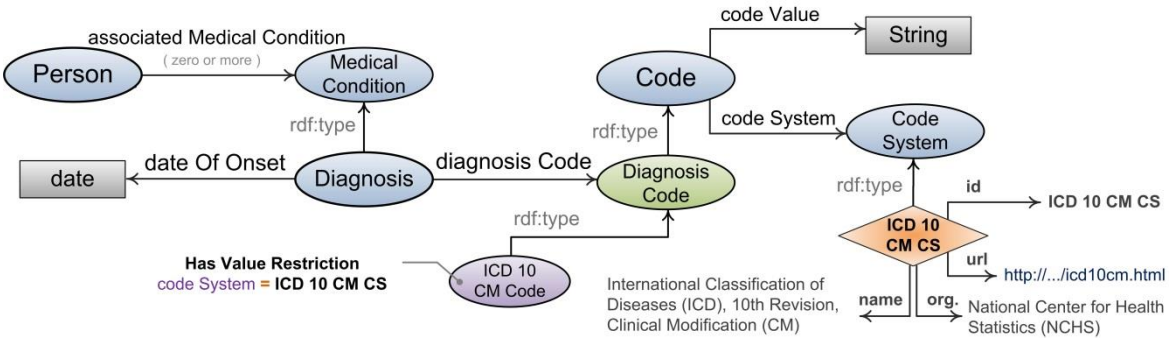
Πίνακας 2: Οι (Υ)ποχρεωτικές και (Π)ροαιρετικές παράμετροι ενός χρονικού περιορισμού.

Το επόμενο κριτήριο (παράδειγμα) δείχνει τη χρήση των παραμέτρων αυτών «Διάγνωση Εμφράγματος του Μυοκαρδίου τουλάχιστον (συγκριτικός τελεστής) 2 φορές (συχνότητα) μέσα (συγκριτικός τελεστής) στους 6 μήνες (διάρκεια) πριν (χρονικός δείκτης) την είσοδο στην κλινική δοκιμή (σημείο αναφοράς)». Όπως φαίνεται, και η περίοδος και η συχνότητα που αναφέρονται έχουν χρησιμοποιηθεί σε συνδυασμό με ένα συγκριτικό τελεστή, ενώ μπορούν να συνδυαστούν περαιτέρω με έναν ή περισσότερους λογικούς τελεστές για την έκφραση πιο περίπλοκων περιορισμών. Φυσικά, σε αρκετές

περιπτώσεις, μόνο ένα υποσύνολο των παραπάνω παραμέτρων είναι απαραίτητο, όπως στο κριτήριο «Ο ασθενής έλαβε αμιοδαρόνη πριν (χρονικός δείκτης) την εισαγωγή του στην κλινική δοκιμή (σημείο αναφοράς)».

Σχήμα 6: Κριτήριο σχετικό με τις διαγνώσεις ενός ασθενούς, τυπικά εκφρασμένο σε SPARQL.

Το Σχήμα 6 παρουσιάζει την τυπική αναπαράσταση του προαναφερθέντος κριτηρίου, χρησιμοποιώντας SPARQL, καθώς επίσης και τα 4 διαφορετικά τεχνολογικά στοιχεία που χρησιμοποιήθηκαν κατά την αναπαράσταση. Πιο συγκεκριμένα, η XML (1) χρησιμοποιήθηκε για τον ορισμό των βασικών παραμέτρων ενός κριτηρίου, ενώ η SPARQL (2) σε συνδυασμό με τους όρους από την οντολογία (3) και τις κωδικοποιήσεις που χρησιμοποιηθήκαν (4) για την τυπική αναπαράσταση του κριτηρίου. Τα οντολογικά στοιχεία που χρησιμοποιήθηκαν για την αναπαράσταση του παραπάνω κριτηρίου, καθώς επίσης και η διασύνδεσή του με τα υπάρχοντα συστήματα κωδικοποίησης και ονοματολογίας παρουσιάζονται στο Σχήμα 7.



Σχήμα 7: Τμήμα της Οντολογίας Κριτηρίων Καταλληλότητας για τη SPARQL αναπαράσταση ενός κριτηρίου.

Σημειώνεται ότι η τυπική έκφραση του κριτηρίου χρησιμοποιώντας τη SPARQL προέκυψε αυτόματα με βάση την XML αναπαράσταση του κριτηρίου, χωρίς να υπάρχει απώλεια πληροφορίας. Κατά τη διάρκεια της διαδικασίας αυτής, τα κατάλληλα triple patterns και FILTER clauses ορίστηκαν στο WHERE clause του αυτομάτως παραγόμενου ASK SPARQL ερωτήματος. Γενικά, τα triple patterns προέρχονται από τα στοιχεία του XML σχήματος που ορίσθηκαν με βάση τα στοιχεία που υπήρχαν στο μοντέλο που αναπτύχθηκε. Από την άλλη, οι Boolean εκφράσεις που ορίσθηκαν στα FILTER clauses προέρχονται από τα στοιχεία εκείνα που ορίσαμε στο XML σχήμα, για να εκφράσουμε τις συνθήκες που θα πρέπει να πληρούν οι αντίστοιχες οντότητες.

Θα πρέπει να σημειωθεί ότι υπάρχει ένα σημασιολογικό χάσμα ανάμεσα στην περίοδο για την οποία ένα κριτήριο ορίζεται (π.χ., δεν έλαβε αμιοδαρόνη τους προηγούμενους 6 μήνες) και την αντίστοιχη περιόδο που συσχετίζεται με τα δεδομένα ενός ασθενούς στο μοντέλο μας (π.χ., περίοδος χορήγησης φαρμάκου: 04/03/2014 – 21/03/2014). Επομένως, οι χρονικοί περιορισμοί ορίσθηκαν ξεκάθαρα στην XML αναπαράσταση του κριτηρίου, ενώ στη SPARQL αναπαράσταση του κριτηρίου οι χρονικοί περιορισμοί εκφράστηκαν λαμβάνοντας υπόψη τη σημασία των εκάστοτε οντολογικών στοιχείων. Για παράδειγμα, στο κριτήριο που περιγράφεται στο Σχήμα 4 η ημερομηνία που «ξέσπασε» μια ασθένεια θα πρέπει να βρίσκεται μέσα στις τελευταίες 30 ημέρες, όπως φαίνεται στο Σχήμα 6. Στο σημείο αυτό, θα πρέπει να τονίσουμε ότι ήταν

απαραίτητο να καθορίσουμε την ημερομηνία στην οποία ξεκίνησε η διαδικασία επιλογής ασθενών (screening). Στην εργασία αυτή υποθέσαμε (με τη σύμφωνη γνώμη των ειδικών στον τομέα της κλινικής έρευνας) ότι η διαδικασία αυτή λαμβάνει χώρα τη στιγμή αυτή (παρόν).

5 *Αξιολόγηση Μοντέλων*

5.1 Αξιολόγηση με βάση Υπάρχοντα Κριτήρια Καταλληλότητας

Για την αξιολόγηση της αναπαράστασης των Κριτηρίων Καταλληλότητας (ΚΚ) που αναπτύχθηκε, 200 ΚΚ επιλέχθηκαν τυχαία (διαφορετικά από τα ΚΚ που χρησιμοποιήθηκαν κατά τη φάση της ανάλυσης) και κατόπιν ο βαθμός στον οποίο μπορούσαμε να τα εκφράσουμε με βάση τα μοντέλα που αναπτύχθηκαν εξετάστηκε, σε συνεργασία με τους 8 ειδικούς στον τομέα της κλινικής έρευνας που συμμετείχαν στην παραπάνω διαδικασία. Θα πρέπει να σημειωθεί ότι η επιλογή των κριτηρίων που χρησιμοποιήθηκαν κατά την αξιολόγηση έγινε με βάση την ποικιλία των όρων όχι μόνο σε επίπεδο κατηγορίας αλλά και τύπου και σκοπού. Σύμφωνα με την ανάλυσή μας, το 87% των κριτηρίων που επιλέχθηκαν μπόρεσαν να εκφραστούν πλήρως με βάση το μοντέλο που αναπτύχθηκε (χαρακτηριστικά παραδείγματα υπάρχουν στην υποενότητα που ακολουθεί). Σχετικά με τα κριτήρια που δεν μπορέσαμε να αναπαραστήσουμε, αυτό συνέβη για δύο κύριους λόγους. Κάποια κριτήρια ήταν ασαφή (π.χ., υπερβολική χρήση καπνού) και επομένως η σημασιολογική τους αναπαράσταση ήταν υποκειμενική και ποικίλλουσα. Επίσης, κάποια κριτήρια (π.χ., θύμα ενός στυγερού εγκλήματος) δεν μπορούσαν να εκφραστούν, καθώς δε συμπεριλαμβάνονταν ανάμεσα στις κατηγορίες που ορίσαμε. Αυτό συνέβη, διότι κατά τον ορισμό των κατηγοριών και των σχετικών

παραμέτρων ακολουθήθηκε μια σειρά προτεραιότητας, με βάση τη συχνότητα της χρήσης τους καθώς και τη σημασία τους στο σχεδιασμό μιας κλινικής δοκιμής.

5.1.1 Παραδείγματα Τυπικής Έκφρασης Κριτηρίων Καταλληλότητας

Ο Πίνακας 3 παρουσιάζει ορισμένα παραδείγματα από υπάρχοντα κριτήρια εισαγωγής/εξαγωγής σε μια κλινική δοκιμή και τον τρόπο που μπορούμε αυτά να τα εκφράσουμε, με βάση το μοντέλο που αναπτύχθηκε. Λόγω έλλειψης χώρου, υιοθετήθηκε μια συγκεκριμένη σημειολογία για την τυπική έκφραση των κριτηρίων στην τελευταία στήλη του πίνακα. Πιο συγκεκριμένα, οι συνθήκες που θα πρέπει να πληρούν ορίζονται χρησιμοποιώντας τα στοιχεία του μοντέλου μας. Θα πρέπει να σημειωθεί ότι η συνάρτηση CV επιστρέφει τον κωδικό της φράσης στο σύστημα ονοματολογίας που επιλέχθηκε (π.χ., ChEBI και/ή ATC για τις δραστικές ουσίες, ICD-10 και/ή DSM-IV για τις διαγνώσεις, κτλ.). Επίσης, για λόγους μείωσης της πολυπλοκότητας της εκάστοτε έκφρασης, τα ποσά και οι χρονικές περίοδοι ορίζονται μέσα σε παρενθέσεις, χωρίς να ορίζονται ρητά οι εσωτερικές τους παράμετροι (π.χ., τιμή και μονάδα μέτρησης για ένα ποσό). Τέλος, ο τελεστής IN δείχνει ότι τα δεδομένα θα πρέπει να βρίσκονται μέσα στο σύνολο ή εύρος τιμών που αναφέρεται, ενώ ο τελεστής ALL δείχνει ότι όλα τα στοιχεία του συνόλου θα πρέπει να υπάρχουν.

Μελέτη	**I/E**	**Κριτήριο**	**Τυπική Έκφραση**
NCT 00260481	I	Be at least 18 years-of-age	Person, i.e., - Age-at-screening >= (18 years)
NCT 00855140	I	Ability to speak, read, and write English	Person Language-Ability-at-screening, i.e., - Language-code = CV(English) - Language-Mode-code ALL { CV(speak), CV(read), CV(write) }
NCT 00394771	E	Pregnancy within the last 3 months	Pregnancy-Data associate-with Person, i.e., - Conception-date IN (3 months prior screening)
NCT	E	Active cigarette smoker	Tobacco-Consumption-Data associated-with

Μελέτη	**I/E**	**Κριτήριο**	**Τυπική Έκφραση**
00014040			Person, i.e., - Consumption-Status-code = CV (active)
NCT 01300364	I	Diagnosis of schizophrenia (DSM-IV criteria)	Disorder associated-with Person, i.e., - Disorder-code = CV(schizophrenia)
NCT 01057706	I	At least 12 week duration of neck and back related disability	Disability associated-with Person, i.e., - Disability-code = CV(neck-and-back-related disability) - Disability-duration >= (12 weeks)
NCT 00302419	I	Continuous chest pain that lasted > 30 minutes within the preceding 12 hours	Symptom associated-with Person, i.e., - Symptom-code = CV(chest pain) - Symptom-duration > (30 minutes) - Symptom-date-of-onset IN (12 hours prior screening)
NCT 00005675	E	Allergy to beef or dairy products	Allergy associated-with Person, i.e., - Allergy-code IN { CV(beef), CV(dairy product)}
NCT 00336544	E	Prior hospitalization within previous 4 weeks	Hospitalization Data associated-with Person, i.e., - Hospitalization-start-date IN (4 weeks before study start)
NCT 00390416	E	Patients who have received previous bevacizumab, docetaxel, or cisplatin	Drug associated-with Person, i.e., - Pharmacological-Substance-Code IN { CV(bevacizumab), CV(docetaxel), CV(cisplatin) } - Prescription-start-date IN (before screening)
NCT 00008489	E	Are unable to take medications by mouth.	Medication able-to-receive, i.e., - Route-of-administration = CV(mouth)
NCT 00022672	E	Previous chemotherapy for metastatic disease	Chemotherapy associated-with Person, i.e., - Chemotherapy-reason = CV(metastatic disease) - Chemotherapy-start-date IN (before screening)
NCT 02174653	E	Previous (within the last 3 months prior to visit 1) or concomitant treatment with	Drug associated-with Person, i.e., - Drug-code = CV(coagulation-inhibiting drug) - Drug-end-date IN (3 months prior visit 1)

Μελέτη	**I/E**	**Κριτήριο**	**Τυπική Έκφραση**
		coagulation-inhibiting drugs such as warfarin	Note: Warfarin is also a coagulation-inhibiting drug.
NCT 01831960	E	Subject had major surgery within 30 days prior to study start or plans to have surgery during the study.	Surgery associated-with Person, i.e., - Surgery-date IN { (30 days prior to study start) , study-period }
NCT 00179595	E	Exhaled nitric oxide levels > 60 ppb	Examination associated-with Person, i.e., - Examination-code = CV(Exhaled nitric oxide) - Examination-value > (60 ppb) - Examination-date = Screening Date
NCT 00895934	I	Serum creatinine =< 1.5 x IULN (within 7 days prior to registration)	Laboratory Examination associated-with Person, i.e., - Examination-code = CV(Serum creatinine) - Examination-value <= 1.5 * UNL - Examination-date IN (7 days prior registration)
NCT 00507663	E	Folstein Mini-Mental State Examination Score < 17	Questionnaire associated-with Person, i.e., - Questionnaire-code = CV(Folstein Mini-Mental State Examination) - Questionnaire -value < 17
NCT 01562301	I	Life expectancy of greater than 6 months	Person, i.e., - Life-Expectancy > (6 months)
NCT 00786825	I	Must be able to provide informed consent	Person, i.e., - Able to Provide Informed Consent = CV(yes)
NCT 00027209	E	Subjects with 1 first degree family member with a history of bilateral breast cancer	Family History Data associated-with Person, i.e., - Relative-degree = CV(1^{st} degree) - Medical Condition = CV(bilateral breast cancer)

Πίνακας 3: Τυπική Έκφραση ορισμένων τυχαίως επιλεγμένων Κριτηρίων Καταλληλότητας Κλινικών Δοκιμών

5.2 Σύγκριση με Υπάρχοντα Μοντέλα

Η αναπαράσταση των ΚΚ αξιολογήθηκε επίσης με βάση το βαθμό διασύνδεσης των όρων του μοντέλου που αναπτύχθηκε με υπάρχοντα πρότυπα και προδιαγραφές. Για το σκοπό αυτό, η ανάλυση ολόκληρου του μοντέλου ΚΚ έλαβε χώρα με σκοπό την καταγραφή των όρων που είχαν συσχετιστεί με υπάρχοντες όρους από τα μοντέλα αναφοράς των CDISC, HL7 και OpenEHR. Η ανάλυση έδειξε ότι ένα σημαντικό μέρος του σημασιολογικού μοντέλου είναι διασυνδεδεμένο με διεθνή πρότυπα (κλάσεις: 34%, συσχετίσεις: 68%, παράμετροι: 62%) με το συνολικό ποσοστό των όρων να ξεπερνά το 50%. Επιπρόσθετα, περισσότερες από 50 ονοματολογίες είναι συνδεδεμένες με την οντολογία που αναπτύχθηκε. Επίσης, κανένας από τους όρους που συμπεριλάβαμε στο μοντέλο μας δεν κρίθηκε περιττός από τους ειδικούς. Επιπλέον, ορισμένα οντολογικά στοιχεία επιλέχθηκαν τυχαία και ο ορισμός τους εξετάστηκε από τους ειδικούς στον τομέα της κλινικής έρευνας. Όλα τα στοιχεία που εξετάστηκαν ήταν καλά ορισμένα με καλή χρήση της Αγγλικής γλώσσας. Τέλος, χρησιμοποιώντας τον Pellet Reasoner [64] βεβαιώθηκε ότι δεν υπάρχει καμία αντίφαση ανάμεσα στα στοιχεία που ορίστηκαν.

Το ΚΚ μοντέλο που αναπτύξαμε το συγκρίναμε και με τα υπάρχοντα μοντέλα που περιγράψαμε στα προηγούμενα κεφάλαια. Πιο συγκεκριμένα, σε στενή συνεργασία με τους ειδικούς στον τομέα της κλινικής έρευνας, τα στοιχεία του μοντέλου συγκρίθηκαν με τους όρους που συμπεριλήφθησαν στο EliXR σημασιολογικό δίκτυο. Η ανάλυση έδειξε ότι όλα τα στοιχεία του EliXR σημασιολογικού δικτύου καλύπτονται από τα στοιχεία του μοντέλου που αναπτύξαμε. Όμως, θα πρέπει να τονίσουμε ότι το μοντέλο μας «χτίστηκε» γύρω από την έννοια του Ανθρώπου και τα Δεδομένα που συσχετίζονται με αυτόν, σε αντίθεση με το EliXR σημασιολογικό δίκτυο που περιγράφει τις παραμέτρους που επηρεάζουν το «σύνολο» των Ασθενών. Κατά συνέπεια, σε ορισμένες περιπτώσεις, τα στοιχεία που ορίστηκαν ήταν διαφορετικά, αλλά συχνά αυτά συνδέονταν με άμεσο τρόπο.

Για παράδειγμα, στο μοντέλο μας ορίστηκε η ημερομηνία στην οποία μια ασθένεια (η γενικά διαταραχή) εντοπίστηκε, ενώ στο EliXR σημασιολογικό δίκτυο μια ασθένεια συσχετίζεται με ένα χρονικό περιορισμό. Σχετικά με το μοντέλο που αναπτύχθηκε από τους Milian κ.ά. [39], αυτό ορίζει τα στοιχεία που χρησιμοποιούνται για την περιγραφή των παραμέτρων ενός προτύπου (π.χ., μοτίβο το οποίο περιγράφεται χρησιμοποιώντας τη φυσική γλώσσα) παρά τις εσωτερικές παραμέτρους που χρησιμοποιούνται. Παρόλα αυτά, τα προτεινόμενα πρότυπα καλύπτονται από το μοντέλο που αναπτύχθηκε. Πιο συγκεκριμένα, τα πρότυπα που ορίστηκαν μπόρεσαν να εκφραστούν, χρησιμοποιώντας ένα ή παραπάνω στοιχεία από το ΚΚ μοντέλο σε συνδυασμό με συγκριτικούς και λογικούς τελεστές, καθώς επίσης και όρους από συγκεκριμένες ονοματολογίες.

Η προσέγγιση που ακολουθήθηκε κατά το σχεδιασμό της παραπάνω οντολογίας διευκολύνει την εισαγωγή νέων παραμέτρων που πιθανώς να χρειάζονται για την κάλυψη των αναγκών μιας συγκεκριμένης κλινικής δοκιμής, οι οποίες δεν υπάρχουν στο μοντέλο μας. Επίσης, οι παράμετροι που έχουν οριστεί μπορούν να διασυνδεθούν με παραπάνω από ένα συστήματα κωδικοποίησης (διαφορετικά από αυτά που έχουν ήδη διασυνδεθεί), τα οποία καλύπτουν τις ανάγκες μιας συγκεκριμένης κλινικής δοκιμής και πιθανόν να διαφέρουν στο πεδίο της γνώσης που καλύπτουν (domain coverage) καθώς επίσης και στη διακριτότητα των όρων τους (granularity of terms), ξεπερνώντας τους περιορισμούς που υπήρχαν στη βασισμένη σε μοντέλο προσέγγιση που ακολουθήθηκε στο ASPIRE έργο.

Η βασισμένη σε μοντέλο προσέγγιση που ακολουθήθηκε για την έκφραση των ΚΚ μιας κλινικής δοκιμής επιτρέπει στους χρήστες να χρησιμοποιήσουν παραπάνω από μία παραμέτρους από κάθε σημασιολογική κλάση, καθώς επίσης και να συνδυάσουν μια ή περισσότερες σημασιολογικές κλάσεις για την έκφραση ιδιαίτερα περίπλοκων ΚΚ (π.χ., ασθενείς που διαγνώστηκαν με μία συγκεκριμένη ασθένεια ή το αποτέλεσμα μιας εργαστηριακής εξέτασης ξεπερνά το ανώτερο όριο της φυσιολογικής της τιμής). Κατά

συνέπεια, ο προτεινόμενος τρόπος αναπαράστασης κριτηρίων είναι πολύ πιο εκφραστικός από υπάρχουσες, βασισμένες σε πρότυπα, αναπαραστάσεις, επιτρέποντας την τυπική έκφραση ιδιαίτερα περίπλοκων ΚΚ, τα οποία χρησιμοποιούν παραμέτρους που γενικά δε χρησιμοποιούνται με την ίδια συχνότητα στις κλινικές δοκιμές και κατά συνέπεια δε συμπεριλαμβάνονται στα πρότυπα που έχουν οριστεί από τους Weng κ.ά. [34], Luo κ.ά. [37], Milian κ.ά. [39] ή Bhattacharya και Cantor [40].

Η σελίδα αυτή είναι σκόπιμα λευκή

6 *Εφαρμογή Αποτελεσμάτων*

6.1 Καθορισμός Κριτηρίων Καταλληλότητας

Η προτεινόμενη αναπαράσταση των Κριτηρίων Καταλληλότητας (ΚΚ) επιτρέπει στους χρήστες – ερευνητές να εκφράσουν με ακρίβεια τις συνθήκες που θα πρέπει να πληρούν οι ασθενείς μέσω της χρησιμοποίησης των στοιχείων – όρων που ορίσθηκαν στο μοντέλο μας. Στο έργο PONTE [65], ο καθορισμός των ΚΚ από τους κλινικούς ερευνητές υποστηρίχθηκε μέσω της χρησιμοποίησης μιας σειράς από φόρμες μέσω του γραφικού περιβάλλοντος που αναπτύχθηκε για τη συγγραφή μιας κλινικής δοκιμής [66]. Ο σχεδιασμός των φορμών καθοδηγήθηκε από τις κλάσεις και παραμέτρους που ορίσθηκαν στην ΚΚ οντολογία, επιτρέποντας στους χρήστες να ορίσουν το σύνολο ή εύρος τιμών στις οποίες θα πρέπει να βρίσκονται τα δεδομένα ενός ασθενούς, τα οποία εσωτερικά ορίζονταν μέσω της προτεινόμενης αναπαράστασης (πιο συγκεκριμένα XML). Επίσης, κατασκευάσαμε ένα εργαλείο [67] που θα μπορούσε να χρησιμοποιηθεί, για να εκφράσουμε τις συνθήκες τις οποίες θα πρέπει να πληρούν τα δεδομένα μας και κατόπιν να χρησιμοποιηθούν για την εύρεση των επιθυμητών δεδομένων από μια RDF βάση. Το εργαλείο αυτό βασίζεται αποκλειστικά στο μοντέλο που υποστηρίζεται και στις ονοματολογίες που έχουν χρησιμοποιηθεί (ορίζεται κατά το σχεδιασμό της εφαρμογής) για τον καθορισμό των κριτηρίων, τα οποία εσωτερικά εκφράστηκαν μέσω SPARQL ερωτημάτων που παράγονταν αυτόματα μέσω των δεδομένων που δόθηκαν από το

χρήστη. Η προσέγγιση που ακολουθήθηκε θα μπορούσε επίσης να χρησιμοποιηθεί για τον ορισμό των ΚΚ, επιτρέποντας στους ερευνητές να εκφράσουν ιδιαίτερα περίπλοκα κριτήρια, στα οποία είναι απαραίτητο να συνδυαστούν παράμετροι από 2 ή περισσότερες κλάσεις. Και στις δύο παραπάνω περιπτώσεις, οι ερευνητές θα πρέπει να συμμετέχουν ενεργά στη διαδικασία έκφρασης των κριτηρίων, παρέχοντας όχι μόνο την περιγραφή του κριτηρίου χρησιμοποιώντας τη φυσική γλώσσα αλλά και την τυπική του αναπαράσταση.

Μια πιο ενδιαφέρουσα προσέγγιση θα ήταν να παράγεται αυτόματα η τυπική έκφραση ενός κριτηρίου μέσω της περιγραφής τους στη φυσική γλώσσα. Η προσέγγιση αυτή θα είχε πολλαπλά οφέλη όχι μόνο κατά το σχεδιασμό μιας νέας κλινικής δοκιμής (π.χ., η τυπική έκφραση των κριτηρίων παράγεται αυτόματα, επιτρέποντας στο χρήστη να επιβεβαιώσει ή ακόμη να αλλάξει την προτεινόμενη έκφραση, αν αυτό ήταν απαραίτητο, μειώνοντας σημαντικά τον απαιτούμενο χρόνο και πολυπλοκότητα του γραφικού περιβάλλοντος) αλλά και για υπάρχουσες κλινικές δοκιμές (που πιθανώς να έχουν ολοκληρωθεί) και επομένως, τα κριτήρια είναι διαθέσιμα μόνο ως κείμενο, χωρίς να υπάρχει η δυνατότητα τυπικής έκφρασης των κριτηρίων από τον ερευνητή (είτε γιατί δεν υποστηρίζεται από το εκάστοτε σύστημα είτε γιατί ο σχετικός ερευνητής δεν είναι πλέον διαθέσιμος). Η προσέγγιση αυτή είναι ιδιαίτερα δύσκολη εξαιτίας της πολυπλοκότητας των ΚΚ [41], καθώς επίσης και της πολυσημίας και ασάφειας της φυσικής γλώσσας [68]. Η χρησιμοποίηση ενός υποσυνόλου της φυσικής γλώσσας [69], το οποίο θα μπορούσε να ερμηνευτεί με ακρίβεια με βάση τους όρους που υπάρχουν στο ΚΚ μοντέλο που αναπτύχθηκε, είναι ένα πολύ ενδιαφέρον θέμα προς εξερεύνηση, αλλά είναι έξω από το πεδίο της εργασίας αυτής.

6.2 Χρησιμοποίηση Κριτηρίων Καταλληλόλητας

Η προτεινόμενη αναπαράσταση των ΚΚ μιας κλινικής δοκιμής διευκολύνει επίσης και την επεξεργασία των δεδομένων των κριτηρίων από τον υπολογιστή και τη

χρησιμοποίησή τους για την εύρεση των ασθενών που πληρούν τα κριτήρια. Η γλώσσα XML χρησιμοποιείται ευρέως στη βιομηχανία των πληροφοριακών συστημάτων και επομένως τα ΚΚ που είναι εκφρασμένα με βάση τη γλώσσα αυτή μπορούν να επεξεργαστούν από ένα μεγάλο εύρος συστημάτων. Στο έργο PONTE, η XML αναπαράσταση των ΚΚ χρησιμοποιήθηκε για την εύρεση πιθανών ασυνεπειών ανάμεσα στα κριτήρια που έχουν ορισθεί και στις άλλες παραμέτρους μιας κλινικής δοκιμής, καθώς επίσης και για τον «εμπλουτισμό» των κριτηρίων με βάση τη σημασία των όρων που χρησιμοποιήθηκαν. Το τελευταίο είναι ιδιαίτερα χρήσιμο για την εύρεση των κατάλληλων ασθενών, καθώς στον ηλεκτρονικό τους φάκελο θα βρούμε, συνήθως, τα συγκεκριμένα προβλήματα με τα οποία διαγνώσθηκε ο ασθενής (π.χ., ανεύρυσμα), ενώ στα ΚΚ, συνήθως, ορίζουμε την ευρύτερη κατηγορία στην οποία αυτά ανήκουν (π.χ., Ισχαιμική Καρδιακή Πάθηση). Επίσης, πρέπει να σημειωθεί ότι η μετάβαση από την XML αναπαράσταση ενός ΚΚ σε SPARQL μπορεί να γίνει με κάποιο αυτόματο τρόπο (όπως ήδη αναφέρθηκε), καθώς οι όροι που χρησιμοποιήθηκαν στην XML αναπαράσταση ορίσθηκαν με βάση τους όρους που έχουν ορισθεί στο ΚΚ μοντέλο. Η SPARQL αναπαράσταση των ΚΚ διευκολύνει τη μετατροπή τους σε ερωτήματα προς μία βάση δεδομένων, με σκοπό τον εντοπισμό των ασθενών που πιθανώς να είναι κατάλληλοι να συμμετέχουν στην κλινική δοκιμή [70]. Επίσης, η SPARQL επιτρέπει τον αποτελεσματικό χειρισμό της απουσίας δεδομένων, χρησιμοποιώντας τον OPTIONAL τελεστή [71].

Επιπλέον, στο έργο PONTE αναπτύχθηκαν εργαλεία για τη διασύνδεση των ΚΚ με τη βάση των ασθενών. Πιο συγκεκριμένα, το OAT [72] επιτρέπει στους χρήστες να ορίσουν με ακρίβεια τη συσχέτιση ανάμεσα στους όρους των μοντέλων και ονοματολογιών, που χρησιμοποιούνται μέσω του γραφικού περιβάλλοντος που αναπτύχθηκε. Επίσης, ένα άλλο εργαλείο / σύστημα / μηχανισμός [73] επιτρέπει τη

«μετάφραση» των ερωτημάτων SPARQL (αρχικά εκφρασμένων μέσω του ΚΚ μοντέλου και των ονοματολογιών που επιλέχθηκαν) στα αντίστοιχα SPARQL ερωτήματα, με βάση τις συσχετίσεις που έχουν ορισθεί με τα μοντέλα και κωδικοποιήσεις που χρησιμοποιούνται στη βάση. Ο μηχανισμός αυτός σε συνδυασμό με κάποιο SPARQL τελικό σημείο (endpoint), όπως D2R εξυπηρετητής [74] και Ontop Framework [75], επιτρέπει την εκτέλεση των ερωτημάτων σε μία σχεσιακή βάση, η οποία, συνήθως, χρησιμοποιείται για την καταγραφή των δεδομένων των ασθενών. Παρ' όλα αυτά, σε ορισμένες περιπτώσεις, δεν επιτρέπεται άμεση πρόσβαση στη βάση μέσω της SQL (π.χ., μέσω HL7 ή συγκεκριμένων προς το χρήστη XML μηνυμάτων). Σε αυτές τις περιπτώσεις, πρέπει να ακολουθηθεί μια διαφορετική προσέγγιση.

Θα πρέπει να σημειωθεί ότι σε ορισμένες περιπτώσεις η σημασία των ΚΚ που έχουν οριστεί μπορεί λίγο να μεταβληθεί, όταν τα κριτήρια εφαρμόζονται σε βάσεις ασθενών. Για παράδειγμα, οι διαφορετικές ονοματολογίες, που πιθανόν να χρησιμοποιούνται κατά την έκφραση των ΚΚ καθώς και κατά την καταγραφή των δεδομένων των ασθενών στις βάσεις, απαιτούν τη μετάβαση από τη μία κωδικοποίηση στην άλλη για τον έλεγχο των συνθηκών που έχουν οριστεί. Επίσης, οι παράμετροι που χρησιμοποιούνται κατά την έκφραση των ΚΚ (π.χ., προσδόκιμο ζωής) μπορεί να μη συνδέονται άμεσα με τις παραμέτρους που υπάρχουν στις βάσεις των ασθενών. Στις περιπτώσεις αυτές, μετάφραση (ή επανέκφραση) των κριτηρίων αυτών είναι απαραίτητη, προκειμένου αυτά να χρησιμοποιηθούν για την επιλογή των ασθενών. Και στις δύο παραπάνω περιπτώσεις, οποιαδήποτε αλλαγή στη σημασία των κριτηρίων που έχουν οριστεί θα πρέπει να καταγράφεται από το αντίστοιχο σύστημα και να ληφθεί σοβαρά υπόψη από τους ειδικούς στον τομέα αυτό κατά την ανάλυση των αποτελεσμάτων και την επιλογή των ασθενών. Το σχήμα που έχει αναπτυχθεί επιτρέπει στους χρήστες να

καταγράψουν τα αρχικά αλλά και μεταφρασμένα κριτήρια, καθώς επίσης και τις αποφάσεις που πάρθηκαν ή τις υποθέσεις που έγιναν κατά την παραπάνω διαδικασία.

Η σελίδα αυτή είναι σκόπιμα λευκή

7 *Συμπέρασμα*

Τα Κριτήρια Καταλληλότητας (ΚΚ) αποτελούν ένα σημαντικό τμήμα μιας κλινικής δοκιμής, καθώς ορίζουν τα χαρακτηριστικά που θα πρέπει να έχουν οι ασθενείς ή γενικότερα τις συνθήκες που θα πρέπει αυτοί να πληρούν, για να λάβουν μέρος στην κλινική δοκιμή. Παρά το γεγονός ότι έχουν δημοσιευτεί αρκετά πρότυπα για την καταγραφή των δεδομένων της κλινικής έρευνας και περίθαλψης από τους οργανισμούς CDISC και HL7 (μεταξύ άλλων), καθώς επίσης και συγκεκριμένες γλώσσες αναπαράστασης και έκφρασης των ΚΚ, δεν υπάρχει κάποια γλώσσα ή γενικά αναπαράσταση που χρησιμοποιείται ευρέως για την τυπική έκφραση των ΚΚ. Στην εργασία αυτή, παρουσιάστηκε ένας καινοτόμος τρόπος για την αναπαράσταση και έκφραση των ΚΚ μιας κλινικής δοκιμής. Η αξιολόγηση που πραγματοποιήθηκε χρησιμοποιώντας έναν περιορισμένο αριθμό τυχαίως επιλεγμένων ΚΚ έδειξε ότι η αναπαράσταση αυτή μπορεί να καλύψει ικανοποιητικά το σκοπό για τον οποίο αναπτύχθηκε. Επίσης, μπορεί να συνδυαστεί με υπάρχοντα εργαλεία και μηχανισμούς που αναπτύχθηκαν για την εύρεση των ασθενών. Επομένως, η προτεινόμενη αναπαράσταση των ΚΚ μπορεί να συμβάλλει σημαντικά στη διαδικασία εύρεσης των ασθενών για ένα μεγάλο εύρος κλινικών δοκιμών, μειώνοντας το χρόνο που απαιτείται για τη στρατολόγηση των ασθενών και, κατά συνέπεια, και το συνολικό κόστος της κλινικής δοκιμής.

Η σελίδα αυτή είναι σκόπιμα λευκή

Β. ΠΡΟΣΒΑΣΗ ΣΤΑ ΔΕΔΟΜΕΝΑ ΣΧΕΣΙΑΚΩΝ ΒΑΣΕΩΝ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΤΟΥ ΣΗΜΑΣΙΟΛΟΓΙΚΟΥ ΙΣΤΟΥ

Στην ενότητα αυτή περιγράφεται η προσέγγιση που ακολουθήθηκε και τα εργαλεία-μηχανισμοί που αναπτύχθηκαν για την πρόσβαση στα δεδομένα μιας ή περισσότερων σχεσιακών βάσεων, χρησιμοποιώντας τις τεχνολογίες του σημασιολογικού ιστού.

Το σύστημα που αναπτύχθηκε βασίζεται στον καθορισμό της συσχέτισης μεταξύ των μοντέλων του χρήστη και των βάσεων δεδομένων, καθώς επίσης και τις ονοματολογίες και κωδικοποιήσεις που υποστηρίζονται. Ειδικότερα, το σύστημα αναλαμβάνει τη μετάφραση των ερωτημάτων του χρήστη στα αντίστοιχα ερωτήματα της βάσης, λαμβάνοντας υπόψη τις συσχετίσεις μεταξύ των όρων των μοντέλων και κωδικοποιήσεων που έχουν καθοριστεί. Επίσης, αναλαμβάνει την περαιτέρω επεξεργασία των δεδομένων που λαμβάνονται από τη βάση έτσι, ώστε αυτά να είναι εκφρασμένα με βάση το μοντέλο και τις κωδικοποιήσεις που χρησιμοποιούνται από τη μεριά του χρήστη.

Στα πλαίσια της εργασίας αυτής, αρχικά, μελετήθηκαν τα υπάρχοντα συστήματα καθορισμού συσχέτισης μεταξύ οντολογιών καθώς και οι υπάρχοντες αλγόριθμοι μετάφρασης ερωτημάτων, τα οποία μπορούν μερικώς να γεφυρώσουν το χάσμα που υπάρχει μεταξύ του χρήστη και της βάσης δεδομένων, λόγω της πληθώρας των δομικών και σημασιολογικών διαφορών που υπάρχουν μεταξύ τους. Για το σκοπό αυτό αναπτύχθηκαν καινοτόμα εργαλεία και μηχανισμοί τόσο για τον καθορισμό της συσχέτισης μεταξύ των όρων δύο οντολογιών όσο και για τη χρησιμοποίηση της συσχέτισης αυτής για την αποτίμηση των ερωτημάτων του χρήστη.

Τα εργαλεία και οι μηχανισμοί που υλοποιήθηκαν χρησιμοποιήθηκαν στο Ευρωπαϊκό έργο ΡΟΝΤΕ για την εύρεση των ασθενών που πληρούν τα Κριτήρια

Καταλληλότητας (ΚΚ) μιας κλινικής δοκιμής, τα οποία ήταν εκφρασμένα με βάση την οντολογία που αναπτύχθηκε για το σκοπό αυτό. Όπως φαίνεται και σε ένα παράδειγμα που υπάρχει στην ενότητα αυτή, η προσέγγιση που ακολουθήθηκε μπορεί να καλύψει ένα μεγάλο εύρος αναντιστοιχιών που μπορεί να παρουσιαστούν κατά την εφαρμογή των ΚΚ στις βάσεις των ασθενών. Ωστόσο, στη γενική περίπτωση, η αποτίμηση των ερωτημάτων του χρήστη, τα οποία είναι εκφρασμένα με διαφορετικά μοντέλα και κωδικοποιήσεις από αυτά που υποστηρίζονται από τη βάση δεδομένων, μπορεί να είναι ιδιαίτερα δύσκολη, αν όχι αδύνατη, ειδικά όταν υπάρχουν σημαντικές δομικές και σημασιολογικές διαφορές, εξαιτίας της περιορισμένης ή καθόλου χρήσης διεθνών προτύπων οργάνωσης και αναπαράστασης της πληροφορίας.

8 *Εισαγωγή*

Οι οντολογίες (ontologies) [76] έχουν ένα διακριτό ρόλο στο σημασιολογικό ιστό (semantic web). Επιτρέπουν στους χρήστες να εκφράσουν τυπικά την πληροφορία ενός συγκεκριμένου πεδίου γνώσης, ενώ μπορούν να μοιραστούν τα δεδομένα τους με τα άλλα μέλη της κοινότητας με τη μορφή των OWL και RDF αρχείων. Όμως, συχνά παρουσιάζουν σημαντικές δομικές και σημασιολογικές διαφορές μεταξύ τους [78], καθώς έχουν αναπτυχθεί ανεξάρτητα η μία από την άλλη και ορισμένες φορές εξυπηρετούν διαφορετικούς σκοπούς [77], εμποδίζοντας την ομαλή επικοινωνία μεταξύ των σχετικών συστημάτων. Επιπρόσθετα, όταν το μέγεθός τους αυξάνει σημαντικά, καταναλώνουν αρκετούς υπολογιστικούς πόρους, ενώ βασικές λειτουργίες, όπως η αναζήτηση δεδομένων, απαιτούν αρκετό χρόνο για την εκτέλεσή τους.

Οι σχεσιακές βάσεις δεδομένων (relational databases) αποτελούν μία δοκιμασμένη και αξιόπιστη λύση για την αποθήκευση δεδομένων, ενώ επιτρέπουν την αποτελεσματική χρήση αυτών, ακόμη και όταν πρέπει να χειριστούμε εκατομμύρια ή δισεκατομμύρια από εγγραφές. Συνεπώς, αρκετά συχνά, η πληροφορία για τις οντότητες ενός οργανισμού καταλήγει σε μία σχεσιακή βάση δεδομένων, ενώ πρόσβαση στα δεδομένα συνήθως παρέχεται μέσω μιας διεπαφής που έχει σχεδιαστεί αποκλειστικά, για να ικανοποιεί τις ανάγκες του εκάστοτε οργανισμού.

Η πρόοδος που έχει υπάρξει τα τελευταία χρόνια στον τομέα του σημασιολογικού ιστού και ειδικότερα η ανάπτυξη εργαλείων, όπως ο D2R εξυπηρετητής [74] και το OnTop σύστημα [75], επιτρέπουν στους χρήστες να έχουν πρόσβαση στα δεδομένα μιας σχεσιακής βάσης χρησιμοποιώντας SPARQL ερωτήματα. Επιπρόσθετα, παρέχουν μια απλή διαδικτυακή διεπαφή για την εξερεύνηση των στοιχείων της βάσης δεδομένων ακολουθώντας τους συνδέσμους που παρέχονται. Όμως, τα μοντέλα που υποστηρίζονται είναι αρκετά κοντά στη δομή της βάσης δεδομένων και τους όρους που χρησιμοποιούνται, ενώ αρκετές φορές δε συνοδεύονται από ικανοποιητική περιγραφή των οντολογικών τους στοιχείων, ιδιαίτερα στις περιπτώσεις εκείνες που το μοντέλο παράγεται αυτόματα με βάση το σχήμα της βάσης.

Ο σχεδιασμός μιας άλλης οντολογίας που παρέχει μια πραγματική εννοιολογική περιγραφή ενός πεδίου γνώσης μπορεί να βελτιώσει σημαντικά την επικοινωνία των χρηστών με τις βάσεις, παρέχοντας όλα τα δεδομένα που χρειάζεται να γνωρίζουν για την συγγραφή των SPARQL ερωτημάτων, συμπεριλαμβανομένης της δομής της βάσης, και τους όρους που χρησιμοποιούνται για την καταγραφή των δεδομένων. Επίσης, για να είναι εφικτή η αποτίμηση των ερωτημάτων των χρηστών, είναι απαραίτητος ο καθορισμός της συσχέτισης μεταξύ των οντολογικών στοιχείων που υπάρχουν στις δύο πλευρές, ώστε να μπορούν έπειτα να χρησιμοποιηθούν για την αναδιατύπωση ή μετάφραση των ερωτημάτων και ενδεχομένως αποτελεσμάτων, χρησιμοποιώντας τους όρους της οντολογίας που υποστηρίζονται στην καθεμία από τις δύο πλευρές του συστήματος.

Ο καθορισμός της συσχέτισης μεταξύ δύο οντολογιών έχει μελετηθεί εκτενώς στη βιβλιογραφία και αρκετά συστήματα και αλγόριθμοι έχουν προταθεί για την ανίχνευση των πιθανών συσχετίσεων καθώς και τον ορισμό νέων. Όμως, τα συστήματα αυτά δεν μπορούν να χειριστούν αποτελεσματικά τις περίπλοκες εκείνες αναντιστοιχίες στις οποίες συμμετέχουν παραπάνω από ένα οντολογικά στοιχεία στο αριστερό και δεξιό μέρος ενός

κανόνα σε συνδυασμό με μία συνάρτηση μετατροπής τιμών. Ο καθορισμός των παραπάνω συσχετίσεων είναι αρκετά συχνό φαινόμενο, ιδιαίτερα στις περιπτώσεις εκείνες που είναι απαραίτητη η επικοινωνία με δύο ή περισσότερες βάσεις δεδομένων (π.χ., βάσεις ασθενών) μέσω μιας «κοινής» γλώσσας. Δεδομένου ότι οι βάσεις αυτές έχουν σχεδιαστεί ανεξάρτητα η μία από την άλλη, παρουσιάζουν σημαντικές διαφορές μεταξύ τους (datasource heterogeneity – παράρτημα 30.2) αφενός στην δομή που χρησιμοποιείται για την καταγραφή των δεδομένων και αφετέρου στις ονοματολογίες ή κωδικοποιήσεις που έχουν επιλεχθεί για την κάλυψη των αναγκών του εκάστοτε οργανισμού. Κατά συνέπεια, οι οντολογίες που υποστηρίζονται από τα SPARQL endpoints έχουν σημαντικές αναντιστοιχίες (ontology mismatches) μεταξύ τους, γεγονός που συμβάλει στην ύπαρξη σημαντικών δομικών και σημασιολογικών διαφορών με το «κοινό» μοντέλο και τις ονοματολογίες/κωδικοποιήσεις που επιλέχθηκαν, ιδιαίτερα όταν ο σχεδιασμός και η επιλογή τους έχει γίνει ανεξάρτητα από τα αντίστοιχα μοντέλα και κωδικοποιήσεις που υποστηρίζονται από τις βάσεις.

Ο καθορισμός και έκφραση των παραπάνω περίπλοκων συσχετίσεων (στη γενική τους περίπτωση), καθώς επίσης και η χρησιμοποίησή τους για την απάντηση ερωτημάτων είναι δύο ιδιαίτερα σημαντικά θέματα για την κοινότητα του σημασιολογικού ιστού. Επίσης, ο σχεδιασμός ενός γραφικού περιβάλλοντος που επιτρέπει στους χρήστες του διαδικτύου να εκφράσουν SPARQL ερωτήματα χωρίς να είναι γνώστες των τεχνολογιών του σημασιολογικού ιστού ή γενικότερα του υπολογιστή, είναι ένα επίσης σημαντικό θέμα, δεδομένου ότι τα δομημένα δεδομένα που δημοσιεύονται συνεχώς αυξάνονται.

Η ενότητα αυτή είναι οργανωμένη ως εξής. Αρχικά στο κεφάλαιο 9 παρουσιάζουμε σχετική βιβλιογραφία στον τομέα του σημασιολογικού ιστού. Στο κεφάλαιο 10 ακολουθεί η περιγραφή του οικοσυστήματος που αναπτύχθηκε για την επικοινωνία με μία ή παραπάνω σχεσιακές βάσεις δεδομένων. Στο κεφάλαιο 11

παρουσιάζουμε το εργαλείο που αναπτύχθηκε για τον καθορισμό των συσχετίσεων μεταξύ των όρων δύο διαφορετικών μοντέλων, ενώ στο κεφάλαιο 12 περιγράφεται λεπτομερώς ο μηχανισμός που υλοποιήθηκε για την αναδιατύπωση των SPARQL ερωτημάτων καθώς και των RDF δεδομένων (εάν αυτό είναι απαραίτητο) μέσω των συσχετίσεων που έχουν καθοριστεί. Στο κεφάλαιο 13 χρησιμοποιήσαμε τα εργαλεία και τους μηχανισμούς που αναπτύχθηκαν για την εύρεση των ασθενών που πληρούν τις επιθυμητές συνθήκες, με βάση τα δεδομένα που έχουν καταγραφεί στη βάση μιας μονάδας περίθαλψης. Στο κεφάλαιο 14 υπάρχει μια συζήτηση σχετικά με την προσέγγιση που ακολουθήθηκε και τα εργαλεία που αναπτύχθηκαν. Τέλος, στο κεφάλαιο 15 συνοψίζουμε τα κύρια σημεία της ενότητας αυτής.

9 *Σχετική Εργασία*

9.1 Εργαλεία, Μηχανισμοί και Γλώσσες Καθορισμού Συσχέτισης μεταξύ Οντολογιών

Για τον ορισμό της συσχέτισης μεταξύ των όρων δύο οντολογιών υπάρχουν αρκετά διαθέσιμα εργαλεία. Το σύστημα SAMBO [79] βοηθάει στη συσχέτιση και συγχώνευση μεταξύ δύο οντολογιών. Παρέχει ένα γραφικό περιβάλλον για την αλληλεπίδραση με τους χρήστες και χρησιμοποιεί μια πληθώρα από τεχνικές για την εύρεση συσχετίσεων, συμπεριλαμβανομένων του ονόματος των όρων (terminological matchers), της δομής της οντολογίας (structural matchers), καθώς και της υπάρχουσας γνώσης, όπως UMLS, και δεδομένων που ελήφθησαν από τη βιβλιογραφία. Οι 1:1 συσχετίσεις που εντοπίζονται παρουσιάζονται στο χρήστη, ο οποίος είναι υπεύθυνος για τον τελικό καθορισμό συσχέτισης. Το Falcon-AO [80] είναι ένα άλλο σύστημα για τον καθορισμό της συσχέτισης μεταξύ δύο οντολογιών. Το εργαλείο αυτό λαμβάνει υπόψη τη δομή των οντολογιών και χρησιμοποιεί γλωσσολογικές τεχνικές (linguistic matchers) για την εύρεση πιθανών 1:1 συσχετίσεων. Επίσης, η μέθοδος διαίρει και βασίλευε (divide and conquer), που έχει υιοθετηθεί, επιτρέπει την εύρεση συσχετίσεων μεταξύ μεγάλων σε αριθμό όρων οντολογιών. Το γραφικό περιβάλλον που παρέχεται επιτρέπει στο χρήστη να ορίσει τις βασικές παραμέτρους του συστήματος και να χειριστεί τις προτεινόμενες συσχετίσεις.

Το OPTIMA [81] είναι ένα εργαλείο γενικού σκοπού για την εύρεση συσχετίσεων. Το εργαλείο αυτό παρέχει ένα γραφικό περιβάλλον για την απεικόνιση και ανάλυση των οντολογιών, ενώ χρησιμοποιεί τη δομή και τα ονόματα των όρων ενός σχήματος ή μοντέλου, για να εντοπίσει πιθανές συσχετίσεις, τις οποίες μπορούμε κατόπιν να αποθηκεύσουμε σε ένα XML αρχείο. Το COMA 3.0 [82] είναι ένα άλλο εργαλείο για την εύρεση συσχετίσεων μεταξύ σχημάτων και οντολογιών. Επιτρέπει στο χρήστη να εισάγει τα σχήματα ή οντολογίες μέσω του γραφικού περιβάλλοντος που παρέχεται, ενώ χρησιμοποιεί γλωσσολογικές και δομικές τεχνικές για τον εντοπισμό πιθανών αντιστοιχίσεων. Ο χρήστης μπορεί να συμμετέχει στην παραπάνω διαδικασία και να ορίσει τις τεχνικές που θα χρησιμοποιηθούν (εναλλακτικά υπάρχουν κάποιες προεπιλεγμένες τεχνικές) για την εύρεση συνωνύμων. Επίσης, ο χρήστης μπορεί μόνος του να ορίσει πιο περίπλοκες συσχετίσεις, στις οποίες συμμετέχουν δύο ή περισσότερα στοιχεία και απαιτείται μετατροπή στα δεδομένα τους.

Το AgreementMaker [83] είναι ένα άλλο εργαλείο για τον καθορισμό συσχέτισης μεταξύ μεγάλων οντολογιών. Διαθέτει ένα γραφικό περιβάλλον μέσω του οποίου ο χρήστης μπορεί να εξετάσει τις οντολογίες που παρουσιάζονται με τη μορφή ενός δένδρου, καθώς επίσης και τις προτεινόμενες συσχετίσεις μεταξύ των όρων. Χρησιμοποιεί μια πληθώρα από τεχνικές, ενώ επίσης επιτρέπει στο χρήστη να ορίσει μόνος του συσχετίσεις που δεν εντοπίστηκαν.

Τα υπάρχοντα εργαλεία εστιάζουν στον εντοπισμό 1:1 συσχετίσεων, ενώ τα περισσότερα απ’ αυτά είτε παρέχουν ένα πολύ απλό γραφικό περιβάλλον είτε δε διαθέτουν καθόλου [84]. Κατά συνέπεια, η αλληλεπίδραση με το χρήστη για τη διαχείριση των προτεινόμενων συσχετίσεων ή για τον καθορισμό νέων είναι δύσκολη. Ειδικότερα, όταν χρειάζεται να καθορίσουμε n:m συσχετίσεις μεταξύ των οντοτήτων, τα

περισσότερα (αν όχι όλα) τα εργαλεία δεν μπορούν να τις χειριστούν ικανοποιητικά τις παραπάνω περιπτώσεις μέσω ενός φιλικού προς το χρήστη γραφικού περιβάλλοντος.

Τα Μοτίβα Αντιστοίχισης (correspondence patterns) [85] και τα Οντολογικά Πρότυπα (ontology patterns) [86] παρέχουν μια ενδιαφέρουσα προσέγγιση για τον καθορισμό συσχετίσεων μεταξύ των οντολογιών. Στην πραγματικότητα, τα Οντολογικά Πρότυπα ξεπερνούν κάποιους από τους περιορισμούς που έχουν τα Μοτίβα Αντιστοίχισης, δίνοντας τη δυνατότητα στους χρήστες να καθορίσουν τα στοιχεία που συμμετέχουν σε ένα κανόνα στην αριστερή και δεξιά του μεριά, ανεξάρτητα το ένα από το άλλο, χρησιμοποιώντας ένα ή περισσότερα μοτίβα. Επίσης, ικανοποιούν μια σειρά από ανάγκες [87], συμπεριλαμβανομένης της εκφραστικότητας, καθορισμού συσχέτισης υπό συνθήκη, συνδυασμού δηλωτικού και διαδικαστικού μέρους, ενώ οι συσχετίσεις που έχουν οριστεί μπορούν να εκφραστούν μέσω της EDOAL [88].

Η γλώσσα EDOAL επεκτείνει το Alignment API [89] ξεπερνώντας τους περιορισμούς του, παρέχοντας μια ιδιαίτερα ευέλικτη γλώσσα για τον αποτελεσματικό χειρισμό πολύπλοκων αναντιστοιχιών που άλλες γλώσσες καθορισμού συσχέτισης, όπως ή χρησιμοποίηση των κατασκευαστών της OWL καθώς επίσης και Context OWL [90], δεν μπορούν να χειριστούν. Επίσης, ο καθορισμός συσχέτισης μεταξύ μοντέλων ή οντολογιών με τη μορφή ερωτημάτων έχει καταγραφεί στη βιβλιογραφία [91]. Η τελευταία επιτρέπει στο χρήστη να ορίσει 1:1 αντιστοιχίες καθώς επίσης και πιο περίπλοκες συσχετίσεις, χρησιμοποιώντας είτε GAV είτε LAV προσεγγίσεις είτε ένα συνδυασμό αυτών (GLAV) [92].

Σχετικά με τη διαδικασία που ακολουθείται για τον εντοπισμό πιθανών συσχετίσεων υπάρχουν αρκετές τεχνικές [93], όπως έχουμε ήδη αναφέρει. Πιο συγκεκριμένα, για τον εντοπισμό της ομοιότητας μεταξύ δύο οντοτήτων μπορούμε να χρησιμοποιήσουμε τεχνικές βασισμένες αποκλειστικά στις συμβολοακολουθίες (strings)

που χρησιμοποιούνται (π.χ. ο Levenshtein [94] edit distance αλγόριθμος), καθώς επίσης και τεχνικές βασισμένες στη γλώσσα αυτή καθ' αυτή. Επίσης, άλλες τεχνικές λαμβάνουν υπόψη τη δομή της οντολογίας και τα αξιώματα που έχουν ορισθεί (π.χ., το πεδίο τιμών, ιεραρχία όρων, κτλ.), επιπρόσθετη γνώση (π.χ., ένα λεξιλόγιο με συνώνυμα) ή ακόμη και μη δομημένα δεδομένα που είναι διαθέσιμα στη βιβλιογραφία.

Τα περισσότερα από τα εργαλεία χρησιμοποιούν ένα συνδυασμό των παραπάνω αλγορίθμων και τεχνικών για τον εντοπισμό πιθανής ομοιότητας μεταξύ των οντολογικών στοιχείων. Όμως, η χρησιμοποίηση των παραπάνω τεχνικών στον εντοπισμό πιο περίπλοκων συσχετίσεων μεταξύ των όρων δύο οντολογιών είναι μια πρόκληση.

9.2 Μηχανισμοί Αναδιατύπωσης Ερωτημάτων – Αποτελεσμάτων

Ο ορισμός της συσχέτισης (correspondence specification) ανάμεσα στους όρους δύο οντολογιών καθώς και η εφαρμογή τους (correspondence consumption) για την υλοποίηση σχετικών εργασιών (π.χ., μετάφραση ερωτημάτων και αποτελεσμάτων, κτλ.) είναι δύο συμπληρωματικές διαδικασίες. Όμως, για να επιτρέψουμε την άμεση χρησιμοποίηση των κανόνων συσχέτισης (mapping rules) που έχουν ορισθεί, τα εργαλεία/συστήματα που χρησιμοποιούνται για την υλοποίηση των δύο παραπάνω διεργασιών θα πρέπει να χρησιμοποιούν την ίδια γλώσσα καθορισμού συσχέτισης μεταξύ οντολογιών (mapping language).

Οι αλγόριθμοι που έχουν υλοποιηθεί από σχετικά συστήματα για τη χρησιμοποίηση των κανόνων αντιστοίχισης που έχουν οριστεί για την υλοποίηση ενός συγκεκριμένου σκοπού (π.χ., απάντηση ερωτημάτων) είναι στενά συνδεδεμένοι με τη γλώσσα συσχέτισης που έχει χρησιμοποιηθεί. Όταν έχουμε να αντιμετωπίσουμε μόνο σημασιολογικές αναντιστοιχίες, οι αντίστοιχοι μηχανισμοί είναι αρκετά απλοί (π.χ. αναλαμβάνουν να αντικαταστήσουν κάθε όρο με το σημασιολογικά αντίστοιχο όρο), όπως ο αλγόριθμος που περιγράφεται στην εργασία [95]. Για τον αποτελεσματικό

χειρισμό πιο περίπλοκων συσχετίσεων, απαιτείται αφενός η χρησιμοποίηση μιας πιο εκφραστικής γλώσσας και αφετέρου ο σχεδιασμός και εφαρμογή πιο περίπλοκων μηχανισμών αναδιατύπωσης ή μετάφρασης.

Μία πιο ενδιαφέρουσα προσέγγιση έχει προταθεί από τους Correndo κ.ά. [96]. Στην προσέγγιση αυτή οι συγγραφείς καθόρισαν τη συσχέτιση μεταξύ των όρων των οντολογιών με τέτοιο τρόπο, ώστε να μπορεί να χρησιμοποιηθεί αποτελεσματικά για την αναδιατύπωση των SPARQL ερωτημάτων. Πιο συγκεκριμένα, το αρχείο που περιέχει τους κανόνες συσχέτισης αποτελείται από μία λίστα με ευθυγραμμίσεις οντολογικών στοιχείων (entity alignment), καθένα απ' τα οποία ορίζει τον τρόπο με τον οποίο θα πρέπει να αναγραφεί μία τριάδα, για να είναι συμβατή με τους όρους της άλλης οντολογίας. Οι παραπάνω αντιστοιχίες χρησιμοποιούνται κατά τη διαδικασία αναδιατύπωσης ενός ερωτήματος, αντικαθιστώντας τις τριάδες του αρχικού SPARQL ερωτήματος με νέες που περιέχουν τους όρους της άλλης οντολογίας. Όμως, το γεγονός ότι ο μηχανισμός αναλαμβάνει την αναδιατύπωση μιας μόνο τριάδας κάθε φορά περιορίζει το είδος των αντιστοιχίσεων που μπορούμε να χειριστούμε αποτελεσματικά.

Για την αναδιατύπωση ερωτημάτων το Healthcare and Life Sciences (HCLS) semantic web interest group [97] βασίστηκε σε συσχετίσεις μεταξύ των οντολογικών στοιχείων εκφρασμένες με τη μορφή των Notation3 (N3) κανόνων [98], τους οποίους κατόπιν χρησιμοποίησαν για την αναδιατύπωση των ερωτημάτων μεταξύ της Οντολογίας Κλινικών Δοκιμών (Clinical Trials Ontology) και της Οντολογίας Κλινικής Πρακτικής (Clinical Practice Ontology). Η προσέγγιση αυτή επιτρέπει μεγαλύτερη ευελιξία εν συγκρίσει με την προηγούμενη προσέγγιση. Ωστόσο, βαρύτητα δόθηκε στην αντιμετώπιση ενός περιορισμένου εύρους αναντιστοιχίσεων.

Και στις δύο παραπάνω περιπτώσεις η συσχέτιση ανάμεσα στους όρους δύο οντολογιών ορίστηκε με τη μορφή των τριάδων, οι οποίες μπορούν άμεσα να

χρησιμοποιηθούν κατά τη μετάφραση ενός ερωτήματος χρησιμοποιώντας τους όρους μιας νέας οντολογίας. Όμως, στην περίπτωση αυτή, οι κανόνες συσχέτισης που έχουν ορισθεί δεν μπορούν να υποστηρίξουν σχετικές εργασίες, όπως η διαδικασία ενοποίησης δύο οντολογιών.

Επίσης, οι συγγραφείς Euzenat κ.ά. [99] χρησιμοποίησαν CONSTRUCT SPARQL ερωτήματα κατά τη μετάφραση των RDF δεδομένων. Και στην περίπτωση αυτή ο ορισμός των συσχετίσεων εξυπηρετεί την υλοποίηση ενός συγκεκριμένου σκοπού. Επιπρόσθετα, η χρησιμοποίηση από το χρήστη, ορισμένων συναρτήσεων κατά την έκφραση της συσχέτισης με την μορφή της SPARQL υποδηλώνει ότι χρησιμοποιείται μια «επέκταση» της SPARQL, η οποία μπορεί να μην υποστηρίζεται από το SPARQL endpoint.

Το πρόβλημα αναδιατύπωσης SPARQL ερωτημάτων εξακολουθεί να αποτελεί ένα ανοιχτό θέμα, καθώς οι υπάρχουσες προσεγγίσεις δεν μπορούν να χειριστούν αποτελεσματικά όλες τις αναντιστοιχίες που πιθανώς να προκύψουν και επομένως δεν μπορούν να εφαρμοστούν σε ένα πραγματικό σενάριο. Επίσης, παρά το γεγονός ότι η αναδιατύπωση των δεδομένων που λαμβάνουμε από τη βάση δεδομένων σχετίζεται άμεσα με την αναδιατύπωση των ερωτημάτων που υποβάλλουμε, ιδιαίτερα στις περιπτώσεις εκείνες στις οποίες οι όροι που χρησιμοποιούνται στις δύο πλευρές είναι διαφορετικοί, τα δύο παραπάνω «προβλήματα» δεν έχουν επαρκώς μελετηθεί στο σύνολό τους. Πιο συγκεκριμένα, οι υπάρχουσες προσεγγίσεις βασίζονται στον καθορισμό της συσχέτισης μεταξύ των οντολογιών σε μορφή που διευκολύνει την αναδιατύπωση είτε ερωτημάτων είτε αποτελεσμάτων.

Σχετικά με την εφαρμογή των κανόνων αντιστοίχισης που έχουν ορισθεί με απώτερο σκοπό την απάντηση των ερωτημάτων του χρήστη, είχαμε προτείνει στο παρελθόν έναν καινοτόμο μηχανισμό για τη μετάφραση των SPARQL ερωτημάτων (και

προαιρετικά των δεδομένων που λαμβάνουμε), ο οποίος αναλάμβανε να κάνει τις απαραίτητες αλλαγές στο αρχικό SPARQL ερώτημα και στα δεδομένα που λαμβάναμε (αν αυτό ήταν απαραίτητο) με βάση τα Μοτίβα Αντιστοίχισης που είχαν χρησιμοποιηθεί [70]. Ένας βασικός περιορισμός της προσέγγισης αυτής είναι το γεγονός ότι ο αλγόριθμος βασίζεται σε έναν περιορισμένο αριθμό μοτίβων αντιστοίχισης, στα οποία τα οντολογικά στοιχεία που συμμετέχουν στο αριστερό και δεξί μέρος του κανόνα είναι συγκεκριμένα. Στην τρέχουσα εργασία, επειδή υποστηρίζεται μια πολύ πιο εκφραστική γλώσσα αντιστοίχισης που επιτρέπει στους χρήστες να καθορίσουν δυναμικά τα οντολογικά στοιχεία που συμμετέχουν στον καθορισμό ενός κανόνα αντιστοίχισης, χρησιμοποιώντας ένα ή περισσότερα οντολογικά πρότυπα, είναι απαραίτητο να ακολουθήσουμε μια αρκετά διαφορετική προσέγγιση, στην οποία να αποφασίζονται δυναμικά οι αλλαγές που πρέπει να γίνουν είτε στο αρχικό ερώτημα είτε στα αποτελέσματα που λαμβάνουμε.

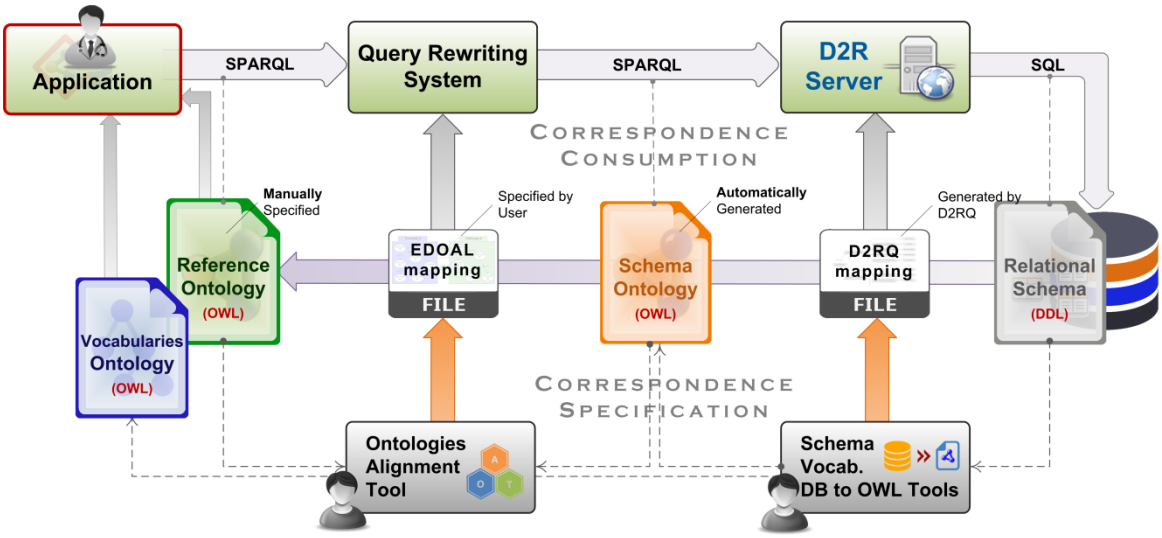
Η σελίδα αυτή είναι σκόπιμα λευκή

10 *Μεθοδολογία – Οικοσύστημα*

10.1Συνοπτική Περιγραφή της Προσέγγισης που Ακολουθείται

Για να επιτρέψουμε στους χρήστες να εκφράσουν περίπλοκα ερωτήματα, στα οποία τα δεδομένα θα πρέπει να ικανοποιούν μία σειρά από συνθήκες χωρίς να είναι απαραίτητο να γνωρίζουν το σχήμα της βάσης και τις κωδικοποιήσεις που χρησιμοποιούνται, είναι απαραίτητη η εννοιολογική περιγραφή των δεδομένων που καταγράφονται από τη βάση δεδομένων καθώς και ο σχεδιασμός μιας διαδικτυακής εφαρμογής που θα επιτρέπει στους χρήστες να περιγράψουν τα δεδομένα που ενδιαφέρονται και τις συνθήκες που θα πρέπει αυτά να ικανοποιούν μέσω ενός γραφικού περιβάλλοντος. Ακολούθως, τα ερωτήματα που ορίζονται από το χρήστη θα πρέπει να εκφραστούν τυπικά χρησιμοποιώντας τη γλώσσα SPARQL και κατόπιν να χρησιμοποιηθούν για την εύρεση των επιθυμητών δεδομένων.

Για να εκμεταλλευτούμε τις δυνατότητες που προσφέρουν οι σχεσιακές βάσεις δεδομένων καθώς και οι τεχνολογίες του σημασιολογικού ιστού, τα αυτομάτως παραγόμενα SPARQL ερωτήματα θα πρέπει να αναδιατυπωθούν και να μεταφραστούν στα αντίστοιχα SQL ερωτήματα χρησιμοποιώντας τους όρους της βάσης. Τα δεδομένα που λαμβάνονται από τη σχεσιακή βάση θα πρέπει να ακολουθήσουν την αντίστροφη διαδικασία, προκειμένου να είναι εκφρασμένα χρησιμοποιώντας τους όρους του χρήστη. Στο Σχήμα 8 παρουσιάζεται συνοπτικά η προσέγγιση που ακολουθείται καθώς και τα εργαλεία/μηχανισμοί που αναπτύχθηκαν για το σκοπό αυτό.



Σχήμα 8: Συνοπτική περιγραφή του Οικοσυστήματος για τη διασύνδεση των Κριτηρίων Καταλληλότητας με μία ή περισσότερες Βάσεις Ασθενών.

Σημειώνουμε ότι, για να καταστεί εφικτή η δυνατότητα εκτέλεσης ερωτημάτων από το χρήστη, είναι απαραίτητη μια φάση αρχικοποίησης του συστήματος (Φάση Α), κατά την οποία πρέπει να καθοριστούν οι οντολογίες και οι κωδικοποιήσεις που χρησιμοποιούνται στις δύο πλευρές καθώς επίσης και η συσχέτιση μεταξύ τους. Κατόπιν, ο χρήστης μπορεί να χρησιμοποιήσει το σύστημα (Φάση Β) για την αποτίμηση των ερωτημάτων του τα οποία είναι εκφρασμένα χρησιμοποιώντας τους όρους, την Οντολογία Αναφοράς και τις Κωδικοποιήσεις που έχουν επιλεχθεί. Ακολούθως, το λογισμικό στοιχείο Mediator που αναπτύχθηκε αναλαμβάνει αρχικά την αναδιατύπωση των SPARQL ερωτημάτων με βάση τις συσχετίσεις που έχουν καθοριστεί, ώστε να είναι συμβατά με την ορολογία που χρησιμοποιείται από τον D2R εξυπηρετητή, ο οποίος αναλαμβάνει τη μετάφρασή τους σε SQL και κατόπιν την εκτέλεσή τους. Ο Mediator αναλαμβάνει επίσης την αναδιατύπωση των αποτελεσμάτων, ώστε αυτά να είναι εκφρασμένα με τους όρους που χρησιμοποιούνται από τη μεριά του χρήστη.

Στα πλαίσια της εργασίας αυτής χρησιμοποιήθηκε ο D2R εξυπηρετητής, ο οποίος επιτρέπει να χειριστούμε μια σχεσιακή βάση δεδομένων ως έναν «εικονικό» RDF γράφο. Ο λόγος που επιλέχθηκε το παραπάνω εργαλείο είναι λόγω της ευρείας χρήσης του στο

σημασιολογικό ιστό. Ακολουθεί μια βήμα προς βήμα περιγραφή της παραπάνω προσέγγισης. Τα εργαλεία που αναπτύχθηκαν θα περιγραφούν αναλυτικά στις επόμενες ενότητες.

10.2 Πρόσβαση στην Βάση χρησιμοποιώντας SPARQL – Μέρος Α

Για την πρόσβαση στη σχεσιακή βάση δεδομένων χρησιμοποιώντας SPARQL ερωτήματα είναι απαραίτητη η οντολογική περιγραφή της βάσης, καθώς και ο καθορισμός των συσχετίσεων μεταξύ των οντολογικών και σχεσιακών στοιχείων κατά τη φάση της αρχικοποίησης του συστήματος. Τα παραπάνω μπορούν να προκύψουν άμεσα μέσω μιας αυτόματης διαδικασίας, όπως περιγράφεται στις επόμενες παραγράφους, ενώ η σαφήνεια των οντολογικών τους στοιχείων μπορεί να βελτιωθεί αισθητά με την ενεργή συμμετοχή των υπευθύνων διαχείρισης της βάσης. Ακολούθως, η αποτίμηση των SPARQL ερωτημάτων είναι εφικτή μέσω της «μετάφρασής» τους σε SQL ερωτήματα.

10.2.1 Οντολογική Περιγραφή της Σχεσιακής Βάσης Δεδομένων

Η οντολογική αναπαράσταση του σχήματος της βάσης (Schema Ontology) προκύπτει αυτόματα χρησιμοποιώντας τον DB-2-OWL αλγόριθμο [100]. Ο αλγόριθμος αυτός, γενικά, παράγει μία OWL κλάση για καθένα από τους πίνακες της βάσης, ενώ παράγει ένα Object ή Datatype Property για καθένα από τα πεδία του πίνακα, λαμβάνοντας υπόψη τους περιορισμούς που έχουν οριστεί. Πιο συγκεκριμένα, εάν ένα πεδίο είναι «ξένο» κλειδί, τότε παράγεται ένα Object Property, διαφορετικά ένα Datatype Property, το πεδίο τιμών του οποίου εξαρτάται από τον αντίστοιχο SQL τύπο. Τα οντολογικά στοιχεία που παράγονται αυτόματα κατά την παραπάνω διαδικασία εξαρτώνται αποκλειστικά από τον ορισμό των αντίστοιχων σχεσιακών στοιχείων και κατά συνέπεια η ονομασία και περιγραφή τους δεν είναι συχνά ικανοποιητική [101]. Για το λόγο αυτό, για καθένα από τα αυτομάτως παραγόμενα οντολογικά στοιχεία είναι απαραίτητη η επικοινωνία με τους

ανθρώπους που είναι υπεύθυνοι για το σχεδιασμό και τη διατήρηση της σχεσιακής βάσης δεδομένων (DB experts), προκειμένου να αποσαφηνιστεί πλήρως η σημασία τους (στα πλαίσια του εκάστοτε οργανισμού), εισάγοντας την κατάλληλη ετικέτα (label) και σχόλια (comments) στην αυτομάτως παραγόμενη οντολογία.

Σημειώνουμε ότι, στην περίπτωση που οι τιμές μιας παραμέτρου προέρχονται από ένα συγκεκριμένο σύνολο τιμών (controlled set of terms), οι τιμές αυτές θα πρέπει να καταγραφούν και η σημασία τους θα πρέπει να προσδιοριστεί, ιδιαίτερα στις περιπτώσεις εκείνες στις οποίες η ονομασία και η περιγραφή τους δε βρίσκεται στη βάση δεδομένων. Για το σκοπό αυτό, αναπτύχθηκαν εργαλεία [102] που επιτρέπουν στους χρήστες να εξάγουν τις διαφορετικές τιμές μιας παραμέτρου (π.χ. τιμές για το φύλο ενός ανθρώπου) με τη μορφή των οντολογιών, καθώς επίσης και των εγγραφών των πινάκων που χρησιμοποιούνται για το σκοπό αυτό (π.χ., πίνακας με τις πιθανές δραστικές ουσίες). Για καθεμία από τις παραπάνω δύο περιπτώσεις ορίζεται μία κλάση που περιλαμβάνει το σύνολο των υπαρκτών τιμών, οι οποίες καταγράφονται ως οντότητες της κλάσης αυτής. Στην πρώτη, όμως, περίπτωση καταγράφεται μόνο το αναγνωριστικό της κάθε τιμής (διακριτή τιμή παραμέτρου – distinct property value), ενώ στη δεύτερη περίπτωση καταγράφονται όλες οι τιμές των παραμέτρων που υπάρχουν για κάθε εγγραφή. Τέλος, σημειώνουμε ότι οι παραπάνω κλάσεις θα πρέπει να συσχετιστούν με τις αντίστοιχες παραμέτρους που παρήχθησαν αυτόματα κατά την οντολογική περιγραφή του σχήματος της βάσης.

10.2.2 Καθορισμός συσχέτισης μεταξύ Οντολογίας και Σχήματος Βάσης

Για να είναι εφικτή η εκτέλεση των SPARQL ερωτημάτων, είναι απαραίτητος ο καθορισμός της συσχέτισης μεταξύ των οντολογικών και σχεσιακών στοιχείων. Δεδομένου ότι η οντολογία παράγεται μέσω μιας αυτόματης διαδικασίας, η συσχέτιση μεταξύ των οντολογικών και σχεσιακών στοιχείων μπορεί επίσης αυτόματα να

καθοριστεί, χρησιμοποιώντας τα εργαλεία που είναι ήδη διαθέσιμα από την D2RQ πλατφόρμα [103].

Ο καθορισμός της συσχέτισης μεταξύ των στοιχείων του σχήματος της σχεσιακής βάσης και των όρων της οντολογίας περιγράφεται μέσω της δηλωτικής γλώσσας D2RQ. Η γλώσσα αυτή αποτελείται από όρους, όπως ClassMaps και PropertyBridges, που περιγράφουν τον τρόπο με τον οποίο τα στοιχεία του σχεσιακού μοντέλου (π.χ., πίνακες) συνδέονται με τα στοιχεία της οντολογίας. Για παράδειγμα, το ClassMaps περιγράφει τον τρόπο με τον οποίο παράγονται τα URIs για τα στοιχεία μιας κλάσης, ενώ το PropertyBridges περιγράφει πώς παράγονται οι παράμετροι μιας οντότητας.

10.2.3 Μετάφραση SPARQL ερωτημάτων σε SQL και Επεξεργασία Δεδομένων

Εφόσον έχει καθοριστεί η συσχέτιση μεταξύ των όρων της οντολογίας και του σχεσιακού μοντέλου, τα SPARQL ερωτήματα του χρήστη μπορούν να εκτελεστούν μέσω του D2R εξυπηρετητή. Το εργαλείο αυτό διαθέτει έναν κειμενογράφο, όπου μπορούμε να εκφράσουμε τα ερωτήματά μας, λαμβάνοντας υπόψη τα οντολογικά στοιχεία που υπάρχουν στις δύο προαναφερθείσες οντολογίες. Ακολούθως, ο D2R εξυπηρετητής αναλαμβάνει τη μετάφρασή τους στα αντίστοιχα SQL ερωτήματα καθώς και επεξεργασία των δεδομένων που λαμβάνει από τη βάση ανάλογα με τον τύπο του SPARQL ερωτήματος που έχει υποβληθεί.

Δεδομένου ότι ορισμένες από τις συναρτήσεις που παρέχονται από τη γλώσσα SPARQL δεν υποστηρίζονται από μία σχεσιακή βάση (π.χ., η χρησιμοποίηση regular expressions), το παραγόμενο SQL ερώτημα δεν είναι πάντα σημασιολογικά ισοδύναμο με το SPARQL ερώτημα. Πιο συγκεκριμένα, το SQL ερώτημα περιλαμβάνει ένα υποσύνολο των συνθηκών που έχουν οριστεί στο WHERE τμήμα του SPARQL ερωτήματος, ενώ τα αποτελέσματα που λαμβάνονται από τη βάση φιλτράρονται, λαμβάνοντας υπόψη τις

υπόλοιπες συνθήκες που έχουν οριστεί. Επίσης, στην περίπτωση που έχει χρησιμοποιηθεί κάποια συνάρτηση που επιδρά στα δεδομένα ενός συνόλου (aggregation function), όπως η συνάρτηση "count" που υπολογίζει το πλήθος των οντοτήτων που πληρούν τις συνθήκες, ο D2R εξυπηρετητής βρίσκει πρώτα τα αντίστοιχα δεδομένα, ενώ η καταμέτρησή τους ακολουθεί. Τέλος, τα δεδομένα επεξεργάζονται, λαμβάνοντας υπόψη τον τύπο του ερωτήματος (π.χ., select/construct SPARQL ερωτήματα) και τους τελεστές που έχουν χρησιμοποιηθεί (π.χ., αναφορικά με τη σειρά παρουσίασης των δεδομένων).

10.3Πρόσβαση στη Βάση χρησιμοποιώντας SPARQL – Μέρος Β

Για την πρόσβαση στη σχεσιακή βάση δεδομένων, χρησιμοποιώντας SPARQL ερωτήματα που έχουν εκφραστεί βασισμένα στην εννοιολογική περιγραφή της πληροφορίας που καταγράφεται από τη βάση δεδομένων, είναι απαραίτητος αρχικά ο καθορισμός των σχετικών οντολογιών, καθώς επίσης και ο καθορισμός της συσχέτισης με τα στοιχεία της βάσης και τα πρωτόκολλα που χρησιμοποιούνται για την πρόσβαση στα δεδομένα τους. Δεδομένης της πληθώρας των θεμάτων που πρέπει να αντιμετωπιστούν κατά τον απευθείας καθορισμό της συσχέτισης των οντολογικών στοιχείων με τους αντίστοιχους όρους της βάσης καθώς και της χρησιμοποίησής τους για την απάντηση των SPARQL ερωτημάτων, η παραπάνω διαδικασία είναι ιδιαίτερα δύσκολη, ειδικά στις περιπτώσεις εκείνες στις οποίες υπάρχουν σημαντικές δομικές και σημασιολογικές διαφορές ανάμεσα στους όρους που χρησιμοποιούνται από το χρήστη και τους αντίστοιχους όρους της βάσης που ενδεχομένως να μην μπορούν να καλυφθούν από τις γλώσσες συσχέτισης μεταξύ οντολογικών και σχεσιακών στοιχείων (object relation mapping). Για το λόγο αυτό, στην εργασία αυτή, η αποτίμηση των SPARQL ερωτημάτων λαμβάνει χώρα σε δύο βήματα. Αρχικά αναδιατυπώνουμε τα SPARQL ερωτήματα χρησιμοποιώντας τους όρους που υποστηρίζονται από τη βάση και κατόπιν εκτελούμε τα αναδιατυπωμένα SPARQL ερωτήματα χρησιμοποιώντας τον D2R εξυπηρετητή.

10.3.1Σχεδιασμός Οντολογίας Αναφοράς και Επιλογή Ονοματολογιών

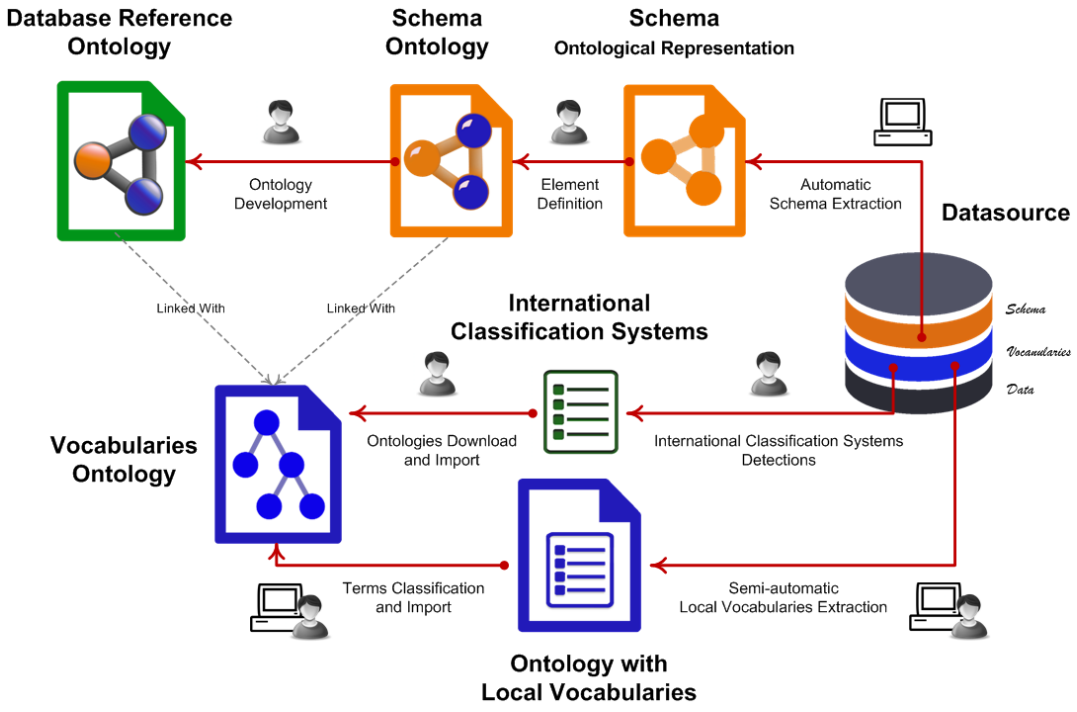
Για την εννοιολογική περιγραφή του πεδίου της γνώσης που καλύπτει μία βάση δεδομένων, χρησιμοποιώντας έναν περιορισμένο αριθμό οντολογικών στοιχείων που περιγράφουν ικανοποιητικά τη βάση δεδομένων και διευκολύνουν το χρήστη στη συγγραφή ενός SPARQL ερωτήματος, είναι απαραίτητος ο σχεδιασμός μιας άλλης οντολογίας επονομαζόμενης Οντολογίας Αναφοράς (Reference Ontology) – Σχήμα 9.

Για το σχεδιασμό της Οντολογίας Αναφοράς, μπορούμε να χρησιμοποιήσουμε δύο διαφορετικές προσεγγίσεις. Πιο συγκεκριμένα, για το σχεδιασμό της μπορούμε να βασιστούμε στην Οντολογία που έχουμε ήδη περιγράψει με βάση το σχήμα της Βάσης (bottom-up approach), παρέχοντας μία «καλύτερη» εννοιολογική περιγραφή του σχήματός της. Ειδικότερα, οι αυτομάτως παραγόμενες κλάσεις μπορούν να οργανωθούν σε τρεις διαφορετικές κατηγορίες. Η πρώτη, επονομαζόμενη «Δεδομένα», περιλαμβάνει τις κλάσεις που χρησιμοποιούνται για την καταγραφή των παραμέτρων μιας οντότητας, όπως για παράδειγμα τις Διαγνώσεις ενός Ασθενούς, τη Συνταγογράφηση Φαρμάκων και τις Ιατρικές Εξετάσεις που έλαβαν χώρα. Η δεύτερη κατηγορία, επονομαζόμενη «Λεξιλόγιο», περιλαμβάνει τις κλάσεις των δεδομένων που αναφέρονται σε ένα πεπερασμένο σύνολο όρων, όπως για παράδειγμα τα πιθανά προβλήματα ή γενικότερα διαταραχές ενός Ασθενούς, τα διαθέσιμα Φάρμακα και/ή τη δραστική τους ουσία, κτλ. Η Τρίτη κατηγορία έχει όνομα «Περίπλοκοι Τύποι Δεδομένων» και περιλαμβάνει τις κλάσεις εκείνες που χρησιμοποιούνται για την αναπαράσταση των δεδομένων τα οποία περιγράφονται πλήρως από δύο ή περισσότερες παραμέτρους, οι οποίες θα πρέπει κάθε φορά να εξεταστούν μαζί, για να προσδιοριστεί επακριβώς η σημασία τους. Υποκατηγορίες περίπλοκων τύπων δεδομένων αποτελούν οι κλάσεις που αναφέρονται σε έναν όρο από κάποιο σύστημα κωδικοποίησης, οι κλάσεις που περιγράφουν μια ποσότητα καθώς και ο ορισμός μιας χρονικής περιόδου.

Αναφορικά με τα Object και Datatype Properties, αυτά μπορούν να οργανωθούν σε ευρύτερες κατηγορίες, λαμβάνοντας υπόψη τη σημασία και το σκοπό για τον οποίο χρησιμοποιούνται. Για παράδειγμα, τα Object Properties που χρησιμοποιούνται για τη διασύνδεση μεταξύ των Δεδομένων μπορούν να διαχωριστούν από τα Object Properties που χρησιμοποιούνται για την καταγραφή μιας τιμής (π.χ., κωδικός προβλήματος, ποσότητα φαρμάκου, κτλ.) που περιγράφονται από έναν περίπλοκο τύπο δεδομένων. Επίσης, τα Datatype Properties που χρησιμοποιούνται για την καταγραφή των παραμέτρων μιας κλάσης μπορούν να οργανωθούν με βάση τον τύπο δεδομένων του πεδίου τιμών τους. Ακόμη, λόγω του μεγάλου αριθμού των αυτομάτως παραγόμενων παραμέτρων που είναι ανάλογος με τον αριθμό των κλάσεων και των πεδίων που αυτές περιέχουν, μία ή περισσότερες παράμετροι με την ίδια σημασία αντικαταστάθηκαν από μια που καλύπτει πλήρως τη σημασία τους, ενώ ο ορισμός της επιτρέπει τη χρησιμοποίησή της για την περιγραφή των οντοτήτων από μία ή περισσότερες από τις διαθέσιμες κλάσεις. Για παράδειγμα, τα Datatype Properties που χρησιμοποιούνται, για να προσδιορίσουν μοναδικά τις οντότητες των κλάσεων, αντικαταστάθηκαν από ένα. Επίσης, τα Datatype Properties που χρησιμοποιούνται για την περιγραφή του κωδικού ενός όρου αντικαταστάθηκαν από μία νέα παράμετρο.

Εναλλακτικά ο σχεδιασμός της Οντολογίας Αναφοράς μπορεί να είναι ανεξάρτητος από τη βάση (top-down approach), παρέχοντας μια εννοιολογική περιγραφή της πληροφορίας που θα θέλαμε να έχουμε στη βάση, βελτιώνοντας τη σημασία και πληρότητα των οντολογικών στοιχείων που ορίζονται. Για παράδειγμα, στην περίπτωση αυτή μπορούμε να παρέχουμε μεγαλύτερη σε βάθος κατηγοριοποίηση των όρων που χρησιμοποιούνται, λεπτομερή περιγραφή των Δεδομένων και κυρίως των παραμέτρων που μας ενδιαφέρουν. Όμως, η προσέγγιση αυτή αυξάνει σημαντικά τη δομική και σημασιολογική απόσταση μεταξύ των όρων των δύο οντολογιών. Πιο συγκεκριμένα, οι

όροι που ορίζονται στις δύο οντολογίες μπορεί να μην είναι σημασιολογικά επικαλυπτόμενοι αλλά σημασιολογικά συσχετίσιμοι, υπό την έννοια ότι η τιμή μιας παραμέτρου μπορεί να προκύψει με βάση την τιμή της άλλης. Για παράδειγμα, η παράμετρος που καθορίζει τη χιλιομετρική ή χρονική απόσταση ενός ασθενούς από ένα νοσοκομείο σχετίζεται άμεσα με την παράμετρο ή παραμέτρους καταγραφής του τόπου διαμονής ενός ανθρώπου. Επιπλέον, είναι αρκετά πιθανό η πληροφορία που ορίζουμε στην Οντολογία Αναφοράς να μην μπορεί να βρεθεί λόγω έλλειψης των αντίστοιχων δεδομένων της βάσης. Για το λόγο αυτό, πολλές φορές, ένας συνδυασμός των δύο παραπάνω προσεγγίσεων είναι η καλύτερη επιλογή, ιδιαίτερα όταν η οντολογία αναφοράς σκοπεύει να χρησιμοποιηθεί, για να παρέχει έναν κοινό τρόπο επικοινωνίας με μία ή παραπάνω βάσεις δεδομένων. Στην περίπτωση αυτή, η χρησιμοποίηση υπαρχόντων, ευρέως χρησιμοποιουμένων, μοντέλων αναφοράς μπορεί να μειώσει σημαντικά τις σημασιολογικές διαφορές που θα συναντήσουμε κατά τον καθορισμό της συσχέτισης μεταξύ των όρων των οντολογιών.



Σχήμα 9: Εννοιολογική Περιγραφή μιας Σχεσιακής Βάσης Δεδομένων

Επιπρόσθετα, οι ορολογίες που χρησιμοποιούνται θα πρέπει επίσης να καταγραφούν και να διασυνδεθούν με τις παραμέτρους της Οντολογίας Αναφοράς. Στην περίπτωση που η οντολογική προσέγγιση που περιγράφουμε πρόκειται να χρησιμοποιηθεί για την πρόσβαση σε αποκλειστικά μία σχεσιακή βάση δεδομένων, η χρησιμοποίηση των ονοματολογιών και κωδικοποιήσεων που χρησιμοποιούνται στη βάση μπορεί να επιταχύνει τη διαδικασία αρχικοποίησης του συστήματος. Όμως, ακόμη και στην περίπτωση αυτή είναι απαραίτητη κάποια προεργασία, όπως αναφέραμε πιο πάνω. Η αυτομάτως παραγόμενη οντολογία με τους όρους που χρησιμοποιούνται στη βάση (ειδικότερα τους τοπικούς όρους, η σημασία των οποίων περιορίζεται στα όρια του εκάστοτε οργανισμού) μπορεί να επεξεργαστεί περαιτέρω από τους ειδικούς στο αντίστοιχο πεδίο γνώσης (π.χ., ασθενειών, φαρμάκων, κτλ.), χρησιμοποιώντας έναν επεξεργαστή οντολογιών, όπως το εργαλείο Protégé [104], για την οργάνωση των όρων σε ευρύτερες κατηγορίες με βάση τη σημασία τους. Επίσης, στην περίπτωση χρησιμοποίησης διεθνών κατηγοριοποιήσεων ή κωδικοποιήσεων, η χρησιμοποίηση των αντίστοιχων αρχείων που είναι διαθέσιμα στο διαδίκτυο μπορεί να βελτιώσει σημαντικά την επικοινωνία, επειδή αφενός παρέχεται η κατηγοριοποίηση των όρων (σε αντίθεση με τους όρους που εξάγονται αυτόματα από τους αντίστοιχους πίνακες της βάσης) και αφετέρου παρέχει πιο πλήρη περιγραφή των όρων της που ορισμένες φορές δεν περιλαμβάνεται στη βάση.

Στην περίπτωση που η οντολογική προσέγγιση θα χρησιμοποιηθεί για την επικοινωνία με παραπάνω από μία βάσεις δεδομένων, οι ονοματολογίες που θα χρησιμοποιηθούν μπορεί να διαφέρουν σημαντικά από τις ονοματολογίες που χρησιμοποιούνται από κάποια από τις βάσεις δεδομένων. Κατά συνέπεια, η συσχέτιση ανάμεσα στους όρους των οντολογιών θα πρέπει να καθοριστεί. Η χρησιμοποίηση διεθνώς αναγνωρισμένων ονοματολογιών και κωδικοποιήσεων τόσο στο σχεδιασμό των

βάσεων όσο και στο μοντέλο αναφοράς είναι επιτακτική, δεδομένου ότι οι συσχετίσεις μεταξύ των όρων τους παρέχονται ήδη από τους εκάστοτε οργανισμούς που είναι υπεύθυνοι για την ανάπτυξη και διαχείρισή τους. Διαφορετικά, η συσχέτιση μεταξύ των όρων θα πρέπει να καθοριστεί από τους ειδικούς στο αντίστοιχο πεδίο γνώσης, μια διαδικασία ιδιαίτερα δύσκολη και χρονοβόρα.

10.3.2 Καθορισμός Συσχέτισης μεταξύ Οντολογιών

Για να καταστήσουμε εφικτή την απάντηση ερωτημάτων με βάση το Μοντέλο Αναφοράς που έχει σχεδιαστεί και τις ονοματολογίες ή κωδικοποιήσεις που έχουν επιλεχθεί, είναι απαραίτητος ο καθορισμός της συσχέτισης με τα αντίστοιχα μοντέλα και κωδικοποιήσεις που υποστηρίζονται από τον D2R εξυπηρετητή. Οι ονοματολογίες και κωδικοποιήσεις που χρησιμοποιούνται από τη μεριά του χρήστη είναι είτε ίδιες με αυτές που χρησιμοποιούνται στη βάση δεδομένων είτε βασίζονται σε διεθνείς κωδικοποιήσεις και συνεπώς η συσχέτιση μεταξύ τους έχει ήδη καθοριστεί. Εξαίρεση αποτελούν οι περιπτώσεις εκείνες στις οποίες έχουν χρησιμοποιηθεί όροι και κωδικοί η σημασία των οποίων περιορίζεται στα σύνορα ενός οργανισμού. Αναφορικά με τις συσχετίσεις μεταξύ των οντολογικών στοιχείων των δύο μοντέλων, στη γενική τους περίπτωση, θα πρέπει να καλύψουμε μια πληθώρα από αναντιστοιχίες [78] και κατά συνέπεια είναι απαραίτητη η χρησιμοποίηση μιας ιδιαίτερα εκφραστικής γλώσσας. Επίσης, δεδομένου ότι τα μοντέλα που χρησιμοποιούνται σχετίζονται άμεσα με τις ονοματολογίες που έχουν επιλεχθεί, χρειάζεται ο αρμονικός συνδυασμός της συσχέτισης μεταξύ μοντέλων και ονοματολογιών, ώστε να είναι εφικτή η μετάβαση από τη μία οντολογία στην άλλη.

Τα Οντολογικά Πρότυπα [86], σε συνδυασμό με τις Συναρτήσεις Μετατροπής Δεδομένων, παρέχουν μια εκφραστική γλώσσα καθορισμού συσχέτισης μεταξύ οντολογικών στοιχείων που μπορεί να χειριστεί αποτελεσματικά τις παραπάνω

περιπτώσεις, ενώ το εργαλείο OAT που αναπτύχθηκε (Κεφάλαιο 4) επιτρέπει στο χρήστη να γεφυρώσει το χάσμα ανάμεσα στις δύο πλευρές μέσω μιας ημιαυτόματης διαδικασίας.

10.3.3Χρησιμοποίηση Συσχέτισης μεταξύ Οντολογιών

Η αναδιατύπωση των SPARQL ερωτημάτων καθώς και των RDF δεδομένων γίνεται μέσω ενός καινοτόμου μηχανισμού που αναπτύχθηκε (Κεφάλαιο 12), ο οποίος βασίζεται στους κανόνες συσχέτισης που έχουν καθοριστεί. Πιο συγκεκριμένα, οι κανόνες συσχέτισης αρχικά χρησιμοποιούνται για την αναδιατύπωση των SPARQL ερωτημάτων που έχουν εκφραστεί χρησιμοποιώντας τους όρους της Οντολογίας Αναφοράς καθώς και τις ονοματολογίες που έχουν επιλεχθεί έτσι, ώστε το ερώτημα να είναι εκφρασμένο χρησιμοποιώντας τους όρους της βάσης και κατά συνέπεια να μπορεί να εκτελεστεί μέσω του D2R εξυπηρετητή. Ακολούθως, τα δεδομένα που λαμβάνουμε από τη βάση δεδομένων θα πρέπει να εκφραστούν χρησιμοποιώντας τους όρους του χρήστη. Για το σκοπό αυτό, οι συσχετίσεις που έχουν καθοριστεί χρησιμοποιούνται για την περαιτέρω επεξεργασία των RDF δεδομένων που λαμβάνουμε ως αποτέλεσμα της εκτέλεσης του αναδιατυπωμένου, για παράδειγμα, construct SPARQL ερωτήματος.

10.4Έκφραση Ερωτημάτων μέσω ενός Γραφικού Περιβάλλοντος

Για την πρόσβαση στη βάση δεδομένων ο χρήστης θα πρέπει να σχηματίσει το κατάλληλο SPARQL ερώτημα, λαμβάνοντας υπόψη τους όρους που υπάρχουν στο Μοντέλο Αναφοράς καθώς και τις ονοματολογίες που έχουν επιλεχθεί. Εναλλακτικά, μπορεί να χρησιμοποιηθεί η διαδικτυακή εφαρμογή που αναπτύχθηκε (παρουσιάζεται αναλυτικά στην ενότητα Γ), η οποία επιτρέπει στο χρήστη να εκφράσει γραφικά SPARQL ερωτήματα με βάση το μοντέλο και τις ονοματολογίες που παρέχονται κατά την αρχικοποίηση του εργαλείου αυτού.

Στα επόμενα δύο κεφάλαια της ενότητας αυτής, έχουμε εστιάσει στα εργαλεία και τους μηχανισμούς που αναπτύχθηκαν, για να καταστήσουμε εφικτή την πρόσβαση στα δεδομένα μιας βάσης δεδομένων με βάση το μοντέλο και τους όρους που υποστηρίζεται από το χρήστη. Πιο συγκεκριμένα, αρχικά παρουσιάζουμε το εργαλείο καθορισμού της συσχέτισης μεταξύ δύο οντολογιών και ακολούθως το μηχανισμό που αναπτύχθηκε για την αυτόματη επανέκφραση των ερωτημάτων με βάση τις συσχετίσεις που έχουν οριστεί. Όπως ήδη αναφέρθηκε, η εφαρμογή για την έκφραση ερωτημάτων προς τη βάση περιγράφεται σε μια νέα ενότητα.

Η σελίδα αυτή είναι σκόπιμα λευκή

11 *Καθορισμός Συσχέτισης Οντολογιών*

Για τον καθορισμό της συσχέτισης μεταξύ των όρων δύο οντολογιών αναπτύχθηκε μια διαδικτυακή εφαρμογή που επιτρέπει στο χρήστη να καθορίσει τις δύο οντολογίες και ακολούθως να ορίσει τις μεταξύ τους συσχετίσεις. Η γλώσσα αναπαράστασης των συσχετίσεων που υποστηρίζεται από το σύστημα, οι αλγόριθμοι που χρησιμοποιούνται για τον εντοπισμό των συσχετίσεων και η αλληλεπίδραση του χρήστη με το εργαλείο αυτό περιγράφονται αναλυτικά στις επόμενες υποενότητες.

11.1 Περιγραφή Συσχέτισης Μεταξύ Οντολογιών

11.1.1 Οντολογικά Πρότυπα και Κανόνες Συσχέτισης

Ο καθορισμός της συσχέτισης μεταξύ οντολογιών (ontology alignment) περιλαμβάνει τον ορισμό ενός ή περισσοτέρων κανόνων συσχέτισης (mapping rule), καθένας απ' τους οποίους καθορίζει επακριβώς τη σχέση μεταξύ των στοιχείων (ontological elements) δύο οντολογιών. Κατά τον ορισμό ενός νέου κανόνα συσχέτισης πρέπει να καθοριστεί μία σειρά από παραμέτρους, συμπεριλαμβανομένων των οντολογικών στοιχείων που συμμετέχουν στον κανόνα καθώς και της μεταξύ τους σχέσης.

Στην εργασία αυτή, για την καταγραφή των οντολογικών στοιχείων που συμμετέχουν στο αριστερό και δεξί μέρος ενός κανόνα βασιστήκαμε στα Οντολογικά Πρότυπα (ΟΠ) [86]. Ένα οντολογικό πρότυπο προσδιορίζει μια η περισσότερες οντότητες

ενός κανόνα συσχέτισης και μπορεί να αναφέρεται σε ένα υπάρχον οντολογικό στοιχείο (π.χ., μία υπάρχουσα κλάση), σε ένα «νέο» που προκύπτει από τον περιορισμό της σημασίας και/ή χρήσης ενός υπάρχοντος στοιχείου (π.χ., περιορίζοντας το εύρος των τιμών μιας παραμέτρου σε ένα υποσύνολο των επιτρεπτών τιμών) ή γενικά σε ένα συνδυασμό από υπάρχοντα οντολογικά στοιχεία (π.χ., ένωση δύο κλάσεων).

Τα Οντολογικά Πρότυπα (ΟΠ) που μπορούν να χρησιμοποιηθούν στο αριστερό ή δεξιό μέρος ενός κανόνα υπάρχουν διαθέσιμα στον ακόλουθο σύνδεσμο [105]. Ενδεικτικά αναφέρουμε ότι το ΟΠ περιορισμού του πεδίου τιμών μιας σχέσης περιλαμβάνει τον ορισμό της σχέσης καθώς και την κλάση στην οποία θα πρέπει να ανήκουν οι τιμές της σχέσης αυτής. Οι δύο αυτές παράμετροι του προτύπου αυτού είναι με τη σειρά τους κάποια άλλα ΟΠ που προσδιορίζουν κάποια σχέση και κλάση αντίστοιχα. Τα ΟΠ μπορεί να αναφέρονται σε ένα υπάρχον στοιχείο μιας οντολογίας (δηλαδή μια σχέση και μια κλάση) ή σε πιο σύνθετα οντολογικά στοιχεία. Για παράδειγμα, για τον περιορισμό του πεδίου τιμών της σχέσης, θα μπορούσαμε να χρησιμοποιήσουμε κάποιο πιο σύνθετο ΟΠ, όπως αυτό που περιορίζει τις οντότητες μιας κλάσης με βάση την τιμή μιας παραμέτρου. Το πρότυπο αυτό με τη σειρά του περιλαμβάνει τον ορισμό της κλάσης, τη συγκεκριμένη παράμετρο της κλάσης καθώς και τη συνθήκη που θα πρέπει να πληρούν τα δεδομένα. Όπως φαίνεται στο παράδειγμα αυτό, τα ΟΠ μπορούν να συνδυαστούν για την έκφραση ιδιαίτερα περίπλοκων συσχετίσεων και ακολούθως να προσδιορίσουμε τις συγκεκριμένες οντότητες οι οποίες αναφέρονται στο αριστερό και δεξιό μέρος ενός κανόνα.

Κατά τον καθορισμό της συσχέτισης μεταξύ των παραμέτρων δύο οντολογιών, ορισμένες φορές, είναι απαραίτητο να ορίσουμε τις αλλαγές που πρέπει να γίνουν στις τιμές τους έτσι, ώστε οι κανόνες να μπορούν να χρησιμοποιηθούν μετέπειτα, για παράδειγμα, για την αναδιατύπωση/μετάφραση των δεδομένων από τη μία οντολογία στην άλλη. Για παράδειγμα, και στις δύο οντολογίες ενδεχομένως να βρούμε μία

παράμετρο που χρησιμοποιείται για τον καθορισμό του μηνιαίου εισοδήματος ενός ανθρώπου, η οποία μπορεί ακόμη και να έχει τον ίδιο τύπο όσον αφορά το πεδίο τιμών της (π.χ., ακέραιος ή δεκαδικός). Όμως, οι τιμές αυτές μπορεί να αναφέρονται σε διαφορετικό νόμισμα (π.χ., Ευρώ στην μία οντολογία και δολάρια ή λίρες Αγγλίας στην άλλη), που θα πρέπει ιδανικά να καταγράφεται στον ορισμό - περιγραφή των στοιχείων αυτών. Σε αυτές τις περιπτώσεις, για να καθορίσουμε επακριβώς τη συσχέτιση ανάμεσά τους έτσι, ώστε να μπορούμε να απαντάμε όχι μόνο σε ερωτήματα σχετικά με το αν ένας άνθρωπος έχει εισόδημα αλλά και αν αυτό είναι πάνω από μία συγκεκριμένη τιμή, θα πρέπει να ορίσουμε τις αλλαγές που πρέπει να γίνουν στις τιμές των παραμέτρων αυτών, όταν οι συσχετίσεις χρησιμοποιούνται για την υλοποίηση σχετικών εργασιών. Εκτός από τον «ευθύ» μετατροπέα τιμών, ίσως χρειάζεται να ορίσουμε και τον «αντίστροφο» μετατροπέα, αλλά αυτό σχετίζεται άμεσα με την κατεύθυνση στην οποία πρόκειται να χρησιμοποιηθούν οι κανόνες συσχέτισης.

Στον Πίνακας 4 έχουμε συνοψίσει τις παραμέτρους που θα πρέπει να καταγραφούν κατά τον ορισμό ενός νέου κανόνα συσχέτισης, μια σύντομη περιγραφή για την καθεμία καθώς και αν αυτές είναι απαραίτητες ή όχι.

Παράμετρος	**Υ/Π**	**Περιγραφή**
Οντότητες 1 και 2 (Entities 1 and 2)	Υ	Ορίζουν τις οντότητες που συμμετέχουν στο αριστερό και δεξιό μέρος ενός κανόνα.
Μετατροπή Τιμών (Data Transformation)	Π	Καθορίζει τις αλλαγές που πρέπει να γίνουν στις τιμές των παραμέτρων που αναφέρονται στις δύο πλευρές
Σχέση (Relation)	Υ	Καθορίζει τη σχέση της οντότητας 1 σε σχέση με την οντότητα 2, π.χ. ισοδύναμοι όροι
Κατεύθυνση (Direction)	Π	Καθορίζει την κατεύθυνση για την οποία ισχύει ο κανόνας, π.χ. Από την οντότητα 1 στην οντότητα 2
Προέλευση (Origin)	Π	Δείχνει την προέλευση του κανόνα (π.χ., ορίστηκε από το χρήστη μέσω ενός γραφικού περιβάλλοντος)

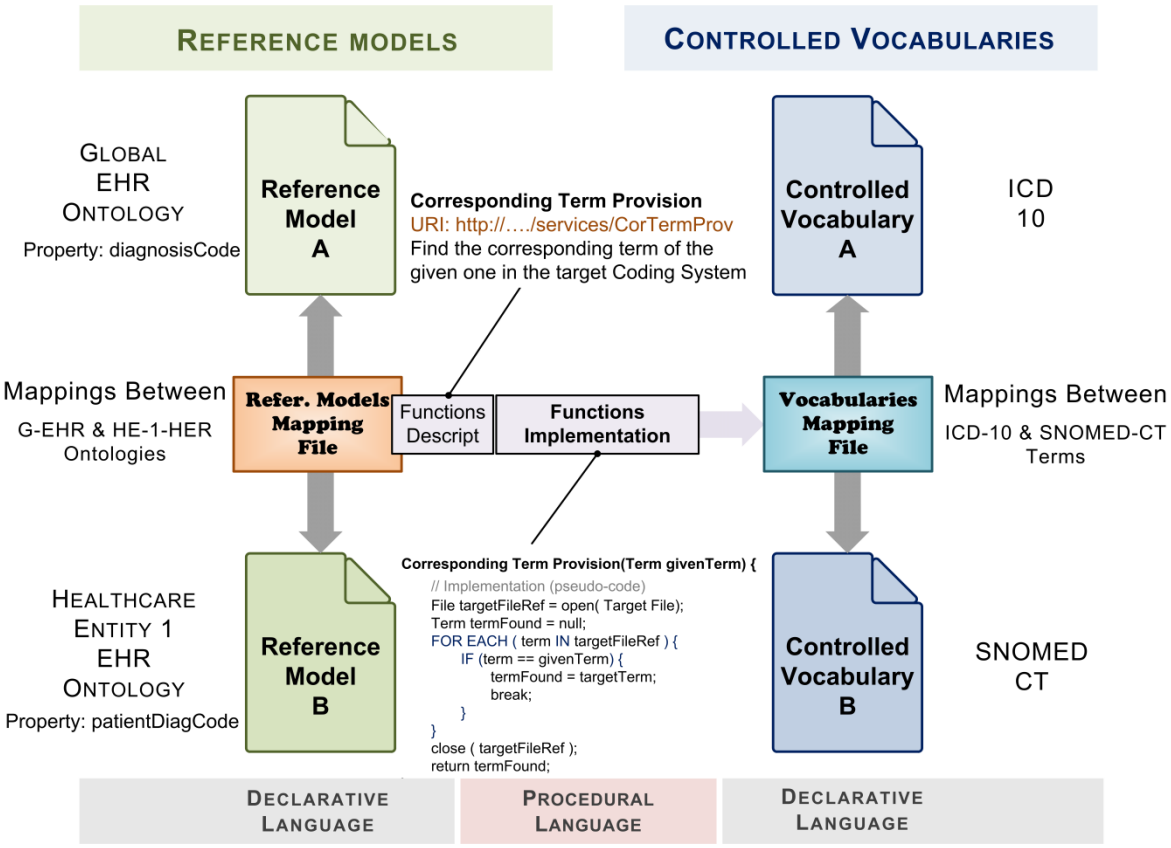
Παράμετρος	**Υ/Π**	**Περιγραφή**
Τιμή Εμπιστοσύνης (Confidence Value)	Π	Δείχνει πόσο σίγουροι είμαστε για τον προτεινόμενο κανόνα τη στιγμή της αποδοχής του.
Περιγραφή (Description)	Π	Παρέχει μια λεπτομερή περιγραφή του κανόνα εκφρασμένη σε φυσική γλώσσα.

Πίνακας 4: Οι (Υ)ποχρεωτικές/(Π)ροαιρετικές Παράμετροι ενός Κανόνα Συσχέτισης.

Στο σημείο αυτό, σημειώνουμε ότι η προέλευση ενός κανόνα (mapping rule origin) δε θα πρέπει να συγχέεται με την προέλευση του συνόλου των κανόνων μεταξύ δύο οντολογιών (mapping file origin). Για παράδειγμα, ο καθορισμός της συσχέτισης μεταξύ δύο οντολογιών μπορεί να έχει προκύψει με κάποιο ημιαυτόματο τρόπο, με τη βοήθεια του εργαλείου αυτού, ωστόσο, κάποιοι από τους κανόνες μπορεί να έχουν προέλθει από την αποδοχή προτεινόμενων από το σύστημα κανόνων, ενώ κάποιοι άλλοι κανόνες μπορεί να έχουν οριστεί εκ του μηδενός από το χρήστη καθώς ο αυτόματος εντοπισμός τους δεν ήταν εφικτός.

11.1.2 Μοντέλα Αναφοράς και Συστήματα Κωδικοποίησης

Η παραπάνω γλώσσα καθορισμού της συσχέτισης μεταξύ δύο οντολογιών μας επιτρέπει να χειριστούμε αποτελεσματικά τις περιπτώσεις εκείνες όπου έχουν χρησιμοποιηθεί διαφορετικές ονοματολογίες και κωδικοποιήσεις για τον προσδιορισμό των παραμέτρων δύο οντοτήτων. Πιο συγκεκριμένα, τα ΟΠ μας επιτρέπουν να προσδιορίσουμε τις αντίστοιχες παραμέτρους, ενώ ο καθορισμός της συνάρτησης την μετάβαση από την μία κωδικοποίηση στην άλλη. Όπως φαίνεται και στο Σχήμα 10, υπάρχει σαφής διαχωρισμός της συσχέτισης μεταξύ των όρων των δυο οντολογιών (μοντέλα αναφοράς) από τον καθορισμό της συσχέτισης μεταξύ των πιθανών τους τιμών (συστήματα κωδικοποίησης).



Σχήμα 10: Καθορισμός της Συσχέτισης μεταξύ διαφορετικών Μοντέλων και Συστημάτων Κωδικοποίησης.

Ο λόγος που επιλέξαμε την προσέγγιση αυτή είναι το γεγονός ότι, σε αρκετές περιπτώσεις, η συσχέτιση μεταξύ των όρων διαφορετικών συστημάτων κωδικοποίησης προϋπάρχει. Ο σύνδεσμος αυτών των δύο συσχετίσεων γίνεται μέσω της χρησιμοποίησης των συναρτήσεων που αναλαμβάνουν τη μετάβαση από το ένα σύστημα κωδικοποίησης στο άλλο. Σημειώνουμε ότι, η συσχέτιση μεταξύ δύο οντολογιών (συμπεριλαμβανομένων των μοντέλων αναφοράς και των συστημάτων κωδικοποίησης) είναι συχνά εκφρασμένη σε κάποια δηλωτική γλώσσα, ενώ οι συναρτήσεις είναι εκφρασμένες σε κάποια διαδικαστική γλώσσα.

11.2 Αλγόριθμος Εντοπισμού Συσχετίσεων Μεταξύ Οντολογιών

11.2.1 Παραγωγή Υποψήφιων Κανόνων Συσχέτισης

Ο καθορισμός της συσχέτισης μεταξύ των όρων των δύο οντολογιών είναι μια χρονοβόρα διαδικασία, καθώς θα πρέπει να εξεταστούν ένα προς ένα όλα τα στοιχεία των δύο οντολογιών, προκειμένου να εντοπιστούν οι πιθανές συσχετίσεις μεταξύ τους. Για το σκοπό αυτό, αναπτύχθηκε ένας αλγόριθμος για τον αυτόματο εντοπισμό πιθανών συσχετίσεων, ο οποίος λαμβάνει υπόψη τον ορισμό των οντολογικών στοιχείων και συγκεκριμένα το όνομά τους (labels), καθώς και τα αξιώματα που έχουν ρητά οριστεί σε καθεμία από τις δύο οντολογίες. Ο αλγόριθμος που υλοποιήθηκε βασίζεται στην ομοιότητα μεταξύ των οντολογικών προτύπων (ΟΠ) και όχι στην ομοιότητα μεταξύ υπαρχόντων οντολογικών στοιχείων, στα οποία βασίζεται η πλειοψηφία των υπαρχόντων αλγορίθμων και τεχνικών ταιριάσματος (matching algorithms and techniques). Πιο συγκεκριμένα, για καθεμιά από τις δύο παρεχόμενες οντολογίες, ο αλγόριθμος παράγει όλα τα πιθανά ΟΠ που θα μπορούσαν να σχηματιστούν με βάση τους όρους της οντολογίας και κατόπιν υπολογίζει την ομοιότητα μεταξύ τους. Σημειώνουμε ότι, κατά την παραγωγή των πιθανών ΟΠ, εστιάσαμε σε αυτά που προκύπτουν, εάν αντικαταστήσουμε τα εσωτερικά τους στοιχεία είτε με κάποιο απλό ΟΠ (προσδιορίζει κάποιο υπάρχον οντολογικό στοιχείο) είτε κάποιο πιο σύνθετο ΟΠ. Ωστόσο, στη δεύτερη περίπτωση, οι εσωτερικές τους παράμετροι θα πρέπει να περιγράφονται από κάποιο απλό ΟΠ. Ο λόγος για την απόφαση αυτή έγκειται στο γεγονός ότι ο αυτόματος εντοπισμός ιδιαίτερα περίπλοκων κανόνων συσχέτισης, στους οποίους συμμετέχουν παραπάνω από 3 οντολογικά στοιχεία σε καθένα από τα 2 μέρη του κανόνα, είναι ιδιαίτερα δύσκολος, ενώ παράλληλα οδηγεί σε αρκετούς, λανθασμένα προτεινόμενους, κανόνες, με βάση τα τεστ που πραγματοποιήθηκαν.

Για τον περιορισμό του αριθμού των παραγόμενων ΟΠ, λαμβάνουμε υπόψη τον ορισμό των οντολογικών στοιχείων που συμμετέχουν σε καθένα από αυτά. Για παράδειγμα, στην περίπτωση μιας παραμέτρου, δημιουργούμε όλα τα πιθανά ΟΠ, λαμβάνοντας υπόψη το πεδίο «δράσης» (γνωστό επίσης ως πεδίο ορισμού) και «τιμών». Πιο συγκεκριμένα, ο αλγόριθμος παράγει αυτόματα τα ΟΠ που περιορίζουν τη χρήση της παραμέτρου αυτής σε ένα στενότερο σύνολο, λαμβάνοντας υπόψη τις υποκλάσεις του υπάρχοντος πεδίου δράσης (κλάσης), εάν και εφόσον έχει οριστεί. Ακολούθως, υπολογίζει την ομοιότητα μεταξύ των παραγόμενων από τις δύο οντολογίες ΟΠ, λαμβάνοντας υπόψη τα οντολογικά στοιχεία που συμμετέχουν σε αυτά καθώς και το ρόλο του καθενός. Για παράδειγμα, ο υπολογισμός της συσχέτισης μεταξύ των δύο παραμέτρων γίνεται λαμβάνοντας υπόψη τα ονόματα των παραμέτρων καθώς και τα πεδία δράσης και τιμών, σύμφωνα με την Εξίσωση 1, όπου κ είναι μία παράμετρος με τιμή μεγαλύτερη από 1, η οποία αποτελεί μέρος της παραμετροποίησης του συστήματος.

$$\text{Similarity} = \frac{\text{Domain\_Sim} + \text{k} * \text{Property\_Label\_Sim} + \text{Range\_Sim}}{2 + \text{k}}$$

Εξίσωση 1: Υπολογισμός ομοιότητας μεταξύ δύο παραμέτρων λαμβάνοντας υπόψη το πεδίο δράσης (domain) και τιμών (range)

Ο λόγος ύπαρξης αυτής της παραμέτρου είναι, για να δοθεί μεγαλύτερη έμφαση στα ονόματα των παραμέτρων παρά στα πεδία δράσης και τιμών τους. Σημειώνουμε ότι το όνομα της παραμέτρου καθώς επίσης και το πεδίο δράσης και τιμών που χρησιμοποιούνται για την εύρεση της ομοιότητας των ΟΠ εξαρτάται από το ΟΠ που έχει χρησιμοποιηθεί σε καθένα από τα δύο μέρη του υποψήφιου κανόνα συσχέτισης. Πιο συγκεκριμένα, στην περίπτωση που μία παράμετρος περιγράφεται από ένα ΟΠ που περιορίζει το εύρος του πεδίου δράσης (Domain) και τιμών (Range) μιας υπάρχουσας παραμέτρου, κατά τον υπολογισμό της ομοιότητας των ΟΠ χρησιμοποιούμε τα «νέα»

πεδία δράσης και τιμών που έχουμε ορίσει στο ΟΠ για την παράμετρο αυτή και όχι το πεδίο δράσης και τιμών που πιθανώς να ορίζεται στην οντολογία.

Οι υποψήφιοι κανόνες συσχέτισης αποτελούνται από τα ζευγάρια εκείνα για τα οποία η ομοιότητα (εκφράζεται μέσω μιας τιμής επονομαζόμενης τιμής εμπιστοσύνης – confidence value) που έχει υπολογιστεί βρίσκεται πάνω από μια προκαθορισμένη τιμή (threshold), η οποία κυμαίνεται στο διάστημα μεταξύ 0 και 1 (συμπεριλαμβανομένων των δύο αυτών τιμών). Όσο πιο μεγάλη είναι η τιμή της τόσο πιο σίγουροι είμαστε για τον προτεινόμενο κανόνα. Δεδομένου ότι η τιμή αυτή αναμένεται να ποικίλλει ανάλογα με τις οντολογίες που συμμετέχουν στη διαδικασία εντοπισμού συσχέτισης, η τιμή αυτή αποτελεί μέρος της παραμετροποίησης του συστήματος. Σημειώνουμε ότι, στην περίπτωση που είναι παραπάνω από μία διαθέσιμες συσχετίσεις για ένα οντολογικό στοιχείο (π.χ., αποτελεί τη βασική παράμετρο παραπάνω από ένα ΟΠ), επιλέγουμε προς παρουσίαση/πρόταση τον κανόνα με τη μεγαλύτερη τιμή εμπιστοσύνης.

11.2.2 Ομοιότητα μεταξύ Συμβολοακολουθιών

Για τον υπολογισμό της ομοιότητας μεταξύ δύο συμβολοακολουθιών (strings) χρησιμοποιήθηκε μια πληθώρα από αλγορίθμους και τεχνικές. Πιο συγκεκριμένα, αρχικά εντοπίσαμε τα επιμέρους τμήματα (tokens) της κάθε συμβολοακολουθίας που διαχωρίζονται με ένα ή περισσότερα κενά, αγνοώντας τα σημεία στίξης (punctuation characters) και τις λέξεις τέλους (stop words), οι οποίες αποτελούνται κατά βάση από λειτουργικές λέξεις (function words) [106] που χρησιμοποιούνται για τη δημιουργία γραμματικά σωστών εκφράσεων και δε συνεισφέρουν στη σημασία μιας φράσης. Ακολούθως, δημιουργήσαμε έναν πίνακα με την ομοιότητα μεταξύ των εναπομεινάντων τμημάτων και κυρίως του πυρήνα τους (stem), χρησιμοποιώντας τον Porter stemming αλγόριθμο [107] έτσι, ώστε να απαλλαχθούμε από τους χαρακτήρες του κάθε τμήματος που οφείλονται στη μορφή (αριθμό, πτώση) με την οποία χρησιμοποιείται μια λέξη μέσα

στη φράση. Για τον υπολογισμό της ομοιότητας των πυρήνων των επιμέρους τμημάτων της κάθε λέξης χρησιμοποιήσαμε ένα συνδυασμό των Levenshtein Distance [94] και N-3-Gram [108]. Τέλος χρησιμοποιήσαμε τον Hungarian αλγόριθμο [109], προκειμένου να μεγιστοποιήσουμε την ομοιότητα μεταξύ των παραπάνω συμβολοακολουθιών, αγνοώντας ουσιαστικά τη σειρά με την οποία εμφανίζονται οι λέξεις στην κάθε φράση.

11.2.3Συμβολή του Χρήστη

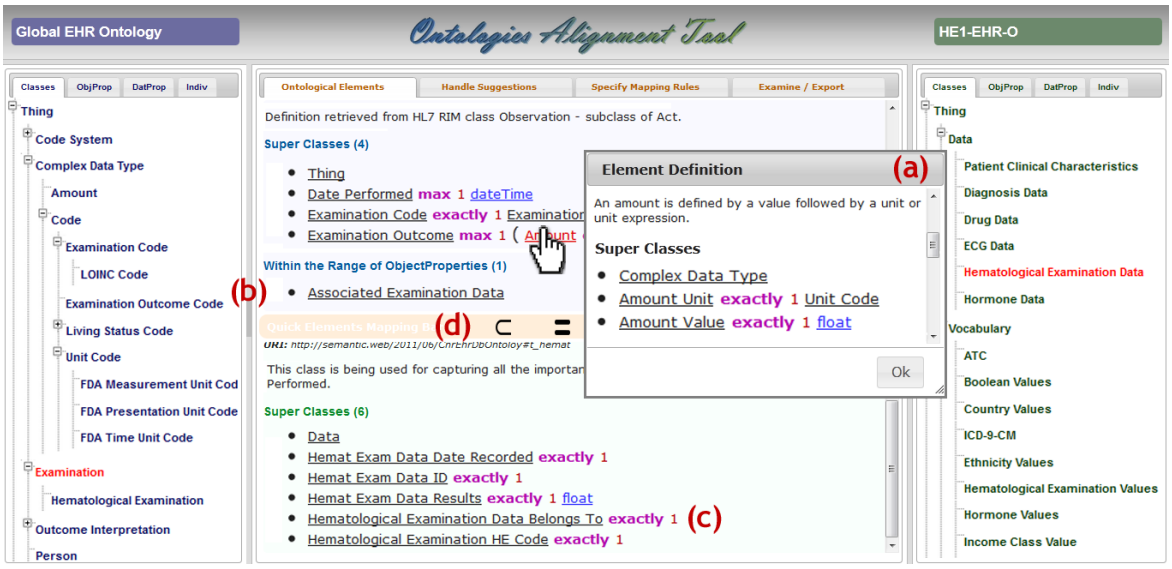
Στην περίπτωση που ο χρήστης συμμετέχει ενεργά στη διαδικασία καθορισμού της συσχέτισης μεταξύ δύο οντολογιών, ο αλγόριθμος λαμβάνει υπόψη τη συμβολή του χρήστη κατά την παραπάνω διαδικασία. Πιο συγκεκριμένα, χρησιμοποιεί τους κανόνες που έχουν ήδη ορισθεί είτε με την αποδοχή ενός προτεινόμενου κανόνα είτε με τον ορισμό του εκ του μηδενός (η ομοιότητα των συγκεκριμένων ΟΠ που συμμετέχουν στο αριστερό και δεξιό μέρος του κανόνα θεωρείται ίση με 1) καθώς επίσης και αυτούς που έχουν απορριφθεί (ομοιότητα ίση με 0) για τη βελτίωση της ακρίβειας των προτεινόμενων κανόνων. Για παράδειγμα, στην περίπτωση που τα πεδία ορισμού (Domain) και τιμών (Range) δύο σχέσεων έχουν ήδη οριστεί ότι είναι ισοδύναμα, η ομοιότητα μεταξύ των παραμέτρων βελτιώνεται σημαντικά εν συγκρίσει με την αρχική τους ομοιότατα σύμφωνα με την Εξίσωση 1. Επομένως, η προσέγγιση που ακολουθείται επιτρέπει τον εντοπισμό επιπρόσθετων συσχετίσεων, η ακρίβεια των οποίων βελτιώνεται σημαντικά, καθώς ο χρήστης συμμετέχει στη διαδικασία καθορισμού των συσχετίσεων μεταξύ των όρων των δύο οντολογιών.

11.3Εργαλείο Καθορισμού Συσχέτισης Μεταξύ Οντολογιών

11.3.1Παρεχόμενες Υπηρεσίες και Αρχιτεκτονική του Συστήματος

Το εργαλείο καθορισμού συσχέτισης οντολογιών (Ontology Alignment Tool - OAT) επιτρέπει στους χρήστες να ορίσουν τις συσχετίσεις ανάμεσα στους όρους των δύο οντολογιών μέσω μιας ημιαυτόματης διαδικασίας, αποδεχόμενοι/απορρίπτοντας τους

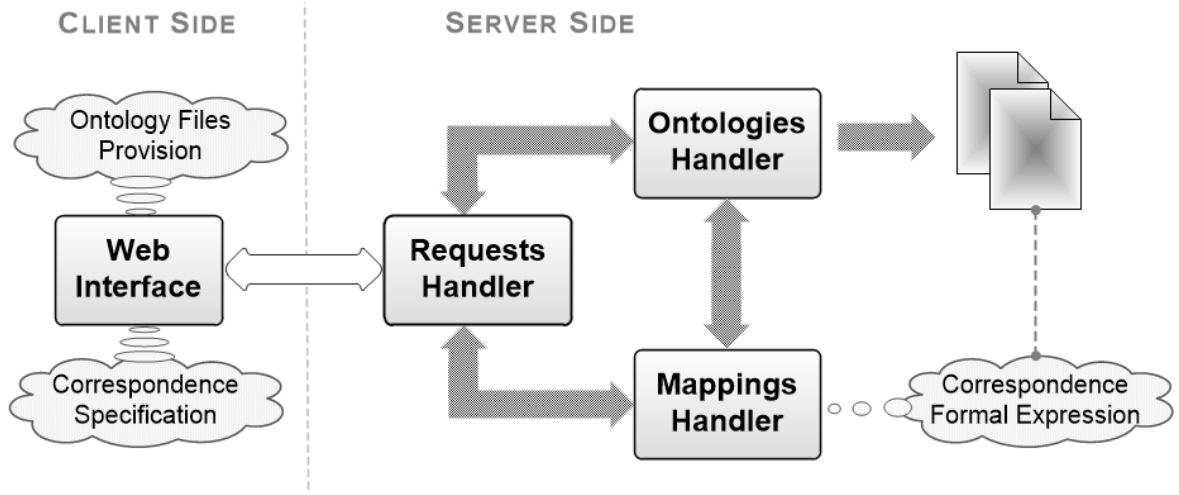
προτεινόμενους κανόνες συσχέτισης ή ορίζοντας εκ του μηδενός όσους δεν μπόρεσαν να ανιχνευτούν. Στη δεύτερη περίπτωση, οι χρήστες μπορούν να καθορίσουν εύκολα τις οντότητες που συμμετέχουν στο αριστερό και δεξιό μέρος, μέσω της χρησιμοποίησης (instantiation) των παρεχόμενων ΟΠ, καθώς επίσης και να ορίσουν τις υπόλοιπες παραμέτρους ενός κανόνα. Το εργαλείο επιτρέπει στο χρήστη να διαχειριστεί επίσης τους κανόνες που έχουν ορισθεί, καθώς επίσης και να τους εξάγει στην επιθυμητή γλώσσα καθορισμού συσχέτισης, ώστε να μπορεί να τους χρησιμοποιήσει σε επόμενο στάδιο για την πραγματοποίηση σχετικών εργασιών.



Σχήμα 11: Στιγμιότυπο της Διεπαφής του Εργαλείου Καθορισμού Συσχέτισης Οντολογιών.

Στο Σχήμα 11 παρουσιάζεται ένα στιγμιότυπο της διεπαφής που αναπτύχθηκε για τον καθορισμό της συσχέτισης μεταξύ δύο οντολογιών. Η Global EHR οντολογία που παρουσιάζεται στο αριστερό μέρος της εικόνας παρέχει την τυπική έκφραση των παραμέτρων που θα θέλαμε να γνωρίζουμε για κάθε ασθενή, ενώ ή Healthcare Entity (HE) 1 EHR οντολογία που παρουσιάζεται στο δεξιό μέρος της εικόνας αυτής αποτελεί την οντολογική αναπαράσταση των δεδομένων των ασθενών (δομή, ονοματολογίες), όπως αυτά έχουν καταγραφεί σε μια συγκεκριμένη μονάδα περίθαλψης.

Η αρχιτεκτονική του συστήματος παρουσιάζεται στο Σχήμα 12. Το σύστημα αποτελείται από μια διαδικτυακή διεπαφή (Web Interface) η οποία χρησιμοποιεί τις υπηρεσίες που παρέχονται από τον εξυπηρετητή (Requests Handler). Το λογισμικό αυτό στοιχείο είναι υπεύθυνο για τη διαχείριση των δύο οντολογιών που παρέχονται από το χρήστη καθώς επίσης και των κανόνων συσχέτισης που έχουν οριστεί. Για το σκοπό αυτό, χρησιμοποιεί τις υπηρεσίες που παρέχονται από άλλα δύο λογισμικά στοιχεία (Ontologies Handler και Mappings Handler) τα οποία βρίσκονται επίσης στον εξυπηρετητή.



Σχήμα 12: Αρχιτεκτονική Συστήματος Καθορισμού Συσχέτισης Οντολογιών.

Οι κανόνες που ορίζονται (στην τρέχουσα έκδοση του συστήματος) αποθηκεύονται στη μεριά του χρήστη και ακολούθως αποστέλλονται στον εξυπηρετητή για την περαιτέρω επεξεργασία τους, όπως, για παράδειγμα, για την τυπική τους έκφραση σε κάποια γλώσσα καθορισμού συσχέτισης. Δεδομένου ότι το εργαλείο αυτό αναπτύχθηκε κυρίως για τον καθορισμό της συσχέτισης μεταξύ διαφορετικών μοντέλων αναπαράστασης δεδομένων, τα οποία περιλαμβάνουν έναν περιορισμένο αριθμό εννοιών εν συγκρίσει με τα συστήματα κατηγοριοποίησης ή κωδικοποίησης που αποτελούνται από χιλιάδες όρους, η παραπάνω απόφαση κρίθηκε λογική. Ωστόσο, αυτό θα μπορούσε να αλλάξει σε μετέπειτα εκδόσεις του εργαλείου αυτού. Οι υπηρεσίες που παρέχει το

εργαλείο καθώς επίσης και η αλληλεπίδρασή του με το χρήστη περιγράφονται αναλυτικά στις επόμενες υποενότητες.

Ένα ιδιαίτερο χαρακτηριστικό της εφαρμογής αυτής είναι ότι επιτρέπει την αποτελεσματική διαχείριση των οντολογικών προτύπων (ΟΠ) και τη συγκεκριμενοποίηση των οντολογικών τους στοιχείων κατά τον καθορισμό της συσχέτισης μεταξύ των όρων των δύο οντολογιών. Τα ΟΠ που υποστηρίζονται από το σύστημα καταγράφηκαν και οργανώθηκαν σε ευρύτερες κατηγορίες (λαμβάνοντας υπόψη τον τύπο των οντολογικών στοιχείων που ορίζουν) κατά το σχεδιασμό της εφαρμογής, καθώς επίσης και οι παράμετροι που πρέπει να παρέχουμε σε καθένα από αυτά. Επιπρόσθετα, ο τρόπος αναπαράστασης των ΟΠ καθορίστηκε κατά την ανάπτυξη της εφαρμογής. Πιο συγκεκριμένα, για καθένα ΟΠ ορίσαμε τον τρόπο αναπαράστασή του σε HTML (για την αλληλεπίδραση με το χρήστη) καθώς και σε JSON για την εσωτερική του αναπαράσταση (για την αλληλεπίδραση με τα υπόλοιπα λογισμικά στοιχεία του συστήματος).

11.3.2 Καθορισμός και Εξερεύνηση Οντολογιών

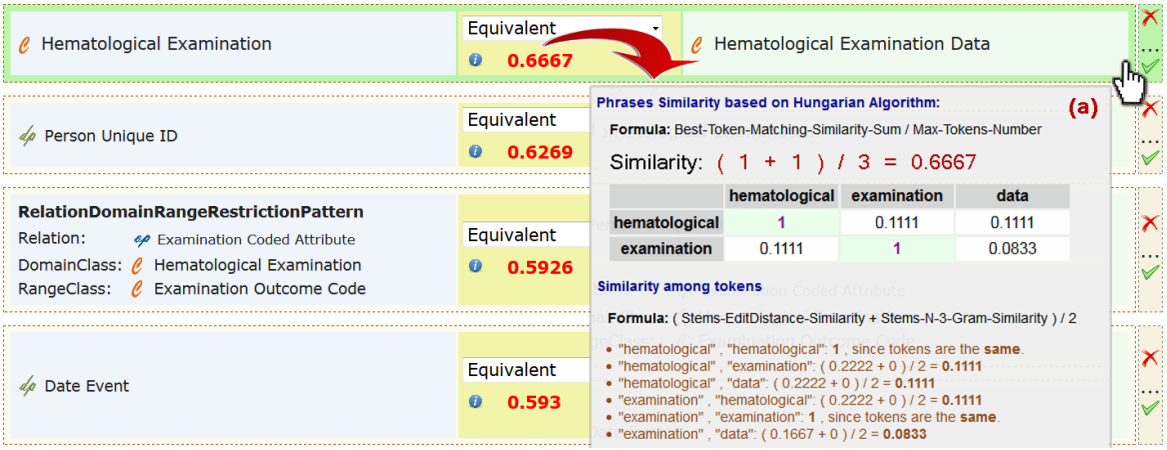
Το εργαλείο επιτρέπει στους χρήστες να καθορίσουν τις δύο οντολογίες, είτε παρέχοντας την τοποθεσία τους στο διαδίκτυο (URL) είτε τα αρχεία των οντολογιών. Αξίζει να σημειωθεί ότι στην περίπτωση που οι οντολογίες χρησιμοποιούν εσωτερικά μια ή περισσότερες οντολογίες, αυτές θα πρέπει επίσης να καθοριστούν κατά τον καθορισμό των δύο οντολογιών. Ακολούθως, τα οντολογικά τους στοιχεία παρουσιάζονται στο αριστερό και δεξιό τμήμα της οθόνης του χρήστη με τη μορφή δένδρου (όμοια με αρκετούς επεξεργαστές οντολογιών, όπως το Protégé [104]) με βάση τα αξιώματα που έχουν οριστεί. Έπειτα, οι χρήστες μπορούν να εξετάσουν ταυτόχρονα τον ορισμό των στοιχείων των δύο οντολογιών, τα οποία παρουσιάζονται στο μεσαίο τμήμα της οθόνης του χρήστη και συγκεκριμένα στην πρώτη καρτέλα. Μέσω της καρτέλας αυτής οι χρήστες μπορούν να εξετάσουν περαιτέρω τα οντολογικά στοιχεία που συμμετέχουν στον ορισμό

τους, ο οποίος παρουσιάζεται σε ένα αναδυόμενο παράθυρο, όπως φαίνεται και στο Σχήμα 11.

Για τα επιλεγμένα οντολογικά στοιχεία, το σύστημα παρέχει όχι μόνο τα αξιώματα που έχουν ρητά οριστεί στην οντολογία (π.χ., πεδίο τιμών των παραμέτρων) αλλά και αυτά που μπορούν άμεσα να εξαχθούν. Για παράδειγμα, στην περίπτωση που το επιλεγμένο στοιχείο είναι μια κλάση, παρέχουμε και τις «εισερχόμενες» παραμέτρους (παράμετροι που μπορούν να «δείχνουν» σε μία οντότητα αυτής της κλάσης) αλλά και τις «εξερχόμενες» παραμέτρους (παράμετροι που μπορούν να εφαρμοστούν σε μία οντότητα αυτής της κλάσης), ώστε ο χρήστης να έχει μια σαφή εικόνα του μοντέλου που έχει οριστεί. Αυτή η πληροφορία είναι ιδιαίτερα χρήσιμη, για την καλύτερη κατανόηση της δομής των δυο οντολογιών και ακολούθως τον εντοπισμό πιθανών συσχετίσεων που δεν μπορούν να εντοπιστούν αυτόματα από το σύστημά.

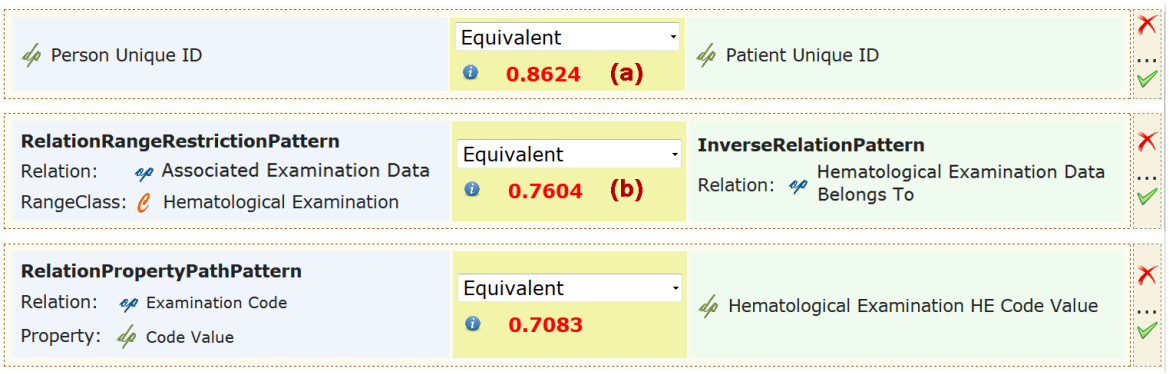
11.3.3Διαχείριση Προτεινόμενων Συσχετίσεων

Μέσω του γραφικού περιβάλλοντος που παρέχεται στη δεύτερη καρτέλα, οι χρήστες μπορούν να διαχειριστούν τους «υποψήφιους» κανόνες που εντοπίζονται αυτόματα από το σύστημα/εργαλείο με βάση τον ορισμό των οντολογικών στοιχείων, όπως περιγράφηκε στην ενότητα 11.2.



Σχήμα 13: Προτεινόμενες Συσχετίσεις (μέσω του γραφικού περιβάλλοντος), όταν κανένας Κανόνας Συσχέτισης δεν έχει ρητά εκφραστεί.

Στο Σχήμα 13 παρουσιάζονται οι πιθανοί κανόνες που εντοπίζονται αυτόματα από το σύστημα αμέσως μετά τη συγκεκριμενοποίηση των δύο οντολογιών. Οι κανόνες παρουσιάζονται ταξινομημένοι με βάση την τιμή εμπιστοσύνης (έναν πραγματικό αριθμό μεταξύ 0 και 1 που δείχνει πόσο σίγουροι/ασφαλείς είμαστε για τον προτεινόμενο κανόνα) που έχει υπολογιστεί. Όπως μπορούμε να δούμε και στην εικόνα αυτή, το σύστημα μας παρέχει επιπρόσθετες πληροφορίες για κάθε προτεινόμενο κανόνα και κυρίως για την τιμή εμπιστοσύνης που έχει υπολογιστεί. Επίσης, επιτρέπει στο χρήστη να εξετάσει περαιτέρω τα οντολογικά στοιχεία που συμμετέχουν σε καθέναν από τους προτεινόμενους κανόνες και κατόπιν να αποδεχτεί ή απορρίψει έναν κανόνα, χρησιμοποιώντας τις επιλογές που υπάρχουν στη δεξιά μεριά. Το εργαλείο επιτρέπει ακόμη τη μαζική αποδοχή ή απόρριψη ενός ή περισσοτέρων κανόνων με βάση την τιμή εμπιστοσύνης που έχει υπολογιστεί.



Σχήμα 14: Προτεινόμενες Συσχετίσεις (μέσω του γραφικού περιβάλλοντος), όταν έχουν ήδη καταγραφεί ορισμένοι Κανόνες Συσχέτισης.

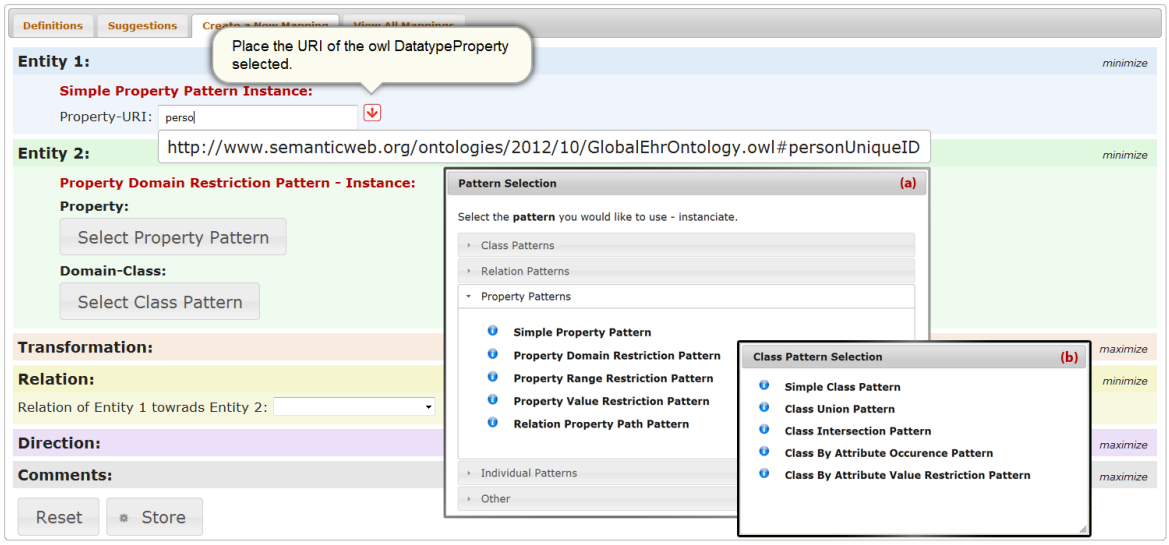
Όπως έχει προαναφερθεί, το εργαλείο λαμβάνει υπόψη τους κανόνες που έχουν οριστεί (είτε με την αποδοχή ενός προτεινόμενου κανόνα είτε με τον εκ του μηδενός ορισμό ενός νέου, όπως παρουσιάζεται στις επόμενες παραγράφους) για τον υπολογισμό των προτεινόμενων συσχετίσεων. Στο Σχήμα 14 παρουσιάζονται οι προτεινόμενοι από το σύστημα κανόνες, όταν έχουν ήδη οριστεί οι δύο παρακάτω: α) οι κλάσεις σχετικά με τις Αιματολογικές Εξετάσεις ενός ασθενούς έχουν ευθυγραμμιστεί, και β) η κλάση ενός

Ανθρώπου έχει ευρύτερη σημασία από την κλάση ενός Ασθενούς. Όπως φαίνεται στην εικόνα, η τιμή εμπιστοσύνης για τον προτεινόμενο κανόνα ανάμεσα στις δύο παραμέτρους, που αναφέρονται στο αναγνωριστικό ενός Ανθρώπου και ενός Ασθενούς αντίστοιχα, έχει αυξηθεί από 0.6296 (Εικόνα 2) σε 0.8624 (Εικόνα 3a), καθώς έχουμε ήδη ορίσει τη συσχέτιση ανάμεσα στις κλάσεις του πεδίου δράσης της προαναφερθείσας παραμέτρου. Επίσης, παρατηρήσαμε ότι το εργαλείο εντοπίζει (στη συγκεκριμένη περίπτωση ορθώς) όχι μόνο 1 προς 1 συσχετίσεις αλλά και πιο περίπλοκες συσχετίσεις, όπως αυτή που καθορίζει τη σχέση μεταξύ ενός Ασθενούς/Ανθρώπου και των Αιματολογικών Εξετάσεων που σχετίζονται με αυτόν στις δύο οντολογίες. Πιο συγκεκριμένα, το εργαλείο εντοπίζει ότι η σχέση που συνδέει τον Ασθενή (υποκατηγορία της κλάσης Άνθρωπος) με τις Αιματολογικές του εξετάσεις είναι ισοδύναμη με την ανάστροφη της σχέσης που συνδέει μια Αιματολογική Εξέταση με έναν Ασθενή στην άλλη οντολογία.

11.3.4 Ορισμός Νέων Συσχετίσεων

Το εργαλείο επιτρέπει τον ορισμό επιπλέον κανόνων που δεν έχουν αυτόματα προταθεί από το σύστημα, μέσω των επιλογών που προσφέρονται από την πρώτη (για τον ορισμό 1:1 συσχετίσεων) και κυρίως την τρίτη καρτέλα (για τον ορισμό n:m συσχετίσεων). Για παράδειγμα, ο κανόνας συσχέτισης μεταξύ των κλάσεων του Ανθρώπου και των Δεδομένων ενός Ασθενούς δεν έχει εντοπιστεί αυτόματα, παρά το γεγονός ότι οι κλάσεις αυτές χρησιμοποιούνται για τον ίδιο σκοπό (καταγραφή δεδομένων των ασθενών – άμεσα ή έμμεσα συνδεδεμένων με τις κλάσεις αυτές, όπως το αναγνωριστικό ενός ασθενούς, δημογραφικά, ασθένειες, αιματολογικές εξετάσεις, κτλ.). Η συσχέτιση αυτή μπορεί να καθοριστεί μέσω της λειτουργικότητας που παρέχεται από την καρτέλα 1 (Σχήμα 11) και κυρίως τις επιλογές (π.χ., ισοδύναμοι όροι) που υπάρχουν στο κέντρο της οθόνης του χρήστη.

Για τον ορισμό περίπλοκων συσχετίσεων, αναπτύχθηκε ένα ιδιαίτερα αλληλεπιδραστικό γραφικό περιβάλλον (Καρτέλα 3) που επιτρέπει στο χρήστη να καθορίσει εύκολα και γρήγορα τις οντότητες που συμμετέχουν στο αριστερό και δεξιό μέρος του κανόνα καθώς επίσης και τις υπόλοιπες παραμέτρους του (Σχήμα 15). Όπως έχει ήδη αναφερθεί, το σύστημα βασίζεται σε ΟΠ για τον καθορισμό των παραμέτρων, δίνοντας τη δυνατότητα στο χρήστη να συνδυάσει δύο ή περισσότερα ΟΠ για την έκφραση περίπλοκων συσχετίσεων. Επίσης, διευκολύνει το συνδυασμό ενός ή περισσοτέρων ΟΠ για την περιγραφή των στοιχείων που συμμετέχουν στο αριστερό ή δεξιό μέρος του κανόνα. Ακόμη, συνεισφέρει σημαντικά στον καθορισμό των στοιχείων που συμμετέχουν σε κάθε ΟΠ μέσω της αυτόματης συμπλήρωσης που παρέχεται ή της επιλογής των αντίστοιχων οντολογικών στοιχείων από το δένδρο που βρίσκεται στην αριστερή ή δεξιά μεριά. Επιπρόσθετα, επιτρέπει τον καθορισμό των μετατροπών που πρέπει να λάβουν χώρα στις τιμές των παραμέτρων κατά τη μετάβαση από τη μία οντολογία στην άλλη, εάν αυτό είναι απαραίτητο. Ένα ιδιαίτερα περίπλοκο παράδειγμα που δείχνει τη συμβολή της καρτέλας αυτής, υπάρχει στο κεφάλαιο 13.



Σχήμα 15: Γραφικό Περιβάλλον για τον Καθορισμό μιας Νέας Συσχέτισης μέσω της χρησιμοποίησης ενός ή περισσοτέρων Οντολογικών Προτύπων.

Κατά τον καθορισμό των μετατροπών που ενδεχομένως να χρειάζεται να λάβουν χώρα κατά τη μετάβαση από το ένα μοντέλο στο άλλο, το σύστημα παρέχει ήδη κάποιες συναρτήσεις (γενικού σκοπού), όπως τη μετατροπή ενός ακεραίου σε string. Εάν, ωστόσο, η μετατροπή που θα πρέπει να λάβει χώρα δεν υποστηρίζεται από το σύστημα, το εργαλείο επιτρέπει στο χρήστη (ειδικό στο συγκεκριμένο τομέα γνώσης – Domain Expert) να περιγράψει, χρησιμοποιώντας τη φυσική γλώσσα, τη διαδικασία που πρέπει να ακολουθηθεί καθώς και τις παραμέτρους που θα ήθελε να ορίσει. Ακολούθως, η παραπάνω διαδικασία θα πρέπει να υλοποιηθεί (από κάποιον άλλο χρήστη που είναι ειδικός σε θέματα ανάπτυξης λογισμικού – IT Expert) σε κάποια διαδικαστική γλώσσα προγραμματισμού έτσι, ώστε οι κανόνες να μπορούν να χρησιμοποιηθούν στην πράξη για την κάλυψη συγκεκριμένων αναγκών.

11.3.5Διαχείριση Υπαρχουσών Συσχετίσεων

Οι κανόνες συσχέτισης που έχουν ήδη οριστεί παρουσιάζονται στην τελευταία καρτέλα, μέσω της οποίας οι χρήστες μπορούν να εξετάσουν περαιτέρω τα οντολογικά στοιχεία που συμμετέχουν σε κάθε κανόνα καθώς και τις υπόλοιπες παραμέτρους του. Επίσης, έχουν τη δυνατότητα να διαγράψουν έναν κανόνα, εάν αυτό κρίνεται απαραίτητο. Τέλος, μπορούν να εξάγουν τις συσχετίσεις που έχουν οριστεί στην επιθυμητή γλώσσα, συμπεριλαμβανομένων των α) JSON, β) XML-EDOAL και γ) HTML.

Το σύστημα επιτρέπει στους χρήστες να παρέχουν άμεσα τους κανόνες που έχουν οριστεί στο παρελθόν, μέσω της καρτέλας αυτής. Η λειτουργικότητα αυτή παρέχεται με την προϋπόθεση ότι οι κανόνες έχουν εκφραστεί με βάση τη γλώσσα JSON [110]. Η γλώσσα αυτή επιτρέπει τη δομημένη αναπαράσταση όλων των παραμέτρων που έχουν καταγραφεί σε έναν κανόνα, σε αντίθεση με την EDOAL XML αναπαράσταση, όπου υπάρχει κάποια απώλεια πληροφορίας. Πιο συγκεκριμένα, τα στοιχεία ενός κανόνα που δεν μπορούμε να εκφράσουμε μέσω των ετικετών που παρέχονται απ' την EDOAL (π.χ.,

περιγραφή ενός κανόνα) τα εισάγουμε με τη μορφή των σχολίων και επομένως η λήψη των αντίστοιχων παραμέτρων είναι δύσκολη. Επίσης η HTML είναι μια γλώσσα που στοχεύει κυρίως στην παρουσίαση των δεδομένων από ένα φυλλομετρητή και όχι στη δομημένη αναπαράστασή τους.

12 *Αναδιατύπωση SPARQL Ερωτημάτων και RDF Δεδομένων*

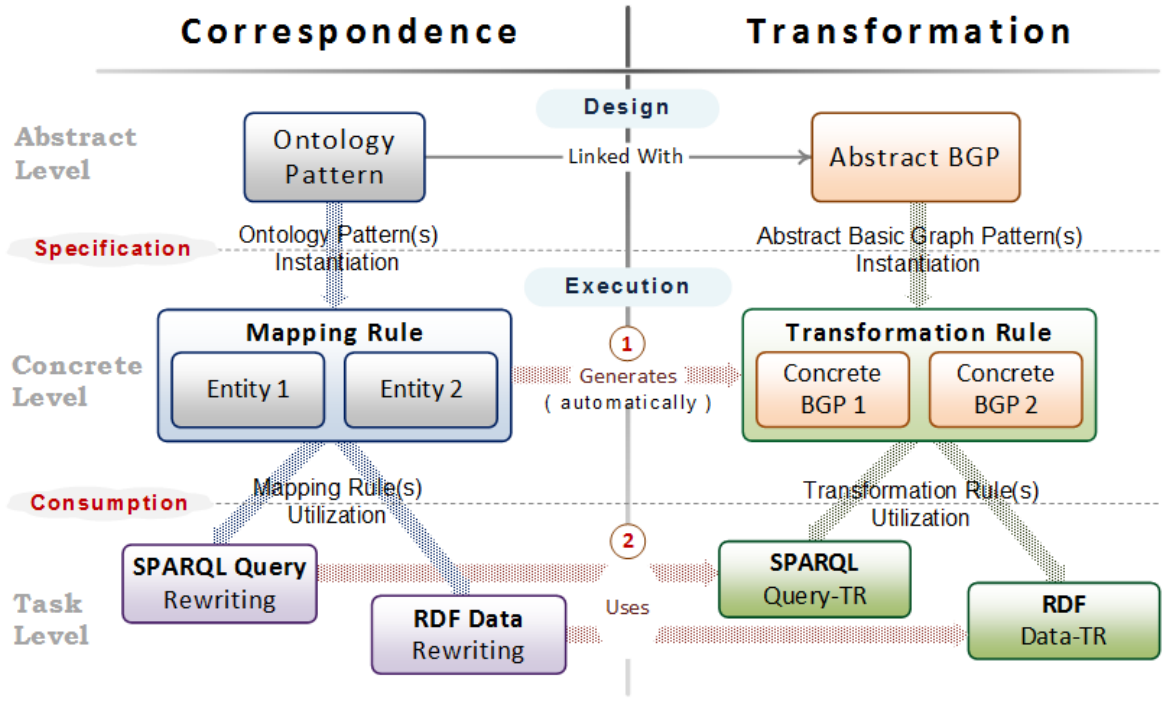
12.1 Οντολογικά Πρότυπα και Κανόνες Μετατροπής

Η γλώσσα που χρησιμοποιήθηκε για τον ορισμό της συσχέτισης μεταξύ των όρων δύο οντολογιών επιτρέπει το συνδυασμό ενός ή περισσοτέρων ΟΠ για τον καθορισμό των στοιχείων που συμμετέχουν στο αριστερό και δεξιό μέρος ενός κανόνα. Λαμβάνοντας υπόψη το γεγονός ότι τα ΟΠ (συμπεριλαμβανομένων και των εσωτερικών ΟΠ) που έχουν χρησιμοποιηθεί στο αριστερό και δεξιό μέρος ενός κανόνα συσχέτισης δεν είναι γνωστά, οι αλλαγές που θα πρέπει να γίνουν στο SPARQL ερώτημα που υποβάλλεται, καθώς επίσης και στα δεδομένα που λαμβάνουμε από την RDF βάση (εάν αυτό είναι απαραίτητο) θα πρέπει να αποφασίζονται δυναμικά, με βάση τα συγκεκριμένα ΟΠ που έχουν χρησιμοποιηθεί για τον καθορισμό των οντολογικών στοιχείων ή γενικότερα παραμέτρων που συμμετέχουν στην καθεμιά από τις δύο πλευρές ενός κανόνα.

Για να καταστεί δυνατή η αυτόματη παραγωγή των Κανόνων Μετατροπής (ΚΜ) που ορίζουν επακριβώς τις αλλαγές που πρέπει να γίνουν στο SPARQL ερώτημα ή στα RDF δεδομένα με βάση τον αντίστοιχο κανόνα συσχέτισης, αρχικά ορίσαμε με έναν «αφηρημένο» - παραμετροποιήσιμο τρόπο αναπαράστασης των Βασικών Γραφικών Μοτίβων (ΒΓΜ) για ένα περιορισμένο σύνολο που αποτελείται από τα «αρχικά» ΟΠ, καθώς επίσης και την διαδικασία που πρέπει να ακολουθηθεί για τη σύνθεση των ΒΓΜ με

βάση τα εσωτερικά ΟΠ που έχουν χρησιμοποιηθεί για την περιγραφή μιας οντότητας σε έναν κανόνα. Κατόπιν, με βάση τους κανόνες που έχουν οριστεί και τα συγκεκριμένα ΟΠ που έχουν χρησιμοποιηθεί σε αυτούς, το σύστημα παράγει αυτόματα τους αντίστοιχους ΚΜ οι οποίοι έπειτα χρησιμοποιούνται για την αναδιατύπωση των SPARQL ερωτημάτων καθώς επίσης και των RDF δεδομένων έτσι, ώστε αυτά να είναι εκφρασμένα με τους όρους που χρησιμοποιούνται σε κάθε πλευρά.

Μια εννοιολογική απεικόνιση της προσέγγισης που ακολουθήθηκε, καθώς και της σύνδεσης μεταξύ των ΟΠ που χρησιμοποιούνται κατά τον καθορισμό της συσχέτισης αλλά και των ΚΜ που χρησιμοποιούνται κατά την εφαρμογή των κανόνων για την κάλυψη συγκεκριμένων αναγκών, απεικονίζεται στο Σχήμα 8.



Σχήμα 16: Συνοπτική Αναπαράσταση του Μηχανισμού Αναδιατύπωσης SPARQL Ερωτημάτων και RDF Δεδομένων.

Όπως φαίνεται στο Σχήμα 16, υπάρχει σαφής διαχωρισμός μεταξύ των οντοτήτων που ορίστηκαν κατά το σχεδιασμό της προσέγγισης που θα ακολουθήσουμε (αφηρημένο επίπεδο) και των οντοτήτων που προκύπτουν από τη χρησιμοποίηση των παραπάνω για ένα συγκεκριμένο σκοπό (συγκεκριμένο επίπεδο), όπως ο ορισμός της συσχέτισης μεταξύ

δύο ή περισσοτέρων οντολογικών στοιχείων καθώς και η χρησιμοποίηση της συσχέτισης αυτής για την κάλυψη συγκεκριμένων αναγκών. Τέλος, η χρησιμοποίηση των παραπάνω οντοτήτων για την αναδιατύπωση ενός SPARQL ερωτήματος καθώς και των δεδομένων που λαμβάνουμε ως απάντηση, εάν αυτό είναι απαραίτητο, τοποθετήθηκε σε ένα νέο επίπεδο επονομαζόμενο επίπεδο εφαρμογής. Οι αλγόριθμοι που υλοποιήθηκαν στο επίπεδο αυτό αναλαμβάνουν την αποτελεσματική διαχείριση και εφαρμογή των αυτομάτως παραγόμενων ΚΜ για την κάλυψη των αναγκών μας και θα περιγραφούν εκτενώς στις υποενότητες που ακολουθούν.

Δεδομένου ότι τα ΟΠ καθώς και η χρησιμοποίησή τους για τον ορισμό των συσχετίσεων έχουν ήδη περιγραφεί, στις ενότητες που ακολουθούν δίνεται ιδιαίτερη έμφαση στην περιγραφή των Αφηρημένων ΒΓΜ και τη χρησιμοποίησή τους για την παραγωγή των ΚΜ καθώς επίσης και την χρησιμοποίηση αυτών για την κάλυψη των αναγκών μας (δηλαδή την περιγραφή των στοιχείων/αλγορίθμων που βρίσκονται κυρίως στο δεξιό μέρος του σχήματος - Σχήμα 16).

12.2Αφηρημένο Επίπεδο

Ένα Αφηρημένο Βασικό Γραφικό Μοτίβο (Α-ΒΓΜ) παρέχει μια τυπική αναπαράσταση ενός ΟΠ με τη μορφή των τριάδων (triples). Το Α-ΒΓΜ για κάποια «απλά» ΟΠ, που προσδιορίζουν μοναδικά ένα υπάρχον οντολογικό στοιχείο μέσω του προσδιορισμού του URI, υπάρχει στον Πίνακας 5. Οι συμβολοακολουθίες που ξεκινούν και τελειώνουν με αγκύλες δηλώνουν ένα μη-ορισμένο οντολογικό στοιχείο (unspecified ontological element), το οποίο θα αντικατασταθεί από το URI του συγκεκριμένου οντολογικού στοιχείου που ορίζεται σε έναν κανόνα αντιστοίχισης μέσω της χρησιμοποίησης των διαθέσιμων ΟΠ. Οι συμβολοακολουθίες που ξεκινούν και τελειώνουν με παρενθέσεις δηλώνουν μια μη-δεσμευμένη οντότητα (unbound entity) και θα αντικατασταθούν από το συγκεκριμένο στοιχείο (π.χ., μια μεταβλητή ή URI) που χρησιμοποιείται σε ένα SPARQL

ερώτημα ή RDF γράφο κατά την εφαρμογή των κανόνων αντιστοίχισης (ειδικότερα των αυτομάτως παραγόμενων κανόνων μετατροπής) για την υλοποίηση ενός συγκεκριμένου σκοπού (π.χ., αναδιατύπωση SPARQL ερωτημάτων).

Οντολογικό Πρότυπο	**Συντό-μευση**	**Εσωτερικά Οντολογικά Στοιχεία**	**Αφηρημένο Βασικό Γραφικό Μοτίβο**
Simple Class Pattern	SCP	Entity Class URI	(subject) rdf:type {class-uri}
Simple Relation Pattern	SRP	Object Property URI	(subject) {property-uri} (object)
Simple Property Pattern	SPP	Datatype Property URI	(subject) {property-uri} (object)

Πίνακας 5: Αφηρημένο Βασικό Γραφικό Μοτίβο για «απλά» Οντολογικά Πρότυπα.

Γενικά, ένα A-ΒΓΜ αποτελείται από ένα πεπερασμένο σύνολο από Αφηρημένες Τριάδες (A-T). Μια A-T είναι μία τριάδα (s,p,o) ∈ (UUAUIUL)×(UUAUI)×(UUAUIUL), όπου U είναι ένα σύνολο αποτελούμενο από τις μη-ορισμένες οντότητες, A είναι ένα σύνολο αποτελούμενο από τις μη-δεσμευμένες οντότητες, I είναι ένα σύνολο με τα Αναγνωριστικά (URIs) και L είναι ένα σύνολο με τα Λεκτικά (Literals).

Οντολογικό Πρότυπο	**Εσωτερικά Οντολογικά Στοιχεία**	**Αφηρημένο Βασικό Γραφικό Μοτίβο**
Relation Domain Restriction (RDR)	Relation: SRP, Domain: SCP	[subject] {property-uri} [object] . [subject] rdf:type {subject-class-uri}
Relation Range Restriction (RRR)	Relation: SRP, Range: SCP	[subject] {property-uri} [object] . [object] rdf:type {object-class-uri}
Relation Property Path (RPP)	Relation: SRP, Property: SPP	[subjectA] {relation-uri} [tmpAB] . [tmpAB] {property-uri} [objectB]

Πίνακας 6: Αφηρημένο Βασικό Γραφικό Μοτίβο «σύνθετων» Οντολογικών Προτύπων, στα οποία οι εσωτερικοί τους παράμετροι εκφράζονται μέσω «απλών» Οντολογικών Προτύπων.

Το Α-ΒΓΜ που αντιστοιχεί στα «περίπλοκα» ΟΠ (δηλαδή τα ΟΠ που εσωτερικά χρησιμοποιούν ένα ή περισσότερα ΟΠ) παράγεται αυτόματα, λαμβάνοντας υπόψη το Α-ΒΓΜ των εσωτερικών ΟΠ καθώς επίσης και το ρόλο τους. Το Α-ΒΓΜ για κάποια περίπλοκα ΟΠ (με την προϋπόθεση ότι τα εσωτερικά ΟΠ είναι απλά) παρουσιάζεται στον Πίνακας 6.

Τα ονόματα που χρησιμοποιούνται για τις μη-δεσμευμένες οντότητες δεν είναι τόσο σημαντικά, καθώς θα αντικατασταθούν από τις συγκεκριμένες οντότητες που χρησιμοποιούνται, για παράδειγμα, σε ένα SPARQL ερώτημα. Παρόλα αυτά, θα πρέπει να τονίσουμε ότι οι μη-δεσμευμένες οντότητες με το ίδιο όνομα αναφέρονται στο ίδιο στοιχείο (π.χ., ίδια μεταβλητή). Πιο συγκεκριμένα, στο Α-ΒΓΜ που αντιστοιχεί στο RDR ΟΠ το «υποκείμενο» της παραμέτρου θα πρέπει να είναι ενός συγκεκριμένου τύπου (κλάσης).

12.3Συγκεκριμένο Επίπεδο

Η παραγωγή των ΚΜ βασίζεται στους ΚΣ που έχουν οριστεί. Πιο συγκεκριμένα, με βάση τα ΟΠ που έχουν χρησιμοποιηθεί για τον ορισμό των στοιχείων που συμμετέχουν στο αριστερό και δεξιό μέρος ενός κανόνα, παράγονται τα συγκεκριμένα ΒΓΜ (Σ-ΒΓΜ), αντικαθιστώντας τα μη-ορισμένα οντολογικά στοιχεία που αναφέρονται στο Α-ΒΓΜ με τα συγκεκριμένα στοιχεία που έχουν χρησιμοποιηθεί στον κανόνα. Συνεπώς, ένα Σ-ΒΓΜ αποτελείται από ένα πεπερασμένο σύνολο από συγκεκριμένες τριάδες (Σ-Τ). Ακολούθως, γίνονται οι απαραίτητες παρεμβάσεις στα μη δεσμευμένα στοιχεία των Σ-ΒΓΜ που παρήχθησαν έτσι, ώστε τουλάχιστον μία από τις μη-δεσμευμένες οντότητες να αναφέρεται στα αντίστοιχα οντολογικά στοιχεία ή γενικότερα δεδομένα. Για παράδειγμα, στην περίπτωση που η συσχέτιση μεταξύ δύο κλάσεων έχει καθοριστεί σε έναν ΚΣ, το υποκείμενο των Σ-ΒΓΜ (δηλαδή η μη-δεσμευμένη οντότητα που βρίσκεται στη θέση του υποκειμένου) που παράγονται από το αριστερό και δεξιό

μέρος του ΚΣ, θα πρέπει να είναι το ίδιο. Επίσης, στην περίπτωση καθορισμού συσχέτισης μεταξύ δύο παραμέτρων, το υποκείμενο και το αντικείμενο των αντίστοιχων Σ-ΒΓΜ θα πρέπει να είναι το ίδιο. Σημειώνουμε ότι οι αντίστροφες σχέσεις αποτελούν μία ειδική περίπτωση, σύμφωνα με την οποία το υποκείμενο και το αντικείμενο της πρώτης θα πρέπει να είναι ίδιο με το αντικείμενο και υποκείμενο της δεύτερης, όπως παρουσιάζεται στην Εξίσωση 2.

$$(\text{subject})\ \text{associated With}\ (\text{object}) \leftrightarrow (\text{object})\ \text{belongs To}\ (\text{subject})$$

Εξίσωση 2: Σχέση μεταξύ των τριάδων δύο αντίστροφων σχέσεων

Ένας ΚΜ μπορεί να συνοδεύεται από μια (συνήθως) ή περισσότερες συνθήκες που θα πρέπει να πληρούν τα στοιχεία που συμμετέχουν στο αριστερό ή δεξιό μέρος ενός ΚΣ, για να ισχύει αυτός, καθώς επίσης και από μία συνάρτηση μετατροπής δεδομένων ή της αντίστροφής της, ανάλογα πως πρόκειται να χρησιμοποιηθεί ο κανόνας. Οι συναρτήσεις που εκφράζουν τις συνθήκες που θα πρέπει να πληρούν τα δεδομένα ή τις αλλαγές που πρέπει να γίνουν, υλοποιήθηκαν σε κάποια διαδικαστική γλώσσα προγραμματισμού με τη μορφή των συναρτήσεων ή διαδικτυακών υπηρεσιών και ορίστηκαν κατά τη χρησιμοποίηση των αντίστοιχων ΟΠ, παρέχοντας ένα αναγνωριστικό (URI) που τις προσδιορίζει μοναδικά. Κατά την παραγωγή των ΚΜ οι παραπάνω συναρτήσεις συσχετίστηκαν με μία (για τις συνθήκες δεδομένων) ή περισσότερες (για τις μετατροπές δεδομένων) μη-ορισμένες οντότητες, που αναφέρονται στα αντίστοιχα δεδομένα (ή γενικότερα σύνολο ή εύρος δεδομένων) που αναφέρονται σε ένα SPARQL ερώτημα ή τα συγκεκριμένα δεδομένα που ορίζονται σε έναν RDF γράφο.

Σημειώνουμε ότι οι συνθήκες λαμβάνουν ως είσοδο ένα σύνολο ή εύρος τιμών και επιστρέφουν αλήθεια (true) ή ψέματα (false) ανάλογα, αν αυτά ικανοποιούν τη συνθήκη. Οι συναρτήσεις μετατροπής λαμβάνουν ως είσοδο μία τιμή και επιστρέφουν την αντίστοιχή της. Στη γενική τους περίπτωση, ορίζουν τις αλλαγές που πρέπει να γίνουν

ανάμεσα σε μία ή περισσότερες παραμέτρους σε καθεμιά από τις δύο πλευρές. Πιο συγκεκριμένα, οι συναρτήσεις αυτές λαμβάνουν ως είσοδο μια λίστα από τιμές που αναφέρονται στις αντίστοιχες παραμέτρους με βάση τη σειρά που έχουν οριστεί και επιστρέφουν μια λίστα με τα αντίστοιχα δεδομένα για τις παραμέτρους που έχουν καθοριστεί στην άλλη μεριά του κανόνα.

Τυπικά, ένας ΚΜ ορίζεται αναδρομικά ως εξής: Εάν Σ-ΒΓΜ1 και Σ-ΒΓΜ2 είναι τα Σ-ΒΓΜ για τις οντότητες 1 και 2 ενός ΚΣ, τότε η έκφραση (C-BGP1 ↔ C-BGP2) είναι ένας ΚΜ, με την προϋπόθεση ότι το σύνολο unb(C-BGP1) ∩ unb(C-BGP2) ≠ ∅, όπου "unb" είναι η συνάρτηση που παρέχει το σύνολο των μη-δεσμευμένων οντοτήτων που αναφέρονται στο Σ-ΒΓΜ που παρέχεται. (2) Εάν Τ είναι ένας ΚΜ και ΣΣ είναι μία δυάδα (u, s) η οποία ανήκει στο U × IC, όπου IC είναι το σύνολο των URIs από γνωστές (δηλαδή δημοσίως διαθέσιμες ή εσωτερικά ορισμένες) συναρτήσεις, τότε το (Τ & ΣΣ) είναι ένας ΚΜ. (3) Εάν Τ είναι ένας ΚΜ και ΣΜ είναι μία δυάδα (u, s) η οποία ανήκει στο U × IDT, όπου IDT είναι το σύνολο των URIs από γνωστές συναρτήσεις/μεθόδους μετασχηματισμού δεδομένων, τότε τα (Τ & Direct: DTP) και (Τ & Inverse: DTP) είναι επίσης ΚΜ με την προϋπόθεση ότι οι «ευθείες» και «αντίστροφες» συναρτήσεις/μέθοδοι μετασχηματισμού δεδομένων δεν έχουν προηγουμένως οριστεί.

12.4Επίπεδο Εφαρμογών

Οι SPARQL και RDF ΚΜ υλοποιήθηκαν ως υπηρεσίες οι οποίες κατά την αρχικοποίησή τους λαμβάνουν ως είσοδο τις παραπάνω παραμέτρους και ακολούθως αναλαμβάνουν να κάνουν τις απαιτούμενες αλλαγές στο SPARQL ερώτημα ή τα RDF δεδομένα, ανάλογα με τον ΚΣ από τον οποίο προέκυψαν. Συνοπτικά, ένας SPARQL ΚΜ κάνει τις απαιτούμενες αλλαγές σε ένα SPARQL ερώτημα, αντικαθιστώντας ένα ή περισσότερα μοτίβα τριάδων (ΤΜ) με τα αντίστοιχα, ενώ κάνει τις απαιτούμενες αλλαγές στο/στα FILTER, εάν αυτό είναι απαραίτητο. Ο RDF ΚΜ αντικαθιστά τα αντίστοιχα Τ με τα νέα,

ενώ κάνει τις απαιτούμενες αλλαγές στα δεδομένα έτσι, ώστε αυτά να είναι εκφρασμένα με τους όρους της άλλης οντολογίας και τις κωδικοποιήσεις που χρησιμοποιούνται.

12.4.1 Αναδιατύπωση SPARQL Ερωτημάτων

Η διαδικασία αναδιατύπωσης ενός SPARQL ερωτήματος βασίζεται εξολοκλήρου στους SPARQL-KM που έχουν αυτόματα παραχθεί με βάση τους ΚΣ που έχουν οριστεί. Συνοπτικά, η παραπάνω διαδικασία περιλαμβάνει τα ακόλουθα βήματα: α) Εντοπισμός των SPARQL-KM που μπορούν να πυροδοτηθούν, β) καθορισμός της σειράς με την οποία θα πρέπει αυτοί να εφαρμοστούν (εάν αυτό είναι απαραίτητο), γ) εφαρμογή των συγκεκριμένων SPARQL-KM (μία ή περισσότερες φορές) και δ) επανεξέταση του αναδιατυπωμένου ερωτήματος και πιθανή αφαίρεση παραμέτρων και συνθηκών που δεν αναδιατυπώθηκαν (λόγω πχ. έλλειψης αντίστοιχων δεδομένων και κανόνων).

Ένας SPARQL-KM μπορεί να πυροδοτηθεί με την προϋπόθεση ότι το Σ-ΒΓΜ που υπάρχει στο αριστερό μέρος του ΚΜ ταιριάζει με τα ΤΜ που υπάρχουν σε ένα SPARQL ερώτημα και ταυτοχρόνως όλες οι συνθήκες που έχουν οριστεί για τις μη-δεσμευμένες οντότητες που υπάρχουν στο αριστερό μέρος του κανόνα (εάν υπάρχουν) ικανοποιούνται.

Ένα Σ-ΤΠ που ανήκει σε έναν Σ-ΒΓΜ ταιριάζει με ένα ΤΜ που ορίζεται στο ΒΓΜ ενός SPARQL ερωτήματος, με την προϋπόθεση ότι τα αντίστοιχα στοιχεία είτε είναι ίδια (στην περίπτωση των URIs και των Λεκτικών) είτε μία συσχέτιση μπορεί να καθοριστεί μεταξύ των μη-δεσμευμένων οντοτήτων και των οντοτήτων που χρησιμοποιούνται σε μία ΤΜ (binding). Ορίζουμε μία σύνδεση b ως τη «μερική» συνάρτηση $b: U \rightarrow (I \cup L \cup V)$ η οποία συνδέει ένα μη-ορισμένο στοιχείο (u) που υπάρχει σε ένα Σ-ΤΠ ενός Σ-ΒΓΜ με τη συγκεκριμένη οντότητα που υπάρχει στο ΤΜ ενός ΒΓΜ. Δεδομένου ότι το πεδίο τιμών της συνάρτησης περιλαμβάνει URIs, Λεκτικά και Μεταβλητές, οι τιμές είναι οι ίδιες, εάν ανήκουν στον ίδιο τύπο και αποτελούνται από την ίδια ακολουθία χαρακτήρων. Δύο συνδέσεις b1 και b2 είναι συμβατές μεταξύ τους, όταν για όλα τα μη-ορισμένα στοιχεία

(u) που ανήκουν στα σύνολα dom(b1) και dom(b2) ισχύει b1(u) = b2(u), όπου dom(b) είναι το υποσύνολο των μη-ορισμένων στοιχείων για τα οποία η συνάρτηση b ορίζεται. Ένα Σ-ΒΓΜ ταιριάζει με τα ΤΠ που υπάρχουν σε ένα ΒΓΜ ενός SPARQL ερωτήματος, όταν όλα τα Σ-ΤΠ ταιριάζουν και οι συνδέσεις που έλαβαν χώρα για τις μη-δεσμευμένες οντότητες είναι συμβατές μεταξύ τους. Για το σκοπό αυτό, εξετάζουμε εάν ταιριάζουν τα Σ-ΤΠ που έχουν οριστεί ένα προς ένα (η σειρά δεν είναι σημαντική), άλλα κάθε φορά λαμβάνουμε υπόψη τις συνδέσεις που έχουν πραγματοποιηθεί στο προηγούμενο στάδιο. Σημειώνουμε ότι ένα Σ-ΒΓΜ μπορεί να υπάρχει στο ΒΓΜ ενός SPARQL ερωτήματος παραπάνω από μία φορά.

Στην περίπτωση που μία ή περισσότερες συνθήκες πρέπει να πληρούνται, θα πρέπει να εντοπίζουμε τα αντίστοιχα δεδομένα (συχνά ορίζονται στο FILTER τμήμα ενός ερωτήματος) και κατόπιν να εξετάσουμε εάν αυτά ικανοποιούν τις συνθήκες ή όχι. Κατά την παραπάνω διαδικασία λαμβάνουμε υπόψη τις συνδέσεις που έχουν ήδη γίνει κατά την «προβολή» ενός Σ-ΒΓΜ στο ΒΓΜ του SPARQL ερωτήματος. Πιο συγκεκριμένα, εντοπίζουμε τις συγκεκριμένες οντότητες (συνήθως μεταβλητές) που σχετίζονται με τις συνθήκες που έχουν οριστεί σε έναν ΚΜ και έπειτα αναζητούμε τα δεδομένα που παρέχονται από το χρήστη στο FILTER τμήμα του SPARQL ερωτήματος. Τέλος, εξετάζουμε αν αυτά ικανοποιούν τις συνθήκες που ορίστηκαν.

Εάν και οι δύο παραπάνω προϋποθέσεις σχετικά με το αριστερό Σ-ΒΓΜ ενός κανόνα και τις συνθήκες που έχουν οριστεί πληρούνται, ο κανόνας αυτός μπορεί να πυροδοτηθεί (όπως ήδη αναφέραμε, ίσως και παραπάνω από μία φορά). Στην περίπτωση αυτή, το Σ-ΒΓΜ θα πρέπει να αντικατασταθεί από το αντίστοιχο Σ-ΒΓΜ που υπάρχει στο δεξιό μέρος του κανόνα, λαμβάνοντας υπόψη τις συνδέσεις που έλαβαν χώρα. Πιο συγκεκριμένα, οι δεσμεύσεις που έγιναν για τις μη-δεσμευμένες οντότητες του αριστερού Σ-ΒΓΜ ισχύουν και για το δεξιό μέρος. Σημειώνουμε ότι, στην περίπτωση που υπάρχει

στο δεξιό μέρος κάποια μη-δεσμευμένη οντότητα που δε χρησιμοποιείται στο αριστερό μέρος, αυτή θα πρέπει να αντικατασταθεί από μια νέα μεταβλητή. Επίσης, στην περίπτωση που χρειάζεται να γίνουν κάποιες αλλαγές στα δεδομένα (δηλαδή έχει οριστεί ο «ευθύς» μετατροπέας δεδομένων), θα πρέπει να εντοπίσουμε τα δεδομένα αυτά και να κάνουμε τις απαραίτητες αλλαγές. Και στην περίπτωση αυτή ισχύουν οι προηγούμενες δεσμεύσεις, ώστε να εντοπίσουμε αρχικά τις σχετικές μεταβλητές και τα αντίστοιχα δεδομένα (συχνά παρέχονται μέσα στο FILTER τμήμα, όπως και οι συνθήκες) και να βρούμε τις νέες τιμές.

Η σειρά με την οποία θα εφαρμοστούν οι SPARQL-KM δεν είναι τόσο σημαντική, εφόσον δεν υπάρχει κάποια επικάλυψη μεταξύ των κανόνων. Για το λόγο αυτό, εξετάζουμε τα TP του SPARQL ερωτήματος καθώς και τους κανόνες που μπορούν να πυροδοτηθούν, για να βρούμε τους κανόνες που επιδρούν πάνω σε αυτά. Στην περίπτωση που παραπάνω από ένας κανόνες χρειάζεται να δράσουν πάνω στα ίδια ΤΟ, τότε εκτελείται πρώτος ο κανόνας που περιλαμβάνει το μεγαλύτερο αριθμό από Σ-ΤΠ στο αριστερό Σ-ΒΓΜ. Επίσης, προτεραιότητα δίνεται σε αυτούς τους ΚΜ στους οποίους τα δεδομένα θα πρέπει να πληρούν κάποιες συνθήκες. Ο λόγος για τις παραπάνω υποθέσεις κατά τη διαδικασία της αναδιατύπωσης είναι ότι στις περιπτώσεις αυτές οι κανόνες είναι πιο συγκεκριμένοι από τους «ανταγωνιστές» τους και επομένως θα πρέπει να εκτελεστούν πρώτοι. Σημειώνουμε ότι, στην περίπτωση που ένα ΤΠ χρησιμοποιείται σε παραπάνω από έναν κανόνες, αυτό θα μετακινηθεί από το ερώτημα μετά την εκτέλεση και του τελευταίου κανόνα που εντοπίσαμε ότι μπορεί να πυροδοτηθεί με βάση αυτό.

Μετά την εφαρμογή των κανόνων, το αρχικό SPARQL ερώτημα θα έχει εκφραστεί με βάση τους όρους που χρησιμοποιούνται στην άλλη μεριά. Δεδομένου ότι η συσχέτιση μεταξύ των όρων των δύο οντολογιών δεν μπορεί να είναι πάντα πλήρης, κάποια από τα TP και κάποιες από τις συνθήκες μπορεί να μην είναι δυνατόν να

αντικατασταθούν από τα αντίστοιχα ΤΠ και τις αντίστοιχες συνθήκες. Στην περίπτωση αυτή, αυτά θα πρέπει να μετακινηθούν από το αναδιατυπωμένο ερώτημα πριν την εκτέλεσή του. Σημειώνουμε ότι η πληροφορία και οι συνθήκες που αγνοήσαμε στο παραπάνω στάδιο θα πρέπει να αναφερθούν στο χρήστη.

12.4.2 Αναδιατύπωση RDF Δεδομένων

Η αναδιατύπωση των RDF δεδομένων βασίζεται στους RDF-KM που έχουν παραχθεί. Όμως, στην περίπτωση αυτή ακολουθείται μια ελαφρώς διαφορετική προσέγγιση. Πιο συγκεκριμένα, αρχικά, εξετάζουμε τους κανόνες που μπορούν να εφαρμοστούν καθώς και τη σειρά αυτών, όπως και στην προηγούμενη περίπτωση. Έπειτα, χρησιμοποιούμε τους ΚΜ, για να αναδιατυπώσουμε τμήματα του γράφου, ενώ ο συνολικά αναδιατυπωμένος γράφος με τα RDF δεδομένα προκύπτει από τη σύνθεση των επιμέρους γράφων.

Η παραγωγή των «ενδιάμεσα» αναδιατυπωμένων τμημάτων του RDF γράφου που λαμβάνουμε από την εκτέλεση ενός CONSTRUCT SPARQL ερωτήματος (και τα δύο τμήματα του ερωτήματος είναι εκφρασμένα με βάση όρους που ορίζονται στην οντολογία της βάσης) βασίζεται στην εκτέλεση των RDF-KM, οι οποίοι βασίζονται στην εκτέλεση ενός νέου CONSTRUCT SPARQL ερωτήματος στον RDF γράφο που λαμβάνουμε, το οποίο με τη σειρά του αναλαμβάνει να αναδιατυπώσει ένα τμήμα του γράφου με βάση τον αντίστοιχο ΚΣ. Το construct τμήμα του ερωτήματος αυτού προκύπτει με βάση το αριστερό Σ-ΒΓΜ του κανόνα, ενώ το where τμήμα με βάση το δεξιό μέρος. Κατά την παραπάνω διαδικασία οι μη-δεσμευμένες οντότητες που υπάρχουν στο δεξιό μέρος του κανόνα αντικαθίστανται από μεταβλητές στο where τμήμα του construct SPARQL ερωτήματος. Σημειώνουμε, επίσης, ότι οι αντικαταστάσεις που έλαβαν χώρα ισχύουν και για τις μη-δεσμευμένες οντότητες με το ίδιο όνομα που υπάρχουν στο αριστερό μέρος του κανόνα, βάσει του οποίου προέκυψε το construct τμήμα. Επισημαίνουμε ότι, στην

περίπτωση που υπάρχει κάποια μη-δεσμευμένη οντότητα η οποία δεν αναφέρεται στο δεξιό μέρος του κανόνα, αυτή αντικαθίσταται από έναν κενό κόμβο (blank node).

Στην περίπτωση που το αποτέλεσμα της εκτέλεσης του παραπάνω ερωτήματος δεν είναι ένας κενός γράφος, τα δεδομένα θα πρέπει να πληρούν τις συνθήκες που βρίσκονται στο δεξιό μέρος του κανόνα (εάν υπάρχουν). Επίσης, στην περίπτωση που έχει οριστεί ένας μετατροπέας τιμών (για την ακρίβεια, ο «αντίστροφος» μετατροπέας τιμών), θα πρέπει να χρησιμοποιηθεί, για να λάβουμε τις νέες τιμές. Ο συνολικός αναδιατυπωμένος γράφος με τα αποτελέσματα της εκτέλεσης του αρχικού ερωτήματος προκύπτει από τη σύνθεση των ενδιάμεσων αποτελεσμάτων που προέκυψαν από την εφαρμογή των κανόνων που μπορούν να πυροδοτηθούν (πρακτικά αυτών που δημιούργησαν έναν μη κενό γράφο που ικανοποιεί τις συνθήκες).

Σημειώνουμε ότι τα δεδομένα, που δεν μπόρεσαν να αναδιατυπωθούν, θα πρέπει να αναφέρονται στο χρήστη. Επίσης, η σειρά εκτέλεσης των κανόνων θα πρέπει να καθοριστεί, εάν αυτό είναι απαραίτητο.

13 *Εφαρμογή Εργαλείων και Μηχανισμών*

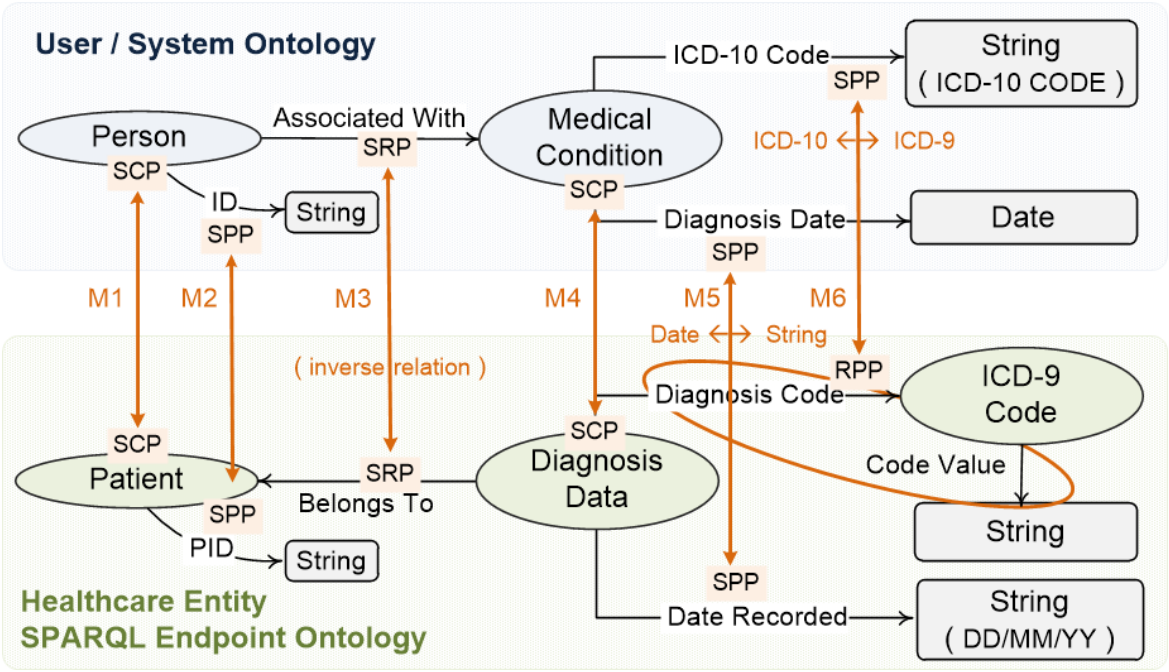
Στο κεφάλαιο αυτό θα δείξουμε πώς χρησιμοποιήθηκαν τα μοντέλα και οι μηχανισμοί που αναπτύχθηκαν, για να βρούμε τους ασθενείς που πληρούν τα κριτήρια του χρήστη σε μία συγκεκριμένη μονάδα περίθαλψης. Στο σημείο αυτό, υποθέτουμε ότι τα Κριτήρια Καταλληλότητας (ΚΚ) μιας κλινικής δοκιμής έχουν ήδη καθοριστεί και εσωτερικά αναπαρασταθεί, χρησιμοποιώντας ένα συνδυασμό από XML στοιχεία (για την οργάνωση των κριτηρίων και των παραμέτρων τους) και SPARQL ερωτήματα (για την τυπική έκφραση των κριτηρίων), με βάση τα μοντέλα που έχουν ήδη αναπτυχθεί για το σκοπό αυτό. Στο Ευρωπαϊκό έργο ΡΟΝΤΕ τα ΚΚ χρησιμοποιήθηκαν για την εύρεση των ασθενών που πληρούν τα κριτήρια που είχε ορίσει ο χρήστης μέσω SPARQL ερωτημάτων, τα οποία προέκυπταν αυτόματα με βάση την τυπική αναπαράσταση των κριτηρίων καταλληλότητας και ακολούθως χρησιμοποιούνταν για την εύρεση των ασθενών από μία ή περισσότερες μονάδες περίθαλψης.

Στη συνέχεια, θα δείξουμε πώς χρησιμοποιήθηκαν τα εργαλεία και οι μηχανισμοί για την εύρεση των ασθενών από μία μόνο μονάδα περίθαλψης. Σημειώνουμε ότι, το μοντέλο που χρησιμοποιείται στη μονάδα περίθαλψης που παρουσιάζουμε για την καταγραφή των δεδομένων των ασθενών, έχει σημαντικές δομικές και σημασιολογικές διαφορές με το μοντέλο που αναπτύχθηκε για την έκφραση των παραμέτρων των κριτηρίων καταλληλότητας (ΚΚ) μιας κλινικής δοκιμής. Για το λόγο αυτό, αρχικά,

χρησιμοποιήθηκε το εργαλείο που αναπτύχθηκε για τον καθορισμό της συσχέτισης μεταξύ των όρων των δύο αυτών μοντέλων και ακολούθως, χρησιμοποιήθηκε ο μηχανισμός που παρουσιάσαμε για την εύρεση των ασθενών που πληρούν τα ΚΚ.

13.1 Καθορισμός Κανόνων Συσχέτισης

Στο Σχήμα 17 υπάρχει η αντιστοίχιση των όρων ενός τμήματος του μοντέλου που χρησιμοποιήθηκε για την έκφραση των κριτηρίων και κατά συνέπεια των ερωτημάτων μας, με τους αντίστοιχους όρους του μοντέλου που χρησιμοποιήθηκε στη μονάδα περίθαλψης για την αποθήκευση των δεδομένων των ασθενών. Στην εικόνα αυτή φαίνονται, επίσης, οι κανόνες συσχέτισης που έχουν οριστεί, τα οντολογικά στοιχεία που συμμετέχουν σε κάθε κανόνα, τα οντολογικά πρότυπα που έχουν χρησιμοποιηθεί καθώς και οι συναρτήσεις μετατροπής τιμής, όπου αυτό ήταν απαραίτητο.



Σχήμα 17: Κανόνες Συσχέτισης μεταξύ των όρων των οντολογιών του Χρήστη για την έκφραση των Κριτηρίων Καταλληλότητας και της Μονάδας Περίθαλψης για την αποθήκευση των δεδομένων των Ασθενών.

Οι κανόνες συσχέτισης που έχουν οριστεί είναι έγκυροι και για τις δύο κατευθύνσεις. Επίσης, η αντιστοίχιση των όρων θεωρήθηκε ότι είναι ισοδύναμη, δεδομένου του σκοπού που εξυπηρετούν, όπως καταγράφηκε κατά τον καθορισμό της συσχέτισης μεταξύ των όρων των δύο αυτών οντολογιών. Για παράδειγμα, για την εύρεση των «Ανθρώπων» που ικανοποιούν τα ΚΚ, θα πρέπει να αναζητήσουμε τους «Ασθενείς» που υπάρχουν στη μονάδα περίθαλψης, παρά το γεγονός ότι οι όροι αυτοί, στη γενική τους περίπτωση, δεν είναι ισοδύναμοι.

13.2 Αναδιατύπωση Ερωτημάτων και Αποτελεσμάτων

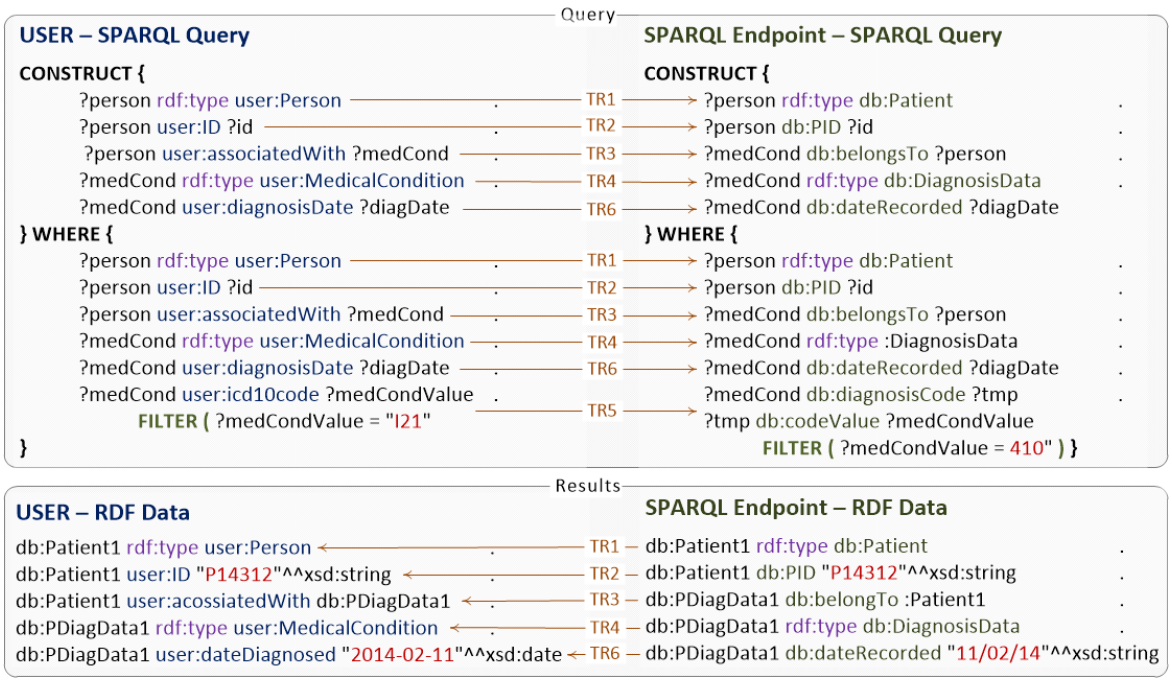
Στον Πίνακας 7 υπάρχουν οι κανόνες μετατροπής (ΚΜ) που προκύπτουν αυτόματα από το σύστημα, με βάση τους κανόνες συσχέτισης (ΚΣ) που έχουν οριστεί. Όπως μπορούμε να δούμε στον πίνακα αυτό, κατά τη μετάβαση από το ένα μοντέλο στο άλλο, απαιτείται μετατροπή των κωδικών που αναφέρονται στις ιατρικές καταστάσεις του ασθενούς, καθώς χρησιμοποιούνται διαφορετικά συστήματα κωδικοποίησης. Πιο συγκεκριμένα, τα ΚΚ είναι εκφρασμένα με βάση τους όρους της ICD-10 κωδικοποίησης, ενώ οι όροι που υπάρχουν στη συγκεκριμένη μονάδα περίθαλψης προέρχονται από την ICD-9 κωδικοποίηση. Συνεπώς, κατά την επανέκφραση των ερωτημάτων (περιλαμβάνουν και τα ΚΚ) είναι επίσης απαραίτητη η εύρεση των αντίστοιχων ICD-9 κωδικών των ιατρικών καταστάσεων που αναφέρονται. Επίσης, δεδομένου ότι η ημερομηνία διάγνωσης μιας κατάστασης είναι διαφορετικά εκφρασμένη σε καθένα από τα δύο μοντέλα, θα πρέπει να γίνουν οι απαραίτητες μετατροπές στα αντίστοιχα δεδομένα που λαμβάνουμε από τη μονάδα περίθαλψης έτσι, ώστε ο τύπος και η μορφή των δεδομένων να ταιριάζουν με αυτά του μοντέλου του χρήστη.

ID	**ΚΣ**	**ΚΜ: Αριστερό Τμήμα**	**ΚΜ: Δεξιό Τμήμα**	**Μετατροπή**
TR1	M1	[sth] rdf:type usr:Person	[sth] rdf:type db:Patient	-
TR2	M2	[sth1] usr:ID [sth2]	[sth1] db:PID [sth2]	-

TR3	M3	[sth1] usr:associatedWith [sth2]	[sth2] db:belongsTo [sth1]	-
TR4	M4	[sth] rdf:type usr:MedicalCondition	[sth] rdf:type db:DiagnosisData	-
TR5	M5	[sth1] user:icd10code [sth2]	[sth1] db:diagnosisCode [tmp] . [tmp] db:codeValue [sth2]	[sth2]:ICD-10 to/from ICD-9
TR6	M6	[sth1] user:diagnosisDate [sth2]	[sth1] user:dateRecorded [sth2]	[sth2]:Date to/from String

Πίνακας 7: Αυτομάτως παραγόμενοι Κανόνες Μετάβασης, με βάση τους Κανόνες Συσχέτισης που έχουν οριστεί.

Το σύστημα χρησιμοποιεί τους κανόνες μετατροπής που έχουν αυτόματα παραχθεί για την επανέκφραση ενός ερωτήματος για την εύρεση των ασθενών που διαγνώστηκαν με οξύ έμφραγμα του μυοκαρδίου (ICD-10 κωδικός: I21), καθώς επίσης και της ημερομηνίας που έγινε η διάγνωση. Στο Σχήμα 18 παρουσιάζεται το αρχικό και αναδιατυπωμένο ερώτημα του χρήστη (στο παράδειγμα αυτό, το ερώτημα περιλαμβάνει ένα μόνο ΚΚ), καθώς και οι κανόνες μετατροπής που χρησιμοποιήθηκαν. Σημειώνουμε ότι η σειρά με την οποία εφαρμόζονται οι κανόνες αυτοί δεν είναι σημαντική, καθώς δεν υπάρχει κάποια επικάλυψη μεταξύ τους. Επίσης, σημειώνουμε ότι ορισμένοι κανόνες εφαρμόζονται παραπάνω από μία φορές, όπως για παράδειγμα ο κανόνας μετατροπής TR1.



Σχήμα 18: Αρχικό και Επανεκφρασμένο (α) SPARQL ερώτημα και (β) RDF δεδομένα καθώς και οι Κανόνες Μετάβασης που έχουν χρησιμοποιηθεί.

Δεδομένου ότι τόσο το τμήμα προβολής (construct clause) όσο και το τμήμα περιορισμού (restriction clause) του αρχικού ερωτήματος έχουν αναδιατυπωθεί, ώστε να είναι εκφρασμένα με βάση το μοντέλο και τους όρους που υποστηρίζονται από τη μονάδα περίθαλψης, τα αποτελέσματα που λαμβάνονται από την εκτέλεση του αναδιατυπωμένου ερωτήματος είναι εκφρασμένα με βάση το μοντέλο και τις κωδικοποιήσεις της συγκεκριμένης μονάδας περίθαλψης. Για το λόγο αυτό, το σύστημα χρησιμοποιεί τους κανόνες μετατροπής για την επανέκφραση των δεδομένων έτσι, ώστε αυτά να είναι συμβατά με το μοντέλο και τις κωδικοποιήσεις του χρήστη, όπως φαίνεται στο Σχήμα 18. Στο σχήμα αυτό παρουσιάζουμε τα αρχικά και αναδιατυπωμένα δεδομένα σε Turtle μορφοποίηση [111] για ένα μόνο ασθενή, λόγω περιορισμού χώρου.

Η σελίδα αυτή είναι σκόπιμα λευκή

14 *Σχετική Συζήτηση*

Για τον εντοπισμό των δεδομένων που πληρούν τα κριτήρια του χρήστη, γενικά, μπορούν να χρησιμοποιηθούν είτε SELECT είτε CONSTRUCT SPARQL ερωτήματα. Ωστόσο, στην περίπτωση που τα δεδομένα που ψάχνουμε προέρχονται από κάποιο σύστημα κωδικοποίησης (στη γενική περίπτωση, διαφορετικό από αυτό που χρησιμοποιείται στη βάση), η χρησιμοποίηση CONSTRUCT SPARQL ερωτημάτων διευκολύνει την περαιτέρω επεξεργασία των δεδομένων της βάσης, καθώς το αποτέλεσμα της εκτέλεσης του ερωτήματος αυτού είναι ένας RDF γράφος, ο οποίος παρέχει μια λεπτομερή και ακριβή περιγραφή των δεδομένων που ικανοποιούν τις συνθήκες που έχουν οριστεί με βάση το μοντέλο και τις κωδικοποιήσεις που υπάρχουν στη βάση, και συνεπώς επιδέχονται περαιτέρω επεξεργασία. Σημειώνουμε ότι η ίδια πληροφορία θα μπορούσε επίσης να βρεθεί με τη χρήση SELECT SPARQL ερωτημάτων, αλλά για τη σωστή ερμηνεία των δεδομένων που λαμβάνουμε (πίνακας) θα πρέπει να λάβουμε υπόψη όχι μόνο τα δεδομένα αυτά αλλά και το αρχικό ερώτημα. Πιο συγκεκριμένα, θα πρέπει να εξετάσουμε το WHERE τμήμα του SPARQL ερωτήματος, για να εντοπίσουμε τη σημασία των μεταβλητών που υπάρχουν στο SELECT τμήμα. Ακολούθως, τα δεδομένα (συμβολοακολουθίες) που λαμβάνουμε θα πρέπει να εξεταστούν και πιθανώς να αλλαχθούν, λαμβάνοντας υπόψη τη σημασία των μεταβλητών και τις κωδικοποιήσεις που υποστηρίζονται από το χρήστη.

Η ορολογία που χρησιμοποιείται από το χρήστη για την έκφραση των ερωτημάτων του είναι στενά συνδεδεμένη με τους Κανόνες Συσχέτισης που έχουν οριστεί, καθώς επίσης και τους Κανόνες Μετάβασης (ΚΜ) που προκύπτουν. Πιο συγκεκριμένα, όταν το μοντέλο και οι κωδικοποιήσεις που χρησιμοποιούνται έχουν σημαντικές δομικές και σημασιολογικές διαφορές με το μοντέλο και τους όρους της βάσης, η πλήρης και ακριβής περιγραφή των συσχετίσεων μεταξύ των όρων αυτών δεν είναι εφικτή, με άμεση επίπτωση στην αποτίμηση των SPARQL ερωτημάτων του χρήστη, όπως, για παράδειγμα, την παράβλεψη ορισμένων τριάδων (triple patterns) και των συνθηκών ή την αντικατάσταση ενός όρου με τον πιο κατάλληλο, λόγω της μη ύπαρξης ενός σημασιολογικά ισοδύναμου όρου στο σύστημα κωδικοποίησης της βάσης δεδομένων. Στις παραπάνω περιπτώσεις, οι σημασιολογικές διαφορές, που προκύπτουν κατά την επανέκφραση του αρχικού ερωτήματος με βάση τους όρους της βάσης, θα πρέπει να αναφέρονται στο χρήστη και να ληφθούν σοβαρά υπόψη για τη σωστή αποτίμηση των αποτελεσμάτων που επιστρέφονται από την εκτέλεση του αναδιατυπωμένου ερωτήματος.

Εάν υποθέσουμε ότι α) όλοι οι όροι του μοντέλου αναφοράς που χρησιμοποιούνται από το χρήστη έχουν διασυνδεθεί με σημασιολογικά ισοδύναμους όρους του μοντέλου της βάσης δεδομένων και β) τα συστήματα κωδικοποίησης που χρησιμοποιούνται από το χρήστη είναι είτε ίδια με αυτά της βάσης είτε υπάρχει ένα προς ένα αντιστοίχιση μεταξύ των όρων τους, τότε η σημασία του αρχικού SPARQL ερωτήματος [112] δεν αλλάζει μετά την εφαρμογή των ΚΜ, καθώς η γενικότερη δομή του ερωτήματος παραμένει η ίδια, ενώ οι τριάδες και τα φίλτρα αντικαθίστανται από σημασιολογικά ισοδύναμες τριάδες και φίλτρα αντίστοιχα.

Εάν τα συστήματα κωδικοποίησης του χρήστη είναι διαφορετικά από αυτά της βάσης, γενικά, δεν υπάρχει ένα προς ένα αντιστοίχιση μεταξύ των όρων. Για παράδειγμα, ένας όρος από το σύστημα κωδικοποίησης του χρήστη μπορεί να είναι ισοδύναμος με την

ένωση δύο όρων του συστήματος κωδικοποίησης της βάσης δεδομένων. Κατά τη διαδικασία επανέκφρασης ενός SPARQL ερωτήματος, το σύστημα πραγματοποιεί τις απαιτούμενες αλλαγές στα φίλτρα του ερωτήματος, αντικαθιστώντας τις αρχικές συνθήκες με σημασιολογικά ισοδύναμες λογικές εκφράσεις χρησιμοποιώντας το λογικό τελεστή «OR». Επίσης, για να είναι τα δεδομένα που λαμβάνουμε από τη βάση κατανοητά από το χρήστη (δηλαδή, εκφρασμένα με βάση το μοντέλο αναφοράς και τα συστήματα κωδικοποίησης που υποστηρίζονται από το χρήστη), θα πρέπει να είναι εφικτή όχι μόνο η επανέκφραση των ερωτημάτων του χρήστη αλλά και των δεδομένων που λαμβάνουμε από τη βάση. Για το λόγο αυτό, το μοντέλο αναφοράς του χρήστη θα πρέπει να σχεδιαστεί προσεκτικά έτσι, ώστε οι τιμές των παραμέτρων να εκφράζονται είτε μέσω κάποιου υπάρχοντος όρου από το σύστημα κωδικοποίησης είτε μέσω ενός συνδυασμού αυτών.

Σχετικά με το μηχανισμό αναδιατύπωσης SPARQL ερωτημάτων, στην εργασία αυτή, έμφαση δόθηκε στους ΚΜ οι οποίοι δεν αλλάζουν τη δομή του ερωτήματος. Ωστόσο, ορισμένες φορές, μπορεί να είναι απαραίτητη η εισαγωγή (ή διαγραφή) ενός προαιρετικού τμήματος. Επίσης, στην περίπτωση που χρειάζεται κάποια αλλαγή στις τιμές των δεδομένων, οι εκφράσεις που υπάρχουν στα φίλτρα ενός SPARQL ερωτήματος θα πρέπει να εξεταστούν προσεκτικά, καθώς οι εκφράσεις που υπάρχουν μπορεί να μην είναι πλέον έγκυρες μετά την αλλαγή, ειδικότερα όταν αλλάζει ο τύπος δεδομένων (π.χ., string σε integer, ή αντίστροφα). Επιπλέον, σε ορισμένες περιπτώσεις, η τιμή μιας παραμέτρου στο ένα μοντέλο εξαρτάται από την τιμή περισσοτέρων από μία παραμέτρων στο άλλο. Για παράδειγμα, η ημερομηνία γέννησης συνδέεται με την ηλικία ενός ασθενούς καθώς επίσης και την ημερομηνία καταγραφής της, και επομένως, η τιμή και των δύο αυτών παραμέτρων θα πρέπει να ληφθεί υπόψη κατά τη μετάβαση από το ένα μοντέλο στο άλλο.

Τα παραπάνω δείχνουν ότι, σε ορισμένες περιπτώσεις, η συσχέτιση των όρων δύο μοντέλων καθώς και η περαιτέρω επεξεργασία τους μπορεί να είναι δύσκολη, αν όχι αδύνατη, εξαιτίας των σημαντικών δομικών και σημασιολογικών διαφορών μεταξύ των όρων τους. Για το λόγο αυτό, χρειάζεται να δοθεί ιδιαίτερη προσοχή κατά το σχεδιασμό των μοντέλων που θα χρησιμοποιηθούν για την αποθήκευση δεδομένων καθώς και να ληφθούν υπόψη υπάρχοντα μοντέλα και κωδικοποιήσεις έτσι, ώστε να περιορίσουν τις πιθανές ασυνέπειες που υπάρχουν μεταξύ τους.

15 *Συμπεράσματα*

Οι οντολογίες έχουν ένα διακριτό ρόλο στην πρόσβαση των δεδομένων μιας σχεσιακής βάσης χρησιμοποιώντας τις τεχνολογίες του σημασιολογικού ιστού, ενώ συχνά παρουσιάζουν δομικές και σημασιολογικές διαφορές με τα αντίστοιχα μοντέλα και όρους που υποστηρίζονται από μία βάση δεδομένων. Στην ενότητα αυτή παρουσιάσαμε ένα καινοτόμο σύστημα που βασίζεται σε οντολογικά πρότυπα (ΟΠ) και συναρτήσεις μετατροπής, για να γεφυρώσουμε το χάσμα που υπάρχει μεταξύ του χρήστη και της βάσης δεδομένων μέσω μιας ημιαυτόματης διαδικασίας. Η προσέγγιση που ακολουθήθηκε, σε συνδυασμό με τα εργαλεία που αναπτύχθηκαν, επιτρέπει τον αποτελεσματικό χειρισμό ενός μεγάλου εύρους αναντιστοιχιών μεταξύ των μοντέλων του χρήστη και της βάσης, και συνεπώς μπορούν να συμβάλλουν στη διασύνδεση των κριτηρίων καταλληλότητας που έχουν εκφραστεί μέσω ενός κοινού (ανεξαρτήτου βάσης) μοντέλου που κατασκευάστηκε με τις βάσεις των ασθενών.

Η σελίδα αυτή είναι σκόπιμα λευκή

Γ. ΣΥΣΤΗΜΑ ΟΠΤΙΚΗΣ ΕΚΦΡΑΣΗΣ ΕΡΩΤΗΜΑΤΩΝ

Στην ενότητα αυτή παρουσιάζουμε ένα εργαλείο που αναπτύχθηκε για την έκφραση περίπλοκων ερωτημάτων, μέσω ενός ιδιαίτερα αλληλεπιδραστικού, φιλικού προς το χρήστη, γραφικού περιβάλλοντος. Το εργαλείο αυτό βασίζεται αποκλειστικά και μόνο στην οντολογική αναπαράσταση ενός συγκεκριμένου πεδίου γνώσης, επιτρέποντας στους χρήστες να εκφράσουν συντακτικά και σημασιολογικά σωστά SPARQL ερωτήματα, χωρίς να είναι γνώστες των τεχνολογιών του σημασιολογικού ιστού.

Στην ενότητα αυτή, αρχικά, μελετήθηκαν τα υπάρχοντα εργαλεία για την Οπτική έκφραση SPARQL ερωτημάτων καθώς επίσης και οι τεχνολογίες που χρησιμοποιούνται σε καθένα από αυτά. Όπως έδειξε η ανάλυσή μας, υπάρχουν αρκετά εργαλεία / συστήματα για το σκοπό αυτό, ωστόσο, είναι ιδιαίτερα περίπλοκα και απαιτούν σημαντική προσπάθεια από τη μεριά του χρήστη ακόμη και για την έκφραση ενός απλού ερωτήματος. Επιπρόσθετα, τα περισσότερα από αυτά απευθύνονται σε ανθρώπους που γνωρίζουν ικανοποιητικά τις τεχνολογίες του σημασιολογικού ιστού, γεγονός που περιορίζει αισθητά το εύρος των χρηστών τους.

Το εργαλείο που αναπτύχθηκε επιτρέπει στους χρήστες να εκφράσουν εύκολα και γρήγορα τα επιθυμητά ερωτήματα, στα οποία τα δεδομένα προς αναζήτηση μπορούν να ικανοποιούν αρκετά κριτήρια, ενώ οι αλγόριθμοι που υλοποιήθηκαν παράγουν αυτόματα το αντίστοιχο SPARQL ερώτημα, λαμβάνοντας υπόψη τη σημασιολογία των όρων της βάσης. Η αρχιτεκτονική του συστήματος καθώς επίσης και η αλληλεπίδραση του χρήστη με αυτό και οι αλγόριθμοι που χρησιμοποιούνται παρουσιάζονται αναλυτικά στην ενότητα αυτή.

Το εργαλείο που παρουσιάζεται στην ενότητα αυτή θα μπορούσε να χρησιμοποιηθεί σε πολλές διαφορετικές εφαρμογές, συμπεριλαμβανομένης της τυπικής

έκφρασης των κριτηρίων καταλληλότητας (ΚΚ) μιας κλινικής μελέτης, ενώ θα μπορούσε να συνδυαστεί με υπάρχοντα εργαλεία και μηχανισμούς για την περαιτέρω επεξεργασία των αυτομάτως παραγόμενων ερωτημάτων.

16 *Εισαγωγή*

Οι οντολογίες επιτρέπουν στους χρήστες να περιγράψουν ένα πεδίο της γνώσης με τη μορφή των κλάσεων, παραμέτρων και σχέσεων [113] [114]. Οι ετικέτες (labels) των όρων που έχουν συμπεριληφθεί στην οντολογία και κυρίως η περιγραφή τους επιτρέπουν στους χρήστες να καταλάβουν τη σημασία των όρων, καθώς επίσης και να εντοπίσουν περισσότερες πληροφορίες για τον καθέναν από αυτούς, με βάση τις μεταξύ τους συσχετίσεις ή γενικότερα τα αξιώματα που έχουν οριστεί. Τα αξιώματα έχουν οριστεί από την W3C ομάδα [115] σε μία μορφή κατανοητή από τον υπολογιστή, γεγονός που επιτρέπει στα λογισμικά συστήματα να επεξεργαστούν περαιτέρω τα δεδομένα που περιέχουν οι οντολογίες και να εξάγουν επιπλέον συσχετίσεις για καθέναν από τους όρους, ακόμη και αν δεν έχουν οριστεί ρητά στην οντολογία.

Οι οντολογίες εστιάζουν στις έννοιες που χρησιμοποιούνται σε ένα συγκεκριμένο πεδίο της γνώσης και κατά συνέπεια μπορούν να συμβάλλουν σημαντικά στη γεμάτη νόημα επικοινωνία του χρήστη με τα αντίστοιχα συστήματα, με την προϋπόθεση ότι τα δεδομένα τους βρίσκονται σε RDF βάσεις. Όμως, αρκετά δεδομένα εξακολουθούν να βρίσκονται σε σχεσιακές βάσεις δεδομένων. Τα πρόσφατα επιτεύγματα στον τομέα του Σημειολογικού Ιστού και κυρίως η ανάπτυξη SPARQL επεξεργαστών, όπως ο D2R εξυπηρετητής [74] και το Ontop εργαλείο [75], αποτελούν ένα σημαντικό βήμα για την ένταξη των σχεσιακών βάσεων στο σημασιολογικό ιστό. Όμως, για να μπορούν οι

χρήστες να έχουν πρόσβαση στα δεδομένα τους, θα πρέπει να γνωρίζουν καλά τις τεχνολογίες του Σημασιολογικού Ιστού και κυρίως OWL και SPARQL. Επιπρόσθετα, θα πρέπει να αφιερώσουν αρκετό χρόνο, για να καταλάβουν τη δομή των δεδομένων και τους όρους που χρησιμοποιούνται, μια διαδικασία ιδιαίτερα δύσκολη (αν όχι αδύνατη), καθώς η οντολογική αναπαράσταση των στοιχείων της σχεσιακής βάσης που υποστηρίζεται από τα παραπάνω εργαλεία (βάσει των οποίων γίνεται η έκφραση των ερωτημάτων) επηρεάζεται συχνά από τις αποφάσεις που πάρθηκαν κατά το σχεδιασμό των βάσεων δεδομένων και τις υποθέσεις που έγιναν [101], ενώ πολλές φορές η περιγραφή των όρων που χρησιμοποιούνται δεν είναι διαθέσιμη. Κατά συνέπεια, είναι απαραίτητη αφενός μεν η σημασιολογική περιγραφή των οντολογικών στοιχείων, όπως έχει ήδη περιγραφεί στο κεφάλαιο 10.3.1, αφετέρου δε η ανάπτυξη καινοτόμων εργαλείων και εφαρμογών που επιτρέπουν στους χρήστες να εκφράσουν γραφικά τα επιθυμητά ερωτήματα, ακόμη και αν δε διαθέτουν κάποιο τεχνολογικό υπόβαθρο.

Η Οπτική Έκφραση Ερωτημάτων (Visual Query Formulation) έχει μελετηθεί εκτενώς στο παρελθόν [116]. Όμως, η δημιουργία συστημάτων και εργαλείων που επιτρέπουν στους χρήστες να εκφράσουν SPARQL ερωτήματα, λαμβάνοντας υπόψη την οντολογική αναπαράσταση των δεδομένων, είναι ένα σχετικά νέο θέμα [117], το οποίο έλαβε ιδιαίτερη προσοχή την τελευταία δεκαετία για δύο κυρίως λόγους: α) η SPARQL έγινε W3C σύσταση για την ερώτηση RDF πηγών δεδομένων από τον Ιανουάριο του 2008 και β) αρκετά συστήματα και εργαλεία αναπτύχθηκαν για τη διευκόλυνση της πρόσβασης σε βάσεις δεδομένων χρησιμοποιώντας τις τεχνολογίες του σημασιολογικού ιστού [118]. Τα παραπάνω συνετέλεσαν στη δημιουργία αρκετών συστημάτων για την Οπτική Έκφραση SPARQL ερωτημάτων, τα οποία περιγράφονται αναλυτικά στην επόμενη ενότητα. Όμως, όπως θα δούμε, τα υπάρχοντα συστήματα δεν καλύπτουν επαρκώς τις ανάγκες του χρήστη, καθώς αρκετά είναι πολύ περίπλοκα, γεγονός που

περιορίζει το εύρος των πιθανών χρηστών, ενώ η έκφραση περίπλοκων ερωτημάτων συχνά δεν υποστηρίζεται.

Για τη γραφική έκφραση SPARQL ερωτημάτων αναπτύχθηκε ένα καινοτόμο σύστημα, το οποίο βασίζεται αποκλειστικά και μόνο στην οντολογική περιγραφή μιας επιστημονικής περιοχής (Μοντέλο Αναφοράς) και τις ονοματολογίες που χρησιμοποιούνται (Λεξιλόγιο), επιτρέποντας στο χρήστη α) να εξετάσει τους όρους που καλύπτει η εννοιολογική αναπαράσταση του συγκεκριμένου πεδίου γνώσης και κατόπιν β) να εκφράσει τα επιθυμητά ερωτήματα, χωρίς να χρειάζεται να γνωρίζει τις τεχνολογίες του σημασιολογικού ιστού. Η εφαρμογή αυτή που αναπτύχθηκε απευθύνεται κυρίως στους ειδικούς ενός συγκεκριμένου πεδίου της γνώσης, που γνωρίζουν σε βάθος τις έννοιες που αυτό καλύπτει και τις συνθήκες που θα πρέπει να πληρούν τα δεδομένα που ψάχνουν [119]. Παρόλα αυτά, ή εφαρμογή αυτή μπορεί, επίσης, να χρησιμοποιηθεί από τους χρήστες του διαδικτύου, ενώ αποτελεί επιπλέον και ένα χρήσιμο εργαλείο για τους ειδικούς σε θέματα του Σημασιολογικού Ιστού.

Η ενότητα αυτή είναι οργανωμένη ως εξής. Αρχικά, στο κεφάλαιο 17 παρουσιάζουμε τα υπάρχοντα εργαλεία για την έκφραση SPARQL ερωτημάτων μέσω ενός γραφικού περιβάλλοντος καθώς και τις αδυναμίες τους. Ακολούθως, στο κεφάλαιο 18 περιγράφουμε το σύστημα που αναπτύχθηκε για την Οπτική Έκφραση SPARQL Ερωτημάτων. Στο κεφάλαιο 19 εστιάσαμε στην αλληλεπίδραση μεταξύ των βασικών οντοτήτων του συστήματος για την έκφραση των ερωτημάτων μέσω του γραφικού περιβάλλοντος που αναπτύχθηκε, ενώ στο κεφάλαιο 20 περιγράφουμε συνοπτικά τον αλγόριθμο που υλοποιήθηκε για την αυτόματη παραγωγή του SPARQL ερωτήματος με βάση τα δεδομένα που παρέχει ο χρήστης. Στο κεφάλαιο 21 υπάρχει μία συζήτηση για θέματα σχετικά με το εργαλείο που αναπτύχθηκε. Τέλος, στο κεφάλαιο 22 συνοψίσαμε τα κύρια σημεία της ενότητας αυτής.

Η σελίδα αυτή είναι σκόπιμα λευκή

17 *Σχετική Εργασία*

17.1 Κατηγοριοποίηση Εργαλείων και Συστημάτων

Για την έκφραση SPARQL ερωτημάτων υπάρχουν αρκετά εργαλεία και συστήματα, τα οποία εντάσσονται σε τρεις ευρύτερες κατηγορίες [120]. Η πρώτη κατηγορία περιλαμβάνει τους Επεξεργαστές Ερωτήματος (Query Editors), οι οποίοι επιτρέπουν στους χρήστες να πληκτρολογήσουν τα επιθυμητά SPARQL ερωτήματα, με την προϋπόθεση ότι γνωρίζουν σε βάθος τις τεχνολογίες του Σημασιολογικού Ιστού (ειδικότερα, RDF και SPARQL). Η δεύτερη κατηγορία περιλαμβάνει τα Οπτικά Συστήματα Έκφρασης Ερωτημάτων (Visual Query Systems), τα οποία επιτρέπουν στους χρήστες να εκφράσουν SPARQL ερωτήματα μέσω της Εικονικής Αναπαράστασης ενός συγκεκριμένου πεδίου γνώσης. Η τρίτη κατηγορία αποτελείται από τις Οπτικές Γλώσσες Έκφρασης Ερωτημάτων (Visual Query Languages), που, επίσης, δίνουν τη δυνατότητα στο χρήστη να εκφράσει SPARQL ερωτήματα. Όμως, τα συστήματα αυτά είναι στενά συνδεδεμένα με το φορμαλισμό που χρησιμοποιείται (Query Language Formalism) και κατά συνέπεια είναι δύσκολο να χρησιμοποιηθούν από μη ειδικούς στον τομέα αυτό.

Ακολουθεί μια συνοπτική περιγραφή των υπαρχόντων συστημάτων για την έκφραση SPARQL ερωτημάτων. Στην εργασία αυτή, ιδιαίτερο ενδιαφέρον δόθηκε στα Οπτικά Συστήματα, καθώς σχετίζονται άμεσα με την εργασία μας και τα εργαλεία που αναπτύχθηκαν. Επίσης, κατά την ανάλυσή μας, δεν εξετάσαμε τις Ελεγχόμενες Φυσικές

Γλώσσες (Controlled Natural Languages) [121], οι οποίες έχουν τη δυνατότητα να παράγουν αυτόματα την τυπική έκφραση των ερωτημάτων με βάση την έκφρασή τους στη φυσική γλώσσα. Οι γλώσσες αυτές επιτρέπουν στο χρήστη να διατυπώσει εύκολα και γρήγορα τα ερωτήματά του. Όμως, ο προτεινόμενος τρόπος τυπικής έκφρασης των ερωτημάτων (π.χ. χρησιμοποιώντας τη SPARQL) δεν είναι πάντα σωστός, ειδικότερα στην περίπτωση που ο χρήστης θα ήθελε να εκφράσει ιδιαίτερα περίπλοκα ερωτήματα που περιλαμβάνουν αρκετές συνθήκες και παραμέτρους.

17.2 Επεξεργαστές Ερωτήματος

Ο Flint Editor [122] είναι μια διαδικτυακή εφαρμογή που διευκολύνει την έκφραση SPARQL ερωτημάτων. Τα στοιχεία ενός SPARQL ερωτήματος, όπως τα ονόματα των μεταβλητών και τα URIs, επισημαίνονται με διαφορετικά χρώματα, συμβάλλοντας σημαντικά στη διαύγεια των ερωτημάτων του χρήστη. Το σύστημα, επίσης, εξετάζει τα SPARQL ερωτήματα για πιθανά συντακτικά λάθη, τα οποία επίσης επισημαίνονται. Η ενσωμάτωση, όμως, του Flint σε υπάρχοντα συστήματα είναι σχετικά δύσκολη, καθώς το εργαλείο αυτό προϋποθέτει ότι τα αντίστοιχα συστήματα είναι Διασταυρούμενης-προέλευσης Χρήση Πόρων (CORS) ενεργά [123], το οποίο δεν υποστηρίζεται από αρκετά συστήματα, όπως ο D2R εξυπηρετητής. Το σύστημα YASGUI [124] είναι μια άλλη διαδικτυακή εφαρμογή που διευκολύνει το χρήστη στην έκφραση SPARQL ερωτημάτων μέσω της υπηρεσίας αυτόματης συμπλήρωσης (auto-complete) που παρέχεται. Τα ερωτήματα του χρήστη μπορούν κατόπιν να σταλούν προς εκτέλεση στα αντίστοιχα συστήματα, χρησιμοποιώντας HTTP GET ή POST αιτήματα που ορίζονται κατά την παραμετροποίηση του συστήματος. Οι χρήστες μπορούν, επίσης, να αποθηκεύσουν τοπικά τα αποτελέσματα της εκτέλεσης των ερωτημάτων τους.

Οι MCCarthy κ.ά. [125] ανέπτυξαν μία διαδικτυακή εφαρμογή που διευκολύνει την έκφραση ενός SPARQL ερωτήματος. Η εφαρμογή περιλαμβάνει μία περιοχή

κειμένου για την πληκτρολόγηση νέων SPARQL ερωτημάτων, ενώ οι αλγόριθμοι που υλοποιήθηκαν βοηθούν το χρήστη κατά την παραπάνω διαδικασία, λαμβάνοντας υπόψη τα οντολογικά στοιχεία που έχουν οριστεί καθώς επίσης και τα στοιχεία του SPARQL ερωτήματος (π.χ., ονόματα μεταβλητών) που έχουν ήδη δοθεί. Η προσέγγιση που ακολουθήθηκε υποστηρίζει την έκφραση ερωτημάτων χωρίς να είναι απαραίτητο ο χρήστης να γνωρίζει τα URIs των οντολογικών στοιχείων, ενώ συμβάλλει σημαντικά στην έκφραση συντακτικά και, σε κάποιο βαθμό, σημασιολογικά σωστών SPARQL ερωτημάτων. Η εφαρμογή αυτή απευθύνεται κυρίως σε ειδικούς σε θέματα σημασιολογικού ιστού και, συνεπώς, δε θα μπορούσε να χρησιμοποιηθεί από χρήστες που δε διαθέτουν ανάλογο υπόβαθρο. Όμως, οι χρήστες δε χρειάζεται να θυμούνται τα ονόματα των οντολογικών στοιχείων (εντοπίζονται αυτόματα καθώς ο χρήστης πληκτρολογεί τους αντίστοιχους χαρακτήρες) και ειδικά τα URIs τους, τα οποία παρέχονται από το σύστημα με βάση τα στοιχεία που έχουν επιλεχθεί.

17.3Οπτικά Συστήματα Έκφρασης Ερωτημάτων

Οι Seneviratne και Sealfon [126] ανέπτυξαν ένα σύστημα επονομαζόμενο QueryMed για τη διευκόλυνση της αναζήτησης δεδομένων σε Βιοϊατρικές πηγές που είναι διαθέσιμες στο Σημασιολογικό Ιστό. Η διαδικτυακή εφαρμογή επιτρέπει την αναζήτηση χρησιμοποιώντας λέξεις κλειδιά (keywords), η οποία είναι ιδιαίτερα χρήσιμη στα πρώτα στάδια της αναζήτησης της πληροφορίας (Information Seeking Process). Επίσης, επιτρέπει στους χρήστες να εκφράσουν πιο περίπλοκα ερωτήματα με βάση τις παραμέτρους που υπάρχουν για κάθε οντότητα. Όμως, το σύστημα δε λαμβάνει υπόψη το πεδίο τιμών των παραμέτρων και, κατά συνέπεια, δε βοηθάει σημαντικά στο σχηματισμό Λογικών Εκφράσεων (Boolean Expression). Επιπρόσθετα, ένας πολύ περιορισμένος τύπος SPARQL ερωτημάτων υποστηρίζεται από το σύστημα.

Οι Groppe κ.ά. [127] παρουσίασαν ένα οπτικό σύστημα για την έκφραση SPARQL ερωτημάτων για την ανάλυση των δεδομένων από τα κοινωνικά δίκτυα που είναι διαθέσιμα στο σημασιολογικό ιστό. Το σύστημα εκτός από έναν επεξεργαστή κειμένου, διαθέτει, επίσης, ένα γραφικό περιβάλλον για το σχηματισμό SPARQL ερωτημάτων και κυρίως των τριάδων (triple patterns) του ερωτήματος. Κατά συνέπεια, το σύστημα δεν είναι ιδιαίτερα χρήσιμο για το μέσο χρήστη, ο οποίος καλείται να έχει επαρκή γνώση της γλώσσας SPARQL και του μοντέλου των δεδομένων. Στην εργασία αυτή παρουσιάστηκε, επίσης, και μια διαφορετική προσέγγιση για την έκφραση ερωτημάτων βασισμένη στην παρουσίαση των διαθέσιμων δεδομένων (Browser-like approach), δίνοντας τη δυνατότητα στο χρήστη να εξετάσει τα δεδομένα που είναι διαθέσιμα για καθεμιά απ' τις οντότητες και κατόπιν να περιορίσει τα δεδομένα που εμφανίζονται εισάγοντας τα κατάλληλα φίλτρα (FILTER clauses), ενώ το σύστημα αναλαμβάνει να διαμορφώσει τα απαραίτητα SPARQL ερωτήματα. Η αναζήτηση με βάση την προσέγγιση αυτή είναι ιδιαίτερα εύκολη για τους χρήστες του διαδικτύου. Όμως, η εφαρμογή της προϋποθέτει ότι τα δεδομένα είναι ήδη διαθέσιμα. Επίσης, η έκφραση περίπλοκων ερωτημάτων (π.χ., χρησιμοποιώντας τελεστές της SPARQL, όπως union) δεν υποστηρίζεται.

Το OntoWiki [128] είναι ένα εργαλείο που συμβάλλει στην ανάπτυξη εφαρμογών για το Σημασιολογικό Ιστό. Το εργαλείο αυτό παρέχει ένα γραφικό περιβάλλον για την προβολή και επεξεργασία RDF δεδομένων και βασίζεται σε δύο διακριτές αλλά στενά συνδεδεμένες διαδικασίες, επονομαζόμενες «προβολή οντότητας» και «προβολή λίστας». Η πρώτη παρουσιάζει όλη τη διαθέσιμη πληροφορία για μία οντότητα, ενώ η δεύτερη παρουσιάζει ένα σύνολο από οντότητες. Ένα σημαντικό στοιχείο του OntoWiki είναι το γεγονός ότι παρέχει αυτόματα μια συνοπτική, φιλική προς το χρήστη, αναπαράσταση μιας οντότητας αντί για το URI, με την προϋπόθεση ότι έχουν χρησιμοποιηθεί υπάρχουσες

οντολογίες για την περιγραφή των οντοτήτων της RDF βάσης δεδομένων. Επίσης, το γραφικό περιβάλλον μπορεί να γίνει αρκετά πιο φιλικό προς το χρήστη και να προσαρμοστεί στις ανάγκες του, εάν και εφόσον είναι γνωστός ο τύπος των δεδομένων (π.χ., δομή και σημασία των όρων) [129]. Οι χρήστες μπορούν να αναζητήσουν τα επιθυμητά δεδομένα, χρησιμοποιώντας λέξεις κλειδιά ή επιλέγοντας κάποια οντότητα από αυτές που ήδη προβάλλονται. Και στις δύο παραπάνω περιπτώσεις το σύστημα, για να εντοπίσει τα δεδομένα, παράγει αυτόματα το αντίστοιχο SPARQL ερώτημα. Το παραπάνω σύστημα αλληλεπιδρά με το χρήστη, επιτρέποντάς του να προσαρμόσει τα ερωτήματα στις ανάγκες του με βάσει τα δεδομένα που λαμβάνει. Όμως, η προσέγγιση αυτή δεν μπορεί πάντα να εφαρμοστεί, καθώς ορισμένες φορές, μόνο μια περιγραφή των δεδομένων είναι διαθέσιμη. Επίσης, το σύστημα δεν επιτρέπει στους χρήστες να εκφράσουν σημασιολογικά σωστά κριτήρια, λαμβάνοντας υπόψη τη σημασία των παραμέτρων. Τέλος, η έκφραση πιο περίπλοκων ερωτημάτων, χρησιμοποιώντας για παράδειγμα τον τελεστή ένωσης, δεν είναι εφικτή.

Οι Haag κ.ά. [130] παρουσίασαν ένα σύστημα για την Οπτική Δημιουργία SPARQL ερωτημάτων, βασισμένο σε μία επέκταση των Φίλτρων (Filter) / Ροών (Flow) Διαγραμμάτων [131]. Τα διαγράμματα αυτά αποτελούνται από κόμβους – φίλτρα και ακμές, το πάχος των οποίων δείχνει την ποσότητα των δεδομένων που ικανοποιούν τον εκάστοτε περιορισμό. Στην ανανεωμένη έκδοση των Διαγραμμάτων Φίλτρων/Ροών, κάθε κόμβος έχει (γενικά) έναν ή περισσότερους υποδοχείς (receptors) και έναν ή περισσότερους εκπομπούς (emitters), επιτρέποντας την έκφραση περίπλοκων λογικών εκφράσεων που περιλαμβάνουν παραπάνω από μία παραμέτρους. Οι Οπτικά Εκφρασμένοι γράφοι επεξεργάζονται κατόπιν απ' το σύστημα που είναι υπεύθυνο για την έκφραση του αντίστοιχου SPARQL ερωτήματος. Η προσέγγιση που ακολουθήθηκε επιτρέπει να εκφράσουν διάφορα SPARQL ερωτήματα, χωρίς να είναι απαραίτητο να

έχουν ιδιαίτερες γνώσεις των τεχνολογιών του Σημασιολογικού Ιστού. Όμως, σημαντική προσπάθεια είναι απαραίτητη απ' τη μεριά του χρήστη για τη δημιουργία ακόμη και ενός απλού ερωτήματος. Η έκφραση περίπλοκων ερωτημάτων, στα οποία οι οντότητες θα πρέπει να ικανοποιούν αρκετές συνθήκες, είναι εφικτή. Ωστόσο, το μέγεθος του γράφου στην περίπτωση αυτή αυξάνει σημαντικά, δυσκολεύοντας το χρήστη τόσο στην κατανόηση όσο και στην περαιτέρω επεξεργασία του ερωτήματος.

Οι Brunetti κ.ά. [132] παρουσίασαν ένα αλληλεπιδραστικό σύστημα, ονομαζόμενο Rhizomer, για την αναζήτηση δεδομένων, το οποίο βασίζεται σε Όψεις (Facets) και Περιστροφές (Pivoting). Μέσω των «όψεων» οι χρήστες μπορούν να φιλτράρουν τα δεδομένα με βάση τις παραμέτρους τους, ενώ χρησιμοποιώντας τη λειτουργία της «περιστροφής» μπορούν να διασχίσουν τους συνδέσμους μεταξύ των δεδομένων για την έκφραση πιο περίπλοκων συνθηκών. Ο συνδυασμός των όψεων και της διαδικασίας της περιστροφής διευκολύνει την υλοποίηση των βασικών διεργασιών για την ανάλυση των δεδομένων που παρουσιάστηκαν από τον Shneiderman [133], καθώς οι χρήστες είναι ήδη εξοικειωμένοι με τις όψεις που χρησιμοποιούνται ευρέως στο διαδίκτυο, ενώ η χρήση της διαδικασίας της περιστροφής συμβάλλει στο να ξεπεράσουμε τους περιορισμούς που σχετίζονται με αυτές. Θα πρέπει να σημειωθεί ότι οι όψεις βασίζονται στην παρουσίαση των διαθέσιμων επιλογών για καθεμιά από τις παραμέτρους μιας οντότητας και, κατά συνέπεια, καταλαμβάνουν αρκετό χώρο στην οθόνη, ιδιαίτερα στις περιπτώσεις εκείνες όπου υπάρχουν αρκετοί παράμετροι. Επίσης, η έκφραση περίπλοκων ερωτημάτων απαιτεί τη χρησιμοποίηση αρκετών όψεων, οι οποίες με τη σειρά τους μπορεί να μπερδέψουν τους χρήστες κατά την παραπάνω διαδικασία.

Οι Soylu κ.ά. [134], επίσης, παρουσίασαν ένα Οπτικό Σύστημα για την έκφραση SPARQL ερωτημάτων, το οποίο αναπτύχθηκε στα πλαίσια του έργου Optique [135]. Η όλη διαδικασία ξεκινά με την επιλογή του τύπου των δεδομένων προς αναζήτηση (key

concept) και στη συνέχεια ο χρήστης ορίζει τις συνθήκες που θα πρέπει να πληρούν οι παράμετροι που είναι διαθέσιμοι για την εκάστοτε οντότητα. Οι σχέσεις με τις άλλες οντότητες παρουσιάζονται σε ένα άλλο τμήμα της οθόνης, απ' όπου ο χρήστης μπορεί αρχικά να επιλέξει τη σχέση που τον ενδιαφέρει και κατόπιν να ορίσει τις συνθήκες που θα πρέπει να πληρούν οι αντίστοιχες οντότητες. Τα δεδομένα που ορίζει ο χρήστης παρουσιάζονται σε ένα τρίτο τμήμα της οθόνης με τη μορφή ενός διαγράμματος, απ' όπου οι χρήστες έχουν τη δυνατότητα να εστιάσουν σε ένα συγκεκριμένο κόμβο – έννοια και στις συνθήκες που αυτός θα πρέπει να ικανοποιεί. Η όλη προσέγγιση είναι ιδιαίτερα ενδιαφέρουσα, καθώς οι δημιουργοί του συστήματος συνδύασαν ομαλά τη χρήση Διαγραμμάτων (Diagrams) και Όψεων (Facets) [136] για την έκφραση SPARQL ερωτημάτων. Όμως, για το σκοπό αυτό οι χρήστες θα πρέπει κάθε φορά να εστιάσουν την προσοχή τους σε αρκετά διαφορετικά σημεία της οθόνης, γεγονός που ίσως μπερδέψει τους μη-έμπειρους διαδικτυακούς χρήστες. Επίσης, οι Όψεις απαιτούν αρκετό χώρο, όπως προαναφέρθηκε, «περιορίζοντας» τις παραμέτρους κάθε κλάσης που είναι άμεσα ορατές στην οθόνη του χρήστη. Επιπρόσθετα, το σύστημα υποστηρίζει την έκφραση συνδετικών ερωτημάτων (conjunctive queries), στα οποία οι οντότητες θα πρέπει να ικανοποιούν όλα τα κριτήρια που έχουν οριστεί, χωρίς να δίνουν τη δυνατότητα στο χρήστη να εκφράσει κάποια διακλάδωση (branching) που πιθανώς να είναι σημαντική. Τέλος, το σύστημα δε λαμβάνει υπόψη τις πιθανές τιμές των παραμέτρων και κυρίως τη σημασία των όρων αυτών. Κατά συνέπεια, τα SPARQL ερωτήματα που προκύπτουν δεν παρέχουν όλα τα σημασιολογικά σωστά αποτελέσματα.

17.4Οπτικές Γλώσσες Έκφρασης Ερωτημάτων

Το NITELIGHT [137] είναι ένα εργαλείο για τη δημιουργία SPARQL ερωτημάτων. Το εργαλείο διαθέτει έναν καμβά για το σχηματισμό SPARQL ερωτημάτων μέσω των οπτικών αναπαραστάσεων των στοιχείων ενός SPARQL ερωτήματος (γνωστά επίσης ως

vSPARQL) [138]. Οι κλάσεις και οι παράμετροι που έχουν οριστεί στην οντολογία παρουσιάζονται στο χρήστη με τη μορφή μιας επίπεδης λίστας, απ' όπου μπορεί να επιλέξει τα επιθυμητά. Το εργαλείο αυτό δίνει τη δυνατότητα στο χρήστη να εκφράσει ιδιαίτερα περίπλοκα SPARQL ερωτήματα. Απευθύνεται, όμως, σε χρήστες που γνωρίζουν τις τεχνολογίες του Σημασιολογικού Ιστού και, κατά συνέπεια, δεν μπορεί να χρησιμοποιηθεί από το μέσο χρήστη του διαδικτύου. Η διεπαφή που αναπτύχθηκε στο TOppS έργο [139] ακολουθεί την ίδια λογική για την έκφραση SPARQL ερωτημάτων. Πιο συγκεκριμένα, διαθέτει έναν καμβά, ενώ ταυτόχρονα δίνει τη δυνατότητα στο χρήστη να εξερευνήσει τα διαθέσιμα οντολογικά στοιχεία. Οι κλάσεις παρουσιάζονται με τη μορφή ενός δένδρου με βάση τα αξιώματα που υπάρχουν στην οντολογία, ενώ οι παράμετροι παρουσιάζονται ως μία επίπεδη λίστα.

Μια οπτική προσέγγιση έκφρασης ερωτημάτων, βασισμένη στα δεδομένα (γνωστή ως DaVinci) έχει προταθεί από τους Zhang κ.ά. [140], η οποία επιτρέπει το σχηματισμό γράφων εύκολα και γρήγορα, με έναν περιορισμένο αριθμό από «κλικ», σε σύγκριση με την κατασκευή τους ακμή-ακμή. Το γραφικό περιβάλλον αποτελείται από γραφικά μοτίβα (επονομαζόμενα canned patters) και ονόματα κόμβων, τα οποία επιτρέπουν στο χρήστη να ορίσει τα δεδομένα προς αναζήτηση μέσω της χρησιμοποίησης ενός ή περισσοτέρων μοτίβων και της συγκεκριμενοποίησης των εσωτερικών τους στοιχείων. Τα μοτίβα που υποστηρίζονται εξήχθησαν αυτόματα από το σύστημα μέσω της ανάλυσης των δεδομένων και, κατά συνέπεια, μπορούν να καλύψουν τα περισσότερα ερωτήματα. Τα ερωτήματα του χρήστη που έχουν εκφραστεί μέσω του γραφικού αυτού προγράμματος μπορούν κατόπιν να εκφραστούν σε μια γλώσσα όπως η SPARQL [141]. Η όλη προσέγγιση που ακολουθήθηκε είναι ιδιαίτερα ενδιαφέρουσα, καθώς βασίζεται στη συχνότητα με την οποία ένας γράφος ή υπό-γράφος χρησιμοποιείται κατά την αναπαράσταση των

δεδομένων. Όμως, το γραφικό περιβάλλον που παρέχεται είναι αρκετά περίπλοκο, καθώς έχει σχεδιαστεί για χρήστες με ιδιαίτερα τεχνικά προσόντα.

17.5Κριτική Αποτίμηση και Επιθυμητά Χαρακτηριστικά

Η παραπάνω ανάλυση έδειξε ότι υπάρχουν αρκετά εργαλεία και συστήματα για την έκφραση SPARQL ερωτημάτων. Όμως, ελάχιστα από αυτά θα μπορούσαν να χρησιμοποιηθούν από χρήστες που δεν είναι γνώστες των τεχνολογιών του σημασιολογικού ιστού, ενώ η έκφραση περίπλοκων ερωτημάτων, στα οποία τα δεδομένα θα πρέπει να ικανοποιούν αρκετές συνθήκες, είναι ιδιαίτερα δύσκολη. Το παραπάνω συμπέρασμα υποστηρίζεται επίσης και από την έρευνα και κατηγοριοποίηση των εργαλείων που σχετίζονται με τις τεχνολογίες του σημασιολογικού ιστού [142], η οποία έδειξε ότι σχεδόν όλα τα εργαλεία που εξετάσθηκαν είναι δύσκολο να χρησιμοποιηθούν από ανθρώπους μη ειδικούς σε θέματα έκφρασης ερωτημάτων βασισμένων σε οντολογίες. Υπάρχοντα μοτίβα αλληλεπίδρασης (interaction paradigms), όπως φόρμες και διαγράμματα, περιπλέκουν τους χρήστες του διαδικτύου, ενώ η χρησιμοποίηση γράφων συχνά θεωρείται ότι δεν ικανοποιεί το λόγο για τον οποίο επιλέχθηκε [143].

Η παραπάνω ανάλυση ανέδειξε ορισμένες παραμέτρους που θα πρέπει να λάβουμε υπόψη κατά το σχεδιασμό ενός νέου συστήματος για την έκφραση SPARQL ερωτημάτων. Πιο συγκεκριμένα, το σύστημα θα πρέπει να παρέχει ένα, φιλικό προς το χρήστη, γραφικό περιβάλλον που θα του επιτρέπει να καθορίσει τις συνθήκες (κριτήρια) τις οποίες θα πρέπει να πληρούν τα δεδομένα, χωρίς οι χρήστες να χρειάζεται να χειριστούν γραφικά μοτίβα, μεταβλητές και URIs. Επίσης, η προσέγγιση που θα ακολουθηθεί θα πρέπει να περιορίζει στο ελάχιστο τη συνεισφορά του χρήστη κατά τη διαδικασία έκφρασης ενός νέου ερωτήματος, ενώ το περιβάλλον που χρησιμοποιείται θα πρέπει να είναι αρκετά αλληλεπιδραστικό έτσι, ώστε να μπορεί να προσαρμοστεί στις ανάγκες έκφρασης του κάθε κριτηρίου. Η αυτόματη συμπλήρωση είναι ιδιαίτερα σημαντική κατά την παραπάνω

διαδικασία, ειδικά στις περιπτώσεις εκείνες όπου οι τιμές μιας παραμέτρου προέρχονται από κάποια ονοματολογία ή σύστημα κωδικοποίησης. Επιπρόσθετα, το σύστημα θα πρέπει να επιτρέπει στους χρήστες να διαμορφώσουν σημασιολογικά έγκυρα ερωτήματα, λαμβάνοντας υπόψη τα αξιώματα που έχουν οριστεί, όπως το πεδίο τιμών των παραμέτρων, τους περιορισμούς πληθικότητας καθώς επίσης και την ιεραρχία των κλάσεων.

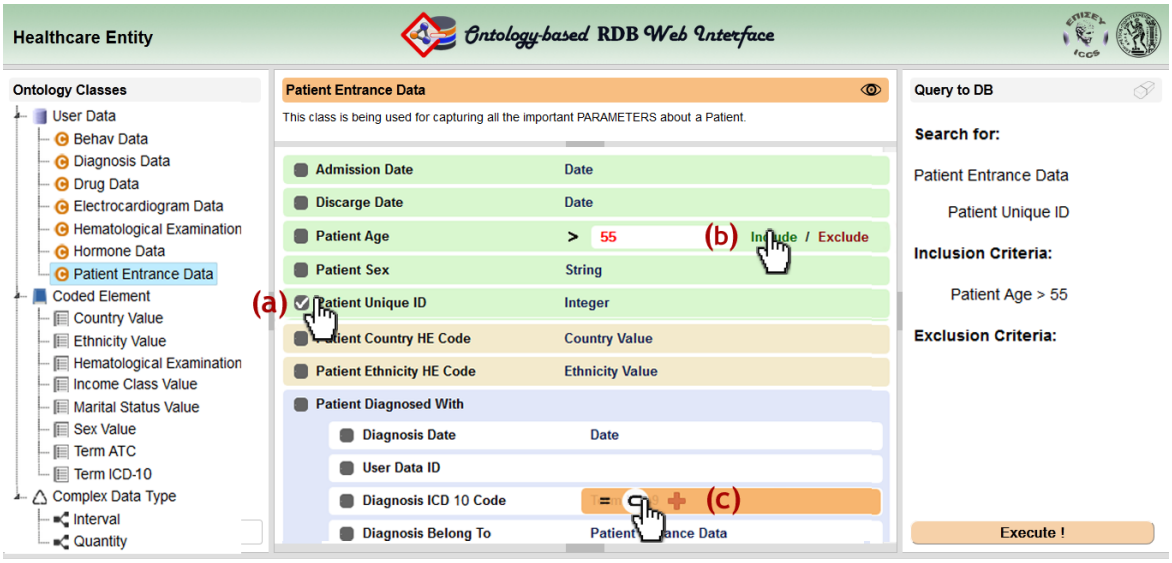
18 *Σύστημα Έκφρασης Ερωτημάτων*

18.1Παρεχόμενες Υπηρεσίες

Για τη διευκόλυνση της χρησιμοποίησης των δυνατοτήτων που προσφέρουν οι τεχνολογίες του Σημασιολογικού Ιστού για την πρόσβαση στα δεδομένα των βάσεων, αναπτύχθηκε μία διαδικτυακή εφαρμογή. Η εφαρμογή παρέχει ένα ιδιαίτερα αλληλεπιδραστικό, φιλικό προς το χρήστη, γραφικό περιβάλλον, που του επιτρέπει, αρχικά, να εξετάσει το πεδίο γνώσης που καλύπτει η βάση δεδομένων και τους όρους που υπάρχουν σε αυτή και, στη συνέχεια, να εκφράσει γραφικά τα επιθυμητά SPARQL ερωτήματα, λαμβάνοντας υπόψη τη δομή των δεδομένων και τις παραμέτρους τους, τα οποία έχουν εκ των προτέρων εκφραστεί, χρησιμοποιώντας τις τεχνολογίες του σημασιολογικού ιστού.

Για τις ανάγκες παρουσίασης του εργαλείου που αναπτύχθηκε και των δυνατοτήτων που αυτό προσφέρει το παραμετροποιήσαμε με τις οντολογίες και ονοματολογίες των όρων που χρησιμοποιούνται σε μία συγκεκριμένη μονάδα περίθαλψης έτσι, ώστε να μπορεί έπειτα να χρησιμοποιηθεί από τους υπεύθυνους στον τομέα αυτό για την εύρεση των ασθενών που πληρούν ορισμένα κριτήρια. Ο κύριος λόγος που διαλέξαμε το παράδειγμα αυτό προέρχεται από το γεγονός ότι οι ασθενείς που ψάχνουμε θα πρέπει να ικανοποιούν αρκετά κριτήρια, τα οποία ορισμένες φορές είναι ιδιαίτερα περίπλοκα [41]. Ως εκ τούτου, το παράδειγμα αυτό μας δίνει την ευκαιρία να αναδείξουμε αρκετά

χαρακτηριστικά της εφαρμογής που αναπτύχθηκε, καθώς επίσης και των αλγορίθμων που χρησιμοποιήθηκαν για την έκφραση του αντίστοιχου SPARQL ερωτήματος.



Σχήμα 19: Ένα στιγμιότυπο από τη Διεπαφή του συστήματος Οπτικής Έκφρασης Ερωτημάτων

Το Σχήμα 19 απεικονίζει ένα στιγμιότυπο της διεπαφής που αναπτύχθηκε. Στη πραγματικότητα, στην εικόνα αυτή ενσωματώσαμε παραπάνω από ένα στιγμιότυπα που προέκυψαν κατά τον ορισμό ενός ή περισσοτέρων κριτηρίων, με σκοπό να δείξουμε τόσο τα επιμέρους τμήματα της διεπαφής όσο και την αλληλεπίδραση του συστήματος με το χρήστη (περιγράφεται αναλυτικά στην επόμενη ενότητα). Στο παράδειγμα αυτό, το Μοντέλο Αναφοράς (Reference Ontology) παρέχει την τυπική περιγραφή των παραμέτρων των δεδομένων που καταγράφονται κατά την επίσκεψη, θεραπεία και αποχώρηση ενός ασθενούς από μία Μονάδα Περίθαλψης (Healthcare Entity), όπως τα δημογραφικά στοιχεία τους ασθενούς (π.χ., ηλικία και φύλο), τις διαγνώσεις που έγιναν (π.χ., «προβλήματα» που εντοπίστηκαν καθώς και την ημερομηνία που έγινε η διάγνωση), φάρμακα που συνταγογραφήθηκαν (π.χ., δραστική ουσία, περίοδος και συχνότητα λήψης) και εργαστηριακές μετρήσεις που έλαβαν χώρα. Η Οντολογία Όρων (Vocabulary Ontology) περιέχει τους όρους που μπορούμε να χρησιμοποιήσουμε για καθεμιά από τις παραμέτρους του μοντέλου μας, εφόσον αυτές προέρχονται από ένα συγκεκριμένο,

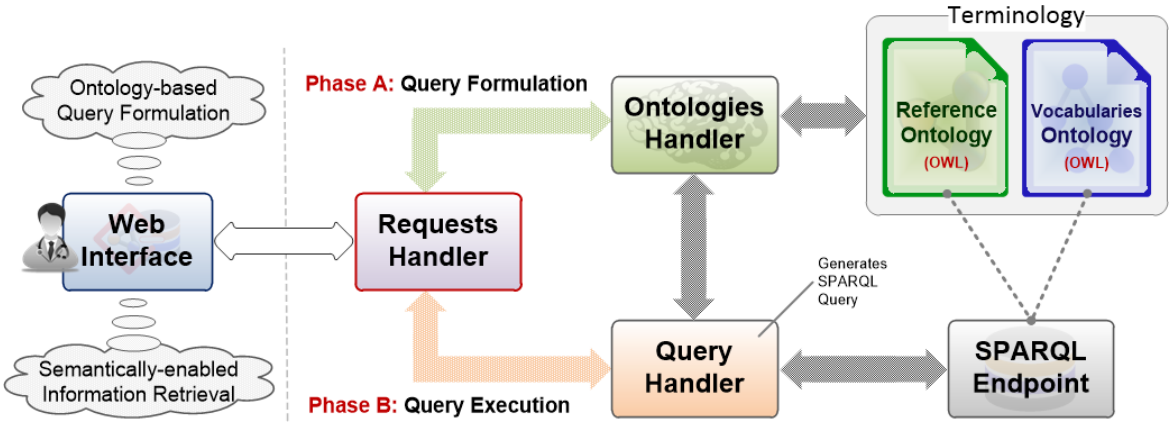
σαφώς ορισμένο σύνολο τιμών. Στην περίπτωσή μας, τα προβλήματα που καταγράφηκαν προέρχονται από τη 10η έκδοση της Διεθνούς Κατηγοριοποίησης των Ασθενειών (ICD-10) που έχει δημοσιευτεί από το Διεθνή Οργανισμό Υγείας (WHO), ενώ οι δραστικές ουσίες που αναφέρονται προέρχονται από το Σύστημα Κατηγοριοποίησης Ανατομικών Φαρμακευτικών Ουσιών (ATC) [144], που έχει δημοσιευτεί από τον παραπάνω οργανισμό.

Μέσω του γραφικού περιβάλλοντος, οι χρήστες έχουν τη δυνατότητα να ορίσουν τις παραμέτρους των δεδομένων που τους ενδιαφέρουν, καθώς επίσης και τις συνθήκες που θα πρέπει να πληρούν, χωρίς να χρειάζεται να γνωρίζουν τη σύνταξη και τη σημασιολογία των γλωσσών OWL και SPARQL. Για τη διευκόλυνση του χρήστη, οι συνθήκες που ορίζονται έχουν χωριστεί σε δύο ευρύτερες κατηγορίες, επονομαζόμενες συνθήκες/κριτήρια εισαγωγής και εξαγωγής αντίστοιχα. Τα δεδομένα που ψάχνουν οι χρήστες θα πρέπει να ικανοποιούν όλα τα κριτήρια εισαγωγής (inclusion criteria) που έχουν οριστεί και ταυτοχρόνως να μην ικανοποιούν κάποιο από τα κριτήρια εξαγωγής (exclusion criteria). Ακολουθώντας την προσέγγιση αυτή, οι χρήστες έχουν τη δυνατότητα να ορίσουν επακριβώς όλες τις συνθήκες που θα πρέπει να πληρούν τα δεδομένα με έναν τρόπο που διευκολύνει την ανάγνωση των συνθηκών που έχουν οριστεί, ως αποτέλεσμα του σαφούς διαχωρισμού τους.

18.2 Αρχιτεκτονική του Συστήματος

Το Σχήμα 20 απεικονίζει την αρχιτεκτονική του συστήματος που αναπτύχθηκε. Το σύστημα αποτελείται από μία διαδικτυακή διεπαφή (Web Interface), η οποία χρησιμοποιεί τις υπηρεσίες που παρέχονται από τον εξυπηρετητή ερωτημάτων (Requests Handler), με την κύρια συνεισφορά του να εντοπίζεται στην παρουσίαση των δεδομένων που λαμβάνει από τον εξυπηρετητή αλλά και στη διαχείριση της αλληλεπίδρασης με τους χρήστες. Ο εξυπηρετητής ερωτημάτων χρησιμοποιεί τις υπηρεσίες που παρέχονται από το χειριστή

οντολογιών (Ontology Handler) για τον εντοπισμό των κλάσεων και των αντίστοιχων παραμέτρων τους που ορίζονται στο Μοντέλο Αναφοράς (Reference Ontology), καθώς επίσης και τους όρους που ορίζονται στην οντολογία όρων (Vocabulary Ontology) που περιέχει τις διαθέσιμες τιμές (όνομα, περιγραφή, κωδικό) των παραμέτρων του Μοντέλου Αναφοράς. Ο χειριστής ερωτημάτων (Query Handler) χρησιμοποιείται, για να εντοπίσουμε πιθανές ασυνέπειες (inconsistencies) μεταξύ των κριτηρίων που έχει ορίσει ο χρήστης (Φάση Α: γραφική αναπαράσταση ερωτημάτων), καθώς επίσης και για την παραγωγή του SPARQL ερωτήματος με βάση τα δεδομένα που έχει ορίσει ο χρήστης (Φάση Β: τυπική έκφραση ερωτημάτων).



Σχήμα 20: Αρχιτεκτονική Συστήματος Οπτικής Έκφρασης Ερωτημάτων

Στον Πίνακας 8 έχουμε συνοψίσει τις υπηρεσίες που παρέχονται από τον εξυπηρετητή ερωτημάτων, συμπεριλαμβανομένων των δεδομένων που παρέχουμε (input) καθώς και των δεδομένων που επιστρέφει (output) σε κάθε περίπτωση. Η αλληλεπίδραση μεταξύ των λογισμικών στοιχείων του χρήστη (client side components) και του εξυπηρετητή (server side components) βασίζεται στην ανταλλαγή JSON [110] μηνυμάτων. Επίσης, η γλώσσα JavaScript σε συνδυασμό με την jQuery βιβλιοθήκη [145] χρησιμοποιήθηκε για τη δημιουργία της διαδικτυακής διεπαφής, ενώ η γλώσσα προγραμματισμού Java μαζί με τα συστήματα και τις αντίστοιχες βιβλιοθήκες του Jena

Framework [146] και Pellet Reasoner [64] για την υλοποίηση των λογισμικών στοιχείων που βρίσκονται στον εξυπηρετητή.

Όνομα	**Είσοδος**	**Έξοδος**	**Σύντομη Περιγραφή**
Classes	-	Classes Tree	Παρέχει τις διαθέσιμες Κλάσεις οργανωμένες με τη μορφή ενός Δένδρου, με βάση τα αξιώματα που έχουν οριστεί.
Properties	Class URI	List of Properties	Παρέχει μία λίστα με τις παραμέτρους που μπορούν να χρησιμοποιηθούν για την περιγραφή των οντοτήτων που ανήκουν στη δοσμένη Κλάση.
Terms	String	List of Terms	Παρέχει μία Λίστα με τους Όρους (όνομα, κωδικός, περιγραφή) που ξεκινούν με τη δοσμένη συμβολοακολουθία.
Consistency Check	Query Data	List of Issues	Παρέχει μία Λίστα με ασυνέπειες που εντοπίστηκαν με βάση τα κριτήρια που έχουν ήδη οριστεί.
Query Execution	Query Data	Response Data	Παράγει την τυπική έκφραση του ερωτήματος σε SPARQL και ακολούθως το εκτελεί και επιστρέφει την απάντηση.

Πίνακας 8: Υπηρεσίες που παρέχονται από τον Εξυπηρετητή για την έκφραση ερωτημάτων μέσω του Γραφικού Περιβάλλοντος

Το σύστημα που αναπτύχθηκε, και κυρίως το γραφικό του περιβάλλον, βασίζεται αποκλειστικά και μόνο στο Μοντέλο Αναφοράς (Reference Ontology) και τις ονοματολογίες που χρησιμοποιούνται (Vocabulary Ontology), τα οποία είναι μέρος της παραμετροποίησης του συστήματος. Ο τύπος των SPARQL ερωτημάτων που θα χρησιμοποιηθούν για την τυπική έκφραση των ερωτημάτων είναι μια άλλη παράμετρος του συστήματος, η οποία έμμεσα επηρεάζει την εμφάνιση και λειτουργικότητα του γραφικού περιβάλλοντος. Πιο συγκεκριμένα, για την έκφραση SELECT ή CONSTRUCT SPARQL ερωτημάτων το γραφικό περιβάλλον θα πρέπει να δίνει τη δυνατότητα στους

χρήστες όχι μόνο να ορίσουν τα κριτήρια που θα πρέπει να πληρούν τα δεδομένα (όπως στην περίπτωση των ASK SPARQL ερωτημάτων) αλλά και τις παραμέτρους που θα θέλαμε να γνωρίζουμε για τις οντότητες που πληρούν τις συνθήκες. Η τοποθεσία (URL) του λογισμικού συστήματος που θα χρησιμοποιηθεί για την αποτίμηση (ή περαιτέρω επεξεργασία) των αυτομάτως παραγόμενων SPARQL ερωτημάτων θα πρέπει επίσης να καθοριστεί. Σε περίπτωση απουσίας του, το παραγόμενο ερώτημα θα επιστρέφεται στο χρήστη. Στον Πίνακας 9 έχουμε συνοψίσει τις υποχρεωτικές και προαιρετικές παραμέτρους του συστήματος.

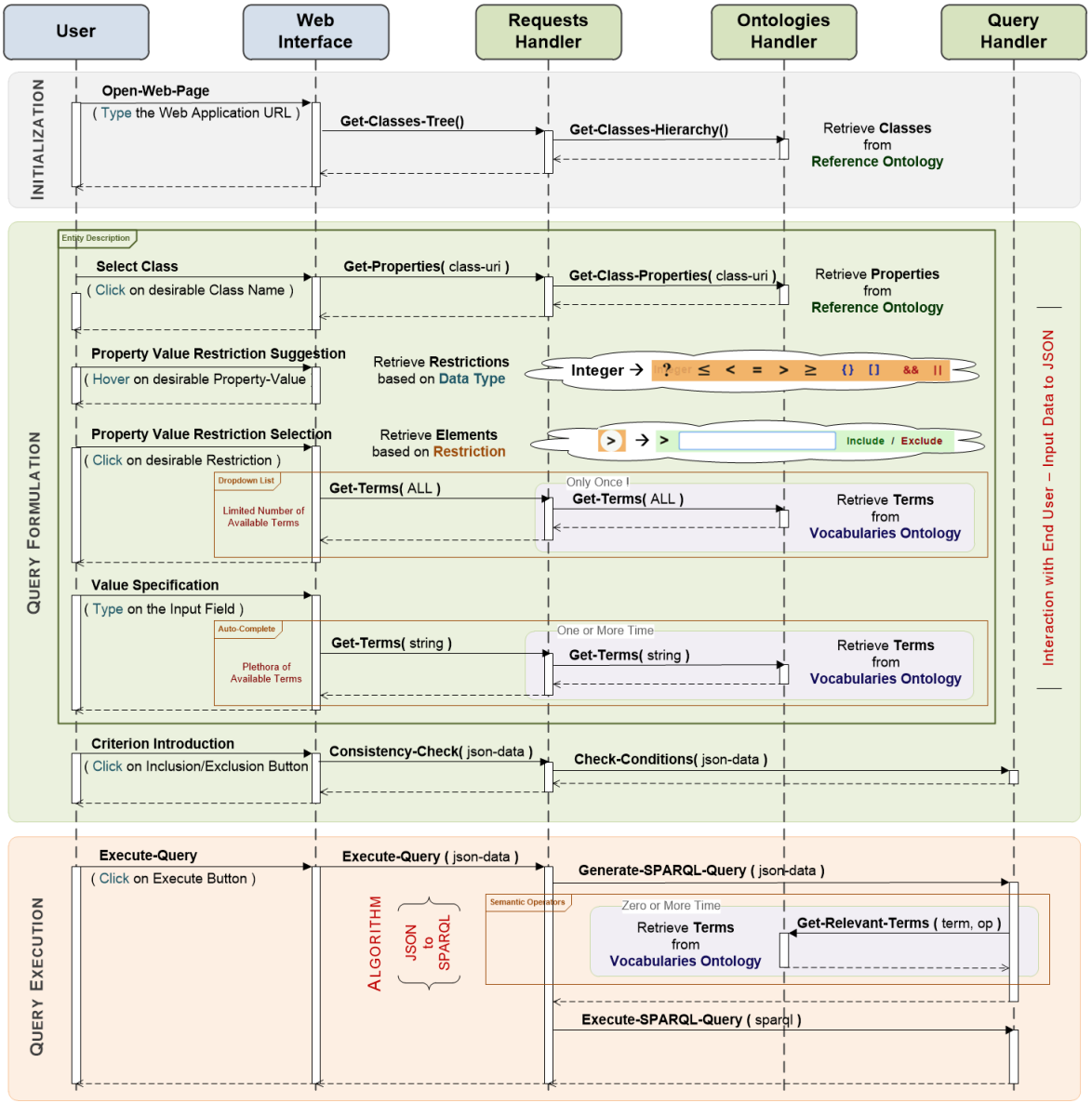
Κατηγορία	**Παράμετρος**	**Υ/Π**	**Σύντομη Περιγραφή**
Ontology	Reference Ontology	Υ	Η οντολογία παρέχει την περιγραφή ενός συγκεκριμένου πεδίου της γνώσης, συμπεριλαμβανομένων των κλάσεων των οντοτήτων που συναντάμε και τις πιθανές παραμέτρους τους.
	Vocabulary Ontology	Π	Η οντολογία παρέχει τους όρους (όνομα, κωδικός, περιγραφή) που χρησιμοποιούνται κατά τον ορισμό των παραμέτρων μιας οντότητας. Ιδανικά οι όροι αυτού θα πρέπει να είναι οργανωμένοι σε ευρύτερες κατηγορίες, με βάση τη σημασία τους.
Property	Query Type	Υ	Ο τύπος του SPARQL ερωτήματος που θα παράγεται αυτόματα από το σύστημα (π.χ., ASK ή SELECT SPARQL ερωτήματα).
	SPARQL Endpoint URL	Π	Το URL της εφαρμογής (SPARQL endpoint) που θα χρησιμοποιηθεί για την αποτίμηση των αυτομάτως παραγόμενων SPARQL ερωτημάτων.

Πίνακας 9: Οι (Υ)ποχρεωτικές και (Π)ροαιρετικές παράμετροι του συστήματος Οπτικής Έκφρασης Ερωτημάτων

19 *Έκφραση Ερωτημάτων μέσω του Γραφικού Περιβάλλοντος*

19.1 Αλληλεπίδραση μεταξύ Οντοτήτων

Όταν ο χρήστης ανοίγει για πρώτη φορά την εφαρμογή πληκτρολογώντας το URL της, το σύστημα λαμβάνει τις κλάσεις που έχουν οριστεί στην OWL οντολογία (μοντέλο αναφοράς) και ακολούθως τις παρουσιάζει στην αριστερή μεριά της οθόνης με τη μορφή ενός δένδρου, λαμβάνοντας υπόψη τα αξιώματα που έχουν οριστεί (αρχικοποίηση συστήματος). Έπειτα, κάθε φορά που ο χρήστης επιλέγει μία κλάση, το σύστημα εντοπίζει τις παραμέτρους της κλάσης και τις παρουσιάζει στο μεσαίο παράθυρο, απ' όπου οι χρήστες έχουν τη δυνατότητα να ορίσουν τα κριτήρια καταλληλότητας και τα δεδομένα προς αναζήτηση. Τα δεδομένα που παρέχει ο χρήστης παρουσιάζονται στη δεξιά μεριά της οθόνης, απ' όπου ο χρήστης μπορεί να εκτελέσει το ερώτημα, χρησιμοποιώντας το «Execute» κουμπί που υπάρχει κάτω δεξιά. Στην περίπτωση αυτή, το σύστημα παράγει αυτόματα την τυπική έκφραση του ερωτήματος σε SPARQL την οποία χρησιμοποιεί, για να αντλήσει δεδομένα από τη βάση, με την προϋπόθεση ότι το αντίστοιχο URL έχει καθοριστεί κατά την παραμετροποίηση του συστήματος.



Σχήμα 21: Αλληλεπίδραση μεταξύ των κύριων Οντοτήτων του συστήματος Οπτικής Έκφρασης Ερωτημάτων

Το Σχήμα 21 απεικονίζει την αλληλεπίδραση μεταξύ των οντοτήτων του συστήματος που συμμετέχουν έμμεσα ή άμεσα στην έκφραση ενός SPARQL ερωτήματος. Κατά την αρχικοποίηση του συστήματος καθώς επίσης και της φάσης έκφρασης του ερωτήματος (φάση Α) η διεπαφή χρησιμοποιεί τις υπηρεσίες που παρέχονται από τον εξυπηρετητή, για να αντλήσει τα απαραίτητα δεδομένα τόσο από το Μοντέλο Αναφοράς όσο και από την Οντολογία με τους Όρους, χρησιμοποιώντας τα αντίστοιχα λογισμικά στοιχεία. Κατά τη φάση της τυπικής έκφρασης του ερωτήματος (φάση Β) χρησιμοποιεί

τον εξυπηρετητή ερωτημάτων και συγκεκριμένα το χειριστή των δεδομένων του ερωτήματος (Query Data Handler), για να παράγει το αντίστοιχο SPARQL ερώτημα. Ο έλεγχος για πιθανές αντιφάσεις, με βάση τα κριτήρια που έχουν οριστεί, αποτελεί μια ειδική περίπτωση, στην οποία συμμετέχουν όλα τα λογισμικά στοιχεία του εξυπηρετητή.

19.2Βασισμένη σε Οντολογίες Έκφραση Ερωτημάτων

Η διαδικασία σχηματισμού ενός ερωτήματος καθοδηγείται από τις κλάσεις και τις παραμέτρους που έχουν οριστεί στο Μοντέλο Αναφοράς (Reference Ontology) και ειδικότερα τα αξιώματα που έχουν οριστεί, όπως ο τύπος των δεδομένων των παραμέτρων τους. Στις περιπτώσεις εκείνες, όπου τα δεδομένα που έχουν καταγραφεί προέρχονται από έναν πρωτεύοντα τύπο δεδομένων (primitive datatype), το σύστημα επιτρέπει στο χρήστη να ορίσει άμεσα τις συνθήκες που θα πρέπει αυτά να ικανοποιούν. Για το σκοπό αυτό, το σύστημα, αρχικά, προτείνει πιθανούς περιορισμούς που μπορούν να χρησιμοποιηθούν, για να ορίσουμε το επιθυμητό εύρος τιμών (παρουσιάζονται σε ένα αναδυόμενο «παράθυρο»), λαμβάνοντας υπόψη τον τύπο των δεδομένων της παραμέτρου. Οι περιορισμοί που μπορούν να χρησιμοποιηθούν για κάθε τύπο δεδομένων καθορίστηκαν εκ των προτέρων, κατά το σχεδιασμό του συστήματος. Για παράδειγμα, στην περίπτωση των «strings», ο χρήστης μπορεί να επιλέξει ότι ψάχνει για τις οντότητες εκείνες για τις οποίες η τιμή της παραμέτρου είτε ταιριάζει απόλυτα με τους χαρακτήρες που παρέχει ο χρήστης είτε απλά ξεκινάει ή τελειώνει με τη δοθείσα συμβολοακολουθία. Ανάλογα με την επιλογή του χρήστη, το γραφικό περιβάλλον ενημερώνεται αυτόματα, δίνοντάς του τη δυνατότητα να παρέχει τις επιθυμητές τιμές μέσω ενός ή περισσοτέρων «input» πεδίων, τα οποία καταγράφονται (είτε ως κριτήριο εισαγωγής είτε εξαγωγής), όταν ο χρήστης πατήσει το αντίστοιχο κουμπί.

Εάν τα δεδομένα ανήκουν σε κάποιο σύνθετο τύπο δεδομένων, γενικά, οι χρήστες θα πρέπει αρχικά να εξετάσουν τις παραμέτρους που καταγράφονται από τις αντίστοιχες

οντότητες και ακολούθως να ορίσουν τις συνθήκες που θα πρέπει αυτές να πληρούν. Για το σκοπό αυτό, το σύστημα επιτρέπει στο χρήστη να επιλέξει τον τύπο των δεδομένων (εάν είναι διαθέσιμες παραπάνω από μία κλάσεις). Ακολούθως, ανάλογα με την κλάση που επιλέχθηκε, το σύστημα επιτρέπει στους χρήστες να εξετάσουν τις παραμέτρους της και να ορίσουν το επιθυμητό εύρος τιμών για μία ή παραπάνω παραμέτρους της κλάσης αυτής. Ο καθορισμός του εύρους τιμών λαμβάνει χώρα είτε άμεσα (στην περίπτωση των datatype παραμέτρων) είτε έμμεσα (στην περίπτωση των object παραμέτρων), ακολουθώντας τα προηγουμένως περιγραφέντα βήματα. Σημειώνουμε ότι οι συνθήκες που ορίζονται κατά τον ορισμό ενός κριτηρίου συνδέονται μεταξύ τους με το λογικό τελεστή «AND» και, επομένως, θα πρέπει να ικανοποιούνται ταυτόχρονα (στην περίπτωση των κριτηρίων εισαγωγής) ή όχι (στην περίπτωση των κριτηρίων εξαγωγής).

Σχετικά με τους ορισμένους από το χρήστη τύπους δεδομένων, όπως είναι οι Όροι (ονόματα συνοδευόμενα από κάποιο κωδικό και το σύστημα κωδικοποίησης απ' όπου προέρχεται), Ποσότητες (μια τιμή συνοδευόμενη από τη μονάδα μέτρησης) και Χρονικές Περίοδοι (ορισμένες είτε παρέχοντας την ημερομηνία αρχής και τέλους είτε καθορίζοντας ένα σημείο αναφοράς και, στη συνέχεια, το χρονικό διάστημα που προηγείται ή ακολουθεί) μπορούμε να ακολουθήσουμε μία παρόμοια διαδικασία κατά τον ορισμό των συνθηκών των παραμέτρων στις οποίες οι πιθανές τιμές προέρχονται από κάποια από τις παραπάνω κλάσεις. Ωστόσο, η παραπάνω διαδικασία αυξάνει την πολυπλοκότητα της διεπαφής, ενώ ταυτόχρονα μπορεί να οδηγήσει σε λάθη. Για το σκοπό αυτό, λαμβάνοντας υπόψη τη σημασία των παραμέτρων αυτών και κυρίως του γεγονότος ότι παράμετροι των παραπάνω κλάσεων συνήθως χρησιμοποιούνται μαζί (π.χ., μία ποσότητα περιγράφεται πλήρως, εάν και εφόσον η τιμή που παρέχουμε ακολουθείται από τη μονάδα μέτρησης) επιτρέπουμε στους χρήστες να ορίσουν ταυτόχρονα τις τιμές τους, χωρίς να αυξάνεται η πολυπλοκότητα της διεπαφής. Στο σημείο αυτό, στην περίπτωση που οι τιμές μιας

παραμέτρου προέρχονται από ένα συγκεκριμένο σύστημα κατηγοριοποίησης (καθορισμένο σύνολο όρων οι οποίοι επιπρόσθετα είναι ιεραρχικά οργανωμένοι) το σύστημα επιτρέπει στους χρήστες να ορίσουν γεμάτα νόημα κριτήρια, χρησιμοποιώντας τους σημασιολογικούς τελεστές που υποστηρίζονται από το σύστημά μας. Οι σημασιολογικοί τελεστές (δυαδικοί τελεστές που βασίζονται στη σημασία των οντοτήτων που ορίζονται), σε συνδυασμό με την ιεραρχία των κλάσεων, επιτρέπουν στους χρήστες να αναζητήσουν τις οντότητες εκείνες στις οποίες οι αντίστοιχες τιμές των παραμέτρων τους είναι σημασιολογικά ισοδύναμες με αυτές που έχει ρητά ορίσει ο χρήστης ή έχουν ευρύτερη ή στενότερη σημασία.

Κάθε φορά που ο χρήστης εισάγει ένα νέο κριτήριο, το σύστημα εξετάζει τα κριτήρια που έχουν ήδη οριστεί για πιθανές ασυνέπειες. Για παράδειγμα, εάν υπάρχει ένας περιορισμός πληθικότητας (cardinality restriction), όπως στην περίπτωση της ηλικίας και του φύλου (max 1 property), γενικά, δεν μπορούμε να ορίσουμε παραπάνω από ένα κριτήρια για την παράμετρο αυτή. Στην περίπτωση αυτή, θα ήταν άσκοπο να αναζητήσουμε τους ανθρώπους οι οποίοι είναι μικρότεροι από 35 χρονών (κριτήριο 1) και ταυτοχρόνως μεγαλύτεροι από 50 χρονών (κριτήριο 2). Στην πραγματικότητα, εάν υπάρχουν παραπάνω από ένα κριτήρια, τότε θα πρέπει η τομή τους να μην είναι το κενό σύνολο. Ωστόσο, αυτό δεν ισχύει, όταν μπορούμε να έχουμε παραπάνω από μια παραμέτρους μιας οντότητας, όπως στην περίπτωση των διαγνώσεων και των φαρμάκων, καθώς ένας άνθρωπος μπορεί να πάσχει από αρκετά προβλήματα (ακόμη και συγχρόνως) και να λαμβάνει παραπάνω από ένα φάρμακα. Στην περίπτωση αυτή, ο έλεγχος εστιάζει στον εντοπισμό αντιφατικών κριτηρίων καταλληλότητας με την προϋπόθεση ότι οι αντίστοιχοι όροι είναι ιεραρχικά ταξινομημένοι. Για παράδειγμα, είναι αντιφατικό να ψάχνουμε για τους ασθενείς που διαγνώστηκαν με οξύ έμφραγμα του μυοκαρδίου (AMI) και ταυτοχρόνως δεν πάσχουν από κάποια ισχαιμική ασθένεια, καθώς το AMI είναι

επίσης μια ισχαιμική ασθένεια, όπως μπορούμε να δούμε στο Σχήμα 25. Ωστόσο, θα πρέπει να εξετάσουμε με προσοχή τα δεδομένα που παρέχει ο χρήστης, καθώς οι διαγνώσεις μπορεί να αναφέρονται σε διαφορετική χρονική περίοδο. Στην περίπτωση αυτή, εφόσον οι χρονικές περίοδοι δε συμπίπτουν, δεν υπάρχει κάποια αντίφαση.

19.3Αλληλεπίδραση με το Χρήστη και Καταγραφέντα Δεδομένα

Στον Πίνακας 10 έχουμε συνοψίσει τις ενέργειες/επεμβάσεις του χρήστη κατά τον ορισμό των κριτηρίων καταλληλότητας και κυρίως τα δεδομένα που καταγράφονται από το σύστημα καθώς επίσης και τις αλλαγές που λαμβάνουν χώρα στο γραφικό περιβάλλον, σε σύγκριση με τα HTML στοιχεία που υπάρχουν ήδη (από το προηγούμενο βήμα) εκεί. Σημειώνουμε ότι το URI των παραμέτρων καθώς επίσης και οι περιορισμοί που τις συνοδεύουν αποθηκεύονται είτε στη λίστα με τα κριτήρια εισαγωγής είτε στη λίστα με τα κριτήρια εξαγωγής, ανάλογα τον τύπο του κριτηρίου που ορίζεται από το χρήστη.

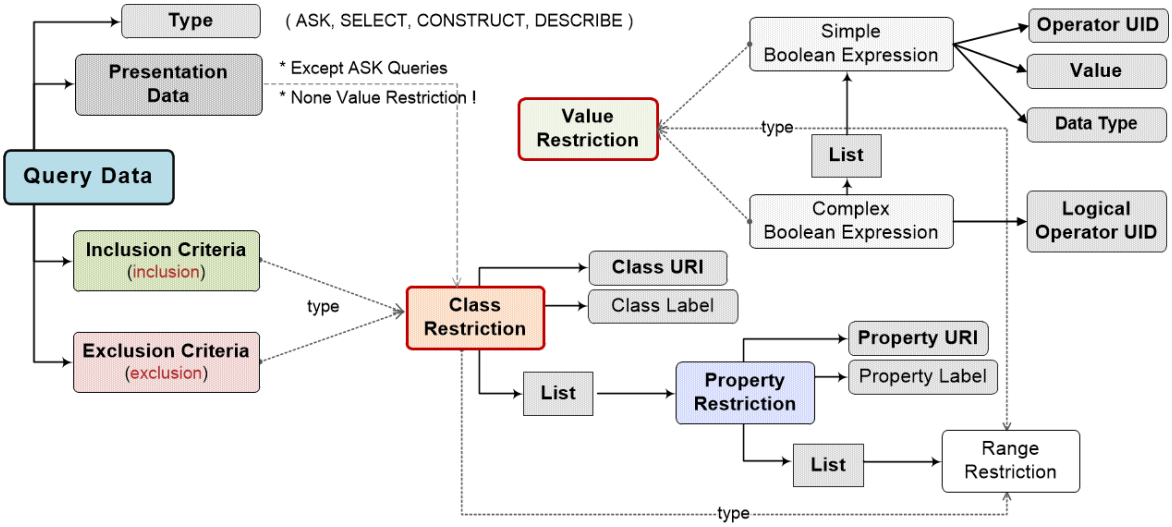
Η διαδικασία σχηματισμού ενός ερωτήματος ξεκινά με την επιλογή του τύπου των δεδομένων προς αναζήτηση (π.χ. Δεδομένα Ασθενών) – ενέργεια UA1 – και συνεχίζει με τον ορισμό των κριτηρίων καταλληλότητας. Στην περίπτωση των datatype χαρακτηριστικών, το URI της παραμέτρου (π.χ. ηλικίας) καθώς επίσης και η συνθήκη που θα πρέπει αυτή να ικανοποιεί (π.χ., να είναι μεγαλύτερη από 35) καταγράφονται από το σύστημα, καθώς ο χρήστης πραγματοποιεί τις ενέργειες UA2, UA3, UA4, και UA5. Επίσης, για κάθε object παράμετρο, το URI της (π.χ., συνδεδεμένη με τον ασθενή διάγνωση) μαζί με τον τύπο στον οποίο θα πρέπει να ανήκουν τα δεδομένα (π.χ., Δεδομένα Διάγνωσης), ακολουθούμενα από τις συνθήκες που θα πρέπει να ικανοποιούν οι παράμετροι της κλάσης αυτής (π.χ., λογικές εκφράσεις σχετικές με το όνομα/κωδικό του προβλήματος, ημερομηνία που έλαβε χώρα, κτλ), καταγράφονται από το σύστημα. Για το σκοπό αυτό, οι ενέργειες UA2, UA3 και UA4 μπορεί να λάβουν χώρα αρκετές

φορές, ενώ η όλη διαδικασία τελειώνει με την ενέργεια UA5, όπου ο χρήστης ορίζει αν το κριτήριο αυτό θα είναι κριτήριο εισαγωγής ή εξαγωγής.

Ενέργεια Χρήστη	**Δεδομένα**	**Αλλαγές στο Γραφικό Περιβάλλον**
ID: UA1 Class Selection (*on mouse click*)	Class-URI	Το μεσαίο τμήμα ανανεώνεται, παρουσιάζοντας τις παραμέτρους της κλάσης που επιλέχθηκε καθώς επίσης και τον τύπο των δεδομένων για καθεμία από αυτές.
ID: UA2 Πρόταση Περιορ. Τιμής Παραμετρ. (*on mouse hover*)	-	Η περιοχή που βρίσκεται δίπλα από κάθε παράμετρο ανανεώνεται προσωρινά, παρουσιάζοντας τους περιορισμούς που μπορεί να χρησιμοποιηθούν.
ID: UA3 Επιλογή Περιορ. Τιμής Παραμέτρ. (*on mouse click*)	Property-URI, Restriction-ID, Class-URI και/ή Datatype-URI	<u>Περίπτωσης Α: «Πρωτόγονοι» Τύποι Δεδομένων</u> Η περιοχή που βρίσκεται στη δεξιά μεριά των παραμέτρων ανανεώνεται με τους κατάλληλους τελεστές και «input» πεδία, ανάλογα με τον περιορισμό που έχει επιλεχθεί. <u>Case B: «Περίπλοκη» Δομή Δεδομένων</u> Εισάγεται μία περιοχή κάτω από την αντίστοιχη παράμετρο που περιέχει τις παραμέτρους της κλάσης (πεδίο τιμών) που έχει επιλεχθεί καθώς επίσης και τα αξιώματα που έχουν οριστεί (π.χ. τύπο τιμών).
ID: UA4 Ορισμός Τιμών (*type or selection*)	-	Τα «input» πεδία γεμίζουν με τα δεδομένα που παρέχει ο χρήστης είτε επιλέγοντας την επιθυμητή τιμή είτε πληκτρολογώντας την.
ID: UA5 Εισαγωγή Κριτηρίου (*on mouse click*)	Input Data: Operators Values	Το γραφικό περιβάλλον επανέρχεται στην αρχική του μορφή, επιτρέποντας στους χρήστες να ορίσουν επιπρόσθετα κριτήρια για τις υπόλοιπες παραμέτρους της κλάσης των δεδομένων που έχει επιλεχθεί στην αρχή (UA1).

Πίνακας 10: Αλληλεπίδραση του χρήστη με τη Διεπαφή για τον ορισμό των Κριτηρίων Καταλληλότητας

Τα δεδομένα που καταγράφονται κατά την παραπάνω διαδικασία σχετικά με τα Κριτήρια Καταλληλότητας αλλά και τα Δεδομένα προς Αναζήτηση, αποθηκεύονται εσωτερικά σε JSON μορφή στην πλευρά (περιηγητή) του κάθε χρήστη. Αρχικά, καταγράφεται το URI της κλάσης που έχει επιλεχθεί και ακολούθως, κάθε φορά που ορίζουμε ένα νέο κριτήριο εισαγωγής ή εξαγωγής, καταγράφουμε το πλήρες μονοπάτι που θα πρέπει να διασχίσουμε καθώς και τις συνθήκες που θα πρέπει να πληρούν τα δεδομένα μας. Στο Σχήμα 22 παρουσιάζουμε τη δομή των δεδομένων που αποθηκεύονται εσωτερικά από το σύστημα. Σημειώνουμε ότι η ίδια δομή χρησιμοποιείται όχι μόνο για τα κριτήρια αλλά και για τα δεδομένα που ψάχνουμε (π.χ., στην περίπτωση των SELECT SPARQL ερωτημάτων). Όμως, στην περίπτωση αυτή δεν καταγράφεται κάποιος περιορισμός όσον αφορά την τιμή της αντίστοιχης παραμέτρου.



Σχήμα 22: Δομή των δεδομένων που καταγράφονται μέσω της Διαδικτυακής Επαφής του συστήματος Οπτικής Έκφρασης Ερωτημάτων

20 *Παραγωγή SPARQL Ερωτημάτων*

20.1Δομή του SRARQL ερωτήματος

Η τυπική έκφραση του ερωτήματος σε SPARQL παράγεται αυτόματα από το σύστημα με βάση τα JSON δεδομένα (Σχήμα 22) που έχουν καταγραφεί από το σύστημα, ως αποτέλεσμα των δεδομένων που έχει δώσει ο χρήστης μέσω του γραφικού περιβάλλοντος. Η διαδικασία αυτή λαμβάνει χώρα στη μεριά του εξυπηρετητή, λαμβάνοντας υπόψη τη σημασία των όρων και ειδικότερα των κριτηρίων που έχουν οριστεί (υποενότητα X). Το τμήμα προβολής (production part) του SPARQL ερωτήματος εξαρτάται από τον τύπο του SPARQL ερωτήματος που θέλουμε να παράγουμε (αποτελεί μέρος της παραμετροποίησης του συστήματος), ενώ το τμήμα με τους περιορισμούς (restriction part) και πιο συγκεκριμένα το «WHERE clause» είναι σε κάθε περίπτωση το ίδιο.

```
/** Automatically Generated SELECT SPARQL Query */
SELECT Export-Select-Variables( SELECT-BGP ) WHERE {
        # Presentation Data
        Produce-GPF ( DATA.SELECT_DATA , ?entityVar )
        # Inclusion Criteria
        Produce-GPF ( DATA.INCLUSION_DATA , ?entityVar )
        # Exclusion Criteria
        FILTER NOT EXISTS {
            Produce-GPF ( DATA.EXCLUSION_DATA , ?entityVar )
        }
}
```

Σχήμα 23: Η δομή του αυτομάτως παραγόμενου SELECT SPARQL ερωτήματος

Το Σχήμα 23 παρουσιάζει τη δομή του αυτομάτως παραγόμενου SELECT SPARQL ερωτήματος. Το «SELECT clause» καθορίζει τα δεδομένα προς παρουσίαση και αποτελείται από τις μεταβλητές εκείνες του «WHERE clause» που αντιστοιχούν στις τιμές των παραμέτρων που ψάχνουμε. Το «WHERE clause» περιέχει επιπρόσθετα την τυπική έκφραση των συνθηκών που πρέπει να πληρούν τα δεδομένα που ψάχνουμε. Δεδομένου ότι τα κριτήρια καταλληλότητας είναι οργανωμένα σε δύο κατηγορίες, το «WHERE clause» περιέχει δύο επιπρόσθετα γραφικά μοτίβα. Το πρώτο αντιστοιχεί στα κριτήρια εισαγωγής, ενώ το δεύτερο στα κριτήρια εξαγωγής. Και στις δύο περιπτώσεις τα γραφικά μοτίβα αποτελούνται από μία ή περισσότερες τριάδες (triple patterns), ακολουθούμενα από ένα ή περισσότερα φίλτρα (filter clause) που εκφράζουν τους περιορισμούς που πρέπει να ικανοποιούν τα δεδομένα μας. Ωστόσο, στην περίπτωση των κριτηρίων εξαγωγής, οι οντότητες δε θα πρέπει να ικανοποιούν ούτε ένα από τα κριτήρια/περιορισμούς που έχουν οριστεί. Αυτό εκφράζεται στο SPARQL ερώτημα μέσω της χρησιμοποίησης του «FILTER NOT EXISTS» τελεστή που είναι διαθέσιμος στην πιο πρόσφατη έκδοση της SPARQL (version 1.1). Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε το μοτίβο «OPTIONAL FILTER ! BOUND», αλλά θα πρέπει επιπρόσθετα να ορίσουμε και τις αντίστοιχες μεταβλητές που θα πρέπει να χρησιμοποιηθούν στη συνάρτηση “BOUND”. Σημειώνουμε ότι τα τρία διαφορετικά τμήματα του «WHERE clause» του SPARQL ερωτήματος αναφέρονται στην ίδια οντότητα και, συνεπώς, τα δεδομένα που παρουσιάζονται ικανοποιούν και τα κριτήρια εισαγωγής και τα κριτήρια εξαγωγής.

20.2Παραγωγή Γραφικών Μοτίβων και Φίλτρων

Το πιο ενδιαφέρον κομμάτι κατά την παραγωγή του SPARQL ερωτήματος είναι ο αλγόριθμος που χρησιμοποιείται για την έκφραση των συνθηκών που θα πρέπει να

πληρούν τα δεδομένα μας. Ο αλγόριθμος αυτός (Σχήμα 24) λαμβάνει το όνομα της μεταβλητής και κυρίως τα JSON δεδομένα που εκφράζουν τις συνθήκες που θα πρέπει αυτά να πληρούν και ακολούθως, παράγει το αντίστοιχο γραφικό μοτίβο (graph pattern) καθώς και τα αντίστοιχα φίλτρα (filter clause).

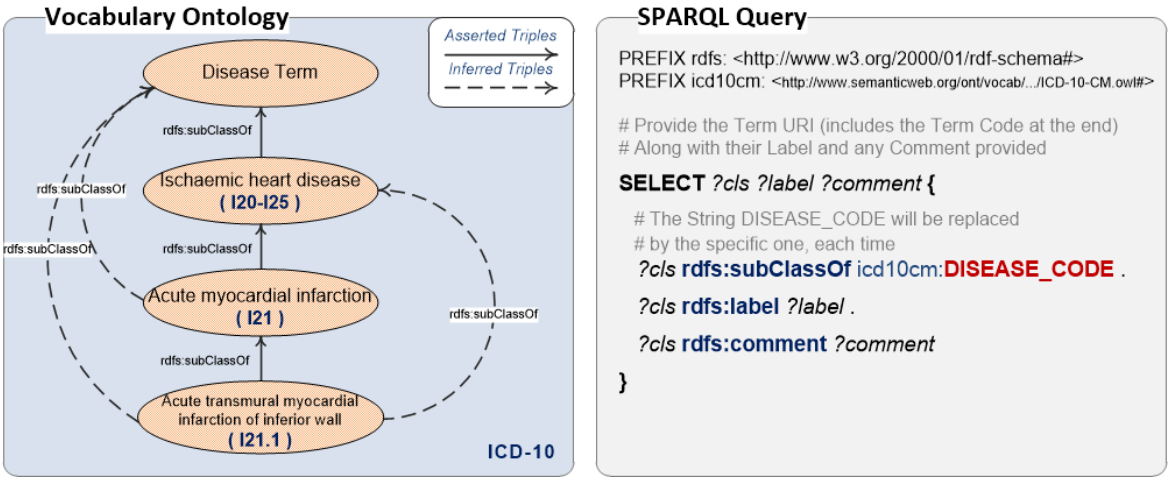
```
/** Produces a Graph Pattern along with zero or more FILTER clauses */
String Produce-GPF ( DATA , VARIABLE ) {
        IF ( DATA IS NULL ) return "" ; // i.e., An empty String
        IF ( Type(DATA) IS Class-Restriction ) {
                IF ( DATA IS-NOT EMPTY ) GPF += . ;
            GPF += VARIABLE rdf:type DATA.CLASS-URI ;
            For-each: DATA.RESTRICTION {
                    IF ( DATA.RESTRICTION[i].PROPERTY-URI IS OPTIONAL ) GPF += " OPTIONAL { "
                GPF += . VARIABLE DATA.RESTRICTION[i].PROPERTY-URI NEW-VARIABLE ;
                GPF += . Produce-GPF ( DATA.RESTRICTION[i].RANGE , NEW-VARIABLE ) ;
                IF ( DATA.RESTRICTION[i].PROPERTY-URI IS OPTIONAL ) GPF += " } "
            }
            return GPF;
        } ELSE {  // Type(DATA) IS Value-Restriction
            IF ( Type(DATA) IS Simple-Boolean-Expression ) {
                return FILTER ( VARIABLE DATA.COMPARISON_OPERATOR DATA.VALUE ) ;
            } ELSE IF ( Type(DATA) IS Complex-Boolean-Expression ) { // AND / OR
                return FILTER ( Produce-GPF ( DATA.EXPR1 , VARIABLE )
                                DATA.LOGICAL_OPERATOR
                                Produce-GPF ( DATA.EXPR2 , VARIABLE ) ) ;
            } ELSE IF ( Type(DATA) IS Special-Boolean-Expression ) {    // Semantic Operators
                return FILTER ( Produce-BE ( DATA.OPERATOR , DATA.VALUE, VARIABLE ) ) ;
            } ELSE {
                    return NULL ; // None condition specified !
            }
        }
}
```

Σχήμα 24: Αλγόριθμος για την αυτόματη παραγωγή των Γραφικών Μοτίβων και των αντίστοιχων Φίλτρων με βάση τα JSON δεδομένα που έχουν καταγραφεί

Ο αλγόριθμος κατασκευάζει σταδιακά το γραφικό μοτίβο με τα αντίστοιχα φίλτρα εισάγοντας είτε τριάδες είτε φίλτρα, ανάλογα με τον τύπο των δεδομένων που του παρέχουμε. Πιο συγκεκριμένα, στην περίπτωση των «Class-Restriction» ο αλγόριθμος εισάγει τις τριάδες που καθορίζουν τον τύπο των οντοτήτων καθώς και τις παραμέτρους

που μας ενδιαφέρουν, υπό την έννοια ότι οι τιμές τους είτε θα χρησιμοποιηθούν στα φίλτρα (στην περίπτωση των "data-type" χαρακτηριστικών) είτε χρησιμοποιούνται ως ένα ενδιάμεσο βήμα, για να φτάσουμε σε αυτές (στην περίπτωση των "object-type" χαρακτηριστικών). Στην περίπτωση των «Value-Restriction» ο αλγόριθμος εισάγει το αντίστοιχο φίλτρο που εκφράζει το επιθυμητό εύρος ή σύνολο τιμών στο οποίο θα πρέπει να ανήκουν τα δεδομένα μας. Σημειώνουμε ότι ο αλγόριθμος είναι αναδρομικός και, συνεπώς, η παραπάνω διαδικασία επαναλαμβάνεται αρκετές φόρες για την παραγωγή του γράφου και των συνθηκών που πρέπει να πληρούν τα δεδομένα που ψάχνουμε.

Ο αλγόριθμος λαμβάνει επίσης υπόψη τους περιορισμούς πληθικότητας (cardinality restriction) που πιθανώς έχουν χρησιμοποιηθεί κατά τον ορισμό των παραμέτρων του μοντέλου αναφοράς. Πιο συγκεκριμένα, εάν μία παράμετρος είναι προαιρετική (max cardinality restriction: one), και οι τριάδες αλλά και τα Φίλτρα που ορίζονται για τη συγκεκριμένη παράμετρο τοποθετούνται μέσα σε "OPTIONAL FILTER clauses" έτσι, ώστε η αποτίμηση των κριτηρίων να λάβει χώρα μόνο, όταν υπάρχουν οι αντίστοιχες τιμές. Ιδιαίτερη προσοχή θα πρέπει να δοθεί, όταν χρησιμοποιείται κατά την έκφραση των κριτηρίων ένας σημασιολογικός τελεστής. Στην περίπτωση αυτή, το σύστημα βρίσκει τους «αποδεκτούς» πιθανούς όρους, λαμβάνοντας υπόψη τις τιμές που έχει ορίσει ο χρήστης, την κατηγοριοποίηση των εννοιών αλλά και τους σημασιολογικούς τελεστές που έχουν χρησιμοποιηθεί, και ακολούθως τις τοποθετεί και τις ενσωματώνει στα φίλτρα. Για το σκοπό αυτό, κατά την υλοποίηση του συστήματος χρησιμοποιήθηκε ο Pellet Reasoner [64] για τη δημιουργία του «πλήρους» γράφου με βάση τα αξιώματα που είχαν οριστεί στην Οντολογία Όρων (αυτό γίνεται την πρώτη φορά που συναντάμε κάποιο σημασιολογικό τελεστή) και στη συνέχεια, το σύστημα βρίσκει τους επιθυμητούς όρους (π.χ., όρους με στενότερη σημασία από την υπάρχουσα) με τη βοήθεια των SPARQL ερωτημάτων.

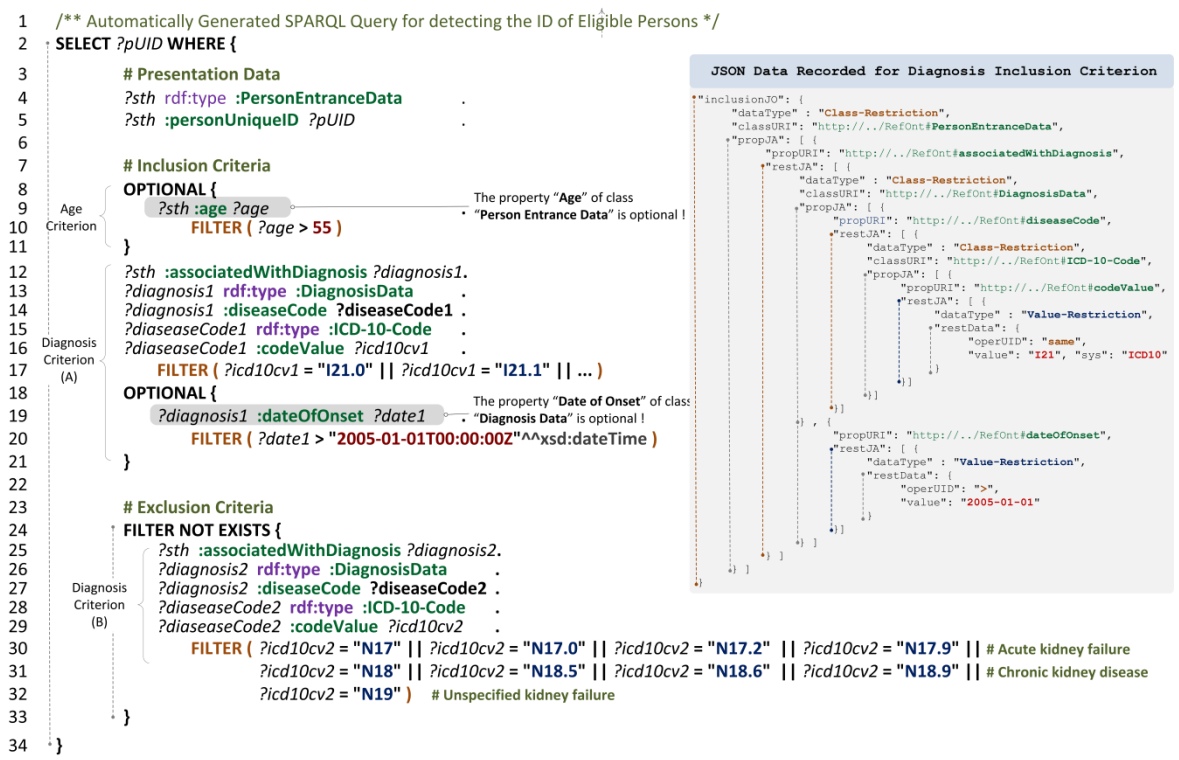


Σχήμα 25: (α) Ο αρχικός (asserted triples) και παραγόμενος (inferred triples) γράφος για ένα υποσύνολο της ICD οντολογίας καθώς και (β) το SPARQL ερώτημα για την εύρεση των όρων με «στενότερη» σημασία.

Το Σχήμα 25 απεικονίζει τον αρχικό και παραγόμενο (inferred) γράφο για ένα πολύ μικρό υποσύνολο ασθενειών που έχουν οριστεί στην οντολογία. Στο σχήμα αυτό παρουσιάζεται επίσης και το SPARQL ερώτημα που θα χρησιμοποιηθεί, για να αντλήσουμε την κατάλληλη πληροφορία από το γράφο αυτό. Όπως μπορούμε να δούμε, το ερώτημα αυτό είναι εύκολα παραμετροποιήσιμο, ώστε να μπορεί να χρησιμοποιηθεί για διαφορετικούς όρους της οντολογίας των ασθενειών.

20.3Ένα παράδειγμα

Σε αυτή την υποενότητα θα παρουσιάσουμε το SPARQL ερώτημα που παράγεται αυτόματα από το σύστημα με βάση τα δεδομένα που έχει δώσει ο χρήστης (εσωτερικά αναπαριστώμενα σε JSON μορφή) μέσω του γραφικού περιβάλλοντος (Σχήμα 19). Πιο συγκεκριμένα, ας υποθέσουμε ότι ο χρήστης είχε ορίσει τα ακόλουθα κριτήρια: (α) Οι ασθενείς θα πρέπει να είναι μεγαλύτεροι από 55 χρονών (δημογραφικά χαρακτηριστικά). (β) Οι ασθενείς θα πρέπει να έχουν διαγνωστεί με Έμφραγμα του Μυοκαρδίου (MI) την προηγούμενη εβδομάδα (κριτήριο εισαγωγής), ενώ δεν πάσχουν από Νεφρική Ανεπάρκεια (κριτήριο εξαγωγής). Επίσης υποθέτουμε ότι ο χρήστης θα ήθελε να εντοπίσει τους ασθενείς (αναγνωριστικό του καθενός) που πληρούν τα κριτήρια αυτά.



Σχήμα 26: Το αυτομάτως παραγόμενο SPARQL ερώτημα για την εύρεση των ασθενών με βάση τα κριτήρια εισαγωγής/εξαγωγής που έχουν οριστεί

Το Σχήμα 26 απεικονίζει το παραγόμενο SPARQL ερώτημα με βάση τα JSON δεδομένα, ένα μικρό μέρος των οποίων επίσης παρουσιάζεται στο σχήμα αυτό. Για λόγους παρουσίασης, θα παρουσιάσουμε συνοπτικά τη διαδικασία που ακολουθήθηκε για την παραγωγή του αντίστοιχου γράφου και φίλτρων μόνο για την περίπτωση του κριτηρίου εισαγωγής που σχετίζεται με τις διαγνώσεις των ασθενών.

Ο αλγόριθμος αρχικά εισάγει μία τριάδα που ορίζει τον τύπο των οντοτήτων (γραμμή 4) και έπειτα εισάγει μία άλλη τριάδα που εκφράζει τη συσχέτιση του ασθενούς με τα δεδομένα διάγνωσης (γραμμή 12). Στην συνέχεια, δεδομένου ότι υπάρχει μία άλλη «Class Restriction», ο αλγόριθμος εισάγει την τριάδα που ορίζει τον τύπο των οντοτήτων αυτών (γραμμή 13) καθώς και τις τριάδες που αναφέρονται στον κωδικό του «προβλήματος» (γραμμή 14) και την ημερομηνία που συνέβη (γραμμή 19). Μιας και η ημερομηνία που ξέσπασε το πρόβλημα είναι προαιρετική, οι αντίστοιχες τριάδες και φίλτρα θα τοποθετηθούν μέσα σε «OPTIONAL clauses» (γραμμές 18 με 21). Σχετικά με

τον κωδικό της διάγνωσης, εισάγεται μία τριάδα που εκφράζει ότι αυτή προέρχεται από έναν ICD-10 κωδικό (γραμμή 15) και μία ακόμη τριάδα που αναφέρεται στον κωδικό αυτό (γραμμή 16). Δεδομένου ότι ο κωδικός ακολουθείται από μία «Value Restriction», εισάγεται το κατάλληλο φίλτρο. Στο σημείο αυτό τονίζουμε ότι έχει χρησιμοποιηθεί ένας σημασιολογικός τελεστής, και επομένως το σύστημα πρώτα βρίσκει τους κωδικούς των ιατρικών συνθηκών που ανήκουν στην κατηγορία του εμφράγματος του μυοκαρδίου και κατόπιν εισάγει τους κωδικούς αυτούς στο φίλτρο, χρησιμοποιώντας το λογικό τελεστή «OR».

Σημειώνουμε ότι το «OPTIONAL» έχει χρησιμοποιηθεί δύο φορές στο SPARQL ερώτημα, καθώς και η ηλικία ενός ασθενούς αλλά και η ημερομηνία που ξέσπασε το πρόβλημα είναι προαιρετικές. Συνεπώς, οι συνθήκες αυτές θα εξεταστούν μόνο για τους ασθενείς που έχουν τα παραπάνω δεδομένα. Σχετικά με τις διαγνώσεις, οι ασθενείς που θα επιλεχθούν θα πρέπει να έχουν διαγνωστεί με Έμφραγμα του Μυοκαρδίου (ICD-10 κωδικός: I21), αλλά να μην έχουν διαγνωστεί με Νεφρική Ανεπάρκεια (ICD-10 κατηγορία/κωδικοί: N17-N19) στο παρελθόν. Στην περίπτωση αυτή, όπως μπορούμε να δούμε και στο σχήμα, εκτός από τη μεταβλητή που αναφέρεται στον ασθενή, όλες οι άλλες μεταβλητές είναι διαφορετικές στα δύο αυτά κριτήρια. Τέλος, σημειώνουμε ότι επιπρόσθετοι όροι/κωδικοί έχουν τοποθετηθεί στα παραπάνω κριτήρια. Ειδικότερα για το κριτήριο εξαγωγής που σχετίζεται με τη νεφρική ανεπάρκεια (μια αρκετά ευρεία κατηγορία) συμπεριλάβαμε αρκετούς όρους έτσι, ώστε να εξάγουμε τους ασθενείς που έχουν διαγνωστεί με Οξεία Νεφρική Ανεπάρκεια (ICD-10 κωδικοί: N17, N17.0, N17.1, N17.2, N17.9), Χρόνια Νεφρική Ανεπάρκεια (ICD-10 κωδικοί: N18, N18.5, N18.6, N18.9) και Μη ορισμένη Νεφρική Ανεπάρκεια (ICD-10 κωδικός: N19).

Η σελίδα αυτή είναι σκόπιμα λευκή

21 *Συζήτηση και Μελλοντική Εργασία*

21.1Χαρακτηριστικά του Συστήματος Έκφρασης Ερωτημάτων

Ένα οπτικό σύστημα έκφρασης ερωτημάτων, σαν αυτό που παρουσιάσαμε στην εργασία αυτή, είναι ένα σύστημα που παράγει τα επιθυμητά ερωτήματα, παρά μία οπτική γλώσσα έκφρασης ερωτημάτων, η οποία βασίζεται στο σαφή ορισμό της σημασιολογίας της γλώσσας με μία οπτική σημειογραφία και σύνταξη [147]. Συνεπώς, το σύστημά μας είναι λιγότερο εκφραστικό από το φορμαλισμό που χρησιμοποιείται για την τυπική έκφραση ερωτημάτων, υπό την έννοια ότι ένα υποσύνολο των ερωτημάτων που μπορούμε να εκφράσουμε με SPARQL υποστηρίζεται από το σύστημά μας. Παρόλα αυτά, το σύστημα που αναπτύχθηκε παρέχει ένα γραφικό περιβάλλον το οποίο είναι αρκετά κατανοητό από το χρήστη και εύκολο στη χρήση του, ακόμη και για αρχάριους σε ένα συγκεκριμένο πεδίο γνώσης, επιτρέποντάς τους να περιγράψουν επακριβώς τα δεδομένα που ψάχνουν χωρίς, να είναι απαραίτητο να διαθέτουν τεχνικά προσόντα.

Το σύστημα που αναπτύχθηκε υποστηρίζει μόνο ένα περιορισμένο αριθμό χαρακτηριστικών από αυτά που θα πρέπει να υποστηρίζει ένα οπτικό σύστημα έκφρασης ερωτημάτων το οποίο βασίζεται σε οντολογίες [120]. Αυτό οφείλεται στο γεγονός ότι στην εργασία μας εστιάσαμε στην περιγραφή της διαδικασίας έκφρασης ερωτημάτων, αγνοώντας κάποια άλλα χαρακτηριστικά που θα πρέπει να παρέχονται, όπως αποθήκευση και ανάκληση των ερωτημάτων που ορίζονται από τους χρήστες έτσι, ώστε να μπορούν

εύκολα να επαναχρησιμοποιηθούν από τους χρήστες. Το χαρακτηριστικό αυτό είναι ιδιαίτερα χρήσιμο, καθώς, σε συνδυασμό με το πεδίο γνώσης του κάθε χρήστη, θα μπορούσε να συμβάλει στην ανάλυση των ερωτημάτων τους και την εξαγωγή χρήσιμων συμπερασμάτων (π.χ., παράμετροι των οντοτήτων που χρησιμοποιούνται κατά κόρον). Ωστόσο, το σύστημά μας παρέχει τη σύνοψη των ερωτημάτων που σχηματίζονται από το χρήστη, είναι ανεξάρτητο του πεδίου γνώσης που καλύπτουν οι οντολογίες και έχει υλοποιηθεί ακολουθώντας μία υπηρεσιοστρεφή αρχιτεκτονική.

Επίσης, το σύστημα (σε συνδυασμό με τις οντολογίες) θα μπορούσε να παρέχει επιπρόσθετη πληροφορία για τις οντότητες που ανήκουν σε ένα συγκεκριμένο πεδίο γνώσης, για να προβλέψει ή να εμποδίσει την έκφραση ερωτημάτων που δεν επιστρέφουν κανένα αποτέλεσμα [148]. Για παράδειγμα, το σύστημα θα μπορούσε να παρουσιάζει τον αριθμό των οντοτήτων που ανήκουν σε κάθε κλάση του Μοντέλου Αναφοράς, με την προϋπόθεση ότι υπήρχε η αντίστοιχη πληροφορία στην οντολογία. Επίσης, το σύστημα θα μπορούσε να ενημερώνει τους χρήστες για τις παραμέτρους με ιδιαίτερη σημασία για κάθε κλάση. Ωστόσο, αυτό προϋποθέτει ότι έχει γίνει ανάλυση των σχετικών εγγράφων που είναι διαθέσιμα στο διαδίκτυο έτσι, ώστε να εντοπίσουμε τις παραμέτρους αυτές, και κατόπιν τις επισημάνουμε κατά το σχεδιασμό των οντολογιών. Στην εργασία μας, υποθέσαμε ότι οι χρήστες της εφαρμογής είναι ειδικοί στο πεδίο γνώσης που καλύπτει η εφαρμογή, και συγκεκριμένα οι οντολογίες, και συνεπώς γνωρίζουν τις παραμέτρους που τους ενδιαφέρουν και ειδικότερα τις τιμές που θα πρέπει αυτές να έχουν. Παρόλα αυτά το σύστημα μπορεί επίσης να χρησιμοποιηθεί από κάθε χρήστη του διαδικτύου, όπως έχει ήδη αναφερθεί.

21.2Σημασιολογικά ενεργή Εξόρυξη Γνώσης και Απουσία Δεδομένων

Οι οντολογίες έχουν έναν διακριτό ρόλο στην εργασία μας. Επιτρέπουν στους χρήστες να αλληλεπιδρούν με τις πηγές δεδομένων σε εννοιολογικό επίπεδο, λαμβάνοντας υπόψη τη σημασία των όρων παρά τη συμβολοακολουθία που χρησιμοποιείται για την αναπαράστασή τους. Για το λόγο αυτό, δεδομένου ότι η επικοινωνία του χρήστη με το σύστημα για την έκφραση των ερωτημάτων βασίζεται σε οντολογίες, θα πρέπει να λάβουμε υπόψη τη σημασία των όρων κατά την τυπική έκφραση των ερωτημάτων στη γλώσσα SPARQL.

Εκ των προτέρων θα πρέπει να τονίσουμε ότι η αποτίμηση των SPARQL ερωτημάτων βασίζεται στις τριάδες (εκφράζουν γεγονότα που ισχύουν στον κόσμο μας) που έχουν ρητά οριστεί στον RDF γράφο και όχι σε αυτές που θα μπορούσαμε να εξάγουμε, εάν λαμβάναμε υπόψη τη σημασία των όρων [149]. Συνεπώς, τα SPARQL ερωτήματα μπορεί να μην επιστρέφουν όλα τα σημασιολογικά σωστά αποτελέσματα, παρά το γεγονός ότι η σημασία τους έχει οριστεί στις οντολογίες. Το θέμα αυτό μπορεί να επιλυθεί χρησιμοποιώντας έναν «reasoner», για να εντοπίσουμε όλα τα γεγονότα (τριάδες) που ισχύουν για τον κόσμο μας και κατόπιν να εκτελέσουμε τα SPARQL ερωτήματα. Εναλλακτικά, μπορούμε να εμπλουτίσουμε το παραγόμενο ερώτημα με επιπρόσθετους όρους έτσι, ώστε να λαμβάνει υπόψη όχι μόνο την πληροφορία που υπάρχει ήδη αλλά και αυτήν που θα μπορούσε να προκύψει. Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη, ειδικά στις περιπτώσεις όπου υπάρχει ένας μεγάλος όγκος δεδομένων, που καθιστά την παραγωγή και διατήρηση του παραγόμενου γράφου ασύμφορη.

Στο εργαλείο που αναπτύχθηκε, επιλέξαμε τη δεύτερη προσέγγιση, εισάγοντας επιπρόσθετους όρους στα φίλτρα κατά την παραγωγή του SPARQL ερωτήματος, με βάση την ιεραρχία των όρων στην Οντολογία. Στο μέλλον, θα μπορούσαμε να

εμπλουτίσουμε το παραγόμενο ερώτημα και με επιπλέον τριάδες, λαμβάνοντας υπόψη περισσότερα αξιώματα που υπάρχουν τόσο στην Οντολογία Όρων όσο και στο Μοντέλο Αναφοράς.

Παρά το γεγονός ότι η αποτίμηση των SPARQL ερωτημάτων βασίζεται στα γεγονότα που ρητά αναφέρονται στον RDF γράφο, η SPARQL παρέχει ορισμένους τελεστές που επιτρέπουν στους χρήστες (σε κάποιες περιπτώσεις) να χειριστούν αποτελεσματικά την έλλειψη δεδομένων. Για παράδειγμα, χρησιμοποιώντας «OPTIONAL clauses» η αποτίμηση ενός κριτηρίου για την ηλικία του ασθενούς θα ελεγχθεί μόνο, εάν και εφόσον υπάρχουν τα δεδομένα αυτά για έναν ασθενή. Ωστόσο, εάν η ηλικία του ασθενούς δε συμπεριληφθεί στα αποτελέσματα ή χρησιμοποιηθεί κάποια συνάρτηση συνάθροισης (aggregation function), δε θα μπορούμε να γνωρίζουμε εάν, όντως, ένας ασθενής ικανοποιεί το κριτήριο ή όχι ή τον ακριβή αριθμό των ασθενών που το ικανοποιούν, εξετάζοντας μόνο τα αποτελέσματα που μας επιστρέφονται από την εκτέλεση του SPARQL ερωτήματος.

Βέβαια, η απουσία μιας ή περισσοτέρων τριάδων, δε σημαίνει απαραίτητα έλλειψη δεδομένων. Για παράδειγμα, το φύλο ενός ασθενούς μπορεί να είναι προαιρετικό (ο περιορισμός πληθικότητας είναι το πολύ ένα), αλλά η έλλειψη των αντίστοιχων τριάδων να μη δηλώνει απαραίτητα ότι δε γνωρίζουμε την τιμή του. Για παράδειγμα (σενάριο 1), εάν οι πιθανές τιμές του φύλου είναι οι εξής δύο, «Αρσενικό» ή «Θηλυκό», τότε απουσία των αντίστοιχων γεγονότων δηλώνει είτε ότι δε γνωρίζουμε την τιμή του είτε ότι τη γνωρίζουμε, αλλά η τιμή αυτή δεν καλύπτεται από το Λεξιλόγιο μας (π.χ., «Ερμαφρόδιτος» που δεν υπάρχει στις πιθανές τιμές που μπορούμε να χρησιμοποιήσουμε για το φύλο). Από την άλλη, η παρουσία τριάδων σχετικά με το φύλο του ασθενούς μπορεί να μη μας παρέχει πληροφορία γι' αυτό. Για παράδειγμα (σενάριο 2), εάν στις πιθανές τιμές του φύλου, συμπεριλάβουμε τις ακόλουθες δύο τιμές, «Άλλο Φύλο» και

«Άγνωστη Τιμή», τότε η παρουσία των αντίστοιχων τριάδων δε μας εγγυάται ότι γνωρίζουμε το φύλο του ασθενούς.

Στην εργασία μας, κατά τον ορισμό του Μοντέλου Αναφοράς και κυρίως των διαθέσιμων Όρων συμπεριλάβαμε όλους τους απαραίτητους όρους έτσι, ώστε η τιμή της παραμέτρου να μπορεί να καταγραφεί με την απαιτούμενη λεπτομέρεια, εάν τη γνωρίζουμε. Για παράδειγμα, η τιμή του φύλου μπορεί να είναι κάποια από τις ακόλουθες τρεις τιμές: «Αρσενικό», «Θηλυκό» και «Άλλο Φύλο». Συνεπώς, απουσία των τριάδων στην περίπτωση του φύλου δείχνει απουσία πληροφορίας και επομένως, κατά την έκφραση των αντίστοιχων SPARQL ερωτημάτων, χρησιμοποιήσαμε OPTIONAL τμήματα έτσι, ώστε τα κριτήρια να ελέγχονται μόνο, όταν υπάρχουν τα αντίστοιχα δεδομένα.

Στην περίπτωση των παραμέτρων, όπως οι διαγνώσεις που σχετίζονται με έναν ασθενή, για τις οποίες μπορεί να υπάρχει παρουσία της ίδιας παραμέτρου παραπάνω από μία φορές κατά την καταγραφή των δεδομένων ενός ασθενούς, απουσία τριάδων δε σημαίνει απαραίτητα ότι δε γνωρίζουμε τα αντίστοιχα δεδομένα, καθώς ο ασθενής θα μπορούσε π.χ. να είναι υγιής. Από την άλλη, παρουσία μιας ή περισσοτέρων τριάδων δε μας εγγυάται ότι όλη η διαθέσιμη πληροφορία υπάρχει στη βάση μας, καθώς ο ασθενής θα μπορούσε π.χ., να έχει διαγνωστεί και με επιπλέον προβλήματα, τα οποία δεν είναι καταγεγραμμένα στη βάση. Στις περιπτώσεις αυτές δε μπορούμε να γνωρίζουμε αν, όντως, υπάρχουν όλα τα δεδομένα ή όχι έτσι, ώστε να είμαστε σίγουροι ότι ο ασθενής ικανοποιεί ή όχι ένα κριτήριο. Στα πλαίσια της εργασίας αυτής υποθέσαμε ότι η βάση μας είναι πλήρης και περιέχει όλη τη διαθέσιμη πληροφορία έτσι, ώστε να μπορούμε να απαντούμε στα ερωτήματα του χρήστη.

Λαμβάνοντας υπόψη την παραπάνω ανάλυση, τα δεδομένα που παρέχονται ως απάντηση της εκτέλεσης του SPARQL ερωτήματος μπορεί να μην ικανοποιούν όλα τα

κριτήρια εξαιτίας της έλλειψης γνώσης. Για το σκοπό αυτό, τα δεδομένα που λαμβάνουμε θα πρέπει να εξετάζονται προσεκτικά, λαμβάνοντας υπόψη το γεγονός ότι η αποτίμηση των ερωτημάτων πραγματοποιείται με βάση τα δεδομένα που έχουν ρητά οριστεί στη RDF βάση. Ιδανικά, τα δεδομένα που επιστρέφονται θα πρέπει να συνοδεύονται από ένα ακόμη έγγραφο με επιπρόσθετη πληροφορία σχετικά με τα κριτήρια που έχουν χρησιμοποιηθεί (π.χ., επιπρόσθετους όρους που συμπεριλάβαμε), αλλά και από τα δεδομένα που χρησιμοποιήθηκαν κατά τον έλεγχό τους.

21.3Χρησιμοποίηση σε Διαφορετικά Πεδία και Διασύνδεση με Βάσεις Δεδομένων

Η εφαρμογή που αναπτύχθηκε βασίζεται στις οντολογίες, αλλά δεν εξαρτάται από το περιεχόμενό τους. Επομένως, η εφαρμογή αυτή θα μπορούσε να χρησιμοποιηθεί άμεσα, σε αρκετά διαφορετικά πεδία γνώσης (π.χ., ηλεκτρονικά μαγαζιά, μηχανές αναζήτησης, κτλ.), δίνοντας τη δυνατότητα στους χρήστες να περιγράψουν με λεπτομέρεια τα δεδομένα που ψάχνουν (π.χ., προϊόντα, έγγραφα, κτλ.) και τα χαρακτηριστικά που αυτά θα πρέπει να πληρούν. Συνεπώς, η εφαρμογή αυτή περιορίζει το χρονικό διάστημα που απαιτείται και το αντίστοιχο κόστος για την κατασκευή της διεπαφής για την έκφραση ερωτημάτων, καθώς ο ρόλος του ειδικού σε θέματα ανάπτυξης λογισμικού περιορίζεται στην οντολογική περιγραφή του αντίστοιχου πεδίου γνώσης και την παραμετροποίηση του συστήματος. Η έκφραση των ερωτημάτων των χρηστών μπορεί να είναι ιδιαίτερα περίπλοκη, ειδικά στις περιπτώσεις εκείνες όπου τα δεδομένα θα πρέπει να ικανοποιούν αρκετές συνθήκες, γεγονός που πιθανόν να αποθαρρύνει τους χρήστες του διαδικτύου στη χρησιμοποίηση της αντίστοιχης εφαρμογής. Ακολουθώντας, όμως, την ίδια κάθε φορά διαδικασία / προσέγγιση για την έκφραση των ερωτημάτων μέσω της παραμετροποίησης του υπάρχοντος συστήματος, η εφαρμογή που αναπτύχθηκε διασφαλίζει εκ των προτέρων

τη δυνατότητα του χρήστη να «επικοινωνήσει» με τα αντίστοιχα συστήματα και να εκμεταλλευτεί στο έπακρο τις δυνατότητες που αυτά του προσφέρουν.

Η διαδικτυακή εφαρμογή που αναπτύχθηκε και κυρίως η προσέγγιση που ακολουθήθηκε θα μπορούσε, επίσης, να χρησιμοποιηθεί στην ανάπτυξη πιο εξειδικευμένων εμπορικών προϊόντων. Για παράδειγμα, η προσέγγιση αυτή θα μπορούσε να χρησιμοποιηθεί για την τυπική αναπαράσταση των κριτηρίων καταλληλότητας, ένα βήμα απαραίτητο για τη βασισμένη σε υπολογιστή αυτόματη επιλογή των ασθενών που είναι κατάλληλοι να συμμετέχουν σε μία κλινική δοκιμή. Η προσέγγιση που ακολουθήθηκε επιτρέπει στους χρήστες να εκφράσουν ιδιαίτερα περίπλοκα κριτήρια, τα οποία κατόπιν μπορούν να εκφραστούν σε ερωτήματα προς τη βάση των ασθενών [70]. Επιπρόσθετα, το γεγονός ότι λαμβάνουμε υπόψη τη σημασιολογία των όρων και τους περιορισμούς που έχουν οριστεί επιτρέπει την εύρεση όλων των ασθενών που πιθανόν να είναι κατάλληλοι να συμμετέχουν σε μία κλινική δοκιμή, παρά το γεγονός ότι, σε ορισμένες περιπτώσεις, κάποια δεδομένα δεν υπάρχουν στη βάση. Αυτό είναι ένα πολύ χρήσιμο χαρακτηριστικό, καθώς πολλές κλινικές μελέτες αντιμετωπίζουν προβλήματα στην εύρεση του αναγκαίου αριθμού των ασθενών, ενώ πολλές από αυτές τερματίζονται πρόωρα για το λόγο αυτό [3].

Η προσέγγιση που ακολουθήθηκε θα μπορούσε, επίσης, να συνδυαστεί με εργαλεία και λογισμικά συστήματα, που έχουμε ήδη αναπτύξει, για τη διευκόλυνση της πρόσβασης στα δεδομένα μιας RDF ή σχεσιακής βάσης, ειδικότερα όταν τα Μοντέλα Αναφοράς και οι Ονοματολογίες που χρησιμοποιούνται εκατέρωθεν είναι διαφορετικές. Στην περίπτωση αυτή, το παραγόμενο SPARQL ερώτημα θα πρέπει να επεξεργαστεί περαιτέρω έτσι, ώστε να είναι εκφρασμένο χρησιμοποιώντας τους όρους και τις ονοματολογίες που υποστηρίζονται από τη βάση δεδομένων. Για το σκοπό αυτό, μπορούμε να χρησιμοποιήσουμε το Ontology Alignment Tool [105], που έχουμε ήδη

αναπτύξει, για τον καθορισμό της συσχέτισης μεταξύ των μοντέλων και όρων που χρησιμοποιούνται εκατέρωθεν. Ακολούθως, μπορούμε να χρησιμοποιήσουμε το λογισμικό σύστημα, που επίσης αναπτύξαμε, το οποίο παρέχει μια υλοποίηση ενός καινοτόμου αλγορίθμου, που και αυτόν παρουσιάσαμε [73], για την μετάφραση του αρχικού ερωτήματος σε ένα σημασιολογικά ισοδύναμο ερώτημα, λαμβάνοντας υπόψη τις παραπάνω διαφορές. Τα δεδομένα που επιστρέφονται από την εκτέλεση του ερωτήματος, σε ορισμένες περιπτώσεις, χρειάζονται περαιτέρω επεξεργασία, προκειμένου να είναι εκφρασμένα με τους όρους του χρήστη, εάν αυτό είναι απαραίτητο. Επίσης, θα πρέπει να συνοδεύονται και με ένα επιπρόσθετο έγγραφο, το οποίο να παρέχει αναλυτική πληροφορία σχετικά με αλλαγές που πραγματοποιήθηκαν τόσο στο ερώτημα όσο και στα δεδομένα, ειδικότερα όταν αυτές είχαν κάποια επίπτωση (έστω μικρή αλλά απαραίτητη) στην αλλαγή της σημασίας του αρχικού ερωτήματος και των δεδομένων της βάσης που επιστρέφονται.

22 *Συμπεράσματα*

Τα επιτεύγματα στον τομέα του Σημασιολογικού Ιστού έχουν αυξήσει το μέγεθος των δεδομένων που έχουν δημοσιευτεί στον Ιστό με τη μορφή των εικονικών ή πραγματικών RDF βάσεων. Ωστόσο, οι χρήστες του διαδικτύου οι οποίοι δεν έχουν επαρκή γνώση των τεχνολογιών του Σημασιολογικού Ιστού, όπως OWL και SPARQL, δεν μπορούν να επωφεληθούν από τον πλούτο των δεδομένων που υπάρχουν στις RDF βάσεις. Σε αυτή την ενότητα παρουσιάσαμε μία διαδικτυακή εφαρμογή που επιτρέπει στους χρήστες να εκφράσουν γραφικά τα επιθυμητά ερωτήματα, βασισμένοι στην οντολογική αναπαράσταση του πεδίου της γνώσης που καλύπτει μία βάση δεδομένων, μέσω ενός ιδιαίτερα αλληλεπιδραστικού, φιλικού προς το χρήστη, περιβάλλοντος που αναπτύχθηκε για το σκοπό αυτό.

Η εφαρμογή που αναπτύχθηκε μπορεί να χρησιμοποιηθεί από όλους τους χρήστες του διαδικτύου, ανεξάρτητα από το πεδίο απ' όπου προέρχονται, ενώ μπορεί εύκολα να επεκταθεί και να χρησιμοποιηθεί σε διαφορετικά πεδία γνώσης, όπως για παράδειγμα για την έκφραση των ΚΚ μιας κλινικής δοκιμής. Επίσης, μπορεί να συνδυαστεί με εργαλεία και μηχανισμούς που αναπτύχθηκαν, τα οποία επιτρέπουν την εκτέλεση των ερωτημάτων του χρήστη (π.χ., εύρεση των ασθενών που πληρούν τα ΚΚ) σε βάσεις δεδομένων που υποστηρίζουν διαφορετικά μοντέλα και κωδικοποιήσεις.

Η σελίδα αυτή είναι σκόπιμα λευκή

Δ. ΣΥΝΤΟΜΕΥΣΕΙΣ ΚΛΙΝΙΚΩΝ ΜΕΛΕΤΩΝ

Στην ενότητα αυτή παρουσιάζουμε το σύστημα που αναπτύχθηκε και τα αποτελέσματα που εξάγαμε από την ανάλυση των συντομεύσεων (abbreviations aka acronym, shorthand, short form) που χρησιμοποιούνται στις κλινικές μελέτες.

Στα πλαίσια της εργασίας αυτής, αρχικά, μελετήσαμε τους υπάρχοντες αλγορίθμους και τεχνικές για τον εντοπισμό της πλήρους μορφής (expansion aka long form) των συντομεύσεων. Ακολούθως, αναπτύξαμε ένα σύστημα για τον εντοπισμό της σημασίας (sense aka meaning) των συντομεύσεων που χρησιμοποιούνται στις κλινικές μελέτες, εφαρμόζοντας καινοτόμους αλγορίθμους και τεχνικές που λαμβάνουν υπόψη τη διακριτική δυνατότητα των λέξεων της πλήρους μορφής τους, προκειμένου να εντοπίσουν σωστά τη σημασία τους. Η αποτίμηση του συστήματος, με βάση τις συντομεύσεις που εμείς ορίσαμε σε ένα σύνολο από κλινικές μελέτες που τυχαία επιλέξαμε, έδειξε ότι το σύστημά μας μπορεί να εντοπίσει με μεγάλη ακρίβεια τη σημασία τους.

Η επεξεργασία των αποτελεσμάτων που προέκυψαν έδειξε ότι οι συντομεύσεις των κλινικών μελετών είναι πολύ λιγότερο ασαφείς σε σύγκριση με τις συντομεύσεις που αναφέρονται στα βιοϊατρικά άρθρα. Ωστόσο, ορισμένες συντομεύσεις εξακολουθούν να έχουν αρκετές διαφορετικές σημασίες, δυσχεραίνοντας το έργο εντοπισμού της σημασίας τους, ιδιαίτερα, όταν η πλήρης μορφή τους δεν παρέχεται μέσα στην περιγραφή της κλινικής μελέτης.

Η σελίδα αυτή είναι σκόπιμα λευκή

23 *Εισαγωγή*

Οι συντομεύσεις (γνωστές επίσης ως ακρώνυμα και αρχικά) αποτελούν ένα σημαντικό τμήμα μιας κλινικής μελέτης. Στόχος τους είναι να παρέχουν μία μικρότερη σε μέγεθος μορφή μιας συχνά μεγάλης λέξης, φράσης ή έκφρασης (γνωστής ως πλήρης μορφή) έτσι, ώστε οι ερευνητές να μπορούν να τη χρησιμοποιήσουν αποτελεσματικά μέσα στο κείμενο. Απ' την άλλη μεριά, ο εντοπισμός της πλήρους μορφής μιας συντόμευσης είναι απαραίτητος για την κατανόηση του κειμένου.

Εξαιτίας του μεγάλου όγκου των διαθέσιμων κλινικών μελετών που υπερβαίνουν τις 200 χιλιάδες, με το μέγεθός τους να αυξάνεται συνεχώς (περίπου 20 χιλιάδες νέες κλινικές μελέτες καταγράφονται κάθε χρόνο), η ανάλυση των δεδομένων τους (π.χ., τα κριτήρια καταλληλότητας μιας μελέτης) συχνά προϋποθέτει την εφαρμογή καινοτόμων μηχανισμών εξόρυξης γνώσης, οι οποίοι μεταξύ των άλλων «προβλημάτων» καλούνται να διαχειριστούν τις συντομεύσεις που ορίζονται μέσα στο κείμενο. Προς αυτή την κατεύθυνση, αρχικά, εξετάσαμε τις συντομεύσεις που ορίζονται σε ένα περιορισμένο σύνολο από κλινικές μελέτες, που τυχαία επιλέξαμε και κατόπιν χρησιμοποιήσαμε τη γνώση που αποκτήσαμε για την ανάπτυξη ενός συστήματος εντοπισμού της σημασίας των συντομεύσεων που χρησιμοποιούνται στις κλινικές μελέτες. Για το σκοπό αυτό υλοποιήθηκαν καινοτόμοι αλγόριθμοι και τεχνικές, οι οποίες λαμβάνουν υπόψη διάφορες παραμέτρους, μεταξύ των οποίων και τη διακριτική δυνατότητα των λέξεων που

συμμετέχουν στην πλήρη έκφραση, για τον εντοπισμό της σημασίας τους. Η εξέταση της ορθότητας του συστήματος και των αλγορίθμων που υλοποιήθηκαν έγινε με βάση τις συντομεύσεις που ρητά ορίσαμε σε 141 κλινικές μελέτες μέσω μιας ημιαυτόματης διαδικασίας, με τη βοήθεια ενός, φιλικού προς το χρήστη, γραφικού περιβάλλοντος που αναπτύχθηκε για το σκοπό αυτό. Τέλος, αναλύσαμε τα δεδομένα που συλλέχθηκαν από όλο το σύνολο των διαθέσιμων κλινικών μελετών, τα οποία παρουσιάζονται στα επόμενα κεφάλαια της ενότητας αυτής.

Η ενότητα αυτή είναι οργανωμένη ως εξής. Αρχικά, στο κεφάλαιο 24 παρουσιάζουμε υπάρχοντες αλγορίθμους και τεχνικές για τον εντοπισμό της σημασίας των συντομεύσεων. Ακολούθως, στο κεφάλαιο 25 περιγράφουμε τη διαδικασία που ακολουθήθηκε για τον ορισμό των συντομεύσεων που χρησιμοποιούνται σε ένα σύνολο από μελέτες που επιλέχθηκαν, καθώς επίσης και τη γνώση που αποκομίσαμε από τη μελέτη αυτή. Στο κεφάλαιο 26 περιγράφουμε αναλυτικά το σύστημα που αναπτύχθηκε και τους αλγορίθμους που υλοποιήθηκαν για τον αυτόματο εντοπισμό της σημασίας των συντομεύσεων σε όλες τις διαθέσιμες κλινικές μελέτες. Στο κεφάλαιο 27 εξετάζουμε την ορθότητα των αλγορίθμων και τεχνικών που χρησιμοποιήθηκαν. Τα αποτελέσματα της μελέτης των συντομεύσεων των κλινικών μελετών παρουσιάζονται στο κεφάλαιο 28, καθώς επίσης και ορισμένα θέματα προς συζήτηση. Στο τελευταίο κεφάλαιο της ενότητας αυτής (κεφάλαιο 29) συνοψίσαμε τα αποτελέσματα της μελέτης μας, καθώς επίσης και κάποια θέματα που χρήζουν περαιτέρω έρευνας.

24 *Αλγόριθμοι και Τεχνικές*

Για την αναγνώριση της σημασίας των συντομεύσεων, υπάρχουν αρκετοί αλγόριθμοι και τεχνικές που μπορούμε να τους κατατάξουμε στις ακόλουθες τέσσερις κατηγορίες [150]: αλγορίθμους ευθυγράμμισης (alignment-based techniques), αλγορίθμους βασισμένους σε κανόνες ή μοτίβα (rule/based / pattern-based techniques), αλγορίθμους βασισμένους σε μηχανική μάθηση (machine learning-based techniques) και αλγορίθμους βασισμένους στη συνύπαρξη εκφράσεων (collocation-based techniques).

Για τον εντοπισμό της πλήρους μορφής των συντομεύσεων προτάθηκε ένας απλός αλγόριθμος από τους Schwartz και Hearst [151], ο οποίος βασίζεται αποκλειστικά και μόνο στους χαρακτήρες που χρησιμοποιούνται. Ο αλγόριθμος αρχικά εντοπίζει τα σημεία εκείνα της πρότασης, όπου μία συντόμευση πιθανώς ακολουθεί την πλήρη μορφή της μέσα σε παρενθέσεις ή αντίστροφα, και ακολούθως ψάχνει για τη μικρότερη φράση (υποψήφια πλήρης έκφραση) που ταιριάζει με τη συντόμευση. Η φράση θα πρέπει να περιέχει όλους τους χαρακτήρες της συντόμευσης με την ίδια σειρά, με τον επιπρόσθετο περιορισμό ότι ο πρώτος χαρακτήρας της συντόμευσης και της πλήρους έκφρασης θα πρέπει να ταιριάζουν. Το μήκος της υποψήφιας πλήρους έκφρασης (δηλαδή ο αριθμός των λέξεων) θα πρέπει να είναι μικρότερος από την τιμή της έκφρασης min(|A| + 5, |A| * 2), όπως και στους Park και Bryd [152], όπου |A| είναι ο αριθμός των χαρακτήρων της συντόμευσης.

Για την αναγνώριση της σημασίας των συντομεύσεων οι Park και Bryd πρότειναν έναν αλγόριθμο που βασίζεται σε μοτίβα. Ο αλγόριθμος χρησιμοποιεί σημεία στίξης (δηλαδή, παρενθέσεις και άγκιστρα) καθώς επίσης και λέξεις ή φράσεις που δείχνουν ότι ακολουθεί ή προηγείται η πλήρης έκφραση για τον εντοπισμό των συντομεύσεων που πιθανώς ορίζονται μέσα στο κείμενο. Ακολούθως, ο αλγόριθμος εξετάζει το κείμενο πριν ή μετά από κάθε συντόμευση, για να βρει τη μικρότερη σε μήκος έκφραση που ταιριάζει με τη συντόμευση. Για το σκοπό αυτό, ο αλγόριθμος αρχικά παράγει ένα βασισμένο στους χαρακτήρες μοτίβο για τη συντόμευση καθώς και την πιθανή πλήρη έκφρασή της και έπειτα ψάχνει σε μία τοπική βάση για την ύπαρξη εγγεγραμμένων κανόνων γι' αυτό το ζευγάρι μοτίβων. Εάν υπάρχουν ένας ή περισσότεροι κανόνες, τότε αυτοί εφαρμόζονται στην υποψήφια έκφραση και το αποτέλεσμά τους συγκρίνεται με τη συντόμευση που υπάρχει. Εάν ταιριάζουν, τότε η έκφραση αποτελεί την πλήρη μορφή της συντόμευσης.

Ο Pustejovsky κ.ά. [153] ανέπτυξαν ένα σύστημα για τον εντοπισμό της πλήρους μορφής των συντομεύσεων που χρησιμοποιούνται στις περιγραφές των Medline άρθρων. Στην εργασία τους χρησιμοποιήθηκαν δύο διαφορετικές προσεγγίσεις, οι οποίες βασίζονταν στον ίδιο αλγόριθμο. Πιο συγκεκριμένα, το σύστημα παρήγαγε μία κανονική έκφραση (regular expression) για κάθε συντόμευση την οποία ακολούθως χρησιμοποιούσε, για να εντοπίσει την πλήρη μορφή της. Στη δεύτερη περίπτωση χρησιμοποιήθηκε μια βελτιστοποιημένη έκδοση του αλγορίθμου αυτού η οποία λάμβανε υπόψη τη σύνταξη των λέξεων. Όπως έδειξαν, η ακρίβεια του αλγορίθμου βελτιωνόταν σημαντικά, όταν λαμβάνανε υπόψη και τη σύνταξη για τον περιορισμό του κειμένου, όπου θα πρέπει να αναζητήσουμε την πλήρη μορφή των συντομεύσεων.

Ο Yu κ.ά. [154] ανέπτυξαν ένα σύστημα που βασίζεται σε ένα σύνολο από κανόνες για τον εντοπισμό της πλήρους μορφής των συντομεύσεων. Το σύστημα, αρχικά, ψάχνει για λέξεις ή φράσεις που εσωκλείονται μέσα σε παρενθέσεις για τον εντοπισμό

υποψήφιων ζευγαριών, ενώ λαμβάνει επίσης υπόψη συγκεκριμένα σημεία στίξης (ερωτηματικό και κόμμα) που ορισμένες φορές ακολουθούν μια συντόμευση. Έπειτα, χρησιμοποιεί τους κανόνες για τον εντοπισμό της μικρότερης έκφρασης που ταιριάζει με τη συντόμευση. Όλοι οι κανόνες που υλοποιήθηκαν βασίζονται στο γεγονός ότι ο πρώτος χαρακτήρας της συντόμευσης και της υπό επεξεργασία έκφρασης ταιριάζουν, ενώ διαφέρουν στον αλγόριθμο που εσωτερικά χρησιμοποιήθηκε για την ευθυγράμμιση των εναπομεινάντων χαρακτήρων της συντόμευσης και της υποψήφιας πλήρους έκφρασης.

Ο Sohn κ.ά. [155] χρησιμοποίησαν διάφορες στρατηγικές για την εύρεση της πιθανής πλήρους μορφής των συντομεύσεων οι οποίες αποτελούνται, το πολύ, από 10 συνεχόμενες λέξεις. Όλες οι στρατηγικές που υλοποιήθηκαν ακολουθούν την ίδια λογική. Πιο συγκεκριμένα, όλες οι στρατηγικές ψάχνουν για τη μικρότερη φράση που περιέχει τους χαρακτήρες της συντόμευσης με την ίδια σειρά. Όμως, οι τεχνικές (κανόνες) που χρησιμοποιήθηκαν εσωτερικά, για να ταιριάξουν τη συντόμευση με την πλήρη μορφή της, είναι διαφορετικές. Για παράδειγμα, στην πιο αξιόπιστη στρατηγική ο αλγόριθμος χρησιμοποιεί μόνο τον αρχικό χαρακτήρα των λέξεων της φράσης. Άλλες, λιγότερο αξιόπιστες στρατηγικές, λαμβάνουν υπόψη τους υπόλοιπους χαρακτήρες της κάθε λέξης, καθώς επίσης και αν αυτές είναι τερματικές λέξεις (stop words) ή όχι. Οι χαρακτήρες στίξης (εάν υπάρχουν) αγνοούνται. Θα πρέπει να τονίσουμε ότι οι στρατηγικές που υλοποιήθηκαν είναι ταξινομημένες με βάση την αξιοπιστία τους και ακολούθως εφαρμόζονται η μία μετά την άλλη για την εύρεση της πλήρους έκφρασης, με την προϋπόθεση ότι οι προηγούμενες έχουν αποτύχει.

Ο Chang κ.ά. [156] χρησιμοποίησαν μία τεχνική επιβλεπόμενης μάθησης για την αναγνώριση των συντομεύσεων. Πιο συγκεκριμένα, εκπαίδευσαν έναν λογιστικής παλινδρόμησης ταξινομητή (logistic regression classifier) [157], χρησιμοποιώντας ένα σύνολο δεδομένων που βασίστηκε στην επεξεργασία ενός περιορισμένου αριθμού

συντομεύσεων από την περιγραφή των Medline άρθρων που εξέτασαν χειροκίνητα. Τα δεδομένα αυτά αποτελούνταν από περίπου 1000 υποψήφια ζευγάρια (παρήχθησαν αυτόματα με βάση τις συντομεύσεις που επιλέχθηκαν) και περιλάμβαναν και σωστές αλλά και λάθος συσχετίσεις. Ο ταξινομητής βασίστηκε σε 9 διαφορετικά χαρακτηριστικά (π.χ., ποσοστό των χαρακτήρων της συντόμευσης που ταιριάζουν) που παρέχουν μία ποσοτική μέτρηση της αντιστοίχισης μεταξύ των χαρακτήρων της συντόμευσης και της υποψήφιας πλήρους μορφής τους για τον υπολογισμό μιας τιμής. Με βάση την τιμή αυτή ο αλγόριθμος αποφάσιζε εάν η αντίστοιχη έκφραση ήταν η πλήρης μορφή της συντόμευσης ή όχι.

Ο Kuo κ.ά. [158] εξέτασαν τέσσερις διαφορετικούς αλγορίθμους μηχανικής μάθησης (Λογιστική παλινδρόμηση, Monte-Carlo Δειγματοληψία Μέγιστης Εντροπίας, Μηχανή Διανυσμάτων Υποστήριξης και naïve Bayes κατηγοριοποίηση) για τον εντοπισμό ζευγαριών συντόμευσης – πλήρους έκφρασης. Στην εργασία τους, χρησιμοποιήθηκαν αρκετά χαρακτηριστικά (features), συμπεριλαμβανομένων μορφολογικών χαρακτηριστικών, καθώς επίσης και χαρακτηριστικά που έχουν να κάνουν με τη σύνθεση της πλήρους έκφρασης. Επιπρόσθετα, χρησιμοποιήθηκαν δύο διαφορετικά σύνολα δεδομένων για λόγους αποτίμησης και επαλήθευσης. Και στα δύο αυτά σύνολα τα «θετικά» ζευγάρια αποτελούνταν από τα ζευγάρια συντόμευσης με την αντίστοιχη πλήρη έκφραση που είχαν οριστεί, ενώ τα «αρνητικά» ζευγάρια παρήχθησαν αυτόματα (δηλαδή συντόμευση – πιθανή πλήρης έκφραση) με βάση το κείμενο που υπήρχε πριν από κάθε συντόμευση. Η εφαρμογή των 4 αλγορίθμων μηχανικής μάθησης έδειξε ότι ή Λογιστική Παλινδρόμηση (Logistic regression) [159] και η Μηχανή Διανυσμάτων Υποστήριξης (Support Vector Machine) [160] είχαν καλύτερη απόδοση. Επίσης, το σύστημα παρείχε καλύτερα αποτελέσματα, όταν χρησιμοποιήθηκαν όλα τα παραπάνω χαρακτηριστικά.

Οι παραπάνω αλγόριθμοι και τεχνικές παρέχουν άμεσα την πλήρη μορφή μιας συντόμευσης, με την προϋπόθεση ότι υπάρχει μια εμφανής συσχέτιση των χαρακτήρων της συντόμευσης με τους χαρακτήρες της πλήρους έκφρασης (ακρωνύμου-τύπου συντομεύσεις). Επίσης, παρά το γεγονός ότι υπάρχουν αρκετές διαφορετικές προσεγγίσεις για τον εντοπισμό της πλήρους μορφής, όλοι οι αλγόριθμοι και οι τεχνικές παρέχουν ικανοποιητικά αποτελέσματα (F-score κοντά στο 90%) [161]. Όμως, όταν η συντόμευση έχει μερική ή καθόλου συσχέτιση με τους χαρακτήρες της πλήρους έκφρασης, οι παραπάνω αλγόριθμοι και τεχνικές δεν μπορούν να εντοπίσουν σωστά την πλήρη μορφή των συντομεύσεων.

Ο Zhou κ.ά. [162] εφάρμοσαν στατιστικές μεθόδους για τον εντοπισμό της πλήρους μορφής των συντομεύσεων. Πιο συγκεκριμένα, το κείμενο που βρισκόταν κοντά στις συντομεύσεις συλλέχθηκε και επεξεργάστηκε για τον εντοπισμό των φράσεων εκείνων που χρησιμοποιούνται αρκετά συχνά. Οι Okazaki και Ananiadou [163] εστίασαν στις περιπτώσεις εκείνες, όπου μία συντόμευση ακολουθεί την πλήρη μορφή της μέσα σε παρενθέσεις. Ακολούθως, χρησιμοποίησαν μία λίγο διαφορετική «έκδοση» της μεθόδου "C-value" [164], η οποία συνδυάζει γλωσσολογικά στοιχεία και στατιστική πληροφορία, για τον υπολογισμό μιας τιμής που εκφράζει την πιθανότητα ένας όρος (αποτελούμενος από μία ή περισσότερες λέξεις) να αποτελεί την πλήρη μορφή της συντόμευσης. Στις περιπτώσεις όπου η τιμή των πιθανών εκφράσεων μιας συντόμευσης ξεπερνούσε ένα κατώτερο όριο, η έκφραση αυτή θεωρούνταν έγκυρη.

Οι τεχνικές που βασίζονται σε στατιστικές μεθόδους χρειάζονται ένα μεγάλο αριθμό από έγγραφα με πολλαπλές εμφανίσεις των ζευγαριών συντόμευσης με πλήρη έκφραση, για να παράγουν αξιόλογα αποτελέσματα, ενώ δεν μπορούν να χειριστούν αποτελεσματικά τις περιπτώσεις εκείνες στις οποίες η συντόμευση ή η σημασία της χρησιμοποιούνται σπάνια. Για το σκοπό αυτό, οι παραπάνω τεχνικές, συχνά,

χρησιμοποιούνται σε συνδυασμό με τις υπάρχουσες τεχνικές αναγνώρισης ακρωνύμου-τύπου συντομεύσεων, συγκεντρώνοντας και εν συνεχεία επεξεργαζόμενες τα δεδομένα εκείνα όπου οι βασισμένοι σε χαρακτήρες αλγόριθμοι αποτυγχάνουν.

Για τον εντοπισμό της σημασίας των συντομεύσεων, όταν η πλήρης έκφρασή τους δεν παρέχεται μέσα στο κείμενο, η απλούστερη προσέγγιση είναι να υποθέσουμε ότι η σημασία τους είναι αυτή με την οποία η συντόμευση χρησιμοποιείται αρκετά συχνά [165]. Όμως, σε αρκετές περιπτώσεις υπάρχει παραπάνω από μία «κυρίαρχη» σημασία και, κατά συνέπεια, είναι απαραίτητο να λάβουμε υπόψη το ευρύτερο πλαίσιο μέσα στο οποίο χρησιμοποιείται η εκάστοτε συντόμευση. Για το σκοπό αυτό, υπάρχουν αρκετές προσεγγίσεις, όπως αυτή που περιγράφεται από τον Xu κ.ά. [166] καθώς επίσης και αυτή του Stevenson κ.ά. [167], οι οποίες, γενικότερα μιλώντας, προσπαθούν να αντιμετωπίσουν το πρόβλημα αυτό ως ένα πρόβλημα αποσαφήνισης της σημασίας μιας λέξης. Οι τεχνικές αυτές χρησιμοποιούν αρκετές παραμέτρους, συμπεριλαμβανομένων των όρων από την MeSH οντολογία, που υπάρχουν στα άρθρα που είναι διαθέσιμα απ' το Medline καθώς επίσης και έννοιες που εντοπίζονται αυτόματα.

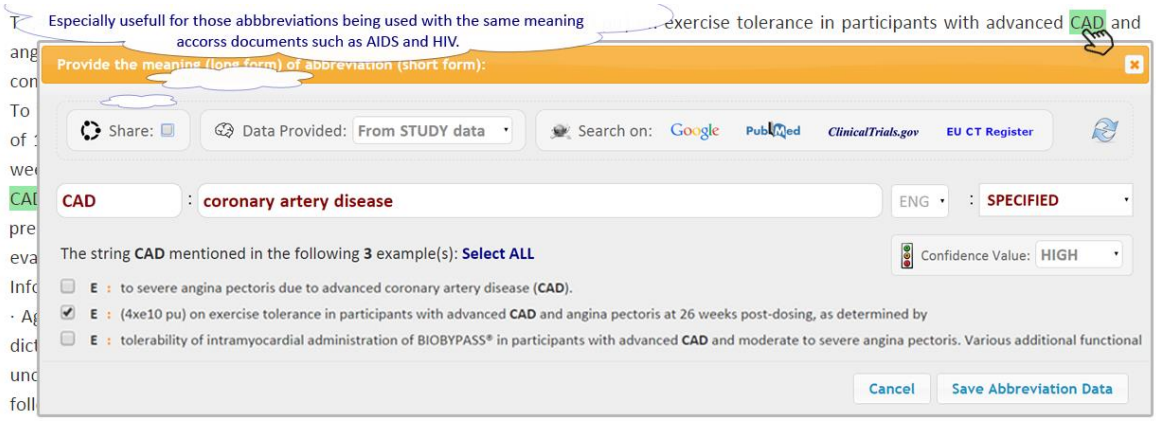
25 *Μελέτη Συντομεύσεων Κλινικών Μελετών*

25.1 Καθορισμός της Σημασίας των Συντομεύσεων

Για τη μελέτη των συντομεύσεων επιλέξαμε 141 κλινικές μελέτες από την τοποθεσία www.clinicaltrialsregister.eu [168] και κατόπιν ορίσαμε τη σημασία των συντομεύσεων που χρησιμοποιούνται σε κάθε μία από αυτές. Για την επιλογή των κλινικών μελετών, αρχικά κατεβάσαμε όλες τις διαθέσιμες κλινικές μελέτες (περίπου 30 χιλιάδες), τις οργανώσαμε σε ευρύτερες κατηγορίες με βάση τον αριθμό των συντομεύσεων που χρησιμοποιήθηκαν σε κάθε μία από αυτές και έπειτα διαλέξαμε τυχαία έναν περιορισμένο αριθμό από κάθε κατηγορία, λαμβάνοντας υπόψη το συνολικό ποσοστό των μελετών που ανήκαν σε κάθε μία από αυτές.

Για τον καθορισμό της σημασίας των συντομεύσεων με τη μικρότερη δυνατή προσπάθεια, αναπτύξαμε μία διαδικτυακή εφαρμογή (Σχήμα 27) η οποία επιτρέπει στους χρήστες να αλληλεπιδρούν με τις συντομεύσεις που χρησιμοποιούνται σε κάθε κλινική μελέτη (επισημαίνονται με πράσινο χρώμα) και ακολούθως να ορίσουν τη σημασία τους καθώς επίσης και τα συγκεκριμένα τμήματα/προτάσεις του κειμένου που χρησιμοποιούνται με τη σημασία αυτή. Γενικά, μία συντόμευση χρησιμοποιείται με την ίδια σημασία σε ένα έγγραφο [169]. Όμως, σε ορισμένες περιπτώσεις, η σημασία της μπορεί να είναι διαφορετική. Ένα χαρακτηριστικό παράδειγμα αποτελεί ο αριθμός «IV» ο

οποίος σε κάποια άλλα σημεία του κειμένου μπορεί να χρησιμοποιείται αντί της λέξης «Intravenous». Επιπρόσθετα, μία συντόμευση μπορεί να χρησιμοποιείται μέσα σε μία άλλη λέξη ή φράση, όπως «HIV-positive» και, συνεπώς, θα θέλαμε να ορίσουμε ρητά ότι η παραπάνω συντόμευση διατηρεί τη σημασία της.



Σχήμα 27: Στιγμιότυπο από τη διαδικτυακή εφαρμογή που αναπτύχθηκε για τον καθορισμό της σημασίας των συντομεύσεων μιας κλινικής δοκιμής.

Το Σχήμα 27 απεικονίζει ένα στιγμιότυπο από τη διαδικτυακή εφαρμογή που αναπτύχθηκε. Η εφαρμογή αυτή παρέχει άμεσα την πλήρη μορφή (Long Form aka Expansion) της συντόμευσης, όταν ορίζεται μέσα στο κείμενο χρησιμοποιώντας τον αλγόριθμο των Schwartz και Hearst [151], διαφορετικά, βοηθάει τους χρήστες να εντοπίσουν τη σημασία της χρησιμοποιώντας πηγές δεδομένων που είναι διαθέσιμες στο διαδίκτυο. Όταν ο χρήστης έχει ορίσει τη σημασία όλων των συντομεύσεων που χρησιμοποιούνται στην κλινική μελέτη, το σύστημα παράγει αυτόματα ένα XML έγγραφο με το αναγνωριστικό (ID) της μελέτης καθώς και τις συντομεύσεις που έχουν οριστεί.

Αξίζει να σημειωθεί ότι, κατά την παραπάνω διαδικασία καθορισμού της σημασίας των συντομεύσεων, το σύστημα επιτρέπει στο χρήστη να παρέχει επιπρόσθετα δεδομένα για κάθε συντόμευση, όπως, εάν αυτή είχε οριστεί ρητά μέσα στο κείμενο ή όχι καθώς επίσης και πόσο ασφαλείς (confident) είναι οι χρήστες για τα δεδομένα (δηλαδή την πλήρη μορφή της) που παρέχουν. Στις περιπτώσεις εκείνες για τις οποίες δεν ήμασταν

σίγουροι για την σημασία μιας ή περισσοτέρων συντομεύσεων επανεξετάσαμε τη σημασία τους σε συνεννόηση με ειδικούς στον τομέα της υγείας και περίθαλψης. Μέσω της παραπάνω διαδικασίας καταφέραμε να ορίσουμε χωρίς αμφιβολία τη σημασία των συντομεύσεων στις κλινικές μελέτες που επιλέχθηκαν [170].

25.2Συντομεύσεις και Εκφράσεις

Η ανάλυση των δεδομένων που συλλέχτηκαν έδειξε ότι πάνω από 10 συντομεύσεις χρησιμοποιούνται κατά μέσο όρο στις κλινικές μελέτες που εξετάσαμε, από τις οποίες το 70% χρησιμοποιούνται χωρίς να έχουμε ορίσει την πλήρη μορφή τους μέσα στο κείμενο. Από την ανάλυσή μας εξαιρέσαμε σκοπίμως τις Λατινικές Συντομογραφίες (π.χ., e.g., i.e., etc.) καθώς επίσης και τις συντομογραφίες των μονάδων μέτρησης (π.χ., mg, kgr, etc.), οι οποίες χρησιμοποιούνται συνήθως σε συντετμημένη μορφή. Επίσης, εξαιρέσαμε και ορισμένες συντομεύσεις οι οποίες εμφανίζονται στον τίτλο των ενοτήτων και, κατά συνέπεια, είναι ίδιες σε όλες τις κλινικές μελέτες.

Στο 94% των περιπτώσεων των συντομεύσεων που ορίζονταν στο κείμενο, αυτή ακολουθούσε την πλήρη μορφή τους μέσα σε παρενθέσεις ή άγκιστρα (μοτίβο 1), ενώ μόλις στο 4% των περιπτώσεων η συντόμευση προηγούνταν της πλήρους έκφρασης (μοτίβο 2). Αξίζει, πάντως, να σημειωθεί ότι μόνο στο 2% των περιπτώσεων η συντόμευση οριζόταν με κάποιον άλλον τρόπο ή μοτίβο (π.χ., προηγούνταν ή ακολουθούσε την πλήρη έκφραση χωρίς να παρεμβάλλεται κάποιο σημείο στίξης). Μια άλλη σημαντική παρατήρηση είναι το γεγονός ότι στο 98% των συντομεύσεων που βρίσκονταν μέσα σε παρενθέσεις ή άγκιστρα (μοτίβο 1) η πλήρης έκφραση βρισκόταν στο κείμενο / πρόταση που προηγούνταν. Αντιθέτως, όταν η συντόμευση χρησιμοποιούνταν μέσα στο κείμενο χωρίς παρενθέσεις, ακολουθούμενη από κάποια έκφραση που βρισκόταν μέσα σε παρενθέσεις ή άγκιστρα (μοτίβο 2), μόνο στο 55% των περιπτώσεων υπήρχε η πλήρης μορφή τους.

Η ανάλυση των κλινικών μελετών έδειξε επίσης ότι οι συντομεύσεις μπορεί να χρησιμοποιούνται είτε σε ενικό είτε σε πληθυντικό αριθμό, ενώ μπορεί να συμμετέχουν σε μία ή περισσότερες εκφράσεις. Πιο συγκεκριμένα, μία συντόμευση μπορεί να χρησιμοποιηθεί μαζί με άλλες λέξεις (π.χ., HIV-positive) ή προθέματα (π.χ., anti-HIV) για το σχηματισμό σύνθετων λέξεων ή φράσεων, η σημασία των οποίων επηρεάζεται άμεσα από τους χαρακτήρες που δεν ανήκουν σε αυτήν. Ο Πίνακας 11 παρουσιάζει τις επικρατέστερες εκφράσεις στις οποίες συμμετέχουν οι συντομεύσεις καθώς και ένα παράδειγμα για κάθε μία από αυτές.

Expression	**Example(s)**	**Expression**	**Example(s)**
ABR-positive	HIV-positive	ABR-induced	NRTI-induced
ABR-negative	CRIM-negative	ABR-specific	YMSM-specific
ABR-related	ATGL-related,	ABR-score	FLIE-score, MELD-score
ABR-associated	CVC-associated	anti-ABR	anti-RET, anti-IFN
ABR-based	MR-based	post-ABR	post-PTA, post-CRT
ABR-containing	MPA-containing	pre-ABR	pre-GCRA, pre-TAVR
ABR-like	LDL-like, BMS-like	non-ABR	non-LDL, non-TCC

Πίνακας 11: Ευρέως χρησιμοποιούμενες εκφράσεις με συντομεύσεις (ABR)

Στην πραγματικότητα, η ανάλυση των δεδομένων που συλλέχθηκαν επεσήμανε το γεγονός ότι ένα σημαντικό ποσοστό των συντομεύσεων συμμετέχουν σε εκφράσεις. Ακολούθως, εξετάσαμε τις συντομεύσεις και ειδικότερα τις εκφράσεις στις οποίες συμμετέχουν οι συντομεύσεις που χρησιμοποιούνται σε όλες τις κλινικές μελέτες που κατεβάσαμε μέσω μιας ημιαυτόματης διαδικασίας. Πιο συγκεκριμένα, αρχικά δημιουργήσαμε ένα μοτίβο για κάθε φράση στην οποία συμμετέχει μία συντόμευση (π.χ., αντικαθιστώντας την συντόμευση με την ίδια πάντα συμβολοακολουθία) και κατόπιν ταξινομήσαμε τα μοτίβα, που εντοπίσαμε με βάση τη συχνότητα χρήσης τους. Έπειτα, εξετάσαμε τα πιο ευρέως χρησιμοποιούμενα μοτίβα καθώς και ορισμένα παραδείγματα που συγκεντρώσαμε για το καθένα από αυτά, προκειμένου να εντοπίσουμε τις επικρατέστερες εκφράσεις στις οποίες συμμετέχουν οι συντομεύσεις.

25.3Συσχέτιση Συντόμευσης με Πλήρη Έκφραση

Η αντιστοίχιση μεταξύ των χαρακτήρων μιας έκφρασης εκφρασμένης σε πλήρη και συντετμημένη μορφή έχει συχνά καθοριστικό ρόλο κατά την αναγνώριση μιας ακρώνυμου-τύπου συντόμευσης. Για το σκοπό αυτό, εξετάσαμε τα δεδομένα που συγκεντρώθηκαν κατά την παραπάνω διαδικασία, με απώτερο στόχο να εντοπίσουμε τον τρόπο σύνδεσης (εάν υπάρχει) μεταξύ των χαρακτήρων της συντόμευσης και των χαρακτήρων των λέξεων της αντίστοιχης έκφρασης. Η ανάλυσή μας έδειξε ότι τα ζευγάρια αυτά μπορούν να ενταχθούν σε τρεις ευρύτερες κατηγορίες, που περιγράφονται στις επόμενες παραγράφους της ενότητας αυτής. Χαρακτηριστικά παραδείγματα για καθεμιά από τις 3 κατηγορίες υπάρχουν στον Πίνακας 12.

Η πρώτη κατηγορία (στενά συνδεδεμένα) περιλαμβάνει τα ζευγάρια εκείνα στα οποία όλες οι λέξεις μιας φράσης συμμετείχαν στη δημιουργία της συντόμευσης. Πιο συγκεκριμένα, ο πρώτος χαρακτήρας της κάθε λέξης υπάρχει στη συντετμημένη μορφή της με την ίδια σειρά, ενώ οι υπόλοιποι χαρακτήρες της συντόμευσης (εάν υπάρχουν) προέρχονται από το εσωτερικό των αντίστοιχων λέξεων (παραδείγματα 1 με 5). Η δεύτερη κατηγορία (χαλαρά συνδεδεμένα) αποτελείται από τα ζευγάρια εκείνα στα οποία τουλάχιστον μία λέξη μιας φράσης δε συμμετείχε στη δημιουργία της συντόμευσης. Οι λέξεις αυτές είναι συνήθως λειτουργικές λέξεις (function words), όπως είναι οι προθέσεις και οι αντωνυμίες, οι οποίες επιτρέπουν στο χρήστη να σχηματίσει γραμματικά σωστές εκφράσεις, χωρίς, όμως, να προσθέτουν επιπλέον πληροφορία (παραδείγματα 7 και 8). Όπως παρατηρήσαμε, όμως, ένας σημαντικός αριθμός από τις λέξεις αυτές δεν ανήκουν στην παραπάνω κατηγορία (παραδείγματα 9 με 12). Τις λέξεις αυτές θα τις εξετάσουμε αναλυτικά στην επόμενη ενότητα.

No	Συντομ.	Πλήρης Έκφραση	Περιγραφή / Σχόλια	
1	CNS	**C**entral **N**ervous **S**ystem	Η συντόμευση αποτελείται από τον αρχικό χαρακτήρα των λέξεων της πλήρους έκφρασης	Στενά Συνδεδεμένα
2	CrCl	**Cr**eatinine **Cl**earance	Η συντόμευση αποτελείται από τους δύο αρχικούς χαρακτήρες των δύο λέξεων της πλήρους έκφρασης	
3	CVA	**C**erebro**v**ascular **A**ccident	Η συντόμευση περιλαμβάνει τον αρχικό χαρακτήρα της κάθε λέξης της πλήρους έκφρασης καθώς επίσης και έναν ακόμη χαρακτήρα απ' την πρώτη λέξη.	
4	TZD	**T**hia**z**oli**d**inedione	Ο πρώτος χαρακτήρας της συντόμευσης και της πλήρους έκφρασης ταιριάζουν, ενώ οι υπόλοιποι χαρακτήρες της συντόμευσης υπάρχουν στο εσωτερικό της πλήρους έκφρασης με την ίδια σειρά.	
5	LTP2	**L**actate **T**urn**p**oint **2**	Οι αρχικοί χαρακτήρες των λέξεων της πλήρους έκφρασης καθώς επίσης και ο αριθμός που αναφέρεται υπάρχουν στη συντόμευση η οποία επιπρόσθετα περιλαμβάνει έναν ακόμη χαρακτήρα απ' τη δεύτερη λέξη.	
6	SD1	**S**tudy **D**ay **one**	Οι λέξεις της πλήρους έκφρασης ταιριάζουν με τους χαρακτήρες της συντόμευσης, λαμβάνοντας υπόψη τους διαφορετικούς τρόπους με τους οποίους μπορούμε να αναπαραστήσουμε έναν αριθμό.	
7	ULN	**U**pper **L**imit , of **N**ormal	Οι συμβολοακολουθίες "of" (πρόθεση) και "," (σημείο στίξης) δε συνεισφέρουν στο σχηματισμό της συντόμευσης.	Χαλαρά Συνδεδεμένα
8	LVLS	**L**ast **V**isit of the **L**ast **S**ubject	Οι λέξεις "of" (πρόθεση) και "the" (άρθρο) δε συνεισφέρουν στον σχηματισμό της συντόμευσης	
9	PDE-5	**P**hospho**die**sterase type **5**	Η λέξη "type" δε συμμετέχει στο σχηματισμό της πλήρους έκφρασης	
10	DSM-IV	**D**iagnostic and **S**tatistical manual of **M**ental disorders, **4th** edition	Οι λέξεις "and", "of" (stop words) , "manual, "disorder" and "edition" δε συμμετέχουν στο σχηματισμό της πλήρους έκφρασης	
11	ACDA	**A**cid **C**itrate **D**extrose solution **A**	Η λέξη "solution" δε συνυπολογίζεται κατά το σχηματισμό της συντόμευσης	
12	CABG	**C**oronary **A**rtery **B**ypass **G**raft procedure	Η λέξη "procedure" δε συνεισφέρει στο σχηματισμό της συντόμευσης, ωστόσο, ορίζει την ευρύτερη κατηγορία στην οποία ανήκει η φράση.	
13	EKG	**E**lectrocardio**g**ram	"EKG" προέρχεται από τη λέξη "**E**lektro**k**ardio**g**ramm" (Γερμανικά)	Μερικώς ή Καθόλου Συνδεδεμένα
14	SUKL	**S**tate Institute for Drug Control	"SUKL" προέρχεται από τη φράση "**S**tátní **ú**stav pro **k**ontrolu **l**éčiv" (Τσέχικα)	
15	DL	**D**azit	"DL" προέρχεται από τη λέξη "**D**es**l**oratadine", "Dazit" είναι η εμπορική ονομασία του φαρμάκου	
16	AZT	Zidovudine	"AZT" προέρχεται από τη λέξη "**Az**ido**t**hymidine", Zidovudine" είναι η Διεθνής Μη-ιδιοκτησίας Ονομασία του φαρμάκου	
17	MDX010	Ipilimumab	"MDX-010" είναι ο κωδικός για το φάρμακο με Διεθνή Μη-ιδιοκτησίας ονομασία "Ipilimumab"	
18	C15	Blood and lymphatic diseases	"C15" είναι ο MeSH Headings κωδικός για τον όρο "Hemic and lymphatic diseases"	

Πίνακας 12: Κατηγοριοποίηση των ζευγαριών συντόμευσης-έκφρασης και παραδείγματα

Η τρίτη κατηγορία (μερικώς ή καθόλου συνδεδεμένα) αποτελείται από τα ζευγάρια εκείνα στα οποία η συντόμευση έχει μερική ή καθόλου ομοιότητα με τους χαρακτήρες της φράσης, καθώς προέρχεται από κάποια διαφορετική φράση από αυτήν που υπάρχει στο κείμενο. Για παράδειγμα, η φράση που αναφέρεται στο κείμενο μπορεί να είναι στα Αγγλικά, ενώ η συντόμευση να προέρχεται από κάποια άλλη, σημασιολογικά ισοδύναμη, φράση εκφρασμένη σε κάποια άλλη γλώσσα (παραδείγματα 13 και 14). Επίσης, ένα φάρμακο έχει αρκετά ονόματα, με αποτέλεσμα, ορισμένες φορές, η συντόμευση που χρησιμοποιείται να μπορεί να προέρχεται από κάποια άλλη ονομασία του φαρμάκου, διαφορετική από αυτήν που ορίζεται ρητά μέσα στο κείμενο (παραδείγματα 13 και 14). Μια συνοπτική περιγραφή της ονοματολογίας των φαρμάκων υπάρχει στο παράρτημα 30.3. Οι κωδικοί αποτελούν μια ειδική περίπτωση και συνήθως ορίζονται αυθαίρετα κατά την εισαγωγή μιας έννοιας σε ένα σύστημα κωδικοποίησης. Αν και δημιουργήθηκαν πρωτίστως για λόγους αναφοράς και επικοινωνίας μεταξύ των συστημάτων λογισμικού, μερικούς από τους κωδικούς αυτούς τους συναντάμε ακόμη και σε κλινικές δοκιμές, όπως κωδικούς φαρμάκων (παράδειγμα 17) καθώς επίσης και κωδικούς από ευρέως διαδεδομένα συστήματα κατηγοριοποίησης (παράδειγμα 18).

Οι αριθμοί έχουν ένα διακριτό ρόλο κατά τη διαδικασία εντοπισμού της πλήρους μορφής μιας συντόμευσης. Γενικά, (εξαιρώντας τους κωδικούς) υπάρχουν και στη συντετμημένη και στην πλήρη μορφή μιας φράσης. Όμως, μπορεί να αναπαρίστανται με Αραβικούς ή Ρωμαϊκούς χαρακτήρες ή ακόμη και λέξεις, όπως στα παραδείγματα 6 και 10 του παραπάνω πίνακα. Τα σημεία στίξης (κενά, παύλες) καθώς επίσης και οι αλλαγές στη μορφή (πχ., από μικρά σε κεφαλαία) και είδος (π.χ., από γράμμα σε αριθμό) των χαρακτήρων της συντόμευσης αρκετές φορές χρησιμοποιούνται σκοπίμως, για να τονίζουν κάποιες ομάδες από συμβολοακολουθίες με ιδιαίτερη σημασία όπως στο παράδειγμα 10.

25.4Διακριτική Δυνατότητα των Λέξεων

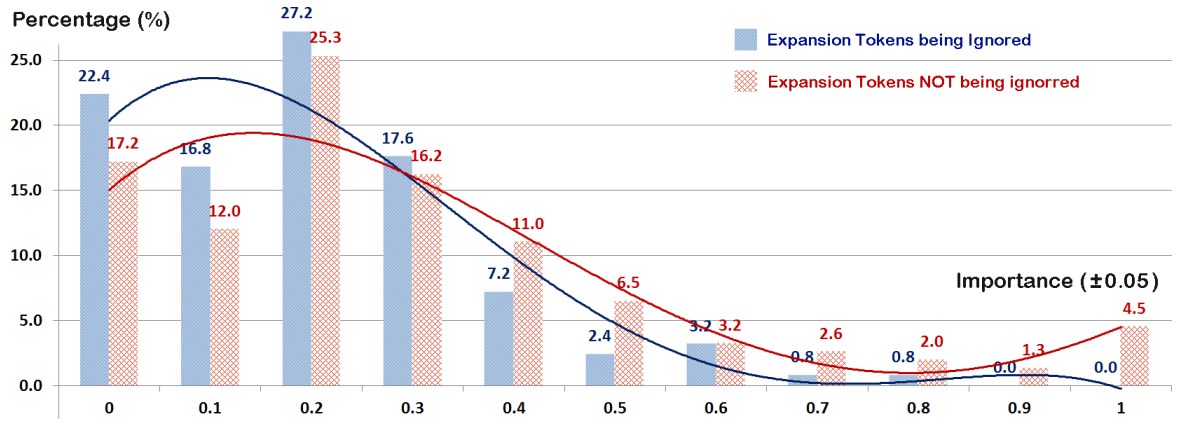
Σχετικά με τα ζευγάρια (συντόμευση, πλήρης έκφραση) που ανήκουν στη δεύτερη κατηγορία παρατηρήσαμε ότι η συντόμευση συνδέεται στενά με την πλήρη έκφραση, εάν αγνοήσουμε μία ή περισσότερες λέξεις. Η ανάλυση των λέξεων αυτών έδειξε ότι η διακριτική τους δυνατότητα είναι γενικά μικρότερη από αυτήν των υπολοίπων λέξεων της πλήρους έκφρασης. Η διακριτική δυνατότητα των λέξεων υπολογίστηκε, λαμβάνοντας υπόψη τη χρησιμοποίηση των λέξεων αυτών σε όλο το φάσμα των διαθέσιμων κλινικών μελετών. Πιο συγκεκριμένα, αρχικά καταγράψαμε τον αριθμό των μελετών που περιέχουν τις λέξεις αυτές (είτε ως έχουν είτε σε κάποια άλλη μορφή) και έπειτα υπολογίσαμε τη διακριτική τους δυνατότητα χρησιμοποιώντας την Εξίσωση 3, σύμφωνα με την οποία όσο πιο συχνά χρησιμοποιείται μία λέξη τόσο μικρότερη είναι η διακριτική της δυνατότητα. Οι τιμές αυτές κατόπιν κανονικοποιήθηκαν, ώστε να ανήκουν στο διάστημα μεταξύ 0 (ελάχιστη διακριτική δυνατότητα) και 1 (μέγιστη διακριτική δυνατότητα).

$$\text{Διακριτική Δυνατότητα Λέξης} = \log_{10}\left(\frac{\text{Συνολικός Αριθμός Κλινικών Μελετών}}{\text{Αριθμός Κλινικών Μελετών με την Λέξη}}\right)$$

Εξίσωση 3: Υπολογισμός Διακριτικής Δυνατότητας Λέξεων με βάση τη συχνότητα εμφάνισης στο σύνολο των διαθέσιμων κλινικών μελετών.

Το Σχήμα 28 απεικονίζει τη διακριτική δυνατότητα των λέξεων που δε συνεισφέρουν στο σχηματισμό της συντόμευσης (μπλε χρώμα) καθώς επίσης και τη διακριτική δυνατότητα των υπολοίπων λέξεων της πλήρους έκφρασης (κόκκινο χρώμα). Η διακριτική δυνατότητα του 98% των λέξεων που δε συνεισφέρουν στο σχηματισμό της συντόμευσης βρίσκεται μεταξύ 0 και 0.6 με τη μέση τιμή τους να βρίσκεται κοντά στο 0.2. Απ' την άλλη, η διακριτική δυνατότητα των υπολοίπων λέξεων καλύπτει όλο το φάσμα, γεγονός που δείχνει ότι στο σχηματισμό μιας συντόμευσης συχνά συμμετέχουν λέξεις ή φράσεις που δεν είναι τόσο σημαντικές, υπό την έννοια ότι η διακριτική τους δυνατότητα είναι μικρή. Όμως, όπως παρατηρήσαμε, στο 77% των περιπτώσεων στις

οποίες ήταν απαραίτητο να «αγνοήσουμε» κάποιες λέξεις της πλήρους έκφρασης, η μέση τιμή της διακριτικής τους δυνατότητας ήταν μικρότερη από τη μέση τιμή της διακριτικής δυνατότητας των υπολοίπων λέξεων της έκφρασης, ενώ, σχεδόν σε όλες τις περιπτώσεις, η πιο «σημαντική» λέξη είχε συνεισφέρει στην κατασκευή της συντόμευσης με έναν ή περισσότερους χαρακτήρες.

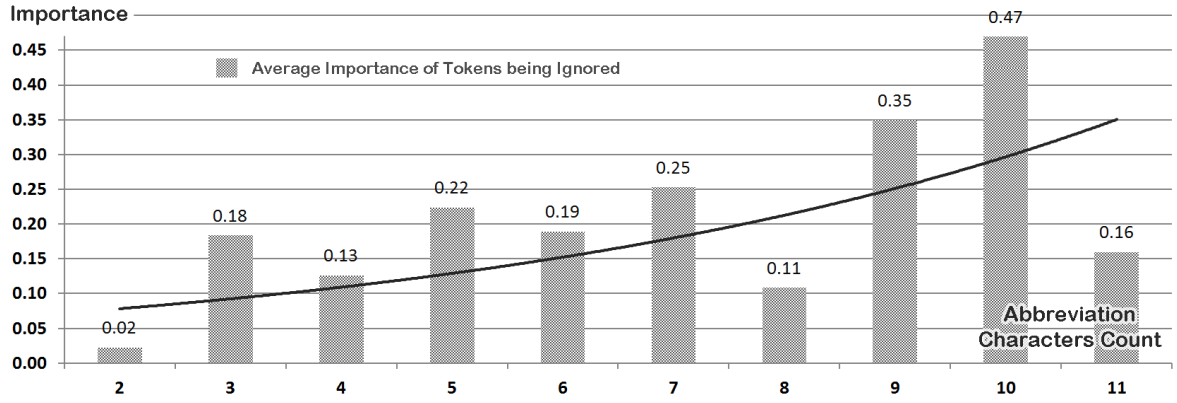


Σχήμα 28: Κατηγοριοποίηση των λέξεων που χρησιμοποιούνται στην πλήρη μορφή μιας συντόμευσης με βάση τη διακριτική τους δυνατότητα.

Η παραπάνω ανάλυση δείχνει καθαρά ότι οι χρήστες συχνά αγνοούν τις λέξεις ή φράσεις με μικρή διακριτική δυνατότητα κατά το σχηματισμό μιας συντόμευσης. Η παρατήρηση αυτή υποστηρίζεται επίσης από το γεγονός ότι στο 95% των περιπτώσεων ο αριθμός των χαρακτήρων της συντόμευσης ήταν 3 ή περισσότεροι. Συνεπώς, εάν οι χρήστες δεν είχαν αγνοήσει μία ή περισσότερες λέξεις κατά το σχηματισμό της συντόμευσης, αυτή θα αποτελούνταν από 4 ή 5 ή και περισσότερους χαρακτήρες και, συνεπώς, θα ήταν δύσκολο να χρησιμοποιηθεί στο υπόλοιπο κείμενο.

Επίσης, όπως μπορούμε να δούμε και στο Σχήμα 29, η διακριτική δυνατότητα των λέξεων που αγνοούνται επηρεάζεται από το μέγεθος της συντόμευσης και της αντίστοιχης έκφρασης. Πιο συγκεκριμένα, η μέση τιμή της διακριτικής δυνατότητας των λέξεων που αγνοούνται αυξάνεται, καθώς αυξάνεται και ο αριθμός των χαρακτήρων της συντόμευσης. Αυτό δείχνει ότι οι χρήστες, για να διατηρήσουν το μέγεθος μιας

συντόμευσης σχετικά μικρό, αγνοούν ορισμένες λέξεις οι οποίες σε κάποιες άλλες εκφράσεις συμμετέχουν ενεργά στο σχηματισμό της συντετμημένης τους μορφής.



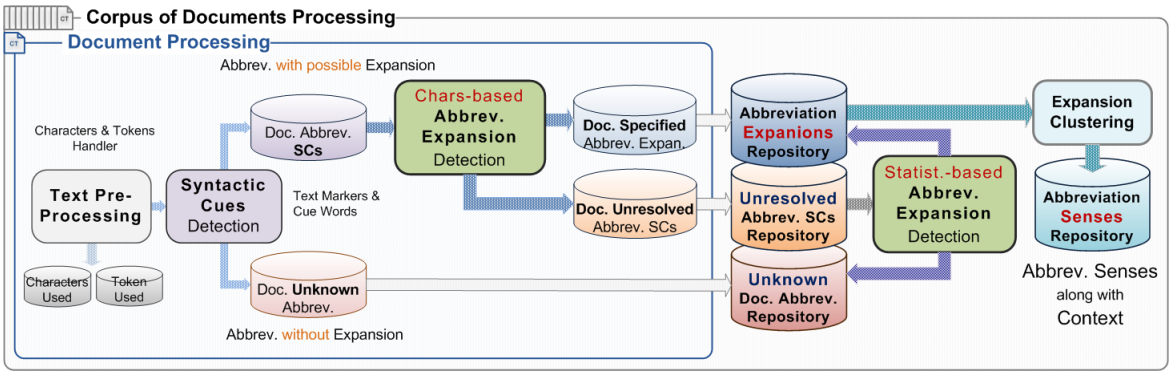
Σχήμα 29: Η μέση τιμή της διακριτικής δυνατότητας των λέξεων που δε συμμετέχουν στο σχηματισμό μιας συντόμευσης με βάση το μέγεθος των συντομεύσεων.

26 *Σύστημα Αναγνώρισης Συντομεύσεων*

26.1Συνολική Προσέγγιση

Το Σχήμα 30 απεικονίζει τη διαδικασία που ακολουθήθηκε και τα επιμέρους λογισμικά συστήματα που αναπτύχθηκαν για τον εντοπισμό της σημασίας των συντομεύσεων που χρησιμοποιούνται σε 190 χιλιάδες κλινικές μελέτες που κατεβάσαμε από την τοποθεσία clinicaltrials.gov [35]. Αρχικά, ήταν απαραίτητη η προεπεξεργασία των δεδομένων μιας κλινικής δοκιμής, προκειμένου να χειριστούμε αποτελεσματικά τις περιπτώσεις εκείνες όπου χρησιμοποιούνται διαφορετικά σύμβολα για το ίδιο σημείο στίξης, καθώς επίσης και πιθανά θέματα διαχωρισμού των επιμέρους τμημάτων μιας πρότασης (tokenization issues). Ακολούθως, εξετάσαμε το περιεχόμενο της κάθε κλινικής μελέτης και ειδικότερα των τμημάτων εκείνων του εγγράφου όπου η πληροφορία είναι εκφρασμένη με βάση τη φυσική γλώσσα, όπως ο τίτλος της μελέτης, η περιγραφή της, τα κριτήρια καταλληλότητας και η σύνοψη. Κατά τη διάρκεια της παραπάνω διαδικασίας, για κάθε έγγραφο καταγράψαμε το αναγνωριστικό της μελέτης, την ημερομηνία όπου ξεκίνησε καθώς επίσης τις ιατρικές καταστάσεις που μελετήθηκαν και τις παρεμβάσεις που έλαβαν χώρα, τα οποία είναι εκφρασμένα με τους όρους της Mesh οντολογίας [52]. Επίσης, για τις συντομεύσεις που χρησιμοποιούνται σε κάθε κλινική μελέτη, ανεξάρτητα αν η πλήρης μορφή τους οριζόταν μέσα στο κείμενο ή όχι, καταγράψαμε όλα τα σημεία στα οποία

χρησιμοποιούνται καθώς επίσης και το κείμενο που προηγούνταν ή ακολουθούσε μέσα στην πρόταση στην οποία εμφανίζονταν.



Σχήμα 30: Συνολική Προσέγγιση για τον εντοπισμό της Σημασίας των Συντομεύσεων των Κλινικών Μελετών.

Για τον εντοπισμό της πλήρους μορφής των συντομεύσεων, αρχικά εντοπίσαμε τις συντομεύσεις που χρησιμοποιούνται μέσα στο κείμενο και ειδικότερα τα σημεία εκείνα του κειμένου στα οποία η συντόμευση και η αντίστοιχη πλήρης έκφραση πιθανώς να υπάρχουν (syntactic cues). Ακολούθως, εξετάσαμε εάν η πλήρης μορφή των συντομεύσεων παρέχεται στην περιγραφή της κλινικής μελέτης ή όχι, χρησιμοποιώντας δύο διαφορετικές τεχνικές. Η πρώτη τεχνική βασίζεται κυρίως στους χαρακτήρες που απαρτίζουν τόσο τη συντόμευση όσο και την πλήρη έκφραση που προηγείται ή ακολουθεί, για να εντοπίσει στοιχεία ή ενδείξεις που δείχνουν ξεκάθαρα ότι υπάρχει μία σύνδεση μεταξύ τους. Η δεύτερη τεχνική βασίζεται αποκλειστικά και μόνο στη συχνότητα με την οποία μία φράση (ή ιδανικά έννοια) συνοδεύει την αντίστοιχη συντόμευση, λαμβάνοντας υπόψη το κείμενο που υπάρχει σε όλες τις διαθέσιμες κλινικές μελέτες.

Οι αλγόριθμοι που βασίζονται στους χαρακτήρες παρέχουν άμεσα την πλήρη μορφή των συντομεύσεων που ορίζονται, με την προϋπόθεση ότι υπάρχει μια εμφανής συσχέτιση μεταξύ τους, σε αντίθεση με τους αλγορίθμους που βασίζονται σε στατιστικά, οι οποίοι προϋποθέτουν την επεξεργασία όλων των διαθέσιμων μελετών, ενώ απαιτείται να επανεξετάσουμε τις μελέτες στις οποίες χρησιμοποιούνται οι συντομεύσεις που

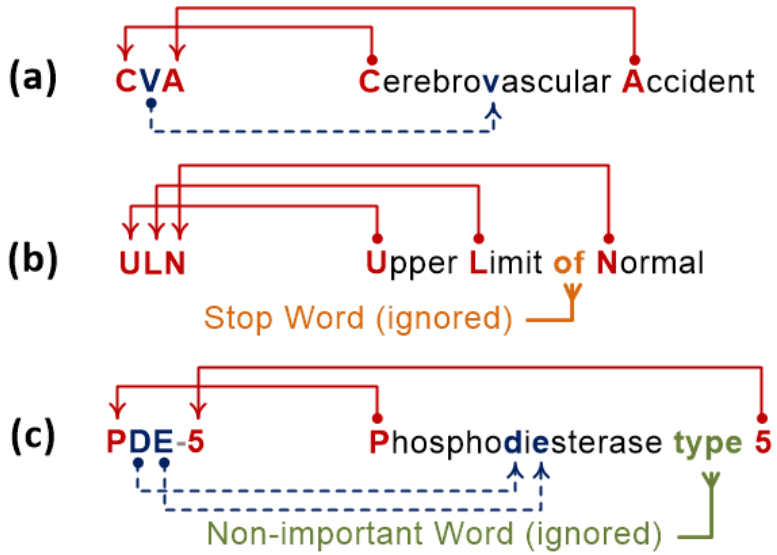
καταφέραμε να εντοπίσουμε τη σημασία τους, για να αποφανθούμε αν, όντως, αυτές ορίζονταν μέσα στο κείμενο ή όχι. Δεδομένης της ύπαρξης πολλών σημασιολογικά ισοδύναμων εκφράσεων, τα δεδομένα που συλλέχθηκαν επεξεργάστηκαν περαιτέρω, προκειμένου να εντοπίσουμε τις πιθανές σημασίες της κάθε συντόμευσης καθώς επίσης και επιπρόσθετη πληροφορία για καθεμία από αυτές, συμπεριλαμβανομένου του αριθμού των μελετών που χρησιμοποιούνται, το ευρύτερο πεδίο γνώσης των εγγράφων στα οποία αναφέρονται με την εκάστοτε σημασία, καθώς και συγκεκριμένες λέξεις ή φράσεις που τις συνοδεύουν. Ακολούθως, επανεξετάσαμε τις κλινικές μελέτες, προκειμένου να εντοπίσουμε τη σημασία των συντομεύσεων που χρησιμοποιούνται, χωρίς να αναφέρεται κάπου στο κείμενο η πλήρης μορφή τους. Τέλος, αναλύσαμε τα δεδομένα που εντοπίσαμε σε όλες τις φάσεις της προαναφερθείσας διαδικασίας.

26.2Αλγόριθμοι και Τεχνικές Αναγνώρισης Συντομεύσεων

26.2.14.2.1. Αλγόριθμοι Ευθυγράμμισης Χαρακτήρων

Σχετικά με την ευθυγράμμιση των χαρακτήρων της συντόμευσης με τους χαρακτήρες της πλήρους έκφρασής της, υλοποιήσαμε τρεις διαφορετικές στρατηγικές οι οποίες αναζητούν τη μικρότερη σε μήκος έκφραση που ταιριάζει με τη συντόμευση. Και οι τρεις στρατηγικές ακολουθούν την ίδια λογική, προκειμένου να εξετάσουν εάν η συντόμευση ταιριάζει με την υποψήφια πλήρη έκφραση. Πιο συγκεκριμένα, αρχικά, εξετάζουν εάν ο πρώτος χαρακτήρας των λέξεων της φράσης χρησιμοποιείται στη συντόμευση και μάλιστα με την ίδια σειρά (βήμα 1). Στην συνέχεια, εξετάζουν εάν οι υπόλοιποι χαρακτήρες της συντόμευσης υπάρχουν στο εσωτερικό των αντίστοιχων λέξεων (βήμα 2). Τέλος, εξετάζουν την ευρύτερη κατηγορία στην οποία ανήκουν οι εναπομείνασες λέξεις της φράσης που δε συνεισφέρουν στο σχηματισμό της συντόμευσης καθώς επίσης και τη διακριτική τους δυνατότητα (βήμα 3). Οι διαφορετικές τεχνικές που υλοποιήθηκαν διαφέρουν στις λέξεις που επιτρέπουν να αγνοήσουμε, προκειμένου να θεωρήσουμε μία

ευθυγράμμιση επιτυχή, με την προϋπόθεση ότι όλοι οι χαρακτήρες της συντόμευσης έχουν εντοπιστεί και μάλιστα με την ίδια σειρά. Πιο συγκεκριμένα, στην πρώτη στρατηγική, πρέπει να συμμετέχουν όλες οι λέξεις της πλήρους έκφρασης. Στη δεύτερη στρατηγική μπορούμε να αγνοήσουμε τις λέξεις τέλους (stop words), ενώ στην τρίτη στρατηγική μπορούμε επιπρόσθετα να αγνοήσουμε τις «μη σημαντικές» λέξεις, λαμβάνοντας υπόψη τη διακριτική τους δυνατότητα. Οι παραπάνω στρατηγικές εφαρμόζονται με τη σειρά που αναφέρουμε, προκειμένου να εντοπίσουμε την πλήρη μορφή της συντόμευσης με βάση το κείμενο που προηγείται ή έπεται, ανάλογα με το μοτίβο στο οποίο συμμετέχει η συντόμευση.



Σχήμα 31: Ευθυγράμμιση των χαρακτήρων της συντόμευσης με τους χαρακτήρες της πλήρους έκφρασης σε τρεις διαφορετικές περιπτώσεις.

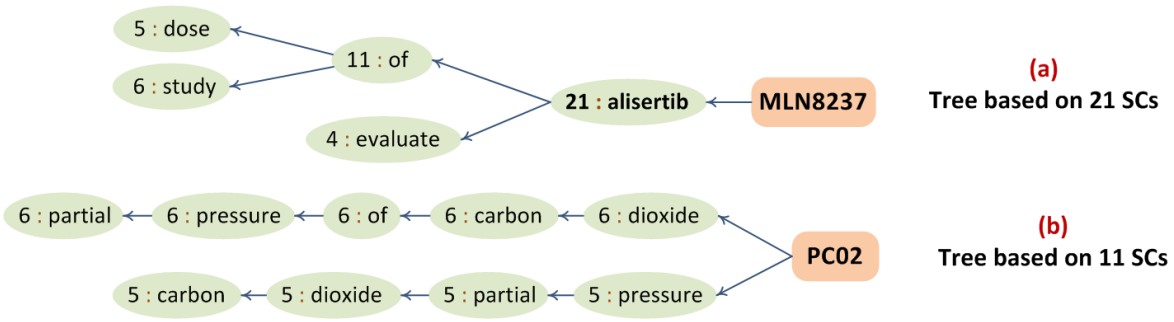
Κατά τα πρώτα στάδια της παραπάνω διαδικασίας, μπορεί να είναι εφικτή παραπάνω από μία ευθυγράμμιση των χαρακτήρων (για την ακρίβεια, ενός υποσυνόλου) της συντόμευσης με τις λέξεις της πλήρους μορφής της. Για το σκοπό αυτό, εξετάσαμε κάθε φορά όλους τους πιθανούς τρόπους ευθυγράμμισης, για να δούμε εάν, όντως, η συντόμευση ταιριάζει με την αντίστοιχη φράση. Επίσης, οι αριθμοί που αναφέρονται τόσο στη συντόμευση όσο και στην πλήρη έκφρασή της θα πρέπει να ταιριάζουν, ανεξάρτητα εάν είναι εκφρασμένοι χρησιμοποιώντας Αραβικά ή Ρωμαϊκά σύμβολα,

συμπεριλαμβανομένων των περιπτώσεων εκείνων όπου οι αριθμοί είναι εκφρασμένοι σε φυσική γλώσσα. Επιπλέον, για τους χαρακτήρες που δεν ανήκουν στην Αγγλική αλφάβητο, επιτρέψαμε την αντιστοίχιση μεταξύ τους με βάση την ομοιότητά τους. Για παράδειγμα, το Ελληνικό γράμμα «β» ταιριάζει με τον Αγγλικό χαρακτήρα «b», γεγονός που μας επιτρέπει να ευθυγραμμίσουμε τη συντόμευση «β-hCG» με την πλήρη έκφρασή της «beta-human chorionic gonadotropin».

26.2.2 Αλγόριθμοι Βασισμένοι στη Συνύπαρξη Εκφράσεων

Για τον εντοπισμό της σημασίας των μη-ακρωνύμου τύπου συντομεύσεων βασιστήκαμε στην επανεμφάνιση του ζευγαριού συντόμευσης – πλήρους έκφρασης σε όλο το σύνολο των διαθέσιμων κλινικών μελετών που εξετάστηκαν. Πιο συγκεκριμένα, για κάθε συντόμευση που βρισκόταν περιστοιχισμένη από παρενθέσεις ή άγκιστρα, αρχικά συγκεντρώσαμε το κείμενο της πρότασης που προηγούνταν καθεμιάς από αυτές σε όλες τις κλινικές μελέτες που τις συναντήσαμε και κατόπιν κατασκευάσαμε ένα δένδρο με βάση τις λέξεις των κειμένων που συλλέχθηκαν και κυρίως τη θέση της κάθε λέξης στο κείμενο αυτό. Το δένδρο αποτελούνταν αρχικά από ένα μόνο κόμβο (ρίζα) και κατόπιν ανανεώθηκε με βάση τις λέξεις / φράσεις που προηγούνταν της εκάστοτε συντόμευσης. Το δένδρο που προέκυψε, δείχνει τις λέξεις που συναντήσαμε πριν τη συντόμευση και κυρίως πόσες φορές η καθεμιά από αυτές εμφανίζεται στην αντίστοιχη θέση πριν τη συντόμευση, με την προϋπόθεση ότι οι λέξεις που προηγούνται είναι ακριβώς οι ίδιες.

Στο Σχήμα 32 μπορούμε να δούμε τα δένδρα που κατασκευάστηκαν σε δύο διαφορετικές περιπτώσεις. Τα δένδρα αυτά έχουν «κλαδευτεί» για τις ανάγκες της παρουσίασης, διαγράφοντας τις ακμές και τους αντίστοιχους κόμβους των υπογράφων, οι οποίοι χρησιμοποιούνται σπάνια (σε λιγότερο από το 10% των περιπτώσεων).



Σχήμα 32: Τα δένδρα που κατασκευάστηκαν με βάση τις λέξεις – φράσεις που προηγούνται των συντομεύσεων (a) "MLN8237" και (b) "PC02".

Το δένδρο που κατασκευάστηκε για τη συντόμευση «MLN8237» βασίστηκε στο κείμενο που προηγούνταν της συντόμευσης αυτής σε 21 διαφορετικές περιπτώσεις. Από το δένδρο αυτό, μπορούμε εύκολα να συμπεράνουμε ότι η πλήρης έκφραση της συντόμευσης αυτής είναι η λέξη «Alisertib». Όμως, στην περίπτωση της «PC02» συντόμευσης δεν μπορούμε να συμπεράνουμε άμεσα την πλήρη μορφή της, καθώς υπάρχουν παραπάνω από μία διαφορετικές εκφράσεις της ίδιας πλήρους έκφρασης, που επηρεάζουν το δένδρο αυξάνοντας το πλάτος του. Για το σκοπό αυτό, επανεξετάσαμε το δένδρο που δημιουργήθηκε, προκειμένου να εντοπίσουμε σημασιολογικά ισοδύναμες λέξεις ή φράσεις. Πιο συγκεκριμένα, για κάθε κόμβο εξετάσαμε αν η φράση που σχηματίζεται, αν διασχίσουμε τους ενδιάμεσους κόμβους μέχρι να φτάσουμε στην ρίζα του δένδρου, είναι σημασιολογικά ισοδύναμη με κάποια άλλη φράση που μπορούμε να δημιουργήσουμε. Στην περίπτωση που εντοπίστηκαν δύο σημασιολογικά ισοδύναμες φράσεις, αντικαταστήσαμε τους αντίστοιχους κόμβους από έναν που περιείχε τις δύο φράσεις καθώς και τον αριθμό εμφάνισης για καθεμία από αυτές.

Τέλος, για τον εντοπισμό της σημασίας της συντόμευσης εξετάσαμε το «ανανεωμένο» δένδρο, για να εντοπίσουμε τις λέξεις ή φράσεις που σχηματίζονται με την προϋπόθεση ότι η συχνότητα εμφάνισής τους ήταν μεγαλύτερη από ένα κατώτερο όριο (threshold). Λαμβάνοντας υπόψη ότι είχαμε ήδη εντοπίσει την πλήρη έκφραση των συντομεύσεων στις οποίες υπάρχει μια εμφανής συσχέτιση μεταξύ τους, το κατώτερο

όριο ορίστηκε να είναι το 75%, το οποίο πρακτικά σημαίνει ότι μπορούμε να εντοπίσουμε το πολύ μία διαφορετική σημασία για κάθε συντόμευση. Ωστόσο, η πλήρης έκφραση μπορεί να εκφράζεται με παραπάνω από μία σημασιολογικά ισοδύναμη φράση. Επίσης, για κάθε συντόμευση θα πρέπει να έχουμε παραπάνω από π.χ. 10 κείμενα (μέρος της παραμετροποίησης του συστήματος) για την εξαγωγή έγκυρων συμπερασμάτων. Τέλος, η πλήρης έκφραση θα πρέπει να περιέχει τουλάχιστον μία λέξη με υψηλή διακριτική δυνατότητα.

26.2.3 Εντοπισμός Σημασιολογικά Ισοδύναμων Εκφράσεων

Για να βρούμε τις πιθανές σημασίες των συντομεύσεων, χρησιμοποιήσαμε έναν απλό αλγόριθμο ομαδοποίησης, ο οποίος επαναληπτικά εξέτασε τις πλήρεις μορφές που καταγράφηκαν για κάθε συντόμευση, για τον εντοπισμό των σημασιολογικά ισοδύναμων εκφράσεων. Πιο συγκεκριμένα, αρχικά, επιλέξαμε τυχαία μία έκφραση και ακολούθως εξετάσαμε τις υπόλοιπες εκφράσεις. Οι εκφράσεις που βρέθηκαν ότι ταιριάζουν τοποθετήθηκαν στην ίδια ομάδα (εκφράζει τις διαφορετικές εκφράσεις που καταγράφηκαν για την ίδια έννοια). Έπειτα, συνεχίσαμε με την επεξεργασία των υπολοίπων εκφράσεων που δεν είχαν ομαδοποιηθεί. Η παραπάνω διαδικασία ολοκληρώθηκε με την τοποθέτηση όλων των εκφράσεων σε κάποια ομάδα, το όνομα της οποίας επιλέχθηκε να είναι η πιο ευρέως χρησιμοποιούμενη έκφραση. Για καθεμιά από τις διαφορετικές σημασίες που εντοπίσαμε, καταγράψαμε τις κλινικές μελέτες στις οποίες αναφέρεται, καθώς επίσης και τις συγκεκριμένες προτάσεις που χρησιμοποιείται με τη σημασία αυτή.

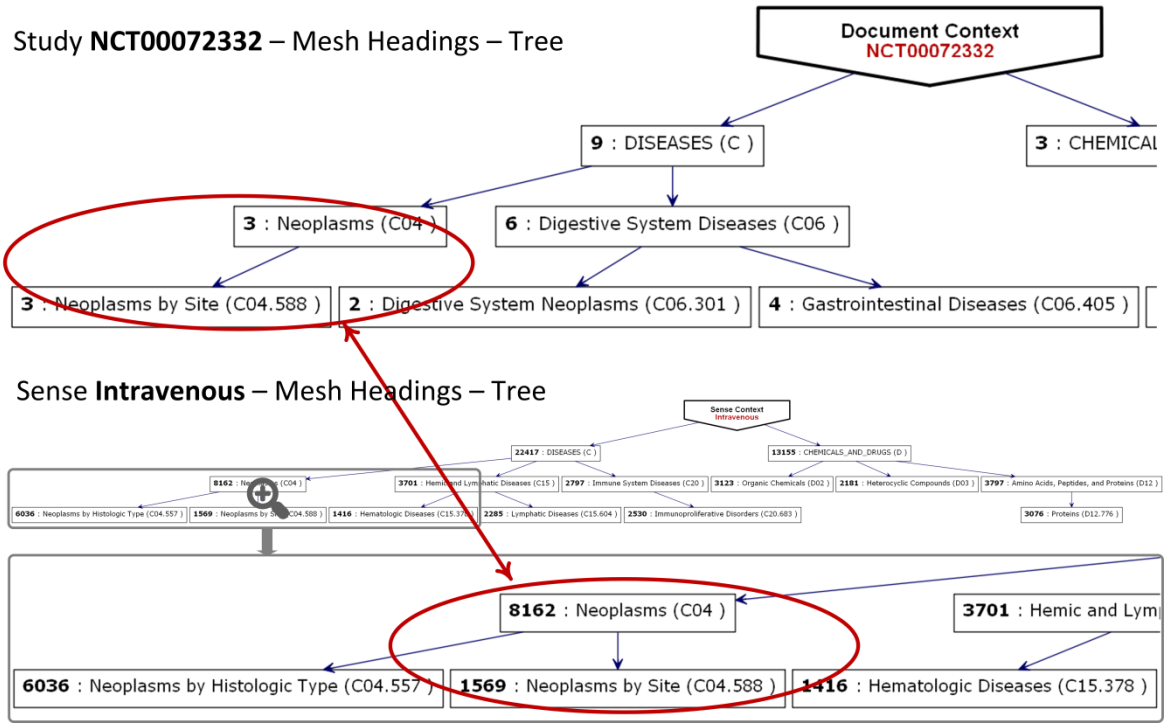
Για τον εντοπισμό των σημασιολογικά ισοδύναμων εκφράσεων μιας συντόμευσης χρησιμοποιήσαμε αρκετές τεχνικές, συμπεριλαμβανομένης της εξάλειψης των σημείων στίξης (elimination) και των λέξεων τέλους (stop words). Επίσης, χρησιμοποιήσαμε τον Porter αλγόριθμο [171], για να εστιάσουμε στη ρίζα της λέξης (stem) και να χειριστούμε

αποτελεσματικά τις διαφορετικές μορφές στις οποίες μπορεί να χρησιμοποιείται μέσα στο κείμενο (π.χ., πληθυντικό αριθμό). Ακόμη, λάβαμε υπόψη τη διακριτική δυνατότητα των λέξεων, γεγονός που μας επέτρεψε να εντοπίσουμε σημασιολογικά ισοδύναμες εκφράσεις, ακόμη και αν δεν είχαν ακριβώς τις ίδιες λέξεις, όπως στην περίπτωση των φράσεων «Upper Limit of Normal» και «Upper Limit of the Normal Range». Επίσης, καταγράψαμε όλες τις πιθανές μορφές ενός αριθμού (π.χ., "two", "ii", "2", "second", "2nd"), γεγονός που μας επέτρεψε να ταιριάξουμε συμβολοακολουθίες που αναφέρονται στους ίδιους αριθμούς. Στην εργασία μας λάβαμε υπόψη πιθανά θέματα διαχωρισμού των λέξεων των πλήρων εκφράσεων που συχνά υπάρχουν (π.χ., «Periacetabular osteotomy» και «Peri Acetabular Osteotomy»), καθώς επίσης και το γεγονός ότι η πλήρης έκφραση μπορεί εσωτερικά να χρησιμοποιεί μία άλλη συντόμευση, όπως στην περίπτωση των εκφράσεων «Continuous Intravenous Infusion» και «Continuous IV Infusion» που καταγράφηκαν για τη συντόμευση «CIVI». Τέλος, χρησιμοποιήσαμε τη Levenshtein μετρική [94], για να εντοπίσουμε εκφράσεις που είναι σχεδόν ίδιες από άποψη χαρακτήρων (π.χ., εξαιτίας κάποιου ορθογραφικού λάθους, όπως στην περίπτωση των συμβολοακολουθιών «electorcardogram» και «electrocardiogram» όπου στην πρώτη περίπτωση λείπει ο χαρακτήρας «i»).

26.2.4 Εύρεση της Σημασίας μη Ορισμένων Συντομεύσεων

Για τον εντοπισμό της σημασίας των συντομεύσεων που δε συνοδεύονταν από την πλήρη έκφρασή τους, αρχικά, εξετάσαμε τη βάση δεδομένων που κατασκευάστηκε, προκειμένου να εντοπίσουμε τις πιθανές σημασίες της συντόμευσης. Ακολούθως, εντοπίσαμε τις επικρατέστερες σημασίες, λαμβάνοντας υπόψη τον αριθμό εμφάνισης των υπόλοιπων ερμηνειών της ίδιας συντόμευσης. Στις περιπτώσεις εκείνες όπου υπήρχε μία μόνο πιθανή σημασία, υποθέσαμε ότι αυτή είναι και η σημασία της συντόμευσης μέσα στο κείμενο. Διαφορετικά, εξετάσαμε το ευρύτερο πεδίο γνώσης μέσα στο οποίο χρησιμοποιείται η

κάθε ερμηνεία, προκειμένου να εντοπίσουμε αυτή με την οποία χρησιμοποιείται μέσα στο κείμενο. Για το σκοπό αυτό, για κάθε πιθανή ερμηνεία, αρχικά κατασκευάσαμε ένα δένδρο με βάση τους όρους που χρησιμοποιούνται στις κλινικές δοκιμές για καθεμία από αυτές και στη συνέχεια το συγκρίναμε με βάση τους όρους που χρησιμοποιούνταν στην εκάστοτε κλινική μελέτη, για να βρούμε το δένδρο που ταιριάζει καλύτερα.



Σχήμα 33: Όροι της μελέτης «NCT00072332» και το ευρύτερο πεδίο γνώσης της πιθανής σημασίας «Intravenous»

Το Σχήμα 33 απεικονίζει το μεγαλύτερο μέρος του δένδρου που κατασκευάστηκε με βάση τους όρους της MeSH οντολογίας που χρησιμοποιήθηκαν για την περιγραφή της κλινικής μελέτης με αναγνωριστικό «NCT00072332». Στην ίδια εικόνα υπάρχει και το δένδρο που δημιουργήθηκε για την έννοια «Intravenous» της συντόμευσης «IV», με βάση τις κλινικές μελέτες στις οποίες η συντόμευση αυτή χρησιμοποιείται με την προαναφερθείσα σημασία. Από το δένδρο αυτό, διαγράψαμε τους κόμβους/όρους εκείνους που δε χρησιμοποιούνται συχνά, σε σύγκριση με τον αριθμό των μελετών στις οποίες χρησιμοποιείται η παραπάνω έννοια. Όπως μπορούμε να δούμε και στην εικόνα

αυτή, το πεδίο γνώσης στο οποίο η συντόμευση «IV» χρησιμοποιείται στη μελέτη «» χωρίς την πλήρη μορφή της ταιριάζει με το πεδίο γνώσης των εγγράφων στα οποία η συντόμευση αυτή χρησιμοποιείται με τη σημασία «Intravenous» και μάλιστα «καλύτερα» από τις υπόλοιπες σημασίες της συντόμευσης αυτής. Συνεπώς, μπορούμε να συμπεράνουμε ότι στην κλινική μελέτη «NCT00072332» η συντόμευση «IV» χρησιμοποιείται αντί της φράσης «Intravenous», παρά το γεγονός ότι η παραπάνω φράση δεν υπάρχει μέσα στο κείμενο.

27 *Αξιολόγηση Συστήματος*

Για να διασφαλίσουμε την ορθότητα του συστήματος και την εγκυρότητα των δεδομένων που συλλέχτηκαν, εξετάσαμε την ορθότητα των αλγορίθμων που υλοποιήθηκαν με βάση τις συντομεύσεις που είχαμε αρχικά ορίσει με ημιαυτόματο τρόπο σε ένα σύνολο από τυχαίες κλινικές μελέτες που διαλέξαμε. Πιο συγκεκριμένα, χρησιμοποιήσαμε το σύστημα που δημιουργήθηκε για τον εντοπισμό της σημασίας των συντομεύσεων που χρησιμοποιούνται σε καθεμιά από τις παραπάνω κλινικές μελέτες και στη συνέχεια συγκρίναμε τη σημασία που βρήκαμε με αυτήν που είχαμε ρητά ορίσει μέσα στο κείμενο, υπολογίζοντας την ακρίβεια (precision), ανάκληση (recall) καθώς και την τιμή F-measure (aka F-score). Η ακρίβεια υπολογίστηκε διαιρώντας τον αριθμό των ζευγαριών συντόμευση – πλήρης έκφραση, που εντοπίσαμε σωστά, με το συνολικό αριθμό των ζευγαριών που εντόπισε το σύστημά μας. Η ανάκληση, υπολογίστηκε διαιρώντας τον αριθμό των ζευγαριών, που εντοπίσαμε σωστά, με τον αριθμό των ζευγαριών, που θα έπρεπε να εντοπίσουμε. Η τιμή F-measure προέκυψε από τον συνδυασμό της ακρίβειας με την ανάκληση, σύμφωνα με την Εξίσωση 4:

$$\text{F measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Εξίσωση 4: Υπολογισμός της τιμής F-measure με βάση την ακρίβεια (precision) και ανάκληση (recall)

Το αποτέλεσμα της αξιολόγησης του συστήματος έδειξε ότι μπορεί να εντοπίσει με ακρίβεια 0.9912 τη σημασία των συντομεύσεων που ορίζονται μέσα στο κείμενο. Η ανάκληση είναι μικρότερη, με την τιμή της να ανέρχεται στο 0.8804. Ωστόσο, η συνολική αξιολόγηση του συστήματος ανέρχεται στο 0.9325 που είναι ικανοποιητική, δεδομένου ότι βρίσκεται πάνω από το μέσο όρο της αξιολόγησης των υπαρχόντων αλγορίθμων και τεχνικών που βρίσκεται γύρω στο 90%. Όπως φαίνεται και στις παραπάνω μετρήσεις, τα ζευγάρια της συντόμευσης με την πλήρη έκφραση που εντοπίζονται αυτόματα από το σύστημά μας είναι σχεδόν σε όλες τις περιπτώσεις σωστά. Σχετικά με τα ζευγάρια που δεν μπορέσαμε να εντοπίσουμε, αρκετά από αυτά δεν προέρχονταν από κάποια από τα γνωστά μοτίβα που υποστηρίζει το σύστημά μας. Για παράδειγμα, σε κάποιες περιπτώσεις η συντόμευση απλά ακολουθούσε την πλήρη έκφρασή της, χωρίς να βρίσκεται μέσα σε παρενθέσεις (π.χ., Alkaline phosphatase ALP > 2.5 x ULN). Επίσης, σε κάποιες άλλες περιπτώσεις υπήρχε κάποια απόσταση από τη συντόμευση και την πλήρη έκφρασή της ή υπήρχε επιπρόσθετη πληροφορία μέσα στις παρενθέσεις, γεγονός που δυσκόλευε το σύστημα να εντοπίσει σωστά τα αντίστοιχα ζευγάρια. Επιπλέον, σε κάποια ζευγάρια κάποιες από τις λέξεις που συνεισφέρουν στο σχηματισμό της συντόμευσης έλλειπαν από την πλήρη έκφραση που υπήρχε μέσα στο κείμενο, όπως στις ακόλουθες δύο περιπτώσεις «HCV: Hepatitis C» και «5-FU: Fluorouracil». Επιπρόσθετα, η πλήρης μορφή κάποιων συντομεύσεων δε χρησιμοποιούνταν αρκετά συχνά, γεγονός που εμπόδισε το σύστημά μας να εντοπίσει την πλήρη έκφρασή τους, χρησιμοποιώντας τεχνικές που βασίζονται στην επανεμφάνιση των ίδιων λέξεων.

28 *Ανάλυση και Επεξεργασία των Δεδομένων*

28.1Συντομεύσεις Κλινικών Μελετών

Η επεξεργασία 190 χιλιάδων κλινικών μελετών έδειξε ότι σε κάθε μελέτη χρησιμοποιούνται κατά μέσο όρο 6.8 διαφορετικές συντομεύσεις, από τις οποίες οι δύο στις τρεις χρησιμοποιούνται χωρίς να έχει αναφερθεί κάπου μέσα στο κείμενο η πλήρης μορφή τους. Από τις παραπάνω μετρήσεις εξαιρέσαμε σκοπίμως τις λατινικές συντομογραφίες (π.χ., i.e., e.g., vs., a.m., p.m., etc.) και τις μονάδες μέτρησης (π.χ., mg, mL, cm), οι οποίες συχνά χρησιμοποιούνται χωρίς να αναφέρεται στο κείμενο η λέξη από την οποία προέρχονται.

Στον Πίνακας 13 μπορούμε να δούμε τις τεχνικές που χρησιμοποιήθηκαν, για να εντοπίσουμε τη σημασία των συντομεύσεων που ορίζονται στις κλινικές μελέτες. Όπως φαίνεται και στον πίνακα, στις περισσότερες περιπτώσεις η συντόμευση είναι «στενά» συνδεδεμένη με την πλήρη έκφρασή της. Στο 58.36 των περιπτώσεων αυτών η συντόμευση προέρχεται από τον πρώτο χαρακτήρα των λέξεων της πλήρους έκφρασης. Όμως, σε αρκετές περιπτώσεις η συντόμευση είναι «χαλαρά» συνδεδεμένη με την πλήρη έκφραση, υπό την έννοια ότι κάποιες λέξεις της πλήρους μορφής τους δεν είχαν συνεισφέρει στη δημιουργία της συντόμευσης. Οι λέξεις αυτές είναι είτε «λέξεις τέλους» είτε λέξεις που δεν είναι τόσο σημαντικές (η διακριτική τους δυνατότητα είναι μικρή) σε

σχέση με τις υπόλοιπες λέξεις της πλήρους έκφρασης. Τέλος, με τη βοήθεια των τεχνικών που βασίζονται σε επανεμφάνιση της ίδιας λέξης ή φράσης, μπορέσαμε να εντοπίσουμε τη σημασία ενός περιορισμένου αριθμού συντομεύσεων σε σχέση με τις συντομεύσεις που εντοπίσαμε με τις τεχνικές που βασίζονται στην ευθυγράμμιση των χαρακτήρων.

ID	**Περιγραφή**	**Ζευγάρια**	**(%)**
T1	Όλες οι λέξεις της πλήρους έκφρασης συμμετείχαν στο σχηματισμό της συντόμευσης με έναν (τον πρώτο) ή παραπάνω (κάποιον από τους υπολοίπους) χαρακτήρες.	55540	72.26
T2	Η συντόμευση συνδέεται «στενά» με την πλήρη έκφραση (T1), εάν αγνοήσουμε μία ή παραπάνω λέξεις τέλους.	10945	14.23
T3	Η συντόμευση συνδέεται «στενά» με την πλήρη έκφραση (T1), εάν αγνοήσουμε μία ή παραπάνω λέξεις με μικρή διακριτική δυνατότητα	7479	9.72
T4	Η συντόμευση ταιριάζει με την πλήρη έκφραση, αν επιπρόσθετα αγνοήσουμε και τη σειρά με την οποία χρησιμοποιούνται οι λέξεις.	2389	3.11
ST	Η συντόμευση ταιριάζει με την πλήρη έκφραση με βάση τη συχνότητα χρήσης τους.	521	0.68

Πίνακας 13: Τεχνικές εντοπισμού συντόμευσης με πλήρη έκφραση και το ποσοστό των ζευγαριών που εντοπίστηκαν από αυτές.

Με βάση τις συντομεύσεις που ορίζονται στις κλινικές μελέτες που εξετάσαμε, κατασκευάσαμε μία βάση δεδομένων η οποία περιλαμβάνει 30 χιλιάδες συντομεύσεις, καθεμιά απ' τις οποίες έχει κατά μέσο όρο 1.8 πιθανές σημασίες. Περαιτέρω επεξεργασία των δεδομένων έδειξε ότι 11 χιλιάδες διαφορετικά ζευγάρια (συντόμευση – πλήρης έκφραση) που υπάρχουν στη βάση χρησιμοποιούνται αποκλειστικά και μόνο στην κλινική μελέτη στην οποία ορίζονται. Εξαιρώντας τα παραπάνω ζευγάρια, παρατηρήσαμε ότι οι πιθανές σημασίες των συντομεύσεων μειώνονται σημαντικά. Πιο συγκεκριμένα, κάθε συντόμευση έχει κατά μέσο όρο 1.25 πιθανές σημασίες, που είναι πολύ λιγότερες από τις

πιθανές σημασίες των βιοϊατρικών συντομεύσεων που ανέρχονται στις 4.61 [156]. Συνεπώς, η βάση δεδομένων που κατασκευάσαμε διευκολύνει τη διαδικασία εύρεσης της πιθανής σημασίας των συντομεύσεων που δεν ορίζονται στις κλινικές δοκιμές, καθώς δε χρειάζεται να εξετάσουμε επιπρόσθετες πιθανές σημασίες των συντομεύσεων οι οποίες κατά κανόνα δε χρησιμοποιούνται στην περιγραφή των κλινικών μελετών.

Η συλλογή και επεξεργασία των συντομεύσεων που δεν ορίζονται στις κλινικές μελέτες έδειξε ότι υπάρχουν 90 χιλιάδες διαφορετικές συντομεύσεις από τις οποίες μόνο οι 10 χιλιάδες υπάρχουν στη βάση δεδομένων που κατασκευάστηκε. Παρά το γεγονός ότι ένας μικρός αριθμός των μη-ορισμένων συντομεύσεων υπάρχει στη βάση δεδομένων μας, οι συντομεύσεις αυτές καλύπτουν το 70% των περιπτώσεων όπου μία συντόμευση χρησιμοποιείται στην κλινική μελέτη, χωρίς να αναφέρεται κάπου μέσα στο κείμενο η πλήρης μορφή της. Επίσης, παρατηρήσαμε ότι από τις 10 χιλιάδες συντομεύσεις που υπάρχουν στη βάση μας, οι 7.5 χιλιάδες έχουν μία μόνο πιθανή σημασία, ενώ οι υπόλοιπες 2.5 χιλιάδες έχουν κατά μέσο όρο 2 περίπου σημασίες. Επιπρόσθετα, οι συντομεύσεις με μία μόνο πιθανή σημασία χρησιμοποιούνται στο 74% των παραπάνω περιπτώσεων και συνεπώς μπορούμε εύκολα να εντοπίσουμε τη σημασία τους. Για την επιλογή της σημασίας των υπολοίπων συντομεύσεων, όπου υπάρχουν παραπάνω από μία σημασίες, χρησιμοποιήσαμε το ευρύτερο πεδίο γνώσης μέσα στο οποίο η εκάστοτε σημασία χρησιμοποιείται, για να επιλέξουμε την πιο κατάλληλη. Ωστόσο, όπως παρατηρήσαμε, σε αρκετές περιπτώσεις, το ευρύτερο πεδίο γνώσης στο οποίο χρησιμοποιούνται οι διαφορετικές σημασίες μιας συντόμευσης δεν μπορεί να μας προσδιορίσει με ακρίβεια ποια από τις διαθέσιμες σημασίες θα πρέπει να διαλέξουμε.

28.2Συντομεύσεις, Πλήρεις Εκφράσεις και Έννοιες

Περαιτέρω επεξεργασία των συντομεύσεων, και ειδικότερα της σημασίας τους, έδειξε ότι, σε αρκετές περιπτώσεις, υπάρχουν παραπάνω από μία συντομεύσεις για την ίδια σημασία,

οι οποίες συνήθως προέρχονται από την ύπαρξη διαφορετικών, σημασιολογικά ισοδύναμων, εκφράσεων [172]. Για παράδειγμα, το φάρμακο azidothymidine (AZT) είναι επίσης γνωστό ως zidovudine (ZDV). Τα παραπάνω ζευγάρια συντόμευσης με πλήρη έκφραση χρησιμοποιήθηκαν σε 131 και 105 κλινικές μελέτες αντίστοιχα. Ωστόσο, παρατηρήσαμε ότι σε 342 μελέτες η συντόμευση «AZT» ακολουθούσε την πλήρη έκφρασή της «zidovudine» (γενικό όνομα φαρμάκου) μέσα σε παρενθέσεις. Σημειώνουμε ότι μπορέσαμε επιτυχώς να το εντοπίσουμε το ζεύγος αυτό, χρησιμοποιώντας τις τεχνικές που βασίζονται σε στατιστικά στοιχεία που εξάγαμε από την ανάλυση του συνόλου των κλινικών μελετών. Επιπρόσθετα, παρατηρήσαμε ότι η συντόμευση «AZT» ακολουθούσε την πλήρη έκφρασή της «Retrovir» (εμπορικό όνομα φαρμάκου) σε 17 κλινικές μελέτες, την οποία, όμως, δεν μπορέσαμε να εντοπίσουμε αυτόματα, καθώς χρησιμοποιείται σε έναν πολύ περιορισμένο αριθμό κλινικών μελετών.

Σε ορισμένες περιπτώσεις παρατηρήσαμε ότι υπάρχουν παραπάνω από μία συντομεύσεις όχι μόνο για την ίδια έννοια αλλά ακόμη και για την ίδια πλήρη έκφρασή της. Για παράδειγμα, καταγράφηκαν πάνω από 10 διαφορετικές συντομογραφίες για την έκφραση «Hepatitis B surface antigen», όπως "HBs-Ag", "HBSAG", "HBsa" και "HepBsAg". Οι διαφορετικές συντομεύσεις που εντοπίστηκαν διαφέρουν στη μορφή (κεφαλαία ή μικρά γράμματα, χρησιμοποίηση σημείων στίξης, όπως κενά και παύλες) ή ακόμη και στον αριθμό των χαρακτήρων που περιέχουν. Για παράδειγμα, η συντόμευση «HBsa» αποτελείται από 4 μόνο χαρακτήρες, ενώ η «HepBsAg» αποτελείται από 7 χαρακτήρες. Οι διαφορές τους οφείλονται στην τεχνική που χρησιμοποίησαν οι χρήστες κατά το σχηματισμό της συντόμευσης και ειδικότερα στους χαρακτήρες που συμπεριέλαβαν από την πλήρη έκφρασή της. Για παράδειγμα, η συντόμευση «HBsa» προέρχεται από τον πρώτο χαρακτήρα της κάθε λέξης, ενώ η συντόμευση «HepBsAg» αποτελείται από τους 3 πρώτους διαδοχικούς χαρακτήρες της λέξης «Hepatitis», το

γράμμα «B», τον πρώτο χαρακτήρα της λέξης «surface» και τον πρώτο χαρακτήρα της κάθε συλλαβής της λέξης «antigen».

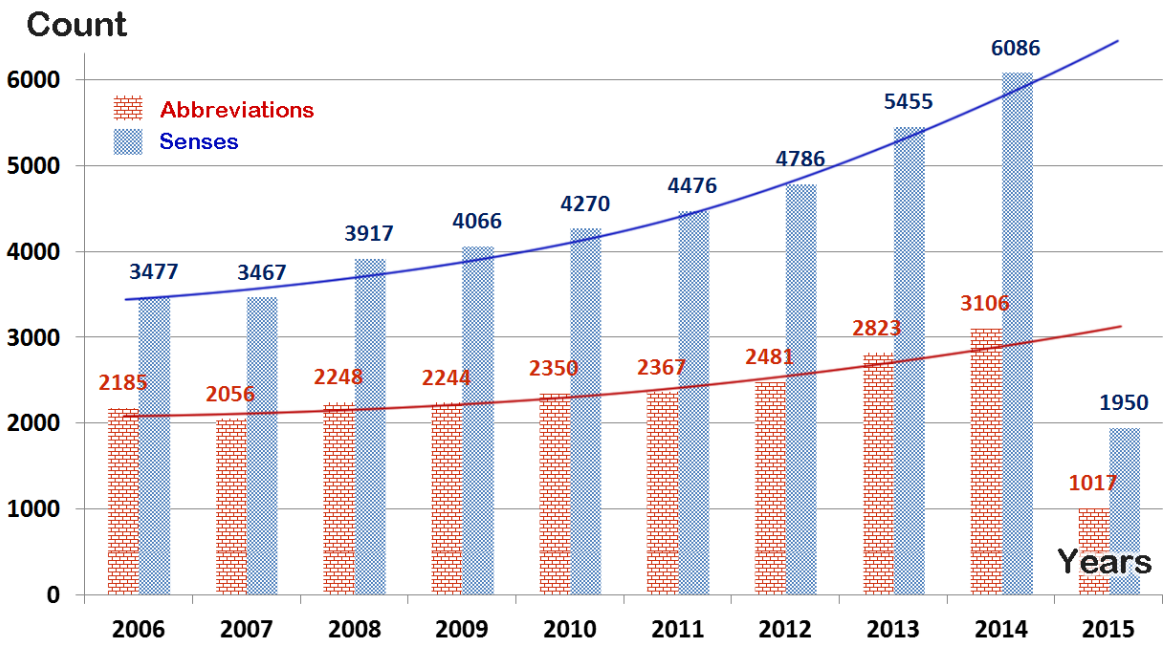
Η παρουσία ενός σημαντικού αριθμού συντομεύσεων για την ίδια έννοια δημιουργεί επιπρόσθετα προβλήματα στα συστήματα επεξεργασίας φυσικού κειμένου, καθώς θα πρέπει αφενός να εντοπίσουν τη σημασία των συντομεύσεων που ορίζονται μέσα στο κείμενο και αφετέρου τη σημασία των υπολοίπων συντομεύσεων που χρησιμοποιούνται, χωρίς να αναφέρεται η πλήρης μορφή τους. Στην πρώτη περίπτωση, η αναντιστοιχία μεταξύ της συντόμευσης και του ονόματος που παρέχεται στο κείμενο κάνει τους αλγορίθμους αναγνώρισης συντομεύσεων που βασίζονται σε τεχνικές ευθυγράμμισης χαρακτήρων να αποτυγχάνουν, ενώ οι τεχνικές που βασίζονται στην επαναχρησιμοποίηση της ίδιας φράσης προϋποθέτουν την εξέταση του συνόλου των διαθέσιμων εγγράφων καθώς επίσης και τη χρησιμοποίηση του ζεύγους σε ένα σημαντικό αριθμό από αυτά, προκειμένου να εξάγουν αξιόλογα συμπεράσματα. Στη δεύτερη περίπτωση, η εύρεση της σημασίας των συντομεύσεων που δεν ορίζονται ρητά μέσα στο κείμενο (δε συνοδεύονται από την πλήρη έκφρασή τους) είναι ακόμη πιο δύσκολη, καθώς η συντόμευση που χρησιμοποιείται στο κείμενο μπορεί να μην υπάρχει στη βάση δεδομένων μας με τη μορφή αυτή, αλλά, ακόμη και αν υπάρχει, μπορεί να έχει παραπάνω από μια πιθανές σημασίες, γεγονός που δυσκολεύει αρκετά τη διαδικασία εύρεσης της σημασίας με την οποία χρησιμοποιείται η συντόμευση μέσα στο κείμενο. Αυτό γίνεται πιο κατανοητό, αν λάβουμε υπόψη ότι ένας μεγάλος αριθμός των «μη ορισμένων» συντομεύσεων δεν υπάρχουν στη βάση που κατασκευάστηκε από την επεξεργασία των δεδομένων των κλινικών δοκιμών, και, συνεπώς, η σημασία τους θα πρέπει να αναζητηθεί σε διαφορετικές βάσεις, όπου οι συντομεύσεις έχουν κατά μέσο όρο πολύ περισσότερες σημασίες και, επομένως, η διαδικασία επιλογής είναι πιθανόν να παρέχει λανθασμένες απαντήσεις, με άμεσες επιπτώσεις στις αποφάσεις που παίρνονται. Ορισμένες δε

συντομεύσεις έχουν αποδειχθεί ιδιαίτερα επικίνδυνες για την υγεία και την ασφάλεια ενός ασθενούς (συντομεύσεις επιρρεπείς σε λάθη) και η χρήση τους θα πρέπει να αποφεύγεται (παράρτημα 30.4).

28.3Νέες Συντομεύσεις και Έννοιες

Μία άλλη παράμετρος που εξετάσαμε είναι η χρησιμοποίηση νέων συντομεύσεων και εννοιών τα τελευταία 10 χρόνια. Για το σκοπό αυτό, αρχικά, οργανώσαμε τις κλινικές μελέτες με βάση την ημερομηνία που έλαβαν χώρα και δημιουργήσαμε μία βάση δεδομένων που αποτελούνταν μόνο από τις συντομεύσεις και έννοιες που εντοπίσαμε στις κλινικές μελέτες που δημοσιεύτηκαν πριν το 2006. Ακολούθως, με βάση τις κλινικές μελέτες που δημοσιεύτηκαν στα επόμενα χρόνια (π.χ., 2006), εντοπίσαμε τις νέες συντομεύσεις και έννοιες, λαμβάνοντας υπόψη τις συντομεύσεις και έννοιες που υπήρχαν ήδη στη βάση μας. Οι νέες συντομεύσεις και έννοιες που εντοπίσαμε για κάθε χρόνο, προστίθεντο στη βάση δεδομένων, προκειμένου να τις λάβουμε υπόψη κατά τη μέτρηση των νέων συντομεύσεων στα χρόνια που ακολούθησαν.

Όπως φαίνεται και στο Σχήμα 34, περισσότερες από 2000 νέες συντομεύσεις και 3500 νέες έννοιες πρώτο-χρησιμοποιήθηκαν το 2006, φτάνοντας τις 2800 και 6000 το 2014 αντίστοιχα, ενώ υπάρχει μία τάση οι αριθμοί αυτοί να αυξάνονται συνεχώς. Επίσης, ο αριθμός των νέων εννοιών αυξάνει ταχύτερα σε σύγκριση με τις νέες συντομεύσεις που εντοπίστηκαν, καθώς ορισμένες από τις νέες έννοιες έχουν προέλθει από υπάρχουσες συντομεύσεις.



Σχήμα 34: Ο αριθμός των νέων συντομεύσεων και εννοιών ανά χρόνο που χρησιμοποιήθηκαν στις κλινικές μελέτες που δημοσιεύτηκαν στο χρονικό διάστημα μεταξύ 2006 και 2015.

Σημειώνουμε ότι πριν το 2006, μόνο 6892 συντομεύσεις και 9628 έννοιες καταγράφηκαν, το οποίο είναι λογικό, καθώς μόνο 25 χιλιάδες κλινικές (το 12.80% των συνολικά διαθέσιμων κλινικών δοκιμών) μελέτες είχαν δημοσιευτεί μέχρι τότε. Επίσης, ο αριθμός των νέων συντομεύσεων και εννοιών που καταγράφηκαν το 2015 είναι ίσος με το 1/3 των νέων συντομεύσεων και εννοιών που καταγράφηκαν τον προηγούμενο χρόνο, το οποίο είναι φυσιολογικό, αν λάβουμε υπόψη ότι η ανάλυσή μας έλαβε χώρα με βάση τις μελέτες που ήταν διαθέσιμες μέχρι το τέλος του Μαρτίου του 2015.

Λαμβάνοντας υπόψη το γεγονός ότι α) παραπάνω από το 50% των νέων εννοιών (ειδικότερα μετά το 2010) «ανήκουν» σε υπάρχουσες συντομεύσεις καθώς επίσης και ότι β) το σύνολο των πιθανών συντομεύσεων που μπορούμε να σχηματίσουμε με έναν περιορισμένο αριθμό χαρακτήρων είναι πεπερασμένο, ο αριθμός των πιθανών εννοιών των συντομεύσεων αναμένεται να αυξηθεί τα επόμενα χρόνια. Ειδικότερα οι πιθανές σημασίες των συντομεύσεων που αποτελούνται από 2 μόνο χαρακτήρες αναμένεται να αυξηθεί σημαντικά, καθώς το 61.83% των συντομεύσεων που μπορούμε να

δημιουργήσουμε με 2 μόνο χαρακτήρες υπάρχουν ήδη, σε αντίθεση με τις συντομεύσεις που αποτελούνται από 3 μόνο χαρακτήρες όπου οι υπάρχουσες συντομεύσεις καλύπτουν μόνο το 15.74%.

29 *Συμπεράσματα*

Η εύρεση της σημασίας των συντομεύσεων που χρησιμοποιούνται στις κλινικές μελέτες είναι ένα απαραίτητο βήμα για την καλύτερη κατανόηση του περιεχομένου τους καθώς επίσης και τη σωστή ερμηνεία και εφαρμογή των αποτελεσμάτων τους. Στην ενότητα αυτή παρουσιάσαμε ένα σύστημα που αναπτύχθηκε για τον εντοπισμό της σημασίας των συντομεύσεων που χρησιμοποιούνται στις κλινικές μελέτες. Η ανάλυση των δεδομένων που συλλέχθηκαν έδειξε ότι υπάρχουν, κατά μέσο όρο, 6.8 συντομεύσεις ανά κλινική μελέτη, από τις οποίες δύο στις τρεις χρησιμοποιούνται χωρίς να αναφέρεται ρητά η πλήρης μορφή τους. Η ασάφεια των συντομεύσεων που χρησιμοποιούνται σε κλινικές μελέτες είναι πολύ μικρότερη της ασάφειας των βιοϊατρικών συντομεύσεων, γεγονός που διευκολύνει την εύρεση της σημασίας τους. Ωστόσο, σε αρκετές περιπτώσεις η εύρεση της σωστής σημασίας είναι δύσκολη και απαιτείται να λάβουμε υπόψη αρκετούς παράγοντες, όπως τη μορφή των συντομεύσεων, τις εκφράσεις στις οποίες μπορεί να συμμετέχουν, τις πιθανές σημασίες των συντομεύσεων, την ευρύτητα με την οποία χρησιμοποιούνται καθώς και το αντίστοιχο πεδίο γνώσης τόσο των πιθανών εννοιών όσο και του ευρύτερου πεδίου γνώσης μέσα στο όποιο χρησιμοποιείται η συντόμευση.

Στην εργασία αυτή εστιάσαμε στους αλγορίθμους και τις τεχνικές που μπορούμε να χρησιμοποιήσουμε, για να εντοπίσουμε τη σημασία των συντομεύσεων που χρησιμοποιούνται ήδη σε μία κλινική μελέτη και ειδικότερα αυτών που ορίζονται ρητά

μέσα στο κείμενο (συνοδεύονται από τη πλήρη έκφρασή τους). Όμως, ο χώρος των κλινικών μελετών είναι «ζωντανός», με αποτέλεσμα νέες έννοιες και συντομεύσεις να δημιουργούνται συνεχώς, οι οποίες μπορεί να συμπίπτουν με υπάρχουσες συντομεύσεις. Συνεπώς, οι πιθανές σημασίες των συντομεύσεων αυξάνονται συνεχώς δυσχεραίνοντας την ήδη δύσκολη διαδικασία επίλυσης της σημασίας τους. Για το σκοπό αυτό, η ανάπτυξη εργαλείων και υπηρεσιών που διευκολύνουν τη δημιουργία της συντετμημένης μορφής μιας φράσης, ενημερώνοντας τους χρήστες για τις πιθανές σημασίες τους σε ένα πεδίο γνώσης αλλά και προτείνοντας πιθανές συντομεύσεις αυτής, είναι απαραίτητη. Ειδικότερα, η δημιουργία μιας «καλής» συντόμευσης μιας φράσης είναι ένα ιδιαίτερα ενδιαφέρον θέμα.

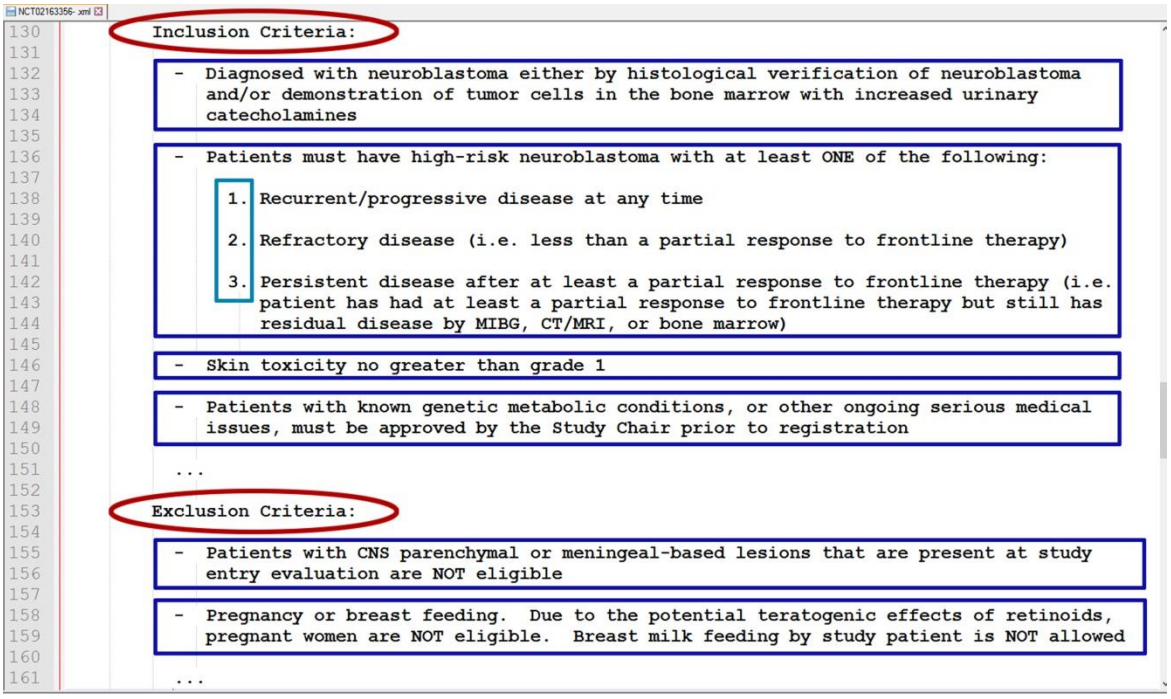
30 *Παράρτημα*

30.1Εξαγωγή Κριτηρίων Καταλληλότητας

Η προσεκτική ανάλυση ορισμένων τυχαίως επιλεγμένων ΚΚ έδειξε ότι, σε αρκετές περιπτώσεις, τα κριτήρια εισαγωγής είναι ξεχωριστά από τα κριτήρια εξαγωγής/απόρριψης, χρησιμοποιώντας συγκεκριμένες λέξεις ή φράσεις. Επίσης, στη συντριπτική τους πλειοψηφία, τα κριτήρια διαχωρίζονται μεταξύ τους με μία κενή γραμμή, ενώ η απόστασή τους από την αρχή της γραμμής (το σύνολο των κενών χαρακτήρων που προηγούνται) είναι συχνά η ίδια. Μερικές όμως φορές, η απόσταση αυτή είναι αισθητά μεγαλύτερη και χρησιμοποιείται, για να δείξει ότι οι παράμετροι που ακολουθούν εντάσσονται στο προηγούμενο κριτήριο. Η ανάλυσή μας επίσης έδειξε ότι, σε αρκετές περιπτώσεις, ο χαρακτήρας της παύλας (dash) προηγείται του ορισμού ενός κριτηρίου, ενώ, σε ορισμένες περιπτώσεις, Ρωμαϊκοί ή Αραβικοί αριθμοί μπορεί να χρησιμοποιηθούν αντ' αυτού. Τέλος, παρατηρήσαμε ότι, σε λίγες περιπτώσεις, τα κριτήρια έχουν τοποθετηθεί το ένα κάτω από το άλλο, χωρίς να προηγείται κάποιο σημείο στίξης ή αριθμός.

Λαμβάνοντας υπόψη τις παραπάνω παραμέτρους, αναπτύχθηκε ένας μηχανισμός που εξάγει τα κριτήρια που έχουν οριστεί σε μια κλινική δοκιμή (δηλαδή, παρέχει μια λίστα από κριτήρια εισόδου και μια άλλη λίστα με τα κριτήρια εξόδου), λαμβάνοντας υπόψη τις λέξεις και φράσεις που χρησιμοποιούνται, για να δείξουν ότι ακολουθεί μια

λίστα από κριτήρια εισόδου ή εξόδου καθώς και τη μορφή που αυτά έχουν. Για το σκοπό αυτό, αρχικά, εντοπίζουμε τις συμβολοακολουθίες (strings) που διαχωρίζονται μεταξύ τους με μία κενή γραμμή καθώς και τη θέση – απόσταση από την αρχή του πρώτου μη κενού χαρακτήρα. Έπειτα, οργανώνουμε τις σύμβολο-ακολουθίες με βάση την απόστασή τους από την αρχή της γραμμής, φτιάχνοντας ένα δένδρο. Θα πρέπει να σημειωθεί ότι όλα τα στοιχεία που ανήκουν στο πρώτο επίπεδο θα πρέπει να έχουν την ίδια μορφή (π.χ. ξεκινούν με το χαρακτήρα της παύλας). Σχετικά με τις επιπλέον παραμέτρους/δεδομένα ενός κριτηρίου (αν αυτά υπάρχουν) θα πρέπει να ακολουθούν μια συγκεκριμένη μορφή (π.χ., χρησιμοποιώντας διαδοχικούς αραβικούς αριθμούς), η οποία μπορεί να είναι διαφορετική από τη μορφή που χρησιμοποιήθηκε στο προηγούμενο επίπεδο.



Σχήμα 35: Υποσύνολο των Κριτηρίων Καταλληλότητας της Κλινικής Δοκιμής με κωδικό "NCT02163356".

Το Σχήμα 35 παρουσιάζει ένα υποσύνολο των ΚΚ που έχουν οριστεί για την κλινική δοκιμή με αναγνωριστικό «NCT02163356». Σε αυτό το παράδειγμα θα πρέπει να σημειώσουμε ότι οι φράσεις «Inclusion Criteria» και «Exclusion Criteria» χρησιμοποιούνται, για να δείξουν ότι μια λίστα από κριτήρια εισαγωγής και εξαγωγής

ακολουθούν αντίστοιχα. Επίσης, και στις δύο περιπτώσεις ο χαρακτήρας της παύλας προηγείται του ορισμού των κριτηρίων. Επιπλέον, το δεύτερο κριτήριο είναι αρκετά περίπλοκο και ο ορισμός του καλύπτει αρκετές γραμμές (παρά το γεγονός ότι αυτές διαχωρίζονται μεταξύ τους με μία κενή γραμμή). Στο κριτήριο αυτό, οι παράμετροι που αναγράφονται ξεκινούν με διαδοχικούς Αραβικούς αριθμούς/σύμβολα. Επομένως, στην κλινική αυτή δοκιμή (λαμβάνοντας υπόψη τα δεδομένα που υπάρχουν στην εικόνα) συνολικά εξάγαμε 4 κριτήρια εισαγωγής και 2 εξαγωγής.

Τα κριτήρια που εξαγάγαμε από καθεμιά από τις διαθέσιμες κλινικές δοκιμές τοποθετήθηκαν σε μια σχεσιακή βάση δεδομένων, καταγράφοντας τον κωδικό της κλινικής δοκιμής στην οποία ανήκει το κάθε κριτήριο, τον τύπο του κριτηρίου (πχ. κριτήριο εισαγωγής ή εξαγωγής), τη μορφή του (π.χ. το κριτήριο ξεκινούσε με το χαρακτήρα της παύλας), τη γλώσσα στην οποία είναι αυτό εκφρασμένο (π.χ., χρησιμοποιώντας τη φυσική γλώσσα) καθώς και το κριτήριο αυτό καθ' αυτό.

Ακολουθώντας την παραπάνω διαδικασία εξάγαμε τα ΚΚ από το 90% των κλινικών δοκιμών που είναι διαθέσιμες, δημιουργώντας μία βάση δεδομένων με 1.9 εκατομμύρια κριτήρια εκφρασμένα με βάση τη φυσική γλώσσα. Η πλειοψηφία των κριτηρίων ξεκινά με το χαρακτήρα της παύλας, ενώ μόνο 5% των κριτηρίων καταλαμβάνουν παραπάνω από μία γραμμή. Κατά την επεξεργασία των ΚΚ δε λάβαμε υπόψη τις περιπτώσεις στις οποίες όλα τα κριτήρια που έχουν ορισθεί βρίσκονται σε μια παράγραφο, καθώς ήταν αρκετά δύσκολος ο εντοπισμός της αρχής και του τέλους του καθενός, ειδικά όταν τα τελευταία είναι εκφρασμένα σε δύο η περισσότερες προτάσεις. Επίσης εστιάσαμε στις κλινικές δοκιμές στις οποίες τα κριτήρια αποτελούνται από δύο κατηγορίες (κριτήρια εισαγωγής και κριτήρια εξαγωγής), τα οποία θα πρέπει να ικανοποιούνται από όλους τους συμμετέχοντες, αγνοώντας τις κλινικές δοκιμές στις

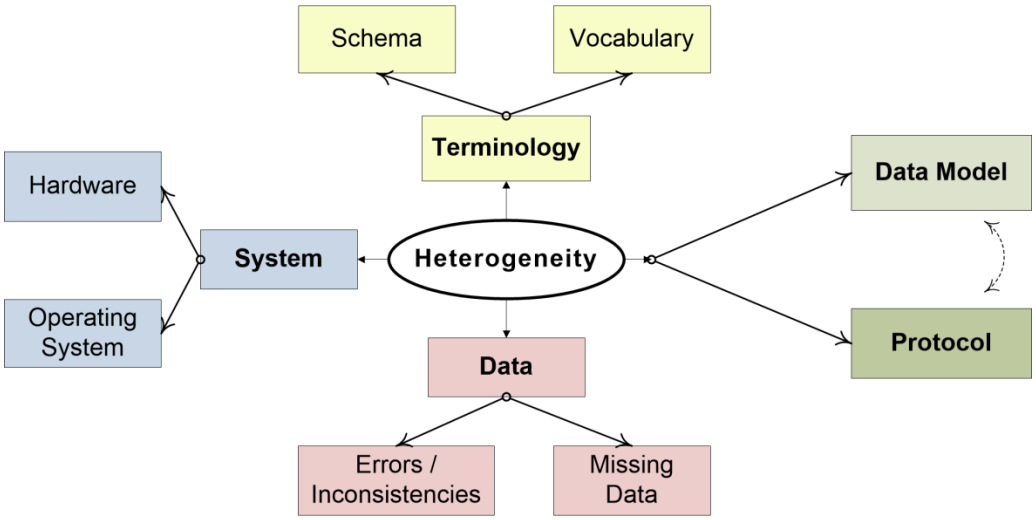
οποίες ένα υποσύνολο του πληθυσμού (π.χ., πληθυσμός ελέγχου) θα πρέπει να ικανοποιεί διαφορετικά κριτήρια.

Σημειώνουμε ότι, εκτός από τα παραπάνω κριτήρια που είναι εκφρασμένα με βάση τη φυσική γλώσσα, χωρίς να υπάρχει σαφής διαχωρισμός μεταξύ τους (στο XML αρχείο), σε καθεμιά από τις κλινικές δοκιμές, ένας πολύ περιορισμένος αριθμός κριτηρίων ορίστηκε ξεχωριστά από τα υπόλοιπα (δομημένη πληροφορία). Τα κριτήρια αυτά καλύπτουν ένα συγκεκριμένο τύπο κριτηρίων (πχ., ηλικία, φύλο), τα δεδομένα των οποίων εξήχθησαν από το XML αρχείο της κλινικής δοκιμής και κατόπιν εισήχθησαν στη βάση δεδομένων με τη μορφή του JSON, για λόγους πληρότητας. Αξίζει να σημειωθεί ότι τα παραπάνω κριτήρια μπορεί να συμπεριλαμβάνονται στο σύνολο των ΚΚ, που ορίσθηκαν με βάση τη φυσική γλώσσα.

30.2Ετερογένεια Πηγών Δεδομένων

Οι βάσεις δεδομένων, ακόμη και αν καλύπτουν το ίδιο πεδίο γνώσης (π.χ. καταγραφή δεδομένων ασθενών) έχουν, συχνά, αρκετές διαφορές μεταξύ τους (ετερογένεια πηγών δεδομένων – datasource heterogeneity), ως αποτέλεσμα του ανεξάρτητου σχεδιασμού τους καθώς επίσης και της περιορισμένης χρήσης διεθνών προτύπων οργάνωσης και αναπαράστασης της πληροφορίας. Σύμφωνα με τον Raji GHAWI [173] τις διαφορές που υπάρχουν μπορούμε να τις εντάξουμε σε τέσσερις κατηγορίες, όπως φαίνεται στο Σχήμα 36.

Η συστημική ετερογένεια (system heterogeneity) έχει να κάνει με τις διαφορές που υπάρχουν στο υλικό (hardware) και το λογισμικό (software) των υπολογιστών, στους οποίους βρίσκονται οι βάσεις δεδομένων. Για παράδειγμα, οι επεξεργαστές μπορεί να διαφέρουν, με άμεσες επιπτώσεις στην απόδοση των συστημάτων και των αντίστοιχων πηγών δεδομένων. Επίσης, το λειτουργικό τους σύστημα μπορεί να είναι διαφορετικό.



Σχήμα 36: Ετερογένεια Πηγών Δεδομένων

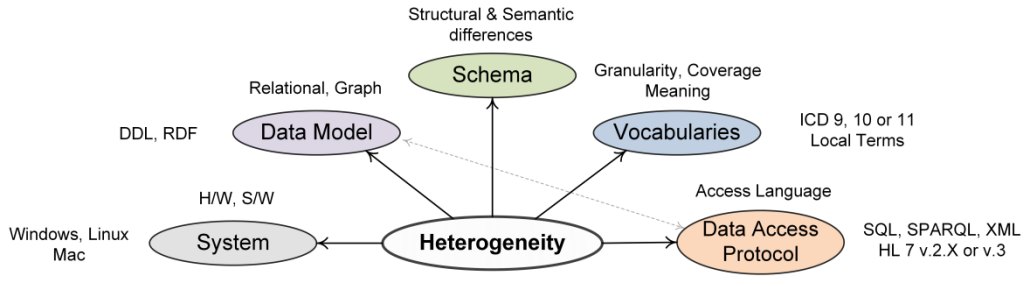
Η συντακτική ετερογένεια (syntactic heterogeneity) σχετίζεται με τα διαφορετικά μοντέλα δεδομένων (data models), τα οποία μπορούν να χρησιμοποιηθούν για την αναπαράσταση της γνώσης μας. Ένα ευρέως διαδεδομένο μοντέλο αναπαράστασης είναι το σχεσιακό μοντέλο (relational model), ωστόσο υπάρχουν και άλλα διαθέσιμα μοντέλα, όπως το αντικειμενοστραφές μοντέλο (object oriented model) και το ιεραρχικό μοντέλο (hierarchical model), καθένα απ' τα οποία έχει τα πλεονεκτήματα και τα μειονεκτήματά του. Ανάλογα με το μοντέλο δεδομένων που έχει επιλεχθεί για την αναπαράσταση των δεδομένων μας, μπορούμε να χρησιμοποιήσουμε και την αντίστοιχη γλώσσα / πρωτόκολλο, για να έχουμε πρόσβαση σε αυτά. Πιο συγκεκριμένα, στην περίπτωση μιας σχεσιακής βάσης μπορούμε να έχουμε πρόσβαση (π.χ., αναζήτηση των επιθυμητών δεδομένων) στα δεδομένα της βάσης, χρησιμοποιώντας SQL, ενώ στην περίπτωση μιας αντικειμενοστραφούς βάσης, χρησιμοποιώντας OQL ερωτήματα.

Η δομική ετερογένεια (structural heterogeneity) προκύπτει από τους διαφορετικούς τρόπους οργάνωσης της γνώσης μας. Μπορούμε να αναπαραστήσουμε την ίδια πληροφορία με αρκετούς διαφορετικούς τρόπους, ακόμα και αν χρησιμοποιούμε το ίδιο μοντέλο δεδομένων. Οι δομικές διαφορές μεταξύ δύο πηγών δεδομένων εντοπίζονται στον τρόπο αναπαράστασης της πληροφορίας (representation heterogeneity) καθώς και

στον τρόπο οργάνωσης της πληροφορίας (schema heterogeneity). Για παράδειγμα, σε μία βάση, η πληροφορία για το όνομα ενός ανθρώπου μπορεί να βρίσκεται μέσα σε ένα μόνο πεδίο (ονοματεπώνυμο), ενώ, σε μία άλλη βάση, η ίδια πληροφορία μπορεί να βρίσκεται μέσα σε δύο πεδία (όνομα και επώνυμο) ή και περισσότερα (π.χ. μεσαίο όνομα - προαιρετικό).

Η σημασιολογική ετερογένεια (semantic heterogeneity) σχετίζεται με την ερμηνεία των στοιχείων μιας βάσης δεδομένων. Ένα παράδειγμα σημασιολογικής ετερογένειας είναι η χρησιμοποίηση διαφορετικών όρων/εννοιών για την αναπαράσταση της ίδιας έννοιας (synonymy). Ένα άλλο παράδειγμα είναι η χρησιμοποίηση των ίδιων όρων για την αναπαράσταση διαφορετικών εννοιών (homonymy). Επιπρόσθετα, τα δεδομένα που υπάρχουν στις βάσεις μπορεί να έχουν ασυνέπειες μεταξύ τους, ενώ οι παράμετροι που καταγράφουν για τις οντότητες ενός οργανισμού μπορεί να διαφέρουν.

Στην εργασία αυτή, εξαιτίας της βασισμένης σε μοντέλο αναπαράστασης των δεδομένων των οντοτήτων που καταγράφονται από έναν οργανισμό (π.χ., μονάδα περίθαλψης) καθώς και της χρησιμοποίησης υπαρχόντων συστημάτων κωδικοποίησης και επικοινωνίας, μπορούμε να οργανώσουμε τις διαφορές που υπάρχουν μεταξύ των βάσεων δεδομένων λίγο διαφορετικά (Σχήμα 37).



Σχήμα 37: Μια διαφορετική αναπαράσταση της Ετερογένειας των Πηγών Δεδομένων

Στο Σχήμα 37 υπάρχει σαφής διαχωρισμός των διαφορών που προκύπτουν από το μοντέλο που χρησιμοποιείται για την αναπαράσταση των δεδομένων (data model) από το πρωτόκολλο που μπορεί να χρησιμοποιηθεί για την πρόσβαση σε αυτά (data access

protocol). Γενικά, το πρωτόκολλο επικοινωνίας εξαρτάται από το μοντέλο αναπαράστασης που έχει χρησιμοποιηθεί. Ωστόσο, ορισμένες φορές, η χρησιμοποίηση του αντίστοιχου πρωτοκόλλου μπορεί να μην επιτρέπεται. Για παράδειγμα, τα δεδομένα ενός ασθενούς μπορεί να αποθηκεύονται σε μία σχεσιακή βάση, ωστόσο, η πρόσβαση του χρήστη στα δεδομένα της βάσης αυτής να μη γίνεται μέσω SQL ερωτημάτων αλλά μέσω της χρησιμοποίησης συγκεκριμένων HL7 v.2 μηνυμάτων. Επίσης, τις δομικές και σημασιολογικές διαφορές που υπάρχουν μεταξύ δύο πηγών δεδομένων, στο σχήμα αυτό, τις χωρίσαμε σε δύο διαφορετικές κατηγορίες. Η πρώτη κατηγορία περιλαμβάνει τις διαφορές που υπάρχουν στο σχήμα/μοντέλο που χρησιμοποιείται για την αναπαράσταση των δεδομένων ενός οργανισμού, ενώ η δεύτερη κατηγορία εστιάζει στις διαφορές που υπάρχουν στα συστήματα κωδικοποίησης. Στην πρώτη περίπτωση υπάρχουν συνήθως σημαντικές δομικές και σημασιολογικές διαφορές μεταξύ των στοιχείων δύο διαφορετικών μοντέλων, εξαιτίας του διαφορετικού σκοπού που εξυπηρετούν καθώς και των υποθέσεων που έγιναν και των αποφάσεων που πάρθηκαν κατά το σχεδιασμό τους. Στην δεύτερη περίπτωση οι διαφορές τους εστιάζονται στο πεδίο της γνώσης που καλύπτουν (domain coverage), στη διακριτικότητα των όρων που περιέχουν (granularity of terms) καθώς και τη σημασία τους.

Στο σημείο αυτό σημειώνουμε ότι οι δομικές και σημασιολογικές διαφορές (ή αντίστοιχα, οι διαφορές στο μοντέλο και στα συστήματα κωδικοποίησης) που υπάρχουν μεταξύ δύο ή περισσοτέρων διαφορετικών βάσεων δεδομένων, οι οποίες καλύπτουν ένα συγκεκριμένο πεδίο γνώσης, εξακολουθούν να υπάρχουν και κατά την οντολογική αναπαράσταση των βάσεων δεδομένων (ontology mismatches), ειδικά, όταν η οντολογική αναπαράσταση είναι στενά συνδεδεμένη με το σχήμα της βάσης και τις κωδικοποιήσεις που υποστηρίζονται από αυτή. Επιπρόσθετα, για την πρόσβαση στα δεδομένα των βάσεων αυτών μέσω μίας κοινής ονοματολογίας (συμπεριλαμβανομένων των όρων του

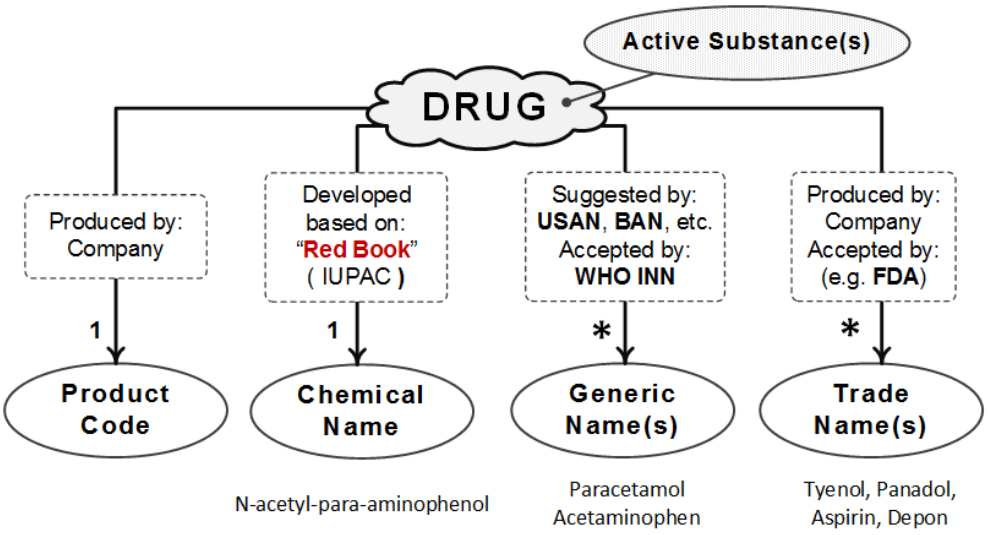
μοντέλου και των συστημάτων κωδικοποίησης), είναι απαραίτητος ο καθορισμός των συσχετίσεων που υπάρχουν μεταξύ των όρων του «κοινού» / «καθολικού» μοντέλου και των συστημάτων κωδικοποίησης που επιλέχθηκαν με τα αντίστοιχα μοντέλα και κωδικοποιήσεις που υποστηρίζονται από την καθεμιά από τις βάσεις δεδομένων. Κατά την παραπάνω διαδικασία, για να γεφυρώσουμε το χάσμα που υπάρχει μεταξύ του χρήστη και των βάσεων, πρέπει να χειριστούμε μια πληθώρα από αναντιστοιχίες που υπάρχουν μεταξύ της νέας «εικονικής» ή «φυσικής» βάσης και της πραγματικής βάσης, στην οποία βρίσκονται τα δεδομένα. Τα προβλήματα που καλούμαστε να αντιμετωπίσουμε είναι παρόμοια με αυτά που ήδη αναφέρθηκαν, με τα σημαντικότερα να εστιάζονται στις δομικές και σημασιολογικές διαφορές που υπάρχουν (τουλάχιστον με μία ή παραπάνω βάσεις), εξαιτίας της χρήσης διαφορετικών μοντέλων αναφοράς και κωδικοποιήσεων, καθώς και στον τρόπο πρόσβασης στα δεδομένα των βάσεων, εξαιτίας της υποστήριξης διαφορετικών πρωτοκόλλων επικοινωνίας από αυτές.

30.3Ονοματολογία Φαρμάκων και Ασθενειών

Ένα φάρμακο έχει αρκετά ονόματα, καθένα απ' τα οποία υπόκειται σε διαφορετικούς κανόνες (Σχήμα 38). Το χημικό όνομα (chemical name) ενός φαρμάκου βασίζεται στη χημική δομή των συστατικών του και δημιουργείται, όταν αναπτύσσεται μια καινοτόμος χημική ή βιολογική οντότητα. Το όνομα αυτό ακολουθεί τους κανόνες ονοματολογίας μη οργανικών χημικών συστατικών, που έχουν δημοσιευθεί από τη Διεθνή Ένωση Καθαρής και Εφαρμοσμένης Χημείας (IUPAC) [174] και αποτελείται από ένα μεγάλο αριθμό χαρακτήρων, που είναι δύσκολο τόσο να τους προφέρουμε όσο και να τους θυμόμαστε. Συνεπώς, δε χρησιμοποιείται συχνά, για να αναφερθούμε σε ένα φάρμακο που είναι ήδη στην αγορά.

Όταν ένα φάρμακο είναι υπό επεξεργασία, η εταιρία που το αναπτύσσει έχει δημιουργήσει έναν κωδικό φαρμάκου έτσι, ώστε οι ερευνητές να μπορούν να

αναφέρονται εύκολα σε αυτό. Ο κωδικός αυτός ακολουθεί συγκεκριμένους κανόνες που ορίζονται από την εταιρία (π.χ., ακρώνυμο της εταιρίας ακολουθούμενο από έναν αριθμό) και συχνά δε δίνει κάποια πληροφορία για το φάρμακο και τη δράση του. Παρόλα αυτά, σε ορισμένες περιπτώσεις, ο κωδικός αυτός εξακολουθεί να χρησιμοποιείται ακόμη και όταν το φάρμακο καταφέρει να μπει στην αγορά.



Σχήμα 38: Ονοματολογία Φαρμάκων

Όταν ένα νέο χημικό προϊόν γίνει αποδεκτό από τη σχετική αρχή (π.χ., η υπηρεσία διαχείρισης τροφίμων και φαρμάκων (FDA) [175] για την Αμερική), παράγεται μία γενική ονομασία (generic name) για το προϊόν αυτό, γνωστή επίσης και ως Διεθνής Μη-Ιδιοκτησίας Ονομασία (INN). Για το σκοπό αυτό, υπάρχουν ορισμένοι ιατρικοί οργανισμοί, όπως το Συμβούλιο Έκδοσης Ονομάτων των Ηνωμένων Πολιτειών (USAN) και το Συμβούλιο Έκδοσης Ονομάτων της Μεγάλης Βρετανίας (BAN), ο ρόλος των οποίων είναι να προτείνουν συντομεύσεις των χημικών ονομάτων των φαρμάκων, τα οποία να μπορούμε να τα θυμόμαστε εύκολα και ταυτόχρονα να είναι γεμάτα σημασία για τους ανθρώπους (τουλάχιστον τους γιατρούς) μιας χώρας. Τα ονόματα ανά χώρα που προκύπτουν για το ίδιο φάρμακο, στη συντριπτική πλειοψηφία των φαρμάκων, είναι ταυτόσημα και οφείλονται σε μία συνεχή συνεργασία μεταξύ των παραπάνω συμβουλίων.

Παρόλα αυτά, εξαιτίας ορισμένων αποκλίσεων στα συστήματα ονοματολογίας τους, είναι πιθανό η γενική ονομασία ενός φαρμάκου να είναι διαφορετική (π.χ., η δραστική ουσία της ασπιρίνης είναι γενικά γνωστή ως "acetaminophen" και "paracetamol"). Η γενική ονομασία του φαρμάκου θα πρέπει να γίνει τελικά αποδεκτή από τη Διεθνή, Μη-ιδιοκτησίας Ονομασία (INN) αρχή του Διεθνούς Οργανισμού Υγείας (WHO).

Η εμπορική ονομασία (trade name) ενός φαρμάκου – γνωστή επίσης και ως εμπορικό σήμα (trademark), όνομα μάρκας (brand name) ή όνομα ιδιοκτησίας (proprietary name) – αναπτύσσεται από την εταιρία (συχνά με τη βοήθεια ενός συμβόλου για την ονομασία φαρμάκων) που φέρνει ένα νέο χημικό προϊόν στην αγορά. Το όνομα αυτό θα πρέπει να είναι μοναδικό (τουλάχιστον σε μια χώρα) και θα πρέπει να δίνει άμεσα ή έμμεσα πληροφορία για το τι κάνει το φάρμακο αυτό. Από τη μεριά της εταιρίας, το όνομα που θα επιλεχθεί θα πρέπει να ενισχύσει τις πωλήσεις του φαρμάκου. Επομένως, θα πρέπει να είναι εύκολο στη μνήμη, διακριτό από τα υπόλοιπα φάρμακα και να δημιουργεί μια αίσθηση της υπηρεσίας (ή ένα μέρος αυτής) που προσφέρει. Με απλά λόγια, η εμπορική ονομασία του φαρμάκου θα πρέπει να είναι «πιασάρικη». Από την οπτική γωνία του οργανισμού τροφίμων και φαρμάκων (FDA), το όνομα θα πρέπει να εστιάζει στην ασφάλεια των ασθενών, αποφεύγοντας παρανοήσεις του φαρμάκου και/ή λανθασμένη χρήση του. Συνεπώς, ο παραπάνω οργανισμός απορρίπτει εμπορικά ονόματα που μοιάζουν πολύ με άλλα φάρμακα (το 1/3 των ονομάτων απορρίπτεται για το λόγο αυτό). Επιπλέον, ο FDA απορρίπτει τα εμπορικά ονόματα που υπονοούν ότι το φάρμακο μπορεί να κάνει κάτι για το οποίο δεν έχει γίνει αποδεκτό ή υπόσχεται παραπάνω απ' ό,τι μπορεί πραγματικά να προσφέρει. Εάν ένα φάρμακο προστατεύεται από την πατέντα ευρεσιτεχνίας (patent license), το φάρμακο είναι γνωστό στην αγορά με ένα μόνο εμπορικό όνομα. Όμως, όταν η παραπάνω άδεια λήξει, το ίδιο φάρμακο μπορεί να είναι

διαθέσιμο στην αγορά με αρκετά διαφορετικά ονόματα. Κατά μέσο όρο, για κάθε φάρμακο υπάρχουν 4 διαφορετικά εμπορικά ονόματα.

Όλα τα παραπάνω ονόματα, συμπεριλαμβανομένων των κωδικών, μπορούν να χρησιμοποιηθούν, για να αναφερθούμε στην ίδια οντότητα (δηλαδή, ενεργή ουσία) που είναι υπεύθυνη για την φαρμακευτική δράση του φαρμάκου. Θα πρέπει να σημειωθεί ότι τα μη ενεργά συστατικά ενός φαρμάκου, τα οποία προσθέτονται κατά τη διαδικασία βιομηχανικής παραγωγής του και είναι υπεύθυνα, για παράδειγμα, για την γεύση του, ποικίλλουν ανά προϊόν και γενικά δεν έχουν κάποια κλινική επίπτωση.

Οι Ιατρικές Καταστάσεις (π.χ., ασθένειες) μπορεί να έχουν παραπάνω από ένα ονόματα. Μερικές ιατρικές καταστάσεις είναι ευρέως γνωστές με βάση την επωνυμία τους. Τα επώνυμα χρησιμοποιούνται συνήθως, για να αναφερθούν στους ερευνητές που περιέγραψαν πρώτοι μία κατάσταση (π.χ., Πάρκινσον) σε κάποιο αξιοσέβαστο ιατρικό περιοδικό, προς τιμήν τους για τη συνεισφορά τους στο συγκεκριμένο πεδίο γνώσης. Μπορεί επίσης να αναφέρονται στον ασθενή που έπασχε από την ιατρική κατάσταση (π.χ., νόσος του Lou Gehrig), στο μέρος που παρατηρήθηκε αρχικά (π.χ., νόσος του Lyme) ή ακόμη και στην κοινωνία ή ομάδα ανθρώπων που πρώτο-εμφανίστηκε (π.χ., νόσος των Λεγεωνάριων). Όμως, η ονομασία των ιατρικών καταστάσεων με τον τρόπο αυτό δεν παρέχει αρκετή πληροφορία για την εν λόγω ιατρική κατάσταση. Συνεπώς, το «πρόβλημα» στο οποίο αναφέρονται δεν είναι εύκολα αντιληπτό, ειδικότερα όταν πρόκειται για ασθένειες που σπάνια συναντάμε.

Τα περιγραφικά ονόματα που παράγονται παρέχουν πολύ περισσότερες πληροφορίες για την εκάστοτε κατάσταση. Πιο συγκεκριμένα, περιγράφουν τις συνθήκες που παρατηρούμε (πρωτεύοντα συμπτώματα ή εργαστηριακά ευρήματα) έτσι, ώστε οι ειδικοί στον τομέα της ιατρικής να έχουν κάποιες ενδείξεις για το τι καλούνται να αντιμετωπίσουν. Για παράδειγμα, η Αμυοτροφική Πλευρική Σκλήρυνση (νόσος του Lou

Gehrig) υποδεικνύει ότι προκαλεί μυϊκή αδυναμία και ατροφία (α-μυο-τροφική) σε όλο το σώμα, λόγω εκφυλισμού των άνω και κάτω κινητικών νευρώνων (πλευρική σκλήρυνση). Τα περιγραφικά ονόματα μπορεί να είναι διαφορετικά ανά χώρα. Για παράδειγμα, η νόσος του Lou Gehrig είναι γνωστή ως «Νόσος του Κινητικού Νευρώνα» στο Ηνωμένο Βασίλειο και την Αυστραλία.

30.4Συντομεύσεις Επιρρεπείς σε Λάθη

Το γεγονός ότι ένας σημαντικός αριθμός συντομεύσεων χρησιμοποιείται στις κλινικές μελέτες, χωρίς να αναφέρεται η πλήρης έκφρασή τους, μπορεί να καταστήσει τη διαδικασία εντοπισμού της σημασίας των συντομεύσεων εξαιρετικά δύσκολη, ειδικότερα όταν υπάρχουν παραπάνω από μία πιθανές σημασίες της μέσα σε ένα συγκεκριμένο πεδίο γνώσης. Τα συμφραζόμενα μπορεί να αποκαλύψουν την πραγματική σημασία των συντομεύσεων αυτών. Ωστόσο, η ερμηνεία τους εξαρτάται από το υπόβαθρο του κάθε ανθρώπου και συνεπώς, υπάρχει πάντα η πιθανότητα μια συντόμευση να αγνοηθεί (άγνωστες συντομογραφίες) ή να ερμηνευθεί λάθος (ασαφείς συντομογραφίες), με άμεση επίπτωση στην ασφάλεια των ασθενών.

Για να μετριάσει τα παραπάνω προβλήματα, η Joint Commission [176] ανέπτυξε το 2004 μία λίστα από συντομεύσεις που μπορεί να οδηγήσουν σε λάθη. Η λίστα αυτή [177] ορίζει τις συντομεύσεις που θα πρέπει να αποφεύγουμε να χρησιμοποιούμε σε χειρόγραφα κείμενα που σχετίζονται με τη φαρμακευτική αγωγή των ασθενών. Για παράδειγμα, δε θα πρέπει να χρησιμοποιούμε τη λατινική συντομογραφία «q.d» (μία φορά την ημέρα) κατά τον καθορισμό της συχνότητας λήψης ενός φαρμάκου, καθώς μπορεί να μπερδευτεί με την λατινική συντομογραφία «q.i.d» (τέσσερις φορές την ημέρα). Επίσης, η χρησιμοποίηση της συντόμευσης IU (International Unit) θα πρέπει να αποφεύγεται, καθώς υπάρχει κίνδυνος να την μπερδέψουμε με τη συντομογραφία IV (Intravenous). Επιπρόσθετα, συντομεύσεις που μοιάζουν πολύ μεταξύ τους θα πρέπει να αποφεύγονται,

όπως οι συντομεύσεις «MSO4» (morphine sulfate) και «MgSO4» (magnesium sulfate). Οι παραπάνω οδηγίες έχουν να κάνουν κυρίως με χειρόγραφα έγγραφα. Ωστόσο, θα πρέπει να αποφεύγουμε τη χρήση των παραπάνω συντομεύσεων ακόμη και στα έντυπα, καθώς οι παραπάνω συντομεύσεις εξακολουθούν να μπερδεύουν, ενώ δημιουργούν εσφαλμένα την εντύπωση ότι οι συντομεύσεις αυτές είναι αποδεκτές. Παρόλα αυτά, εξακολουθούν να χρησιμοποιούνται στα έγγραφα, αλλά, όπως έδειξε η έρευνα που πραγματοποιήθηκε από την Joint Commission, με φθίνουσα τάση [178].

Το ινστιτούτο Ασφαλών Φαρμακευτικών Πρακτικών (ISMP) [179] ανέπτυξε, επίσης, μία λίστα από συντομεύσεις που είναι επιρρεπείς σε λάθη. Η λίστα αυτή [180] αποτελείται από συντομεύσεις, σύμβολα και ονομασίες φαρμάκων τα οποία δε θα πρέπει να χρησιμοποιούνται κατά την ανταλλαγή πληροφορίας σχετικής με τη φαρμακευτική αγωγή των ασθενών. Ωστόσο, αρκετές από τις παραπάνω συντομεύσεις χρησιμοποιούνται στις κλινικές μελέτες αλλά σε πολύ μικρά ποσοστά.

Η σελίδα αυτή είναι σκόπιμα λευκή

31 *Συντομογραφίες*

Στον παρακάτω πίνακα παραθέτουμε τους όρους και συντομογραφίες που χρησιμοποιήθηκαν στη διατριβή:

Όρος	**Ερμηνεία**
aka	also known as
API	Application Programming Interface
ASPIRE	Agreement on Standardized Protocol Inclusion Requirements for Eligibility
ATC	Anatomical Therapeutic Chemical Classification
BRIDG	Biomedical Research Integrated Domain Group
BAN	British Approved Name
ChEBI	Chemical Entities of Biological Interest
CDISC	Clinical Data Interchange Standards Consortium
CORS	Cross-Origin Resource Sharing
DB	Database
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, 4th Edition
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
EHR	Electronic Health Record
EDOAL	Expressive and Declarative Ontology Alignment Language
EliXR	Eligibility criteria extraction and representation
XML	EXtensible Markup Language
ERGO	Eligibility Rule Grammar and Ontology
FDA	Food and Drug Administration
GLAV	Global and Local As View
GAV	Global as View
HL7	Health Level Seven International
HTML	HyperText Markup Language
IT	Information technology
ISMP	Institute for Safe Medication Practices
ICD	International Classification of Diseases
ICD-9	International Classification of Diseases 9th Revision
ICD-10	International Classification of Diseases 10th Revision
ICD-10-CM	International Classification of Diseases, 10th Revision, Clinical Modification

INN	International Nonproprietary Name
IUPAC	International Union of Pure and Applied Chemistry
JSON	JavaScript Object Notation
LAV	Local as View
LOINC	Logical Observation Identifiers Names and Codes
MeDRA	Medical Dictionary for Regulatory Activities
MESH	Medical Subject Headings
MLM	Medical Logic Module
OAT	Ontology Alignment Tool
OBA	Open Biomedical Annotator
OCL	Object Constraint Language
ODM	Operational Data Model
PRM	Protocol Representation Model
RDFS	RDF Schema
RIM	Reference Information Model
RDF	Resource Description Framework
SPARQL	SPARQL Query Language for RDF
SQL	Structured Query Language
SDTM	Study Data Tabulation Model
SDM	Study/Trial Design Model
SNOMED-CT	Systematized Nomenclature of Medicine - Clinical Terms
UML	Unified Modeling Language
URI	Uniform Resource Identifier
USAN	United States Adopted Names
OWL	Web Ontology Language
WHO	World Health Organization

32 *Βιβλιογραφικές Αναφορές*

[1] DiMasi, J.A., Hansen, R.W. and Grabowski, H.G. (2003) 'The price of innovation: new estimates of drug development costs', Journal of Health Economics, Vol. 22 No. 2, pp. 151-185.

[2] Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C. and Greyson, D. (2011) 'The cost of drug development: A systematic review', Health Policy, Vol. 100 No. 1, pp. 4-17.

[3] Dickson, M. and Gagnon, J.P. (2004) 'The cost of new drug discovery and development', Discovery Medicine, Vol. 4 No. 22, pp. 172-9.

[4] Kola, I. and Landis, J. (2004) 'Can the pharmaceutical industry reduce attrition rates?', Nature Reviews Drug Discovery, Vol. 3 No. 8, pp. 711-5.

[5] Rothwell, P.M. (2005) 'External validity of randomised controlled trials: to whom do the results of this trial apply?', Lancet, Vol. 365 No. 9453, pp. 82-93.

[6] Van Spall, H.G., Toren, A., Kiss, A. and Fowler, R.A. (2007) 'Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review', Journal of the American Medical Association, Vol. 297 No. 11, pp. 1233-40.

[7] Clinical Data Interchange Standards Consortium, available at http://www.cdisc.org

[8] Health Level Seven International, available at http://www.hl7.org

[9] CDISC Operational Data Model, available at http://www.cdisc.org/odm

[10] CDISC Study Data Tabulation Model, available at http://www.cdisc.org/sdtm

[11] Schadow, G., Mead, C.N. and Walker, D.M. (2006) 'The HL7 reference information model under scrutiny', Studies in Health Technology and Informatics, Vol. 124, pp. 151-6.

[12] Haber, M.W., Kisler, B.W., Lenzen, M. and Wright, L.W. (2007) 'Controlled terminology for clinical research: a collaboration between CDISC and NCI enterprise vocabulary services', Drug Information Journal., Vol. 41 No. 3, pp. 405-412.

[13] HL7 v3 Namespaces (Code Systems and Value Sets), available at http://www.hl7.org/fhir/terminologies-v3.html

[14] HL7 EHR Clinical Research Functional Profile, available at http://www.hl7.org/implement/standards/product_brief.cfm?product_id=16

[15] International Classification of Diseases (ICD), available at http://www.who.int/classifications/icd

[16] World Health Organization, available at http://www.who.int

[17] Brown, E.G., Wood, L. and Wood, S. (1999) 'The medical dictionary for regulatory activities (MedDRA)', Drug Safety, Vol. 20 No. 2, pp. 109-17.

[18] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, available at http://www.ich.org

[19] McDonald, C.J., Huff, S.M., Suico, J.G., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., Williams, W., Case, J. and Maloney, P. (2003) 'LOINC, a universal standard for identifying laboratory observations: a 5-year update', Clinical Chemistry, Vol. 49 No. 4, pp. 624-33.

[20] Regenstrief Institute, available at http://www.regenstrief.org

[21] Willoughby, C., Fridsma, D., Chatterjee, L., Speakman, J., Evans, J. and Kush R. (2007) 'A Standard Computable Clinical Trial Protocol: The Role of the BRIDG Model', Drug Information Journal, Vol. 41 No. 41, pp. 383-392.

[22] Fridsma, D.B., Evans, J., Hastak, S. and Mead, C.N. (2008) 'The BRIDG project: a technical report', Journal of the American Medical Informatics Association, Vol. 15 No. 2, pp. 130-7.

[23] Bodenreider, O. (2004) 'The Unified Medical Language System (UMLS): integrating biomedical terminology', Nucleic Acids Research, Vol. 32(Database issue), pp. D267-D270.

[24] Chatterjee, L., Gemzik, D., Gertel, A., Wold, D., Evans, J., Kush, R.D. and Niland J. (2009) 'Optimising Clinical Research via Standardisation – The Clinical Data Interchange Standards Consortium Protocol Representation Model', Drug Development.

[25] CDISC Study/Trial Design Model, available at http://www.cdisc.org/study-trial-design

[26] Weng, C., Tu, S.W., Sim, I. and Richesson, R. (2010) 'Formal representation of eligibility criteria: a literature review', Journal of Biomedical Informatics, Vol. 43 No. 3, pp. 451-67.

[27] Pryor, T.A. and Hripcsak, G. (1993) 'The Arden syntax for medical logic modules, International Journal of Clinical Monitoring and Computing, Vol. 10 No. 4, pp. 215-24.

[28] Sordo, M., Ogunyemi, O., Boxwala, A.A. and Greenes, R.A. (2003) 'GELLO: an object-oriented query and expression language for clinical decision support', Proceedings of the AMIA Symposium, pp. 1012.

[29] Cabot, J. and Gogolla, M. (2012) 'Object constraint language (OCL): a definitive guide', Proceedings of the 12th international conference on Formal Methods for the Design of Computer, Communication, and Software Systems: formal methods for model-driven engineering (SFM'12), pp. 58-90.

[30] Rumbaugh, J., Jacobson, I. and Booch, G. (2004) Unified Modeling Language Reference Manual, 2nd ed., Pearson Higher Education.

[31] Tu, S., Peleg, M., Carini, S., Bobak, M., Rubin, D. and Sim, I. (2008) 'ERGO: A TemplateBased Expression Language for Encoding Eligibility Criteria', The Human Studyome Project.

[32] Tu, S.W., Peleg, M., Carini, S., Bobak, M., Ross, J., Rubin, D. and Sim I. (2011) 'A practical method for transforming free-text eligibility criteria into computable criteria', Journal of Biomedical Informatics, Vol. 44 No. 2, pp. 239-50.

[33] Niland, J. and Cohen, E. (2007) 'ASPIRE: Agreement on Standardized Protocol Inclusion Requirements for Eligibility'.

[34] Weng, C., Wu, X., Luo, Z., Boland, M.R. Theodoratos, D. and Johnson S.B. (2011) 'EliXR: an approach to eligibility criteria extraction and representation', Journal of the American Medical Informatics Association, Vol. 18 Suppl 1, pp. 1116-24.

[35] ClinicalTrials.Gov, available at https://clinicaltrials.gov

[36] Chamberlin, D.D. and Boyce, R.F. (1974) 'SEQUEL: A Structured English Query Language', Proceedings of the ACM SIGFIDET '74 workshop, pp. 249-264.

[37] Luo, Z., Johnson, S.B. and Weng, C. (2010) 'Semi-Automatically Inducing Semantic Classes of Clinical Research Eligibility Criteria Using UMLS and Hierarchical Clustering', Proceedings of the AMIA Symposium, pp. 487-91

[38] Luo, Z., Yetisgen-Yildiz, M. and Weng, C. (2011) 'Dynamic categorization of clinical research eligibility criteria by hierarchical clustering', Journal of Biomedical Informatics, Vol. 44 No. 6, pp. 927-35.

[39] Milian, K., Hoekstra, R., Bucur, A., Ten Teije, A., Van Harmelen, F. and Paulissen J. (2015) 'Enhancing reuse of structured eligibility criteria and supporting their relaxation', Journal of Biomedical Informatics, Vol. 56, pp. 205-19.

[40] Bhattacharya, S. and Cantor, M.N. (2013) 'Analysis of eligibility criteria representation in industry-standard clinical trial protocols', Journal of Biomedical Informatics, Vol. 46 No. 5, pp. 805-813.

[41] Ross, J., Tu, S.W., Carini, S. and Sim, I. (2010) 'Analysis of Eligibility Criteria Complexity in Clinical Trials', Summit on Translational Bioinformatics, pp. 46–50.

[42] Luo, Z., Johnson, S.B., Lai, A.M. and Weng C. (2011) 'Extracting temporal constraints from clinical research eligibility criteria using conditional random fields', Proceedings of the AMIA Symposium, pp. 843-52.

[43] Electronic Medical Records and Genomics (eMERGE), available at https://emerge.mc.vanderbilt.edu

[44] Conway, M., Berg, R.L., Carrell, D., Denny, J.C., Kho, A.N., Kullo, I.J., Linneman, J.G., Pacheco, J.A., Peissig, P., Rasmussen, L., Weston, N., Chute, C.G. and Pathak, J. (2011) 'Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms', Proceedings of the AMIA Symposium, pp. 274-83.

[45] Chondrogiannis, E., Andronikou, V., Tagaris, A., Karanastasis, E., Varvarigou, T. and Tsuji, M. (2017) A novel semantic representation for eligibility criteria in clinical trials, Journal of Biomedical Informatics, Vol. 69,pp. 10-23. DOI= https://doi.org/10.1016/j.jbi.2017.03.013.

[46] Eligibility Criteria Database (EC-DB), available at http://ponte.grid.ece.ntua.gr:8080/EligibilityCriteriaDB/

[47] Chondrogiannis, E., Karanastasis, E., Andronikou, V. and Varvarigou, T. (2017) 'Building a Repository for Inferring the Meaning of Abbreviations Used in Clinical Studies', Journal of Computers, Vol. 12 No. 1, pp. 76-88.

[48] PubMed, available at http://www.ncbi.nlm.nih.gov/pubmed

[49] Chondrogiannis, E., Andronikou, V., Karanastasis, E. and Varvarigou, T. (2015) 'Meaning Inference of Abbreviations Appearing in Clinical Studies', Proceedings of the SLATE 2015 - Symposium on Languages, Applications and Technologies, 18-19 June, Madrid, Spain, pp. 127-136

[50] Jonquet, C., Shah, N.H. and Musen, M.A. (2009) 'The Open Biomedical Annotator', Summit on Translational Bioinformatics, pp. 56–60.

[51] Wang, A.Y., Sable, J.H. and Spackman, K.A. (2002) 'The SNOMED clinical terms development process: refinement and analysis of content', Proceedings of the AMIA Symposium, pp. 845-9.

[52] Lipscomb, C.E. (2000) 'Medical Subject Headings (MeSH)', Bulletin of the Medical Library Association, Vol. 88 No. 3, pp. 265-266.

[53] Dai, M., Shah, N.H., Xuan, W., Musen, M.A., Watson, S. J., Athey, B.D. and Meng, F. (2008) 'An efficient solution for mapping free text to ontology terms', Summit on Translational Bioinformatics, San Francisco, CA.

[54] Banko, M. and Etzioni, O. (2008) 'The tradeoffs between open and traditional relation extraction', Proceedings of the Association for Computational Linguistics (ACL)-08: Human Language Technologies (HLT), 15-20 June, Columbus, Ohio, USA, pp. 28-36.

[55] Eichelberg, M., Aden, T., Riesmeier, J., Dogac, A. and Laleci, G.B. (2005) 'A survey and analysis of Electronic Healthcare Record standards', ACM Computing Surveys, Vol. 37 No. 4, pp. 277-315.

[56] Souza, T., Kush, R. and Evans, J.P. (2007) 'Global clinical data interchange standards are here!', Drug Discovery Today, Vol. 12 No. 3-4, pp. 174-81.

[57] Beale, T. and Heard, S. (2008) 'openEHR Architecture Overview', The openEHR Foundation. available at http://www.openehr.org/releases/1.0.2/architecture/overview.pdf

[58] Degtyarenko, K., de-Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner M. (2008) 'ChEBI: a database and ontology for chemical entities of biological interest', Nucleic Acids Research, Vol. 36(Database issue), pp. D344-D350.

[59] Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., and Sattler U. (2008) 'OWL 2: The next step for OWL', Web Semantics, Vol. 6 No. 4, pp. 309-322.

[60] Eligibility Criteria Representation (EC-R), available at http://ponte.grid.ece.ntua.gr:8080/EligibilityCriteriaRepresentation

[61] Bray, T., Paoli, J. and Sperberg-McQueen C.M. (1998) 'Extensible Markup Language (XML) 1.0 (Fifth Edition)', W3C recommendation.

[62] Prud'hommeaux, E. and Seaborne A. (2008) 'SPARQL Query Language for RDF', W3C Recommendation.

[63] Diagnostic and Statistical Manual of Mental Disorders - Fifth Edition, available at http://www.dsm5.org

[64] Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., and Katz, Y. (2007) 'Pellet: A practical OWL-DL reasoner', Web Semantics, Vol. 5 No. x2, pp. 51-53.

[65] Efficient Patient Recruitment for Innovative Clinical Trials of Existing Drugs to other Indications (PONTE) project, available at http://www.ponte-project.eu

[66] Tagaris, A., Andronikou, V., Karanastasis, E., Chondrogiannis, C., Tsirmpas. T., Varvarigou, T. and Koutsouris, D. (2014) 'PAT: an intelligent authoring tool for facilitating clinical trial design', Proceedings of the 25th European Medical Informatics Conference (MIE 2014), 31 August - 3 September, Istanbul, Turkey, pp. 970-4.

[67] Chondrogiannis, E., Andronikou, V., Karanastasis, E. and Varvarigou, T. (2015), 'A Novel Framework for User-Friendly Ontology-Mediated Access to Relational Databases', WASET, International Science Index 99, International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol. 9 No. 3, pp. 685 - 694.

[68] Friedman, C. and Hripcsak, G. (1999) 'Natural language processing and its future in medicine', Academic Medicine, Vol. 74 No. 8, pp. 890-5.

[69] Fuchs, N.E., Kaljur, K. and Shneider, G. (2006) 'Attempto Controlled English Meets the Challenges of Knowledge Representation, Reasoning, Interoperability and User Interfaces', Presented at the 19th International Florida Artificial Intelligence Research Society Conference (FLAIRS-06), 11–13 May, Melbourne Beach, Florida, US.

[70] Chondrogiannis, E., Andronikou, V., Mourtzoukos, K., Tagaris, A. and Varvarigou, T. (2012) 'A novel query rewriting mechanism for semantically interlinking clinical research with electronic health records', Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS'12), 13-15 June, Craiova, Romania, Article 48, pp. 12.

[71] Angles, R. and Gutierrez, C. (2008) 'The Expressive Power of SPARQL' Proceedings of the 7th International Semantic Web Conference (ISWC 2008), 26-30 October, Karlsruhe, Germany, pp. 114-129.

[72] Chondrogiannis, E., Andronikou, V., Karanastasis, E. and Varvarigou, T. (2014) 'An intelligent ontology alignment tool dealing with complicated mismatches', Presented at the 7th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2015), 9-11 December, Berlin, Germany.

[73] Chondrogiannis, E., Andronikou, V., Karanastasis, E. and Varvarigou, T. (2015) 'An Advanced Query and Result Rewriting Mechanism for Information Retrieval Purposes from RDF datasources', Proceedings of the 6th International Conference on Knowledge Engineering and Semantic Web (KESW 2015), 30 September - 2 October, Moscow, Russia, pp. 32-47.

[74] Bizer, C. and Cyganiak, R. (2006) 'D2R Server – Publishing Relational Databases on the Semantic Web', Poster at the 5th International Semantic Web Conference (ISWC 2006), 5-9 November, Athens, Georgia, USA.

[75] Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M. and XIao, G. (2016) 'Ontop: Answering SPARQL Queries over Relational Databases', Accept for publication in the Semantic Web – Interoperability, Usability, Applicability an IOS Press Journal.

[76] Gruber, T.R. (1993) 'A translation approach to portable ontology specifications', Knowledge Acquisition, Vol. 5 No. 2, pp. 199-220.

[77] Guarino, N. (1997) 'Understanding, building and using ontologies', Int. J. Hum.-Comput. Stud., Vol. 46 Is. 2-3, pp. 293-310.

[78] Klein, M. (2001) 'Combining and relating ontologies: an analysis of problems and solutions' in IJCAI-2001 Workshop on Ontologies and Information Sharing, Seattle, WA, pp. 53-62.

[79] Lambrix, P., Tan, H. (2006) 'SAMBO-A System for Aligning and Merging Biomedical Ontologies', Web Semant., Vol. 4 Is. 3, pp. 196-206.

[80] Ningsheng, J., Wei, H., Gong, C., Yuzhong, Q. (2005) 'Falcon-AO: Aligning Ontologies with Falcon' in K-Cap 2005 Workshop on Integrating Ontologies, Banff, Alberta, Canada, pp. 87-93.

[81] Kolli, R., Doshi, P. (2008) 'OPTIMA: Tool for Ontology Alignment with Application to Semantic Reconciliation of Sensor Metadata for Publication in SensorMap' in Proceedings of the 2008 IEEE International Conference on Semantic Computing (ICSC '08), IEEE Computer Society, Washington, DC, USA, pp. 484-485.

[82] Massmann, S., Raunich, S., Aumueller, D., Arnold, P., Rahm, E. (2011) 'Evolution of the coma match system' in Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany.

[83] Cruz, I.F., Antonelli, F.P., Stroe, C. (2009) 'Agreement Maker Efficient Matching for Large Real-World Schemas and Ontologies' in Proceeding of International Conference on Very Large Databases, pp. 1586-1589.

[84] Shvaiko, P., Euzenat, J. (2013) 'Ontology Matching: State of the Art and Future Challenges', IEEE Transactions on Knowledge and Data Engineering, Vol. 25 No. 1, pp. 158-176.

[85] Scharffe, F., Fensel. D. (2008) 'Correspondence Patterns for Ontology Alignment' in Proceedings of the 16th international conference on Knowledge Engineering: Practice and Patterns, Springer-Verlag, Berlin, Heidelberg, pp. 83-92.

[86] Šváb-Zamazal, O., Svátek, V., Scharffe, F., David, J. (2011) 'Detection and Transformation of Ontology Patterns' in Fred, A., Dietz, J.L.G., Liu, K., Filipe, J. (eds.) IC3K 2009. CCIS, vol. 128, Springer, Berlin, Heidelberg, pp. 210--223.

[87] Scharffe, F., de Bruijn, J. (2005) 'A Language to Specify Mappings between Ontologies' in Proceedings of the Internet Based Systems IEEE Conference (SITIS05), Yandoue, Cameroon.

[88] EDOAL: Expressive and Declarative Ontology Alignment Language, available at http://alignapi.gforge.inria.fr/edoal.html

[89] David, J., Euzenat, J., Scharffe, F., Trojahn dos Santos, C. (2011) 'The Alignment API 4.0.', Semantic Web, Vol. 2 No. 1, pp. 3-10.

[90] Bouquet, P., Giunchiglia, F., Harmelen, F.V., Seraf, L., Stuckenschmidt, H. (2003) 'C-OWL: Contextualizing ontologies', in International Semantic Web Conference, Springer, Berlin, pp. 164-179.

[91] Zhang, H. (2014) 'A Query Driven Method of Mapping from Global Ontology to Local Ontology in Ontology-based Data Integration', Journal of Software, Vol. 9 Is. 3, pp. 738-742.

[92] Lenzerini, M. (2002) 'Data Integration: A Theoretical Perspective' in Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM, New York, USA, pp. 233-246.

[93] Granitzer, M., Sabol, V., Onn, K.W., Lukose, D., Tochtermann, K. (2010) 'Ontology Alignment—A Survey with Focus on Visually Supported Semi-Automatic Techniques', Future Internet, Vol. 2 No. 3, pp. 238-258.

[94] Algorithms and Theory of Computation Handbook, CRC Press LLC. 1999. Levenshtein distance. In Dictionary of Algorithms and Data Structures [online] https://xlinux.nist.gov/dads/HTML/Levenshtein.html

[95] Ghawi, R., Poulain, T., Gomez, G., Cullot N. (2007) 'OWSCIS: Ontology and Web Service Based Cooperation of Information Sources' in Proceedings of the Third International IEEE Conference on Signal-Image Technologies and Internet-Based Systems (SITIS 2007), pp. 246-253.

[96] Correndo, G., Salvadores, M., Millard, I., Glaser, H., Shadbolt, N. (2010) 'SPARQL query rewriting for implementing data integration over linked data' in Proceedings of the 2010 EDBT/ICDT Workshops (EDBT '10), ACM, New York, NY, USA, pp. 4:1-4:11.

[97] W3C - Semantic Web Health Care and Life Sciences (HCLS) Interest Group, available at http://www.w3.org/blog/hcls/

[98] Notation3 (N3): A readable RDF syntax, available at https://www.w3.org/TeamSubmission/n3/

[99] Euzenat, J., Polleres, A., Scharffe, F. (2008) 'Processing ontology alignments with sparql' in Second International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2008). pp. 913-917.

[100] Cullot, N., Ghawi, R., Yétongnon, K. (2007) 'DB2OWL: A Tool for Automatic Database-to-Ontology Mapping' in Proceedings of 15th Italian Symposium on Advanced Database Systems (SEBD 2007), pp. 491–494.

[101] Spyns, P., Meersman, R., Jarrar, M. (2002) 'Data modelling versus ontology engineering', SIGMOD Rec., Vol. 31 No. 4, pp. 12-17.

[102] DB to OWL Tools, available at http://ponte.grid.ece.ntua.gr:8080/DbToOwl/

[103] The D2RQ Platform: Accessing Relational Databases as Virtual RDF Graphs, available at http://d2rq.org/

[104] Protégé, available at http://protege.stanford.edu/

[105] Ontology Alignment Tool, available at http://ponte.grid.ece.ntua.gr:8080/OntoMapping

[106] Common English Stop Words, available at http://www.textfixer.com/resources/common-english-words.php

[107] Willett, P. (2006) 'The Porter stemming algorithm: then and now', Program, Vol. 40 No. 3, pp. 219-223.

[108] Brown, P.F., deSouza, P.V., Mercer, R.L., Della Pietra, V.J., Lai, J.C. (1992) 'Class-based N-gram Models of Natural Language', Comput. Linguist., Vol. 18 No. 4, pp. 467-479.

[109] Kuhn, H.W. (1955) 'The Hungarian method for the assignment problem', Naval Research Logistics Quarterly, Vol. 2 No. 1-2, pp.83-97.

[110] JavaScript Object Notation (JSON), available at http://www.json.org/

[111] RDF 1.1 Turtle, available at https://www.w3.org/TR/turtle/

[112] Pérez, J., Arenas, M., Gutierrez, C. (2009) 'Semantics and complexity of SPARQL', ACM Trans. Database Syst., Vol. 34 No. 3, pp. 1–45.

[113] Gruber, T.R., Toward principles for the design of ontologies used for knowledge sharing ?, International Journal of Human-Computer Studies, 4 (5), 1995, 907-928.

[114] Noy, N.F. and McGuinness, D.L., Ontology development 101: A guide to creating your first ontology, Tech. Rep., 2001.

[115] World Wide Web Consortium (W3C), available at https://www.w3.org/

[116] Catarci, T., Costabile M.F., Levialdi S. and Batini, C., Visual Query Systems for Databases, Journal of Visual Languages & Computing, 8 (2), 1997, 215-260.

[117] García, R., Paulheim, H. and Di Maio, P., Special issue on semantic web interfaces. Semant. Web, 6(8), 2015, 213–214.

[118] Gwani, R. and Cullot, N., Database-to-Ontology Mapping Generation for Semantic Interoperability, In Proceedings of the Third International Workshop on Database Interoperability, (2007).

[119] Kuhlthau, C.C., Inside the search process: Information seeking from the user's perspective, Journal of the American Society for Information Science, 42 (5), 1991, 361–371.

[120] Soylu, A. and Giese, M., Qualifying Ontology-Based Visual Query Formulation, In Proceedings of the Flexible Query Answering Systems, (2015), 243–255.

[121] Kuhn, T., A survey and classification of controlled natural languages, Computational Linguistics, 40 (1) 2014, 121-170.

[122] Flint SPARQL editor, http://openuplabs.tso.co.uk/demos/sparqleditor

[123] Kesteren A. (editor). Cross-origin resource sharing, W3C Recommendation 16 January 2014. https://www.w3.org/TR/cors/

[124] Rietveld, L. and Hoekstra, R., YASGUI: Not Just Another SPARQL Client, In Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI'06), (2013), 78–86.

[125] McCarthy, L., Vandervalk, B. and Wilkinson, M., SPARQL Assist language-neutral query composer, BMC Bioinformatics, 13(Suppl 1):S2, 2012.

[126] Seneviratne, O. and Sealfon, R., QueryMed: An Intuitive Federated SPARQL Query Builder for Biomedical RDF Data, (2010).

[127] Groppe, J., Groppe, S. and Schleifer, A., Visual query system for analyzing social semantic web, In Proceedings of the 20th international conference companion on World wide web, (2011), 217-220.

[128] Heino, N., Dietzold, S., Martin, M. and Auer, S., Developing semantic web applications with the ontowiki framework, In: Networked Knowledge - Networked Media, Springer, 2009, 61–77.

[129] Frischmuth, P., Martin, M., Tramp, S., Riechert, T. and Auer, S., OntoWiki - An authoring, publication and visualization interface for the Data Web. Semantic Web, 6 (3), 2015, 215-240.

[130] Haag, F., Lohmann, S., Bold, S. and Ertl, T., Visual SPARQL querying based on extended filter/flow graphs, In Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, (2014), 305-312.

[131] Young, D. and Shneiderman, B., A graphical filter/flow representation of Boolean queries: a prototype implementation and evaluation, Journal of the American Society for Information Science and Technology, 44 (6), 1993, 327-339.

[132] Brunetti, J.M., García, R. and Auer, S., From Overview to Facets and Pivoting for Interactive Exploration of Semantic Web Data, International Journal on Semantic Web and Information Systems, 9 (1), 2013, 1-20.

[133] Shneiderman, B., The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In Proceedings of the 1996 IEEE Symposium on Visual Languages (VL '96). (1996).

[134] Soylu, A., Giese, M., Jimenez-Ruiz, E., Kharlamov, E., Zheleznyakov, D. and Horrocks, I., OptiqueVQS: Towards an Ontology-based Visual Query System for Big Data, In Proceedings of the Fifth International Conference on Management of Emergent Digital EcoSystems, (2013), 119-126.

[135] Optique: Scalable End-user Access to Big Data, available at http://optique-project.eu/

[136] Tunkelang, D., Faceted search. San Rafael, CA: Morgan & Claypool, (2009).

[137] Russell, A., Smart, P.R., Braines, D. and Shadbolt, N.R., NITELIGHT: A Graphical Tool for Semantic Query Construction, In Proceedings of Semantic Web User Interaction Workshop (SWUI), (2008).

[138] Smart, P.R., Russell, A., Braines, D., et al., A Visual Approach to Semantic Query Design Using a Web-Based Graphical Query Designer. In Proceedings of the 16th international conference on Knowledge Engineering: Practice and Patterns (EKAW '08), (2008), 275-291.

[139] Kabir, A.F.M.S. and Mamun, M.S.I., Graphical Query Builder in Opportunistic Sensor Networks to discover Sensor Information, International Journal of Computer Applications, 15 (1), 2011, 11-25.

[140] Zhang, J., Bhowmick, S.S., Nguyen, H.H., Choi, B. and Zhu, F., DAVINCI: Data-driven Visual Interface Construction for Subgraph Search in Graph Databases, In Proceedings of the IEEE 31st International Conference on Data Engineering, (2015).

[141] Bhowmick, S.s., Choi, B. and Dyreson, C., Data-driven Visual Graph Query Interface Construction and Maintenance: Challenges and Opportunities, Journal Proceedings of the VLDB Endowment, 9 (12), 2016, 984-992.

[142] Ivanova T., Approaches and Tools for Viewing Browsing and Querying Semantic Web Data and Ontologies, International Journal of Scientific & Engineering Research, 6 (5), 2015.

[143] Schraefel, M.C. and Karger, D., The Pathetic Fallacy of RDF, In Proceedings of International Workshop on the Semantic Web and User Interaction (SWUI), (2006).

[144] Anatomical Therapeutic Chemical (ATC) classification system, available http://www.whocc.no/atc/

[145] jQuery JavaScript Library, available at https://jquery.com/

[146] Apache Jena, available at http://jena.apache.org/

[147] Epstein, G.R., The TableTalk query language. Journal of Visual Languages and Computin. 2, (2), 1991.

[148] Godfrey, P., Minimization in cooperative response to failing database queries, International Journal of Cooperative Information Systems, 6, 1997, 95–149.

[149] Arenas, M. and Pérez J. Querying semantic web data with SPARQL. In Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '11). ACM, New York, NY, USA, (2011), 305-316.

[150] Torii, M., Hu, Z., Song, M., Wu, C. H., & Liu, H. (2007). A comparison study on algorithms of detecting long forms for short forms in biomedical text. BMC Bioinformatics, 8(Suppl. 9):S5.

[151] Schwartz, A., and Hearst, M. 2003. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. Pacific Symposium on Biocomputing, 4(8), 451-462. DOI= 10.1142/9789812776303_0042

[152] Park, Y., and Byrd, R. J. 2001. Hybrid Text Mining for Finding Abbreviations and their Definitions. In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, 126–133.

[153] Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M., and Morrell, M. 2001. Automatic extraction of acronym-meaning pairs from MEDLINE databases. Stud Health Technol Inform., 84(Pt 1), 371-5.

[154] Yu, H., Hripcsak, G., and Friedman, C. 2002. Mapping Abbreviations to Full Forms in Biomedical Articles. Journal of the American Medical Informatics Association, 9(3), 262–272. DOI= 10.1197/jamia.M0913

[155] Sohn, S., Comeau, D. C., Kim, W., and Wilbur, W. J. 2008. Abbreviation definition identification based on automatic precision estimates. BMC Bioinformatics, 9(1), 402. DOI= 10.1186/1471-2105-9-402

[156] Chang, J. T., Schutze, H., and Altman, R. B. 2002. Creating an Online Dictionary of Abbreviations from MEDLINE. Journal of the American Medical Informatics Association, 9(6), 612–620. DOI= 10.1197/jamia.M1139

[157] Hastie, T., Friedman, J., and Tibshirani, R. 2001. The Elements of Statistical Learning. Springer Series in Statistics, New York, USA. DOI= 10.1007/978-0-387-21606-5

[158] Kuo, C., Ling, M., Lin, K., and Hsu, C. 2009. BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. BMC Bioinformatics, 10(Suppl. 15):S7. DOI= 10.1186/1471-2105-10-S15-S7

[159] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.

[160] Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. Journal of machine learning research, 2(Nov), 45-66.

[161] Gawlik, M. 2010. Comparison of abbreviation recognition algorithms.

[162] Zhou, W., Torvik, V. I., and Smalheiser, N. R. 2006. ADAM: another database of abbreviations in MEDLINE. Bioinformatics, 22(22), 2813-2818. DOI= 10.1093/bioinformatics/btl480

[163] Okazaki, N., and Ananiadou, S. 2006. Building an Abbreviation Dictionary Using a Term Recognition Approach. Bioinformatics, 22(24), 3089-3095. DOI= 10.1093/bioinformatics/btl534

[164] Frantzi, K., Ananiadou, S., and Mima, H. 2000. Automatic recognition of multi-word terms: The C-value/NC-value method. International Journal on Digital Libraries, 3(2), 115-130. DOI= 10.1007/s007999900023

[165] McCarthy, D., Koeling, R., Weeds, J. and Carroll, J. 2004. Finding predominant word senses in untagged text. In Proceedings of ACL'04, Stroudsburg, PA, USA, 280-287. DOI= 10.3115/1218955.1218991

[166] Xu H., Fan J. W., Hripcsak G., Mendonça E. A., Markatou M., and Friedman C. 2007. Gene symbol disambiguation using knowledge-based profiles. Bioinformatics, 23(8), 1015-1022. DOI= 10.1093/bioinformatics/btm056

[167] Stevenson, M., Guo, Y., Amri, A. A., and Gaizauskas, R. 2009. Disambiguation of biomedical abbreviations. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP '09), Association for Computational Linguistics, Stroudsburg, PA, USA, 71-79. DOI= 10.3115/1572364.1572374

[168] EU Clinical Trials Register, available at www.clinicaltrialsregister.eu

[169] Gale, W.A., Church, K.W., Yarowsky, D.: One sense per discourse. In: Proceedings of the workshop on Speech and Natural Language HLT '91, pp. 233-237. New York (1992)

[170] Abbreviations-annotated corpus of clinical studies, available at http://ponte.grid.ece.ntua.gr:8080/AbbrAnnotatedCorpus/

[171] Porter, M. F. 1980. An algorithm for suffix stripping. Program. 14(3), 130-137. DOI= 10.1108/eb046814

[172] Chondrogiannis, E., Andronikou, V., Karanastasis, E. and Varvarigou, T. (2017) Semantically-enabled Context-aware Abbreviations Expansion in the Clinical Domain. In Proceedings of the 9th International Conference on Bioinformatics and Biomedical Technology (ICBBT 2017), 14-16 May, Lisbon, Portugal.

[173] Ghawi, R. (2010) Ontology-based cooperation of information systems: contributions to database-to-ontology mapping and XML-to-ontology mapping. Diss. Dijon.

[174] International Union of Pure and Applied Chemistry (IUPAC), available at https://iupac.org/

[175] US Food and Drug Administration (FDA), available at http://www.fda.gov/

[176] The Joint Commission, https://www.jointcommission.org/

[177] Joint Commission, Official "Do not Use" List, available at http://www.jointcommission.org/assets/1/18/dnu_list.pdf

[178] Brunetti, L., Santell, J. P., and Hicks, R. W. 2007. The impact of abbreviations on patient safety. The Joint Commission Journal on Quality and Patient Safety, 33(9), 576-583. DOI= 10.1016/S1553-7250(07)33062-6

[179] Institute for Safe Medication Practices (ISMP), available at http://www.ismp.org/

[180] List of Error-Prone Abbreviations, Symbols, and Dose Designations, available at https://www.ismp.org/tools/errorproneabbreviations.pdf

Η σελίδα αυτή είναι σκόπιμα λευκή