Ανάπτυξη Οντολογίας Βιοϊατρικών Όρων

Βαγγελάτος Α.[1], Ορφανός Γ. [2], Τσαλίδης Χ.[2], Καλαμαρά Χ.[3]

[1] Ερευνητικό Ακαδημαϊκό Ινστιτούτο Τεχνολογίας Υπολογιστών (ΕΑΙΤΥ)
[2] Neurosoft A.E.
[3] Ευρωκλινική Αθηνών

Λέξεις Κλειδιά: Επεξεργασία Φυσικής Γλώσσας, Οντολογίες, Εξόρυξη Δεδομένων, Βιοϊατρική ορολογία

Υπεύθυνος Αλληλογραφίας: Δρ. Αριστείδης Βαγγελάτος, Μηχανικός Ηλεκτρονικών Υπολογιστών και Πληροφορικής, Ε. Α. Ινστιτούτο Τεχνολογίας Υπολογιστών, Ακταίου 11, 11851 Θησείο, Αθήνα.
Τηλ.: 210 3416220, Fax: 210 3416250
E-mail: vagelat@cti.gr

Implementation of a Biomedical Ontology

Vagelatos A.[1], Orfanos G.[2], Tsalidis Ch.[2], Kalamara Ch.[3]

[1] Research Academic Computer Technology Institute (RACTI)
[2] Neurosoft S.A.
[3] Athens's Euroclinic

Keywords: Natural Language Processing, ontologies, data mining, biomedical terminology

Corresponding Author: Dr. Aristides Vagelatos, R.A. Computer Technology Institute, Aktaiou 11, GR 11851 Thiseio, Athens, Greece.
Tel.: 210 3416220, Fax: 210 3416250
E-mail: vagelat@cti.gr

Περίληψη

Το μέγεθος της ιατρικής - βιοϊατρικής πληροφορίας που παράγεται καθημερινά στο σύγχρονο κόσμο είναι τεράστιο και ίσως δύσκολο ακόμα και να εκφραστεί με τρόπο απλό. Οι πηγές είναι πολλές (φορείς υγείας, εκπαίδευση, ερευνητικοί φορείς, θεσμικοί φορείς, κλπ). Τα μέσα διακίνησης, άλλαξαν με την πάροδο του χρόνου και η ψηφιακή μορφή τείνει να επικρατήσει ολοσχερώς. Ο κυβερνοχώρος (WWW) και σε αυτή την περίπτωση είναι το μέσο που υπερισχύει με βάση τα μεγέθη. Το αναπόφευκτο ερώτημα είναι αν και με ποιο τρόπο είναι δυνατό να αξιοποιηθεί πλέον αυτή η υπερπληθώρα πληροφορίας και να μετατραπεί σε γνώση.

Σήμερα, ο απλός χρήστης (είτε ερευνητής, είτε ασθενής, είτε άλλης κατηγορίας) έχει ιδιαίτερα απλοϊκά εργαλεία ως υποβοήθηση στην προσπάθειά του να πλοηγηθεί ή να αναζητήσει ιατρική πληροφορία. Απλές μηχανές αναζήτησης γενικού σκοπού με μόνη δυνατότητα κάποιες λογικές εκφράσεις. Ειδικά δε για την Ελληνική γλώσσα με τις ιδιαιτερότητές της (πλούσια μορφολογία, μακρόχρονη εξέλιξη, κλπ) τα προβλήματα είναι ακόμα πιο μεγάλα.

Με βάση τα παραπάνω, κρίνεται ως απαραίτητη η ανάπτυξη κατάλληλης υποδομής γλωσσικής τεχνολογίας για το βιοϊατρικό τομέα και την Ελληνική γλώσσα, που θα υποβοηθήσει το μέσο χρήστη, τον ερευνητή αλλά και κάθε ενδιαφερόμενο να βελτιώσει τις δυνατότητες αναζήτησης, πλοήγησης, αλλά και κάθε μορφή επεξεργασίας βιοϊατρικών δεδομένων.

Η παρούσα αναφορά περιγράφει/προδιαγράφει το έργο ΙΑΤΡΟΛΕΞΗ[1], δηλαδή το σχεδιασμό της ανάπτυξης μιας οντολογίας βιοϊατρικών όρων αλλά και των απαραίτητων εργαλείων πληροφορικής τα οποία θα αποτελέσουν πολύτιμα όπλα για την ερευνητική και όχι μόνο κοινότητα. Βασικό παραδοτέο και εργαλείο επίδειξης του έργου θα είναι μια εξειδικευμένη μηχανή αναζήτησης που θα αξιοποιεί την οντολογία για να δίνει στο χρήστη καλύτερα και ποιοτικότερα αποτελέσματα. Πέραν τούτου, η οντολογία και τα εργαλεία που θα κατασκευαστούν θα είναι διαθέσιμα ώστε να μπορούν να προσαρμοστούν, να επεκταθούν και να εμπλουτιστούν, αποτελώντας έτσι τη βάση και για άλλες χρήσεις (δεικτοδότηση ιατρικών κειμένων, χαρακτηρισμός κειμένων, επεξεργασία δεδομένων που παράγονται από φορείς υγείας, μέτρηση ποιότητας υπηρεσιών υγείας, κλπ).

[1] Το έργο «ΙΑΤΡΟΛΕΞΗ» έχει εγκριθεί για χρηματοδότηση από την Γενική Γραμματεία Έρευνας και Τεχνολογίας (κωδ. 9) στα πλαίσια του Μέτρου 3.3 του Επιχειρησιακού Προγράμματος «Κοινωνία της Πληροφορίας».

1. Εισαγωγή

Η βιοϊατρική πληροφορία που υπάρχει σήμερα διαθέσιμη σε ψηφιακή μορφή, έχει ήδη χαρακτηριστεί «ζούγκλα πληροφορίας» (information jungle - G. Eysenbach), αφηγηματικής (narrative) μορφής με βασικό πλέον (αρνητικό) χαρακτηριστικό την ιδιαίτερη δυσκολία αξιοποίησής της. Ο στόχος της επιστημονικής – ερευνητικής προσπάθειας είναι η μετάβαση από την κατάσταση αυτή σε μια νέα όπου η πληροφορία θα είναι δομημένη με τρόπο ώστε να μπορεί να χαρακτηριστεί ως συλλογή – δεξαμενή γνώσης (knowledge repository) και η οποία θα επιτρέπει τη χρήση προηγμένων τεχνολογιών διαχείρισης γνώσης (knowledge management).

Αντικείμενο του έργου ΙΑΤΡΟΛΕΞΗ (www.iatrolexi.cti.gr) είναι η δημιουργία της απαραίτητης γλωσσικής υποδομής για την Ελληνική γλώσσα, που θα επιτρέψει σε πρώτη φάση την καλύτερη διαχείριση και επεξεργασία της βιοϊατρικής πληροφορίας που υπάρχει σε ψηφιακή μορφή. Αυτό θα καταστεί δυνατό με τη δημιουργία και διάθεση στους χρήστες εξελιγμένων εργαλείων για την αναζήτηση, τη συσχέτιση και το χαρακτηρισμό των βιοϊατρικών κειμένων. Σε επόμενη φάση, είναι δυνατή η επέκταση των αποτελεσμάτων του και η διαμόρφωσή του ώστε να εξυπηρετεί και άλλους σκοπούς (π.χ. υλοποίηση δυνατοτήτων semantic web σε βιοϊατρικούς ιστοχώρους, κλπ).

Πιο συγκεκριμένα, το έργο αποσκοπεί στη δημιουργία περιβάλλοντος ανάπτυξης, πιστοποίησης και αξιοποίησης οντολογίας βιοϊατρικών όρων με ιδιαίτερο χαρακτηριστικό τη δυνατότητα διαχείρισης πολυλεκτικών όρων. Η Οντολογία αποτελεί τον πλέον ενδεδειγμένο και αξιοποιήσιμο μηχανισμό αποτύπωσης γνώσης μίας θεματικής περιοχής. Καθορίζει τον τρόπο αναπαράστασης με τον οποίο οι έννοιες τα αντικείμενα και οι μεταξύ τους σχέσεις αποτυπώνονται. Οι κυριότεροι λόγοι για την κατασκευή μιας οντολογίας είναι 1) Αναπαράσταση της γνώσης με ένα μοντέλο εξίσου καταληπτό σε ανθρώπους και μηχανές. 2) Καταγραφή των παραδοχών. 3) Επαναχρησιμοποίηση της αποτυπωμένης γνώσης. 4) Ανάλυση της γνώσης. 5) Προηγμένη ευρετηρίαση, ανάκτηση και παρουσίαση πληροφορίας, δυνατότητα εξόρυξης δεδομένων.

Παρά την αδιαμφισβήτητη σπουδαιότητα των οντολογιών και τη μεγάλη ανάπτυξη που γνωρίζουν τα τελευταία χρόνια, δεν υπάρχει καθολικά αποδεκτή μεθοδολογία για την κατασκευή τους. Η μεθοδολογία που θα υποστηριχτεί από το προτεινόμενο προς υλοποίηση έργο είναι συνδυασμός δύο μεθοδολογιών από διαφορετικούς επιστημονικούς χώρους. Η πρώτη προέρχεται από τον χώρο της τεχνολογίας λογισμικού (software engineering) και περιγράφει τον τρόπο ανάπτυξης λογισμικού ενώ η δεύτερη από τον χώρο της γλωσσολογίας και περιγράφει τον τρόπο κατασκευής λεξικών. Τα βασικά σημεία που χαρακτηρίζουν τις δύο αυτές μεθοδολογίες είναι ο κύκλος ανάπτυξης ενός προϊόντος λογισμικού για την πρώτη μεθοδολογία και η χρήση corpus (συλλογής κειμένων) για την ανάπτυξη γλωσσικών πόρων από τη δεύτερη μεθοδολογία. Το σύστημα που θα αναπτυχθεί θα υποστηρίξει τις δύο

αυτές μεθοδολογίες με ένα σύνολο εργαλείων ανάλυσης κειμένων, διαχείρισης γλωσσικών πόρων, καθώς και παρουσίασης της οντολογίας.

Συγκεκριμένα, τα παραγόμενα του έργου θα είναι:
Α) Μεθοδολογία ανάπτυξης οντολογιών. Η μεθοδολογία θα περιλαμβάνει: 1) τον καθορισμό της αρχικής ταξινομίας πάνω στην οποία θα βασιστεί μια οντολογία, 2) τη συλλογή κειμένων της θεματικής περιοχής-στόχου, 3) την ανίχνευση του ειδικού λεξιλογίου που χρησιμοποιούν τα κείμενα, 4) τον καθορισμό μορφοσυντακτικών κανόνων περιγραφής των όρων, 5) την εξαγωγή υποψήφιων όρων, 6) τον εμπλουτισμό της οντολογίας με επιλεγμένους όρους και σχέσεις και 7) ένα συνεχή κύκλο των βημάτων 4, 5 και 6. Αν και η εφαρμογή της μεθοδολογίας θα γίνει για την κατασκευή βιοϊατρικής οντολογίας, μπορεί να εφαρμοσθεί στην κατασκευή οντολογιών σε οποιαδήποτε θεματική περιοχή.
Β) Υπολογιστικοί πόροι και εργαλεία που θα υποστηρίζουν τη μεθοδολογία ανάπτυξης οντολογιών. Θα εμπλουτισθούν/επεκταθούν/ανάπτυχθούν/προσ-αρμοστούν πόροι και εργαλεία με εξειδίκευση στην ανάλυση κειμένων βιοϊατρικού περιεχομένου, στην εξαγωγή βιοϊατρικών όρων και στην κατασκευή οντολογίας, οι οποίοι αναλυτικά είναι οι εξής: 1) Μορφολογικό Λεξικό εμπλουτισμένο με βιοϊατρικούς όρους. 2) Μορφοσυντακτικός Σχολιαστής (Morphosyntactic Tagger), ο οποίος θα χαρακτηρίζει μορφοσυντακτικά κάθε λέξη ενός κειμένου. 3) Αναγνωριστής Ονοματικών Φράσεων (Noun-Phrase Chunker), ο οποίος θα ανιχνεύει υποψήφιους πολυλεκτικούς όρους σε κείμενα με χρήση μορφοσυντακτικών κανόνων. 4) Περιβάλλον Ανάπτυξης Οντολογιών βασισμένο στο state-of-the-art σύστημα Protégé (http://protege.stanford.edu), το οποίο θα συνδυάζεται με το Μορφολογικό Λεξικό της Neurosoft ώστε να είναι σε θέση να χειρίζεται τη μορφολογική ποικιλότητα των ελληνικών βιοϊατρικών όρων. 5) Σημασιολογικός Σχολιαστής (Semantic Tagger), ο οποίος θα επισυνάπτει πληροφορία από την οντολογία σε όρους που αναγνωρίζονται σε κείμενα. 5) Μηχανισμός δεικτοδότησης βιοϊατρικών κειμένων βασισμένος κυρίως στους (μονολεκτικούς και πολυλεκτικούς) ιατρικούς όρους που εμφανίζονται μέσα σε αυτά (και όχι σε όλες –άκριτα– τις λέξεις των κειμένων).
Γ) Σώμα (corpus) βιοϊατρικών κειμένων, το οποίο αφενός θα οδηγήσει την όλη διαδικασία εξαγωγής ορολογία και ανάπτυξης της οντολογίας και αφετέρου θα είναι διαθέσιμο στους χρήστες/ερευνητές για ανάκτηση/εξόρυξη πληροφορίας.
Δ) Εφαρμογές Ιστού για την παρουσίαση και χρησιμοποίηση των αποτελεσμάτων και εργαλείων του έργου, οι οποίες θα αποτελούν και το **τελικό παραδοτέο** του έργου. Όλες οι τεχνολογίες, εργαλεία και πόροι που θα αναπτυχθούν στα πλαίσια του έργου αυτού θα διατεθούν για χρήση μέσα από ένα δικτυακό τόπο. Με εφαρμογές ιστού ο εξειδικευμένος και μη χρήστης του διαδικτύου θα μπορεί: 1) Να ελέγχει την ορθογραφία μίας λέξης (βιοϊατρικός όρος ή μέρος βιοϊατρικού όρου). 2) Να περιηγείται στην οντολογία πραγματοποιώντας σύνθετες αναζητήσεις για σχέσεις που διέπουν όρους της οντολογίας. 3) Να εισάγει ένα νέο κείμενο και να λαμβάνει ως αποτέλεσμα το κείμενο με μορφοσυντακτικούς και σημασιολογικούς σχολιασμούς των όρων του κειμένου (annotations). 4) Να αναζητά κείμενα με

συγκεκριμένους όρους ή συνδυασμό όρων ή σχέσεων που ισχύουν σε όρους. Η αναζήτηση θα μπορεί να γίνει α) στο σώμα κειμένων που θα συλλεχθεί στα πλαίσια του έργου και β) σε κείμενα του Παγκόσμιου Ιστού μέσω ειδικού meta-search engine που θα αναπτυχθεί για το σκοπό αυτό και το οποίο θα χρησιμοποιεί το Google (http://www.google.com). Για τις παραπάνω λειτουργίες, θα αναπτυχθούν και αντίστοιχες υπηρεσίες ιστού (web services), με τη βοήθεια των οποίων ο δικτυακός τόπος θα προσφέρει τη δυνατότητα αξιοποίησης των αποτελεσμάτων από λογισμικό τρίτων στον ερευνητικό χώρο.

2. Τεχνική Περιγραφή

Στην πρώτη ενότητα εργασίας θα καθοριστούν οι προδιαγραφές για την κατάρτιση σώματος (corpus) ιατρικών κειμένων και θα σχεδιαστεί η οντολογία βιοϊατρικών όρων. Παράλληλα, θα σχεδιαστεί το μοντέλο αναπαράστασης της οντολογίας, βασισμένο στη γλώσσα περιγραφής οντολογιών OWL (Web Ontology Language, http://www.w3.org/2004/OWL), το οποίο θα παρέχει τη δυνατότητα για: α) διαχείριση πολυλεκτικών όρων (η πολυλεκτικότητα αποτελεί βασικό χαρακτηριστικό όχι μόνο των όρων της βιοϊατρικής αλλά και όλων των ειδικών πεδίων), β) καταχώριση ερμηνευμάτων σε ελληνικά και αγγλικά, γ) ορισμό σχέσεων συνωνυμίας, ομωνυμίας κτλ., πέραν των κλασικών οντολογικών σχέσεων (υπερωνυμίας, υπωνυμίας, μερωνυμίας), δ) καταχώριση της φωνητικής μεταγραφής των όρων (ώστε μελλοντικά να μπορούν να διαβαστούν αυτόματα) και ε) δυνατότητα σύνδεσης των όρων με πολυμεσική πληροφορία (εικόνα, ήχο, βίντεο). Μέρος της σχεδίασης του μοντέλου της οντολογίας αποτελεί και ο καθορισμός του τρόπου έκφρασης κανόνων μεταξύ εννοιών και σχέσεων. Οι κανόνες αυτοί θα μπορούν να λειτουργήσουν ως μηχανισμός για: α) τον έλεγχο της ορθότητας των αντικειμένων της οντολογίας, β) τον περιορισμό του χώρου αποθήκευσης της οντολογίας και γ) την επαγωγή γνώσης από τα στοιχεία της οντολογίας.

Παράλληλα με τις προδιαγραφές του corpus και τη σχεδίαση της οντολογίας, θα ξεκινήσει η δεύτερη ενότητα εργασίας, η οποία αφορά στην ανάπτυξη πληροφοριακού συστήματος για τη διαχείριση και επεξεργασία των κειμένων του corpus. Τα εργαλεία και οι γλωσσικοί πόροι των οποίων η ανάπτυξη θα ξεκινήσει στην ενότητα αυτή είναι: 1) Αποθήκη Εγγράφων, όπου θα αποθηκεύονται τα κείμενα και τα μεταδεδομένα τους. 2) Προσκομιστής Εγγράφων, ο οποίος θα συλλέγει τα κείμενα του corpus, κατά κανόνα από το Διαδίκτυο. 3) Μετατροπέας Εγγράφων, ο οποίος θα μετατρέπει τα συλλεγόμενα έγγραφα από τις διάφορες μορφές τους (html, pdf κτλ.) σε μια ενιαία μορφή (xml). 4) Αναγνωριστής Στοιχείων, ο οποίος θα κερματίζει κάθε κείμενο στα συστατικά του (λέξεις, σημεία στίξης, αριθμοί, σύμβολα κτλ.). 5) Μορφολογικό Λεξικό, το οποίο θα είναι υπεύθυνο για την κλίση των όρων. 6) Μορφοσυντακτικός Σχολιαστής, ο οποίος θα επισυνάπτει μορφοσυντακτική πληροφορία στις λέξεις των κειμένων. 7) Αναγνωριστής Όρων, ο οποίος θα αναγνωρίζει μονολεκτικούς και πολυλεκτικούς όρους στα κείμενα. 8)

Σημασιολογικός Σχολιαστής, ο οποίος θα επισυνάπτει στους όρους σημασιολογική πληροφορία από την οντολογία.

Η Τρίτη Ενότητα Εργασίας αφορά στη συλλογή και επεξεργασία βιοϊατρικών κειμένων αλλά και στην συγκρότηση της πρώτης ταξινομίας βιοϊατρικών εννοιών, πάνω στην οποία θα στηριχτεί η κατασκευή της οντολογίας. Η συλλογή θα γίνει από κείμενα που βρίσκονται στο διαδίκτυο είτε σε ιατρικούς κόμβους, είτε σε ελληνικά περιοδικά είτε αλλού (πανεπιστήμια, κλπ). Όταν, σύμφωνα με τις προδιαγραφές συγκρότησης του corpus, συμπληρωθεί ο ελάχιστος ικανός αριθμός κειμένων, θα αρχίσει η διαδικασία εξαγωγής της ορολογίας. Στη φάση αυτή το σύστημα ξεκινά την λειτουργία του με πρώτο στόχο τον εμπλουτισμό των γλωσσικών πόρων (μορφολογικό λεξικό, συντακτικοί κανόνες αναγνώρισης όρων) που υποβοηθούν τη λειτουργία των εργαλείων. Μικτή ομάδα γλωσσολόγων και γιατρών θα αναλύει τα αποτελέσματα της επεξεργασίας των κειμένων και θα «διορθώνει» τα αδύνατα σημεία του συστήματος. Οι άγνωστες λέξεις θα αξιολογούνται και μέρος αυτών θα εισάγονται στο μορφολογικό λεξικό. Θα αξιολογείται η απόδοση των συντακτικών κανόνων που περιγράφουν πολυλεκτικούς όρους. Θα διορθώνονται τα προβληματικά τους σημεία και θα εμπλουτίζεται το σύνολο με νέους κανόνες για τους όρους που το σύστημα απέτυχε να βρει. Ταυτόχρονα, θα καταγράφονται οι νέοι όροι που ανακαλύπτονται, οι μορφοσυντακτικές παραλλαγές των υπαρχόντων όρων αλλά και το γλωσσικό περιβάλλον που δύο ή περισσότεροι όροι συνεμφανίζονται. Η συνεμφάνιση αυτή θα σηματοδοτήσει πιθανές σχέσεις μεταξύ των όρων.

Αντικείμενο της τέταρτης ενότητας εργασίας είναι η κατασκευή οντολογίας με τους όρους που συλλέγονται κατά τη διάρκεια της τρίτης ενότητας εργασίας. Η τέταρτη ενότητα εργασίας θα ξεκινήσει αμέσως μετά τον καθορισμό του μοντέλου αναπαράστασης οντολογιών (ο οποίος θα λάβει χώρα στην πρώτη ενότητα εργασίας), με εναρκτήρια δράση την προσαρμογή του open-source περιβάλλοντος ανάπτυξης οντολογιών Protégé κατά τρόπο ώστε να υποστηρίζει το εν λόγω μοντέλο. Στη συνέχεια, οι συλλεχθέντες όροι θα αξιολογούνται και όσοι επιλεχθούν θα εισάγονται στην οντολογία. Η συνεμφάνιση των όρων στο ίδιο κειμενικό περιβάλλον (context) θα εξετάζεται για την ύπαρξη πιθανής συσχέτισής τους. Στην περίπτωση αυτή, οι σχέσεις θα εισάγονται στην οντολογία. Μετά τον εμπλουτισμό της οντολογίας με όρους και σχέσεις, τα κείμενα θα ευρετηριάζονται με βάση τους όρους και τις σχέσεις που εμφανίζονται στο περιεχόμενό τους. Τα δεδομένα του ευρετηρίου θα αναλύονται από αλγορίθμους εξόρυξης δεδομένων οι οποίοι θα παράγουν συσχετίσεις μεταξύ των όρων.

Στην πέμπτη ενότητα εργασίας θα αναπτυχθούν οι εφαρμογές που θα επιτρέπουν την αξιοποίηση των αποτελεσμάτων του έργου. Ένα υποσύνολο των εργαλείων που θα κατασκευαστούν στις προηγούμενες ενότητες εργασίας θα επεκταθούν κατάλληλα ώστε να αποτελέσουν εφαρμογές ιστού (web

applications), οι οποίες θα φιλοξενηθούν στον ιστοχώρο του έργου (*βλ.* επόμενη §). Οι εφαρμογές αυτές είναι οι εξής: 1) Περιηγητής Οντολογίας, ο οποίος θα επιτρέπει την αναζήτηση και την πλοήγηση στην αναπτυχθείσα οντολογία. 2) Μηχανή Αναζήτησης βιοϊατρικής πληροφορίας στο σώμα κειμένων του έργου και στον Παγκόσμιο Ιστό, όπου η σύνθεση ερωτημάτων θα κατευθύνεται από την Οντολογία. 3) Ορθογραφικός Διορθωτής όρων της βιοϊατρικής. 4) Μορφοσυντακτικός και Σημασιολογικός Σχολιαστής βιοϊατρικών κειμένων, ο οποίος θα επισυνάπτει στις λέξεις/φράσεις των κειμένων μορφοσυντακτική πληροφορία (από το Μορφολογικό Λεξικό) και σημασιολογική πληροφορία (από την Οντολογία).

4. Αναλυτική Περιγραφή ενοτήτων υλοποίησης

Αναλυτικά τα βήματα υλοποίησης είναι τα παρακάτω:

4.1 Σχεδιασμός Σώματος Βιοϊατρικών Κειμένων και Οντολογίας

Ο σχεδιασμός ενός σώματος κειμένων υπόκειται σε ποσοτικές προδιαγραφές (π.χ. αντιπροσωπευτικότητα) και σε ποιοτικές προδιαγραφές σχετικές με το είδος των κειμένων που πρέπει να συλλεχθούν (π.χ. κείμενα εκπαιδευτικού χαρακτήρα, επιστημονικά κείμενα, επιστημονικά άρθρα εκλαϊκευτικού χαρακτήρα κτλ.). Κι ενώ οι προδιαγραφές μπορούν να συλλεχθούν εύκολα από τη διεθνή βιβλιογραφία (*βλ.* Biber 1993, Bowker 1996), το δύσκολο είναι ο εντοπισμός των πηγών από τις οποίες θα συλλεχθούν κείμενα που να πληρούν τις προδιαγραφές. Μια από τις δράσεις της ενότητας αυτής θα αφορά στον εντοπισμό πηγών ελληνικών βιοϊατρικών κειμένων (ιστοχώροι, ηλεκτρονικές βιβλιοθήκες, εκδοτικοί οίκοι κτλ), ώστε το σχέδιο συγκρότησης σώματος κειμένων που θα παραχθεί, εκτός από τις προδιαγραφές, να αναφέρει και τις πηγές.

Παράλληλα θα ξεκινήσει ο σχεδιασμός του μοντέλου αναπαράστασης της οντολογικής γνώσης (επί τη βάσει των διεθνών προτύπων OWL/RDF) και της μεθοδολογίας ανάπτυξης της οντολογίας. Ιδιαίτερο χαρακτηριστικό του μοντέλου της οντολογίας θα είναι η δυνατότητα ταυτοποίησης των μορφολογικών παραλλαγών των όρων, χαρακτηριστικό απαραίτητο για γλώσσες με έντονη μορφολογία (όπως είναι η Ελληνική), το οποίο θα υποστηριχθεί κατόπιν σύνδεσης του μοντέλου με το Μορφολογικό Λεξικό της Neurosoft. Επίσης, θα καθορισθεί το είδος των σχέσεων μεταξύ των εννοιών καθώς και η μηχανισμός έκφρασης κανόνων. Σχέσεις και κανόνες αποτελούν τον επαγωγικό μηχανισμό επέκτασης της γνώσης από τις υπάρχουσες έννοιες/όρους. Η δύναμη του μηχανισμού αυτού μπορεί να ποικίλει από τις «όψεις» (views) ενός Συστήματος Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) έως τις γλώσσες προγραμματισμού με λογικές προτάσεις (Horn clauses).

4.2 Υλοποίηση Εργαλείων Διαχείρισης Σώματος Κειμένων και αυτόματης εξαγωγής ορολογίας

Στο πλαίσιο της ενότητας αυτής θα υλοποιηθούν/επεκταθούν τα εξής εργαλεία:

1) Αποθήκη Εγγράφων (Document Warehouse): είναι ο χώρος αποθήκευσης των κειμένων αλλά και των παραγόμενων από αυτά μεταδεδομένων. Θα υλοποιηθεί πάνω σε ένα Σύστημα Διαχείρισης Σχεσιακών Βάσεων Δεδομένων (ΣΔΣΒΔ - RDBMS).
2) Προσκομιστής Εγγράφων (Crawler): διατρέχει συγκεκριμένους δικτυακούς τόπους ή τοπικούς δίσκους και προσκομίζει τα κείμενα προς επεξεργασία.
3) Αναγνωριστής Στοιχείων (Tokenizer): κερματίζει ένα κείμενο σε μία σειρά στοιχείων (tokens: λέξεις, σημεία στίξης, αριθμοί, σύμβολα κτλ.) με τα οποία τροφοδοτούνται οι επόμενες φάσεις επεξεργασίας του κειμένου.
4) Μορφοσυντακτικός Σχολιαστής (Morphosyntactic Tagger): επισυνάπτει μορφοσυντακτικά μεταδεδομένα (μέρος του λόγου, γένος, αριθμός πτώση, κλπ.) σε κάθε λέξη του κειμένου (που έχει αναγνωρίσει ο Tokenizer). Αυτό γίνεται με τη βοήθεια του Μορφολογικού Λεξικού.
5) Μορφολογικό Λεξικό. Η Neurosoft έχει αναπτύξει μορφολογικό λεξικό ~90.000 λημμάτων, το οποίο περιέχει και περιορισμένο αριθμό βιοϊατρικών όρων. Το λεξικό αυτό θα εμπλουτισθεί με τις άγνωστες λέξεις-όρους που θα βρεθούν στα βιοϊατρικά κείμενα. Το υπάρχον σύστημα περιγραφής της κλίσης των λημμάτων καλύπτει τον ορισμό της μορφοσυντακτικής πληροφορίας μονολεκτικών όρων.
6) Μηχανισμός Κλίσης Πολυλεκτικών Όρων. Το μοντέλο περιγραφής της κλίσης πολυλεκτικών όρων είναι πιο πολύπλοκο από αυτό των μονολεκτικών, διότι η κλίση πολυλεκτικών όρων δεν ακολουθεί μία τυποποιημένη διαδικασία. Υπάρχουν πολυλεκτικοί όροι στους οποίους κλίνεται μόνο η μία συστατική λέξη ενώ οι υπόλοιπες παραμένουν άκλιτες, π.χ. οξείδιο του αζώτου, οξείδια του αζώτου. Σε άλλες περιπτώσεις, όλες οι συστατικές λέξεις ακολουθούν την ίδια κλιτική φόρμα, π.χ. γλουταμινικό νάτριο, γλουταμινικού νατρίου. Θα επεκταθεί ο μηχανισμός κλίσης μονολεκτικών όρων (στοιχείο 5 της παρούσας λίστας) για να υποστηρίξει τον ορισμό πολυλεκτικών όρων. Η εφαρμογή που θα υλοποιήσει το μοντέλο αυτό θα παρέχει στο χρήστη τη δυνατότητα εύκολου ορισμού της κλίσης πολυλεκτικών όρων.
7) Αναγνωριστής Όρων: συμβουλεύεται το Μορφολογικό Λεξικό, όπου περιγράφεται η μορφολογία των μονολεκτικών και πολυλεκτικών όρων, καθώς και κανόνες που περιγράφουν τη σύνταξη πολυλεκτικών όρων και αναγνωρίζει τους όρους αυτούς σε κείμενα, σε όποια κλιτική μορφή κι αν βρίσκονται. Στην περίπτωση των πολυλεκτικών όρων, πρέπει να είναι σε θέση να αναγνωρίζει δομικές παραλλαγές (π.χ. οξείδια *του* Αζώτου, οξείδια Αζώτου, αζωτούχα οξείδια) και συντμήσεις (π.χ. NOx).

Όλα τα παραπάνω εργαλεία θα συνδεθούν σε ένα ολοκληρωμένο πληροφοριακό σύστημα. Η πληροφορία που θα διαχειρίζεται το σύστημα αφορά: α) διευθύνσεις δικτυακών τόπων και παραμέτρους άντλησης εγγράφων από αυτούς, β) κείμενα και μεταδεδομένα που έχουν επισυναφθεί στα κείμενα από τα εργαλεία γλωσσικής επεξεργασίας, γ) κανόνες κλίσης μονολεκτικών και πολυλεκτικών όρων, δ) μορφοσυντακτικοί κανόνες αναγνώρισης όρων στα κείμενα και ε) ευρετήριο των κειμένων βασισμένο στους βιοϊατρικούς όρους που αυτά περιέχουν. Η πληροφορία αυτή αποθηκεύεται σε μία κοινή βάση δεδομένων. Ένα από τα κύρια χαρακτηριστικά της βάσης δεδομένων θα είναι η διατήρηση όλων των εκδόσεων των δεδομένων (versioning), επιτρέποντας στους ερευνητές να παρακολουθήσουν την πορεία εξέλιξης ενός αντικειμένου/πόρου.

4.3 Συλλογή Βιοϊατρικών Κειμένων και Εξαγωγή Ορολογίας

Η ενότητα αυτή θα ξεκινήσει αμέσως μόλις παραδοθούν από την ενότητα 2 ο Προσκομιστής Εγγράφων, ο Αναγνωριστής Στοιχείων και ο Μορφοσυντακτικός Σχολιαστής, εργαλεία τα οποία είναι απαραίτητα για να αρχίσει η συλλογή και στοιχειώδης επεξεργασία βιοϊατρικών κειμένων. Η πηγές των κειμένων καθορίζονται στο «Σχέδιο Συγκρότησης Σώματος Βιοϊατρικών Κειμένων».

Από την επεξεργασία των κειμένων θα προκύπτουν (στην αρχή αρκετές) «άγνωστες» λέξεις οι οποίες, λόγω της υψηλής κάλυψης του Μορφολογικού Λεξικού σε όρους του κοινού λεξιλογίου, με μεγάλη πιθανότητα θα είναι όροι ή μέρος όρων της Βιοϊατρικής. Μικτή ομάδα γλωσσολόγων και γιατρών θα αναλύει τις άγνωστες λέξεις και θα τις κατατάσσει σε μία από τις τρεις κατηγορίες: α) βιοϊατρικοί όροι, β) όχι-βιοϊατρικοί όροι και γ) ορθογραφικά λάθη. Οι λέξεις των κατηγοριών (α) και (β) θα εμπλουτίζουν το Μορφολογικό Λεξικό και η διαδικασία θα επαναλαμβάνεται στα νεοεισερχόμενα κείμενα, έως ότου οι άγνωστες λέξεις ανήκουν όλες στην κατηγορία (γ).

Επόμενο βήμα είναι η εξαγωγή πολυλεκτικών όρων από τα κείμενα. Ομάδα γλωσσολόγων θα ορίσει τους κανόνες σύνταξης πολυλεκτικών όρων, π.χ. ο όρος *γλουταμινικό οξύ* περιγράφεται από ένα κανόνα της μορφής:

ΕΠΙΘ_βιοϊατρ(γένος=Γ, αριθμός=Α, πτώση=Π) + ΟΥΣ_βιοϊατρ(γένος=Γ, αριθμός=Α, πτώση=Π),

ενώ ο όρος *δίαυλος ασβεστίου* περιγράφεται από ένα κανόνα της μορφής:

ΟΥΣ_βιοϊατρ(γένος=?, αριθμός=?, πτώση=?)+ ΟΥΣ_βιοϊατρ(πτώση="Γεν").

Οι κανόνες σύνταξης πολυλεκτικών όρων θα ενσωματωθούν στον Αναγνωριστή Όρων, ο οποίος θα επεξεργάζεται τα κείμενα και θα προτείνει πολυλεκτικούς όρους. Ομάδα γλωσσολόγων και ιατρών θα αξιολογεί τους προτεινόμενους πολυλεκτικούς όρους. Οι αποδεκτοί όροι θα εμπλουτίζουν το Μορφολογικό Λεξικό, με τη βοήθεια του Μηχανισμού Κλίσης Πολυλεκτικών Όρων. Οι μη αποδεκτοί όροι θα ωθούν τους γλωσσολόγους στην αναθεώρηση κάποιων συντακτικών κανόνων.

Όλη η παραπάνω διαδικασία είναι επαναλαμβανόμενη, με την έννοια της συνεχόμενης ανακύκλωσης των λειτουργιών: επεξεργασία κειμένων ➔ ενημέρωση Μορφολογικού Λεξικού ➔ διόρθωση/βελτίωση των κανόνων και των εργαλείων επεξεργασίας ➔ επεξεργασία κειμένων ➔ ... Στο τέλος της ενότητας 3, τα εργαλεία επεξεργασίας θα έχουν φτάσει στο μέγιστο της αποδοτικότητάς τους και οι χρησιμοποιούμενοι πόροι (Μορφολογικό Λεξικό και Συντακτικοί Κανόνες) θα έχουν φτάσει στο μέγιστο της πληρότητάς τους, όσον αφορά το πεδίο της βιοϊατρικής ορολογίας.

Τέλος στην ενότητα αυτή, θα καθοριστούν οι βασικές έννοιες της οντολογίας καθώς και η ιεράρχηση των εννοιών, παίρνοντας υπόψη τις ήδη υπάρχουσες βιοϊατρικές ταξινομίες από το διεθνή χώρο (π.χ. ICD10, Νόσοι - Διαγνώσεις, MeSH, κτλ.) με παράλληλη αναφορά στο δημιουργηθέν corpus . Η ταξινομία βιοϊατρικών όρων που θα προκύψει θα χρησιμοποιηθεί αφενός για να δοκιμαστεί η βασική λειτουργικότητα του μοντέλου της οντολογίας και αφετέρου για να χτιστεί πάνω σε αυτή η πλήρης οντολογία με όρους που θα εξαχθούν από κείμενα.

4.4 Επεξεργασία Βιοϊατρικών Όρων και Δημιουργία Οντολογίας

Μετά τον ορισμό του Μοντέλου Αναπαράστασης Οντολογίας, η ενότητα αυτή θα ξεκινήσει υλοποιώντας το μοντέλο αυτό με προσαρμογή/επέκταση του περιβάλλοντος ανάπτυξης οντολογιών Protégé. Στη συνέχεια, η Αρχική Ταξινομία Βιοϊατρικών Όρων θα εισαχθεί στο Protégé και πάνω σε αυτή θα στηριχθεί η ανάπτυξης της οντολογίας. Η ομάδα εξειδικευμένων γιατρών και έμπειρων γλωσσολόγων θα εξετάζει:

1) τους βιοϊατρικούς όρους που συλλέχθηκαν από τα κείμενα,
2) το περιβάλλον (context) που βρέθηκαν και κυρίως το περιβάλλον που συνεμφανίστηκαν οι όροι.
3) αν οι υποψήφιοι όροι είναι παραλλαγές όρων της οντολογίας ή νέοι όροι,
4) αν οι έννοιες που ορίσθηκαν στην οντολογία είναι επαρκείς για τον εισαγωγή ενός όρου στην οντολογία ή πρέπει να επεκταθεί η Οντολογία με νέες έννοιες,
5) αν η συνεμφάνιση όρων σηματοδοτεί κάποια σχέση που πρέπει να εισαχθεί στην οντολογία.

Επιπλέον στην οντολογία θα εισαχθεί η απαραίτητη πληροφορία σχετικά με τις συσχετίσεις των όρων, την αγγλική και ελληνική ερμηνεία καθώς και τη φωνητική μεταγραφή.

Όλες οι παραπάνω ενέργειες θα υποστηρίζονται από το Σύστημα Διαχείρισης Σώματος Κειμένων.

4.5 Υλοποίηση Εφαρμογών Ιστού

Οι εφαρμογές ιστού που θα αναπτυχθούν στην ενότητα αυτή θα φιλοξενηθούν στον ιστοχώρο του έργου. Η τελική μορφή που θα πάρει ο ιστοχώρος του έργου θα περιέχει τις ακόλουθες εφαρμογές ιστού:

1) Περιηγητής Οντολογίας, ο οποίος θα επιτρέπει στο χρήστη την περιήγηση στους όρους της οντολογίας. Ο χρήστης θα μπορεί επίσης να κάνει σύνθετες ερωτήσεις για όρους και σχέσεις της οντολογίας με γραφικό και απλό τρόπο.
2) Μηχανή Αναζήτησης στο σώμα βιοϊατρικών κειμένων, τα οποία θα έχουν ευρετηριαστεί επι τη βάσει σημασιολογικών κατηγοριών από την οντολογία, αλλά και σε κείμενα του Παγκόσμιου Ιστού. Στην περίπτωση αναζήτησης στο σώμα βιοϊατρικών κειμένων, ο χρήστης θα μπορεί να διαμορφώνει ερωτήματα με όρους που προέρχονται από την Οντολογία, με τη βοήθεια του Περιηγητή Οντολογίας. Στην περίπτωση αναζήτησης στον Παγκόσμιο Ιστό, η μηχανή αναζήτησης θα λειτουργεί ως προθάλαμος (meta-search) του Google. Θα παίρνει δηλαδή την ερώτηση (η οποία πιθανόν να περιέχει όρους της οντολογίας) από το χρήστη, θα τη στέλνει στο Google και στη συνέχεια θα αναλύει/φιλτράρει τα κείμενα/απαντήσεις του Google για εμφάνιση των αναζητηθέντων όρων.
3) Ορθογραφικός Διορθωτής, ο οποίος θα ελέγχει την ορθογραφία ενός κειμένου ή λέξης χρησιμοποιώντας το πλούσιο Μορφολογικό Λεξικό βιοϊατρικών όρων που θα έχει δημιουργηθεί.
4) Μορφοσυντακτικός και Σημασιολογικός Σχολιαστής, ο οποίος θα επισυνάπτει μορφοσυντακτικά (από το Μορφολογικό Λεξικό) και σημασιολογικά (από την Οντολογία) σχόλια σε λέξεις/στοιχεία ενός κειμένου. Ο σχολιασμός θα πραγματοποιείται με τη χρήση επισημειώσεων εκφρασμένων ως XML tags.

5. Αναμενόμενα Αποτελέσματα

Ο βασικός στόχος του έργου για την υλοποίηση υποδομής που θα βελτιώσει την δυνατότητα διαχείρισης βιοϊατρικής πληροφορίας, είναι σίγουρο ότι έχει πληθώρα ωφελουμένων είτε στη μορφή που θα προκύψει στα πλαίσια της παρούσας πρότασης, είτε με βελτιώσεις προσθήκες που θα μπορεί κάποιος να κάνει με βάση πιο εξειδικευμένη κατεύθυνση.

Πιο συγκεκριμένα, η επιστημονική κοινότητα του κλάδου της βιοϊατρικής, θα αποκτήσει μια σειρά εργαλείων που θα υποβοηθήσουν σε μεγάλο βαθμό το έργο της: από την καλύτερη αναζήτηση σε ιατρικές πηγές, έως την αυτόματη δεικτοδότηση, χαρακτηρισμό ή άλλη επεξεργασία των κειμένων. Παράλληλα ο απλός χρήστης μέσω μιας εξειδικευμένης πύλης αναζήτησης ιατρικών θεμάτων, θα μπορεί να εντοπίσει ευκολότερα πιο αξιόπιστη και χρήσιμη πληροφορία για θέματα υγείας που τον απασχολούν. Επιπλέον φορείς υγείας που παράγουν πληροφορία ιατρικού περιεχομένου (νοσοκομεία, δημόσια

διοίκηση, κλπ) θα μπορούν να επεξεργαστούν τα δεδομένα τους με πληθώρα τρόπων με στόχο την ικανοποίηση των σκοπών λειτουργίας τους (π.χ. μελέτες συσχέτισης νόσων – ιατρικών εξετάσεων για τα νοσοκομεία, κλπ).

6. Βιβλιογραφία

[1] http://www.neurosoft.gr/en/download/main.asp?actcat=download&actbul=dld Το site της Neurosoft Α.Ε. απ΄ όπου μπορεί κανείς να βρει και να χρησιμοποιήσει κάποια γλωσσικά εργαλεία (ορθογράφος και συλλαβιστής).

[2] http://www.nlm.nih.gov/research/umls/ Στον κόμβο αυτό της Εθνικής Ιατρικής Βιβλιοθήκης των ΗΠΑ, υπάρχει το περιβάλλον UMLS. Επίσης υπάρχουν και μια σειρά από ενδιαφέρουσες δημοσιεύσεις σχετικά με την ολοκλήρωση του UMLS.

[3] http://www.nlm.nih.gov/mesh/meshhome.html Medical Subject Headings της Εθνικής Ιατρικής Βιβλιοθήκης των ΗΠΑ.

[4] http://protege.stanford.edu/ Στον κόμβο αυτό του πανεπιστημίου του Στάνφορντ, χει το περιβάλλον "protégé" για την υλοποίηση οντολογιών.

[5] http://www.hon.ch/ Health on the Net Foundation.

[6] http://www.who.int/classifications/icd/en/ International Classification of Diseases : Διεθνής κωδικοποίηση Νόσων.

[7] Biber, D. 1993. "Representatives in corpus design", *Literary and Linguistic Computing*, 8, pp. 243-257.

[8] Bowker, L. 1996. "Towards a corpus-based approach to terminology", *Terminology*, 3, pp. 27-52.

[9] Eysenbach G.. *The semantic Web and healthcare consumers: a new challenge and opportunity on the horizon?* Int. J. Healthcare Technology and Management, vol. 59, n. 3/4/5, 2003.

[10] Ceusters W, Smith B, Flanagan J. *Ontology and Medical Terminology: Why description Logics are not enough*. TERP 2003, San Antonio, USA.

[11] Boyer et al. *HONselect: a multilingual and intelligent search tool integrating heterogeneous web resources*. Int. J. of Medical Informatics, (64) 2001.

[12] Gobel G. et al. *A MeSH based intelligent search intermediary for Consumers Health Information Systems*. Int. J. of Medical Informatics, (64) 2001.

Abstract

This paper presents the design of terminological and specialized textual resources that will be produced in the framework of the national R&D project "IATROLEXI".

The aim of IATROLEXI is to create the critical infrastructure for the Greek language, i.e. linguistic resources and tools, to be used in high level Natural Language Processing (NLP) applications in the domain of Biomedicine, i.e. information extraction, data mining, etc.

The project will build upon existing resources that have been developed by the project partners, i.e. a Greek morphological lexicon of about 100.000 words, and language processing tools such as a lemmatiser and a morphosyntactic tagger, and it will further develop new resources such as a specialised corpus of biomedical texts and an ontology of medical terminology.