ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΚΟΙΝΩΝΙΑ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΠΡΑΞΗ

«ΕΠΕΞΕΡΓΑΣΙΑ ΕΙΚΟΝΩΝ, ΗΧΟΥ ΚΑΙ ΓΛΩΣΣΑΣ»

στο πλαίσιο του ΜΕΤΡΟΥ 3.3

«Έρευνα και Τεχνολογική Ανάπτυξη στην Κοινωνία της Πληροφορίας»

ΕΡΓΟ - **9: ΙΑΤΡΟΛΕΞΗ**

Π8: Αρχική Ταξινομία Βιοϊατρικών Όρων

(στο παραδοτέο περιλαμβάνεται οπτικός δίσκος που περιέχει την ταξινομία)

Ημερομηνία:	*28.02.2007*
Έκδοση:	*Final*
Τύπος:	*Εμπιστευτικό*

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

Το έργο ΙΑΤΡΟΛΕΞΗ έχει στόχο τη δημιουργία της κρίσιμης υποδομής για την ελληνική γλώσσα, η οποία θα αποτελέσει τη βάση για εξελιγμένες εφαρμογές επεξεργασίας φυσικής γλώσσας (ΕΦΓ) στο θεματικό πεδίο της βιοϊατρικής, όπως ευρετηρίαση κειμένων (text indexing), εξαγωγή και ανάκτηση πληροφορίας (information extraction and retrieval), εξόρυξη δεδομένων, συστήματα ερωταποκρίσεων κτλ.
Τα προσδοκώμενα αποτελέσματα του έργου είναι εργαλεία που θα απευθύνονται στον τελικό χρήστη, όπως είναι ο *ορθογραφικός διορθωτής ελληνικών βιοϊατρικών όρων*, αλλά και εργαλεία που θα ενισχύσουν την επεξεργασία βιοϊατρικών κειμένων στα ελληνικά και θα βελτιώσουν την αναζήτηση και την ανάκτηση βιοϊατρικών δεδομένων, όπως *μορφοσυντακτικός σχολιαστής* (morphosyntactic tagger) κατάλληλα προσαρμοσμένος στις ιδιαιτερότητες της βιοϊατρικής υπογλώσσας, και *οντολογία* της βιοϊατρικής ορολογίας, που θα χρησιμοποιηθεί ως τελικό προϊόν για αναζήτηση ορολογικών και εννοιολογικών πληροφοριών αλλά και ως ενδιάμεσος γλωσσικός πόρος, ως **βάση γνώσης** (knowledge base), για την τροφοδότηση ενός *σημασιολογικού σχολιαστή βιοϊατρικών κειμένων*.

Το παρόν παραδοτέο **Π8: «Αρχική Ταξινομία Βιοϊατρικών Όρων»** περιλαμβάνει:
1. αρχείο δεδομένων με την αρχική ταξινομία της οντολογίας ΙΑΤΡΟΛΕΞΗ (περιλαμβάνεται στον επισυναπτόμενο οπτικό δίσκο),
2. την παρούσα, σύντομη τεκμηρίωση της αρχικής ταξινομίας.

2 ΚΑΤΑΡΤΙΣΗ ΑΡΧΙΚΗΣ ΤΑΞΙΝΟΜΙΑΣ

Όπως ήδη έχει αναφερθεί στο Παραδοτέο Π2 «Μοντέλο Αναπαράστασης της Οντολογίας», στο πλαίσιο του παρόντος έργου, η οντολογία αποτελεί την **τυπική περιγραφή της γνώσης του πεδίου της βιοϊατρικής**, για την κατασκευή της οποίας υλοποιήθηκαν μέχρι στιγμής οι εξής υποεργασίες:

1. **Καθορισμός μοντέλου αναπαράστασης οντολογίας[1]:** αφορά αφενός τον προσδιορισμό του θεωρητικού πλαισίου και των σημασιολογικών μηχανισμών που θα χρησιμοποιηθούν για την ανάλυση, την περιγραφή και την οργάνωση των **εννοιών** του βιοϊατρικού τομέα και αφετέρου τον προσδιορισμό των πληροφοριών που θα χρησιμοποιηθούν για την οργάνωση των **όρων** που κατασημαίνουν κάθε έννοια. Για την επιλογή του βασικού μοντέλου αναπαράστασης, εξετάσθηκαν μια σειρά από σχετικές οντολογίες – κωδικοποιήσεις (ICD10, MESH, Νόσοι – Διαγνώσεις, UMLS). Κατόπιν μελέτης των μοντέλων αυτών τόσο από την ιατρική όσο και από τη γλωσσική πλευρά τους βασικό μοντέλο αναπαράστασης των βιοϊατρικών εννοιών του ΙΑΤΡΟΛΕΞΗ υιοθετήθηκε το *μοντέλο αναπαράστασης* του Σημασιολογικού Δικτύου του UMLS (http://umlsinfo.nlm.nih.gov/), που αποτελεί την πιο εκτεταμένη από άποψη όγκου και ποικιλίας δεδομένων πηγή οντολογικής γνώσης για το θεματικό πεδίο της βιοϊατρικής.

2. **Κατάρτιση αρχικής ταξινομίας:** για την κατάρτιση της αρχικής ταξινομίας των βιοϊατρικών όρων ακολουθήθηκε η *από πάνω προς τα κάτω* (top-down) και η *από κάτω προς τα πάνω* (bottom-up) προσέγγιση ανάπτυξης:

   - Στο πλαίσιο της *από πάνω προς τα κάτω* (top-down) προσέγγισης προσαρμόστηκαν στα ελληνικά από τους ειδικούς (ιατρούς γλωσσολόγους) οι κατηγορίες βιοϊατρικών εννοιών που περιλαμβάνονται στις δύο ιεραρχίες του Σημασιολογικού Δικτύου του UMLS.
   - Στο πλαίσιο της *από κάτω προς τα πάνω* (bottom-up) προσέγγισης, η οποία θα εφαρμόζεται σταδιακά μέχρι και την ολοκλήρωση της οντολογίας, η αρχική ταξινομία εμπλουτίζεται με ειδικότερες κατηγορίες εννοιών, που είτε απαντώνται στο Σώμα Κειμένων του ΙΑΤΡΟΛΕΞΗ είτε αφορούν την ευκρινέστερη διάκριση των σημασιολογικών τύπων του UMLS με βάση τις ιεραρχικές σχέσεις “is_a” και “part_of”, δηλαδή των εννοιών που αναφέρονται σε είδη και μέρη, διάκριση που δεν εφαρμόζεται συστηματικά στις κατηγοριοποιήσεις του Σημασιολογικού Δικτύου και της υπάρχουσας αρχικής ταξινομίας.

[1] Τα αποτελέσματα αυτής της υποεργασίας καταγράφονται αναλυτικά στο Π2: «Μοντέλο Αναπαράστασης της Οντολογίας».

3 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΗΣ ΑΡΧΙΚΗΣ ΤΑΞΙΝΟΜΙΑΣ ΕΝΝΟΙΩΝ

Ευρύτητα ταξινομικών κατηγοριών: Οι κατηγορίες εννοιών που συγκροτούν την αρχική ταξινομία αφορούν ένα ευρύ φάσμα υποπεδίων της βιοϊατρικής (π.χ. ανατομία, βιολογία, ασθένειες, συμπτώματα, χημικές ουσίες, φάρμακα κτλ.), γεγονός που εξυπηρετεί αποτελεσματικότερα τις ανάγκες ενός έργου υποδομής όπως είναι το ΙΑΤΡΟΛΕΞΗ. Επιπλέον, η ευρύτητα των κατηγοριών, εν αντιθέσει με άλλες οντολογίες που είναι περιορισμένες σε ένα μόνο πεδίο της βιοϊατρικής (π.χ. GO[2] ή FMA[3]), εξασφαλίζει συνοχή στη γνώση του σχετικού θεματικού πεδίου, η οποία μπορεί να γενικευτεί και να επαναχρησιμοποιηθεί σε ποικιλία άλλων υποπεδίων.

Επεκτασιμότητα γενικών σημασιολογικών κατηγοριών: Η τρέχουσα έκδοση της αρχικής ταξινομίας περιέχει 135 σημασιολογικούς τύπους και η ευρύτητά των περιγραφόμενων κατηγοριών επιτρέπει την αξιοποίησή της ως οντολογίας πυρήνα για την κωδικοποίηση εννοιών από διαφορετικά πεδία της βιοϊατρικής. Ωστόσο, επειδή η ιεραρχία των εννοιών δομείται μέσω της σχέσης “is_a”, δηλαδή μέσω της σχέσης γένους-είδους, είναι δυνατή η εξειδίκευση των γενικών σημασιολογικών κατηγοριών μέσω της καθιέρωσης πιο ειδικών εννοιών, χωρίς να ανατρέπεται ούτε η υπάρχουσα δομή του μοντέλου ούτε το περιεχόμενο της αρχικής ταξινομίας. Αυτό το χαρακτηριστικό επιτρέπει τον εμπλουτισμό των δύο ιεραρχιών με ειδικότερες κατηγορίες, που μπορεί να απαντηθούν στο Σώμα Κειμένων του ΙΑΤΡΟΛΕΞΗ, για τη λεπτομερέστερη περιγραφή εννοιών σε πεδία που το UMLS περιορίζεται σε γενικές κατηγοριοποιήσεις (π.χ. ανατομία, ασθένειες κτλ.).

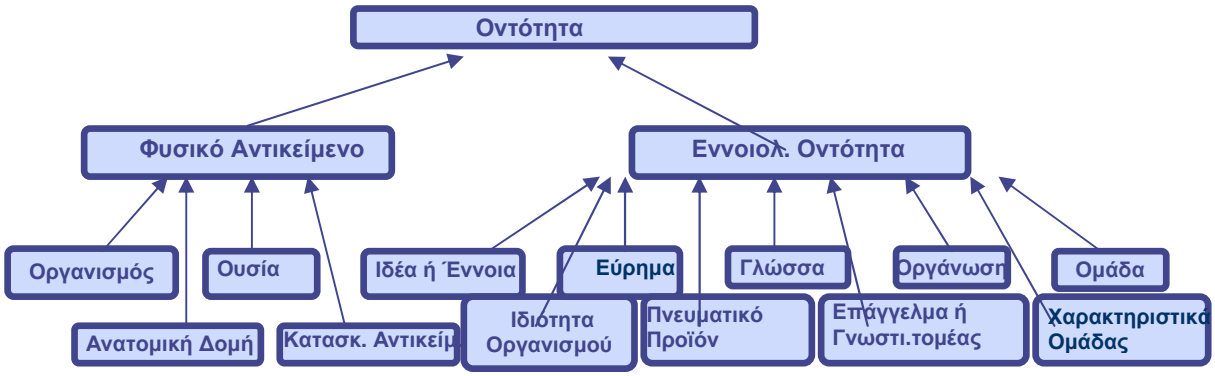
Επεξεργασιμότητα από υπολογιστικά εργαλεία: Αν και το Σημασιολογικό Δίκτυο του UMLS δε θεωρείται ένα αυστηρά τυπικό μοντέλο αναπαράστασης οντολογιών που θα μπορούσε να στηρίξει άμεσα και χωρίς σχεδιαστικές προσαρμογές τη δημιουργία μηχανών συμπερασμού, ωστόσο, η καθαρότητα και η απλότητα που χαρακτηρίζει τη δομή των δύο ιεραρχιών του επιτρέπουν την εύκολη μεταφορά του σε σύγχρονους φορμαλισμούς αναπαράστασης (π.χ. OWL) και την επεξεργασία του από πλατφόρμες ανάπτυξης οντολογιών (π.χ. Protégé).

[2] http://www.geneontology.org/

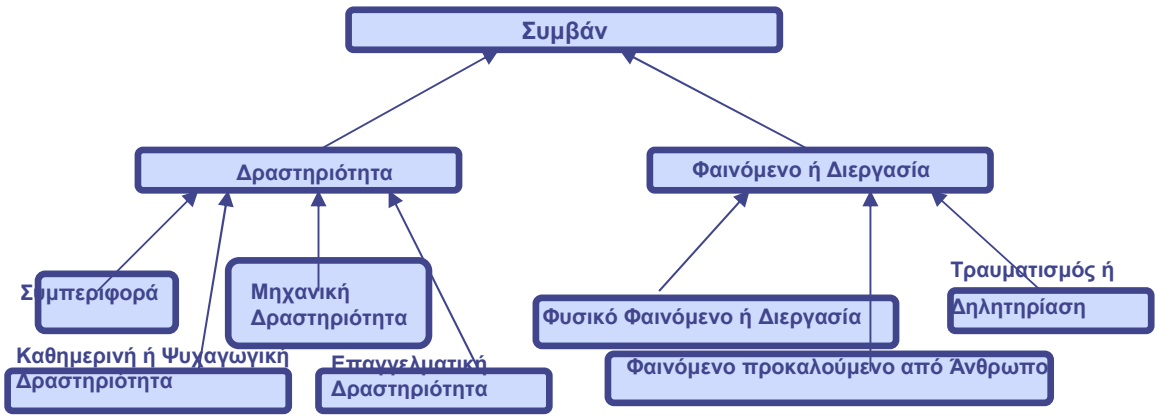
[3] http://sig.biostr.washington.edu/projects/fm/AboutFM.html

4 ΚΑΤΗΓΟΡΙΕΣ ΕΝΝΟΙΩΝ ΤΗΣ ΑΡΧΙΚΗΣ ΤΑΞΙΝΟΜΙΑΣ

Η αρχική ταξινομία περιλαμβάνει δύο επιμέρους ιεραρχίες: την ιεραρχία των οντοτήτων (entities) και την ιεραρχία των συμβάντων (events). Η ιεραρχία των οντοτήτων περιλαμβάνει 100 και η ιεραρχία των συμβάντων 35 κατηγορίες εννοιών. Οι κυριότεροι κόμβοι των δύο ιεραρχιών παρουσιάζονται σχηματικά στα ακόλουθα σχήματα:



Σχήμα 1. Σχηματική παρουσίαση των κυριότερων κόμβων στην Ταξινομία των Οντοτήτων



Σχήμα 2. Σχηματική παρουσίαση των κυριότερων κόμβων στην Ταξινομία των Γεγονότων

- **Ανατομία**
- **Αντικείμενα**
- **Γεωγραφικές Περιοχές**
- **Γονίδια και Μοριακές Ακολουθίες**
- **Διαδικασίες**
- **Διαταραχές**
- **Δραστηριότητες και Συμπεριφορές**
- **Έμβια Όντα**
- **Έννοιες και Ιδέες**
- **Επαγγελματικοί τομείς**
- **Οργανώσεις**
- **Φαινόμενα**
- **Φυσιολογία**
- **Χημικά και Φάρμακα**

Οι κατηγορίες εννοιών που περιλαμβάνονται στις δύο ιεραρχίες μελετήθηκαν λεπτομερώς από τους ειδικούς του έργου και διαμορφώθηκαν μέσω των ακόλουθων **15** σημασιολογικών ομάδων σε ένα λιγότερο πολύπλοκο εννοιολογικό σχήμα, που αποτυπώνει με συνοπτικό τρόπο την εννοιολογική δομή του πεδίου της βιοϊατρικής:

5 ΠΑΡΑΣΤΑΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΗΝ ΑΡΧΙΚΗ ΤΑΞΙΝΟΜΙΑ(ΑΡΧΕΙΟ ΔΕΔΟΜΕΝΩΝ)

Στο αρχείο δεδομένων του Π8 (βλέπε επισυναπτόμενο οπτικό δίσκο αλλά και επόμενο κεφάλαιο), παρουσιάζονται υπό τη μορφή πίνακα η ταξινομία των Οντοτήτων (Entities) και η Ταξινομία των Γεγονότων (Events).

Οι σχετικοί πίνακες περιλαμβάνουν τις εξής πληροφορίες:

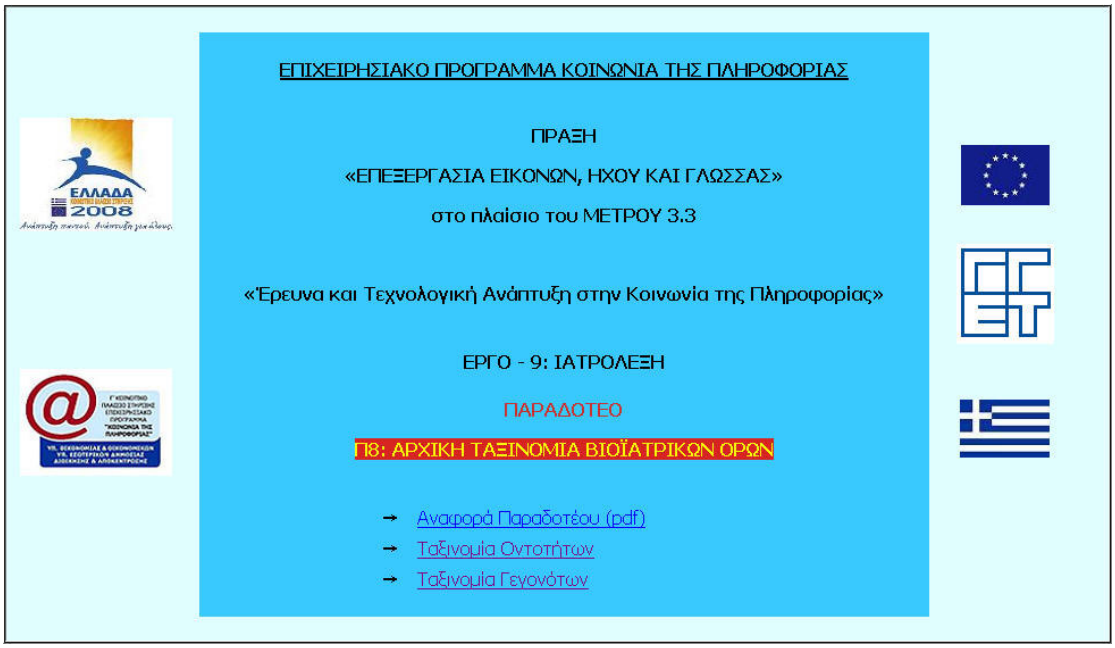
1. Στην αριστερή στήλη δίνεται ένας κωδικός με βάση το αλφαριθμικό σύστημα ταξινόμησης. Η Ιεραρχία των Οντοτήτων ξεκινά με το γράμμα Α ενώ η Ιεραρχία των Γεγονότων με το γράμμα Β και έπεται ακολουθία ψηφίων σύμφωνα με το δεκαδικό σύστημα ταξινόμησης (Α, Α1, Α1.1, Α1.2 … Β, Β1, Β1.1, Β1.2 κτλ.).
2. Στη μεσαία στήλη δίνεται το όνομα της έννοιας που αποτελεί τον εκάστοτε κόμβο της ταξινομίας.
3. Στην τελευταία στήλη παρουσιάζεται σύντομη περιγραφή της περιγραφόμενης κατηγορίας εννοιών.

Οι δύο κατάλογοι διατάσσονται κατά ιεραρχική στάθμη, σε μορφή «κατά βάθος διάσχισης», δηλ. εμφανίζονται πρώτα οι διαδοχικά υποτασσόμενοι σημασιολογικοί τύποι και στη συνέχεια οι παρατασσόμενοι κατά την αλφαριθμική τάξη του κωδικού τους.

6 ΟΔΗΓΙΕΣ ΧΡΗΣΗΣ ΤΟΥ ΟΠΤΙΚΟΥ ΔΙΣΚΟΥ

Στο παρόν παραδοτέο περιλαμβάνεται οπτικός δίσκος (CD) ο οποίος περιέχει την αρχική ταξινομία με δυνατότητα πλοήγησης σε αυτή.

Μόλις εισαχθεί ο οπτικός δίσκος στη μονάδα οπτικού δίσκου του υπολογιστή θα ανοίξει αυτόματα ο φυλλομετρητής με την παρουσίαση του παραδοτέου (εικόνα 1).



Εικόνα 1: Στιγμιότυπο παρουσίασης του παραδοτέου.

Στην πρώτη σελίδα παρουσιάζονται στοιχεία του έργου και του συγκεκριμένου παραδοτέου καθώς και τρεις υπερσύνδεσμοι (links):

- Ο πρώτος οδηγεί στην αναφορά του παραδοτέου (παρούσα αναφορά) σε μορφή PDF.
- Ο δεύτερος οδηγεί στην ταξινομία οντοτήτων και
- Ο τρίτος, οδηγεί στην ταξινομία γεγονότων.