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**

ΠΑΡΑΔΟΤΕΟ

ΠΣ: **Σύστημα Διαχείρισης Σώματος Κειμένων και Εξαγωγής Ορολογίας (λογισμικό)**

(στο παραδοτέο περιλαμβάνεται οπτικός δίσκος που περιέχει το λογισμικό)

Ημερομηνία:	*28.05.2007*
Έκδοση:	*Final*
Τύπος:	*Εμπιστευτικό*

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

Το παρών παραδοτέο αποτελεί το **Π3:** ***«Σύστημα Διαχείρισης Σώματος Κειμένων και Εξαγωγής Ορολογίας»*** του έργου ΙΑΤΡΟΛΕΞΗ. Πρόκειται για ένα λογισμικό (στην πραγματικότητα μια σειρά εργαλείων) που υλοποιήθηκε στο πλαίσιο της Ενότητας Εργασίας 2, με τίτλο: *«Υλοποίηση Εργαλείων Ανάπτυξης Διαχείρισης Σώματος Κειμένων και Αυτόματης Εξαγωγής Ορολογίας»*.

Τα εργαλεία αυτά έγιναν για να υποστηρίξουν την ανάπτυξη των γλωσσικών πόρων του έργου, αλλά και την ολοκλήρωση των τελικών στόχων του.

Τα εργαλεία που ολοκληρώθηκαν δεν υλοποιήθηκαν όλα από την αρχή. Με δεδομένο ότι οι δύο φορείς διαθέτουν μακράν εμπειρία στην ανάπτυξη εργαλείων γλωσσικής τεχνολογίας, κάποια αναπτύχθηκαν εκ του μηδενός, κάποια βασίστηκαν πάνω σε ήδη υπάρχοντα προπλάσματα και κάποια εμπλουτίστηκαν. Πιο συγκεκριμένα στα πλαίσια της παρούσας ΕΕ, ολοκληρώθηκαν τα παρακάτω συστημάτων:

1) Αποθήκη Εγγράφων (Document Warehouse): είναι ο χώρος αποθήκευσης των κειμένων αλλά και των παραγόμενων από αυτά μεταδεδομένων. Υλοποιήθηκε με βάση την open source βάση δεδομένων MySQL.
2) Προσκομιστής Εγγράφων (Crawler): διατρέχει συγκεκριμένους δικτυακούς τόπους και προσκομίζει τα κείμενα προς επεξεργασία. Υλοποιήθηκε σε Java (JDK 6.0).
3) Μετατροπέας Εγγράφων σε απλό κείμενο (Document Converter): μετατρέπει τα έγγραφα HTML ή PDF που προσκομίζει ο Crawler σε έγγραφα TXT (δηλ. σε απλό κείμενο). Υλοποιήθηκε σε Java (JDK 6.0). Το εργαλείο αυτό δεν είχε προβλεφθεί αρχικά μια και είχε γίνει η εκτίμηση ότι θα βρεθούν έγγραφα έτοιμα σε μορφή κειμένου, κάτι που όπως αποδείχθηκε εκ των υστέρων δεν ήταν σωστό.
4) 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. Για την υλοποίησή του εξελίχθηκε ένα πρωτόλειο σχετικό σύστημα που είχε αναπτυχθεί στο παρελθόν για ερευνητικούς σκοπούς.
5) 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 πραγματοποιείται με τη βοήθεια του Μορφολογικού Λεξικού. Υλοποιήθηκε σε Java (JDK 6.0).

6) Μορφολογικό Λεξικό. Η Neurosoft Α.Ε. έχει αναπτύξει μορφολογικό λεξικό της Ελληνικής με περίπου 90.000 λήμματα, το οποίο περιέχει και περιορισμένο αριθμό βιοϊατρικών όρων. Το λεξικό αυτό εμπλουτίστηκε με τις άγνωστες λέξεις-όρους (περίπου 7.250) που συλλέχθηκαν στο πλαίσιο της ΕΕ3. Το υπάρχον σύστημα περιγραφής της κλίσης των λημμάτων καλύπτει τον ορισμό της μορφοσυντακτικής πληροφορίας μονολεκτικών όρων.
7) Μηχανισμός Κλίσης Πολυλεκτικών Όρων. Στην ΕΕ3 επεκτάθηκε ο μηχανισμός κλίσης μονολεκτικών όρων ώστε να υποστηρίζει τον ορισμό πολυλεκτικών όρων. Για την υποστήριξη του μοντέλου αυτού, αναπτύχθηκε σχετική εφαρμογή, που παρέχει στο χρήστη τη δυνατότητα εύκολου ορισμού της κλίσης πολυλεκτικών όρων.
8) Αναγνωριστής Όρων. Αναπτύχθηκε σχετικό σύστημα το οποίο συμβουλεύεται το Μορφολογικό Λεξικό, καθώς και τους κανόνες που περιγράφουν τη σύνταξη πολυλεκτικών όρων και αναγνωρίζει τους όρους αυτούς σε κείμενα, σε όποια κλιτική μορφή κι αν βρίσκονται.

Σχεδόν όλα τα παραπάνω εργαλεία, μπορούν να χρησιμοποιηθούν αυτόνομα. Επιπρόσθετα, για την καλύτερη και ευκολότερη χρήση τους από τρίτους (ερευνητές ή μη) ολοκληρώθηκαν (εκτός του crawler) κάτω από ένα κοινό περιβάλλον που καθοδηγεί το χρήστη σε διαδοχικές ενέργειες/βήματα.

Στο παραδοτέο **Π4: Εγχειρίδιο χρήσης**, περιγράφονται αναλυτικά τα υποσυστήματα αυτά. Ο τρόπος περιγραφής που έχει επιλεγεί, περιλαμβάνει τρία βασικά τμήματα για κάθε υποσύστημα:

- Εισαγωγή: περιγράφεται γενικά το εργαλείο – υποσύστημα.
- Τεχνικά χαρακτηριστικά – εγκατάσταση: δίνονται τα τεχνικά χαρακτηριστικά του και ο τρόπος εγκατάστασης.
- Εγχειρίδιο χρήσης: Περιγράφεται ο τρόπος χρήσης του συστήματος.

Τέλος στο ίδιο παραδοτέο (Π4), περιγράφεται και ένα επιπλέον εργαλείο που αποφασίστηκε να υλοποιηθεί με βάση τις απαιτήσεις των ειδικών (ιατρών – γλωσσολόγων): Ένας **Συλλογέας Λεξιλογικών Συνάψεων** (Concordancer), που όταν χρησιμοποιηθεί με βάση τη συλλογή κειμένων, δίνει πολύ καίρια και σημαντικά στοιχεία στον επιστήμονα σε σχέση με τη λεξικογραφία ενός όρου.

Στο παρών, επισυνάπτεται οπτικός δίσκος (CD) στον οποίο περιέχεται το λογισμικό που αναφέρθηκε παραπάνω. Η διαδικασία εγκατάστασης και κάθε σχετική πληροφορία για το λογισμικό, περιλαμβάνεται στο παραδοτέο Π4.

2 ΣΥΜΠΕΡΑΣΜΑΤΑ

Το παραδοτέο αυτό (Π3) περιλαμβάνει το «Σύστημα Διαχείρισης Σώματος Κειμένων και Εξαγωγής Ορολογίας» του έργου ΙΑΤΡΟΛΕΞΗ. Το παραδοτέο, είναι αποτελέσματα της Ενότητας Εργασίας 2: *«Υλοποίηση Εργαλείων Ανάπτυξης Διαχείρισης Σώματος Κειμένων και Αυτόματης Εξαγωγής Ορολογίας»*.

Βασικό συμπέρασμα της ενότητας συνολικά, είναι ότι η υλοποίηση των εργαλείων αποδείχθηκε περισσότερο δύσκολη και χρονοβόρα από ότι είχε προβλεφθεί. Ένας λόγος είναι ίσως ο γενικός νόμος που ισχύει στα έργα ανάπτυξης λογισμικού και τα φέρνει να είναι σχεδόν πάντα καθυστερημένα σε σχέση με το αρχικό χρονοδιάγραμμα. Ένας δεύτερος λόγος είναι η απαίτηση της ανάπτυξης εργαλείων από δύο φορείς, τα οποία στη συνέχεια έπρεπε να συνεργάζονται και να λειτουργούν συνεκτικά. Τέλος ένας ακόμα λόγος είναι το γεγονός ότι κάποιες απαιτήσεις για το λογισμικό και τις λειτουργίες του, έρχονταν από τους γλωσσολόγους και τους ιατρούς της ομάδας εργασίας, μόνον όταν οι τελευταίο, έπαιρναν μια πρόγευση (χρησιμοποιούσαν) των εργαλείων.

Συμπερασματικά τα εργαλεία που υλοποιήθηκαν πέτυχαν το στόχο τους: να βοηθήσουν στην ανάπτυξη των υποδομών του έργου. Πιστεύουμε και ευελπιστούμε ότι θα ικανοποιήσουν και τον τελικό σκοπό τους: Την επίτευξη των τελικών στόχων του έργου!

3 ΒΙΒΛΙΟΓΡΑΦΙΑ

3.1 ΔΗΜΟΣΙΕΥΣΕΙΣ ΣΤΑ ΠΛΑΙΣΙΑ ΤΟΥ ΕΡΓΟΥ

Α. Βαγγελάτος, Γ. Ορφανός, Χ. Τσαλίδης, Χρ. Καλαμαρά: Ανάπτυξη Οντολογίας Βιοϊατρικών Όρων. 8ο Πανελλήνιο Επιστημονικό Συνέδριο Management Υπηρεσιών Υγείας και κοινωνικής Φροντίδας, Ερέτρια, Οκτώβριος 2006.

M. Pantazara, E. Mantzari, A. Vagelatos, Ch. Kalamara, A. Iordanidou: Development of a Greek biomedical corpus, 11th Panhellenic Conference on Informatics (PCI 2007), Patras, Greece, May 2007.

Ch. Tsalidis, E. Mantzari, M. Pantazara, Ch. Diolis, A. Vagelatos: Developing a Greek biomedical corpus for text mining, Accepted for presentation at: Corpus Linguistics Conference 2007 (CL 2007), Birmingham, UK, July 2007.

A. Vagelatos, E. Mantzari, G. Orphanos, Ch. Tsalidis, M. Pantazara, Ch. Kalamara, Ch. Diolis: Biomedical data mining for the Greek language, Accepted for presentation at: MEDNET 2007, Leipzig, Germany, October 2007.

Α. Ιορδανίδου, Μ. Πανταζάρα, Ε. Μάντζαρη, Α. Βαγγελάτος, Γ. Ορφανός, Β. Παπαπαναγιώτου: Ζητήματα αναγνώρισης των πολυλεκτικών σύνθετων όρων στον τομέα της βιοϊατρικής. Έγινε δεκτό για παρουσίαση στο: 6ο Συνέδριο «Ελληνική Γλώσσα και Ορολογία» (2007), Αθήνα, Νοέμβριος 2007.

3.2 ΣΧΕΤΙΚΟ ΛΟΓΙΣΜΙΚΟ

The source for Java developers: http://java.sun.com/

MySQL: The world's most popular open source database: http://www.mysql.com/

Open Source Initiative: http://www.opensource.org/

Ritmark FS (Filesystem Storage Engine for MySQL): http://www.ritmark.com