ΙΑΤΡΟΛΕΞΗ
Ανάπτυξη Υποδομής Γλωσσικής Τεχνολογίας για το Βιοϊατρικό Τομέα

Neurosoft A.E. --- EAITY

ΓΓΕΤ, ΚτΠ, Πρόγραμμα «ΕΠΕΞΕΡΓΑΣΙΑ ΕΙΚΟΝΩΝ, ΗΧΟΥ ΚΑΙ ΓΛΩΣΣΑΣ»

Σχεδιασμός & Συλλογή Σώματος Βιοϊατρικών Κειμένων (ΣΒΚ)

Ομιλητής: Χρ. Διολής

Αντικείμενο παρουσίασης

1. Εντοπισμός πηγών στο διαδίκτυο
2. Αξιολόγηση ιστοτόπων βάσει περιεχομένου
3. Συλλογή κειμένων μέσω Προσκομιστή εγγράφων
4. Κατηγοριοποίηση κειμένων
5. Εισαγωγή κειμένων σε μια Αποθήκη Εγγράφων

1. Εντοπισμός πηγών στο διαδίκτυο

Στόχος: Αναζήτηση ιστοτόπων στο διαδίκτυο που περιέχουν ελληνικά ιατρικά περιοδικά και άρθρα ελληνικών ιατρικών συνεδρίων

Επίπεδο 1 (Ερευνητικοί & Ακαδημαϊκοί ιστότοποι**)**

- Ελληνικό Ιατρικό Δίκτυο - MedNet Hellas (http://www.mednet.gr)
- Εθνικό Κέντρο Τεκμηρίωσης (http://antigua.ekt.gr/portal/dt)
- Βιβλιοθήκη Πανεπιστημίου Μακεδονίας (http://www.lib.uom.gr/index.php)

Επίπεδο 2 (Μηχανές Αναζήτησης)

- **Google**
- **Yahoo**

2. Αξιολόγηση ιστοτόπων βάσει περιεχομένου

Στόχος: Επιλογή κατάλληλων ιστοτόπων με βάση τη συγκέντρωση, κατηγοριοποίηση και δόμηση της πληροφορίας παρουσίασης των κειμένων.

Κριτήρια Επιλογής:

- Επιστημονικά περιοδικά, ενημερωτικά ιατρικά περιοδικά και άρθρα συνεδρίων
- Επεξεργάσιμη μορφή (html, pdf, txt)
- Θεματική κατηγοριοποίηση

Αποτελέσματα Αξιολόγησης ιστοτόπων

ΚΑΤΗΓΟΡΙΑ	ΙΣΤΟΤΟΠΟΙ
Επιστημονικά Περιοδικά	33
Ενημερωτικά Ιατρικά Περιοδικά	12
Ιατρικά Συνέδρια	2
Σύνολο	**47**

3. Συλλογή κειμένων μέσω Προσκομιστή εγγράφων

Στόχος: Σχεδιασμός και υλοποίηση ενός Προσκομιστή εγγράφων ειδικού σκοπού για την περισυλλογή των κειμένων από το σύνολο των προεπιλεγμένων ιστοτόπων.

Χαρακτηριστικά Προσκομιστή

- Διαχείριση ελληνικών κειμένων
- Είσοδος από XML αρχείο που περιέχει το σύνολο των ιστοτόπων
- Παραγωγή XML αρχείου με πληροφορίες της συγκεκριμένης πηγής
- Λαμβάνει υπόψη το Robots.txt της πηγής, αν υπάρχει

Αποτελέσματα Συλλογής Κειμένων

- Περιλήψεις Άρθρων
- Πρακτικά Συνεδρίων
- Πλήρη Άρθρα
- Κείμενα με πολλά άρθρα σε ένα αρχείο

ΣΥΝΟΛΟ ΚΕΙΜΕΝΩΝ	HTML	PDF
6.276 **(11,5 εκατ. λέξεις)**	4.380 / 69,8%	1.896 / 30,2%

4. Κατηγοριοποίηση κειμένων

Πρόβλημα: Τα κείμενα που βρέθηκαν στο διαδίκτυο δεν ήταν κατηγοριοποιημένα ως προς θεματική ενότητα

Αποτέλεσμα: Μη λειτουργικό Σώμα Κειμένων

Αντιμετώπιση: Κατηγοριοποίηση των κειμένων από ειδικούς (Ιατρούς) με βάση τις ιατρικές ειδικότητες

Θεματική Ενότητα (Ειδικότητα)	Κείμενα	Θεματική Ενότητα (Ειδικότητα)	Κείμενα
Αλλεργιολογία	4	Νευρολογία	78
Αναισθησιολογία	12	Νευροχειρουργική	104
Καρδιολογία	454	Οφθαλμολογία	137
Κυτταρολογία	6	Ορθοπεδική	162
Δερματολογία	1265	ΩΡΛ	231
Ενδοκρινολογία	29	Παθολογική Ανατομική	612
Ιατροδικαστική	2	Παιδιατρική	324
Γαστρεντερολογία	143	Πνευμονολογία	525
Γενική Ιατρική	4	Ψυχιατρική	26
Γενετική	4	Ιατρική Απεικόνιση	341
Γυναικολογία-Μαιευτική	403	Ρευματολογία	15
Αιματολογία	20	Κοινωνική Ιατρική	14
Ιατρικά Θέματα (Γενικά)	810	Χειρουργική	283
Μικροβιολογία	19	Ουρολογία	163
Νεφρολογία	14		

ΚτΠ, Μέτρο 3.3, Πρόγραμμα «ΕΠΕΞΕΡΓΑΣΙΑ ΕΙΚΟΝΩΝ, ΗΧΟΥ ΚΑΙ ΓΛΩΣΣΑΣ»

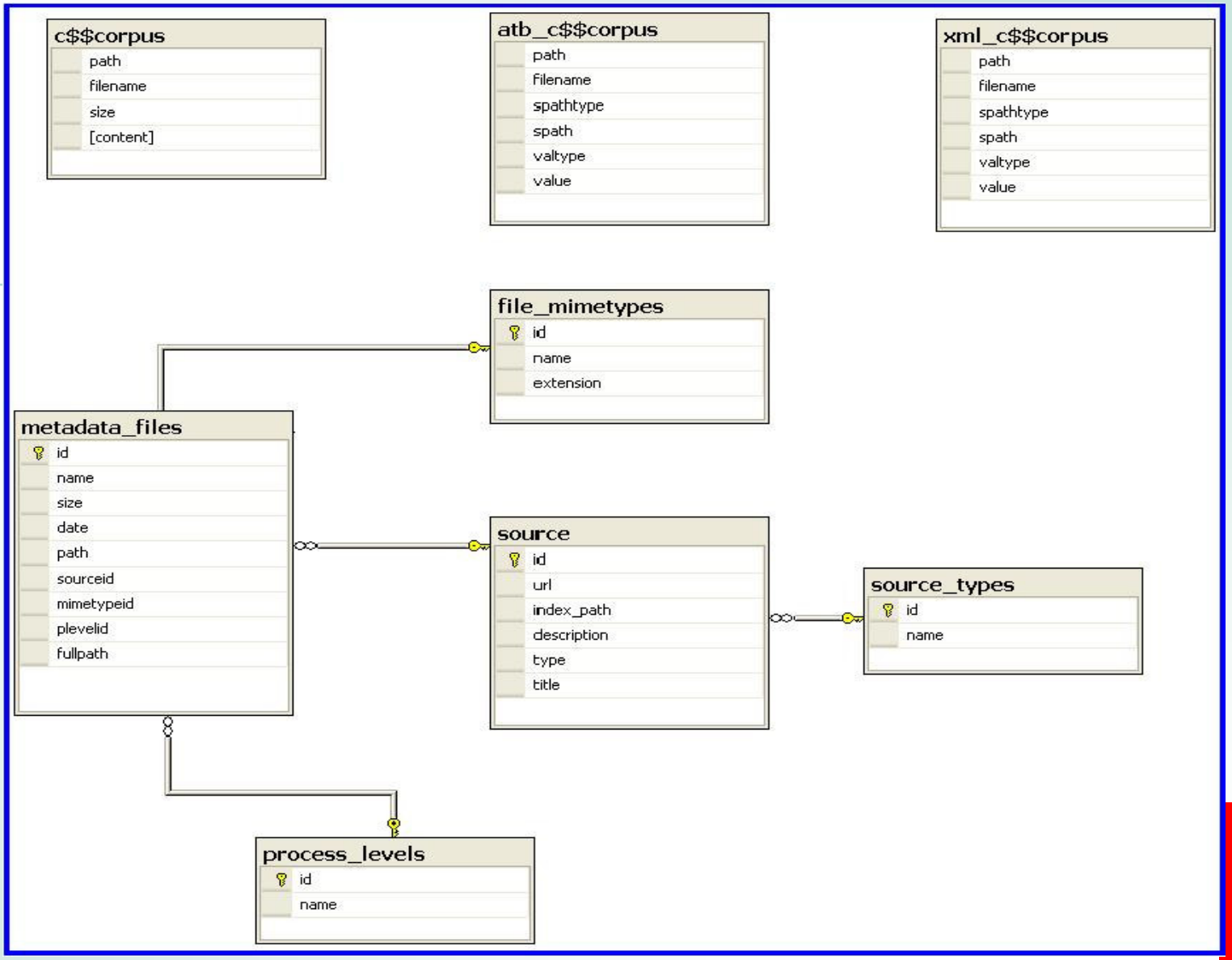
5. Εισαγωγή κειμένων σε μια Αποθήκη Εγγράφων

Στόχος: Εισαγωγή των κειμένων και των μεταδεδομένων τους σε μια αποθήκη εγγράφων

Αποθήκη Εγγράφων → Σύστημα Διαχείρισης Σχεσιακής Βάσης Δεδομένων

- Προσπέλαση δεδομένων από πολλούς χρήστες ταυτόχρονα
- Ανεξαρτησία λογικών δομών
- Διατήρηση εκδόσεων των δεδομένων

E.R.


c$$corpus
path
filename
size
[content]
atb_c$$corpus
path
filename
spathtype
spath
valtype
value
xml_c$$corpus
path
filename
spathtype
spath
valtype
value
file_mimetypes
id
name
extension
metadata_files
id
name
size
date
path
sourceid
mimetypeid
plevelid
fullpath
source
id
url
index_path
description
type
title
source_types
id
name
process_levels
id
name

Επόμενα Βήματα

- Μετατροπή Κειμένων (htm , pdf) σε απλό κείμενο (txt)
- Σχολιασμός Κειμένων
- Εύρεση Μονολεκτικών Βιοϊατρικών Όρων
- Εύρεση Πολυλεκτικών Βιοϊατρικών Όρων