Ταυτόχρονη αναζήτηση σε πολλαπλές πηγές δεδομένων με χρήση λογισμικού ανοιχτού κώδικα και εργαλείου εξαγωγής περιεχομένου από ιστοσελίδες

Κωνσταντίνος Ντονάς

Βιβλιοθήκη & Κέντρο Πληροφόρησης του Πανεπιστημίου Μακεδονίας

Γ' ΚΠΣ / ΕΠΕΑΕΚ ΙΙ / ΕΝΕΡΓΕΙΑ 2.1.3 δ
Συγχρηματοδότηση κατά 75% Ευρωπαϊκή Ένωση (ΕΚΤ) και 25% Εθνικοί πόροι (ΥΠΕΠΘ/ΕΥΔ ΕΠΕΑΕΚ)

Εισαγωγή (1/2)

- Τεράστιος και συνεχώς αυξανόμενος όγκος πληροφορίας στον Παγκόσμιο Ιστό.
- Βαθύς Ιστός (deep Web): μη ορατές σελίδες στις μηχανές αναζήτησης, κυρίως βάσεις δεδομένων.
- Οι μηχανές αναζήτησης δεν επαρκούν.
- Μεγάλος αριθμός ηλεκτρονικών πηγών.
- Απλότητα, ευκολία στη χρήση και ταχύτητα: Google!

Εισαγωγή (2/2)

- Προβληματική η παραδοσιακή πρακτική αναζήτησης στις πηγές που παρέχει μια βιβλιοθήκη.
  - Επαναλαμβανόμενη, χρονοβόρα και επίπονη προσπάθεια
- Κάθε βάση δεδομένων έχει διαφορετικά χαρακτηριστικά και επιλογές αναζήτησης.
  - Εξοικείωση με το περιβάλλον αναζήτησης κάθε πηγής
- Ταυτόχρονη αναζήτηση σε πολλαπλές πηγές δεδομένων από ενιαίο περιβάλλον (federated search).
  - Ταχύτητα
  - Ευκολία
  - One stop shopping

Ταυτόχρονη Αναζήτηση

- Πολύ δημοφιλής υπηρεσία μεταξύ των χρηστών.
- Παράκαμψη δυσκολιών που προκύπτουν από τον πολύ μεγάλο αριθμό διαθέσιμων πηγών.
- Διευκολύνει σημαντικά τον εντοπισμό χρήσιμων πληροφοριών.
- Φέρνει τους χρήστες πιο κοντά στις πηγές.

Ζητήματα και Προβληματισμός

- Μεγάλος όγκος πληροφορίας (information overload).
- Σχετικά αργή απόδοση.
- Σειρά εμφάνισης και σχετικότητα.
- Απουσία δυνατότητας απαλοιφής των διπλότυπων.
- Ποιότητα των επιστρεφόμενων αποτελεσμάτων.
- Πολυπλοκότητα και δυσκολία υλοποίησης.

Κριτική Ανάλυση

- Δεν υποκαθιστά:
  - δεξιότητες και ικανότητα εκτίμησης των αποτελεσμάτων.
  - τη διεπαφή αναζήτησης κάθε πηγής (native interface).
- Πολύ καλό σημείο εκκίνησης της έρευνας.
- Ποιότητα αποτελεσμάτων: συνάρτηση των όρων αναζήτησης. Επιστρέφει αρκετά καλά αποτελέσματα.
- Εξυπηρετεί κυρίως τους νέους χρήστες.
- Δεν αποτελεί πανάκεια.

Στοιχεία Έργου

- Έργο: «Πλοηγίς: Από την πληροφορία στη γνώση» (ΕΠΕΑΕΚ ΙΙ).
- Ανάπτυξη εργαλείου ταυτόχρονης αναζήτησης για:
  - ηλεκτρονικό κατάλογο της βιβλιοθήκης
  - σημαντικές βιβλιογραφικές βάσεις δεδομένων
  - συλλογές δεδομένων ανοιχτής ή μη πρόσβασης
  - άλλες πηγές πληροφοριών
- Έναρξη εργασιών υλοποίησης υπηρεσίας ταυτόχρονης αναζήτησης: Σεπτέμβριος 2006.
- Έπειτα από αναζήτηση σχετικών εργαλείων λογισμικού ανοικτού κώδικα (open source software tools), επιλέχθηκε το dbWiz.
- dbWiz: μέλος της οικογένειας προγραμμάτων reSearcher.

dbWiz

- Εργαλείο ταυτόχρονης αναζήτησης ανοικτού κώδικα.
- Προσπελάσιμο μέσω προγράμματος πλοήγησης Ιστού (web browser).
  - περιβάλλον αναζήτησης (βασική & σύνθετη)
  - περιβάλλον διαχείρισης (administration)
- Υλοποίηση στη γλώσσα προγραμματισμού Perl.
- Μηχανή παράλληλης αναζήτησης.
- Λειτουργικές μονάδες αναζήτησης (search modules).

Λειτουργικές Μονάδες Αναζήτησης

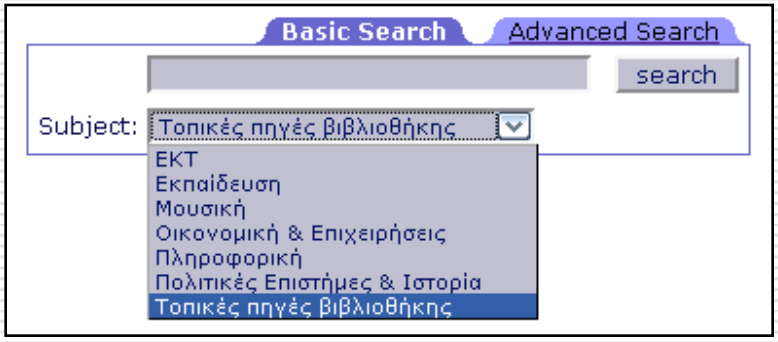
- Ενότητα κώδικα σε Perl που αναλαμβάνει την ανάκτηση των αποτελεσμάτων από μια πηγή.
- Το dbWiz διαθέτει μεγάλο αριθμό από έτοιμες προ-εγκατεστημένες λειτουργικές μονάδες.
- Είναι δυνατή η προσθήκη νέων πηγών.
- Η κατασκευή και η συντήρηση των λειτουργικών μονάδων αναζήτησης εμπεριέχουν δυσκολίες.

Νέες Λειτουργικές Μονάδες

- Ηλεκτρονικός κατάλογος (OPAC)
- Ιδρυματικό Αποθετήριο (ΨΗΦΙΔΑ)
- Πύλη Θεματικής Αναζήτησης (ΘΥΡΑ)
- Εθνικό Κέντρο Τεκμηρίωσης (ΕΚΤ)
- Silverplatter (EconLit, PsycINFO, INSPEC, κ.α.)

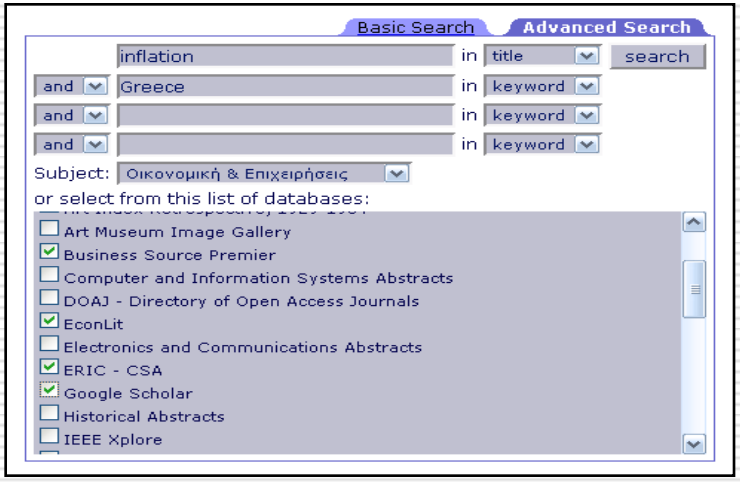
Βασική Αναζήτηση

- Αναζήτηση με λέξεις κλειδιά σε μία ομάδα προεπιλεγμένων πηγών.
- Δυνατότητα προβολής αποτελεσμάτων για κάθε μεμονωμένη πηγή.
- Επιστρέφει προκαθορισμένο πλήθος από τα πρώτα αποτελέσματα κάθε πηγής.



Σύνθετη Αναζήτηση

- Ο χρήστης μπορεί να δώσει συνδυασμό κριτηρίων και να ψάξει
  - είτε μια ομάδα προεπιλεγμένων πηγών
  - είτε μεμονωμένες πηγές
- Αναζήτηση μέχρι και με τέσσερις όρους συνδεδεμένους μεταξύ τους με λογικούς τελεστές (AND, OR, NOT).



Αποτελέσματα Αναζήτησης

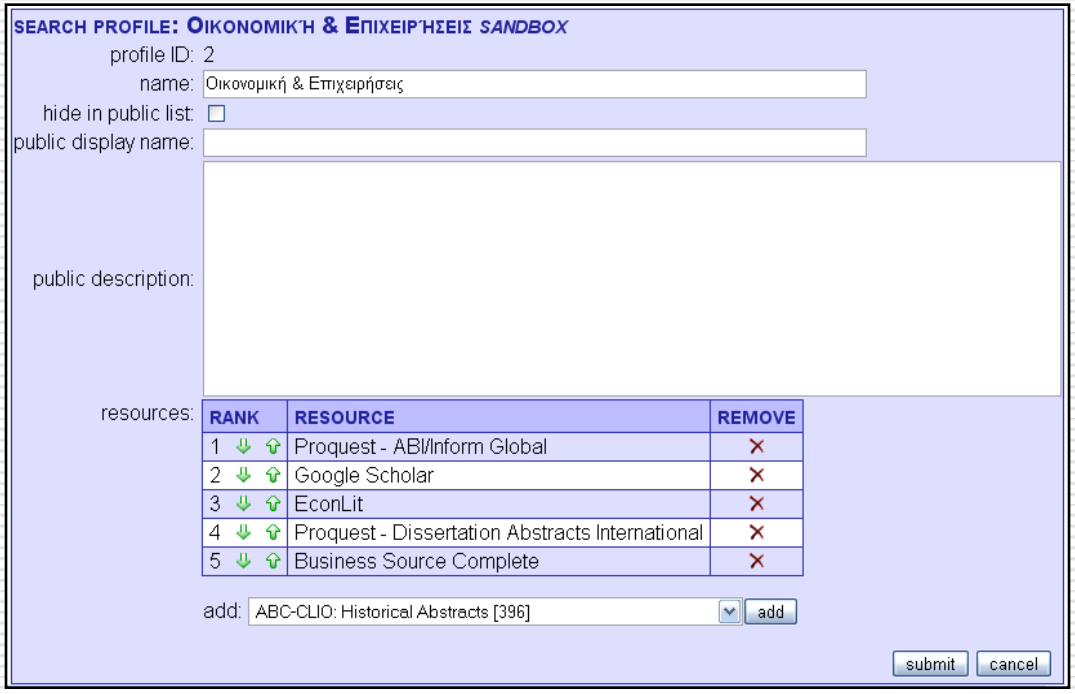
- Κάθε εγγραφή (record) εμφανίζει μία σύντομη αλλά περιεκτική ποσότητα πληροφορίας.
- Κάθε εγγραφή περιέχει:
  - υπερσύνδεσμο που δείχνει στην ιστοσελίδα με τη λεπτομερή περιγραφή
  - σύνδεσμο προς τη σελίδα του δικτυακού τόπου της πηγής από την οποία εξήχθη το αποτέλεσμα (native URL) και
  - openURL σύνδεσμο, κατάλληλο για τον link resolver της βιβλιοθήκης, για εντοπισμό του πλήρους κειμένου

Διαχείριση

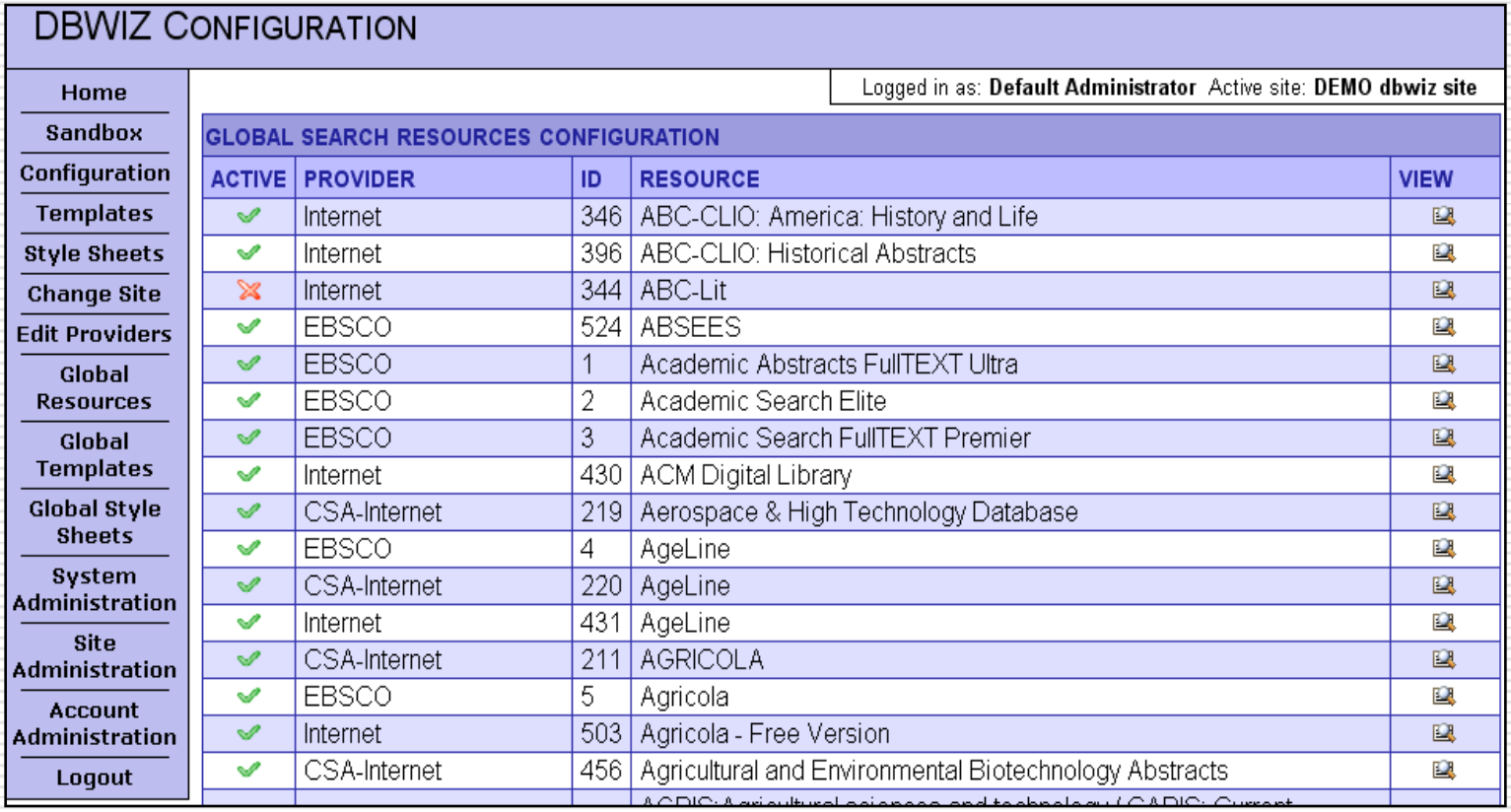
- Δημιουργία κατηγοριών αναζήτησης
- Προσθήκη / διαγραφή πηγών
- Παράμετροι και ρυθμίσεις
- Ενεργοποίηση πηγών
- Πρότυπα μορφοποίησης (templates)

Κατηγορίες Αναζήτησης

ID	PROFILE	ACTIVE	TRANSFER	SANDBOX
7	Οικονομική & Επιχειρήσεις			
8	Μουσική			
9	Εκπαίδευση			
10	Πολιτικές Επιστήμες & Ιστορία			
11	Πληροφορική			
13	Τοπικές πηγές βιβλιοθήκης			
15	ΕΚΤ			
add			submit	cancel



Διαθέσιμες Πηγές



Πρόσβαση στις Πηγές

- Διεπαφή ιστού (web interface)
  - προσομοίωση αλληλεπίδρασης χρήστη με το δικτυακό τόπο της εκάστοτε πηγής.
  - επιρρεπής σε αλλαγές, απαιτείται συντήρηση.

- Προγραμματιστική διεπαφή (API)
  - συνήθως μέσω του πρωτοκόλλου Z39.50 και του σχετικού API (ZOOM Perl module).
  - πιο αξιόπιστη μεθοδολογία.

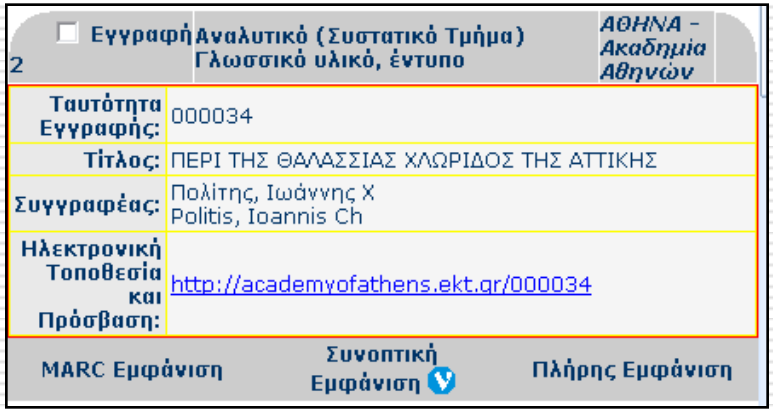
Πρόσβαση μέσω Διεπαφής Ιστού

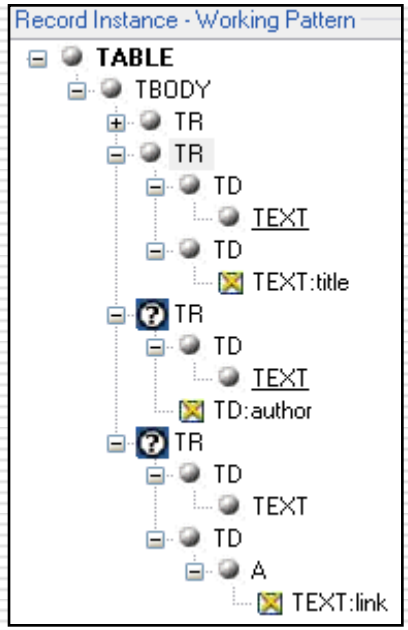
- Αυτόματη πλοήγηση στον δικτυακό τόπο της πηγής στο παρασκήνιο.
- Συμπλήρωση φόρμας αναζήτησης και υποβολή ερωτήματος χρήστη.
- Πλοήγηση στις σελίδες αποτελεσμάτων.
- Εντοπισμός και εξαγωγή επιστρεφόμενων τεκμηρίων σε κατάλληλη μορφή.

Εντοπισμός και Εξαγωγή Αποτελεσμάτων

- Κανόνας εξαγωγής (extraction rule): περιγράφει τη δομή των στιγμιότυπων επιθυμητής πληροφορίας.
- dbWiz: χρήση κανονικών εκφράσεων.
  - τυπική μέθοδος περιγραφής προτύπων κειμένου
  - χρονοβόρα και δύσκολη η κατασκευή τέτοιων κανόνων
- W3C DOM: παρέχει μία δενδροειδή αναπαράσταση των ιστοσελίδων.
- Αναπτύχθηκε εργαλείο λογισμικού για κατασκευή δενδροειδών κανόνων εξαγωγής.

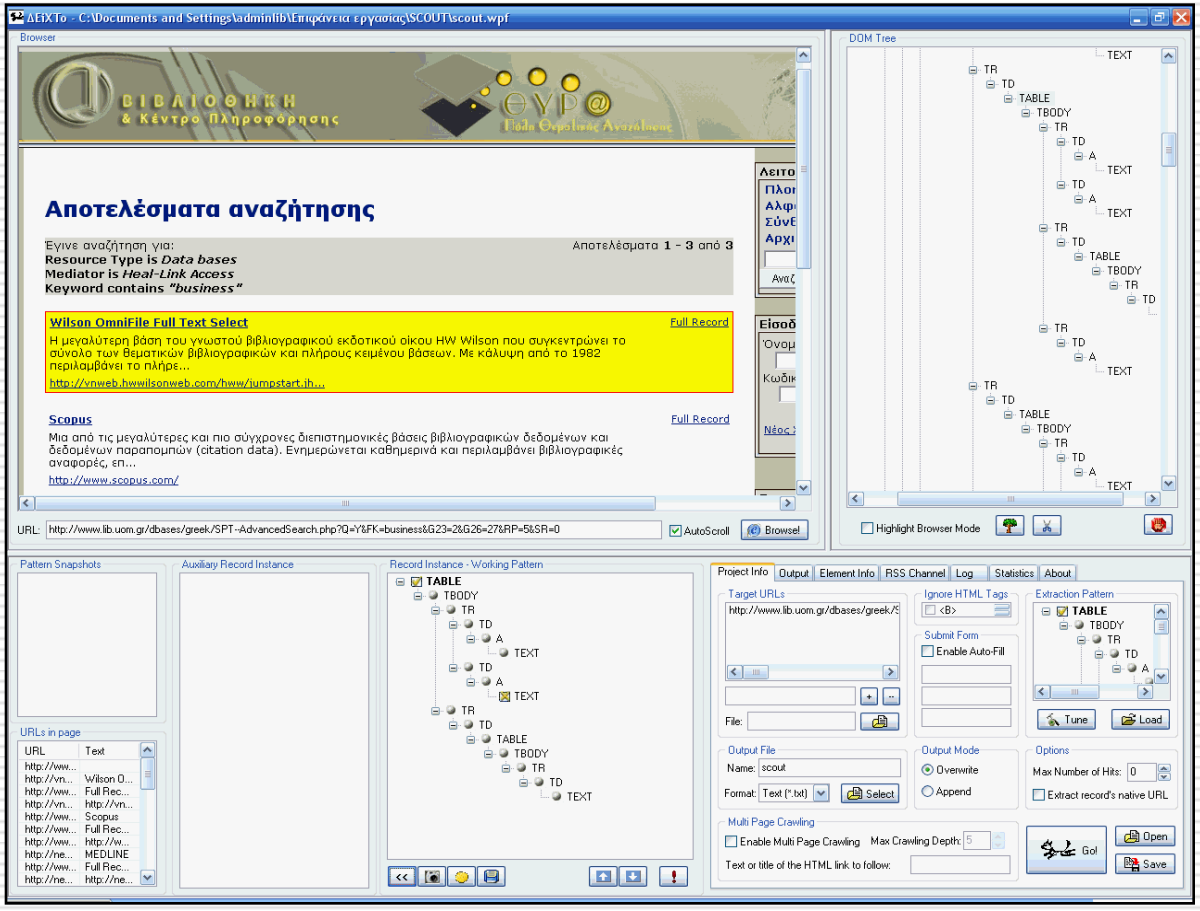
Τυπική Εγγραφή






```
<table …> <tr> <td …> Ταυτότητα Εγγραφής: </td>
<td …>000034</td> </tr> <tr> <td …> Τίτλος:
</td> <td …>ΠΕΡΙ ΤΗΣ ΘΑΛΑΣΣΙΑΣ ΧΛΩΡΙΔΟΣ ΤΗΣ
ΑΤΤΙΚΗΣ </td> </tr> <tr> <td …> Συγγραφέας:
</td> <td …>Πολίτης, Ιωάννης Χ <br> Politis,
Ioannis Ch </td> </tr> <tr> <td …> Ηλεκτρονική
Τοποθεσία και Πρόσβαση: </td> <td …><a
…>http://academyofathens.ekt.gr/000034</a> </td>
</tr> </table>
```

Εργαλείο Εξαγωγής Περιεχομένου (1/3)



Εργαλείο Εξαγωγής Περιεχομένου (2/3)

- Εύχρηστο γραφικό περιβάλλον χρήστη (GUI).
- Υπόδειξη του προτύπου μέσω mouse εντός του ενσωματωμένου προγράμματος πλοήγησης.
- Λειτουργία επισήμανσης στοιχείου (highlight mode).
- Ανάθεση σημασιολογικών ετικετών (labels).
- Αλγόριθμος συμφωνίας προτύπου: υψηλή ακρίβεια (precision) και σχετικότητα (recall).

Εργαλείο Εξαγωγής Περιεχομένου (3/3)

- Αποθήκευση κανόνα σε XML αρχείο.
- Perl λειτουργική μονάδα με συναρτήσεις για εκτέλεση δενδροειδών κανόνων μέσα στο dbWiz.
- Δουλεύει ήδη με επιτυχία για πηγές του ΕΚΤ και τη ΘΥΡΑ.

Συμπεράσματα

- Ταυτόχρονη αναζήτηση: ισχυρό εργαλείο αναζήτησης.
- dbWiz: πολύ αξιόλογο λογισμικό ανοικτού κώδικα.
- Κατασκευάστηκαν αρκετές νέες λειτουργικές μονάδες αναζήτησης καθώς και έχουν τροποποιηθεί κατάλληλα κάποιες από τις έτοιμες.
- Τα πρώτα δείγματα γραφής του συστήματος είναι ενθαρρυντικά και η απόδοση κρίνεται ικανοποιητική.

Προκλήσεις

- Κατασκευή νέων λειτουργικών μονάδων αναζήτησης και συντήρηση των υπαρχόντων.
- Εκπαίδευση των χρηστών.
- Προώθηση της χρήσης της υπηρεσίας.
- Συνεργατική λειτουργία με άλλα συστήματα λογισμικού.
- Μοντέλο ανοικτού λογισμικού.