

# 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



ΥΠΟΥΡΓΕΙΟ ΕΘΝΙΚΗΣ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ  
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΕΑΕΚ



ΕΥΡΩΠΑΪΚΗ ΕΝΩΣΗ  
ΣΥΓΧΡΗΜΑΤΟΔΟΤΗΣΗ  
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ



Η ΠΑΙΔΕΙΑ ΣΤΗΝ ΚΟΡΥΦΗ  
Επιχειρησιακό Πρόγραμμα  
Εκπαίδευσης και Αρχικής  
Επαγγελματικής Κατάρτισης

Γ' ΚΠΣ / ΕΠΕΑΕΚ II / 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: «Πλοηγός: Από την πληροφορία στη γνώση» (ΕΠΕΑΕΚ II).
- 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.

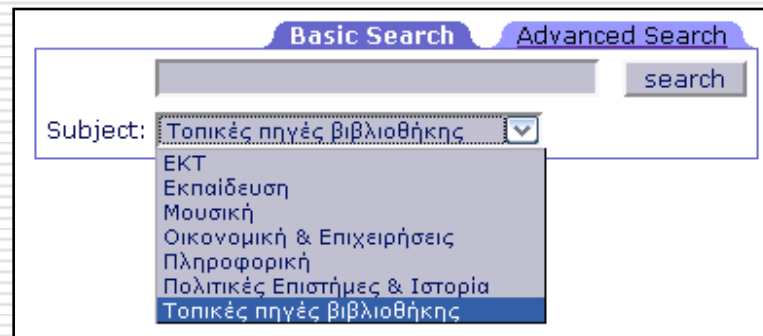
# Νέες Λειτουργικές Μονάδες

---

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

# Βασική Αναζήτηση

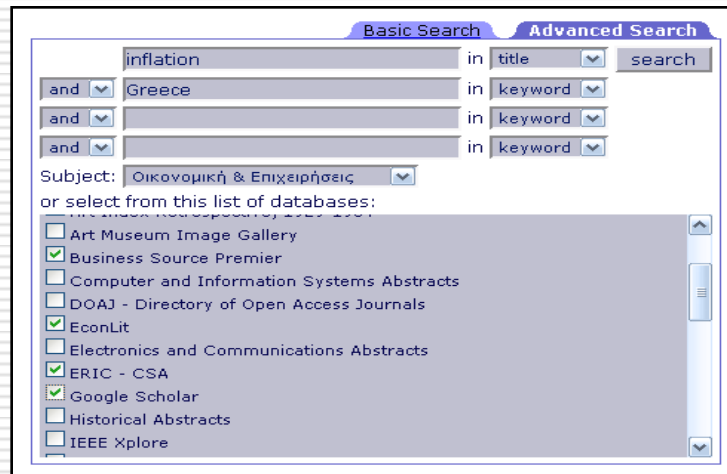
- Αναζήτηση με λέξεις κλειδιά σε μία ομάδα προεπιλεγμένων πηγών.
- Δυνατότητα προβολής αποτελεσμάτων για κάθε μεμονωμένη πηγή.
- Επιστρέφει προκαθορισμένο πλήθος από τα πρώτα αποτελέσματα κάθε πηγής.



The screenshot displays a search interface with two tabs: "Basic Search" (selected) and "Advanced Search". Below the tabs is a search input field with a "search" button. Underneath the input field, there is a "Subject:" label followed by a dropdown menu. The dropdown menu is open, showing a list of subject categories: "Τοπικές πηγές βιβλιοθήκης", "ΕΚΤ", "Εκπαίδευση", "Μουσική", "Οικονομική & Επιχειρήσεις", "Πληροφορική", "Πολιτικές Επιστήμες & Ιστορία", and "Τοπικές πηγές βιβλιοθήκης". The last option is highlighted in blue.

# 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).



The screenshot shows a search interface with two tabs: "Basic Search" and "Advanced Search". The "Basic Search" tab is active. The search query is "inflation" in the "title" field, "and" operator, "Greece" in the "keyword" field, "and" operator, and another empty "keyword" field. The "Subject" is set to "Οικονομική & Επιχειρήσεις". Below the search fields, there is a section titled "or select from this list of databases:" with a list of databases and checkboxes:

- Art Museum Image Gallery
- Business Source Premier
- Computer and Information Systems Abstracts
- DOAJ - Directory of Open Access Journals
- EconLit
- Electronics and Communications Abstracts
- ERIC - CSA
- Google Scholar
- Historical Abstracts
- IEEE Xplore

# 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

[Expensive Living: The Greek Experience under the Euro.](#)  
Pelagidis, Theodore | *May2007*  
Intereconomics : vol. 42 - issue 3 - pp. 167-176  
[Where can I get this?](#)  
([search Business Source Premier](#))

# 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	X		X
8	Μουσική	X		X
9	Εκπαίδευση	X		X
10	Πολιτικές Επιστήμες & Ιστορία	X		X
11	Πληροφορική	X		X
13	Τοπικές πηγές βιβλιοθήκης	X		X
15	ΕΚΤ	X		X
add	<input type="text"/>		<input type="button" value="submit"/>	<input type="button" value="cancel"/>

**SEARCH PROFILE: ΟΙΚΟΝΟΜΙΚΗ & ΕΠΙΧΕΙΡΗΣΕΙΣ SANDBOX**

profile ID: 2

name:

hide in public list:

public display name:

public description:

resources:

RANK	RESOURCE	REMOVE
1	Proquest - ABI/Inform Global	X
2	Google Scholar	X
3	EconLit	X
4	Proquest - Dissertation Abstracts International	X
5	Business Source Complete	X

add:

# Διαθέσιμες Πηγές

DBWIZ CONFIGURATION					
					Logged in as: <b>Default Administrator</b> Active site: <b>DEMO dbwiz site</b>
<b>GLOBAL SEARCH RESOURCES CONFIGURATION</b>					
	ACTIVE	PROVIDER	ID	RESOURCE	VIEW
Home	✓	Internet	346	ABC-CLIO: America: History and Life	
Sandbox	✓	Internet	396	ABC-CLIO: Historical Abstracts	
Configuration	✗	Internet	344	ABC-Lit	
Templates	✓	EBSCO	524	ABSEES	
Style Sheets	✓	EBSCO	1	Academic Abstracts FullTEXT Ultra	
Change Site	✓	EBSCO	2	Academic Search Elite	
Edit Providers	✓	EBSCO	3	Academic Search FullTEXT Premier	
Global Resources	✓	Internet	430	ACM Digital Library	
Global Templates	✓	CSA-Internet	219	Aerospace & High Technology Database	
Global Style Sheets	✓	EBSCO	4	AgeLine	
System Administration	✓	CSA-Internet	220	AgeLine	
Site Administration	✓	Internet	431	AgeLine	
Account Administration	✓	CSA-Internet	211	AGRICOLA	
Logout	✓	EBSCO	5	Agricola	
	✓	Internet	503	Agricola - Free Version	
	✓	CSA-Internet	456	Agricultural and Environmental Biotechnology Abstracts	
				AGRIC: Agricultural science and technology (CAPIC: Current	



# 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.

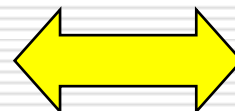
# Εντοπισμός και Εξαγωγή Αποτελεσμάτων

---

- 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.

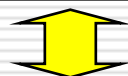
# Τυπική Εγγραφή

<input type="checkbox"/> Εγγραφή Αναλυτικό (Συστατικό Τμήμα) 2 Γλωσσικό υλικό, έντυπο		ΑΘΗΝΑ - Ακαδημία Αθηνών
Ταυτότητα Εγγραφής:	000034	
Τίτλος:	ΠΕΡΙ ΤΗΣ ΘΑΛΑΣΣΙΑΣ ΧΛΩΡΙΔΟΣ ΤΗΣ ΑΤΤΙΚΗΣ	
Συγγραφέας:	Πολίτης, Ιωάννης Χ Politis, Ioannis Ch	
Ηλεκτρονική Τοποθεσία και Πρόσβαση:	<a href="http://academyofathens.ekt.gr/000034">http://academyofathens.ekt.gr/000034</a>	
MARC Εμφάνιση	Συνοπτική Εμφάνιση	Πλήρης Εμφάνιση



Record Instance - Working Pattern

TABLE
TBODY
TR
TR
TD
TEXT
TD
TEXT:title
TR
TD
TEXT
TD:author
TR
TD
TEXT
TD
A
TEXT:link



```
<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)

The screenshot displays the SCOUT web crawler interface. The top window shows a browser view of a search results page from the National Library of Greece. The page title is "Αποτελέσματα αναζήτησης" (Search Results). The search criteria are: "Resource Type is Data bases", "Mediator is Heal-Link Access", and "Keyword contains 'business'". The results show two entries: "Wilson OmniFile Full Text Select" and "Scopus".

The right window shows the DOM Tree of the page, highlighting the structure of the search results table. The tree shows a TABLE element containing a TBODY with multiple TR elements, each containing TD and A (Anchor) elements.

The bottom window shows the SCOUT configuration panel. It includes sections for "Pattern Snapshots", "Auxiliary Record Instance", and "Record Instance - Working Pattern". The "Record Instance" section shows a tree view of the selected table structure. The "Target URLs" section shows the current URL: "http://www.lib.uom.gr/dbases/greek/SPT-AdvancedSearch.php?Q=Y&FK=business&G23=2&G26=27&RP=5&SR=0". The "Extraction Pattern" section shows a tree view of the selected table structure. The "Output File" section shows the output file name "scout" and the format "Text (\*.txt)". The "Options" section includes "Max Number of Hits: 0" and "Extract record's native URL".

## 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 δένδροειδών 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.