

# Archiving the Web sites of Athens University of Economics and Business

Βασίλης Πλαχούρας

Χρυσόστομος Καπέτης

Μιχάλης Βαζιργιάννης



Οικονομικό Πανεπιστήμιο Αθηνών

3/11/2003

# Κίνητρα & Στόχοι εργασίας

- ▶ Απώλεια της πληροφορίας από τους ιστοτόπους του πανεπιστημίου
  - Ανάγκη για μακροπρόθεσμη διατήρηση
  - Προστασία της φήμης του ιδρύματος
- ▶ Απουσία δραστηριοτήτων αρχειοθέτησης ιστοπεριεχομένου στον ελληνικό χώρο
- ▶ Αρχειοθέτηση ιστοπεριεχομένου από τους ιστοτόπους του ΟΠΑ
  - Απαιτήσεις σε υλικό
  - Ανάλυση των δεδομένων
  - Συσχέτιση με τους στόχους της βιβλιοθήκης

# Περιγραφή παρουσίασης

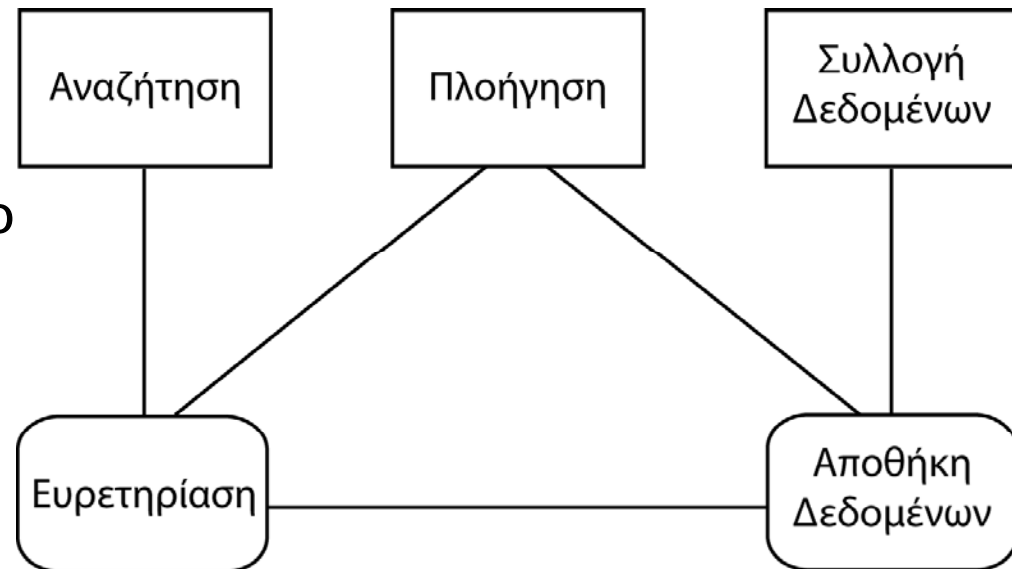
- ▶ Αρχαιοθήκη ιστοπεριεχομένου
- ▶ Παρουσίαση συστήματος
- ▶ Χαρακτηριστικά δεδομένων
- ▶ Ρόλος της βιβλιοθήκης
- ▶ Επεκτάσεις και συμπεράσματα

# Αρχειοθέτηση Ιστοπεριεχομένου

- ▶ Απώλεια περιεχομένου από ιστότοπους
  - Αλλαγές σε ιστοσελίδες
  - Διακοπή συντήρησης ιστότοπων
  - Αστοχία υλικού
- ▶ Αναγκαιότητα για αρχειοθέτηση ιστοπεριεχομένου
- ▶ Πρωτοβουλίες από φορείς σε διαφορετικά επίπεδα
  - Internet Archive, μη-κερδοσκοπικός οργανισμός
  - Εθνικές βιβλιοθήκες
  - Μεμονωμένοι οργανισμοί

# Παρουσίαση συστήματος

- ▶ Βασισμένο σε ελεύθερο λογισμικό – λογισμικό ανοικτού κώδικα
- ▶ 3 υπηρεσίες
  - Αναζήτηση με βάση το URL των ιστοσελίδων
  - Αναζήτηση με λέξεις κλειδιά
  - Πλοήγηση στις αρχειοθετημένες ιστοσελίδες
- ▶ <http://archive.aueb.gr>



# Συλλογή δεδομένων

- ▶ Χρησιμοποιεί το λογισμικό Heritrix
  - Crawler σχεδιασμένος απο το Internet Archive για την αρχειοθέτηση ιστοπεριεχομένου
- ▶ Αποθήκευση δεδομένων σύμφωνα με το πρότυπο WARC (ISO 88500 2009)
  - Συμπιεσμένα αρχεία με πολλαπλές εγγραφές
  - Αποθήκευση όλων των τύπων αρχείων
- ▶ Συλλογή δεδομένων με εκκίνηση από 82 URLs
- ▶ Σεβασμός στην ομαλή λειτουργία των εξυπηρετητών.
  - Ανάκτηση ενός URL
    - ανά 10 δευτερόλεπτα από τον ίδιο εξυπηρετητή
    - με αναμονή 10 φορές το χρόνο που χρειάστηκε η τελευταία ανάκτηση

# Πλοήγηση & αναζήτηση με βάση το URL

- ▶ Δημιουργία ευρετηρίου με βάση το URL και την ημερομηνία συλλογής κάθε URL
  - Βασισμένο στο λογισμικό Wayback Machine
- ▶ Ερωτήσεις με χρονικό περιορισμό

**AUEB Web Archive**  
Experimental

Type text  Search Type URL  Take me back

From date  To date

Searched for <http://www.db-net.aueb.gr/>

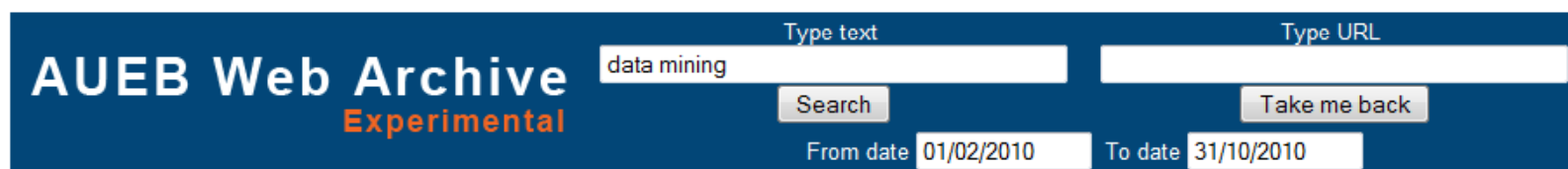
Φεβ 2010	Μαρ 2010	Απρ 2010	Μαΐ 2010	Ιουν 2010	Ιουλ 2010	Αυγ 2010
2 pages	3 pages	0 pages	1 page	2 pages	2 pages	1 page
<a href="#">Φεβ 19, 2010</a> * <a href="#">Φεβ 26, 2010</a>	<a href="#">Μαρ 5, 2010</a> <a href="#">Μαρ 12, 2010</a> <a href="#">Μαρ 20, 2010</a>		<a href="#">Μαΐ 13, 2010</a> *	<a href="#">Ιουν 11, 2010</a> <a href="#">Ιουν 24, 2010</a>	<a href="#">Ιουλ 2, 2010</a> <a href="#">Ιουλ 29, 2010</a>	<a href="#">Αυγ 6, 2010</a>

Powered by 



# Αναζήτηση με λέξεις κλειδιά

- ▶ Αναζήτηση στο πλήρες κείμενο των αρχειοθετημένων ιστοσελίδων
  - Βασισμένο στο λογισμικό NutchWax



AUEB Web Archive  
Experimental

Type text: data mining  
Type URL:

Search  
Take me back

From date: 01/02/2010  
To date: 31/10/2010

Search took 0.441 seconds. Hits 1-10 (out of about 943 total matching pages):

## [DB-NET - Research team on Data & Web Mining](#)

... DB-NET - Research team on Data & Web Mining > DB-NET Courses Diploma theses External Services Jobs - Opportunities News ...  
addresses the whole life cycle of data ...  
<http://www.db-net.aueb.gr/> [html] (16796 bytes) - 2010-02-19 - [other versions](#) - [more from www.db-net.aueb.gr](#)

## [eClass του Οικονομικού Πανεπιστημίου Αθηνών | ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΒΑΣΕΙΣ ΔΕΛΔΟΜΕΝΩΝ ΚΑΙ ΤΟΝ ΠΑΓΚΟΣΜΙΟ](#)

... Quality Assessment and Uncertainty Handling in Data Mining", Springer Verlag, LNAI Series, 2003. - Tom Mitchell. "Machine Learning", McGraw  
... Γιατσιδης Π. Μαγδαληνος (pmagdal@aub.gr) X. Γιατσιδης Λέξεις Κλειδιά: ...  
<http://eclass.aueb.gr/courses/INF131/> [html] (11298 bytes) - 2010-02-19 - [other versions](#) - [more from eclass.aueb.gr](#)



# Χαρακτηριστικά δεδομένων (1)

## ▶ 4 συλλογές δεδομένων

Crawl	C1	C2	C3	C4
Started at	2010-02-26	2010-03-20	2010-04-26	2010-05-13
Not Fetched	2882	2684	17097	8661
2xx (Successful)	139212	96495	113811	141084
3xx (Redirection)	3200	2364	3020	3135
4xx (Client error)	10867	8248	10141	10808
5xx (Server Error)	24	24	15	16

## ▶ Πιο συχνοί τύποι αρχείων

- HTML, JPEG, GIF, PDF
- Αντιστοιχούν σε περισσότερα από 88% των URLs
- Παρόμοια κατανομή στους ιστότοπους των ΕΚΠΑ, ΕΜΠ

## Χαρακτηριστικά δεδομένων (2)

- ▶ Συλλογή δεδομένων για τους ιστότοπους που δεν υπήρξε αλλαγή στις ρυθμίσεις

Crawl	C1	C2	C3	C4
Started at	2010-02-26	2010-03-20	2010-04-26	2010-05-13
Not Fetched	1730	1604	1786	1895
2xx (Successful)	68825	67826	67768	66497
3xx (Redirection)	3048	2241	2775	2887
4xx (Client error)	7572	7238	7358	6504
5xx (Server Error)	25	25	15	14

# Απαιτήσεις σε αποθηκευτικό χώρο

- ▶ Δεδομένα από το δίκτυο:
  - μεταξύ 10 και 15GB
- ▶ Αποθήκευση σε συμπιεσμένη μορφή:
  - μεταξύ 8 και 10GB
- ▶ Αποθήκευση των URLs που αλλάζουν μόνο:
  - λιγότερα από 2GB

# Μεταβολές στο Ιστοπεριεχόμενο

- ▶ 94% των αλλαγών αφορούν δυναμικές HTML σελίδες στο C1/C2

Crawls	C1/C2	C2/C3	C3/C4
All Web sites			
URL in $C_i \setminus C_{i+1}$	50747	6017	13990
URL in $C_{i+1} \setminus C_i$	8030	23333	41263
URL in $C_i \cap C_{i+1}$	88465	90578	99821
URL same	56749	63865	59805
URL changes	31716	26613	40016
Excluding reconfigured Web sites			
URL in $C_i \setminus C_{i+1}$	3848	3847	5347
URL in $C_{i+1} \setminus C_i$	2849	3429	4076
URL in $C_i \cap C_{i+1}$	64977	64339	62421
URL same	54691	55127	53193
URL changes	10286	9212	9228

# Ρόλος της Βιβλιοθήκης

- ▶ Στόχοι της βιβλιοθήκης
  - Αρχαιοθήκη και διατήρηση ιστοπεριεχομένου του πανεπιστημίου
  - Ολοκλήρωση με το ψηφιακό αποθετήριο της βιβλιοθήκης
- ▶ Θεματικές συλλογές για την πανεπιστημιακή κοινότητα
- ▶ Ζητήματα που προκύπτουν
  - Πνευματικά δικαιώματα
  - Υλικοτεχνική υποδομή
  - Πολιτικές και στρατηγικές σχετικά με
    - Επιλογή πληροφορίας
    - Υιοθέτηση προτύπων για μεταδεδομένα και δεικτοδότηση

# ΕΠΕΚΤΑΣΕΙΣ

- ▶ Βελτιστοποίηση συλλογής δεδομένων (crawling)
  - Αυξημένη συχνότητα
  - Δείκτες ποιότητας/πληρότητας δεδομένων
- ▶ Συλλογή δεδομένων μέσω φορμών (hidden Web)
- ▶ Συμπύεση αποθηκευμένης πληροφορίας
- ▶ Ανάπτυξη επιπλέον υπηρεσιών



# Συμπεράσματα

- ▶ Η αρχειοθέτηση ιστοπεριεχομένου στο ΟΠΑ αποτελεί μια βιώσιμη διαδικασία
  - Περιορισμένες υλικοτεχνικές απαιτήσεις
- ▶ Η αρχειοθέτηση ιστοπεριεχομένου ως στόχος της βιβλιοθήκης του πανεπιστημίου
  - Μακροπρόθεσμη πρόσβαση στην πληροφορία
- ▶ Ανάλογα μεγέθη για τους ιστότοπους άλλων ιδρυμάτων

# Ερωτήσεις