Ενισχύοντας **σημασιολογικά** τις **διαδικασίες αναζήτησης** σε ένα περιβάλλον **μετα-αναζήτησης**

Μιχάλης Σφακάκης & Σαράντος Καπιδάκης

Εργαστήριο Ψηφιακών Βιβλιοθηκών και Ηλεκτρονικής Δημοσίευσης
Τμήμα Αρχειονομίας - Βιβλιοθηκονομίας
Ιόνιο Πανεπιστήμιο, Κέρκυρα
{sfakakis, sarantos}@ionio.gr

16ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών
1-3 Οκτωβρίου 2007, Πειραιάς

Περιβάλλον Μετα-αναζήτησης: Χαρακτηριστικά

- Κεντρικό σημείο ομοιόμορφης αναζήτησης σε διαφορετικά περιβάλλοντα αναζήτησης
  - Δε συνθέτουν κεντρική βάση δεδομένων
  - Κατανεμημένα, μεταδίδουν την επερώτηση στα επιμέρους περιβάλλοντα και συνθέτουν τα αποτελέσματα από αυτά σε ενιαία μορφή
  - Είναι ετερογενή τόσο ως προς τα συστήματα αναζήτησης, όσο και ως προς το περιεχόμενο που διαθέτουν
  - Ποικίλουν στις παρεχόμενες δυνατότητες αναζήτησης

Περιβάλλον Μετα-αναζήτησης: Βιβλιοθήκες

- Κυρίως αναζητήσεις σε
  - Καταλόγους βιβλιοθηκών μέσω του Z39.50 πρωτοκόλλου
  - Συστήματα διάθεσης ηλεκτρονικών βιβλίων και περιοδικών
  - Βιβλιογραφικές βάσεις και βάσεις πλήρους κειμένου
  - Τυπικές μηχανές αναζήτησης (Google, AltaVista, Yahoo, κλπ.)
  - Αποθετήρια
  - ...

Περιβάλλον Μετα-αναζήτησης: Απόδοση, αποτελεσματικότητα

- Καθορίζονται από
  - Τη δυνατότητα αναζήτησης σε πολλές πηγές ταυτόχρονα
  - Την ποιότητα αναζήτησης (γενική / εξειδικευμένη)
  - Πόσο κοντά θα είναι η μεταγραφή της επερώτησης σε αυτή που υποστηρίζει κάθε επιμέρους περιβάλλον
  - Τη δυνατότητα σύνθεσης των αποτελεσμάτων (όπως ταύτιση – ενοποίηση όμοιων αποτελεσμάτων και ενιαία εμφάνιση)

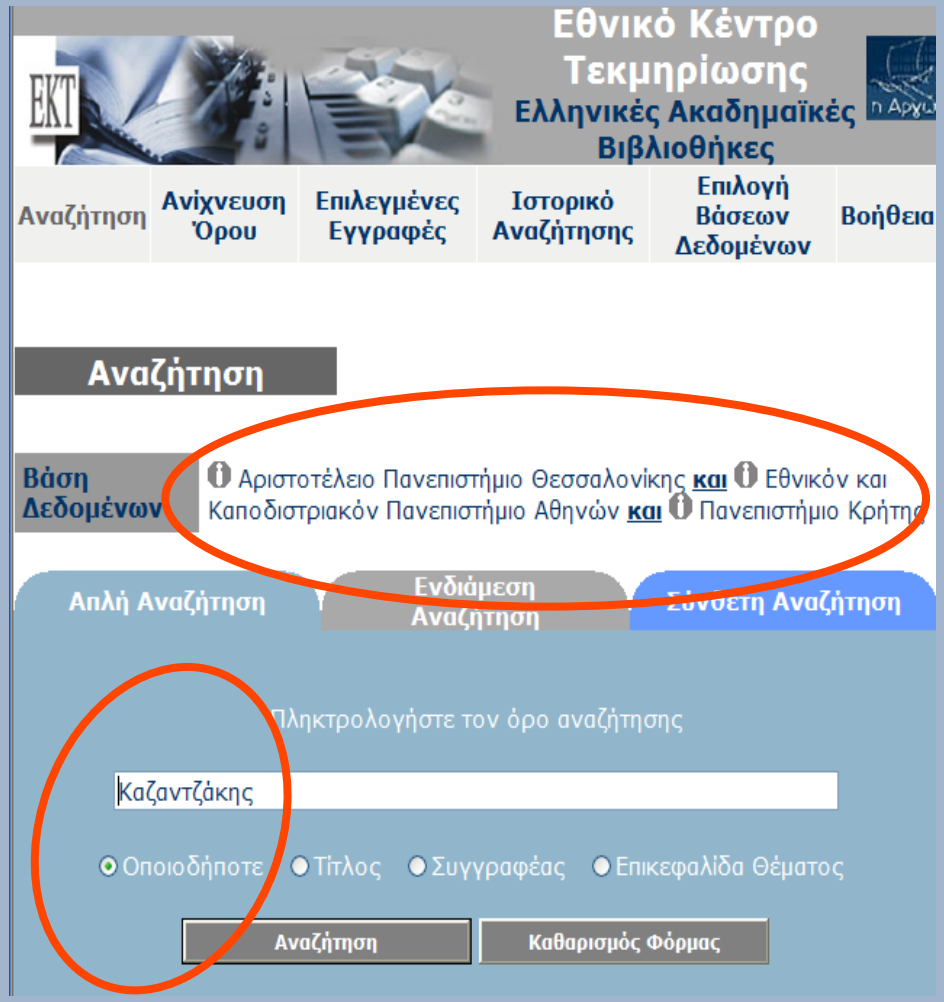
- Περιορίζονται από τις δυνατότητες κάθε περιβάλλοντος

Περιβάλλον Μετα-αναζήτησης: Z39.50 πηγές

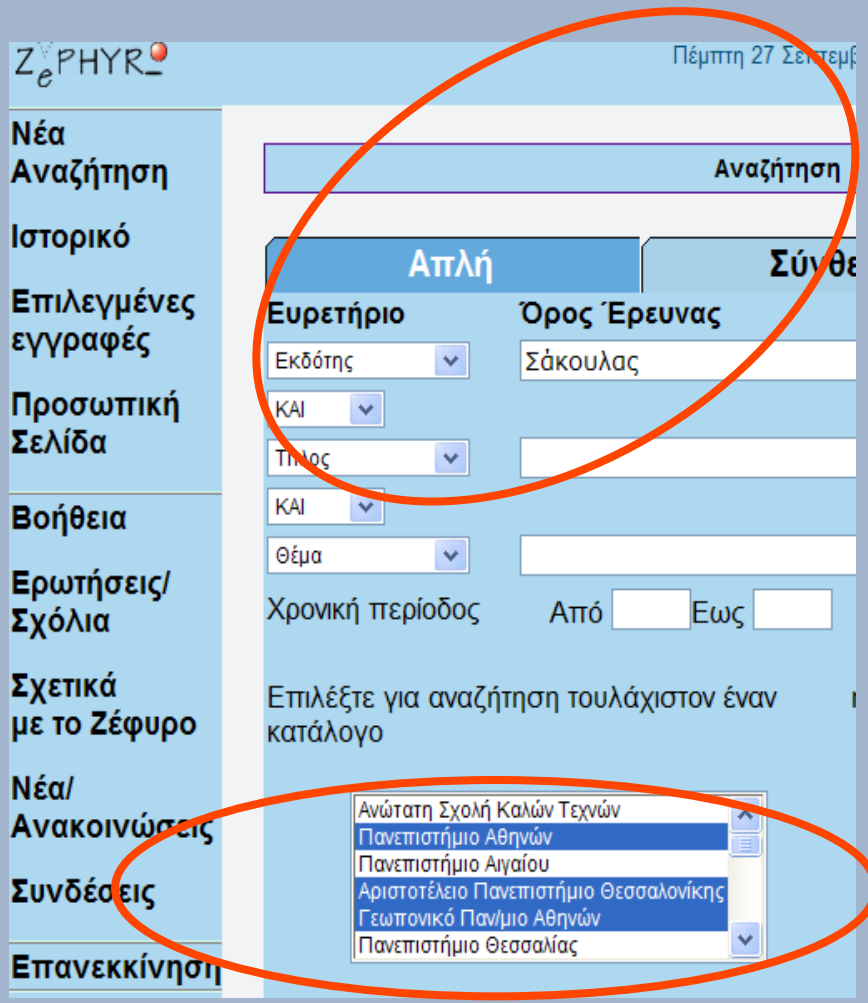
- Είναι δεδομένη η εφαρμογή του ως σύστημα αναζήτησης από τις βιβλιοθήκες
- Είναι πρωτόκολλο αναζήτησης – ανάκτησης πληροφοριών με σαφώς ορισμένες διαδικασίες
- Υπάρχει πολύ μεγάλος αριθμός διαθέσιμων πηγών (κατάλογοι βιβλιοθηκών, βιβλιογραφικών βάσεων δεδομένων, κλπ.)
- Στα περιβάλλοντα μετα-αναζήτησης των βιβλιοθηκών οι Z39.50 πηγές αποτελούν τεράστια ενότητα

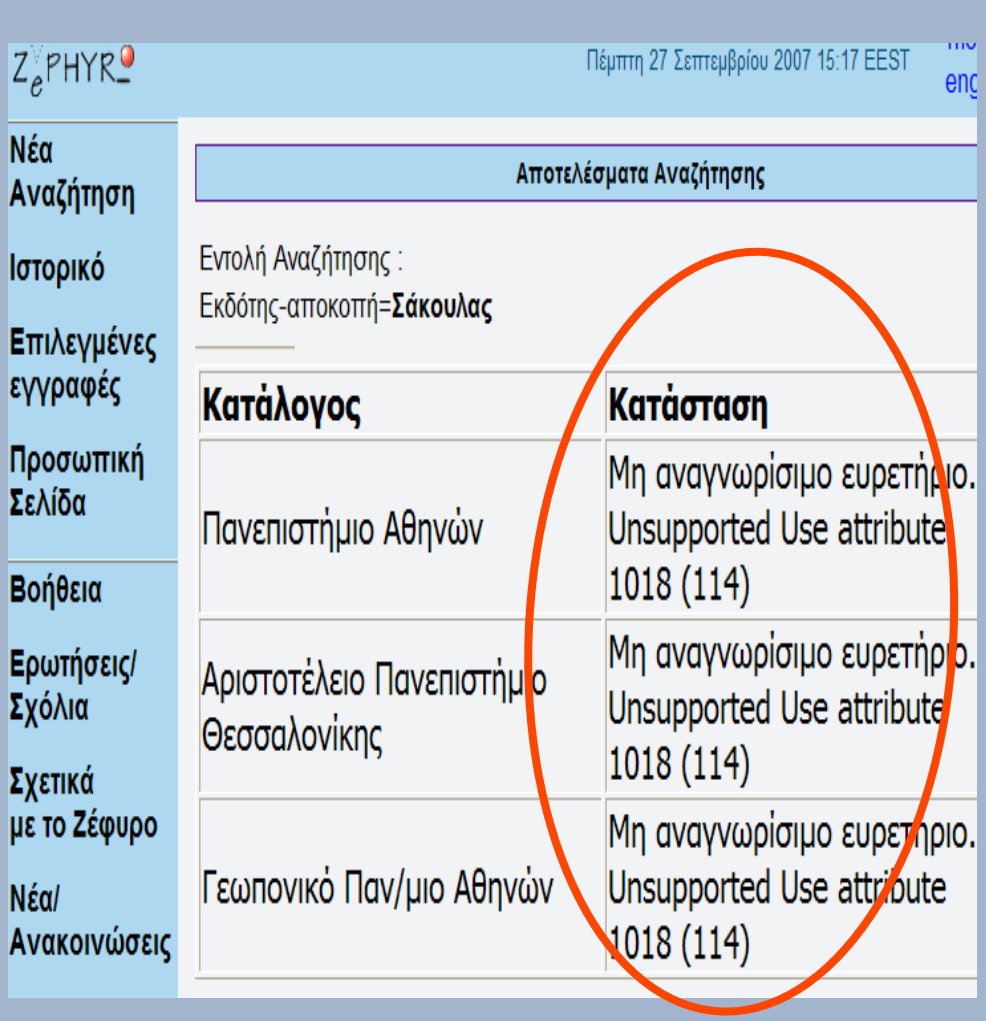
- Κυρίαρχο πρόβλημα
  - Η υποστήριξη διαφορετικών χαρακτηριστικών αναζήτησης από επιμέρους Z39.50 υλοποιήσεις

Αναζήτηση από το περιβάλλον "η Αργώ"

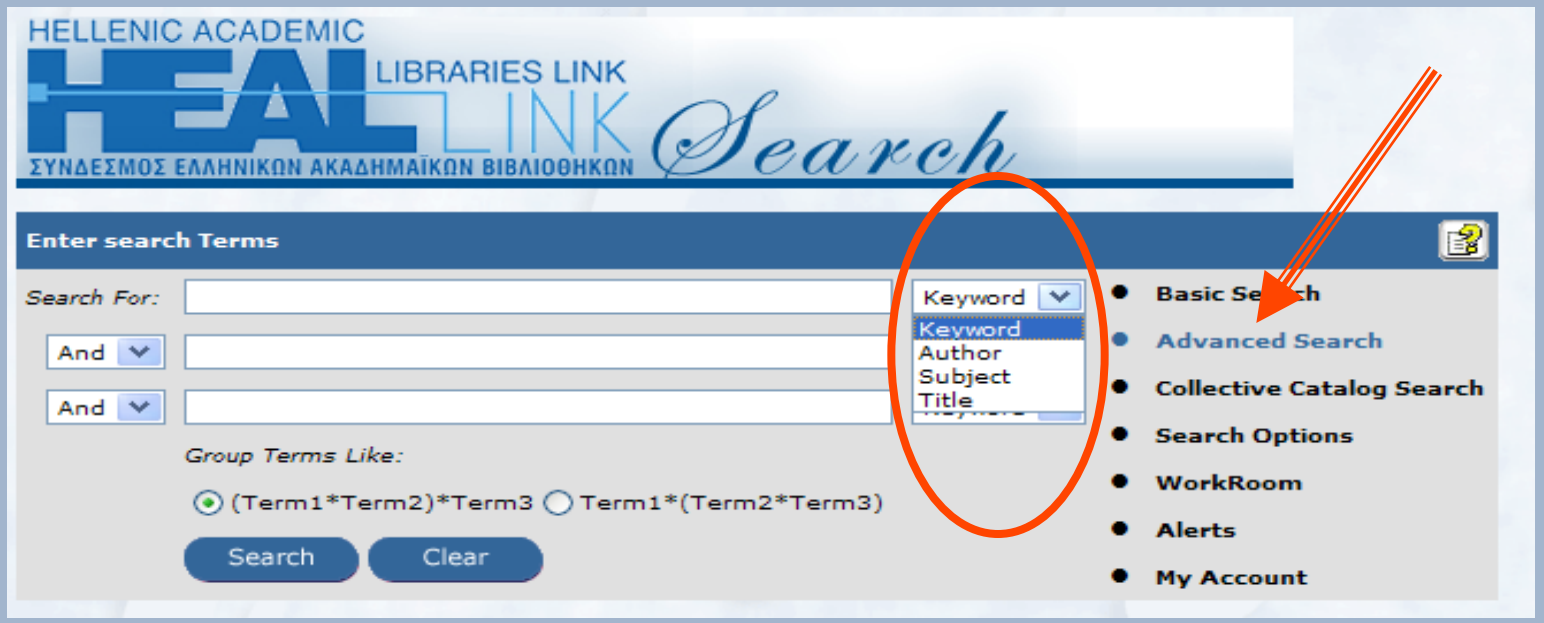


Αναζήτηση από το περιβάλλον “Ζέφυρος”





Αναζήτηση από το περιβάλλον "HEAL Link Search"



- Περιορισμός των ευρετηρίων αναζήτησης!

Συνέπειες από μη υποστηριζόμενα Σημεία Πρόσβασης

- Αποτυχημένες επερωτήσεις
  - Η Z39.50 πηγή απορρίπτει την επερώτηση και στέλνει διαγνωστικό μήνυμα λάθους (π.χ. MELVYL, COPAC)
  - Ο χρήστης δεν παίρνει αποτελέσματα
- Ασυνεπείς απαντήσεις
  - Η Z39.50 πηγή αντικαθιστά αυθαίρετα το μη υποστηριζόμενο χαρακτηριστικό με κάποιο άλλο που υποστηρίζει (π.χ. Library of Congress)
  - Ο χρήστης θα λάβει απάντηση χωρίς να γνωρίζει πώς προήλθε
  - Το σπουδαιότερο: ο χρήστης δε θα ενημερωθεί για το ότι έγινε η αντικατάσταση

- Στην Ελλάδα 24 πηγές που ελέγχθηκαν απορρίπτουν την επερώτηση με αποστολή διαγνωστικού μηνύματος λάθους

Πόσο υπαρκτό είναι το πρόβλημα της μη υποστήριξης Σημείων Πρόσβασης - Ι

- Στατιστικά της IndexData σχετικά με τα πιο δημοφιλή υποστηριζόμενα Σημεία Πρόσβασης από:

  3.018 πηγές σε διεθνές επίπεδο, όπου 1.697 υλοποιούν την υπηρεσία αναζήτησης (search service)

  - Δεν υπάρχει Σημείο Πρόσβασης που να το υποστηρίζουν όλες οι πηγές
  - Title (use attribute 4) το υποστηρίζουν 1.591 ή **93,75%**
  - Subject Heading (use attribute 21) το υποστηρίζουν 1.555 ή **91,63%**
  - Author (use attribute 1003) το υποστηρίζουν 1.549 ή **91,28%**

Πόσο υπαρκτό είναι το πρόβλημα της μη υποστήριξης Σημείων Πρόσβασης - ΙΙ

- Σε αντίστοιχη έρευνά μας σε δείγμα 24 Z39.50 πηγών στην Ελλάδα
  - Υπάρχει ένα μόνο Σημείο Πρόσβασης το οποίο υποστηρίζουν όλες οι πηγές: Author (use attribute 1003)
  - Subject Heading (use attribute 21) και Title (use attribute 4): κάθε ένα υποστηρίζεται από 23 διαφορετικές πηγές

- Η κατάσταση στην Ελλάδα εμφανίζεται καλύτερη από τη διεθνή (μεγαλύτερος βαθμός συμβατότητας), παρά το ότι η σειρά κατάταξης των Σημείων Πρόσβασης είναι διαφορετική

Αντικατάσταση μη υποστηριζόμενων Σημείων Πρόσβασης;

- Πώς θα μπορούσε μία μηχανή μετα-αναζήτησης να αντικαταστήσει το μη υποστηριζόμενο Σημείο Πρόσβασης με κάποια άλλα;

- Θα αλλάξει σημασιολογικά η αρχική επερώτηση, πόσο;

Σημασιολογική Αντικατάσταση μη υποστηριζόμενων Σημείων Πρόσβασης

Στο Εργαστήριο Ψηφιακών Βιβλιοθηκών και Ηλεκτρονικής Δημοσίευσης του ΤΑΒ - Ιονίου Πανεπιστημίου αναπτύξαμε:

- Μεθοδολογία αντικατάστασης των μη υποστηριζόμενων Σημείων Πρόσβασης από την πηγή
  - Χρησιμοποιώντας το γράφο συσχέτισης των Σημείων Πρόσβασης
  - Επεκτείνοντας ή περιορίζοντας τη σημασιολογία της αρχικής επερώτησης
- Σύστημα υλοποίησης των μεθόδων αντικατάστασης: **http://dlib.ionio.gr/zSAPN**

Σημείο Πρόσβασης

- Σε ένα κατάλογο ή σε μία βάση δεδομένων
  - Μπορεί να θεωρηθεί οποιοδήποτε μέρος κάθε εγγραφής με το οποίο μπορούμε να αναζητήσουμε, είτε ακόμα και να ταυτίσουμε, τις οντότητες που περιγράφονται
    - Ένα όνομα, μία θεματική επικεφαλίδα, ένας ταξινομικός αριθμός, κλπ
- Σε ένα πληροφοριακό σύστημα
  - Τα πεδία (ή ευρετήρια) αναζήτησης που προέρχονται από την ομαδοποίηση των σημείων πρόσβασης
  - Χρησιμοποιούνται στις επερωτήσεις κατά τις διαδικασίες αναζήτησης.

Σημασιολογία σημείων πρόσβασης στο περιβάλλον Z39.50

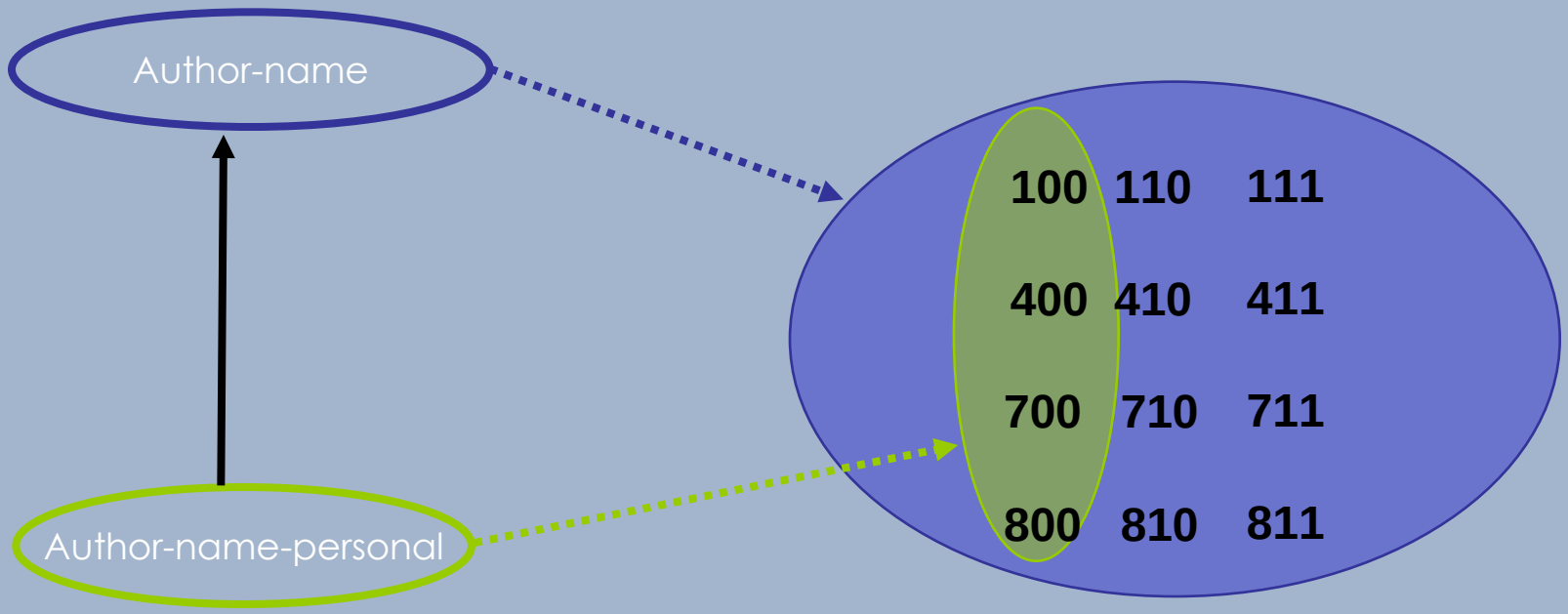
- Η σημασιολογία των Σημείων Πρόσβασης είναι ορισμένη στο *"Attribute Set BIB-1 (Z39.50-1995): Semantics"*
  - Αντιπροσωπεύει γενική συναίνεση μεταξύ των μελών του Z39.50 Implementors Group (ZIG)
  - Συντηρείται ως επίσημο τεκμήριο της Z39.50 Maintenance Agency
  - Ορίζει τη σημασιολογία κάθε Σημείου Πρόσβασης χρησιμοποιώντας ετικέτες των αντιπροσωπευτικών πεδίων MARC bibliographic format

- Ένα παράδειγμα
  - Το Σημείο Πρόσβασης *Author-name-Personal* (ή χαρακτηριστικό *use* με τιμή 1004) περιλαμβάνει τα δεδομένα από τα πεδία με MARC 21 ετικέτες {100, 400, 700, 800}

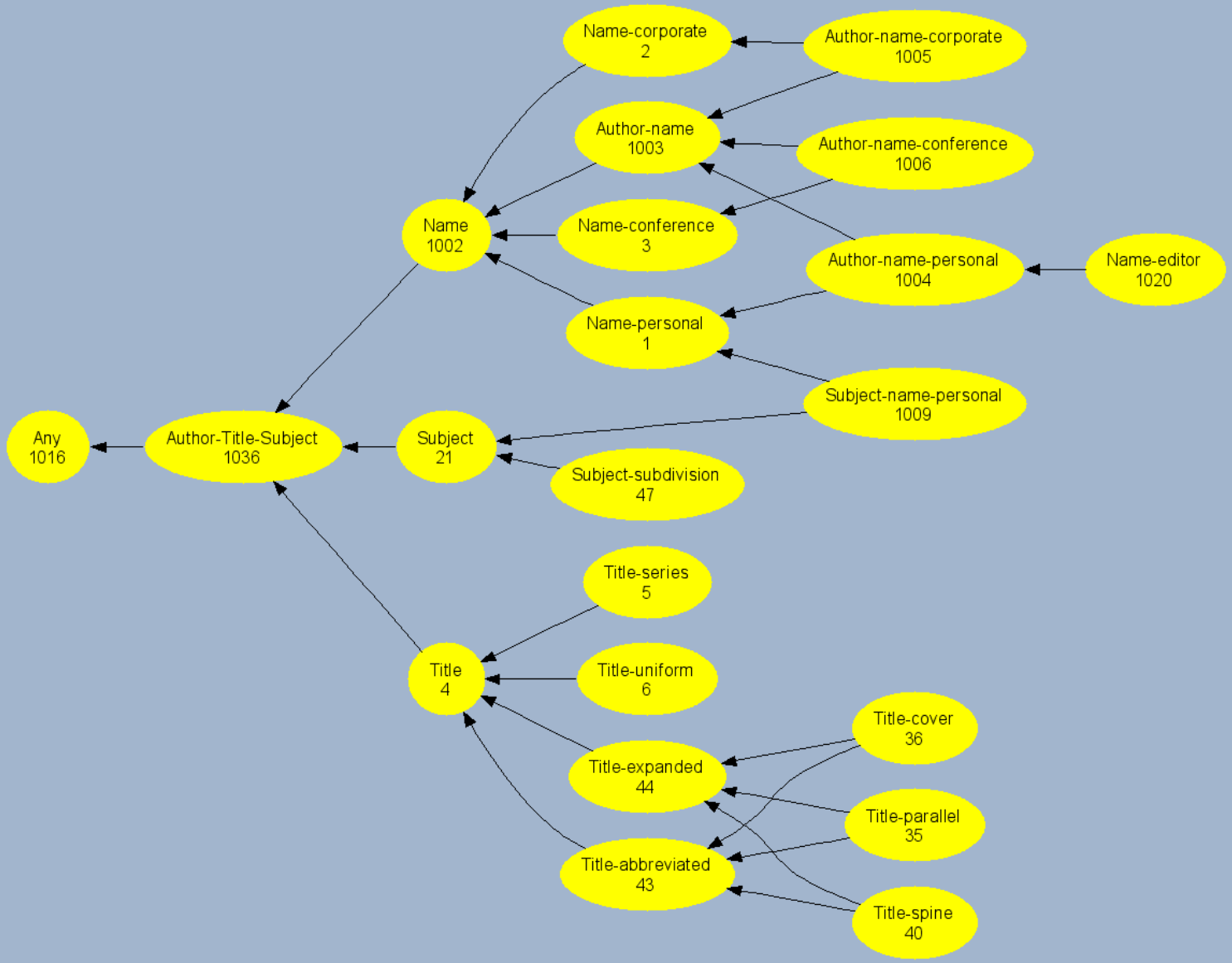
Συσχέτιση υποσυνόλου Σημείων Πρόσβασης

- Ένα Σημείο Πρόσβασης θεωρείται υποσύνολο ενός άλλου εάν το σύνολο των πεδίων που ορίζονται για να το δημιουργήσουν είναι υποσύνολο των πεδίων που ορίζονται για να δημιουργήσουν το άλλο
  - Ένα παράδειγμα
    - *Author-name* = {100, 110, 111, 400, 410, 411, 700, 710, 711, 800, 810, 811}
    - *Author-name-personal*={100, 400, 700, 800}
    - Το Σημείο Πρόσβασης Access Point *Author-name- personal* θεωρείται ως υποσύνολο του *Author-name*

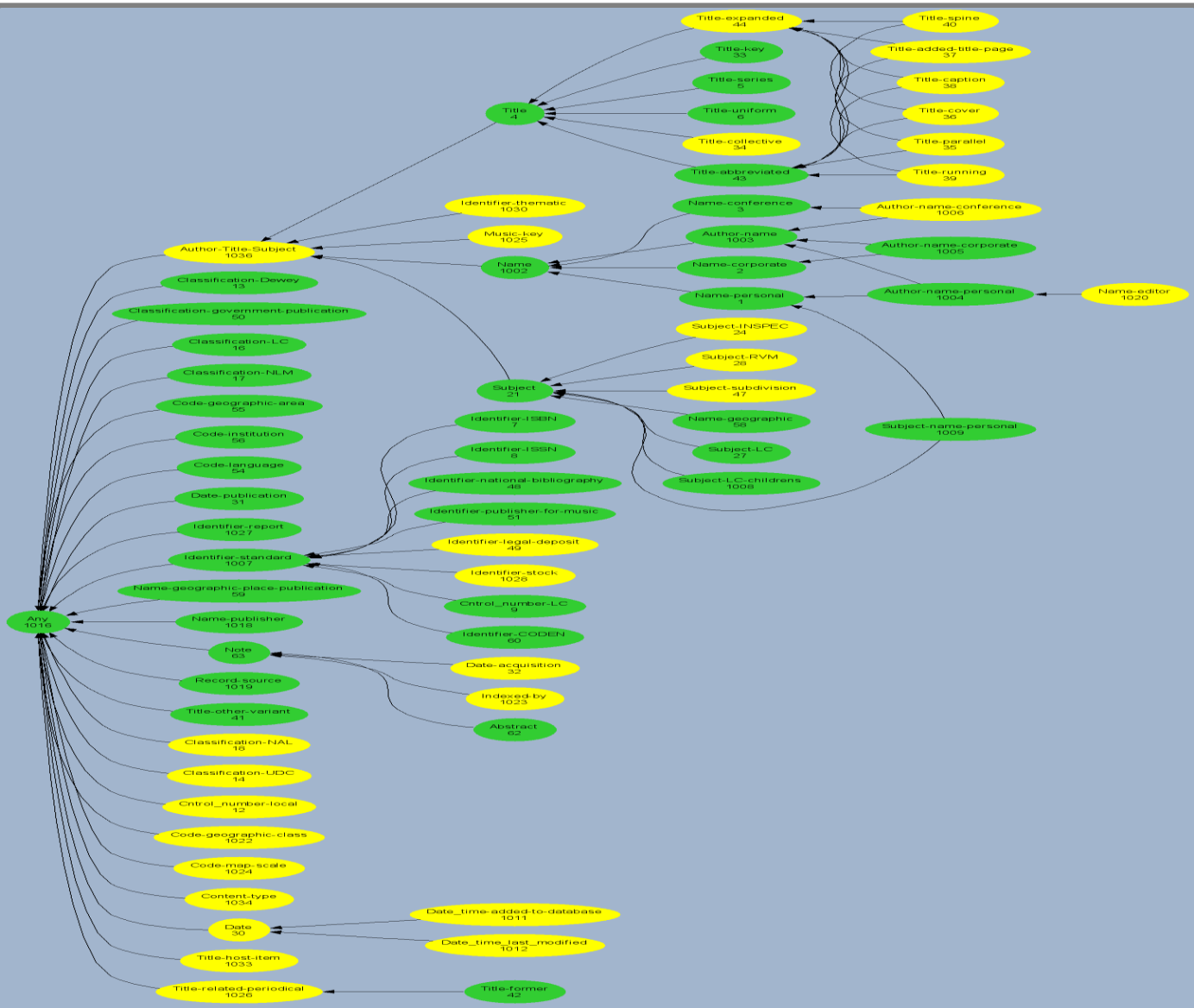
Συσχέτιση υποσυνόλου Σημείων Πρόσβασης



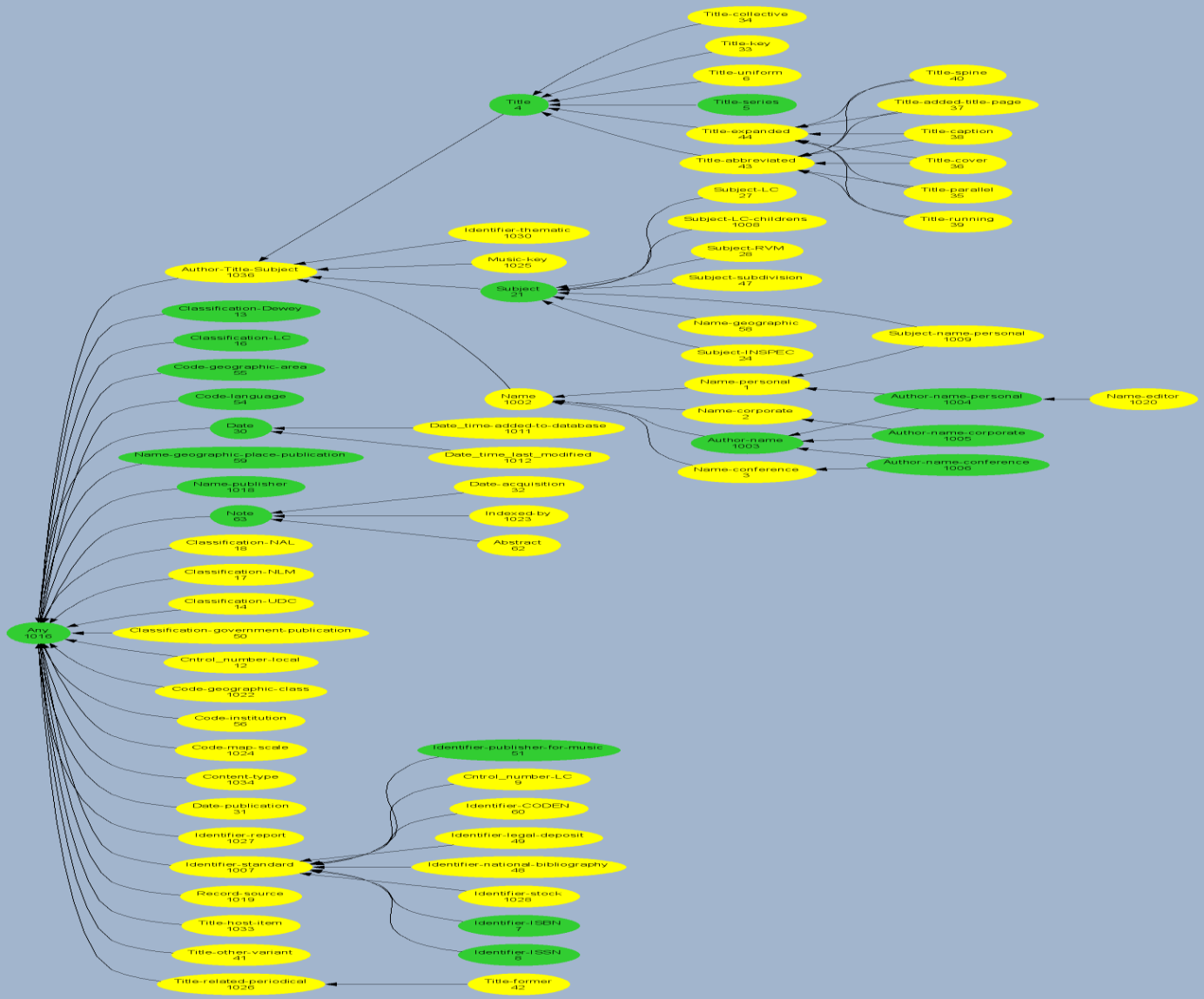
Ενδεικτικός γράφος συσχέτισης υποσυνόλου Σημείων Πρόσβασης



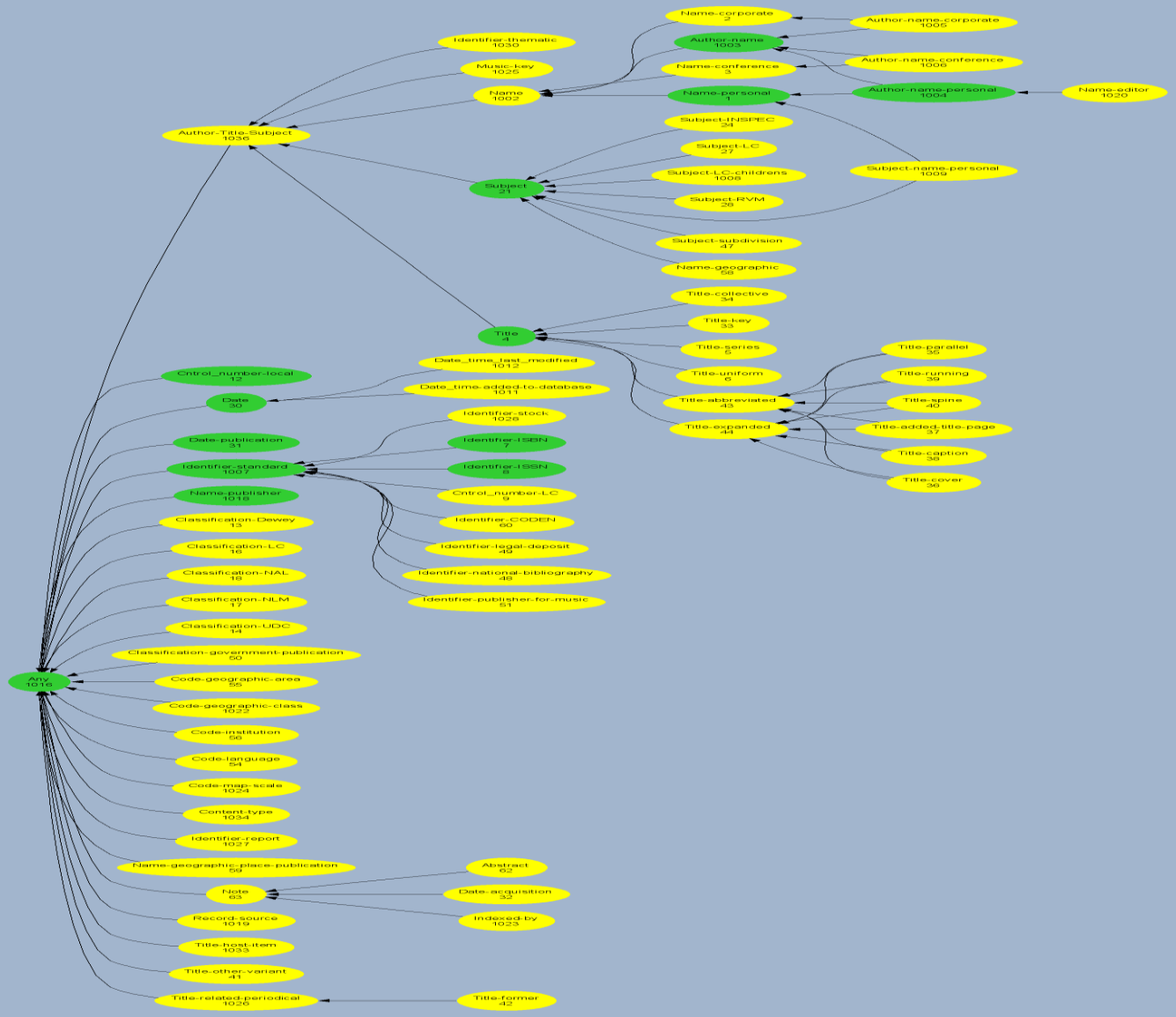
Γράφος συσχετίσεων με υποστηριζόμενα Σημεία Πρόσβασης - LC



Γράφος συσχετίσεων με υποστηριζόμενα Σημεία Πρόσβασης - ΣΚΕΑΒ



Γράφος συσχετίσεων με υποστηριζόμενα Σημεία Πρόσβασης – Πανεπιστήμιο Κρήτης



Ένα παράδειγμα σημασιολογικής αντικατάστασης

- Θέλουμε όλα τα συνέδρια της IEEE και μόνο
  - Προσοχή! Όχι τεχνικές εκθέσεις τις IEEE ούτε εγγραφές με θέμα συνέδρια της IEEE

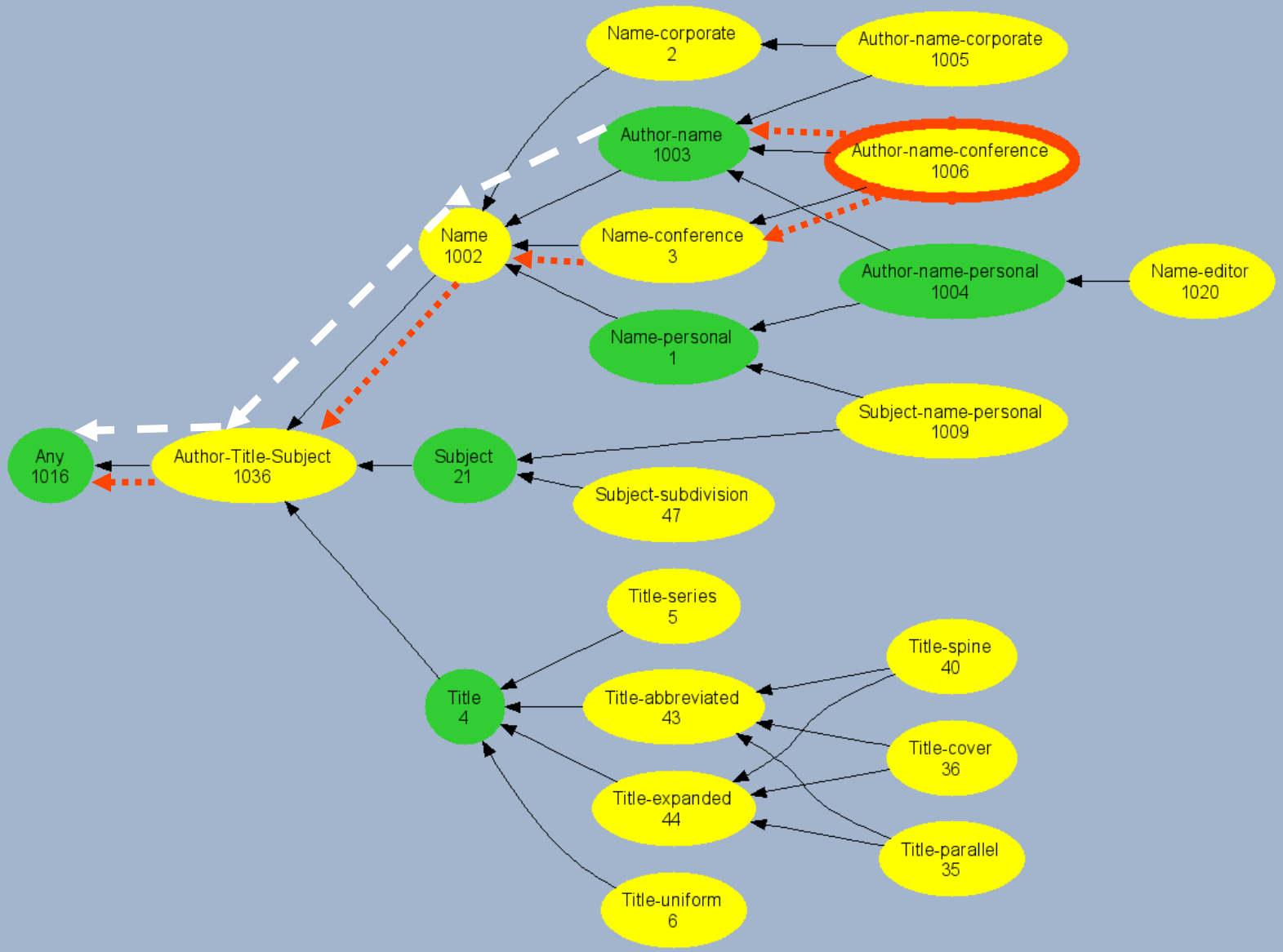
- Καταλληλότερο Σημείο Πρόσβασης:
  - *Author-name-conference-1006* = {111, 411, 711, 811}

! Σχεδόν καμία μηχανή αναζήτησης δε δίνει τη δυνατότητα χρήσης του!

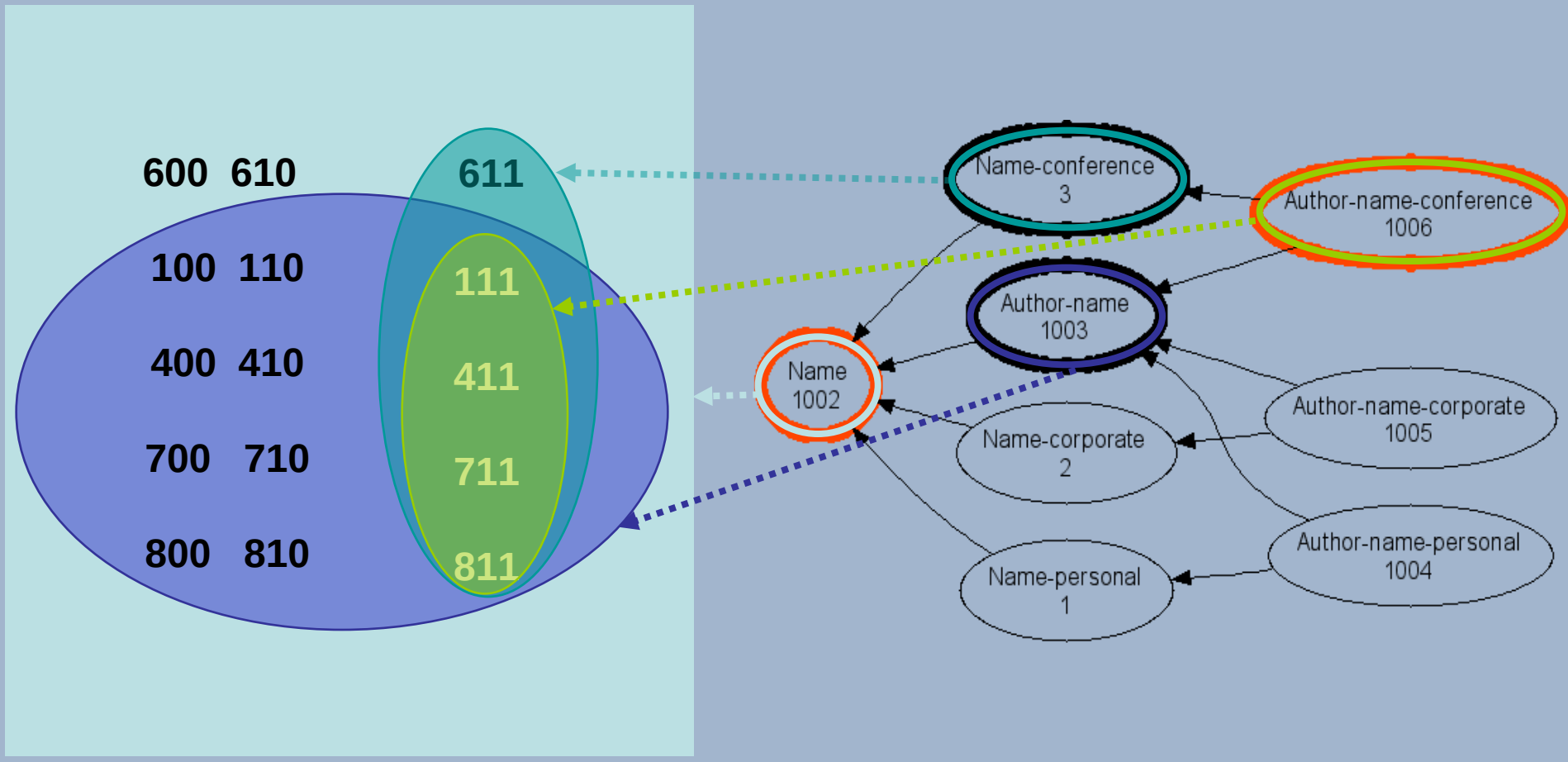
Αντικατάσταση μη υποστηριζόμενων Σημείων Πρόσβασης: Z39.50 πηγή ΣΚΕΑΒ

Αντικατάσταση μη υποστηριζόμενων Σημείων Πρόσβασης: Z39.50 πηγή LC

Αντικατάσταση μη υποστηριζόμενων Σημείων Πρόσβασης: Z39.50 πηγή Π-Κρήτης



Σημασιολογία αποτελεσμάτων αντικαταστάσεων – I



Σημασιολογία αποτελεσμάτων αντικαταστάσεων – ΙΙ

- Library of Congress: αποτελέσματα ίδια με την περίπτωση που θα το υποστήριζε
  - Το ίδιο ακριβή αποτελέσματα με το ΣΚΕΑΒ που το υποστηρίζει
- Πανεπιστήμιο Κρήτης
  - Αποτελέσματα πολύ κοντά στην αρχική επερώτηση
  - Ο θόρυβος περιλαμβάνει και άλλες εγγραφές της IEEE
  - Δεν περιλαμβάνει εγγραφές που έχουν ως θέμα την IEEE

Μη επιθυμητή εγγραφή

110 20|**a**American National Standards Institute.

245 10|**a**Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications :|**b**IEEE standards for local area networks /|**c**sponsor Technical Committee on Computer Communications of the IEEE Computer Society.

260 0 |**a**New York, NY, USA :|**b**Distributed in cooperation with Wiley-Interscience,|**c**c1985.

300 |**a**143 p. :|**b**ill. ;|**c**27 cm.

500 |**a**Cover title: Local area networks.

500 |**a**At head of title: An American national standard.

500 |**a**Spine title: 802.3 CSMA/CD

500 |**a**"ANSI/IEEE, Std 802.3-1985."

650 0|**a**Local area networks (Computer networks)|**x**Standards|**z**United States.

710 20|**a**IEEE Computer Society.|**b**Technical Committee on Computer Communications.

740 01|**a**Local area networks.

740 01|**a**802.3 CSMA/CD.

852 0 |**a**GrHePKT|**b**HERSC|**c**HBCS|**h**TK5105.7|**i**.A46 1985b

Η ερώτηση στο σύστημα – Χωρίς Αντικατάσταση

Source	Hits	Query
Library of Congress	7969	Author-name-conference_1006 = IEEE
National and Kapodistrian University of Athens	Error: Unsupported Use attribute	Author-name-conference_1006 = IEEE
University of Crete	Error: Unsupported Use attribute	Author-name-conference_1006 = IEEE
Hellenic Academic Libraries Union Catalogue	1036	Author-name-conference_1006 = IEEE

Η ερώτηση στο σύστημα – Με Αντικατάσταση

Source	Hits	Query
Library of Congress	1661	Author-name-conference_1006 = IEEE From the Access Point Substitution... Author-name_1003 Name-conference_3
National and Kapodistrian University of Athens	95	Author-name-conference_1006 = IEEE From the Access Point Substitution... Author-name_1003
University of Crete	339	Author-name-conference_1006 = IEEE From the Access Point Substitution... Author-name_1003
Hellenic Academic Libraries Union Catalogue	1036	Author-name-conference_1006 = IEEE The Access Point is supported by the source

Σύγκριση αποτελεσμάτων

- Επερώτηση: *Author-name-conference_1006 = IEEE*

Z39.50 πηγή	Χωρίς Αντικατάσταση	Με Αντικατάσταση
Library of Congress	7969	1661
National and Kapodistrian University of Athens	Error: Unsupported Use attribute	95
University of Crete	Error: Unsupported Use attribute	339
Hellenic Academic Libraries Union Catalogue	1036	1036

Σύνοψη - Ι

- Η απόδοση και αποτελεσματικότητα ενός περιβάλλοντος μετα-αναζήτησης περιορίζονται σημαντικά από τις δυνατότητες των επιμέρους περιβαλλόντων
- Ειδικότερα, στα περιβάλλοντα μετα-αναζήτησης των βιβλιοθηκών όπου οι Z39.50 πηγές αποτελούν τεράστια ενότητα
  - Κυρίαρχο πρόβλημα: η υποστήριξη διαφορετικών χαρακτηριστικών αναζήτησης από τις επιμέρους Z39.50 υλοποιήσεις
  - Μπορεί να βελτιωθεί σημαντικά με τη χρήση της σημασιολογίας των σημείων πρόσβασης
    - Επεκτείνοντας ή περιορίζοντας τη σημασιολογία της αρχικής επερώτησης
- Η προτεινόμενη μέθοδος μπορεί να βελτιώσει και τις δυνατότητες της πηγής αποφεύγοντας τις αυθαίρετες αντικαταστάσεις μη υποστηριζόμενων χαρακτηριστικών στις επερωτήσεις

Σύνοψη - ΙΙ

- Η λειτουργικότητα αυτή παρέχεται ως ελεύθερη υπηρεσία από το ***Εργαστήριο Ψηφιακών Βιβλιοθηκών και Ηλεκτρονικής Δημοσίευσης*** του ΤΑΒ - Ιονίου Πανεπιστημίου:

  http://dlib.ionio.gr/zSAPN

- Η χρήση του συστήματος θα μας δώσει σημαντικά στοιχεία για την επέκταση και βελτίωσή του

Τέλος!

Σας ευχαριστώ για την προσοχή σας

Μιχάλης Σφακάκης

Εργαστήριο Ψηφιακών Βιβλιοθηκών και Ηλεκτρονικής Δημοσίευσης
Τμήμα Αρχειονομίας - Βιβλιοθηκονομίας
Ιόνιο Πανεπιστήμιο, Κέρκυρα
sfakakis@ionio.gr