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

Κωνσταντίνος Ντονάς
Βιβλιοθήκη και Κέντρο Πληροφόρησης του Πανεπιστημίου Μακεδονίας
Εγνατίας 156, Θεσσαλονίκη
2310891830

donas@uom.gr

Εισαγωγή

Η αξία των μηχανών αναζήτησης για τον εντοπισμό πληροφοριών είναι αδιαμφισβήτητη αλλά είναι απλώς αδύνατο να κατασκευαστεί ευρετήριο (index) για ολόκληρο τον Παγκόσμιο Ιστό. Αυτό οφείλεται στην ταχύτατη ανάπτυξη του Ιστού καθώς και στο ότι πολλές ιστοσελίδες (web pages) ενημερώνονται τακτικά. Ο όγκος της διαθέσιμης πληροφορίας στον Ιστό είναι πραγματικά τεράστιος και μάλιστα συνεχώς αυξάνεται. Οι πληροφορίες αυτές όμως βρίσκονται κατακερματισμένες σε μεγάλο πλήθος διαφορετικών ηλεκτρονικών τοποθεσιών. Το πρόβλημα είναι ακόμα μεγαλύτερο διότι τεράστιος αριθμός σελίδων δεν είναι ορατός στα προγράμματα «σάρωσης» που χρησιμοποιούν οι μηχανές αναζήτησης (spiders ή web crawlers). Οι σελίδες αυτές αποτελούν το λεγόμενο *βαθύ Ιστό* (deep Web), το μεγαλύτερο μέρος του οποίου αποτελείται από online βάσεις δεδομένων ελεγχόμενης χρήσης (proprietary databases). Συνεπώς, μεγάλη ποσότητα πληροφοριακών τεκμηρίων που βρίσκεται αποθηκευμένη σε βάσεις δεδομένων δεν είναι προσβάσιμη από τις παραδοσιακές μηχανές αναζήτησης.

Ωστόσο, οι απαιτήσεις των καιρών (ή μάλλον των χρηστών) είναι η απλότητα, η ευκολία στη χρήση και η ταχύτητα. Αυτοί είναι και οι βασικοί λόγοι που η μεγάλη πλειοψηφία των φοιτητών και ερευνητών χρησιμοποιεί κατά κύριο λόγο τη μηχανή αναζήτησης *Google* που σε γενικές γραμμές επιστρέφει «αρκετά καλά» αποτελέσματα. Παρά το γεγονός ότι οι ακαδημαϊκές βιβλιοθήκες παρέχουν πρόσβαση σε μεγάλο αριθμό από ηλεκτρονικές πηγές και ξοδεύουν μεγάλα ποσά για συνδρομές σε παροχείς δεδομένων, οι περισσότεροι χρήστες στρέφονται στον ανοιχτό Ιστό. Στατιστικές έρευνες έχουν δείξει πως πολλοί χρήστες δε χρησιμοποιούν τις πηγές που τους παρέχει μια βιβλιοθήκη λόγω δυσκολίας στην επιλογή, πλοήγηση και αναζήτηση στις εκατοντάδες διαθέσιμες πηγές.

Σύμφωνα με την παραδοσιακή πρακτική αναζήτησης, ο χρήστης είναι υποχρεωμένος να επιλέξει μία βάση δεδομένων, να κάνει την αναζήτησή του, να συλλέξει και να αξιολογήσει τα επιστρεφόμενα δεδομένα και έπειτα να επαναλάβει τη διαδικασία με άλλη βάση δεδομένων. Επιπρόσθετα, κάθε βάση δεδομένων έχει διαφορετικά χαρακτηριστικά και επιλογές αναζήτησης, γεγονός που αναγκάζει το χρήστη να αφιερώσει χρόνο ώστε να εξοικειωθεί με το περιβάλλον αναζήτησης που αυτή παρέχει. Μάλιστα έχει παρατηρηθεί το φαινόμενο να ψάχνουν οι φοιτητές σε λίγες μόνο πηγές με τις οποίες είναι εξοικειωμένοι άσχετα αν αυτές είναι οι πλέον κατάλληλες.

Εδώ και αρκετά χρόνια έχει αναγνωριστεί η ανάγκη ύπαρξης ενός κοινού τρόπου αναζήτησης σε πολλές πηγές. Μία νέα τεχνολογία που έχει εξελιχθεί για το σκοπό αυτό είναι η *ταυτόχρονη ή παράλληλη αναζήτηση* σε πολλαπλές πηγές δεδομένων από ενιαίο περιβάλλον (federated search ή metasearch ή cross-search). Δίνει τη δυνατότητα στις βιβλιοθήκες να ανταποκριθούν στις σύγχρονες προκλήσεις με την παροχή επαρκούς και αποτελεσματικής πρόσβασης σε πληθώρα πηγών πληροφορίας και προσφέρει την ελπίδα να κερδίσουν ξανά ορισμένους χρήστες (Luther 2003).

Παράλληλη Αναζήτηση

Η ανερχόμενη τεχνολογία της παράλληλης αναζήτησης παρέχει τη δυνατότητα στο χρήστη να ψάξει ταυτόχρονα πολλές ανεξάρτητες πηγές κάνοντας μία και μόνο ερώτηση και επιστρέφει ενιαίο σύνολο αποτελεσμάτων. Έτσι, εξοικονομείται επαναλαμβανόμενη επίπονη προσπάθεια και εξασφαλίζεται περισσότερος χρόνος για αξιοποίηση και επεξεργασία των επιστρεφόμενων τεκμηρίων. Ουσιαστικά, μια μηχανή ταυτόχρονης αναζήτησης μετασχηματίζει το ερώτημα του χρήστη σε μια μορφή κατανοητή από κάθε επιθυμητή πηγή, επικοινωνεί με αυτές και ανακτά τα αποτελέσματα. Επίσης, εξαλείφεται η ανάγκη εξοικείωσης με πολλά διαφορετικά περιβάλλοντα αναζήτησης και μεταπήδησης από τη μια διεπαφή (interface) αναζήτησης σε άλλη. Τα πλεονεκτήματα λοιπόν είναι προφανή.

Η ανάπτυξη και η χρήση τέτοιων πληροφοριακών συστημάτων σε βιβλιοθήκες τα τελευταία χρόνια έχει δείξει πως είναι από τις δημοφιλέστερες υπηρεσίες μεταξύ των χρηστών και είναι το βασικό εργαλείο που χρησιμοποιούν οι φοιτητές για αναζήτηση στις διαθέσιμες πηγές μιας βιβλιοθήκης (Boyd et al. 2006). Η δημοτικότητα τους οφείλεται κυρίως στη δυνατότητα παράκαμψης των δυσκολιών που προκύπτουν από τον πολύ μεγάλο αριθμό διαθέσιμων πηγών πληροφοριών αφού κάθε πηγή έχει τους δικούς της κανόνες και τις ιδιαιτερότητες της ενώ οι διεπαφές των πηγών παρουσιάζουν αρκετές διαφορές μεταξύ τους. Επίσης, τα συστήματα ταυτόχρονης αναζήτησης διευκολύνουν σημαντικά τον εντοπισμό χρήσιμων πληροφοριών και βοηθούν ιδιαίτερα τα άτομα που δε γνωρίζουν σε ποιες ακριβώς βάσεις δεδομένων να ψάξουν. Αυτό επιτυγχάνεται μέσω της οργάνωσης των διαθέσιμων πηγών ανά θεματική κατηγορία.

Ωστόσο, υπάρχουν και αρκετά θέματα που προκαλούν προβληματισμό (Marshall, Herman & Rajan 2006). Ένα ζήτημα που προκύπτει από τη χρήση ενός συστήματος ταυτόχρονης αναζήτησης είναι ο μεγάλος όγκος πληροφορίας (information overload) που επιστρέφεται, ο οποίος δημιουργεί την ανάγκη διαχείρισης, οργάνωσης και ταξινόμησης των αποτελεσμάτων. Ένα άλλο πρόβλημα, εγγενές στις μηχανές ταυτόχρονης αναζήτησης, είναι η σχετικά αργή απόδοση τους καθώς ψάχνουν ταυτόχρονα πολλές πηγές, οπότε η πιο αργή πηγή καθορίζει τη συνολική ταχύτητα. Επιπλέον, προβλήματα παρουσιάζουν η ανάκτηση και παρουσίαση των σχετικότερων αποτελεσμάτων πρώτα. Η σειρά εμφάνισης και η σχετικότητα των αποτελεσμάτων είναι θέματα που επιδέχονται πολλής συζήτησης. Επίσης, δεν υπάρχει δυνατότητα απαλοιφής των διπλότυπων. Αξίζει ακόμα να σημειωθεί η πολυπλοκότητα και η δυσκολία υλοποίησης τέτοιων συστημάτων.

Παρά τα σημαντικά πλεονεκτήματα της παράλληλης αναζήτησης, προβληματισμός επικρατεί, στις τάξεις των βιβλιοθηκονόμων κυρίως, για την ποιότητα των επιστρεφόμενων αποτελεσμάτων και την παραίσθηση που σου δίνουν τέτοια συστήματα ότι μπορεί να εντοπίσει κανείς εύκολα και γρήγορα τις ζητούμενες πληροφορίες. Ορισμένοι πιστεύουν πως τέτοια συστήματα επιστρέφουν σε γενικές γραμμές σχετικά καλά αποτελέσματα αλλά δεν είναι ίσως αυτά που οι χρήστες έχουν πραγματικά ανάγκη (Cox 2003). Στην πράξη όμως πολλοί χρήστες έχουν χρησιμοποιήσει εργαλεία παράλληλης αναζήτησης αρκετά ώστε να ψάχνουν αποτελεσματικά και με επιτυχία.

Σίγουρα πάντως τα εργαλεία αυτά έχουν μέλλον και θα βελτιωθούν ακόμα περισσότερο στην πορεία. Φέρνουν τους χρήστες πιο κοντά στις διαθέσιμες πηγές που σε άλλη περίπτωση θα τις έχαναν. Βέβαια τα εργαλεία αυτά δεν αντικαθιστούν τις δεξιότητες που πρέπει να έχει κάποιος καλός ερευνητής ούτε και την ικανότητα εκτίμησης και κριτικής ανάλυσης των αποτελεσμάτων. Σαν συμπέρασμα η τεχνολογία της ταυτόχρονης αναζήτησης δεν αποτελεί πανάκεια αλλά είναι ένα πολύ βοηθητικό και χρήσιμο εργαλείο στα χέρια της ακαδημαϊκής κοινότητας. Είναι ένα πολύ καλό σημείο εκκίνησης της έρευνας για τον εντοπισμό πηγών με σχετικές εγγραφές. Η ποιότητα αποτελεσμάτων εξαρτάται άμεσα από τους όρους αναζήτησης που έχει εισάγει ο χρήστης. Η σχετικότητα και η ακρίβεια ενός τέτοιου πληροφοριακού συστήματος κυμαίνονται συνήθως σε υψηλά επίπεδα. Σε γενικές

γραμμές τα αποτελέσματα είναι τουλάχιστον αποδεκτά. Δεν αντικαθιστά όμως ούτε τη διεπαφή αναζήτησης κάθε πηγής ούτε την εκτενή έρευνα. Δε βελτιώνει τις δυνατότητες αναζήτησης που προσφέρει μια πηγή, απλά τις χρησιμοποιεί, ούτε επιστρέφει καλύτερα αποτελέσματα. Εξυπηρετεί κυρίως τους νέους χρήστες που αναζητούν μερικές καλές πηγές πληροφορίας. Το ζητούμενο πάντως είναι να ικανοποιεί τις απαιτήσεις τόσο των αρχάριων όσο και των έμπειρων χρηστών που διεξάγουν έρευνα. Είναι τεχνολογία που έχει συγκεντρώσει πολύ μεγάλο ενδιαφέρον στη βιβλιοθηκονομική κοινότητα και αναπτύσσεται συνεχώς. Έχουν κυκλοφορήσει ορισμένα εμπορικά προϊόντα (Chen 2006) αλλά και ελεύθερο λογισμικό ανοικτού κώδικα, που συνιστά αξιοπρόσεχτη εναλλακτική λύση χαμηλότερου κόστους αλλά υψηλής λειτουργικότητας.

Λογισμικό Ανοικτού Κώδικα - dbWiz

Στο πλαίσιο του έργου ΠΛΟΗΓΙΣ[1], η Βιβλιοθήκη του Πανεπιστημίου Μακεδονίας ανέλαβε την ανάπτυξη εργαλείου ταυτόχρονης αναζήτησης στον ηλεκτρονικό κατάλογο της βιβλιοθήκης, σε σημαντικές βιβλιογραφικές βάσεις δεδομένων, συλλογές δεδομένων ανοιχτής ή μη πρόσβασης και άλλες πηγές πληροφοριών. Έπειτα από αναζήτηση σχετικών εργαλείων λογισμικού ανοιχτού κώδικα (open source software tools), επιλέχθηκε το *dbWiz* (Mah & Stranack 2005), το οποίο αποτελεί μέλος της οικογένειας προγραμμάτων reSearcher, μιας πραγματικά αξιόλογης προσπάθειας της βιβλιοθήκης του Simon Fraser University. Το έχουν υιοθετήσει και συνεργάζονται στην ανάπτυξή του πολλές βιβλιοθήκες και κοινοπραξίες στον Καναδά.

Το dbWiz είναι προσπελάσιμο μέσω του Ιστού, οπότε μπορεί κανείς να χρησιμοποιήσει το περιβάλλον αναζήτησής του με χρήση ενός προγράμματος πλοήγησης (web browser). Το περιβάλλον αναζήτησης αλλά και αυτό της διαχείρισης του είναι υλοποιημένα στη γλώσσα προγραμματισμού *Perl*. Για την εγκατάσταση και λειτουργία του απαιτείται ένας σύγχρονος σταθμός εργασίας με λειτουργικό σύστημα *Linux* και αξιόπιστη και γρήγορη σύνδεση στο Διαδίκτυο. Οι τεχνικές προδιαγραφές απαιτούν τον διακομιστή Ιστού *Apache*, το σύστημα διαχείρισης βάσεων δεδομένων *MySQL*, το πακέτο λογισμικού *YAZ* για αναζήτηση με το πρωτόκολλο *Z39.50* και πληθώρα Perl λειτουργικών μονάδων.

Το dbWiz διαθέτει μια μηχανή παράλληλης αναζήτησης (parallel search engine) ώστε να πραγματοποιείται αναζήτηση σε πολλαπλές πηγές ταυτόχρονα. Για καθεμία από τις πηγές που ενδιαφέρεται ο χρήστης να ψάξει, δημιουργείται μια νέα παράλληλη διεργασία (process) η οποία διεκπεραιώνει την επικοινωνία και την ανάκτηση πληροφορίας από τη συγκεκριμένη πηγή. Όταν ολοκληρωθεί η αναζήτηση σε όλες τις πηγές, τότε τα αποτελέσματα συνδυάζονται και παρουσιάζονται στο χρήστη ως ένα ενιαίο σύνολο. Απαραίτητη προϋπόθεση για την αναζήτηση σε μια επιθυμητή πηγή πληροφορίας είναι η ύπαρξη αντίστοιχης *λειτουργικής μονάδας αναζήτησης* (search module ή plug-in).

Μια λειτουργική μονάδα αναζήτησης του dbWiz είναι μία ενότητα κώδικα στην γλώσσα προγραμματισμού Perl που αναλαμβάνει την επικοινωνία με την πηγή, την υποβολή του ερωτήματος και την ανάκτηση των αποτελεσμάτων. Το dbWiz έχει μεγάλο αριθμό από έτοιμες προ-εγκατεστημένες λειτουργικές μονάδες για πληθώρα πηγών και παροχέων πληροφορίας. Φυσικά, είναι δυνατή η προσθήκη νέων πηγών, αρκεί να κατασκευαστεί σχετική λειτουργική μονάδα αναζήτησης και να γίνουν οι κατάλληλες ρυθμίσεις στο διαχειριστικό περιβάλλον του dbWiz. Βέβαια η κατασκευή και η συντήρηση των

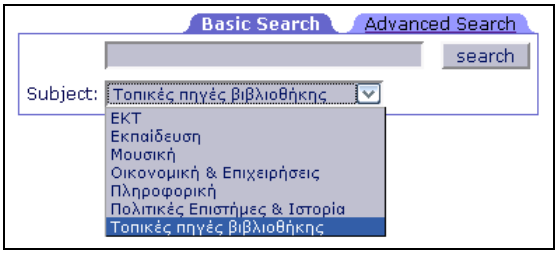
[1] Το Πανεπιστήμιο Μακεδονίας Οικονομικών και Κοινωνικών Επιστημών υλοποιεί το έργο «ΠΛΟΗΓΙΣ: ΑΠΟ ΤΗΝ ΠΛΗΡΟΦΟΡΙΑ ΣΤΗ ΓΝΩΣΗ: Ενίσχυση διάθεσης ηλεκτρονικού περιεχομένου και παροχή πληροφοριακής παιδείας στη Βιβλιοθήκη του Πανεπιστημίου Μακεδονίας», το οποίο εκτελείται στα πλαίσια του Επιχειρησιακού Προγράμματος Εκπαίδευσης και Αρχικής Επαγγελματικής Κατάρτισης ΙΙ (ΕΠΕΑΕΚ ΙΙ-Ενέργεια 2.1.3, Κατηγορία Πράξεων δ «Ενίσχυση και Εμπλουτισμός Ακαδημαϊκών Βιβλιοθηκών») με χρονική διάρκεια από 01/07/2000 μέχρι 30/09/2008 και συγχρηματοδοτείται μέσω του ΥΠΕΠΘ από την Ευρωπαϊκή Ένωση [3ο Κοινοτικό Πλαίσιο Στήριξης κατά 75% Κοινοτική Συμμετοχή (ΕΚΤ) και 25% Εθνικοί πόροι (ΥΠΕΠΘ/ΕΥΔ ΕΠΕΑΕΚ)].

λειτουργικών μονάδων αναζήτησης εμπεριέχουν σημαντικές δυσκολίες και απαιτούν εξειδικευμένες γνώσεις πληροφορικής.

Για τις ανάγκες της βιβλιοθήκης του Πανεπιστημίου Μακεδονίας κατασκευάστηκαν νέες λειτουργικές μονάδες αναζήτησης για τον κατάλογο της βιβλιοθήκης, το ιδρυματικό αποθετήριο (ΨΗΦΙΔΑ), την πύλη θεματικής αναζήτησης (ΘΥΡΑ), δημοφιλείς βάσεις δεδομένων του Εθνικού Κέντρου Τεκμηρίωσης (ΕΚΤ) και σημαντικές βιβλιογραφικές βάσεις δεδομένων (π.χ. πηγές στις οποίες είναι δυνατή η πρόσβαση μέσω της Silverplatter). Δηλαδή, προστέθηκε δυνατότητα αναζήτησης για ελληνικές πηγές που δεν είναι προ-εγκατεστημένες σε κάποιο εμπορικό ή μη προϊόν παράλληλης αναζήτησης. Επίσης, τροποποιήθηκαν κατάλληλα ορισμένες λειτουργικές μονάδες από τις έτοιμες του dbWiz που παρουσίαζαν κάποια προβλήματα.

Υπάρχουν δύο βασικοί μηχανισμοί πρόσβασης στις πηγές στόχους. Ο πρώτος είναι μέσω της διεπαφής ιστού (web interface) και ο δεύτερος μέσω προγραμματιστικής διεπαφής (API), συνήθως μέσω του πρωτοκόλλου αναζήτησης και ανάκτησης πληροφορίας Z39.50. Στην πρώτη περίπτωση προσομοιώνεται η αλληλεπίδραση ενός χρήστη με το δικτυακό τόπο της πηγής και τα αποτελέσματα αντί να προβληθούν σε ένα πρόγραμμα πλοήγησης (web browser) σαρώνονται από το dbWiz. Η αναζήτηση πραγματοποιείται μέσω *αυτόματης* πλοήγησης στους δικτυακούς τόπους των πηγών, συμπλήρωσης και υποβολής της σχετικής φόρμας, εντοπισμού, εξαγωγής και συνδυασμού των αποτελεσμάτων σε κατάλληλη μορφή. Οι λειτουργικές μονάδες αναζήτησης (search modules) της κατηγορίας αυτής επιτυγχάνουν τον εντοπισμό των επιθυμητών δεδομένων στις ιστοσελίδες των πηγών στόχων με χρήση *κανονικών εκφράσεων* (regular expressions), οι οποίες συνιστούν μία τυπική μέθοδο περιγραφής προτύπων κειμένου. Ο μηχανισμός όμως αυτός είναι επιρρεπής σε αλλαγές στη διεπαφή αναζήτησης των πηγών και στον τρόπο εμφάνισης των αποτελεσμάτων με συνέπεια να απαιτείται συστηματική συντήρηση των λειτουργικών μονάδων αναζήτησης. Όσο για τις πηγές στόχους της δεύτερης περίπτωσης, είναι συνήθως δυνατή η ανάκτηση των δεδομένων μέσω του πρωτοκόλλου επικοινωνίας Z39.50 και χρήσης του σχετικού API, μεθοδολογία που είναι περισσότερο αξιόπιστη.

Η σύνταξη και οι δυνατότητες αναζήτησης είναι διαφορετικές σε κάθε πηγή. Για την υποστήριξη ενός βασικού συνόλου λειτουργιών αναζήτησης, επιλέχθηκαν από τους δημιουργούς του dbWiz τα συνηθέστερα χαρακτηριστικά στις περισσότερες πηγές. Στη βασική αναζήτηση (Εικόνα 1**: Κατηγορίες πηγών στη βασική αναζήτηση**) ο χρήστης μπορεί να ψάξει μόνο με λέξεις κλειδιά σε μία ομάδα ή θεματική κατηγορία πηγών, η οποία περιλαμβάνει κάποιες προκαθορισμένες από το διαχειριστή πηγές. Για παράδειγμα η κατηγορία *'Τοπικές πηγές βιβλιοθήκης'* που φαίνεται στην ακόλουθη εικόνα περιλαμβάνει τον κατάλογο, το ψηφιακό αποθετήριο και την πύλη θεματικής αναζήτησης της Βιβλιοθήκης του Πανεπιστημίου Μακεδονίας.

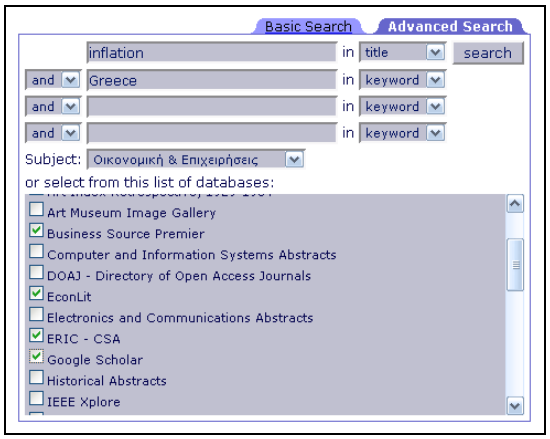


Εικόνα 1: Κατηγορίες πηγών στη βασική αναζήτηση του dbwiz

Στη σύνθετη αναζήτηση (Εικόνα 2**: Επιλογή μεμονωμένων πηγών στη σύνθετη αναζήτηση**) ο χρήστης μπορεί να δώσει συνδυασμό κριτηρίων (τίτλο, συγγραφέα, θεματική ενότητα και λέξη κλειδί) και να ψάξει είτε πάλι μια ομάδα πηγών είτε μεμονωμένες πηγές

που έχει επιλέξει ο ίδιος στη λίστα με τις διαθέσιμες πηγές. Είναι δυνατή η αναζήτηση μέχρι και με τέσσερις όρους συνδεδεμένους μεταξύ τους με λογικούς τελεστές (*AND*, *OR*, *NOT*).

Η αναζήτηση επιστρέφει προκαθορισμένο πλήθος από τα πρώτα αποτελέσματα από κάθε πηγή, αριθμός που είναι παραμετροποιήσιμος από το διαχειριστή. Όσο μεγαλύτερος βέβαια ο αριθμός, τόσο περισσότερος χρόνος απαιτείται για την ολοκλήρωση της αναζήτησης. Επίσης, οι δημιουργοί του dbWiz έλαβαν υπόψη τους το χρόνο απόκρισης του συστήματος. Για το σκοπό αυτό υπάρχει μέγιστος χρόνος αναμονής των αποτελεσμάτων από μια πηγή (timeout). Σε περίπτωση που η πηγή δεν έχει απαντήσει μέσα σε αυτό το διάστημα τότε απλά διακόπτεται η προσπάθεια ανάκτησης δεδομένων και επιστρέφεται κατάλληλο μήνυμα για τη συγκεκριμένη πηγή.

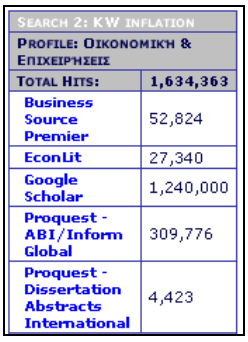


Εικόνα 2: Επιλογή μεμονωμένων πηγών στη σύνθετη αναζήτηση του dbwiz

Όταν ολοκληρωθεί μία αναζήτηση, τότε συλλέγονται τα επιστρεφόμενα αποτελέσματα και παρουσιάζονται ενιαία στο χρήστη. Κάθε εγγραφή (record) από το σύνολο των αποτελεσμάτων εμφανίζει μία σύντομη αλλά περιεκτική ποσότητα πληροφορίας που περιλαμβάνει τίτλο, συγγραφέα, ημερομηνία και το όνομα της πηγής από την οποία προήλθε (Εικόνα 3**: Ενδεικτικό αποτέλεσμα αναζήτησης**). Επίσης, κάθε εγγραφή περιέχει υπερσυνδέσμο που δείχνει στην ιστοσελίδα με τη λεπτομερή περιγραφή του αποτελέσματος και σύνδεσμο προς τη σελίδα του δικτυακού τόπου της πηγής από την οποία εξήχθη το αποτέλεσμα καθώς και openURL σύνδεσμο, κατάλληλο για τον link resolver της βιβλιοθήκης, για εντοπισμό του πλήρους κειμένου.

Εικόνα 3: Ενδεικτικό αποτέλεσμα αναζήτησης στο dbwiz

Επίσης, στο αριστερό τμήμα της διεπαφής αναζήτησης εμφανίζεται πίνακας με το πλήθος των αποτελεσμάτων για κάθε πηγή στην οποία πραγματοποιήθηκε αναζήτηση (Εικόνα 4**: Πίνακας με το πλήθος αποτελεσμάτων**). Να σημειωθεί πως υπάρχει δυνατότητα προβολής αποτελεσμάτων για κάθε μεμονωμένη πηγή, αρκεί ο χρήστης να επιλέξει την πηγή στον πίνακα με τα αποτελέσματα. Ακόμα, κρατιέται ιστορικό αναζητήσεων και συνεπώς ο χρήστης μπορεί να ανατρέξει όποτε το θελήσει στα αποτελέσματα προηγούμενων αναζητήσεων του.

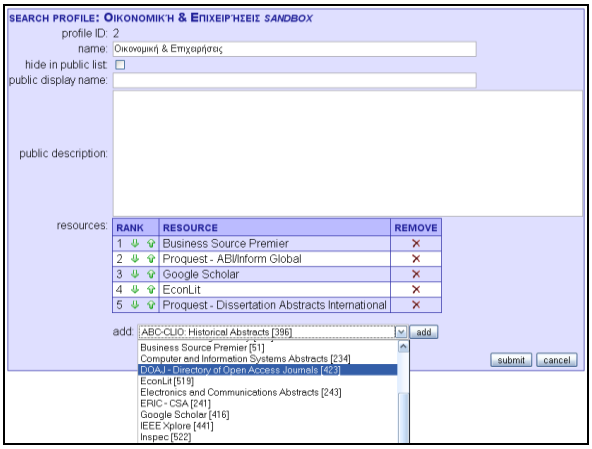


SEARCH 2: KW INFLATION	
PROFILE: ΟΙΚΟΝΟΜΙΚΗ & ΕΠΙΧΕΙΡΗΣΕΙΣ	
TOTAL HITS:	1,634,363
Business Source Premier	52,824
EconLit	27,340
Google Scholar	1,240,000
Proquest - ABI/Inform Global	309,776
Proquest - Dissertation Abstracts International	4,423

Εικόνα 4: Πίνακας με το πλήθος αποτελεσμάτων κάθε πηγής έπειτα από αναζήτηση στο dbwiz

Τέλος, ιδιαίτερη έμφαση έχει δοθεί στο κομμάτι της διαχείρισης (administration) του συστήματος. Το περιβάλλον διαχείρισης επιτρέπει στις βιβλιοθήκες να δημιουργήσουν τις δικές τους κατηγορίες αναζήτησης (search categories) και να προσθέτουν ή να διαγράφουν γρήγορα και εύκολα πηγές (resources) από το dbWiz προφίλ τους. Οι παράμετροι και οι ρυθμίσεις του περιβάλλοντος διαχείρισης αποθηκεύονται σε βάση δεδομένων. Η δημιουργία κατηγοριών είτε βάσει θεματικής ενότητας ή άλλου κριτηρίου γίνεται εύκολα, απλά με την εισαγωγή του ονόματος της κατηγορίας από το διαχειριστή και επιλογή των πηγών που αυτή θα περιλαμβάνει (Εικόνα 5**: Προσθήκη πηγών σε μια κατηγορία αναζήτησης**).

Ο χρήστης μετά την είσοδο του στο περιβάλλον διαχείρισης μπορεί να προσπελάσει τη λίστα με όλες τις διαθέσιμες πηγές του dbWiz και να ενεργοποιήσει μόνο αυτές που θέλει να συμπεριλάβει στη συλλογή πηγών που ενδιαφέρουν τη βιβλιοθήκη. Τέλος, η διεπαφή αναζήτησης έχει κατασκευαστεί με το *Perl Template Toolkit*, γεγονός που σημαίνει πως γίνεται χρήση προτύπων μορφοποίησης, οπότε είναι δυνατή η εκτεταμένη παραμετροποποίηση της εμφάνισης του δικτυακού τόπου του dbWiz σύμφωνα με τις ανάγκες κάθε βιβλιοθήκης. Οι αλλαγές αυτές γίνονται εύκολα μέσα από το περιβάλλον διαχείρισης.



Εικόνα 5: Προσθήκη πηγών σε μια κατηγορία αναζήτησης

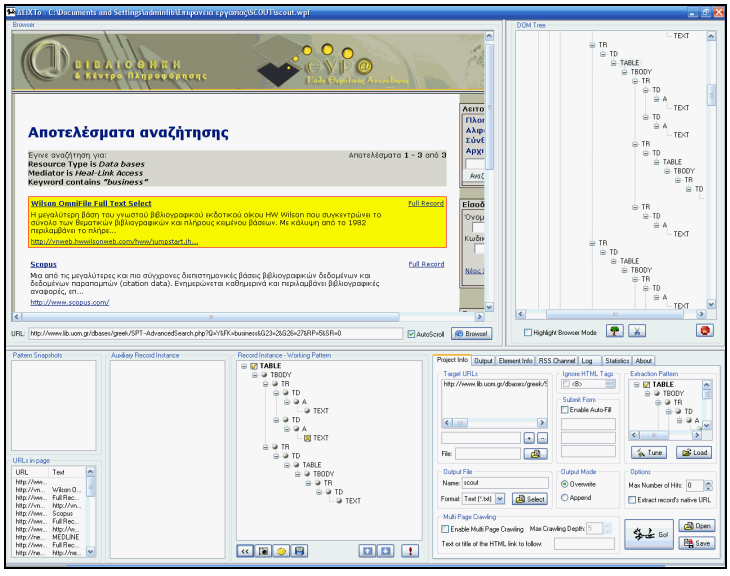
Εξαγωγή περιεχομένου από ιστοσελίδες βάσεων δεδομένων

Στο εργαλείο ταυτόχρονης αναζήτησης dbWiz, για τις περισσότερες επιθυμητές πηγές πληροφορίας η ανάκτηση των αποτελεσμάτων για το ερώτημα του χρήστη γίνεται μέσω της διεπαφής του δικτυακού τους τόπου. Η μεθοδολογία που χρησιμοποιεί το dbWiz σε αυτές τις περιπτώσεις βασίζεται στις κανονικές εκφράσεις και στον πηγαίο κώδικα των σελίδων. Ο μηχανισμός αυτός ωστόσο βελτιώθηκε σημαντικά με την ενσωμάτωση δυνατότητας εξαγωγής της επιθυμητής πληροφορίας από ιστοσελίδες με δενδροειδείς κανόνες εξαγωγής οι οποίοι κατασκευάζονται σχετικά εύκολα με πρωτότυπο εργαλείο λογισμικού που αναπτύχθηκε για το σκοπό αυτό.

Ένας *κανόνας εξαγωγής* (extraction rule) είναι μία απεικόνιση (mapping) που επιτρέπει την πλήρωση ενός αποθετηρίου δεδομένων (data repository) με πληροφορίες που περιέχονται σε μια ιστοσελίδα (Laender 2002) και λειτουργεί ουσιαστικά σαν *πρότυπο* (pattern) αναζήτησης χρήσιμης πληροφορίας εντός μιας σελίδας. Η σχετική εφαρμογή αναπτύχθηκε στο περιβάλλον αντικειμενοστραφούς (object-oriented) και οπτικού (visual) προγραμματισμού *Turbo Delphi Explorer*, το οποίο είναι ένα εντυπωσιακό εργαλείο ανάπτυξης λογισμικού για MS Windows με μεγάλες δυνατότητες. Στην Εικόνα 6**: Εργαλείο κατασκευής δενδροειδών κανόνων εξαγωγής** που ακολουθεί φαίνονται ορισμένα από τα τμήματα του παραθύρου της εφαρμογής αυτής.

Η ευκολία δημιουργίας κανόνων έγκειται στο εύχρηστο γραφικό περιβάλλον του εργαλείου και στην οπτικοποίηση (visualization) της διαδικασίας υπόδειξης του προτύπου εξαγωγής μέσω του ενσωματωμένου προγράμματος πλοήγησης. Οι παραγόμενοι κανόνες εξαγωγής είναι αρκετά ευέλικτοι και παρουσιάζουν βελτιωμένη ανθεκτικότητα σε παραλλαγές της δομής των επιθυμητών αντικειμένων πληροφορία. Επίσης, οι κανόνες επιδεικνύουν ανεκτικότητα σε μικρές μεταβολές των ιστοσελίδων. Στην πλειοψηφία των περιπτώσεων είναι δυνατή αλλά και εύκολη η επεξεργασία και τροποποίηση ενός υπάρχοντος κανόνα εξαγωγής, ο οποίος για κάποιους λόγους σταμάτησε να λειτουργεί απόλυτα ικανοποιητικά.

Ένας κανόνας εξαγωγής παραγόμενος από το εργαλείο αποθηκεύεται σε ένα XML αρχείο το οποίο περιγράφει την HTML δομή που ακολουθούν τα στιγμιότυπα επιθυμητής πληροφορίας. Για τη χρήση τέτοιων κανόνων και εξαγωγή των επιθυμητών δεδομένων χρειάστηκε η κατασκευή μιας Perl λειτουργικής μονάδας, ικανής να εκτελεί τους κανόνες. Με τον τρόπο αυτό κατέστη δυνατή η χρήση τέτοιων κανόνων σε λειτουργικές μονάδες αναζήτησης του dbWiz, ώστε να εξάγονται τα αποτελέσματα από τις σελίδες των πηγών.



Εικόνα 6: Εργαλείο κατασκευής δενδροειδών κανόνων εξαγωγής

Το εργαλείο, βασίζεται στο *Μοντέλο Αντικειμένου Εγγράφου* (Document Object Model ή DOM) το οποίο αποτελεί Σύσταση της Κοινοπραξίας για τον Παγκόσμιο Ιστό (W3C). Διευκολύνει αρκετά τη δημιουργία, δοκιμή, ρύθμιση και τη συντήρηση αποτελεσματικών κανόνων εξαγωγής για τις ιστοσελίδες των επιθυμητών πηγών και το κυριότερο επιτυγχάνει υψηλή ακρίβεια αποτελεσμάτων. Αξίζει να σημειωθεί πως για πηγές του Εθνικού Κέντρου Τεκμηρίωσης (ΕΚΤ) και την πύλη θεματικής αναζήτησης (ΘΥΡΑ) της βιβλιοθήκης χρησιμοποιήθηκε το εργαλείο αυτό και κατασκευάστηκαν αντίστοιχοι κανόνες εξαγωγής.

Συμπεράσματα

Στο πλαίσιο ανάπτυξης εφαρμογής ταυτόχρονης αναζήτησης σε πολλαπλές ανεξάρτητες πηγές δεδομένων για τη Βιβλιοθήκη του Πανεπιστημίου Μακεδονίας, επιλέχθηκε το εργαλείο dbWiz το οποίο είναι ένα ιδιαίτερα αξιόλογο έργο λογισμικού ανοικτού κώδικα. Το εργαλείο έχει εγκατασταθεί σε σταθμό εργασίας της βιβλιοθήκης και βρίσκεται σε πιλοτική εφαρμογή. Το σύστημα είναι ακόμα υπό ανάπτυξη και δοκιμάζονται τα χαρακτηριστικά, οι παράμετροι και οι λειτουργίες του ώστε να καλυφθούν πλήρως οι ανάγκες της βιβλιοθήκης. Έχουν ήδη κατασκευαστεί αρκετές νέες λειτουργικές μονάδες αναζήτησης για πηγές που δεν είναι προ-εγκατεστημένες στο dbWiz καθώς και έχουν τροποποιηθεί κατάλληλα κάποιες από τις έτοιμες. Στην τρέχουσα φάση υποστηρίζεται η αναζήτηση σε αρκετές από τις δημοφιλέστερες πηγές πληροφορίας και μελλοντικά θα προστεθούν όσο το δυνατόν περισσότερες. Τα πρώτα δείγματα γραφής του συστήματος πάντως είναι ενθαρρυντικά και η απόδοση του κρίνεται αρκετά ικανοποιητική.

Η μεγάλη πρόκληση για τη βιβλιοθήκη είναι η εκπαίδευση των χρηστών. Για το σκοπό αυτό η υπηρεσία παράλληλης αναζήτησης πρόκειται να ενταχθεί στο πρόγραμμα πληροφοριακής παιδείας της βιβλιοθήκης ώστε να ενημερωθούν οι χρήστες για το εργαλείο και να εκπαιδευθούν στη χρήση του. Πρέπει επίσης να δοθεί ιδιαίτερο βάρος στον τομέα του μάρκετινγκ ώστε να γίνει γνωστή η αξιόλογη αυτή υπηρεσία της βιβλιοθήκης. Αξίζει να δοθεί έμφαση στη δυνατότητα μελλοντικής συνεργατικής λειτουργίας του με άλλα συστήματα λογισμικού. Η πληρότητα και ο βαθμός ολοκλήρωσης του συστήματος ενδεχομένως να βοηθήσουν στην ενοποίηση (integration) υπαρχόντων υπηρεσιών που τώρα λειτουργούν σε αυτόνομη βάση. Τέλος, το μοντέλο ανοικτού λογισμικού σίγουρα θα μπορούσε να ωφελήσει σημαντικά τις ελληνικές ακαδημαϊκές βιβλιοθήκες οι οποίες θα μπορούσαν να συνεργαστούν μεταξύ τους για να καλύψουν πολλές κοινές τους ανάγκες για επαρκή και αποτελεσματική πρόσβαση σε σημαντικές πηγές πληροφορίας.

Βιβλιογραφικές Παραπομπές

Boyd, J., Hampton, M., Morrison, P., Pugh, P. & Cervone, F. (2006) The One-Box Challenge: Providing a Federated Search That Benefits the Research Process. *Serials Review*, 32 (4) December, pp.247-254.

Chen, X. (2006) Metalib, WebFeat, and Google: The strengths and weaknesses of federated search engines compared with Google. *Online Information Review*, 30 (4), pp.413–427.

Cox, A. (2003) Choosing a library Portal System. *VINE: The Journal of Information and Knowledge Management Systems*, 33 (1), pp.37-41.

Laender A., Ribeiro-Neto, B., da Silva, A.S. & Teixeira, J. (2002). A Brief Survey of Web Data Extraction Tools. *ACM SIGMOD Record*, 31 (2), pp.84-93.

Luther, J. (2003) Trumping Google? Metasearching's promise. *Library Journal*, 128 (16), pp. 36–39.

Mah, C. & Stranack, K. (2005) dbWiz: Open Source Federated Searching for Academic Libraries. *Library Hi Tech*, 23 (4), pp.490-503.

Marshall, P., Herman, S. & Rajan, S. (2006) In Search of More Meaningful Search. *Serials Review*, 32 (3) September, pp.172-180.