Καθιστώντας μια υπηρεσία θεματικής πλοήγησης στο διαδίκτυο συμβατή με τις τεχνολογίες των συνδεδεμένων δεδομένων

Κωνσταντίνος Κυπριανός[1] και Ιωάννης Παπαδάκης[1]
[1] Ιόνιο Πανεπιστήμιο, Τμήμα Αρχειονομίας & Βιβλιοθηκονομίας
Ιωάννου Θεοτόκη 72, Κέρκυρα, 49100
k.kyprianos@gmail.com, papadakis@ionio.gr

Περίληψη

Στις μέρες μας, με την εμφάνιση του σημασιολογικού ιστού ένας μεγάλος αριθμός βιβλιοθηκών παρέχουν μέρος των πληροφοριών τους ως ανοιχτά συνδεδεμένα δεδομένα (linked open data – LOD). Θεματικές επικεφαλίδες καθώς και όροι θησαυρών παρέχονται πλέον online μέσω του διαδικτύου με τη χρήση κατάλληλων σημασιολογικών τεχνολογιών, όπως είναι τα triplestores των συνδεδεμένων δεδομένων και οι αντίστοιχες υπηρεσίες αναζήτησης με ερωτήματα SPARQL. Ακολουθώντας λοιπόν, το παράδειγμα βιβλιοθηκών του εξωτερικού, αποφασίστηκε η μετατροπή της υπηρεσίας διαδραστικής θεματικής πλοήγησης που παρέχεται από την ψηφιακή βιβλιοθήκη Μεταπτυχιακών και Διδακτορικών Διατριβών του Πανεπιστημίου Πειραιά, σε υπηρεσία θεματικής πλοήγησης με τεχνολογίες συμβατές με τα συνδεδεμένα δεδομένα. Η υπηρεσία αυτή μέσω του γραφικού περιβάλλοντος δίνει τη δυνατότητα στους χρήστες να πλοηγηθούν στις θεματικές επικεφαλίδες που έχει η ψηφιακή βιβλιοθήκη μέσω ευρύτερων και στενότερων όρων καθώς και μέσω των τυχόν κοινών υποδιαιρέσεων που μπορεί αυτές να έχουν. Παράλληλα, οι χρήστες μπορούν να πάρουν πληροφορίες και από άλλα αποθετήρια που παρέχουν τα δεδομένα τους ως συνδεδεμένα δεδομένα (π.χ. άρθρα των New York Times).
Η παρούσα εργασία διαρθρώνεται ως εξής: Αρχικά, αναλύονται οι βασικές έννοιες και τα δομικά συστατικά που είναι απαραίτητα για τη συμμετοχή στην κοινότητα των συνδεδεμένων δεδομένων. Στη συνέχεια παρουσιάζεται το παράδειγμα της Βιβλιοθήκης του Κογκρέσου, η οποία από το 2008 παρέχει τη θεματική της πληροφορία ως ανοιχτά συνδεδεμένα δεδομένα. Κατόπιν, αναλύεται η διαδικασία που ακολουθήθηκε για τη δημιουργία του τοπικού αποθετηρίου των συνδεδεμένων δεδομένων, καθώς και η σύνδεσή του με άλλες πηγές. Έπειτα, παρουσιάζεται το γραφικό περιβάλλον της εφαρμογής και ολοκληρώνεται η εργασία με μερικά γενικά συμπεράσματα.

Λέξεις κλειδιά: Ανοιχτά συνδεδεμένα δεδομένα, Ψηφιακή Βιβλιοθήκη Πανεπιστημίου Πειραιά, New York Times, Θεματικές Επικεφαλίδες της Βιβλιοθήκης του Κογκρέσου

Abstract

Nowadays, an ever-increasing amount of libraries provide their data as linked open data - LOD. Subject headings and thesauri terms are provided online via Internet with the employment of adequate semantic technologies, such as triplestores of LOD and their corresponding SPARQL endpoints. Following the examples of major libraries worldwide, we decided to convert the existing service of interactive information retrieval that is provided by the University of Piraeus digital library of Thesis and Dissertations into a service compatible with LOD. The service through the Graphical User Interface – GUI gives the opportunity to the end-users to navigate to the subject headings of the digital library through the employment of broader and narrower terms and as well through the employment of any common subdivision that may share. At the same time, the users can find information from other repositories that provide their data as well as LOD (i.e. the articles' database of New York Times).

This paper is structured as follows: Initially, the basic principles and the components that are necessary to participate in the LOD cloud are analyzed. Then, the example of Library of Congress is presented, which provides subject-based information as LOD since 2008. The next section presents the process that was followed to create the local LOD triplestore and demonstrates its connection with other LOD-compliant data sources. Finally, the GUI of the service is presented, followed by some general conclusions.

Keywords: Linked Open Data, University of Piraeus digital library, New York Times, Library of Congress Subject Headings

1. Εισαγωγή

Στις μέρες μας, ένας μεγάλος αριθμός σημασιολογικής πληροφορίας των βιβλιοθηκών παρέχεται ως πληροφορία βασισμένη στα ανοιχτά συνδεδεμένα δεδομένα. Πολλά αποθετήρια γνώσης και ειδικότερα βιβλιοθήκες συνειδητοποίησαν από νωρίς τα ωφέλη από την υιοθέτηση τεχνολογιών βασισμένων στα συνδεδεμένα δεδομένα. Έτσι, οι οργανισμοί αυτοί θα επωφεληθούν όχι μόνο από την εξεύρεση λύσεων όσον αφορά στο πρόβλημα της διαλειτουργικότητας που ταλανίζει την κοινότητα των βιβλιοθηκών εδώ και πολλές δεκαετίες [4], αλλά και από τη δημιουργία της απαραίτητης υποδομής για την εισαγωγή νέων υπηρεσιών προστιθέμενης αξίας. Αυτές οι υπηρεσίες θα δίνουν τη δυνατότητα στις βιβλιοθήκες να ενσωματώνουν τη σημασιολογική πληροφορία που υπάρχει στις

παραδοσιακές τους υπηρεσίες, όπως είναι ο κατάλογος ανοιχτής πρόσβασης (OPAC – Open Public Access Catalog) με πληροφορίες που προέρχονται από τρίτους, όπως είναι τα αρχεία βίντεο και εικόνων, ή βάσεις γνώσης (π.χ. DBpedia[1], Freebase[2] κλπ.)

Ένα σημαντικό είδος σημασιολογικής πληροφορίας που υπάρχει στις βιβλιοθήκες είναι η θεματική περιγραφή των συλλογών τους. Παραδοσιακά, για τη θεματική περιγραφή του υλικού που υπάρχει στις βιβλιοθήκες χρησιμοποιούνται δύο συστήματα θεματικής επεξεργασίας: οι θεματικές επικεφαλίδες και οι θησαυροί. Οι θεματικές επικεφαλίδες γενικά είναι δομημένες φράσεις που συνιστούν ένα ελεγχόμενο λεξιλόγιο. Κάθε τεκμήριο της συλλογής περιγράφεται θεματικά από μία ή περισσότερες θεματικές επικεφαλίδες. Οι θεματικές επικεφαλίδες καθορίζουν σχέσεις μεταξύ τους για να καταδείξουν τη συνωνυμία και τη θεματική συγγένεια. Από την άλλη μεριά, οι θησαυροί είναι και αυτοί ελεγχόμενα λεξιλόγια που έχουν ιεραρχική δομή και καθορίζουν σχέσεις μεταξύ των όρων τους (περιγραφείς). Πιο συγκεκριμένα, οι σχέσεις μεταξύ των όρων ενός θησαυρού καταδεικνύουν όχι μόνο τη συνωνυμία και τη θεματική συγγένεια, αλλά και σχέσεις γένους-είδους και όλου-μέρους [1].

Στην παρούσα εργασία, περιγράφεται μια υπηρεσία θεματικής πλοήγησης βασισμένη στις τεχνολογίες των ανοιχτών συνδεδεμένων δεδομένων που είναι ικανή να ενσωματώνει πληροφορίες από διάφορα αποθετήρια. Πιο συγκεκριμένα, η εργασία αυτή περιγράφει μια υπηρεσία για την ψηφιακή βιβλιοθήκη μεταπτυχιακών και διδακτορικών διατριβών του Πανεπιστημίου Πειραιώς, η οποία όχι μόνο παρέχει τη θεματική πληροφορία (θεματικές επικεφαλίδες) στην ευρύτερη κοινότητα των ανοιχτών συνδεδεμένων δεδομένων, αλλά παρέχει στους τελικούς της χρήστες πηγές που προέρχονται από άλλα αποθετήρια (π.χ. άρθρα από τους New York Times) μέσω της συσχέτισης των θεματικών δεδομένων των δύο αποθετηρίων.

Η υπόλοιπη εργασία έχει την ακόλουθη δομή: στην επόμενη ενότητα αναλύονται οι βασικές έννοιες και τα απαραίτητα συστατικά που απαιτούνται για τη συμμετοχή στην κοινότητα των ανοιχτών συνδεδεμένων δεδομένων. Στη συνέχεια, παρουσιάζεται το παράδειγμα της Βιβλιοθήκης του Κογκρέσου, η οποία από το 2008 παρέχει τη θεματική της πληροφορία ως ανοιχτά συνδεδεμένα δεδομένα. Έπειτα, παρουσιάζεται η διαδικασία που ακολουθήθηκε για τη δημιουργία του τοπικού αποθετηρίου των συνδεδεμένων δεδομένων. Στη συνέχεια, παρουσιάζεται το γραφικό περιβάλλον της εφαρμογής και δίνονται περισσότερες λεπτομέρειες σχετικά με συγκεκριμένες τεχνολογίες των ανοικτών συνδεδεμένων δεδομένων όπως είναι το αποθετήριο τριπλετών – triplestore και το αντίστοιχο σημείο ερωτημάτων – SPARQL endpoint. Τέλος, η εργασία ολοκληρώνεται με μερικά γενικά συμπεράσματα.

[1] DBpedia Διαθέσιμο στο: http://dbpedia.org/ Ημερομηνία πρόσβασης: 20/09/2012

[2] Freebase Διαθέσιμο στο: http://www.freebase.com/ Ημερομηνία πρόσβασης: 20/09/2012

2. Αποσαφήνιση εννοιών

Ο όρος 'συνδεδεμένα δεδομένα' αναφέρεται σε ένα σύνολο αρχών και τεχνικών για τη δημοσίευση δομημένων δεδομένων στο διαδίκτυο έτσι ώστε αυτά να είναι πιο εύκολα προσβάσιμα και πιο χρήσιμα [11]. Τα συνδεδεμένα δεδομένα βασίζονται πάνω σε δύο τεχνολογίες: α) στα Ενιαία Αναγνωριστικά Πόρων (Uniform Resource Identifiers – URIs) και β) στο Πρωτόκολλο Μεταφοράς Υπερκειμένου (HyperText Transfer Protocol – HTTP). Τα URIs και το HTTP υλοποιούνται στα συνδεδεμένα δεδομένα με την τεχνολογία του προτύπου Resource Description Framework – RDF[3]. Το RDF είναι το πρότυπο για την κωδικοποίηση μεταδεδομένων και γενικά της γνώσης στον σημασιολογικό ιστό. Παρέχει ένα γενικό μοντέλο δεδομένων με το οποίο δομούνται και συνδέονται τα δεδομένα που περιγράφουν αντικείμενα στον κόσμο. Η κωδικοποίηση των δεδομένων, σύμφωνα με το RDF, γίνεται με τη μορφή τριπλετών (triples): 'υποκείμενο, κατηγόρημα, αντικείμενο'. Στη γενική τους μορφή, και τα τρία συστατικά της τριπλέτας αναπαρίστανται ως URIs. Το κατηγόρημα καθορίζει το πώς συνδέονται μεταξύ τους το υποκείμενο και το αντικείμενο. Πολλές φορές, το αντικείμενο μπορεί να είναι απλό κείμενο. Λόγω του γεγονότος ότι το RDF παρέχει ένα γενικό μοντέλο δεδομένων, έχουν δημιουργηθεί διάφορα λεξιλόγια για την περιγραφή αντικειμένων του πραγματικού κόσμου (π.χ. για την περιγραφή της θεματικής πληροφορίας έχει δημιουργηθεί η οντολογία Simple Knowledge Organization System – SKOS[4]). Επίσης, επειδή το RDF δεν περιγράφει ένα μορφότυπο δεδομένων αλλά ένα μοντέλο δεδομένων, για να δημοσιευθεί στο διαδίκτυο θα πρέπει να σειριοποιηθεί. Κάποια πρότυπα που έχουν δημιουργηθεί για τη σειριοποίηση είναι το RDF/XML[5], το RDFa[6], το Turtle[7] και το N-triple[8]. Τα δεδομένα που έχουν δημιουργηθεί με το RDF αποθηκεύονται σε αποθετήρια τριπλετών (triplestores). Η κυρίαρχη γλώσσα ερωτημάτων για την ανάκτηση πληροφοριών από τα triplestores είναι η SPARQL[9].

Στη συνέχεια, περιγράφεται η υπηρεσία ανοιχτών συνδεδεμένων δεδομένων της Βιβλιοθήκης του Κογκρέσου.

---

[3] Resource Description Framework (RDF) Διαθέσιμο στο: http://www.w3.org/RDF/ Ημερομηνία πρόσβασης: 25/08/2012

[4] SKOS Simple Knowledge Organization System Διαθέσιμο στο: http://www.w3.org/2004/02/skos/ Ημερομηνία Πρόσβασης: 25/08/2012

[5] RDF/XML Syntax Specification (Revised) Διαθέσιμο στο: http://www.w3.org/TR/REC-rdf-syntax/ Ημερομηνία πρόσβασης: 25/08/2012

[6] Rich Structured Data Markup for Web Documents Διαθέσιμο στο: http://www.w3.org/TR/xhtml-rdfa-primer/ Ημερομηνία πρόσβασης: 25/08/2012

[7] RDF Primer — Turtle version Διαθέσιμο στο: http://www.w3.org/2007/02/turtle/primer/ Ημερομηνία πρόσβασης: 25/08/2012

[8] Notation 3 Logic Διαθέσιμο στο: http://www.w3.org/DesignIssues/Notation3.html Ημερομηνία Πρόσβασης: 25/08/2012

[9] SPARQL Query Language for RDF Διαθέσιμο στο: http://www.w3.org/TR/rdf-SPARQL-query/ Ημερομηνία Πρόσβασης: 25/08/2012

3. Το παράδειγμα της Βιβλιοθήκης του Κογκρέσου

Το πιο γνωστό και ευρέως χρησιμοποιούμενο εργαλείο για τη θεματική περιγραφή του υλικού των βιβλιοθηκών είναι οι Θεματικές Επικεφαλίδες της Βιβλιοθήκης του Κογκρέσου (Library of Congress Subject Headings - LCSH[10]). Οι LCSH αποτελούν ένα ελεγχόμενο λεξιλόγιο, που υποστηρίζεται από τη Βιβλιοθήκη του Κογκρέσου, για τη θεματική περιγραφή των βιβλιογραφικών εγγραφών. Οι LCSH βρίσκουν εφαρμογή σε κάθε τεκμήριο και γενικά στο υλικό που υπάρχει στη συλλογή μιας βιβλιοθήκης και βοηθούν τους χρήστες στο να έχουν πρόσβαση σε τεκμήρια που έχουν το ίδιο θέμα. Οι LCSH υποστηρίζονται ενεργά από το 1898. Λόγω της συνέπειας που έχουν δείξει μέσα στο χρόνο, οι LCSH έχουν υιοθετηθεί από πολλές βιβλιοθήκες των ΗΠΑ, αλλά και από άλλες βιβλιοθήκες ανά τον κόσμο, πολλές φορές σε μετάφραση. Είναι εμφανές, λοιπόν, ότι οι LCSH κυριάρχησαν τον τελευταίο αιώνα στον τομέα των βιβλιοθηκών ως το «ντε φάκτο» εργαλείο για τη θεματική περιγραφή των τεκμηρίων και γενικά των πηγών που υπάρχουν στις βιβλιοθήκες [2]. Ακόμα και σήμερα, παρά την εμφάνιση πολλών θησαυρών και άλλων ελεγχόμενων λεξιλογίων που θεωρούνται πιο κατάλληλα για χρήση στο online περιβάλλον [3], ακόμα μεγάλος αριθμός OPAC και ψηφιακών βιβλιοθηκών ανά τον κόσμο υιοθετούν τις LCSH. Μάλιστα, το γεγονός ότι από το 2008, οι LCSH δημοσιεύονται και ως ανοιχτά συνδεδεμένα δεδομένα[11] είναι ενδεικτικό της σημασίας που δίνει η Βιβλιοθήκη του Κογκρέσου στην υιοθέτηση νέων τεχνολογιών που θα τη βοηθήσει στην παροχή καλύτερων και ποιοτικότερων υπηρεσιών προς τους χρήστες και τις άλλες βιβλιοθήκες. Επίσης, σε πειραματικό στάδιο λειτουργεί και το SPARQL endpoint[12] των LCSH, το οποίο είναι σχεδιασμένο να χρησιμοποιείται από άλλες εφαρμογές που χρησιμοποιούν τις LCSH. Με αυτόν τον τρόπο οι βιβλιοθήκες μπορούν να πάρουν έτοιμα, αξιόπιστα και μοναδικά URI για κάθε θεματική επικεφαλίδα. Η βιβλιοθήκη του Κογκρέσου, για τη δημοσίευσή των θεματικών επικεφαλίδων ως ανοιχτά συνδεδεμένα δεδομένα, χρησιμοποιεί το λεξιλόγιο της οντολογίας SKOS [5].

4. Η υπηρεσία

Ακολουθώντας το παράδειγμα της Βιβλιοθήκης του Κογκρέσου και άλλων μεγάλων βιβλιοθηκών, δημιουργήθηκε μια υπηρεσία θεματικής πλοήγησης βασισμένης στα ανοιχτά συνδεδεμένα δεδομένα[13]. Η υπηρεσία αυτή είναι ικανή να ενσωματώνει πηγές που προέρχονται από την ψηφιακή βιβλιοθήκη Μεταπτυχιακών και Διδακτορικών Διατριβών του Πανεπιστημίου Πειραιά[14] που βασίζεται στο DSpace και από τη βάση άρθρων των New

---

[10] Library of Congress Authorities Διαθέσιμο στο: http://authorities.loc.gov Ημερομηνία Πρόσβασης: 27/08/2012

[11] LC Linked Data Service: Authorities and Vocabularies Διαθέσιμο στο: http://id.loc.gov/ Ημερομηνία πρόσβασης: 25/08/2012

[12] LSCH-info provided by the Talis platform. Διαθέσιμο στο: http://api.talis.com/stores/lcsh-info/services/SPARQL Ημερομηνία πρόσβασης: 25/08/2012

[13] Υπηρεσία θεματικής πλοήγησης βασισμένης στα ανοιχτά συνδεδεμένα δεδομένα. Διαθέσιμο στο: http://neel.cs.unipi.gr/entry/ Ημερομηνία πρόσβασης: 25/08/2012

[14] Πανεπιστήμιο Πειραιώς: Ψηφιακή Βιβλιοθήκη: Dspace. Διαθέσιμο στο: http://digilib.lib.unipi.gr/dspace/ Ημερομηνία πρόσβασης: 25/08/2012

York Times – NYT[15]. Η υπηρεσία αυτή παρέχει ένα Γραφικό Περιβάλλον Χρήσης (Graphical User Interface – GUI) για την πλοήγηση στις θεματικές επικεφαλίδες που συγκροτούν το αποθετήριο των ανοιχτών συνδεδεμένων δεδομένων.

Καθώς ο χρήστης επιλέγει μια συγκεκριμένη θεματική επικεφαλίδα, εμφανίζεται μια λίστα, η οποία περιέχει τις μεταπτυχιακές και διδακτορικές διατριβές καθώς και τα άρθρα των NYT που περιγράφονται με τη συγκεκριμένη θεματική επικεφαλίδα. Η θεματική πλοήγηση καταφέρνει όχι μόνο να εμφανίσει τις θεματικές επικεφαλίδες ως ανοιχτά συνδεδεμένα δεδομένα αλλά επιτυγχάνει στο να παρέχει στους χρήστες και επιπλέον σχετικές πληροφορίες και πηγές που προέρχονται από ένα εξωτερικό triplestore (π.χ. NYT triplestore).

Η ψηφιακή βιβλιοθήκη που βασίζεται στο DSpace αναφέρεται στην ψηφιακή βιβλιοθήκη Μεταπτυχιακών και Διδακτορικών Διατριβών του Πανεπιστημίου Πειραιά. Η ψηφιακή βιβλιοθήκη περιέχει περίπου 3,400 θεματικές επικεφαλίδες στα ελληνικά και στα αγγλικά. Οι θεματικές επικεφαλίδες σχετίζονται μεταξύ τους χρησιμοποιώντας εκτός από την παραδοσιακή συνδετική δομή που υπάρχει ανάμεσα στις θεματικές επικεφαλίδες (π.χ. ευρύτερος όρος (Broader Term – BT), στενότερος όρος (Narrower Term – NT) και σχετικός όρος (Related Term – RT), και την εκτεταμένη συνδετική δομή. Όπως αναφέρεται στο [7], η εκτεταμένη συνδετική δομή εκφράζεται ως η σχέση που υπάρχει ανάμεσα σε θεματικές επικεφαλίδες που μοιράζονται την ίδια υποδιαίρεση. Επομένως, όποιες θεματικές επικεφαλίδες μοιράζονται την ίδια υποδιαίρεση σχετίζονται μεταξύ τους με μια σχέση που το όνομά της είναι ίδιο με αυτής της κοινής υποδιαίρεσης.

Οι πηγές που απαρτίζουν την ψηφιακή βιβλιοθήκη χαρακτηρίζονται από μια ή περισσότερες θεματικές επικεφαλίδες σύμφωνα με το περιεχόμενό τους. Οι θεματικές επικεφαλίδες ακολουθούν τις οδηγίες των LCSH και είναι στα ελληνικά, στα αγγλικά ή και στις δύο γλώσσες. Η συλλογή της ψηφιακής βιβλιοθήκης αποτελείται από μεταπτυχιακές και διδακτορικές διατριβές των τμημάτων του Πανεπιστημίου του Πειραιά. Πιο συγκεκριμένα, οι θεματικές περιοχές που καλύπτει είναι: α) Οικονομία, β) Οργάνωση και Διοίκηση Επιχειρήσεων, γ) Στατιστική, δ) Βιομηχανική Διοίκηση και Τεχνολογία, ε) Χρηματοοικονομική και Τραπεζική, στ) Ναυτιλιακά, ζ) Πληροφορική και η) Διεθνείς και Ευρωπαϊκές Σπουδές.

Σε μια προσπάθεια εμπλουτισμού των πηγών της ψηφιακής βιβλιοθήκης, αποφασίστηκε η χρήση του εργαλείου NYT api[16] έτσι ώστε να παρέχονται στον τελικό χρήστη άρθρα των NYT σχετικά με το θέμα που έχει επιλέξει. Για να επιτευχθεί αυτό, οι θεματικές επικεφαλίδες της ψηφιακής βιβλιοθήκης συσχετίστηκαν/ ευθυγραμμίστηκαν με τις θεματικές επικεφαλίδες των NYT. Η ηλεκτρονική βάση δεδομένων των NYT αποτελείται από άρθρα που χρονολογούνται από το 1981. Σχετικά πρόσφατα, οι NYT δημιούργησαν ένα ευρετηριασμένο λεξιλόγιο, το οποίο είναι διαθέσιμο στο ευρύ κοινό ως ανοιχτά συνδεδεμένα

[15] The New York Times. Διαθέσιμο στο: http://www.nytimes.com/ Ημερομηνία πρόσβασης: 25/08/2012

[16] NYT api tool. Διαθέσιμο στο: http://prototype.nytimes.com/gst/apitool/index.html Ημερομηνία πρόσβασης: 25/08/2012

δεδομένα[17]. Αυτό το λεξιλόγιο περιέχει ένα μεγάλο αριθμό (περίπου 10.000) θεματικών επικεφαλίδων που περιγράφει τη συλλογή άρθρων των NYT καθώς και το Αρχείο τους που περιέχει μεγάλο αριθμό φωτογραφιών, γραφικών, ηχητικών ντοκουμέντων, βίντεο κ.α.

Στις επόμενες ενότητες, παρουσιάζεται η διαδικασία που ακολουθήθηκε για τη δημιουργία των URIs, των σχέσεων ανάμεσα στις θεματικές επικεφαλίδες με τη βοήθεια του λεξιλογίου SKOS και του triplestore με το αντίστοιχο SPARQL endpoint του.

4.1. Μοναδικά URIs

Όπως ειπώθηκε από τον Tim Berners-Lee [6], βασική προϋπόθεση για τη συμμετοχή στο κίνημα των ανοιχτών συνδεδεμένων δεδομένων είναι η χρήση μοναδικών URIs. Επομένως, κάθε μια θεματική επικεφαλίδα της περιγραφόμενης υπηρεσίας θα πρέπει να αντιστοιχεί σε ένα και μοναδικό URI. Η πλειονότητα των θεματικών επικεφαλίδων που υπάρχουν στην ψηφιακή βιβλιοθήκη ανήκει στις LCSH. Για αυτές τις θεματικές επικεφαλίδες αποφασίστηκε η διατήρηση του μοναδικού URI που είχε αρχικά ανατεθεί από τη Βιβλιοθήκη του Κογκρέσου, αντί να δημιουργηθεί ένα νέο τοπικό URI. Η απόφαση αυτή πάρθηκε για να αποφύγουμε τα προβλήματα που μπορεί να δημιουργηθούν από τη χρήση του <owl:sameas> [8].

Επίσης, για τις ανάγκες της ευρετηρίασης του υλικού της ψηφιακής βιβλιοθήκης, δημιουργήθηκαν επιπλέον θεματικές επικεφαλίδες. Παρ' όλο που αυτές οι θεματικές επικεφαλίδες ακολουθούν τους κανόνες των LCSH, δεν υπάρχουν αυτούσιες στο λεξιλόγιο των LCSH. Ως εκ τούτου, σε κάθε μια από αυτές τις θεματικές επικεφαλίδες δόθηκε ένα νέο URI. Αυτά τα URIs ανήκουν στον τοπικό χώρο ονομάτων (μετάφ. namespace) της υπηρεσίας (http://id.lib.unipi.gr/authorities/subjects/) και διευκρινίζονται (μετάφ. dereferenced) μέσω του τοπικού SPARQL endpoint[18] της υπηρεσίας.

4.2. Ανοιχτά συνδεδεμένα δεδομένα που βασίζονται στο SKOS

Για τις ανάγκες τις περιγραφόμενης υπηρεσίας, αποφασίστηκε η υιοθέτηση του SKOS λεξιλογίου για τη δημιουργία του αποθετηρίου των τοπικών ανοιχτών συνδεδεμένων δεδομένων. Το αποθετήριο αυτό αποτελείται από τις ευθυγραμμισμένες θεματικές επικεφαλίδες και τις συσχετίσεις τους που ανήκουν στους χώρους ονομάτων των NYT, των LCSH και των τοπικών θεματικών επικεφαλίδων. Το λεξιλόγιο SKOS επιλέχθηκε λόγω της δυνατότητας εφαρμογής του στη μοντελοποίηση θεματικής πληροφορίας και στην ευρεία χρήση του από άλλες συναφείς υπηρεσίες.

Πιο συγκεκριμένα, σύμφωνα με τις προδιαγραφές του SKOS, οι θεματικές επικεφαλίδες μοντελοποιούνται ως έννοιες. Προτιμώμενοι και μη προτιμώμενοι όροι μπορούν να εκφραστούν σε οποιαδήποτε γλώσσα και είναι όλοι στιγμιότυπα της ίδιας έννοιας [9]. Οι

---

[17] The New York Times: Linked Open Data. Διαθέσιμο στο: http://data.nytimes.com Ημερομηνία πρόσβασης: 25/08/2012

[18] SPARQL endpoint της ψηφιακής βιβλιοθήκης του Πανεπιστημίου Πειραιώς. Διαθέσιμο στο: http://neel.cs.unipi.gr/endpoint/ Ημερομηνία πρόσβασης: 25/08/2012

θεματικές επικεφαλίδες οργανώνονται ιεραρχικά σύμφωνα με τη συνδετική τους δομή, όπως δηλώνεται από τις LCSH. Ο πίνακας 1 συνοψίζει την ορολογία που χρησιμοποιήθηκε για τις ανάγκες της δημιουργίας του συγκεκριμένου αποθετηρίου ανοιχτών συνδεδεμένων δεδομένων.

Πίνακας 1: Λεξιλόγιο SKOS

Θεματικές Επικεφαλίδες	SKOS Λεξιλόγιο
Broader Term – Ευρύτερος Όρος	skos:broader
Narrower Term – Στενότερος Όρος	skos:narrower
Related Term – Σχετικός Όρος	skos:related
Use – Χρησιμοποίησε	skos:prefLabel
Use For – Χρησιμοποίησε Αντί	skos:altLabel

Επίσης, για την ευθυγράμμιση των θεματικών επικεφαλίδων της ψηφιακής βιβλιοθήκης και των NYT απαιτούνται δύο ακόμα σχέσεις από το SKOS λεξιλόγιο, του <skos:exactMatch> και του <skos:closeMatch>, οι οποίες θα αναλυθούν στη συνέχεια.

4.3. Διαδικασία ευθυγράμμισης

Για την ευθυγράμμιση των θεματικών επικεφαλίδων που προέρχονται από την τοπική ψηφιακή βιβλιοθήκη και από τους NYT, αποφασίστηκε να χρησιμοποιηθεί ο αλγόριθμος αντιστοίχισης οντολογιών που περιγράφεται στο [10]. Σύμφωνα με τον αλγόριθμο αυτό, δύο όροι θεωρούνται ισοδύναμοι, όταν είναι ακριβώς ίδιοι. Σε αυτή την περίπτωση, η ορολογία SKOS που χρησιμοποιείται για να περιγράψει μια τέτοια σχέση είναι η <skos:exactMatch>. Για παράδειγμα, η θεματική επικεφαλίδα "Physics" υπάρχει και στο χώρο ονομάτων των NYT και στο χώρο ονομάτων των LCSH. Επομένως, θα δημιουργηθεί η ακόλουθη τριπλέτα (triple):

```
<http://id.loc.gov/authorities/subjects/sh85101653>
<http://www.w3.org/2004/02/skos/core#exactMatch>
<http://data.nytimes.com/48662872284940183040> .
```

Επιπλέον, υπάρχουν πολλές περιπτώσεις στις οποίες δύο θεματικές επικεφαλίδες θεωρούνται μερικώς ισοδύναμες:

- ◊ Μια θεματική επικεφαλίδα είναι στον πληθυντικό και η άλλη στον ενικό (π.χ. η θεματική επικεφαλίδα "Bankruptcies" στους NYT και "Bankruptcy" στο DSpace).
- ◊ Παραλλαγή καταλήξεων, επιθεμάτων κλπ., μερικοί όροι μπορεί να είναι διαφορετικοί με την χρήση κάποιων επιπλέον χαρακτήρων (π.χ. η θεματική επικεφαλίδα "Iraq war (2003-)" στους NYT και η θεματική επικεφαλίδα "Iraq war, 2003-" στο DSpace).
- ◊ Μια θεματική επικεφαλίδα μπορεί να έχει διαφορετική σειρά λέξεων (π.χ. η θεματική επικεφαλίδα "Colleges and Universities" στους NYT και η θεματική επικεφαλίδα "Universities and colleges" στο DSpace).

◊ Μια θεματική επικεφαλίδα μπορεί να αντιστοιχεί σε μια σύνθετη θεματική επικεφαλίδα ή και το αντίστροφο (π.χ. η θεματική επικεφαλίδα "Advertising and marketing" στους NYT οι θεματικές επικεφαλίδες "Advertising" και "Marketing" στο DSpace).

Το SKOS λεξιλόγιο που χρησιμοποιείται για τη δήλωση αυτής της σχέσης είναι η <skos:closeMatch>. Μετά την εφαρμογή του αλγόριθμου, 207 θεματικές επικεφαλίδες βρέθηκαν κοινές (111 "closematch"και 96 "exactmatch") και στους τρεις χώρους ονομάτων (NYT, DSpace και LCSH).

4.4. Δημιουργία αποθετηρίου με τη βοήθεια του λογισμικού 4store

Το triplestore των θεματικών επικεφαλίδων δημιουργήθηκε με τη βοήθεια του λογισμικού 4store[19]. Το 4store παρέχει ένα αποθετήριο δεδομένων και μια μηχανή ερωτημάτων και είναι σχεδιασμένο για να λειτουργεί σε συστήματα LINUX. Πιο συγκεκριμένα, το 4store περιλαμβάνει έναν SPARQL HTTP protocol server, ο οποίος μπορεί να απαντήσει σε SPARQL ερωτήματα χρησιμοποιώντας το πρωτόκολλο ερωτημάτων SPARQL HTTP. Το 4store είναι σχεδιασμένο για την αποθήκευση και ανάκτηση τριπλετών. Επομένως, όλες οι θεματικές επικεφαλίδες αναπαρίστανται σε τριπλέτες.

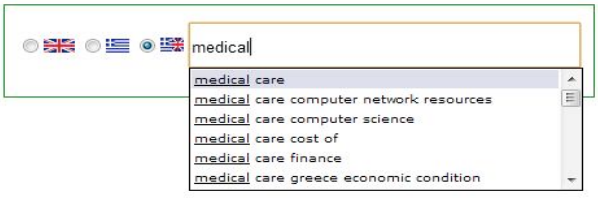
4.5. Γραφικό περιβάλλον χρήστη ή SPARQL endpoint;

Ο τελικός χρήστης έχει δύο επιλογές προβολής των αποτελεσμάτων των θεματικών επικεφαλίδων της υπηρεσίας: α) μέσω του γραφικού περιβάλλοντος χρήστη και β) μέσω του SPARQL endpoint.

α) γραφικό περιβάλλον χρήστη

Οι χρήστες αλληλεπιδρούν με την υπηρεσία μέσω ενός γραφικού περιβάλλοντος χρήσης που βασίζεται στις τεχνολογίες Ajax. Το γραφικό περιβάλλον βασίζεται σε προηγούμενη έκδοσή του που περιγράφεται στην εργασία [7]. Πιο συγκεκριμένα, το γραφικό περιβάλλον αποτελείται από τα ακόλουθα 3 δομικά συστατικά: α) το πεδίο αναζήτησης αυτόματης συμπλήρωσης (autosuggest search box), β) την πλοήγηση θεματικών επικεφαλίδων και γ) τα αποτελέσματα αναζήτησης από το DSpace και τους NYT.

Αρχικά, οι χρήστες καλούνται να εκφράσουν τις πληροφοριακές τους ανάγκες στα ελληνικά ή/ και στα αγγλικά (βλέπε εικ. 1).



Εικόνα 1: Πεδίο αυτόματης συμπλήρωσης

[19] 4store Διαθέσιμο στο: http://4store.org Ημερομηνία πρόσβασης: 26/08/2012

Όταν ο χρήστης επιλέξει μια θεματική επικεφαλίδα από τις προτεινόμενες, δημιουργείται ένα πλαίσιο κάτω από το πεδίο αναζήτησης αυτόματης συμπλήρωσης που περιέχει την επιλεγμένη θεματική επικεφαλίδα (βλέπε εικ. 2). Στη δίγλωσση έκδοση της υπηρεσίας, το πλαίσιο αποτελείται από την αγγλική ετικέτα της επιλεγμένης θεματικής επικεφαλίδας μαζί με την ελληνική της μετάφραση καθώς και με τους προτιμώμενους και μη προτιμώμενους όρους στα ελληνικά και στα αγγλικά αντίστοιχα.

Εικόνα 2: Πλαίσιο θεματικής επικεφαλίδας "Medical care" και θεματικό μενού (context menu) στενότερων όρων

Κάτω από τις ετικέτες των θεματικών επικεφαλίδων υπάρχουν δύο ετικέτες με το σήμα του μεγεθυντικού φακού για τη λειτουργία της «μεγέθυνσης» (+) και της «σμίκρυνσης» (-). Αυτές οι δύο ετικέτες αντιστοιχούν στους ευρύτερους και στενότερους όρους της επιλεγμένης θεματικής επικεφαλίδας. Με την επιλογή μιας από αυτές τις ετικέτες, ένα θεματικό μενού (context menu) σχεδιάζεται δίπλα στην ετικέτα, το οποίο περιέχει αυτές τις ευρύτερες ή στενότερες θεματικές επικεφαλίδες αντίστοιχα.

β) SPARQL endpoint

Στο SPARQL endpoint της υπηρεσίας (βλέπε εικ. 3) ο χρήστης θέτοντας τα κατάλληλα ερωτήματα σε SPARQL μπορεί να πάρει τα αντίστοιχα αποτελέσματα σε μορφή τριπλετών.

Εικόνα 3: Η αρχική οθόνη του SPARQL endpoint της υπηρεσίας

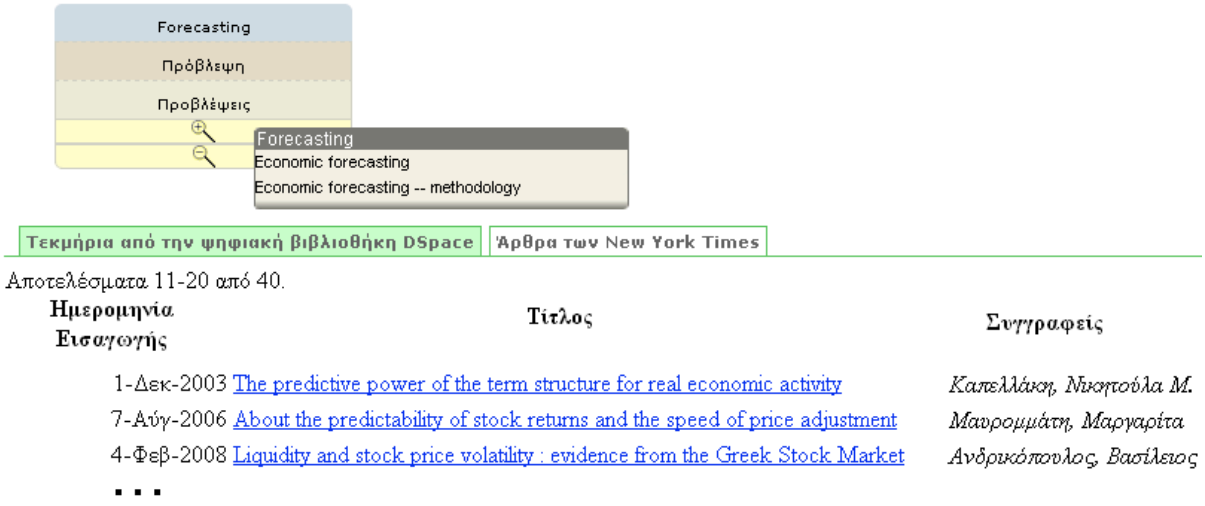
Για παράδειγμα, εάν ο χρήστης θέλει όλες τις θεματικές επικεφαλίδες που σχετίζονται με τη θεματική επικεφαλίδα "Forecasting" δεν έχει παρά να θέσει το ακόλουθο ερώτημα στο SPARQL endpoint:

```
SELECT DISTINCT * WHERE {{
<http://id.loc.gov/authorities/subjects/sh85050485> ?p ?o .}
UNION {?s ?pre <http://id.loc.gov/authorities/subjects/sh85050485> .}}
```

Η υπηρεσία θα του δώσει την ακόλουθη απάντηση:

<?xml version="1.0"?>
<SPARQL xmlns="http://www.w3.org/2005/SPARQL-results#">
...
1. <results><result>
<binding name="s"><uri>http://id.loc.gov/authorities/subjects/sh2009124595</uri></binding>
<binding name="pre"><uri>http://www.w3.org/2004/02/skos/core#broader</uri></binding>
2. </result><result>
<binding name="s"><uri>http://id.loc.gov/authorities/subjects/sh85040814</uri></binding>
<binding name="pre"><uri>http://www.w3.org/2004/02/skos/core#broader</uri></binding>
3. </result><result>
<binding name="p"><uri>http://www.w3.org/2004/02/skos/core#broader</uri></binding>
<binding name="o"><uri>owl:Thing</uri></binding>
4. </result><result>
<binding name="p"><uri>http://www.w3.org/2004/02/skos/core#altLabel</uri></binding>
<binding name="o"><literal xml:lang="GR">Προβλέψεις</literal></binding>
5. </result><result>
<binding name="p"><uri>http://www.w3.org/2004/02/skos/core#prefLabel</uri></binding>
<binding name="o"><literal xml:lang="GR">Πρόβλεψη</literal></binding>
6. </result><result>
<binding name="p"><uri>http://www.w3.org/2004/02/skos/core#prefLabel</uri></binding>
<binding name="o"><literal xml:lang="EN">Forecasting</literal></binding>
</result></results></SPARQL>

Ανάλογα με τη διάδραση που θα επιλέξει ο χρήστης κάθε φορά, η παραπάνω πληροφορία αναπαρίσταται γραφικά από το αντίστοιχο γραφικό περιβάλλον.



Εικόνα 4: Γραφικό περιβάλλον χρήστη με εμφάνιση και των σχετικών τεκμηρίων από το DSpace

Πιο συγκεκριμένα, τα αποτελέσματα 1 και 2 του ερωτήματος αντιστοιχούν σε 2 θεματικές κεφαλίδες που βρίσκονται κάτω από την ευρύτερη θεματική 'Forecasting' (skos:broader). Στο γραφικό περιβάλλον οι ετικέτες των 2 αυτών κεφαλίδων εμφανίζονται αφού επιλεγεί το σήμα της μεγέθυνσης (+). Ακολούθως, το αποτέλεσμα 3 του ερωτήματος αντιστοιχεί σε μια θεματική κεφαλίδα που περιέχει τη 'Forecasting' ως στενότερη θεματική (skos:narrower). Στο γραφικό περιβάλλον η ετικέτα της κεφαλίδας αυτής εμφανίζεται αφού επιλεγεί το σήμα της σμίκρυνσης (-). Τέλος, τα αποτελέσματα 4, 5 και 6 του ερωτήματος αντιστοιχούν στους καθιερωμένους και μη όρους στα ελληνικά και αγγλικά (μπλε χρώμα για τα αγγλικά, καφέ χρώμα για τα ελληνικά).

5. Συμπεράσματα

Στην παρούσα εργασία, παρουσιάστηκε μια υπηρεσία θεματικής πλοήγησης βασισμένη στα ανοιχτά συνδεδεμένα δεδομένα, η οποία είναι ικανή να ενσωματώνει πηγές από διαφορετικά αποθετήρια (ψηφιακή βιβλιοθήκη του Πανεπιστημίου Πειραιά και συλλογή άρθρων των NYT). Η ενσωμάτωση αυτή έγινε μέσω της ευθυγράμμισης των θεματικών επικεφαλίδων που υπάρχουν στα δύο αποθετήρια. Η ευθυγράμμιση βασίστηκε σε έναν αλγόριθμο αντιστοίχησης θεματικών επικεφαλίδων. Η προτεινόμενη υπηρεσία με λίγες τροποποιήσεις μπορεί να εφαρμοστεί σε οποιαδήποτε υπηρεσία που βασίζεται σε ανοιχτά συνδεμένα δεδομένα.

Βιβλιογραφία

1. Τσάφου, Σ., & Χατζημαρή, Σ. Θησαυροί και θεματική ευρετηρίαση στις Ελληνικές βιβλιοθήκες, 2001. Στο: 10ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών, Θεσσαλονίκη, 228-242 (2001)
2. Stone, A. T.: The LCSH Century: A Brief History of the Library of Congress Subject Headings, and Introduction to the Centennial Essays, Cataloging And Classification Quarterly, VOL 29; PART 1/2, 1-16, Haworth Press Inc (2000) ISSN: 0163-9374
3. Summers, E., Isaac, A., Redding, C., Krech, D.: LCSH, SKOS and linked data. Στο: Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications (DCMI '08). Dublin Core Metadata Initiative 25-33 (2008)
4. Paepcke, A., Chang, C. K., Winograd, T., Garcia-Molina, H.: Interoperability for digital libraries worldwide. Commun. ACM 41(4), 33-42 (1998) DOI=10.1145/273035.273044
5. Miles, A., Perez-Aguera, J. R.: SKOS: Simple knowledge organisation for the web. Cataloging & Classification Quarterly, 43(3-4), 69–84 (2007) doi: 10.1300 /J104v43n03_04
6. Berners-Lee, T.: Linked Data - Design Issues. (2006) Διαθέσιμο στο: http://www.w3.org/DesignIssues/LinkedData.html Ημ. πρόσβασης: 28/11/2011
7. Papadakis, I., Stefanidakis, M., Kyprianos, K., Mavropodi,R.: Subject-based Information Retrieval within Digital Libraries Employing LCSHs. Dlib Magazine (www.dlib.org), 15(9/10), doi:10.1045/september2009-papadakis, (2009) Διαθέσιμο στο: http://www.dlib.org/dlib/september09/papadakis/09papadakis.html
8. Halpin, H., Hayes, P. J., McCusker, J. P., McGuinness, D. L., Thompson, H. S.: When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. Στο: International Semantic Web Conference (2010)
9. Isaac, A. Summers, E.: SKOS Simple Knowledge Organization System Primer. W3C Group Note, (2009) Διαθέσιμο στο: http://www.w3.org/TR/skos-primer/
10. Papadakis, I., Kyprianos, K.: Merging Controlled Vocabularies for More Efficient Subject-based Search. International Journal of Knowledge Management -IJKM, IGI Global 7(3), 74-90 (2011)
11. Bizer, Christian; Heath, Tom; Berners-Lee, Tim. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5(3), 1-22 (2009)