ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ ΓΙΑ ΤΟ ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2014-2015

Επιβλέπων καθηγητής: Νικόλαος Μήτρου

Εργασία 1: Εξαγωγή περιεχομένων σχεσιακών βάσεων δεδομένων ως ανοιχτά διασυνδεδεμένα δεδομένα

Η δημοσιοποίηση δεδομένων ως «Ανοιχτά Διασυνδεδεμένα Δεδομένα» (ΑΔΔ) [1] είναι μια από τις πλέον δημοφιλείς πρακτικές διάθεσης δεδομένων στο κοινό. Ωστόσο, η πλειονότητα των συστημάτων που διαχειρίζονται πληροφορία εξακολουθεί να ακολουθεί πιο συμβατικές και ώριμες τεχνολογίες όπως οι σχεσιακές βάσεις δεδομένων.

Στα πλαίσια της εργασίας αυτής θα πραγματοποιηθεί καταρχάς μελέτη των δυνατοτήτων διαφόρων μεθόδων και τεχνολογιών για την αντιστοίχιση δεδομένων από σχεσιακές βάσεις δεδομένων σε οντολογίες με στόχο τη διάθεση των δεδομένων ως ΑΔΔ. Στις εν λόγω τεχνολογίες περιλαμβάνονται τα πρότυπα RDF [2], R2RML [3], και SPARQL [4].

Στη συνέχεια, θα πραγματοποιηθεί αντιστοίχιση και μετατροπή των περιεχομένων από σχεσιακή βάση δεδομένων σε γράφο RDF, και φιλοξενία του αποτελέσματος σε μορφή ΑΔΔ.

Για το σκοπό αυτό θα χρησιμοποιηθεί και θα επεκταθεί το υπάρχον λογισμικό αντιστοίχισης R2RML Parser [5], με στόχο (α) την πληρέστερη υποστήριξη του προτύπου R2RML και (β) φιλοξενία του αποτελέσματος σε RDF server, όπως ο Fuseki [6] του Jena Framework [7] ώστε να είναι δυνατή η υποβολή ερωτημάτων SPARQL σε αυτό.

Αναφορές:

[1] http://en.wikipedia.org/wiki/Linked_data
[2] http://www.w3.org/RDF/
[3] http://www.w3.org/TR/r2rml/
[4] http://www.w3.org/TR/sparql11-overview/
[5] https://github.com/nkons/r2rml-parser
[6] http://jena.apache.org/documentation/serving_data/
[7] http://jena.apache.org/

Σχετικές τεχνολογίες: Java, Maven, Git, Mysql/Postgres, RDF, Jena

Άτομα: 1

Επικοινωνία: Νίκος Κωνσταντίνου (nkons@cn.ntua.gr)

Εργασία 2: Αξιοποίηση ανοικτών δεδομένων

Τα τελευταία χρόνια, έχει παρατηρηθεί μια παγκόσμια στροφή προς τα ανοικτά δεδομένα, μετά από πρωτοβουλίες εθνικών κυβερνήσεων, όπως η αμερικανική [1], η βρετανική [2], αλλά και η ελληνική [3], οι οποίες διαθέτουν ελεύθερα τα δεδομένα τους για χρήση από τρίτους, ενώ και άλλοι οργανισμοί και επιχειρήσεις «ανοίγουν» τα δεδομένα τους στο πλαίσιο αυτής της νέας τάσης. Ο συνδυασμός ανοικτών συνόλων δεδομένων από πλήθος διαφορετικών πηγών επιτρέπει την εξαγωγή συμπερασμάτων που πριν δεν ήταν εφικτά, καθώς και την ανάπτυξη καινοτόμων εφαρμογών που ενημερώνουν καλύτερα τον πολίτη και βελτιώνουν την ποιότητα ζωής του (παραδείγματα [4]-[7]).

Στο πλαίσιο της εργασίας αυτής, θα διερευνηθεί η δυνατότητα συνδυασμού και διασύνδεσης ανοικτών συνόλων δεδομένων από τον ελληνικό και διεθνή χώρο σε ένα συγκεκριμένο τομέα ενδιαφέροντος επιλογής του φοιτητή. Τα επιλεγμένα σύνολα δεδομένων θα εκφραστούν σύμφωνα με το μοντέλο περιγραφής RDF [8] (στην περίπτωση που δεν είναι ήδη εκφρασμένα σε αυτό) και θα χρησιμοποιηθούν υπάρχοντα εργαλεία για την εύρεση διασυνδέσεων μεταξύ τους [9][10].

Ενδεικτικά παραδείγματα τομέων ενδιαφέροντος είναι ο χώρος των βιβλιοθηκών, όπου μπορούν να αξιοποιηθούν σύνολα δεδομένων όπως τα [11]-[13] και ο χώρος της μουσικής, όπου μπορεί να εξεταστεί η διασύνδεση μουσικών μεταδεδομένων με δεδομένα κοινωνικών δικτύων [14]-[16].

Στόχος της διπλωματικής είναι η ανάπτυξη μιας εφαρμογής αποκλειστικά βασισμένης σε ανοικτά δεδομένα που θα αναδείξει τη χρησιμότητά τους στη νέα ψηφιακή εποχή.

Αναφορές:

[1] http://data.gov
[2] http://data.gov.uk
[3] http://data.gov.gr/
[4] http://wheredoesmymoneygo.org/
[5] http://crashmap.okfn.gr/
[6] http://www.donteat.at/
[7] http://traintimes.org.uk/map/london-buses
[8] http://www.w3.org/TR/rdf11-primer/
[9] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/
[10] http://aksw.org/Projects/LIMES.html
[11] https://openlibrary.org/
[12] http://www.oclc.org/data/data-sets-services.en.html
[13] http://viaf.org
[14] http://linkedbrainz.org/
[15] http://dbtune.org/
[16] http://www.last.fm/

Σχετικές τεχνολογίες: Java, RDF, SPARQL

Άτομα: 1

Επικοινωνία: Δημήτρης Σπανός (dspanos@cn.ntua.gr)

Εργασία 3: Δυναμική πρόσβαση σε σχεσιακές ΒΔ μέσω SPARQL

Παραδοσιακά, οι χρήστες και οι εφαρμογές που επικοινωνούν με μια σχεσιακή βάση δεδομένων, χρειάζεται να θέσουν κατάλληλα SQL ερωτήματα έχοντας γνώση του σχήματος της ΒΔ. Ένας εναλλακτικός τρόπος πρόσβασης στα δεδομένα μιας σχεσιακής ΒΔ, ο οποίος δε χρειάζεται γνώση του σχεσιακού σχήματος, χρησιμοποιεί τη γλώσσα SPARQL [1], η οποία αποτελεί μια γλώσσα ερωτημάτων για RDF γράφους [2]. Η έκφραση των σχεσιακών δεδομένων μιας ΒΔ ως RDF γράφο μέσω της γλώσσας ορισμού αντιστοιχιών R2RML [3] επιτρέπει όχι μόνο την πρόσβαση σε αυτά μέσω SPARQL, αλλά και την ευκολότερη ολοκλήρωση με άλλες ΒΔ, καθώς και το συνδυασμό με άλλα δεδομένα διαθέσιμα σε μορφή Συνδεδεμένων Δεδομένων [4].

Στο πλαίσιο αυτής της διπλωματικής εργασίας, θα μελετηθούν αλγόριθμοι (π.χ. [5][6]) επανεγγραφής SPARQL ερωτημάτων σε ισοδύναμα SQL, δεδομένης μιας R2RML αντιστοιχίας και θα επεκταθεί η υλοποίηση σχετικού υπό ανάπτυξη εργαλείου της ερευνητικής ομάδας Επικοινωνιών Πολυμέσων και Τεχνολογιών Παγκόσμιου Ιστού. Παράλληλα, θα πραγματοποιηθούν συγκριτικές μετρήσεις απόδοσης με σχετικά εργαλεία (π.χ. [7][8]).

Αναφορές:

[1] http://www.w3.org/TR/sparql11-query/
[2] http://www.w3.org/TR/rdf11-primer/
[3] http://www.w3.org/TR/r2rml/
[4] http://linkeddata.org/
[5] Chebotko, A., Lu, S., and Fotouhi, F. 2009. Semantics Preserving SPARQL-to-SQL Translation. In *Data & Knowledge Engineering*, 68, 10, 973–1000
[6] Rodriguez-Muro, M., Kontchakov, R., Zakharyaschev, M. 2013. Ontology-Based Data Access: Ontop of Databases. In *Proceedings of the 12th International Semantic Web Conference (ISWC'13)*, Sydney, Australia
[7] http://d2rq.org/
[8] http://ontop.inf.unibz.it/

Σχετικές τεχνολογίες: Java, SQL, RDF, SPARQL

Άτομα: 1

Επικοινωνία: Δημήτρης Σπανός (dspanos@cn.ntua.gr)

Εργασία 4: Δημιουργία Εφαρμογής επισημείωσης κειμένων με χρήση ελεγχόμενων λεξιλογίων

Τα ελεγχόμενα λεξιλόγια (controlled vocabularies) αποτελούν μια προσεκτικά επιλεγμένη αλληλουχία από λέξεις και φράσεις οι οποίες χρησιμοποιούνται τόσο στον τομέα των Ψηφιακών Βιβλιοθηκών όσο και γενικότερα στο πεδίο της Επιστήμης των Πληροφορίων για την επισημείωση τμημάτων πληροφορίας που μπορεί να αποτελούν κομμάτι μιας ιστοσελίδας ή ενός εγγράφου, έτσι ώστε τα παραπάνω να μπορούν να ανακτηθούν με έναν εύκολο τρόπο στα πλαίσια μιας αναζήτησης [1], [2]. Τα ελεγχόμενα λεξιλόγια χρησιμοποιούνται ευρέως στους παραπάνω τομείς καθώς η χρήση τους έχει οδηγήσει στον περιορισμό του προβλήματος της αμφισημίας και της πολυσημίας, όπου η ίδια έννοια μπορεί να αποδίδεται με περισσότερους από έναν τρόπους, ενισχύοντας έτσι την συνέπεια του περιεχομένου ενός εγγράφου.

Από την άλλη πλευρά οι οντολογίες [3] μπορεί να θεωρηθούν ώς ένας από τους πυλώνες του Σημασιολογικού Ιστού, καθώς μπορεί να θεωρηθούν ως ένας ειδικός τύπος λεξιλογίου ή μερικές φορές ακόμα και ως μια συλλογή από URIs [4] τα οποία μπορεί να οδηγούν σε κάποια περιγραφή. Οι οντολογίες συνήθως ακολουθούνται από κάποιο έγγραφο γραμμένο σε κάποια γλώσσα περιγραφής οντολογιών όπως οι RDF [5] και OWL [6].

Στην παρούσα εργασία θα υλοποιηθεί εφαρμογή επισημείωσης κειμένων σε μορφή doc, pdf ή κάποιο άλλο μορφότυπο με τη χρήση ελεγχόμενων λεξιλογίων. Τα ελεγχόμενα λεξιλόγια θα χρησιμοποιηθούν για την επισημείωση των κειμένων με θεματικές επικεφαλίδες (subject headings). Αρχικά, θα πρέπει να γίνει μια ανασκόπηση των υπαρχόντων πρακτικών και να καθοριστούν τα κριτήρια με τα οποία θα επιλεγούν τα κατάλληλα λεξιλόγια ή οι οντολογίες που θα χρησιμοποιηθούν ανάλογα και με υπάρχοντα δεδομένα. Στη συνέχεια, θα πραγματοποιηθεί αντιστοίχηση και μετατροπή των μεταδεδομένων των εγγράφων σε Γράφο RDF, και φιλοξενία του αποτελέσματος σε triple store π.χ.Virtuoso [7], CouchDB [8], για πρόσβαση και διαχείριση των δεδομένων.

Αναφορές:

[1] http://boxesandarrows.com/what-is-a-controlled-vocabulary/
[2] http://en.wikipedia.org/wiki/Controlled_vocabulary
[3] http://semanticweb.org/wiki/Ontology
[4] http://www.w3.org/Addressing/
[5] http://www.w3.org/RDF/
[6] http://www.w3.org/2001/sw/wiki/OWL
[7] http://virtuoso.openlinksw.com/
[8] http://couchdb.apache.org/

Σχετικές τεχνολογίες: Java, SparQL, Virtuoso Universal Server, RDF, Jena, CouchDB

Άτομα: 1

Επικοινωνία: Ελένη Γιαννοπούλου (egiann@cn.ntua.gr)

Εργασία 5: Αυτόματη εξαγωγή φράσεων κλειδιών (keyphrase extraction) από κείμενα

Ένας από τους πιο γνωστούς αλγορίθμους εξαγωγής φράσεων κλειδιών από κείμενα είναι ο αλγόριθμος KEA [1]. Ο συγκεκριμένος αλγόριθμος μπορεί να χρησιμοποιηθεί είτε για την ευρετηρίαση μέσω φράσεων που επιλέγονται μέσα από το ίδιο το κείμενο ή μέσω ενός ελεγχόμενου λεξιλογίου [2]. Ο αλγόριθμος KEA έχει ενσωματωθεί στο MAUI indexer [3] το οποίο είναι ένα εργαλείο ανοικτού κώδικα υλοποιημένο σε JAVA.

Τα ελεγχόμενα λεξιλόγια (controlled vocabularies) αποτελούν μια προσεκτικά επιλεγμένη αλληλουχία από λέξεις και φράσεις οι οποίες χρησιμοποιούνται για την επισημείωση τμημάτων πληροφορίας, έτσι ώστε τα παραπάνω να μπορούν να ανακτηθούν με έναν εύκολο τρόπο στα πλαίσια μιας αναζήτησης [2]. Τα ελεγχόμενα λεξιλόγια χρησιμοποιούνται ευρέως καθώς η χρήση τους έχει οδηγήσει στον περιορισμό του προβλήματος της αμφισημίας και της πολυσημίας, όπου η ίδια έννοια μπορεί να αποδίδεται με περισσότερους από έναν τρόπους, ενισχύοντας έτσι την συνέπεια του περιεχομένου ενός εγγράφου. Οι οντολογίες [8] μπορεί να θεωρηθούν ώς ένας από τους πυλώνες του Σημασιολογικού Ιστού, καθώς μπορεί να θεωρηθούν ως ένας ειδικός τύπος λεξιλογίου ή μερικές φορές ακόμα και ως μια συλλογή από URIs [4] τα οποία μπορεί να οδηγούν σε κάποια περιγραφή.

Στην παρούσα εργασία θα υλοποιηθεί εφαρμογή εξαγωγής φράσεων από κείμενα σε JAVA με χρήση της βιβλιοθήκης JENA [7]. Αρχικά, θα πρέπει να γίνει μια ανασκόπηση των υπαρχόντων πρακτικών και να καθοριστούν τα κριτήρια με τα οποία θα επιλεγούν τα κατάλληλα λεξιλόγια ή οι οντολογίες που θα χρησιμοποιηθούν ανάλογα και με υπάρχοντα δεδομένα. Στη συνέχεια, θα πραγματοποιηθεί αντιστοίχηση και μετατροπή των μεταδεδομένων των εγγράφων σε Γράφο RDF, και φιλοξενία του αποτελέσματος σε triple store π.χ.Virtuoso [6], για πρόσβαση και διαχείριση των δεδομένων μέσω των προτύπων RDF [5], SQARQL [4]. Παρέχεται η δυνατότητα να χρησιμοποιηθεί το εργαλείο ανοικτού λογισμικού Maui indexer [3] ή κάποιο άλλο σύστημα διαχείρισης δεδομένων όπως CouchDB [9].

Αναφορές:

[1] https://code.google.com/p/kea-algorithm/
[2] http://en.wikipedia.org/wiki/Controlled_vocabulary
[3] https://code.google.com/p/maui-indexer/
[4] http://www.w3.org/TR/sparql11-overview/
[5] http://www.w3.org/RDF/
[6] http://virtuoso.openlinksw.com/
[7] http://jena.apache.org/
[8] http://semanticweb.org/wiki/Ontology
[9] http://couchdb.apache.org/

Σχετικές τεχνολογίες: Java, SparQL, Virtuoso Universal Server, RDF, Jena, CouchDB

Άτομα: 1

Επικοινωνία: Ελένη Γιαννοπούλου (egiann@cn.ntua.gr)

Εργασία 6: Δημιουργία Εφαρμογής διαχείρισης δεδομένων Ψηφιακού Αποθετηρίου

Η έννοια του εμπλουτισμού των μεταδεδομένων [1] έχει να κάνει με τη μετατροπή ή την ενίσχυση των συμβολοσειρών των μεταδεδομένων με τη χρήση ελεγχόμενων λεξιλογίων [2]. Τα ελεγχόμενα λεξιλόγια μπορεί να χρησιμοποιηθούν ως URI για συγκεκριμένα πεδία εγγραφών ενός Ψηφιακού Αποθετηρίου. Τα ελεγχόμενα λεξιλόγια (controlled vocabularies) αποτελούν μια προσεκτικά επιλεγμένη αλληλουχία από λέξεις και φράσεις οι οποίες χρησιμοποιούνται για την επισημείωση τμημάτων πληροφορίας, έτσι ώστε να μπορούν να ανακτηθούν με έναν εύκολο τρόπο στα πλαίσια μιας αναζήτησης. Οι οντολογίες [3] μπορεί να θεωρηθούν ώς ένας από τους πυλώνες του Σημασιολογικού Ιστού, καθώς μπορεί να θεωρηθούν ως ένας ειδικός τύπος λεξιλογίου ή μερικές φορές ακόμα και ως μια συλλογή από URIs [4] τα οποία μπορεί να οδηγούν σε κάποια περιγραφή.

Ένα Ψηφιακό Αποθετήριο αποτελεί ένα σύστημα που χρησιμεύει στην ηλεκτρονική απόθεση, διαχείριση και ανάδειξη ψηφιακού περιεχομένου που χρήζουν μακροχρόνιας διατήρησης [7]. Μπορεί να προσφέρει υπηρεσίες αναζήτησης, πλοήγησης, πρόσβασης στο περιεχόμενο μέσω μονίμων προσδιοριστών, κατάθεσης και διαχείρισης περιεχομένου καθώς και ασφαλούς διαφύλαξης και διατήρησης του ψηφιακού υλικού.

Στην παρούσα εργασία θα υλοποιηθεί εφαρμογή διαχείρισης δεδομένων Ψηφιακού Αποθετηρίου. Στόχος είναι ο εμπλουτισμός των μεταδεδομένων των εγγραφών του αποθετηρίου με αυτόματο τρόπο χρησιμοποιώντας ελεγχόμενα λεξιλόγια ή θησαυρούς αναζητώντας τρόπους για τον εμπλουτισμό των μεταδεδομένων με έμφαση στην παροχή πιο ποιοτικών πληροφορίων, παρέχοντας ταυτόχρονα και άλλους μηχανισμούς πρόσβασης στα δεδομένα (αυξάνοντας τη διακριτότητα και τον τρόπο εμφάνισης των δεδομένων) [6]. Η εφαρμογή μπορεί να υλοποιεί είτε ως πρόσθετο σε κάποιο σύστημα διαχείρισης των εγγραφών ενός Ψηφιακού Αποθετηρίου όπως DSPACE [5], είτε ως αυτόνομη (standalone) εφαρμογή. Ακόμα, θα πρέπει να υλοποιηθεί η δυνατότητα σημασιολογικής αναζήτησης επί των εμπλουτισμένων δεδομένων μέσω κατάλληλης διεπαφής. Η εφαρμογή θα πρέπει να επικοινωνεί με triple store π.χ. μέσω Virtuoso [8], για πρόσβαση και διαχείριση των δεδομένων.

Αναφορές:

[1]http://www.iks-project.eu/resources/automatic-metadata-enrichment-liferay-stanbol
[2] http://en.wikipedia.org/wiki/Controlled_vocabulary
[3] http://semanticweb.org/wiki/Ontology
[4] http://www.w3.org/Addressing/
[5] http://www.dspace.org/
[6] http://www.ala.org/alcts/mgrps/taskforces/metadataenrichment
[7] http://www.epset.gr/el/content/psifiako-apothetirio
[8] http://virtuoso.openlinksw.com/

Σχετικές τεχνολογίες: Java, SparQL, Virtuoso Universal Server, RDF, Jena

Άτομα: 1

Επικοινωνία: Ελένη Γιαννοπούλου (egiann@cn.ntua.gr)