ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΓΝΩΣΗΣ

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2020 **ΘΕΜΑ**

1 Περιγραφή

Σε αυτό το θέμα καλείστε να κατασκευάσετε μια σημασιολογική βάση γνώσης που θα περιέχει δεδομένα που σχετίζονται με μετακινήσεις με μέσα μεταφοράς. Συγκεκριμένα, θα πρέπει να κατασκευάσετε μια οντολογία για τα μέσα μεταφοράς η οποία να μοντελοποιεί επαρκώς το συγκεκριμένο πεδίο, να συγκεντρώσετε κατάλληλα δεδομένα, να τα αναπαραστήσετε ως στιγμιότυπα των εννοιών, ρόλων και ιδιοτήτων τύπων δεδομένων της οντολογίας ώστε να φτιάξετε μια ολοκληρωμένη βάση γνώσης, και τέλος να εισαγάγετε τα δεδομένα σε μια αποθήκη τριάδων.

Η οντολογία που θα αναπτύξετε θα είναι σε μορφή OWL2, και θα πρέπει να είναι σχεδιασμένη έτσι ώστε το λεξιλόγιο και οι εκφραστικές δυνατότητές της να καλύπτουν τις ανάγκες περιγραφής διαφόρων μέσων μεταφοράς (αστικά μέσα μεταφοράς, υπεραστικά λεωφορεία, τρένα, αεροπλάνα, πλοία) καθώς και των σχετιζόμενων με αυτά ζητημάτων (π.χ. διαδρομές, πίνακες δρομολογίων, εκτέλεση δρομολογίων, συγκοινωνιακές εταιρείες, οχήματα). Η οντολογία θα είναι προσανατολισμένη στην περιγραφή των μέσων, των διαδρομών και των δρομολογίων τους, και δεν προορίζεται για σύστημα κρατήσεων ή έκδοσης εισιτηρίων. Με βάση το λεξιλόγιο της οντολογίας θα πρέπει να μπορεί να μοντελοποιηθεί και στατική πληροφορία, αλλά και χρονική πληροφορία (π.χ. χρόνοι εκτέλεσης δρομολογίων). Για την διευκόλυνση της ανάπτυξης της οντολογίας προτείνεται η χρήση του γραφικού εργαλείου ανάπτυξης και επεξεργασίας οντολογιών Protege (χρήσιμες οδηγίες υπάρχουν εδώ).

Αφού αναπτύξετε την οντολογία θα πρέπει να την εμπλουτίσετε με δεδομένα (ABox). Λόγω του μεγάλου εύρους του πεδίου, προτείνεται να ασχοληθείτε με δεδομένα που αφορούν μια συγκεκριμένη κατηγορία μέσων μεταφοράς και κάποια ή κάποιες συγκεκριμένες περιοχές (π.χ. πόλεις), χωρίς αυτό να σημαίνει ότι ο σχεδιασμός της οντολογίας θα πρέπει είναι προσανατολισμένος στις συγκεκριμένες ανάγκες ενός μέσου ή περιοχής.

Στο πλαίσιο αυτό, ως ελάχιστο θα πρέπει να κατασκευάσετε δεδομένα για τα αστικά μέσα μεταφοράς και την πόλη της Αθήνας χρησιμοποιώντας ως πηγή δεδομένων τα Δρομολόγια Αστικών Συγκοινωνιών Αθήνας, τα οποία παρέχονται στο σχήμα δεδομένων του GTFS (τα δεδομένα δεν είναι επικαιροποιημένα, αλλά αυτό δεν έχει σημασία). Αφού μελετήσετε το σχήμα των δεδομένων, θα πρέπει να τα μετατρέψετε σε ένα σύνολο δεδομένων RDF με βάση το λεξιλόγιο της οντολογίας που έχετε σχεδιάσει, το οποίο και θα εισαγάγετε σε μία αποθήκη RDF. Η μετατροπή θα συνίσταται ουσιαστικά στη δημιουργία των κατάλληλων URI και των δηλώσεων RDF που αποδίδουν ιδιότητες ή σχετίζουν πόρους. Για να επιτύχετε τη μετατροπή θα χρειαστεί να γράψετε κώδικα, σε Java ή Python, για την κατάλληλη επεξεργασία των δεδομένων αρχείων. Ο κώδικας που θα γράψετε θα πρέπει να παράγει αρχεία δηλώσεων RDF είτε σε μορφή N-Triples/N-Quads είτε σε μορφή Turtle/Trig. Για τις ανάγκες παραγωγής των δηλώσεων RDF στην Java μπορείτε, προαιρετικά, να χρησιμοποιήσετε κάποια από τις βιβλιοθήκες Apache Jena ή RDF4J. Ως αποθήκη τριάδων RDF για την αποθήκευση των δεδομένων που θα παραγάγετε προτείνεται το OpenLink Virtuoso, καθώς μπορεί να υποστηρίξει χωρίς προβλήματα μεγάλους όγκους δεδομένων.

Επειδή τα δεδομένα περιέχουν γεωγραφική πληροφορία (π.χ. τοποθεσίες στάσεων) θα πρέπει να μεριμνήσετε για την καταλληλότερη αναπαράστασή τους ως δεδομένων RDF. Γενικά, για την αναπαράσταση γεωγραφικών τοποθεσιών συντεταγμένων υπάρχει το λεξιλόγιο WGS84 Geo Positioning και το πρότυπο GeoSPARQL. Αν χρησιμοποιήσετε το OpenLink Virtuoso προτείνεται να χρησιμοποιήσετε μια αναπαράσταση σαν αυτή που έχουν υιοθετήσει τα LinkedGeoData (βλ. εδώ για παραδείγματα) και κάνει χρήση του τύπου δεδομένων http://www.openlinksw.com/schemas/virtrdf#Geometry για την αναπαράσταση χωρικών σχημάτων. Πληροφορίες για την αναπαράσταση και τις πράξεις που μπορούν να εφαρμοστούν στον χωρικό τύπο δεδομένων δίνονται εδώ και εδώ. Σαν παράδειγμα, ένα γεωγραφικό σημείο με γεωγραφικό πλάτος 51.3466 και γεωγραφικό μήκος 12.3831 μπορεί να αναπαρασταθεί ως το λεκτικό "POINT(12.3830858 51.3465518)"^^<http://www.openlinksw.com/schemas/virtrdf#Geometry>.

Αφού ολοκληρώσετε την κατασκευή της βάσης γνώσης θα πρέπει να επιδείξετε τις δυνατότητές της κατασκευάζοντας και λαμβάνοντας απαντήσεις σε διάφορα χρήσιμα ερωτήματα. Ενδεικτικά ερωτήματα θα μπορούσαν να είναι για παράδειγμα τα εξής: Ποια μέσα μεταφοράς διέρχονται πλησιέστερα από ένα δεδομένο σημείο; Μπορεί να μεταβεί κάποιος από μία στάση σε κάποια άλλη χωρίς αλλαγή μέσου, ή με μία αλλαγή, δύο αλλαγές; Τα ερωτήματα θα πρέπει να διατυπωθούν σε SPARQL και να εκτελεστούν επί της αποθήκης τριάδων που έχετε κατασκευάσει. Επειδή όμως τα δεδομένα που θα έχετε εισαγάγει στην αποθήκη τριάδων καλύπτουν μόνο ένα υποσύνολο του πεδίου ενδιαφέροντος που περιγράφει η οντολογία, για λόγους επίδειξης θα πρέπει να διατυπώσετε και επιπλέον ενδιαφέροντα ερωτήματα που χρησιμοποιούν τις εκφραστικές δυνατότητες της μοντελοποίησης σας, δεν έχουν όμως απαντήσεις λόγω του ελλιπούς ABox.

2 Αξιολόγηση

Το θέμα είναι ατομικό. Για την αξιολόγησή του θα παραδώσετε α) μια αναφορά όπου θα περιγράφετε θέματα που αφορούν τον σχεδιασμό και την υλοποίηση της βάσης γνώσης, β) τα αρχεία της οντολογίας και τα αρχεία των δεδομένων RDF που κατασκευάσατε γ) τον κώδικα που αναπτύξατε για τις ανάγκες της υλοποίησης. Στην αναφορά θα πρέπει:

1. Να παρουσιάσετε την οντολογία που αναπτύξατε, σχολιάζοντας τις σχεδιαστικές αποφάσεις που λάβατε και τις πιθανές παραδοχές που κάνατε.
2. Να περιγράψετε τη διαδικασία μετατροπής των δεδομένων από τις πρωτογενείς πηγές σε τριάδες RDF.
3. Να παρουσιάστε ενδεικτικά αποσπάσματα των μετασχηματισμένων δεδομένων (δηλαδή μικρά τμήματα του συνόλου δεδομένων RDF είτε σε N-Triples/N-Quads είτε σε Turtle/Trig), στα οποία να αποτυπώνονται οι επιλογές που κάνατε για το σύνολο των ειδών δεδομένων που επεξεργαστήκατε.
4. Να παρουσιάστε και να περιγράψετε τα ερωτήματα SPARQL που κατασκευάσατε μαζί με αποσπάσματα των αντίστοιχων πινάκων αποτελεσμάτων τους.
5. Να σχολιάσετε τυχόν προβλήματα που αντιμετωπίσατε και περιορισμούς του συστήματος που αναπτύξατε.

Η βαθμολόγηση του θέματος θα γίνει ως εξής:

1. (40%) Κατασκευή οντολογίας μέσων μεταφοράς.
2. (40%) Μετασχηματισμός δεδομένων των Αστικών Συγκοινωνιών Αθήνας σε RDF και εισαγωγή τους στην αποθήκη τριάδων.
3. (20%) Κατασκευή και εκτέλεση ερωτημάτων SPARQL.

Παρατηρήσεις για το OpenLink Virtuoso

1. Αναλυτικές πληροφορίες για την εγκατάσταση και τη χρήση του OpenLink Virtuoso υπάρχουν στη διεύθυνση http://vos.openlinksw.com/owiki/wiki/VOS.
2. Μετά την εγκατάσταση μπορείτε να εκκινήσετε το OpenLink Virtuoso πηγαίνοντας στον κατάλογο bin της εγκατάστασης και εκτελώντας την εντολή virtuoso-t +foreground +configfile ../database/virtuoso.ini. Με τον τρόπο αυτό θα βλέπετε απευθείας στην γραμμή εντολών μηνύματα σχετικά με τη λειτουργία και τυχόν σφάλματα. Αν η εγκατάσταση έχει δημιουργήσει κάποιο service του λειτουργικού συστήματος που εκτελείται αυτόματα θα πρέπει πρώτα να το απενεργοποιήσετε αν θέλετε να ελέγχετε χειροκίνητα την εκκίνηση και τον τερματισμό. (Στα Windows αυτό γίνεται από το Control Panel (Πίνακας Ελέγχου)/Administrative Tools (Εργαλεία Διαχείρισης)/Services (Υπηρεσίες): επιλέγετε την υπηρεσία OpenLink Virtuoso Server [vos], διακόπτετε τη λειτουργία [με την επιλογή που δίνεται επάνω αριστερά], και με δεξί κλικ/Properties (Ιδιότητες)/General (Γενικά) στο μενού Startup type επιλέγετε Manual (Μη αυτόματα)). Τερματίζετε το OpenLink Virtuoso πατώντας Ctrl+C στο παράθυρο από τον οποίο το εκκινήσατε.
3. Το OpenLink Virtuoso είναι μεταξύ άλλων μια αποθήκη τριάδων RDF και περιλαμβάνει δύο server, έναν database server και έναν web server. Ο πρώτος επιτρέπει την απευθείας διαχείριση της βάσης δεδομένων μέσω μιας γραμμής εντολών και ο δεύτερος παρέχει μια γραφική διεπαφή διαχείρισης και υποβολής ερωτημάτων SPARQL.
4. Ο web server ακούει (με βάση τις προκαθορισμένες ρυθμίσεις) στη διεύθυνση http://localhost:8890/. Η διεπαφή υποβολής ερωτημάτων SPARQL είναι η http://localhost:8890/sparql στην οποία μπορείτε να εκτελέσετε απευθείας οποιοδήποτε ερώτημα SPARQL. Η διαχειριστικές δυνατότητες προσφέρονται αφού κάνετε login πατώντας τον σύνδεσμο conductor επάνω αριστερά στη σελίδα http://localhost:8890/. Το προκαθορισμένο username και password ειναι dba και dba, αντίστοιχα. Μόλις κάνετε login εμφανίζονται οι επιλογές διαχείρισης. Κυρίως ενδιαφέρει το tab Linked Data. Εκεί, στο tab SPARQL μπορείτε να υποβάλλεται ερωτήματα SPARQL, στο tab Graphs/Graphs μπορείτε να δείτε (και να διαγράψετε) τους ονοματισμένους γράφους που περιέχει η βάση τριάδων, και από το Quad Store Upload μπορείτε να ανεβάσετε αρχεία με δεδομένα RDF (n-triples, turtle, rdf/xml) προς εισαγωγή στη βάση τριάδων.
5. Όλες τις λειτουργίες που μπορείτε να κάνετε μέσω του γραφικού περιβάλλοντος που προσφέρει ο web server μπορείτε να τις κάνετε και μέσω του database server μέσω γραμμής εντολών. Εισέρχεστε στο περιβάλλον του database server εκτελώντας την εντολή isql από τη γραμμή εντολών στον κατάλογο bin της εγκατάστασης. Στη γραμμή εντολών του database server, για αν εκτελέστε κάποιο ερώτημα ή άλλη εντολή SPARQL πρέπει να γράψετε πρώτα SPARQL και να συνεχίστε με το ερώτημα. Η γραμμή εντολών isql είναι διαθέσιμη και μέσω του περιβάλλοντος διαχείρισης του web server μέσω του συνδέσμου Interactive SQL (ISQL).
6. Στο virtuoso.ini ορίζονται διάφορες ρυθμίσεις για τη λειτουργία του OpenLink Virtuoso. Ενδιαφέρουν ιδιαίτερα:

(α΄) Το πεδίο Parameters/DirsAllowed που έχει ως τιμές (χωρισμένες μεταξύ τους με κόμματα) τους καταλόγους από τους οποίους επιτρέπεται να διαβαστούν αρχεία δεδομένων προκειμένου να εισαχθούν στη βάση γνώσης.

(β΄) Το πεδίο Paremeters/ServerPort που έχει ως τιμή την πόρτα στον οποία θα ακούει ο database server.

(γ΄) Το πεδίο HTTPServer/ServerPort που έχει ως τιμή την πόρτα στον οποία θα ακούει ο web server και το sparql endpoint.

(δ΄) Το πεδίο SPARQL/ResultSetMaxRows που έχει ως τιμή το μέγιστο πλήθος αποτελεσμάτων που επιτρέπεται να επιστρέψει ένα ερώτημα SPARQL.

7. Σε περίπτωση κακού τερματισμού του server και αδυναμίας επανεκκίνησής του, δοκιμάστε να σβήσετε το αρχείο virtuoso.lck στον κατάλογο database της εγκατάστασης.

8. Σε περίπτωση που θέλετε να αρχικοποιήσετε τη βάση γνώσης του virtuoso (να διαγράψετε δηλαδή ότι έχετε εισαγάγει) μπορείτε, αφού τερματίσετε τον server, να διαγράψετε όλα τα αρχεία από το κατάλογο database εκτός από το virtuoso.ini. Την επόμενη φορά που θα επανεκκινήσετε τον server θα δημιουργηθεί μια κενή βάση.

9. Για την εισαγωγή και διαχείριση μεγάλων όγκων δεδομένων καταλληλότερος είναι ο database server. Από τη γραμμή εντολών του database server μπορείτε να δίνετε εντολές εισαγωγής και διαγραφής δεδομένων, αλλά και να κάνετε ερωτήματα. Επειδή το OpenLink Virtuoso είναι υλοποιημένο πάνω σε σχεσιακή βάση, η διαχείριση των δεδομένων γίνεται με τη μεσολάβηση σχεσιακών πινάκων.

10. Για οδηγίες για τη μαζική εισαγωγή αρχείων δεδομένων RDF, συμβουλευτείτε τη σελίδα http://vos.openlinksw.com/owiki/wiki/VOS/VirtBulkRDFLoader που εξηγεί αναλυτικά τη διαδικασία.

11. Για ευκολότερη διαχείριση των δεδομένων σας προτείνεται να τα εντάξετε σε έναν ή περισσότερους ονοματισμένους γράφους, ώστε να μπορείτε π.χ. να τα διαγράψετε εύκολα μαζικά αλλά και επιλεκτικά αν χρειαστεί. Η διαγραφή των περιεχομένων ενός ονοματισμένου γράφου γίνεται μέσω της εντολής SPARQL CLEAR GRAPH <graph-name>; στη γραμμή εντολών του database server. Βλ. και http://vos.openlinksw.com/owiki/wiki/VOS/VirtTipsAndTricksGuideDeleteLargeGraphs. Η διαγραφή γίνεται εύκολα και μέσω του web server.

12. Αν θελήσετε να εκτελέσετε ομόσπονδα ερωτήματα SPARQL θα πρέπει να δώσετε προηγουμένως στη γραμμή εντολών του database server τις εξής δύο εντολές:
grant execute on ”DB.DBA.SPARQL_SINV_IMP” to ”SPARQL”; και
grant select on ”DB.DBA.SPARQL_SINV_2” to ”SPARQL”;