ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ Η/Υ

ΤΜΗΜΑ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΣΧΟΛΗΣ ΝΑΥΤΙΛΙΑΣ

&ΒΙΟΜΗΧΑΝΙΑΣ ΤΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΕΙΡΑΙΩΣ

ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΤΕΧΝΟΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ

ΔΙΕΠΙΣΤΗΜΟΝΙΚΟ – ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

Ολοκληρωμένη Διαχείριση Σημασιολογικών Δεδομένων στον Παγκόσμιο Ιστό

(INTEGRATED SEMANTIC DATA MANAGEMENT ON THE WORLD WIDE WEB)

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΑΙΝΕΣΗ Π. ΧΑΛΑΜΠΙΑ

ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ
ΑΔΑΜΟΠΟΥΛΟΥ ΕΥΓΕΝΙΑ

ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ Η/Υ

ΤΜΗΜΑ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΣΧΟΛΗΣ ΝΑΥΤΙΛΙΑΣ

&ΒΙΟΜΗΧΑΝΙΑΣ ΤΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΕΙΡΑΙΩΣ

ΔΙΕΠΙΣΤΗΜΟΝΙΚΟ – ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

Ολοκληρωμένη Διαχείριση Σημασιολογικών Δεδομένων στον Παγκόσμιο Ιστό

(INTEGRATED SEMANTIC DATA MANAGEMENT ON THE WORLD WIDE WEB)

Επιβλέπουσα: Αδαμοπούλου Ευγενία

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 1η Μαρτίου 2024

Αδαμοπούλου Ευγενία	Δεμέστιχας Κων/νος	Ευστάθιος Συκάς
Ε.ΔΙ.Π.	Επίκουρος καθηγητής Γ.Π.Α.	Ομότιμος Καθηγητής Ε.Μ.Π
Ε.Μ.Π.		

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αθήνα Μάρτιος 2024

Χαραλαμπία, Π. Παινέση

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Πανεπιστημίου Πατρών

Περίληψη

Η συγκεκριμένη μεταπτυχιακή διπλωματική εργασία ασχολείται με την διεργασία εξέλιξης του Παγκόσμιου Ιστού στο Σημασιολογικό Ιστό και τους τρόπους με τους οποίους μπορεί να βελτιωθεί η αναζήτηση του χρήστη στον Παγκόσμιο Ιστό.

Στα πρώτα βήματα του Παγκόσμιου Ιστού το σημαντικότερο ίσως πρόβλημα για τους χρήστες που ήθελαν να αναζητήσουν πληροφορίες σε αυτό ήταν η έλλειψη πολλών και χρήσιμων πηγών. Σταδιακά, αλλά με ιδιαίτερα γρήγορους ρυθμούς ο Παγκόσμιος Ιστός μετατράπηκε σε μία από τις μεγαλύτερες πηγές πληροφοριών που χρησιμοποιεί ο άνθρωπος καθώς όλο και περισσότεροι εισάγουν δεδομένα για κάθε είδους δραστηριότητα και θέμα. Το πρόβλημα των χρηστών λοιπόν που αναζητούν πληροφορίες ανάχθηκε στη γρήγορη εξαγωγή των χρήσιμων, από τον τεράστιο όγκο των παρεχόμενων, πληροφοριών.

Δεδομένης αυτής της ανάγκης των χρηστών δημιουργήθηκαν όροι και τεχνικές όπως Data Mining (Εξόρυξη Δεδομένων), Information Retrieval (Ανάκτηση Πληροφορίας), Knowledge Management (Διαχείριση Γνώσης), οι οποίες προσπαθούν να καλύψουν τις ανάγκες των χρηστών. Επιπλέον, στην προσπάθεια για καλύτερη ποιότητα των παρεχόμενων αποτελεσμάτων στο χρήστη σημαντικό ρόλο διαδραμάτισε η εκμετάλλευση των ιδιαίτερων στοιχείων που μπορούν να εξαχθούν για τα ενδιαφέροντά του, τόσο στο στάδιο της διαπέρασης, όπου συγκεντρώνονται σελίδες συγκεκριμένης θεματολογίας (topic-focused crawling), όσο και στο στάδιο της αναζήτησης μέσα από αυτές των πιο σημαντικών για τον εκάστοτε χρήστη (personalization). Παράλληλα, καθώς ο Παγκόσμιος Ιστός σταδιακά μετεξελίσσεται στο Σημασιολογικό Παγκόσμιο Ιστό (Semantic Web) νέα μοντέλα και πρότυπα (XML, RDF, OWL) αναπτύσσονται για την προώθηση αυτής της διαδικασίας. Η έκφραση, μετάδοση και αναζήτηση πληροφοριών με χρήση αυτών των προτύπων ανοίγει νέους ορίζοντες στη χρήση του Διαδικτύου. Το βασικό αντικείμενο της εργασίας αυτής είναι η αξιοποίηση των παρεχόμενων μοντέλων και προτύπων του Σημασιολογικού Ιστού σε συνδυασμό με ήδη εφαρμοσμένες ιδέες και αλγορίθμους στον απλό Παγκόσμιο Ιστό ώστε να είναι εφικτή η ταχύτερη και ακριβέστερη ανάκτηση και επεξεργασία πληροφοριών. Δόθηκε επίσης προσπάθεια στην αξιοποίηση τεχνικών που εκμεταλλεύονται τις ιδιαίτερες προτιμήσεις κάθε χρήστη, και στη διερεύνηση της χρήσης των νέων μοντέλων και προτύπων του Σημασιολογικού Ιστού για την προώθηση της διαδικασίας αυτής.

Λέξεις – κλειδιά : Παγκόσμιος Ιστός , διασυνδεδεμένα δεδομένα , Σημασιολογικός Ιστός , γεωχωρικές βάσεις δεδομένων , οντολογίες

ABSTRACT

This master's thesis deals with the process of the evolution of the World Wide Web into the Semantic Web and the ways in which the user's search of the World Wide Web can be improved.

In the early days of the World Wide Web perhaps the most important problem for users who wanted to search for information on it was the lack of many and useful sources. Gradually, but at a very fast pace, the World Wide Web turned into one of the largest sources of information used by man as more and more people enter data on all kinds of activities and topics. So the problem of users looking for information is reduced to the quick extraction of the useful information from the huge amount of information provided.

Given this need of the users, terms and techniques such as Data Mining, Information Retrieval, Knowledge Management were created, which try to meet the needs of the users. In addition, in the effort to improve the quality of the results provided to the user, an important role was played by the exploitation of the particular elements that can be extracted for their interests, both in the penetration stage, where pages of a specific topic are gathered (topic-focused crawling), and in the search stage through those most important for each user (personalization). At the same time, as the World Wide Web is gradually evolving into the Semantic Web, new models and standards (XML, RDF, OWL) are being developed to promote this process. The expression, transmission and retrieval of information using these standards opens up new horizons in the use of the Internet. The main object of this work is to exploit the provided models and standards of the Semantic Web in combination with already implemented ideas and algorithms in the simple World Wide Web to enable faster and more accurate information retrieval and processing. Efforts were also made to exploit techniques that take advantage of the particular preferences of each user, and to explore the use of new models and standards of the Semantic Web to promote this process.

Keywords: World Wide Web, linked data, Semantic Web, geospatial databases, ontologies

Πρόλογος

Καθώς αυτή η εργασία αποτελεί το επιστέγασμα των προσπαθειών μου για την ολοκλήρωση των μεταπτυχιακών μου σπουδών θα ήθελα να ευχαριστήσω όλους όσους με στήριξαν στην προσπάθειά μου αυτή, τόσο για να ξεκινήσω το μεταπτυχιακό όσο και να το ολοκληρώσω.

Αμέριστη ήταν η συμπαράσταση και η βοήθεια που είχα κατά το σχεδιασμό και την υλοποίηση της μεταπτυχιακής διπλωματικής εργασίας από την επιβλέπουσα κ.Αδαμοπούλου, όπως επίσης ιδιαίτερη στήριξη μου παρείχε σε όλη τη διάρκεια της συγγραφής αυτής της εργασίας και ο Διδακτορικός Φοιτητής Νίκος Πεππές, τους οποίους ευχαριστώ από καρδιάς.

Θα ήθελα ακόμα να ευχαριστήσω τον σύζυγό μου Φώτη , που ήταν εκείνος που με παρότρυνε για την εκπόνηση αυτού του μεταπτυχιακού και με στήριξε με κάθε μέσο και όχι απλώς μοιράστηκε αλλά επωμίστηκε όλη την αγωνία μου για την ολοκλήρωση των μεταπτυχιακών μου σπουδών μη αφήνοντας τίποτα να με αποσπάσει από το στόχο μου σε όλη τη διάρκεια των σπουδών μου. Για το λόγο αυτό του αφιερώνω και την παρούσα εργασία.

Πίνακας Περιεχομένων

1 Εισαγωγή

Η ψηφιακή εποχή έχει διεισδύσει σε όλες τις πτυχές της ανθρώπινης δραστηριότητας και τις μεταμορφώνει με έναν ιδιαίτερο και επαναστατικό τρόπο. Ο Σημασιολογικός Ιστός (Semantic Web) τα τελευταία χρόνια είναι μία ενεργή πρωτοβουλία του W3C (The World Wide Web Consortium) που αποτελεί έμπνευση του δημιουργού του ίδιου του Διαδικτύου, του Tim Berners-Lee. Το W3C έχει αναπτύξει ένα σύνολο προτύπων και εργαλείων για την υποστήριξη του. Η ανάπτυξή του οφείλεται στη συνεργασία της Επιτροπής του Παγκόσμιου Ιστού (The World Wide Web Consortium \W3C) με πολλούς ερευνητές και ανθρώπους από τη βιομηχανία.

Ο Σημασιολογικός Ιστός (Semantic Web) είναι η εξέλιξη του Παγκόσμιου Ιστού (World Wide Web \(www)) στον οποία τα δεδομένα και το περιεχόμενο του Ιστού έχει καθοριστεί επακριβώς, επιτρέποντας του να κατανοεί καλύτερα τα αιτήματα των χρηστών και των μηχανών αναζήτησης.

Βασίζεται στο RDF (Resource Description Framework), μια μέθοδο περιγραφής ή μοντελοποίησης των πόρων που υπάρχουν στο διαδίκτυο. Πιο συγκεκριμένα, ο Σημασιολογικός Ιστός απαιτεί κοινή μορφή δεδομένων, τα οποία προέρχονται από διαφορετικές πηγές του διαδικτύου. Επίσης, απαιτεί μία κοινή γλώσσα περιγραφής των δεδομένων του διαδικτύου και σύνδεσής τους με τα αντικείμενα του πραγματικού κόσμου τα οποία αντιπροσωπεύουν. Τα δύο αυτά στοιχεία επιτρέπουν σε ένα χρήστη ή μία μηχανή που χρησιμοποιεί το διαδίκτυο να μετακινείται από βάση δεδομένων σε βάση δεδομένων οι οποίες δεν επικοινωνούν επειδή είναι συνδεδεμένες με καλώδιο ή μέσω κάποιου δικτύου, αλλά επειδή αναφέρονται στο ίδιο αντικείμενο με τον ίδιο τρόπο.

Στην εικόνα 2.1 παρουσιάζεται η ιεραρχία και η σχέση των στοιχείων που απαρτίζουν το Σημασιολογικό Ιστό, όπως προτείνει ο Tim Berners-Lee. Δεν θα αναλύσουμε κάθε ένα από αυτά τα στοιχεία ξεχωριστά, απλά θα αναφέρουμε ότι ο Σημασιολογικός Ιστός συνδυάζει τις δυνατότητες που του δίνουν εργαλεία όπως: η XML, XML Schema, RDF, RDF Schema και η OWL. Στο συγκεκριμένο σχήμα φαίνεται η τέταρτη έκδοση (V4) της αρχιτεκτονικής του Σημασιολογικού Ιστού από τις τέσσερις που έχει προτείνει ο Tim Berners-Lee σε διαλέξεις που έχει δώσει κατά καιρούς. Η πρώτη έκδοση (V1) παρουσιάστηκε το 2000, η V2 το 2003, η V3 το 2005 και η V4 τον Ιούλιο του 2006.

Η διπλωματική αυτή πραγματεύεται τη χρήση των τεχνολογιών του Σημασιολογικού Ιστού με άμεση εφαρμογή σε βάσεις δεδομένων για την άντληση πληροφοριών από αυτές. Έχει ως στόχο την εμβάθυνση στις τεχνολογίες Σημασιολογικού Ιστού, στην ανάδειξη των πλεονεκτημάτων του και στην ανάπτυξη μιας καινοτόμου αναζήτησης πρακτικά μέσα από τη βάση δεδομένων.

Στο 1ο κεφάλαιο γίνεται μια εισαγωγή στην εργασία. Παρουσιάζεται ο σκοπός της παρούσας εργασίας μαζί με κάποιες εισαγωγικές έννοιες απαραίτητες για την κατανόησή της , καθώς επίσης και η δομή της.

Στο δεύτερο κεφάλαιο γίνεται αναφορά στις σημαντικές έννοιες με τις οποίες θα ασχοληθούμε. Δίνονται ορισμοί, όπου αυτό είναι εφικτό καθώς και παραδείγματα χρήσης του Σημασιολογικού Ιστού, μέσα από τα οποία γίνεται εμβάθυνση της έννοιάς του και ομαλή ένταξη στο παράδειγμα που θα αναπτυχθεί στη συνέχεια.

Στο 3ο κεφάλαιο γίνεται η ανάπτυξη του παραδείγματος μέσα από κατάλληλη βάση δεδομένων και απεικόνιση της χρήσης του Σημασιολογικού Ιστού με πραγματικά δεδομένα Συγκεκριμένα, παρουσιάζεται η μεθοδολογία που ακολουθήσαμε προκειμένου να αναπτύξουμε την οντολογία και η λεπτομερής ανάλυσή της.

Στο 4ο κεφάλαιο γίνεται εξαγωγή των συμπερασμάτων Κι πιθανές εφαρμογές της οντολογίας που αναπτύξαμε καθώς και διάφορα συμπεράσματα σχετικά με την έρευνα που πραγματοποιήθηκε στη συγκεκριμένη εργασία.

Τέλος, στην εικόνα παρατηρούμε μια σύνοψη της δομής της εργασίας αποτυπωμένη σε εικόνα (Εικόνα 1).

1.1 Η σημασία της ολοκληρωμένη διαχείρισης στον παγκόσμιο ιστό

Η ολοκληρωμένη διαχείριση δεδομένων στον Παγκόσμιο Ιστό αποτελεί κρίσιμο ζήτημα δεδομένου του όγκου της πολυπλοκότητας των δεδομένων που δημιουργούνται, ανταλλάσσονται και αποθηκεύονται καθημερινά σε διάφορες πλατφόρμες στον Παγκόσμιο Ιστό. Η ολοκληρωμένη διαχείριση δεδομένων συνεπάγεται την ακόλουθη σειρά βημάτων:

Συλλογή δεδομένων: Η αρχή όλης της διαδικασίας είναι η συλλογή των δεδομένων. Τα δεδομένα αυτά μπορεί να είναι προσωπικά, επιχειρησιακά αισθητήρων και άλλα. Πρωταρχικός στόχος είναι να προσδιοριστεί ο σκοπός για τον οποίο κρίνεται απαραίτητη η συλλογή των δεδομένων και η επιλογή	**Προστασία δεδομένων**: Είναι αναγκαίο πλέον σύμφωνα με την ισχύουσα νομοθεσία να υιοθετούνται μέτρα ασφαλείας όπως η κρυπτογράφηση των δεδομένων και η προστασία τους από κακόβουλες επιθέσεις. Συμμόρφωση σύμφωνα με τον κανονισμό GDPR ή σχετικές νομοθεσίες περί απορρήτου.

των πηγών από τις οποίες θα αντληθούν τα δεδομένα. Η διαδικασία συλλογής δεδομένων μπορεί και να αυτοματοποιηθεί μέσω web scraping ή άλλες τεχνολογίες.	
Αποθήκευση των δεδομένων: Όλα αυτά τα δεδομένα που συλλέχθηκαν πρέπει να αποθηκεύονται σε ασφαλείς βάσεις δεδομένων. Για το σκοπό αυτό χρησιμοποιούνται συστήματα διαχείρισης βάσεων δεδομένων (DBMS) και αποθηκευτικές λύσεις στο νέφος (cloud).	**Ανταλλαγή δεδομένων**: Τα δεδομένα μπορούν να ανταλλάσσονται στον Παγκόσμιο Ιστό μέσω πρωτοκόλλων επικοινωνίας και του Διαδικτύου.
Επεξεργασία δεδομένων: Μετά την συλλογή και την αποθήκευση των δεδομένων είναι απαραίτητη η επεξεργασία τους. Αυτή μπορεί να περιλαμβάνει την αφαίρεση αδιάφορων πληροφοριών, την αντιστοίχιση και των χαρακτηρισμό των δεδομένων. Με τη χρήση αλγορίθμων μηχανικής μάθησης και τεχνικών ανάλυσης δεδομένων οι πληροφορίες υποβάλλονται σε επεξεργασία προκειμένου να ανακαλυφθούν προτεραιότητες, πρότυπα και συσχετίσεις μεταξύ τους.	**Διαχείριση Δεδομένων**: Περιλαμβάνει τη διαχείριση βάσεων δεδομένων, τη διαχείριση της απόθεσης, την επικαιροποίηση, τον έλεγχο της ποιότητας και την αποθήκευση ανάλυσης δεδομένων.

Πίνακας 1 : Η ολοκληρωμένη διαχείριση δεδομένων στον Παγκόσμιο Ιστό

Οι μηχανές αναζήτησης του Διαδικτύου έχουν γίνει πλέον αναπόσπαστο κομμάτι της καθημερινής ζωής όσων ασχολούνται με την τεχνολογία της πληροφορικής. Η χρήση τους επεκτείνεται από απλή, όπως για προσωπική χρήση στο σπίτι, έως το γραφείο και την εργασία τους και από εκεί στο Παγκόσμιο Δίκτυο και τα εσωτερικά Δίκτυα μεγάλων επιχειρήσεων (*e-business*), την αγορά (*e-shopping*), την διασκέδαση (*e-entertainment*) και την εκπαίδευση (*e-*

learning). Ωστόσο, παρόλο την ευρεία χρήση τους και τις ανάγκες της κοινωνίας η αποδοτικότητα των μηχανών αυτών ως συστήματα ανάκτησης πληροφοριών (IR – information retrieval) παραμένει απογοητευτικά χαμηλή σύμφωνα με το Working Group on Government Information Navigation, το οποίο προσδιορίζει τα προβλήματα των μηχανών αναζήτησης στο Παγκόσμιο Δίκτυο ως εξής:

Κίνδυνος να μη βρεθούν ιστοσελίδες αν χρησιμοποιούν για απεικόνιση των δεδομένων τους κείμενο διαφορετικό από HTML.	Οι μηχανές αναζήτησης δεν έχουν πάντα τα πιο πρόσφατα αρχεία αφού ανανεώνουν τις βάσεις τους σε συγκεκριμένα χρονικά διαστήματα.
Οι μηχανές αναζήτησης δεν ελέγχουν κάθε σελίδα, αλλά μόνο αυτές που βρίσκονται στα δύο – τρία πρώτα ιεραρχικά επίπεδα, χάνοντας πληροφορίες που μπορεί να βρίσκονται σε επόμενα.	Επιστρέφουν άσχετη πληροφορία που επιστρέφεται καθώς δεν είναι σε θέση να αναγνωρίσουν τις σημαντικές πληροφορίες από αυτές που τυχαία εμφανίζονται στο κείμενο.

Πίνακας 2: Τα προβλήματα των μηχανών αναζήτησης στο Παγκόσμιο Δίκτυο

Πιο συγκεκριμένα η ουσία του προβλήματος εντοπίζεται ακόμα στο γεγονός ότι η_αναζήτηση στον Παγκόσμιο Ιστό αντιπροσωπεύεται από την αντιστοίχιση επερωτήσεων (*queries*) και εγγράφων στο επιφανειακό επίπεδο της γλωσσολογικής ανάλυσης λέξεων κλειδιών αντί για το βαθύτερο επίπεδο που περιβάλει και δίνει ξεχωριστό νόημα σε κάθε ερώτηση και τη σημασία της.

Είναι γεγονός ότι η πολυσημία και η συνωνυμία των λέξεων παρουσιάζονται συχνά στη φυσική γλώσσα, με αποτέλεσμα η χρήση λέξεων κλειδιών να καθιστά δύσκολη την προσπάθεια να κατανοηθεί η ανάγκη του χρήστη που θέτει το ερώτημα στην μηχανή αναζήτησης.

Λύση στο παραπάνω πρόβλημα έρχεται να δώσει το όραμα του Σημασιολογικού Ιστού με το οποίο θα ασχοληθούμε εκτενέστερα στο Κεφάλαιο 2 της παρούσας εργασίας. Σύμφωνα με τον Σημασιολογικό Ιστό τα δεδομένα δεν θα αποτυπώνονται απλώς στην καθημερινή γλώσσα, αλλά σε μια υβριδική μορφή που θα περιλαμβάνει τμήματα κειμένου άμεσα συνδεδεμένα με επεξηγηματικές σημασιολογικές ετικέτες που θα αναφέρονται σε έννοιες τυπικά ορισμένες σε οντολογίες προσβάσιμες στο Δίκτυο.

Με τον όρο οντολογία στον Παγκόσμιο Ιστό αναφερόμαστε στη χρήση της Σημασιολογικής Ιστολογίας, μιας τεχνολογίας που σχετίζεται με την αναζήτηση και την οργάνωση πληροφοριών στον Παγκόσμιο Ιστό και ειδικότερα της εκχώρησης σημασίας στις δεδομένες πληροφορίες. Μία οντολογία μπορεί να περιλαμβάνει μια ιεραρχία εννοιών και τις σχέσεις μεταξύ τους, που επιτρέπουν στις υπολογιστικές συστοιχίες να παράγουν σημαντικές συσχετίσεις και συμπεράσματα από τα δεδομένα του Παγκόσμιου Ιστού.

Οι οντολογίες μπορούν να κατηγοριοποιηθούν σε τέσσερις κατηγορίες, οι οποίες είναι οι εξής:

Γενικές οντολογίες : καθορίζουν υψηλού επιπέδου γενικές έννοιες κοινές στα περισσότερα θεματικά πεδία.
Ειδικές οντολογίες : ορίζουν εξειδικευμένες έννοιες για περιορισμένες θεματικές ενότητες.
Γλωσσολογικές οντολογίες : καθορίζουν έννοιες, οι οποίες βασίζονται στην ύπαρξη λέξεων που αναφέρονται σε μία ή περισσότερες φυσικές γλώσσες.
Καθαρά εννοιολογικές οντολογίες : καθορίζουν έννοιες αποκλειστικά βασιζόμενες στη χρησιμότητα που θα έχουν αυτές οι έννοιες στη διευκόλυνση αυτοματοποιημένων εργασιών λήψης αποφάσεων από τους έξυπνους πράκτορες (*agents*).

Πίνακας 3 : Κατηγοριοποίηση οντολογιών

Στον Σημασιολογικό Ιστό οι χρήστες θα μπορούν να αναζητούν τις πληροφορίες που χρειάζονται χρησιμοποιώντας όρους από μία ή περισσότερες οντολογίες αντί για λέξεις κλειδιά. Έτσι, οι νέες μηχανές αναζήτησης θα αντιστοιχούν το περιεχόμενο των εγγράφων με τις ανάγκες του χρήστη στο βαθύτερο επίπεδο της σημασιολογικής ανάλυσης, βελτιώνοντας πολύ την ακρίβεια των αποτελεσμάτων τους.

Τα τελευταία χρόνια έχουν γίνει σημαντικά βήματα για τη μετατροπή του Παγκόσμιου Ιστού σε μία παγκόσμια βάση δεδομένων χρήσιμων τόσο για τους χρήστες, δηλαδή τους ανθρώπους, όσο και για τα προγράμματα λογισμικού που λειτουργούν ως πράκτορες. Ακόμη, έχουν δημιουργηθεί πολλές γενικού και ειδικού περιεχομένου οντολογίες , οι οποίες είναι διαθέσιμες

στον Παγκόσμιο Ιστό, αλλά ταυτόχρονα έχουν σχηματιστεί και εργαλεία που είναι απαραίτητα για τη συντήρησή τους.

1.1.1 Η έννοια του Παγκόσμιου Ιστού και η εξέλιξή του στο χρόνο

Ο παγκόσμιος ιστός είναι μια μεγάλη και επιτυχημένη προσπάθεια αν αναλογιστεί κανείς την ποσότητα της διαθέσιμης πληροφορίας, και το ρυθμό αύξησης των ανθρώπων που τον χρησιμοποιούν. Έχει αρχίσει να διαπερνά τις περισσότερες πλευρές της καθημερινής μας ζωής, αλλά και της ζωής των επιχειρήσεων και οργανισμών. Η επιτυχία του βασίζεται στην απλότητα του. Η απλότητα του HTTP και της HTML που στοχεύουν στη μεταφορά και προβολή υπερκειμένων, επέτρεψε στους κατασκευαστές λογισμικού, στους φορείς παροχής πληροφοριών και τους τελικούς χρήστες, την εύκολη προσπέλαση στο νέο μέσο και συνέβαλε στο να αναπτύξουν σε πολύ μικρό χρονικό διάστημα μια σημαντική κρίσιμη μάζα.

Ο Παγκόσμιος Ιστός βασίζεται στην ιδέα των κατανεμημένων πληροφοριών. Αντί όλες οι πληροφορίες να φιλοξενούνται σε ένα σημείο, κάθε οντότητα που διαθέτει πληροφορίες τις οποίες θέλει να μοιραστεί με άλλους τις αποθηκεύει στο δικό της υπολογιστή και επιτρέπει στους χρήστες του Internet να τις προσπελάζουν. Με άλλα λόγια, αποτελεί μία συλλογή από έγγραφα πολυμέσων.

1.1.2 Ιστορική διαδρομή Διαδικτύου

Στα τέλη της δεκαετίας του 1960, η επικοινωνία μεταξύ δύο υπολογιστών κατέστη δυνατή μέσω ενός δικτύου υπολογιστών. Στις αρχές της δεκαετίας του 1980 εισήχθη το Πρωτόκολλο Ελέγχου Μετάδοσης/Πρωτόκολλο Διαδικτύου (Transmission Control Protocol/Internet Protocol; TCP/IP), επιτρέποντας την εμπορική χρήση του Διαδικτύου στα τέλη της δεκαετίας του 1980. Αργότερα, ο Παγκόσμιος Ιστός (WorldWide Web; WWW) έγινε διαθέσιμος το 1991, γεγονός που έκανε το Διαδίκτυο πιο δημοφιλές και τόνωσε την ταχεία ανάπτυξή του. Έπειτα, οι κινητές συσκευές άρχισαν να συνδέονται με το Διαδίκτυο και σχημάτισαν το Κινητό-Διαδίκτυο. Με την εμφάνιση της κοινωνικής δικτύωσης, οι χρήστες άρχισαν να συνδέονται μέσω του Διαδικτύου. Το τελευταίο βήμα ήταν το Διαδίκτυο των Πραγμάτων, όπου τα αντικείμενα γύρω μας μπορούν να συνδεθούν μεταξύ τους (π.χ., μηχανή με μηχανή) και να επικοινωνούν μέσω του Διαδικτύου.

Ο όρος "Διαδίκτυο των Πραγμάτων" εισήχθη για πρώτη φορά από τον Kevin Ashton, το 1998. Αναφέρει ότι το Διαδίκτυο των Πραγμάτων έχει τη δυνατότητα να αλλάξει τον κόσμο, όπως ακριβώς έκανε το Διαδίκτυο, και ίσως ακόμη περισσότερο. Αργότερα, το IoT εισήχθη επίσημα από τη Διεθνή Ένωση Τηλεπικοινωνιών (International Telecommunication Union; ITU) το 2005.

Τα τελευταία χρόνια η εξέλιξη του Παγκόσμιου Ιστού, αλλά και του Διαδικτύου των Πραγμάτων γεννούν όλο και περισσότερες πηγές δεδομένων, κάτι που οδηγεί στην ανάγκη καθιέρωσης μίας προηγμένης προσέγγισης διαχείρισης της πληροφορίας. Ταυτόχρονα, κρίνεται αναγκαία η ενσωμάτωση του συνόλου των δεδομένων, καθώς με τον τρόπο αυτό είναι δυνατή η σφαιρική κατανόηση της κατανεμημένης πληροφορίας, η συγχώνευση των πληροφοριών (information fusion) και οι σχέσεις που επικρατούν μεταξύ των ποικίλων τμημάτων της. Μία τεχνολογία, η οποία διασφαλίζει την αποδοτική διαχείριση και ενσωμάτωση ετερογενών δεδομένων και, κατ' επέκταση, την αποτελεσματική διασύνδεση εφαρμογών και συστημάτων, είναι η σημασιολογική περιγραφή των υποκείμενων δεδομένων και διαδικασιών. Αυτό συνεπάγεται την περιγραφή των προηγουμένων έτσι, ώστε το νόημά τους να είναι κατανοητό από μηχανές. Αυτή η θεμελιώδης ιδέα του σημασιολογικού εμπλουτισμού των δεδομένων οδήγησε στην εμφάνιση του Σημασιολογικού Ιστού (Semantic Web). Η ιδέα αυτή εισήχθη το 2001 από τον δημιουργό του Παγκόσμιου Ιστού (World Wide Web), Tim Berners-Lee.

1.2 Στόχοι της Παρούσας Διπλωματικής

Παρόλο που έχει τεθεί το πλαίσιο της λειτουργίας του Σημασιολογικού Ιστού, η μετατροπή του υπάρχοντος Διαδικτύου δεν μπορεί να γίνει από τη μία μέρα στην άλλη. Η μετατροπή αυτή θα είναι μία αργή και σταδιακή διαδικασία. Συγκεκριμένα, πρέπει πρώτα να δημιουργηθούν πιο σύνθετες βάσεις δεδομένων, τυπική γλώσσα και υψηλού επιπέδου προγραμματιστικές - τεχνικές ικανότητες από τους χρήστες του Διαδικτύου που ανεβάζουν πληροφορίες στον Παγκόσμιο Ιστό.

Με άλλα λόγια το Διαδίκτυο έγινε ευρέως χρησιμοποιούμενο καθώς η σύνταξη ενός HTML εγγράφου για αναζήτηση πληροφοριών αποτελεί απλά την πληκτρολόγηση κειμένου, τη μορφοποίηση και την κατανόηση πλοήγησης μέσω υπερσυνδέσμων. Αντίθετα, στον Σημασιολογικό Ιστό ο σχολιασμός εγγράφων από τις υπάρχουσες οντολογίες απαιτεί διαχείριση γνώσης (knowledge engineering) που θα δυσκολέψει το μέσο σχεδιαστή ιστοσελίδας.

Στη παρούσα εργασία έγινε μια προσπάθεια να μελετηθούν και να κατανοηθούν οι δυνατότητες του Σημασιολογικού ιστού προς το σκοπό της ταχύτερης διαπέρασης ιστοσελίδων του Σημασιολογικού ιστού οι οποίες σχετίζονται με συγκεκριμένο θέμα προτίμησης του χρήστη που ζήτησε τη διαπέραση αυτή.

Επιπλέον, μπόρεσαν να εξαχθούν χρήσιμα συμπεράσματα για την αποδοτικότητα της προτεινόμενης μεθόδου και να ενισχυθεί η πεποίθηση πως παρόμοια πειράματα θα διεξαχθούν

με μεγαλύτερη αποτελεσματικότητα στο μέλλον όπου το όραμα του Σημασιολογικού ιστού πλέον υπάρχει μέσα στην καθημερινότητα κάθε χρήστη του Παγκόσμιου Ιστού. Προς το παρόν, έχουν κάνει την εμφάνισή τους κυρίως εναλλακτικές μέθοδοι αύξησης της αποδοτικότητας των μηχανών αναζήτησης με μερική χρήση των τεχνολογιών που αναπτύσσονται για τη μετατροπή του διαδικτύου στο Σημασιολογικό ιστό προτού αυτή η αλλαγή γίνει πραγματικότητα.

2.1 Η Έννοια των δεδομένων και των πληροφοριακών συστημάτων στο διαδίκτυο

Τα δεδομένα αποτελούν μεμονωμένα γεγονότα, τα οποία μπορεί να είναι στατιστικά στοιχεία ή στοιχεία πληροφοριών και συχνά είναι αριθμητικού τύπου. Αν χρησιμοποιήσουμε μια έννοια η οποία είναι πιο τεχνική, τα δεδομένα αντιπροσωπεύουν ένα σύνολο τιμών οι οποίες μπορεί να είναι είτε ποιοτικές, είτε ποσοτικές μεταβλητές για ένα ή περισσότερα πρόσωπα ή αντικείμενα, ενώ ένα δεδομένο είναι μια ενιαία τιμή μιας μεμονωμένης μεταβλητής.

Τα δεδομένα συνεισφέρουν στην λήψη των αποφάσεων, καθώς αποτελούν μικρότερες μονάδες από πραγματικές πληροφορίες οι οποίες μπορούν να χρησιμοποιηθούν ως βάση για συζήτηση και υπολογισμούς. Διακρίνονται στις παρακάτω κατηγορίες, οι οποίες είναι οι εξής:

Τα ακατέργαστα δεδομένα αποτελούν μια συλλογή χαρακτήρων ή αριθμών πριν αυτή «καθαριστεί» και πάρει την επιθυμητή μορφή από τους ερευνητές.

Τα ανεπεξέργαστα δεδομένα πρέπει να διορθωθούν και να αφαιρεθούν τυχόν ακραίες τιμές ή σφάλματα, όπως για παράδειγμα είναι μια λανθασμένη ένδειξη θερμομέτρου.

Η διαδικασία της επεξεργασίας δεδομένων πραγματοποιείται συνήθως κατά στάδια και τα «επεξεργασμένα δεδομένα» περνάνε στην φάση όπου πλέον, μπορούν να θεωρηθούν τα «ακατέργαστα δεδομένα» του επόμενου σταδίου.

Τα δεδομένα μετρούνται, συλλέγονται, αναλύονται και χρησιμοποιούνται για την δημιουργία οπτικοποιημένων γραφημάτων, πινάκων ή εικόνων.

Πιο συγκεκριμένα, ο όρος δεδομένα στον Παγκόσμιο Ιστό αναφέρεται στις πληροφορίες που ανακαλούνται μέσω του Παγκόσμιου Ιστού ή απλά του Διαδικτύου. Τα δεδομένα στο Διαδίκτυο μπορούν να περιλαμβάνουν κείμενο, εικόνες ήχο, κωδικούς πηγής, πληροφορίες σχετικά με τοποθεσίες και άλλα. Όλα αυτά τα δεδομένα αποθηκεύονται σε διάφορους διακομιστές και είναι προσβάσιμα μέσω του προγράμματος περιήγησης στο Διαδίκτυο.

Η ολοκληρωμένη διαχείριση δεδομένων στον Παγκόσμιο Ιστό αποτελεί κρίσιμο ζήτημα δεδομένου του όγκου της πολυπλοκότητας των δεδομένων που δημιουργούνται, ανταλλάσσονται και αποθηκεύονται σε διάφορες πλατφόρμες στον Παγκόσμιο Ιστό. Η ολοκληρωμένη διαχείριση δεδομένων συνεπάγεται την ακόλουθη σειρά βημάτων:

- **Συλλογή δεδομένων**: Η αρχή όλης της διαδικασίας είναι η συλλογή των δεδομένων. Τα δεδομένα αυτά μπορεί να είναι προσωπικά, επιχειρησιακά αισθητήρων και άλλα.

- **Αποθήκευση των δεδομένων**: Όλα αυτά τα δεδομένα που συλλέχθηκαν πρέπει να αποθηκεύονται σε ασφαλείς βάσεις δεδομένων. Για το σκοπό αυτό χρησιμοποιούνται συστήματα διαχείρισης βάσεων δεδομένων (DBMS) και αποθηκευτικές λύσεις στο νέφος.
- **Επεξεργασία δεδομένων**: Με τη χρήση αλγορίθμων μηχανικής μάθησης και τεχνικών ανάλυσης δεδομένων οι πληροφορίες υποβάλλονται σε επεξεργασία προκειμένου να ανακαλυφθούν προτεραιότητες, πρότυπα και συσχετίσεις μεταξύ τους.
- **Προστασία δεδομένων**: Υιοθετούνται μέτρα ασφαλείας όπως η κρυπτογράφηση και η προστασία από κακόβουλες επιθέσεις ,αφού είναι απαραίτητη η διαφύλαξη των δεδομένων.
- **Ανταλλαγή δεδομένων**: Τα δεδομένα μπορούν να ανταλλάσσονται στον Παγκόσμιο Ιστό μέσω πρωτοκόλλων επικοινωνίας και του Διαδικτύου.
- **Συμμόρφωση σύμφωνα με τον κανονισμό GDPR** ή σχετικές νομοθεσίες περί απορρήτου.
- **Διαχείριση Δεδομένων**: Περιλαμβάνει τη διαχείριση βάσεων δεδομένων, τη διαχείριση της απόθεσης, την επικαιροποίηση `σ, τον έλεγχο της ποιότητας και την αποθήκευση ανάλυσης δεδομένων.

Ένας συνεχώς αυξανόμενος αριθμός δεδομένων διατίθενται μέσω του Διαδικτύου λόγω της ευρείας χρήσης του και προκύπτει ως ένα βαθμό από την προσπάθεια να καλυφθούν οι πληροφοριακές ανάγκες ποικίλων ομάδων χρηστών. Με σκοπό, λοιπόν, την κάλυψη συγκεκριμένων αναγκών έχουν δημιουργηθεί πληροφοριακά συστήματα, τα οποία χαρακτηρίζονται από μεγάλο βαθμό αυτονομίας σε ποικίλα επίπεδα, όπως είναι οι διαφορετικές δυνατότητες αναζήτησης και ανάκτησης δεδομένων σε κάθε ένα από αυτά.

Η αυτονομία στο σχεδιασμό πληροφοριακών συστημάτων οδηγείσαι στην εμφάνιση ετερογένειας σε τέσσερα διαφορετικά επίπεδα, τα οποία είναι τα εξής:

- **Ετερογένεια συστημάτων**: πρόκειται για διαφορετικές πλατφόρμες, λειτουργικά συστήματα και πρωτόκολλα δικτύου.
- **Ετερογένεια στη σύνταξη**: πρόκειται για διαφορές στην κωδικοποίηση, στα πρωτόκολλα επικοινωνίας , στα formats των δεδομένων.
- **Ετερογένεια σχημάτων**: είναι το αποτέλεσμα της χρήσης διαφορετικών μοντέλων δεδομένων, δομών δεδομένων και σχημάτων κωδικοποίησης τους ανάμεσα σε πηγές.
- **Σημασιολογική ετερογένεια**: προκύπτει από τις σημασιολογικές αντιθέσεις, οι οποίες προκύπτουν όταν η σημασία δεδομένων μπορεί να εκφραστεί με διαφορετικούς τρόπους.

Η ετερογένεια αυτή στα πληροφοριακά συστήματα προκαλεί δυσκολίες στους χρήστες που αναζητούν μία ενιαία αναζήτηση με στόχο να καλύψουν τις πληροφοριακές τους ανάγκες. Προκειμένου, λοιπόν, να καλυφθούν οι ανάγκες για μία ενιαία αναζήτηση δημιουργήθηκαν τα συστήματα ολοκλήρωσης δεδομένων (*data integration systems*), τα οποία παρέχουν στον χρήστη πρόσβαση σε ομάδα αυτόνομων πηγών, οι οποίες όμως λειτουργούν ως σύνολο μία πηγή δεδομένων.

Στα συστήματα ολοκλήρωσης δεδομένων υπάρχει ένα σχήμα διαμεσολάβησης ή καθολικό σχήμα που ο χρήστης υποβάλλει το ερώτημά του. Στο καθολικό αυτό σχήμα υπάρχει ένας πυρήνας έκφρασης των δεδομένων που διαθέτουν οι πηγές. Οι σχέσεις μεταξύ του καθολικού σχήματος και των τοπικών πηγών(*local sources*) εκφράζονται μέσα από όψεις(*views*). Για να επιλυθεί το όποιο πρόβλημα σημασιολογικής ετερογένειας ανάμεσα στα τοπικά δεδομένα και το καθολικό σχήμα ορίζονται κανόνες συσχέτισης (*mapping rules*).

Παρόλο όμως που οι μηχανές αναζήτησης του Διαδικτύου έχουν γίνει αναπόσπαστο κομμάτι της καθημερινής ζωής όσων ασχολούνται με την τεχνολογία της Πληροφορικής, η αποδοτικότητά τους ως συστήματα ανάκτησης πληροφοριών παραμένει απογοητευτικά χαμηλή.

Οι μηχανές αναζήτησης λοιπόν έχουν μικρή ακρίβεια επιστρέφοντας υπερβολικά πολλά άσχετα έγγραφα και ταυτόχρονα δεν εντοπίζουν το σύνολο των χρήσιμων. Στην ουσία ο πυρήνας του προβλήματος έγκειται στην αναζήτηση που κάνουν οι χρήστες στον Παγκόσμιο Ιστό που αντιπροσωπεύεται από την αντιστοίχιση ερωτήσεων (*queries*) στο επιφανειακό επίπεδο της γλωσσολογικής ανάλυσης λέξεων κλειδιών, αντί για το βαθύτερο επίπεδο που λαμβάνει υπόψη το πλαίσιο που περιβάλει και δίνει ξεχωριστό νόημα σε κάθε ερώτηση και στη σημασία της.

Εξαιτίας της συνωνυμίας και της πολυσημίας των λέξεων καθίσταται δύσκολη η χρήση λέξεων κλειδιών που θα εξυπηρετήσει τις ανάγκες του κάθε χρήστη και θα απαντήσει την ερώτησή του. Λύση σε αυτό το πρόβλημα δίνει η έννοια του Σημασιολογικού Ιστού.

2.1.1 Ο Σημασιολογικός Ιστός

Ο Παγκόσμιος Ιστός ανακαλύφθηκε από τον Lee, του οποίου το αρχικό όραμα του ήταν πιο φιλόδοξο από την πραγματικότητα του υπάρχοντος Ιστού. Συγκεκριμένα, ο Lee είχε αναφέρει:

«Ένας από τους στόχους του Ιστού ήταν η απάντηση στην ερώτηση, εάν η αλληλεπίδραση μεταξύ του ατόμου και του υπερκειμένου μπορούσε να είναι τόσο διαισθητική, ώστε ο αναγνώσιμος από μηχανή χώρος πληροφοριών να έδινε μια ακριβή αναπαράσταση της

κατάστασης των σκέψεων των ανθρώπων, τις αλληλεπιδράσεις και των μοντέλων εργασίας τους. Τότε η μηχανική ανάλυση θα μπορούσε να γίνει πολύ ισχυρό διοικητικό εργαλείο, παρακολουθώντας τα μοντέλα εργασίας μας και διευκολύνοντας τη συνεργασία μας από τα συνήθη προβλήματα που περιστοιχίζουν τη διοίκηση των μεγάλων οργανισμών»

Ως τώρα οι ιστοσελίδες χρησιμοποιούν μέσα που στηρίζονται σε γλώσσες, όπως HTML και χρησιμοποιούν πρωτόκολλα που επιτρέπουν στις μηχανές αναζήτησης να αναπαράγουν πληροφορίες στους ανθρώπινους αναγνώστες. Όμως τα προβλήματα που δημιουργούνται στην πρόσβαση και στην επεξεργασία των διαθέσιμων πληροφοριών οδήγησαν στην προσπάθεια για να αντιπροσωπευθεί το περιεχόμενο του Ιστού με μια μορφή που είναι ευκολότερα επεξεργάσιμη από τη μηχανή. Αποτέλεσμα αυτής της προσπάθειας ήταν η δημιουργία του Σημασιολογικού Ιστού.

Ο Σημασιολογικός Ιστός είναι ειδικότερα ένας Ιστός από πληροφορίες που είναι δυνατό να διαβιβαστούν από τις μηχανές και η έννοια των οποίων είναι σαφώς καθορισμένη από πρότυπα: χρειάζεται απολύτως τη δια λειτουργική υποδομή που μόνο τα παγκόσμια τυποποιημένα πρωτόκολλα παρέχουν.

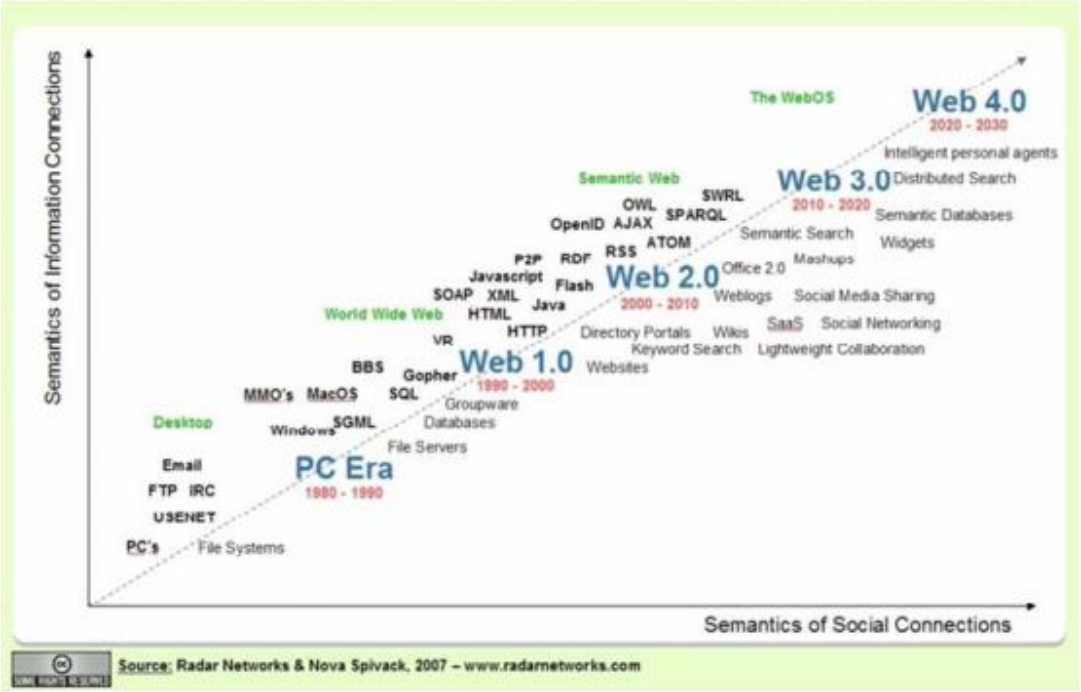
Ο Σημασιολογικός Ιστός, λοιπόν, αποτελεί μια επέκταση του σημερινού Ιστού, η οποία έχει ως σκοπό την αυτοματοποίηση των λειτουργιών και των εφαρμογών του διαδικτύου. Η αυτοματοποίηση αυτή μπορεί να πραγματοποιηθεί μόνο εάν η γνώση και η πληροφορία που υπάρχει αποθηκευμένη και δημοσιευμένη αυτή την στιγμή στον σημερινό Παγκόσμιο Ιστό αποκτήσει τυπικό νόημα και σημασιολογία και δομηθεί με έναν τέτοιο τρόπο ώστε να γίνεται κατανοητή από τις μηχανές που την επεξεργάζονται. Υπάρχει, πλέον, παγκοσμίως η διάθεση συνεργασίας μεταξύ των, ανά τον κόσμο, χρηστών του ιστού με σκοπό να διασπείρουν πληροφορίες εμπλουτισμένες από τη γνώση όχι μόνο ενός ατόμου, αλλά πολλών.

Σύμφωνα με το όραμα των εμπνευστών του Σημασιολογικού Ιστού, η προσθήκη σημασίας στην πληροφορία του Διαδικτύου, θα απελευθερώσει πλήθος δυνατοτήτων για την πιο ευφυή εκμετάλλευση της πληροφορίας αυτής. Ένας χρήστης του Διαδικτύου θα μπορεί, για παράδειγμα, μεταξύ άλλων δυνατοτήτων, να πραγματοποιεί ευφυείς αναζητήσεις, να λαμβάνει δηλαδή από μια μηχανή αναζήτησης αποτελέσματα τα οποία να είναι πιο σχετικά με αυτό που πραγματικά αναζητά.

Από την άλλη πλευρά, η αποθήκευση μεγάλου όγκου πληροφοριών σε βάσεις δεδομένων οδήγησε στην εμφάνιση ενός προβλήματος. Το πρόβλημα αυτό ήταν ότι καθιερώθηκε και διατηρήθηκε η σημασιολογία των δεδομένων που είναι αποθηκευμένα στις βάσεις δεδομένων. Αυτό το πρόβλημα σημασιολογίας δεδομένων παρέμενε ελέγξιμο όσο ελέγξιμες ήταν και οι αλλαγές που θα μπορούσαν να προκληθούν στη κάθε βάση δεδομένων. Δηλαδή, οι

πληροφορίες που ήταν διαθέσιμες, μπορούσαν να αλλοιωθούν μόνο από συγκεκριμένο αριθμό ατόμων και συγκεκριμένο αριθμό εφαρμογών.

Με την εμφάνιση όμως του Παγκόσμιου Ιστού το τοπίο αυτό άλλαξε. Μεταβλήθηκε κατά πολύ ο τρόπος με τον οποίο διαρθρώνεται η επικοινωνία μεταξύ των ανθρώπων και ειδικά ο τρόπος με τον οποίο η πληροφορία που υπάρχει διαθέσιμη ανά τον κόσμο, διαδίδεται και ανακτάται. Αμέτρητοι πλέον χρήστες και εφαρμογές μπορούν και έχουν πρόσβαση στις βάσεις δεδομένων που είναι διαθέσιμες στον Ιστό. Υπό αυτές τις συνθήκες, η σημασιολογία κάθε πληροφορίας πρέπει να είναι διαθέσιμη στον κάθε χρήστη μαζί με την ίδια την πληροφορία. Όταν ως χρήστης εννοείται κάποιο φυσικό πρόσωπο, αυτό μπορεί να επιτευχθεί με την κατάλληλη επιλογή κάποιας σχηματικής παρουσίασης για σημασιολογικά δεδομένα. Όταν, όμως, πρόκειται για κάποια εφαρμογή που θα αποκτήσει πρόσβαση στη βάση, η σημασιολογία πρέπει να είναι δομημένη σε μορφή που θα είναι προσπελάσιμη και κατανοητή από τη μηχανή που θα αναλάβει την επεξεργασία της. Επομένως, κρίθηκε απαραίτητη η επέκταση του σημερινού Ιστού, η οποία κατέληξε στη δημιουργία του Σημασιολογικού Ιστού, γνωστό και ως Semantic Web3.0.



Εικόνα 1: Η Εξέλιξη του Σημασιολογικού Ιστού από την εποχή των προσωπικών υπολογιστών (PC) έως το Web 4.0.

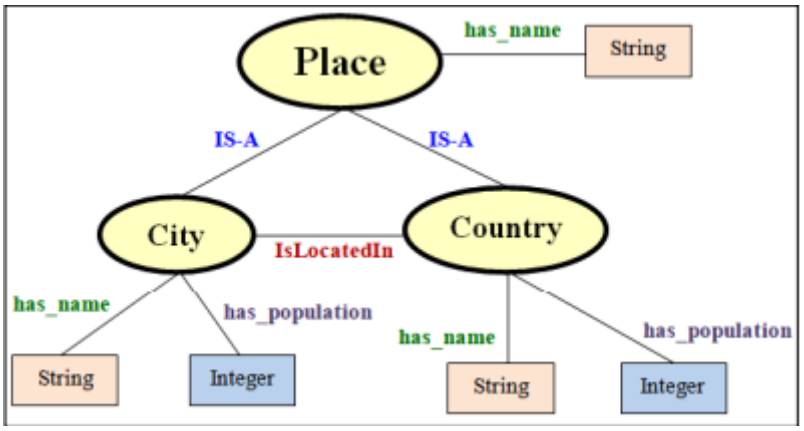
Πηγή: http://www.novaspivack.com/technology/web-3-0-the-best-official-definition-imaginable.

2.1.2 Οντολογίες

Οντολογία (*Ontology*) είναι η αυστηρά μαθηματική περιγραφή ενός πεδίου γνώσης και περιλαμβάνει ένα πεπερασμένο σύνολο από όρους και τις σημασιολογικές συσχετίσεις μεταξύ αυτών. Οι όροι περιγράφουν κλάσεις αντικειμένων, δηλαδή έννοιες σχετικές με αντικείμενα και οι συσχετίσεις αφορούν ιεραρχικές σχέσεις μεταξύ των όρων αυτών.

Τα συστατικά μέρη μιας οντολογίας είναι:

- οι κλάσεις (*classes*), που αναπαριστούν έννοιες.
- οι συσχετίσεις (*relations*) οι οποίες εκφράζουν ένα είδος αλληλεπίδρασης μεταξύ των εννοιών ενός πεδίου με κάποια ιεραρχία.
- οι συναρτήσεις (*functions*), που εκπροσωπούν μια ειδική σχέση, στην οποία το ν-οστό στοιχείο της σχέσης προσδιορίζεται μοναδικά από τα ν-1 προηγούμενα στοιχεία.
- τα αξιώματα (*axioms*), τα οποία χρησιμοποιούνται για προτάσεις που είναι πάντοτε αληθείς.
- και τα στιγμιότυπα (*instances*), που εκφράζουν συγκεκριμένα στοιχεία.



Εικόνα 2: Τα Βασικά Συστατικά Μέρη μιας Οντολογίας

Πηγή: Σχεδιασμός Οντολογίας για τη Σημασιολογική Διερεύνηση των Μικρών και Μεσαίων Έξυπνων Πόλεων στον Μεσογειακό Χώρο,Παναγιωτοπούλου

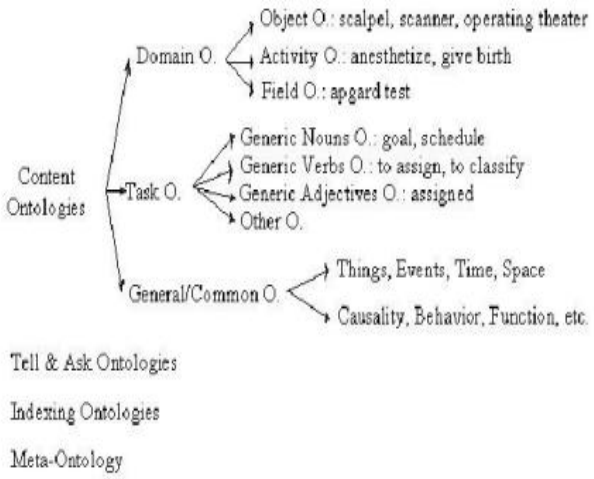
Με τις οντολογίες μπορούμε να επιτύχουμε την κοινή κατανόηση ενός πεδίου γνώσης με την χρήση μιας κοινής οντολογίας ή με την διασύνδεση μεταξύ οντολογιών, έτσι ώστε να καταφέρουμε να ξεπεράσουμε προβλήματα που οφείλονται στην χρήση διαφορετικής ορολογίας. Ο πιο διαδεδομένος, ίσως, ορισμός για την οντολογία είναι αυτός που διατυπώθηκε από τον Gruber το 1993, σύμφωνα με τον οποίο : «μια οντολογία είναι μια σαφής περιγραφή μιας εννοιολογικής αναπαράστασης»– («an ontology is an explicit specification of a conceptualization») (Gruber 1993). Ο όρος εννοιολογική αναπαράσταση αναφέρεται σε μια

αφηρημένη και απλοποιημένη εικόνα του κόσμου που κάποιος θέλει να μοντελοποιήσει, η οποία προκειμένου να αποτελεί οντολογία - σύμφωνα με τον παραπάνω ορισμό- πρέπει να αναπαρασταθεί σε μορφή κατανοητή και επεξεργάσιμη από την υπολογιστική μηχανή. Ο όρος ρητή σημαίνει ότι οι έννοιες που χρησιμοποιούνται, καθώς επίσης και οι περιορισμοί που αφορούν στη χρήση τους, έχουν προσδιοριστεί με σαφήνεια. Αργότερα, το 1997, ο Borst τροποποίησε ελαφρώς τον ορισμό του Gruber διατυπώνοντάς τον ως εξής: ("... μια οντολογία είναι μια τυπική περιγραφή μιας κοινής εννοιολογικής αναπαράστασης) – (... an ontology is a formal specification of a shared conceptualization") (Borst 1997). Στη συγκεκριμένη περίπτωση, ο όρος κοινή (shared) αναφέρεται στο ότι μια οντολογία πρέπει να αποτυπώνει γνώση κοινής αποδοχής στο πλαίσιο μιας κοινότητας, να είναι δηλαδή αποτέλεσμα ομοφωνίας παρά προσωπικής άποψης και να είναι επεξεργάσιμη από τους Η/Υ (Guarino). Το 1998, οι Studer και άλλοι συνδύασαν τους παραπάνω ορισμούς και διατύπωσαν ότι ("... μια οντολογία αποτελεί μια τυπική, σαφή περιγραφή μιας κοινής εννοιολογικής αναπαράστασης)- (.... an ontology is a formal, explicit specification of a shared conceptualization") (Studer και άλλοι 1998).

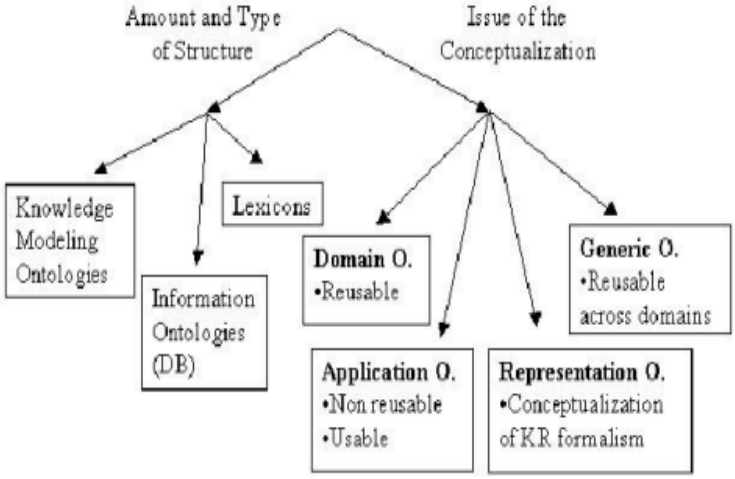
Οι οντολογίες μπορούν να ταξινομηθούν σύμφωνα με τον βαθμό λεπτομέρειας και το βαθμό εξάρτησης σε τρεις κατηγορίες :

- Οντολογίες ανώτερου επιπέδου (*top - level ontology*), που περιγράφουν πολύ γενικές έννοιες όπως χώρος, χρόνος, ύλη, κτλ. και δεν σχετίζονται με κάποιο συγκεκριμένο πρόβλημα ή πεδίο. Είναι χρήσιμες για μεγάλες κοινότητες με πολλούς χρήστες.
- Οντολογίες πεδίου και οντολογίες έργου (*domain & task ontologies*), για την περιγραφή των όρων του λεξιλογίου που σχετίζεται με ένα γενικό πεδίο (π.χ. ιατρική) ή με ένα γενικό έργο ή δραστηριότητα (π.χ. διάγνωση) και είναι πιο εξειδικευμένες.
- Οντολογίες εφαρμογής (*application ontology*), οι οποίες περιγράφουν έννοιες που εξαρτώνται από ένα ορισμένο πεδίο και από ένα έργο και συνήθως αποτελούν εξειδικεύσεις και των δύο σχετικών οντολογιών. Οι έννοιες αυτές αντιστοιχούν συνήθως στους ρόλους που έχουν οι οντότητες του πεδίου όταν πραγματοποιούν μία συγκεκριμένη δραστηριότητα (π.χ. θεραπεία).

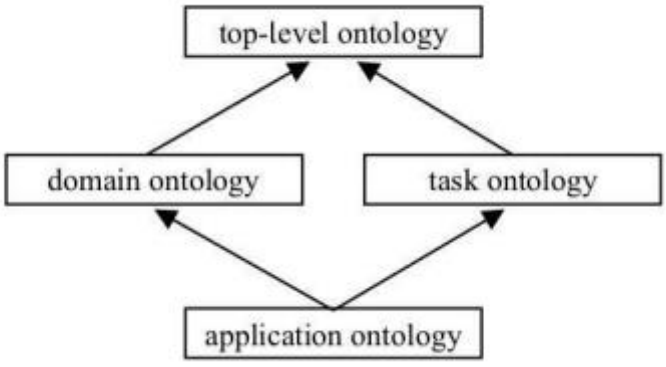
Επιπλέον, μπορούν να διακριθούν και να κατηγοριοποιηθούν κατά Mizoguchi, κατά Van Heijst, κατά Guarino και κατά Lassila και McGuinness. Συγκεκριμένα, οι κατηγοριοοποιήσεις σύμφωνα με τους παραπάνω επιστήμονες παρουσιάζονται στον παρακάτω πίνακα:



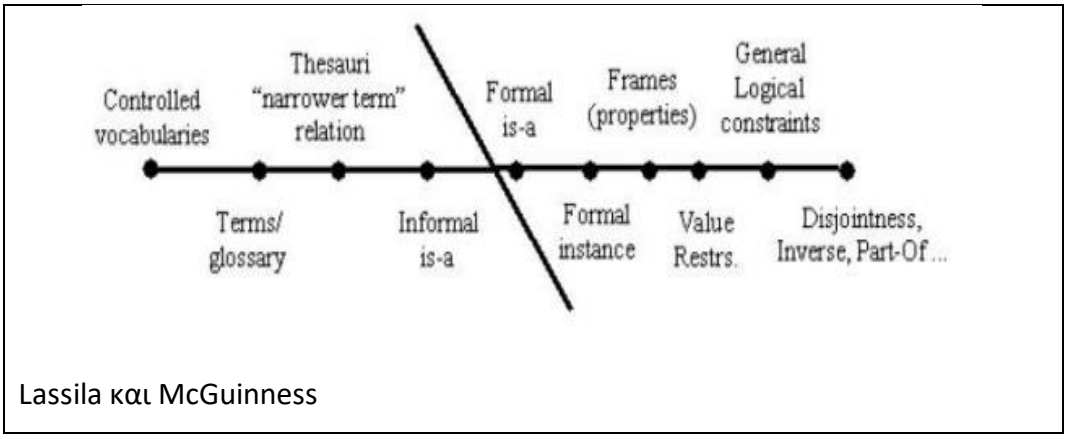
Mizoguchi



Van Heijst



Guarino



Εικόνα 3: Κατηγοριοποίησεις

Πηγή: Εφαρμογή τεχνολογιών του Σημασιολογικού Ιστού στη Διαχείριση Γνώσης στα πλαίσια του Ηλεκτρονικού Εμπορίου,Καπαντζάκης

2.1.3 Οντολογίες και Ιστός

Τα έγγραφα στον Παγκόσμιο Ιστό χαρακτηρίζονται κυρίως από τις λέξεις που εξάγουμε γι' αυτά και από κάποιο βαθμό σημαντικότητας που λαμβάνει υπόψη τη συνδεσμολογία όλου του. Η ομοιότητα ανάμεσα στα έγγραφα, ή ανάμεσα στις ερωτήσεις των χρηστών και τα έγγραφα, βασίζεται στην απόλυτη λεξική ομοιότητα των όρων που αντιστοιχούν σ' αυτά περιορίζοντας έτσι σημαντικά τις αναζητήσεις. Για παράδειγμα, ένα έγγραφο d1 που χαρακτηρίζεται από τις λέξεις: d1={φίδι, έρημος} δε θα θεωρηθεί ποτέ σχετικό με ένα έγγραφο d2 που χαρακτηρίζεται από τις λέξεις: d2={οχιά, Σαχάρα}. Είναι προφανές ότι οι δύο λίστες (και άρα και τα αντίστοιχα έγγραφα) σχετίζονται, καθώς μια "οχιά" είναι "φίδι" και η "Σαχάρα" είναι "έρημος" και συνεπώς, το έγγραφο d2 πραγματεύεται τις ίδιες έννοιες με το έγγραφο d1, απλά είναι πιο εξειδικευμένο. Με αντικατάσταση των λέξεων με έννοιες και μάλιστα σε μια ιεραρχία εννοιών (π.χ. σε μια οντολογία), έχουμε μια πιο ευέλικτη διαδικασία εύρεσης της ομοιότητας από το δυαδικό ταίριασμα (*binary matching*), που διαχειρίζεται τις γενικεύσεις και εξειδικεύσεις των εννοιών.

Οι μηχανισμοί επεξεργασίας των εγγράφων αναλαμβάνουν την εξαγωγή λέξεων από τα περιεχόμενα των εγγράφων, τους εισερχόμενους συνδέσμους προς αυτά κτλ. Η αντιστοίχιση των λέξεων, που εξάγονται από τους συνδέσμους, σε έννοιες μιας οντολογίας δεν αποσκοπεί να προσδιορίσει την αντικειμενική σημασία των συνδέσμων. Το κυριότερο όφελος έγκειται στο γεγονός ότι οι λέξεις που εξάγονται από τους συνδέσμους, και που δεν μας ενδιαφέρουν στο σύνολό τους, αντιστοιχίζονται σε ένα στενότερο σύνολο από έννοιες, οργανωμένες σε μια δομή που απεικονίζει τον τομέα ενδιαφέροντός μας (οντολογία). Οι ερωτήσεις των χρηστών με όμοιο τρόπο αντιστοιχίζονται σε έννοιες της οντολογίας. Με τον τρόπο αυτό μειώνονται οι διαστάσεις του προβλήματος εύρεσης εγγράφων που ικανοποιούν τις ερωτήσεις των χρηστών και επιπλέον η απαίτηση για ακριβή λεξική ομοιότητα μετατρέπεται σε προσεγγιστική

ομοιότητα εννοιών. Οι έλεγχοι σε επίπεδο συμβολοσειρών, που έπρεπε να συγκρίνουν εξαντλητικά τις λέξεις του ερωτήματος με τις λέξεις του κάθε εγγράφου, μετατρέπονται σε αναζητήσεις για την πλησιέστερη έννοια στο γράφο της οντολογίας.

Η αποσαφήνιση της έννοιας των λέξεων (*word sense disambiguation*) που εξάγονται από ένα έγγραφο είναι ένα πολύ σημαντικό και δύσκολο πρόβλημα και οι αλγόριθμοι που ασχολούνται με την επίλυση τέτοιων προβλημάτων θα μπορούσαν να διαχωριστούν στις παρακάτω κατηγορίες :

- σε αυτούς που απαιτούν κάποιο σύνολο προ-χαρακτηρισμένων εγγράφων για προπαίδευση και
- σε αυτούς που βασίζονται σε λεξικογραφικές πηγές (π.χ. λεξικά, θησαυρούς κτλ.) και αλγορίθμους που επεξεργάζονται εξ αρχής κάποιο σύνολο εγγράφων ενός εξειδικευμένου τομέα.Τα έγγραφα αυτά χρησιμοποιούν συγκεκριμένη ορολογία και οι λέξεις που εμφανίζονται σε αυτά έχουν περιορισμένο σημασιολογικό εύρος.

Στην πρώτη κατηγορία, της επιβλεπόμενης αποσαφήνισης (*supervised disambiguation*), ανήκει η δουλειά των Brown κ.ά. που έχουν χρησιμοποιήσει τον αλγόριθμο Flip-Flop για να αποσαφηνίσουν τις έννοιες γαλλικών λέξεων. Στην περίπτωση αυτή αναλύονται τα έγγραφα μιας συλλογής ως προς τη δομή και τη σειρά εμφάνισης των λέξεων σε αυτά και εντοπίζονται τα χαρακτηριστικά εκείνα που καθορίζουν τη σημασία των αμφισβητούμενων λέξεων (*indicators*). Για παράδειγμα η λέξη ”per” ή η ύπαρξη ενός αριθμού πριν τη λέξη "cent" καθορίζουν τη σημασία που έχει η λέξη (αν είναι ποσοστό ή αριθμός).

Στην ίδια κατηγορία της επιβλεπόμενης αποσαφήνισης ανήκει και η προσπάθεια των Gale κ.ά να χρησιμοποιήσουν την προσέγγιση Bayes για να αποσαφηνίσουν έννοιες λέξεων. Αντί να προσδιορίσουν ένα συγκεκριμένο χαρακτηριστικό που οδηγεί στην αποσαφήνιση μιας λέξης, θεωρούν ότι πολλά χαρακτηριστικά επηρεάζουν τη σημασία μιας λέξης, καθένα σε διαφορετικό βαθμό, ανάλογα με τη συχνότητα εμφάνισής του στο περιβάλλον της λέξης.

Στα πλεονεκτήματα αυτών των αλγορίθμων συγκαταλέγονται τα μεγάλα ποσοστά επιτυχίας, που φτάνουν ακόμη και το 90% σε ορισμένες περιπτώσεις για την Bayes. Οι αλγόριθμοι αυτοί απαιτούν ένα σύνολο χαρακτηρισμένων εγγράφων (πληροφοριακό περιεχόμενο) για να λειτουργήσουν και επιπλέον δεν εκμεταλλεύονται τα ιεραρχικά δέντρα κατηγοριοποίησης εννοιών όπως αυτό του Wordnet.

Στην δεύτερη κατηγορία, ο αλγόριθμος που έχει προταθεί από τον Lesk αποσαφηνίζει τις έννοιες κάνοντας χρήση λεξικού. Εντοπίζει δηλαδή την πιθανότερη έννοια με τη βοήθεια του ερμηνευτικού ορισμού της σημασίας κάθε έννοιας, όπως αυτή παρέχεται από κάποιο λεξικό ή

άλλη πηγή. Βασίζεται στην απλή ιδέα ότι τα στοιχεία του ορισμού μιας λέξης είναι πιθανόν καλή ένδειξη της χρησιμοποιούμενης σημασίας της λέξης.

2.1.4 Σημασιολογικός Ιστός και συνδεδεμένα δεδομένα

Τα linked data (διασυνδεμένα δεδομένα) περιγράφουν μια μέθοδο έκδοσης δομημένων δεδομένων, τα οποία διασυνδέονται ώστε να γίνουν πιο χρήσιμα. Στηρίζονται στα πρότυπα τεχνολογιών του Web, όπως η HTTP και τα URIs, αλλά αντί να δίνουν τις ιστοσελίδες στον άνθρωπο, τις επεκτείνουν για να μοιραστεί η πληροφορία με τέτοιο τρόπο ώστε να διαβαστεί από την μηχανή. Αυτό επιτρέπει σε δεδομένα από διαφορετικές πηγές να συνδεθούν και να εξεταστούν.

Ο Lee περιγράφει τέσσερις βασικές αρχές των διασυνδεδεμένων δεδομένων (*Linked data*), με την προοπτική να κάνουν τον Ιστό να λειτουργήσει σωστά :

- Χρησιμοποίησε URIs για να προσδιορίσεις αντικείμενα και έννοιες.
- Χρησιμοποίησε HTTP URIs για να αναζητηθούν οι προσδιοριστές, δηλαδή να επιτρέπεις στα άτομα να λάβουν μια περιγραφή του αντικειμένου που τυποποιείται από το URI.
- Όταν αναζητείται κάποιο URI, να παρέχονται πληροφορίες μέσω σχετικών προτύπων (RDF/ XML), δηλαδή να οδηγείσαι σε περισσότερα χρήσιμα δεδομένα.
- Χρησιμοποίησε συνδέσμους προς άλλα URIs για να υπάρχει πρόσβαση σε επιπλέον πληροφορία.

2.1.5 Τα συστατικά του Σημασιολογικού Ιστού

Ο Σημασιολογικός Ιστός στηρίζεται από τις ακόλουθες γλώσσες και πρότυπα:

- XML (Extensible Markup Language): Είναι μια γλώσσα περιγραφής δεδομένων τα οποία είναι εύκολο να διαβαστούν και να επεξεργαστούν από ανθρώπους και προγράμματα. Δεν επιβάλλει κανέναν σημασιολογικό περιορισμό στα δεδομένα που περιγράφει.
- XML Schema: Είναι μια γλώσσα η οποία περιορίζει τη δομή των εγγράφων XML. • RDF: Είναι ένα μοντέλο περιγραφής και επεξεργασίας μεταδεδομένων.
- RDF Schema: Είναι ένας μηχανισμός περιγραφής πόρων και των σχέσεων ανάμεσα τους και αποτελεί σημασιολογική επέκταση του RDF.
- OWL: Παρέχει έναν τρόπο περιγραφής όρων και σχέσεων γύρω από ένα πεδίο ενδιαφέροντος, προσφέροντας πιο ισχυρό συντακτικό από τις RDF και RDF Schema, καθώς και πιο ισχυρή σημασιολογία που βασίζεται στη λογική (logic–based semantics).

Τα δεδομένα που ακολουθούν το **RDF**, καθώς και το RDFS, είναι εκφρασμένα ως τριπλέτες ή δηλώσεις (statements), οι οποίες έχουν τη μορφή υποκείμενο - κατηγόρημα - αντικείμενο (object – predicates - subject).

Κάθε δομικό στοιχείο μίας τριπλέτας αναγνωρίζεται μοναδικά από ένα Ενιαίο Αναγνωριστικό Πόρων (Uniform Resource Identifier; URI), γεγονός το οποίο διευκολύνει την ενσωμάτωση κατανεμημένων δεδομένων αποφεύγοντας τον κίνδυνο συγκρούσεων. Θα πρέπει να σημειωθεί ότι τέτοιου είδους μορφές δεδομένων φιλοξενούνται σε αποθετήρια τριπλετών (triplestore). Πρόσβαση στα οντολογικά δεδομένων των αποθετηρίων τριπλετών δίνεται μέσω της SPARQL, της γλώσσας ερωτημάτων για οντολογικά δεδομένα.

Η **OWL** αποτελεί μία γλώσσα αναπαράστασης γνώσης για το Σημασιολογικό Ιστό. Λόγω των περιορισμών στη λειτουργικότητα του RDFS, ο οργανισμός W3C εισήγαγε την πρώτη έκδοση του προτύπου της OWL το 2004, ενώ το 2012 προτυποποιήθηκε η δεύτερη έκδοση της OWL (OWL 2), η οποία είναι και η τρέχουσα.

Η γλώσσα OWL δεν περιορίζεται απλά στην αναπαράσταση εννοιών και των συνδέσεών τους, αλλά στοχεύει στην ανάπτυξη οντολογιών, οι οποίες μπορεί να αξιοποιηθούν στα πλαίσια διαδικασιών λήψης αποφάσεων. Ταυτόχρονα, επωφελείται των δυνατοτήτων του RDF ώστε να επιτρέπει στις οντολογίες να περιέχουν αναφορές σε όρους άλλων οντολογιών OWL, δημιουργώντας τα Ανοικτά Συνδεδεμένα Δεδομένα (Linked Open Data; LOD).

Ο Tim Berners-Lee όρισε την προϋποθέσεις που πρέπει να πληρούν τα δεδομένα, για να μπορούν να χαρακτηριστούν ως LOD, ως εξής :

1. Χρήση URIs για τη μοναδική αναγνώριση των πραγμάτων.
2. Χρήση HTTP URIs ώστε να μπορούν να ανακαλυφθούν.

Συνεπώς, η OWL χρησιμοποιείται για την ανάπτυξη οντολογιών, δηλαδή για τον ορισμό εννοιών και των συσχετίσεων που υπάρχουν μεταξύ τους. Διαθέτει τη δυνατότητα δημιουργίας κλάσεων, ιδιοτήτων και στιγμιότυπων κλάσεων. Οι προαναφερθείσες ιδιότητες χωρίζονται σε ιδιότητες αντικειμένων (*object property*), οι οποίες συνδέουν στιγμιότυπα κλάσεων, και σε ιδιότητες δεδομένων (*data property*), οι οποίες αναθέτουν κάποια τιμή σε ένα στιγμιότυπο.

Σε αντίθεση με το RDFS, το εύρος των δυνατοτήτων της γλώσσας OWL επεκτείνεται περαιτέρω, προσφέροντας λειτουργίες έκφρασης τομής και ένωσης κλάσεων και περιορισμών τιμών ή πληθικότητας. Συγκεκριμένα, η OWL 2 προσφέρεται για εφαρμογές που χρησιμοποιούν οντολογίες με μεγάλο αριθμό κλάσεων ή/και ιδιοτήτων. Οι γλώσσες οντολογιών ιστού και γενικά οι τεχνολογίες σημασιολογικού ιστού χρησιμοποιούνται ευρέως και για την αναπαράσταση δεδομένων και οντοτήτων του IoT.

2.1.6 Παραδείγματα Εφαρμογών του Σημασιολογικού Ιστού

Η χρήση τεχνολογιών Σημασιολογικού Ιστού χρησιμοποιείται ευρέως από εταιρείες παγκόσμιας εμβέλειας και είναι διαρκώς αυξανόμενη. Εταιρείες όπως η Adobe, η Oracle, η Microsoft και η Google προσφέρουν εργαλεία Σημασιολογικού Ιστού ή συστήματα που βασίζονται σε αντίστοιχες τεχνολογίες και είναι ενεργά μέλη του W3C, βοηθώντας στην ανάπτυξη και επέκταση του Σημασιολογικού Ιστού. Στη συνέχεια, θα περιγράψουμε σύντομα κάποιες εφαρμογές των τεχνολογιών Σημασιολογικού Ιστού.

Swoogle

Το Swoogle είναι ένα σύστημα ανάκτησης και δημιουργίας ευρετηρίου εγγράφων RDF ή OWL, το οποίο έχει αναπτυχθεί από το Πανεπιστήμιο Maryland Baltimore Country. Το συγκεκριμένο σύστημα χωριζεται σε τέσσερα βασικά συστατικά: την εύρεση των εγγράφων του Σημασιολογικού Ιστού, τη δημιουργία μεταδεδομένων, την ανάλυση δεδομένων και τη διεπαφή, τα οποία εργάζονται ανεξάρτητα και αλληλεπιδρούν μεταξύ τους μέσω μιας βάσης δεδομένων.

2.1.7 Friend of a Friend (FOAF) project

Το εγχείρημα FOAF (http://www.foaf-project.org/), δηλαδή φίλος – ενός – φίλου, επιτρέπει στους ανθρώπους να δημοσιεύουν απλές βιογραφικές πληροφορίες για τον εαυτό τους και τους φίλους τους σε προσωπικές ιστοσελίδες. Η υλοποίησή του FOAF έχει γίνει με την χρήση RDF και OWL, έτσι ώστε τα δεδομένα να είναι αναγνώσιμα και από πράκτορες λογισμικού. Το έργο ξεκίνησε το 2000 από τους Libby Miller και Dan Brickley, αποτελώντας για πολλούς την πρώτη εφαρμογή Σημασιολογικού Ιστού, για αυτό και κρίθηκε σκόπιμο να αναφερθεί στην παρούσα μελέτη.

2.1.8 BBC – Διαχείριση ΜΜΕ

Το BBC (British Broadcasting Corporation) είναι η μεγαλύτερη επιχείρηση ραδιοφωνίας και τηλεόρασης στον κόσμο, αποτελεί δημόσιο ΜΜΕ και χρηματοδοτείται από τους κατοίκους της Μεγάλης Βρετανίας. Το περιεχόμενο αυτό διαμοιράζεται σε διακριτές HTML σελίδες, οι οποίες δεν είναι σωστά διασυνδεδεμένες, με αποτέλεσμα να μην είναι χρηστικές, αν κάποιος τηλεθεατής έχει ενδιαφέροντα τα οποία εκτείνονται σε πολλές ξεχωριστές σελίδες (*domains*) και περιορίζεται η αλληλεπίδραση του χρήστη με την πληροφορία.

Η διαδικασία αποθήκευσης, διαλογής και παρουσίασης των μέσων ενημέρωσης μπορεί να επωφεληθεί σε πολλούς τομείς από τη χρήση των τεχνολογιών του Σημασιολογικού Ιστού. Οι ιστοσελίδες πρέπει να μεταβάλλονται συνεχώς για να παραμένουν ενδιαφέρουσες, τόσο όσον αφορά το περιεχόμενο, αλλά και όσον αφορά τη σχεδίασή τους και πρέπει να υπάρχουν

συνδέσεις μεταξύ σελίδων, βίντεο, πολυμέσων blogs, κ.λπ., οι οποίες να αλλάζουν με την πάροδο του χρόνου.

Με την χρήση αναγνωριστικών, με σελίδες HTML και ροές πληροφοριών (*feeds*), αναγνώσιμων από μηχανές με βάση τεχνολογίες όπως RDF/XML, JSON, απλή XML, έγινε δυνατή η διασύνδεση του περιεχομένου. Το περιεχόμενο του BBC μπορεί να παρουσιαστεί με πολλούς διαφορετικούς τρόπους και οι ομάδες οργάνωσης του, έχουν ένα κομβικό σημείο γύρω από το οποίο μπορούν να το διαμορφώσουν και να το επεκτείνουν.

Μία μηχανή αναζήτησης είναι μια εφαρμογή που εξυπηρετεί στην εύρεση πληροφοριών στο Διαδίκτυο. Διαθέτει μία Βάση Δεδομένων, στην οποία αποθηκεύονται τεράστιες συλλογές κειμένων και αρχείων διαφόρων ειδών, που έχουν δημοσιευθεί και διατίθενται στον ιστό. Η αναζήτηση γίνεται με τη χρήση κλειδιών, ενώ τα αποτελέσματα που επιστρέφονται αποτελούνται από μια λίστα διευθύνσεων ιστοσελίδων με τις λέξεις του ζητήματος ταξινομημένη με βάση την πολιτική της εκάστοτε μηχανής αναζήτησης.

Ο μηχανισμός ανάκτησης πληροφορίας μέσω μιας μηχανής αναζήτησης επιτυγχάνεται μέσω ενός πράκτορα λογισμικού, του ανιχνευτή διαδικτύου (Web Crawler ή Web Spider ή Web Robot). Ο ανιχνευτής μετακινείται στο διαδίκτυο και επισκέπτεται μία λίστα URL's που του δίνεται από έναν URL Server. Σκοπός είναι να δημιουργήσει αντίγραφα όλων των ιστοσελίδων διαδικτύου, καθώς και ένα αρχείο με όλους τους υπερσυνδέσμους που βρίσκονται σε αυτές. Μετά το πέρας της διαδικασίας συλλογής (διαδικασία Web Crawling) γίνεται η εξαγωγή των λέξεων όλων των ιστοσελίδων μέσω του προγράμματος indexer και ολοκληρώνεται η καταγραφή των URL's. Στη συνέχεια δημιουργείται ένα ευρετήριο (index) με όλους τους όρους του διαδικτύου. Οι σελίδες που ανακτώνται διατηρούνται προσωρινά σε ένα αποθετήριο ιστοσελίδων, ενώ πραγματοποιείτε και μία διατήρηση των σελίδων που χρησιμοποιήθηκαν πρόσφατα στην cache. Κατόπιν ολοκληρώνεται το ταίριασμα λέξεων -κλειδιών του ευρετηρίου και τα αποτελέσματα κατατάσονται και παραδίδονται από ένα πρόγραμμα ταξινόμησης.

Το μοντέλο αυτό είναι καθολικό, όμως θα πρέπει να επισημάνουμε ότι κάθε μηχανή αναζήτησης διαθέτει διαφορετικούς αλγορίθμους και μηχανισμούς ανάκτησης πληροφορίας που βασίζονται όμως σε αυτό το γενικό μοντέλο αρχιτεκτονικής.

2.1.9 Αρχιτεκτονική Συστημάτων Ανάκτησης

Τα Συστήματα Ανάκτησης πληροφοριών είναι συστήματα διαχείρισης πληροφοριών (*βάσεων δεδομένων*), τα στοιχεία των οποίων έχουν με μορφή κειμένου (*εγγράφου*) φυσικής γλώσσας (*Free Text Data Bases*).

Μια τέτοια βάση αποτελείται από ένα σύνολο τμημάτων από κείμενα (*documents, segments*). Ένα τμήμα μπορεί να είναι ένα αρχείο κειμένου, μια παράγραφος, μια σελίδα, ή οποιαδήποτε

άλλη ενότητα κειμένου. Κάθε τμήμα αποτελείται από όρους (*terms*) που είναι τα βασικά στοιχεία πληροφορίας τα οποία μπορεί να χειριστεί ένα σύστημα ανάκτησης. Η ανάκτηση των αποθηκευμένων τμημάτων επιτυγχάνεται μέσω ερωτήσεων (*queries*) που εισάγονται σε φυσική γλώσσα από το χρήστη.

Οι ερωτήσεις αυτές αναλύονται σε απλούς όρους (λέξεις), οι οποίοι χρησιμοποιούνται για να προσπελάσουν την αποθηκευμένη πληροφορία. Ως εδώ, τα συστήματα ανάκτησης δεν φαίνεται να παρουσιάζουν σημαντικές διαφορές (εκ πρώτης όψεως τουλάχιστον) σε σχέση με τα παραδοσιακά συστήματα διαχείρισης βάσεων δεδομένων (DBMS).

Ωστόσο έχουν αρκετά διαφορετική προσέγγιση με την οποία θα ασχοληθούμε στα επόμενα. Για παράδειγμα μια βάση δεδομένων είναι αυστηρά ορισμένη, δέχεται ερωτήματα σε μια αυστηρή γλώσσα (SQL) και οι απαντήσεις είναι εξ ίσου αυστηρές δηλαδή πληρούν ή δεν πληρούν το ερώτημα. Σε αντίθεση σε ένα σύστημα ανάκτησης τα κείμενα δεν είναι αυστηρά ορισμένα, οι ερωτήσεις είναι σε φυσική γλώσσα και κατά συνέπεια και οι απαντήσεις δεν είναι αυστηρές, για παράδειγμα κάποιες απαντήσεις δεν είναι σχετικές με το ερώτημα.

Αυτό που οδήγησε στην μελέτη γύρω από τα συστήματα ανάκτησης, είναι η ύπαρξη τεράστιων ποσοτήτων δεδομένων με μορφή κειμένου (*textual data*) σε οργανισμούς, υπηρεσίες, βιβλιοθήκες, πανεπιστήμια. Όλοι συμφωνούν ότι τέτοιου είδους δεδομένα, απαιτούν διαφορετική αντιμετώπιση και η απάντηση στο πρόβλημα φιλοδοξούν να είναι τα συστήματα ανάκτησης πληροφοριών τα οποία ενσωματώνουν χαρακτηριστικά που έχουν ως στόχο την αποτελεσματικότερη επεξεργασία των κειμένων. Σε γενικές γραμμές ένα σύστημα ανάκτησης και επεξεργασίας κειμένων αποτελείται από τον πυρήνα του συστήματος, ο οποίος αποτελείται από δύο ενότητες :

- Την βάση δεδομένων και
- Έναν μηχανισμό αναζήτησης.

Εκτός του πυρήνα υπάρχουν μια σειρά από ενότητες που υποστηρίζουν την λειτουργία του πυρήνα. Η βάση δεδομένων αποτελείται από τα τμήματα κειμένου τα οποία αποθηκεύονται, και μια σειρά από δομές δεδομένων που επιταχύνουν την πρόσβαση και επεξεργασία των κειμένων. Η διαδικασία αναζήτησης στηρίζεται σε όλες τις πληροφορίες (*Full Text Retrieval*) που αντλούνται αυτόματα από το κείμενο κατά την δημιουργία των ευρετηρίων της βάσης δεδομένων (εκτός από ορισμένες λέξεις που αποδεδειγμένα δεν περιέχουν χρήσιμη πληροφορία) καθώς και σε πληροφορίες που είναι δυνατό να μην υπάρχουν στην βάση, αλλά να είναι σχετικές με άλλες ήδη αποθηκευμένες.

Αυτό αποτελεί ένα από τα πιο ισχυρά χαρακτηριστικά των συστημάτων ανάκτησης κειμένων και βασίζεται τόσο στα μαθηματικά μοντέλα που χρησιμοποιούνται για το σκοπό αυτό (και τα

οποία θα αναλύσουμε παρακάτω), όσο και στην λεξική ανάλυση (*lexical analysis*) και επεξεργασία του κειμένου, που γίνεται κατά την δημιουργία της βάσης δεδομένων. Όταν κάνουμε ερωτήσεις σε μια βάση δεδομένων που αποτελείται από τμήματα κειμένου, το σύστημα ανάκτησης κειμένων, ανταποκρίνεται εντοπίζοντας και αξιολογώντας τα τμήματα της βάσης που σχετίζονται εννοιολογικά με την ερώτηση. Για να μπορεί να γίνει κάτι τέτοιο γίνονται δύο βασικές παραδοχές:

Θεωρούμε ότι, αν ένα τμήμα κειμένου είναι σχετικό με μια έννοια, τότε θα περιέχει λέξεις ή τμήματα λέξεων σχετικά με την έννοια αυτή.

Θεωρούμε ότι, λεξική συγγένεια συνεπάγεται και εννοιολογική συγγένεια.

Παρά το γεγονός ότι είναι πιθανό να υπάρχουν περιπτώσεις, στις οποίες να μην ισχύουν απόλυτα τα παραπάνω, είναι αποδεκτό ότι στις περισσότερες περιπτώσεις, οι λέξεις ενός τμήματος κειμένου μπορούν να προσδιορίζουν σε μεγάλο βαθμό τα περιεχόμενά του.

Έτσι για παράδειγμα, σε μια βάση που αποτελείται από επιστημονικά άρθρα σχετικά με την πληροφορική, αν η ερώτηση είναι «συστήματα στήριξης αποφάσεων», το αποτέλεσμα θα είναι μια λίστα των άρθρων που σχετίζονται με τα συστήματα στήριξης αποφάσεων, σε φθίνουσα σειρά σπουδαιότητας. Η σπουδαιότητα ενός άρθρου, - στην προκειμένη περίπτωση - υπολογίζεται από το σύστημα και είναι η απάντησή του στην ερώτηση πόσο σχετικό (relevant) είναι, κάθε ένα από τα άρθρα που θεωρήθηκαν ως σχετικά με την ερώτηση του χρήστη. Η διαδικασία της διαβάθμισης ή ταξινόμησης των κειμένων σύμφωνα με το κατά πόσο είναι σχετικά ή συναφή με το αντίστοιχο ερώτημα είναι η βασική λειτουργία του μηχανισμού αναζήτησης. Οι τεχνικές και η θεωρητικές αρχές πίσω από αυτήν την διαδικασία, θα παρουσιαστούν στο επόμενο κεφάλαιο. Τα βασικά πλεονεκτήματα των συστημάτων ανάκτησης, έναντι άλλων γνωστών προσεγγίσεων στο πρόβλημα της αναζήτησης και ανάκτησης πληροφοριών, συνοψίζονται σε τρία κυρίως σημεία.

Σε πληροφορίες που αποθηκεύονται υπό την μορφή κειμένου σε φυσική γλώσσα, δεν μπορούν να επιτύχουν ικανοποιητικό χειρισμό τα παραδοσιακά συστήματα DBMS, κυρίως λόγω του προσανατολισμού τους σε δομημένα δεδομένα.

Οι πληροφορίες που αποθηκεύονται στη βάση δεδομένων ενός συστήματος ανάκτησης, είναι το μοναδικό κριτήριο που χρησιμοποιείται για την αναζήτηση. Αντίθετα στις κλασικές προσεγγίσεις των βάσεων δεδομένων, ένα αντικείμενο του πραγματικού κόσμου (π.χ. ένα βιβλίο) αναπαρίσταται από μια δομή δεδομένων (π.χ. μια εγγραφή με τα στοιχεία του βιβλίου), η οποία στην συνέχεια αποτελεί την παρουσία του αντικειμένου και με βάση αυτήν γίνεται η όποια αναφορά στο αντικείμενο (π.χ. δεν μπορεί κανείς να βρει το βιβλίο με βάση ένα θεματικό όρο που δεν είχε κριθεί εκ των προτέρων σκόπιμο να καταχωρηθεί).

Ο πυρήνας ενός τέτοιου συστήματος είναι εξαιρετικά μικρός, με αποτέλεσμα να μπορεί να ενσωματωθεί σε ένα άλλο, μεγαλύτερο σύστημα με χαρακτηριστική ευκολία.

Ένα βασικό μειονέκτημα ωστόσο των συστημάτων ανάκτησης πληροφορίας είναι ότι η επίδοσή τους δεν βρίσκεται ακόμη στο επιθυμητό επίπεδο. Πολλές φορές η απάντηση από μια ερώτηση δεν είναι σχετική με την ερώτηση. Έτσι, στην ερώτηση «συστήματα στήριξης αποφάσεων» μπορεί η λίστα των άρθρων που θα πάρουμε (και τα οποία υποτίθεται ότι είναι σχετικά με την ερώτηση), να περιέχει άρθρα τα οποία είναι (πολύ ή λίγο) άσχετα με αυτό που εμείς επιθυμούσαμε.

Στην χειρότερη περίπτωση, είναι δυνατό, τα άσχετα άρθρα να καταλαμβάνουν πολύ καλή θέση στην κατάταξη σχετικότητας, υποσκελίζοντας άρθρα που είναι σχετικότερα. Πρέπει ωστόσο να τονίσουμε, ότι και το ερώτημα που υποβάλλεται σε ένα σύστημα ανάκτησης είναι πιο «χαλαρό» και όχι τόσο αυστηρό όσο ένα SQL ερώτημα γι' αυτό και τα αποτελέσματα δεν είναι τα αναμενόμενα. Πολλές προσπάθειες γίνονται για την βελτίωση αυτού του προβλήματος, και μια σειρά λύσεων έχει προταθεί, όπως για παράδειγμα η μέθοδος βελτίωσης της ερώτησης με αυτόματη εισαγωγή σχετικών όρων (relevance feedback).

Έχοντας δώσει μια εικόνα του τι είναι τα συστήματα ανάκτησης, μπορούμε να περάσουμε στο θεωρητικό υπόβαθρο που σχετίζεται με αυτά, καθώς και την περιγραφή των μεθοδολογιών για την ανάπτυξή τους, που ακολουθεί στη συνέχεια. Εδώ θα περιγράψουμε τη βασική δομή, τα χαρακτηριστικά και τη τεχνολογία πάνω στην οποία στηρίζεται ένα τυπικό σύστημα ανάκτησης κειμένων. Παρά το γεγονός ότι δεν υπάρχει ακόμη κάποιου είδους πρότυπο για την δημιουργία τέτοιων συστημάτων, υπάρχουν απόψεις και προσεγγίσεις που τείνουν να καθιερωθούν.

3.1 Οπτικοποιήσεις

Η οπτικοποίηση πληροφορίας δεν είναι σημερινή ανακάλυψη και έχει τις ρίζες της σε παλαιότερες ιστορικές περιόδους. Χαρακτηριστικά παραδείγματα είναι οι αστρονομικοί χάρτες, η χαρτογράφηση περιοχών, η αναπαράσταση στατιστικών δεδομένων, καθώς επίσης και οι επινοήσεις γραφικών παραστάσεων και η πύκνωση της χρήσης, παραγωγής και αναπαραγωγής τους που κάνουν την εμφάνισή τους κατά τον 18ο αιώνα. Από τα τέλη του 18ου αιώνα μέχρι τις αρχές του 19ου αιώνα κάνουν την εμφάνισή τους απεικονίσεις δεδομένων με τη χρήση διανυσματικών γραφικών, ενώ τον 19ο αιώνα δε, λόγω της ανάγκης διαχείρισης μεγάλου όγκου αριθμητικών δεδομένων, η οποία καθιστά αναγκαία την απεικόνιση των δεδομένων αυτών, διεισδύει σε τομείς όπως η πολεοδομία, η αστική ανάπτυξη, το εμπόριο, οι μεταφορές, η βιομηχανία και η επιδημιολογία.

Σήμερα, η χρήση οπτικοποιήσεων είναι αναπόσπαστο κομμάτι του Data Science. Πολλές ιστοσελίδες και γενικότερα εφαρμογές χρησιμοποιούν το συγκεκριμένο μέσο, προκειμένου να παρουσιάζουν τα δεδομένα τους στους χρήστες με έναν πιο ευχάριστο και κατανοητό τρόπο. Στη συνέχεια του κεφαλαίου εξετάζονται ορισμένες οπτικοποιήσεις Crawl που αξίζει να αναφερθούν διότι έχουν άμεση σχέση με το θέμα και τα δεδομένα που πραγματεύεται η πτυχιακή εργασία.

3.1.1 Ο WebSphinx crawler

Ο WebSphinx (*Website-Specific Processors for HTML Information eXtraction*) είναι μια βιβλιοθήκη σε JAVA και ένα διαδραστικό περιβάλλον για web crawlers. Ο WebSphinx αποτελείται από δύο μέρη: τον Crawler Workbench και την WebSphinx class library.

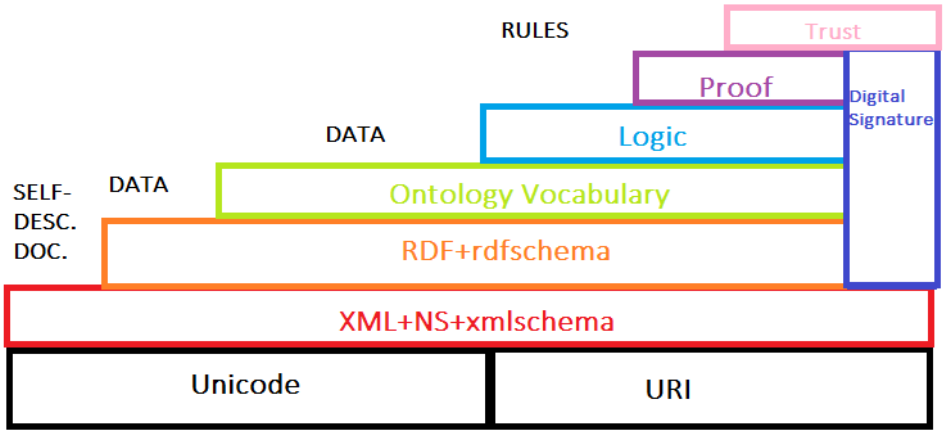
3.1.2 Crawler Workbench

Η Crawler Workbench αποτελεί μια γραφική διεπαφή χρήστη (GUI) η οποία επιτρέπει να προσαρμόσουμε και να ελέγξουμε έναν προσαρμόσιμο web crawler. Παρέχει τις παρακάτω δυνατότητες:

Οπτικοποίηση μιας συλλογής από σελίδες σε έναν γράφο.
Αποθήκευση σελίδων στον τοπικό δίσκο για offline περιήγηση και επεξεργασία.
Συνένωση σελίδων για προβολή και εκτύπωση σαν ένα ενιαίο έγγραφο.
Εξαγωγή όλου του κειμένου το οποίο ακολουθεί κάποιο πρότυπο για μια συλλογή σελίδων.

Ανάπτυξη ενός crawler σε javaή Javascript ο οποίος μπορεί να επεξεργαστεί τις σελίδες όπως επιθυμούμε

3.1. Πίνακας 4: Οι δυνατότητες της διεπαφής crawler.



Εικόνα 4: Απεικόνιση των επιπέδων του Σημασιολογικού Ιστού, σε σχέση με τις εμπλεκόμενες τεχνολογίες.

Πηγή: http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html

---

Ένας Web Crawler ή αλλιώς Spider ή Spider Bot είναι ένα Internet Bot που συστηματικά περιηγείται στον Παγκόσμιο Ιστό, συνήθως για λόγους εύρεσης δεδομένων (*Web Indexing*). Ο τρόπος συλλογής των δεδομένων από τον Crawler είναι αυτόματος, ακολουθεί τη σύνδεση που υπάρχει σε μια ιστοσελίδα και συλλέγει τα δεδομένα που υπάρχουν αποθηκεύοντας αυτά, έως ότου ολοκληρώσει τη συλλογή των δεδομένων ή φτάσει σε αδιέξοδο. Στην τελευταία περίπτωση, συνεχίζει με μια νέα περιήγηση.

Οι Crawlers χρησιμοποιούνται κυρίως από τις διάφορες μηχανές αναζήτησης, προκειμένου να συλλέξουν στοιχεία για τη βάση δεδομένων τους, ώστε να εμφανίζουν ταχύτερα και πιο σχετικά αποτελέσματα αναζήτησης. Στην πραγματικότητα, όταν ένας χρήστης εισάγει κάποιον όρο αναζήτησης οι μηχανές αυτές δεν αναζητούν σε ολόκληρο το διαδίκτυο, αλλά εξετάζουν την βάση δεδομένων των ιστοσελίδων που έχουν συλλέξει οι Crawlers τους.

Ακόμη, υπάρχουν και μηχανισμοί για τις ιστοσελίδες που δεν επιθυμούν να ανιχνευθούν από κάποιο Crawler. Για παράδειγμα, μια ιστοσελίδα μπορεί να ζητήσει από τον Crawler να

Η ευχερέστερη αναζήτηση και επεξεργασία πληροφοριών με χωρική διάσταση.
Ο συνδυασμός δυνατοτήτων Συσχετισμένων Δεδομένων και Συστημάτων GIS σε ένα γεωγραφικά εμπλουτισμένο Linked Data Web.

προσθέσει στο ευρετήριο του μόνο τμήματα αυτής ή ακόμη και κανένα τμήμα της.

3.2 Σημασιολογία Γεωγραφικών Εννοιών

Ο Γεωχωρικός Σημασιολογικός Ιστός είναι ένα όραμα να συμπεριληφθούν οι γεωχωρικές

πληροφορίες στον πυρήνα του Σημασιολογικού Ιστού. Αποτελεί μέρος της γεωγραφικής επιστήμης των πληροφοριών. Τα κίνητρα της γεωχωρικής έρευνας στον τομέα του

Πίνακας 5: Τα κίνητρα της γεωχωρικής έρευνας.

Η σημασιολογία των γεωγραφικών εννοιών (*geospatial semantics*) αποτελεί τον επιστημονικό τομέα που βρίσκεται στο "σημείο συνάντησης" τριών άλλων επιστημών: 1)της επιστήμης των Συστημάτων Γεωγραφικών Πληροφοριών (ΣΓΠ), 2)της επιστήμης των υπολογιστών και της μηχανικής της γνώσης (Ballatore 2016). Μελετά τον τρόπο δημοσιοποίησης, ανάκτησης, επαναχρησιμοποίησης και ολοκλήρωσης των χωρικών δεδομένων, το πως περιγράφεται η γεωχωρική πληροφορία μέσα από εννοιολογικά μοντέλα, καθώς και το πως αναπτύσσονται τυπικές προδιαγραφές δομών δεδομένων με στόχο τη μείωση των περιπτώσεων ασυμβατότητας (Janowicz και άλλοι 2013). Η σημασιολογία των γεωγραφικών οντοτήτων αποτελεί μια ιδιαίτερη και ενδιαφέρουσα περίπτωση καθώς οι οντότητες αυτές παρουσιάζουν συγκεκριμένες ιδιαιτερότητες, όπως (Kavouras & Kokla 2008) :

Πολυπλοκότητα και δυσκολία στην ανάλυση και τυποποίησή τους, σε αντίθεση με οντότητες της καθημερινής εμπειρίας (οργανισμοί και τεχνητά αντικείμενα).
Δεν έχουν πάντα καθορισμένα - σαφή όρια (Laurini 2012, Laurini & Kazar 2016). Εκτός από ασαφή, τα όρια μπορεί να είναι εποχιακά (seasonal) ή βαθμωτά/κλιμακωτά (gradual) (Ballatore 2016)
Δεν ανήκουν πάντα σε σαφείς κατηγορίες. Πολλές φορές προκύπτει το ερώτημα για το αν υπάρχουν χαρακτηριστικές ιδιότητες των οντοτήτων που να τις ορίζουν

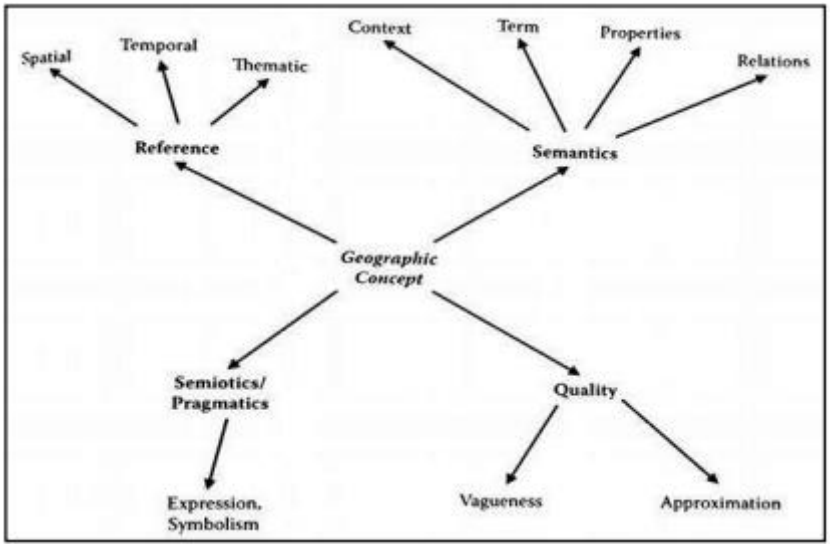
σαφώς και με μοναδικό τρόπο. Χαρακτηριστικό παράδειγμα αποτελεί η σύγχυση που μπορεί να παρατηρηθεί ανάμεσα στις έννοιες του λόφου και του βουνού.

Πίνακας 6 : Ιδιαιτερότητες των γεωγραφικών οντοτήτων

3.3 Διαστάσεις γεωχωρικών εννοιών

Η όποια αναφορά στη γεωγραφική πραγματικότητα περιλαμβάνει έννοιες που περιγράφουν οντότητες του φυσικού κόσμου (πχ, βουνό κ.λπ.), καθώς επίσης και έννοιες που σχετίζονται με κατασκευασμένες/τεχνητές γεωγραφικές οντότητες (γέφυρα, σιδηροδρομικό δίκτυο, κ.λπ.). Σε αυτό το σημείο είναι χρήσιμο να αναλυθεί το πλαίσιο των κύριων διαστάσεων που χρησιμοποιούνται για τη διαχείριση των γεωγραφικών εννοιών. Οι διαστάσεις αυτές μπορούν να οργανωθούν σε τέσσερις κύριες κατηγορίες οι οποίες είναι (Kavouras & Kokla 2008):

1. το πλαίσιο αναφοράς,
2. η σημασιολογία,
3. η σημειολογία
4. και η ποιότητα.



Εικόνα 5: Διαστάσεις Γεωχωρικών Εννοιών

Πηγή: Kavouras & Kokla 2008

Ειδικότερα :

Όσον αφορά το πλαίσιο αναφοράς (reference):

Χωρικό (spatial frame): σύστημα αναφοράς για τον προσδιορισμό της θέσης των γεωχωρικών οντοτήτων, ιδιοτήτων και σχέσεων.
Χρονικό (temporal frame): χρονικό σύστημα αναφοράς για την περιγραφή των χρονικών ιδιοτήτων και σχέσεων.
Θεματικό (thematic frame): θεματικό σύστημα αναφοράς για την περιγραφή των θεματικών ιδιοτήτων και σχέσεων.

Πίνακας 7: Είδη Πλαίσιων αναφοράς

Όσον αφορά τη Σημασιολογία (semantics):

Γενικό πλαίσιο - περιβάλλον (context): η οπτική βάσει της οποίας δομούνται οι έννοιες και η πληροφορία ερμηνεύεται και αποκτά κάποιο νόημα.
Όρος (term): το όνομα της έννοιας το οποίο θεωρείται τμήμα της ταυτότητάς της.
Σημασιολογικές ιδιότητες (internal properties): οι ιδιότητες οι οποίες χαρακτηρίζουν τις γεωχωρικές έννοιες, είναι ανεξάρτητες από άλλες έννοιες και μπορεί να είναι χωρικές (σχήμα, θέση, δομή, έκταση, κ.λπ.) ή μη χωρικές (σκοπός, υλικό, χρόνος, περιοδικότητα, κ.λπ.)
Σημασιολογικές σχέσεις (external relations): οι σχέσεις των γεωχωρικών εννοιών με άλλες έννοιες οι οποίες μπορεί να είναι χωρικές (πχ. σχετική θέση, απόσταση, προσανατολισμός, εγγύτητα, γειτνίαση, κ.λπ.) ή μη χωρικές

Πίνακας 8: Σημασιολογία

Όσον αφορά τη Σημειολογία (semiotics):

Έκφραση - συμβολισμός (expression - symbolism): οι γεωγραφικές έννοιες σχετίζονται με σήματα (εικόνες, λέξεις, σύμβολα) που εκφράζουν το νόημά τους.

Πίνακας 9: Σημειολογία

Όσον αφορά την ποιότητα (quality):

Ασάφεια (vagueness): αναφέρεται στο βαθμό ανακρίβειας και αβεβαιότητας των γεωχωρικών εννοιών, των ιδιοτήτων ή των μεταξύ τους σχέσεων (χωρικών ή μη).
Προσέγγιση (approximation): αναφέρεται στο βαθμό ανάλυσης.

Πίνακας 10: Ποιότητα

3.4 Σημασιολογικά Προβλήματα

Η εκτενής αναφορά στη σημασιολογία γίνεται εφόσον, ιδιαίτερα τα τελευταία χρόνια, η πρόσβαση στην γεωχωρική πληροφορία, καθώς και η χρήση της, έχουν αλλάξει δραματικά. Σήμερα, τα γεωχωρικά δεδομένα μπορούν να ανακτηθούν από οποιοδήποτε σημείο του πλανήτη και να συσχετιστούν μεταξύ τους, ξεφεύγοντας με αυτό τον τρόπο από κάθε έννοια τοπικού πλαισίου (Kuhn 2005).

Επιπλέον, έχει προκύψει ιδιαίτερα επιτακτική ανάγκη για:

1. απαλλαγή του χρήστη από τη διαδικασία της συλλογής δεδομένων που χαρακτηρίζεται από πολυπλοκότητα και μεγάλες απαιτήσεις σε πόρους (χρήμα και χρόνος),
2. χρήση υφιστάμενων δεδομένων - αξιοποίηση της αυξανόμενης τάσης για διάθεση χωρικών δεδομένων
3. ενοποίηση χωρικών δεδομένων που παράγονται ανεξάρτητα από διαφορετικούς φορείς.

Με βάση τα παραπάνω, αναδύεται μια σειρά πολύ σημαντικών και συσχετιζόμενων προβλημάτων που αφορούν στην παραγωγή, επικοινωνία, διάδοση, ανταλλαγή, επαναχρησιμοποίηση και συσχέτιση χωρικών δεδομένων που παράγονται από ετερογενείς πηγές.

3.5 Γεωγραφικές Οντολογίες

Οι γεωγραφικές ή γεωχωρικές οντολογίες (geographic, geospatial ontologies ή geoontologies) αποτελούν οντολογίες πεδίου - ονοματολογικές (terminological) (κάθε έννοια περιγράφεται από έναν όρο, έναν ορισμό σε φυσική γλώσσα και τις ιεραρχικές σχέσεις μεταξύ των εννοιών), που επικεντρώνονται στην περιγραφή των γεωγραφικών οντοτήτων (Kavouras & Kokla 2008, Laurini 2012, 2015, Laurini & Kazar 2016).

Ενσωματώνουν και τις δύο προαναφερθείσες προσεγγίσεις (Φιλοσοφία και Πληροφορική) και στοχεύουν στη συστηματοποίηση της γνώσης του γεωγραφικού χώρου, των οντοτήτων, φαινομένων, διαδικασιών, καθώς και των μεταξύ τους συσχετίσεων. Επιπλέον, συνεισφέρουν στην :

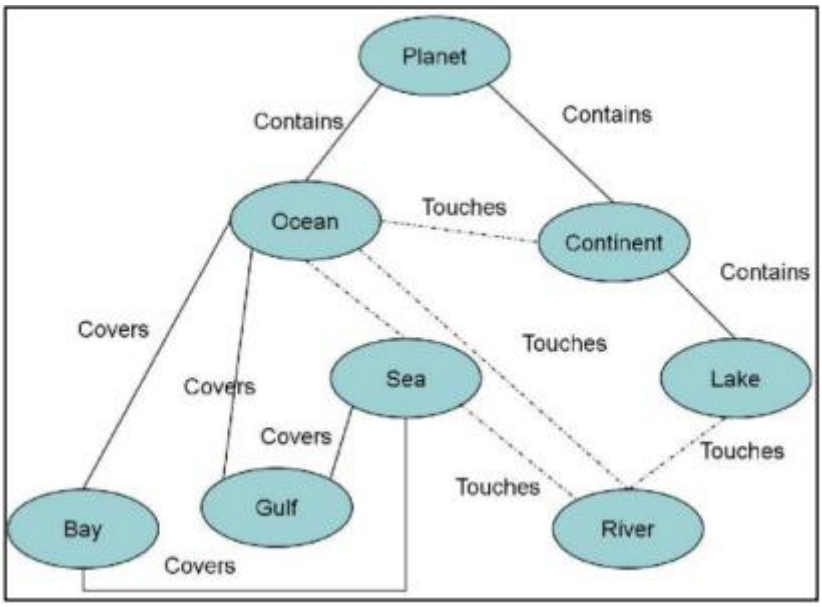
Κατανόηση των διαφορών που παρατηρούνται στην αντίληψη και τον ορισμό των γεωγραφικών οντοτήτων.
Αποδοτικότερη σημασιολογική περιγραφή και αναπαράσταση του γεωγραφικού χώρου.

Αποτελεσματικότερη ανταλλαγή γεωγραφικών πληροφοριών (ανάπτυξη και επέκταση προτύπων ανταλλαγής χωρικών δεδομένων).
Επαναχρησιμοποίηση της γνώσης.

Πίνακας 11: Συνεισφορά Γεωγραφικών Οντολογιών

Κατά το παρελθόν, η οργάνωση των γεωγραφικών οντοτήτων, στο πλαίσιο μιας οντολογίας, υλοποιούνταν μέσα από την θεμελίωση συμβατικών σχέσεων (IS-A, PART-OF, HAS-A, κ.λπ.) ανάμεσά τους. Παρόλα αυτά, έγινε γρήγορα αντιληπτό ότι ενώ μια τέτοια προσέγγιση οργανώνει μεν σημασιολογικά τις εμπλεκόμενες γεωγραφικές έννοιες, αγνοεί δε παντελώς τη χωρική τους υπόσταση, γεγονός που καθιστά αναγκαία την εισαγωγή αποδοτικότερων τρόπων αναπαράστασης του χώρου.

Για την επίτευξη καλύτερης και πληρέστερης χωρικής αναπαράστασης ενσωματώθηκαν στις γεωγραφικές οντολογίες οι χωρικές σχέσεις, με πιο σημαντικές και ευρέως διαδεδομένες από αυτές, τις τοπολογικές (Allen 1983, Egenhofer & Franzosa 1991, Randell και άλλοι 1992, Egenhofer 1994) . Αξίζει να σημειωθεί πως παρά το γεγονός ότι οι τοπολογικές σχέσεις είναι αυτές που χρησιμοποιούνται κατά κόρον, θα πρέπει να λαμβάνονται υπόψη παράλληλα και άλλες χωρικές και γεωγραφικές σχέσεις (σχέσεις κατεύθυνσης, σχετικής θέσης, σχέσεις εγγύτητας, αποστάσεις, κ.λπ.).



Εικόνα 6: Οργάνωση Γεωγραφικών Οντοτήτων Βάσει Τοπολογικών Σχέσεων

Πηγή: Laurini 2015, Laurini & Kazar 2016

3.6 Σύνδεση δεδομένων στον Ιστό

Η σύνδεση δεδομένων στον Ιστό διακρίνεται σε εσωτερική και εξωτερική. Η εσωτερική σύνδεση υπάρχει όταν συνδέονται οι πόροι μεταξύ τους μέσα στην ίδια πηγή και η εξωτερική υπάρχει όταν οι πόροι που συνδέονται μεταξύ τους είναι σε διαφορετικές πηγές. Όπως έχουμε αναφέρει η σύνδεση των δεδομένων γίνεται μέσα σε μία τριπλέτα RDF και οι πόροι είναι το υποκείμενο και το αντικείμενο οι οποίοι παρουσιάζονται με τα αντίστοιχα URIs τους, ενώ το κατηγόρημα συνδέει τους δύο προηγούμενους πόρους επίσης με ένα URI. Μπορούμε να διακρίνουμε τρία είδη συνδέσεων οι οποίοι είναι οι εξής:

- Σύνδεσμοι Σχέσης: Οι συγκεκριμένοι σύνδεσμοι συνδέουν δύο οντότητες σε πραγματικό χρόνο, όπως για παράδειγμα ο χρόνος γέννησης ενός ατόμου.

- Σύνδεσμοι ταυτότητας: Οι σύνδεσμοι ταυτότητας συνδέουν δύο URIs τα οποία υπάρχουν σε διαφορετικά σύνολα δεδομένων ώστε να δείξουν ότι αναφέρονται στον ίδιο πόρο.

- Σύνδεσμοι λεξιλογίου: Χρησιμοποιούνται για την εννοιολογική σύνδεση των δεδομένων, για να δημιουργείται νέα γνώση από τα συμπεράσματα που προκύπτουν κάθε φορά. Ορίζονται σύνδεσμοι λεξιλογίου όπως είναι οι : owl:equivalentClass και owl:equivalentProperty ,οι οποίοι χρησιμοποιούνται για την ισοδυναμία δύο όρων ή οι rdfs:subClassOf, rdfs:subPropertyOf που χρησιμοποιούνται για να δηλώσουν τις ιεραρχίες και τέλος οι skos:broadMatch, και skos:narrowMatch που χρησιμοποιούνται για να δηλώσουν μικρότερη ισοδυναμία μεταξύ των δεδομένων.

Για να δημιουργηθούν οι σύνδεσμοι RDF ανάμεσα στους πόρους από δύο σύνολα δεδομένων, χρειάζεται να γίνει αυτό είτε χειροκίνητα, είτε αυτόματα. Η αυτόματη σύνδεση των πόρων μπορεί να γίνει είτε με τη χρήση SPARQL endpoint είτε με φυλλομετρητή Συνδεδεμένων Δεδομένων, αλλά και με URIs ευρετήρια όπως είναι το Sindice και το Falcons, με τη βοήθεια των οποίων γίνεται η σύνδεση των URIs με τη διαδικασία της αναζήτησης συγκεκριμένων λέξεων κλειδιών. Από την άλλη πλευρά η χειροκίνητη σύνδεση εφαρμόζεται για πολύ μικρή ποσότητα δεδομένων και όχι για εκατομμύρια URIs από διαφορετικές πηγές. Επίσης, με τη χειροκίνητη σύνδεση συνήθως συνδέεται ένα URI με το κείμενο το οποίο περιγράφει το αντικείμενο και όχι με το URI αυτό καθεαυτό.

3.7 Ομοαναφορά

Η ομοαναφορά (Co-reference) είναι το πρόβλημα που υπάρχει όταν δύο ονόματα αποτελούν την ίδια οντότητα. Η ομοναναφορά μπορεί να δημιουργηθεί σε τρεις περιπτώσεις:

- Όταν η οντότητα είναι γνωστή με δύο ονόματα που αναφέρονται στο ίδιο πρόσωπο.
- Όταν υπάρχει λάθος στο όνομα.

- Όταν χρησιμοποιείται διαφορετικό αλφάβητο.

Στο χώρο των βιβλιοθηκών έχει αναπτυχθεί η προσπάθεια σύνδεσης ονομαστικών καταλόγων και πηγών (authority files) από διαφορετικές βιβλιοθήκες, η οποία συναντάται στην εφαρμογή Virtual International Authority File (VIAF) .

Το πρόβλημα που υπάρχει στην ομοαναφορά είναι η χρήση διαφορετικών URIs στο Σημασιολογικό Ιστό από τα διαφορετικά σύνολα δεδομένων. Όταν δεν υπάρχει γνώση πάνω στο αντικείμενο της ομοαναφοράς μπορεί εύκολα να μην υπάρξει η σωστή και κατάλληλη σύνδεση δεδομένων, κάτι που έχει σαν αποτέλεσμα την ελλιπή δημιουργία συμπερασμάτων. Από την άλλη πλευρά οι διπλές εγγραφές μέσα στην ίδια βάση ή όταν υπάρχει σύνδεση εγγραφών από διαφορετικές βάσεις δημιουργεί την επιβεβαίωση του προβλήματος της ομοναφοράς, κάτι που συναντάμε στις βάσεις δεδομένων. Η σύνδεση των εγγραφών μέσα σε RDF διαφορετικούς γράφους λαμβάνει χώρα όταν υπάρχει ένα κοινό χαρακτηριστικό τουλάχιστον ανάμεσα τους.

3.8 Πρόσβαση στα Συνδεδεμένα Δεδομένα

Η πρόσβαση στα Συνδεδεμένα Δεδομένα από τους ανθρώπους ή τις μηχανές γίνεται με τους παρακάτω τρόπους:

Linked Data Browser: Είναι μία εφαρμογή που μοιάζει με τους Web Browsers και ο χρήστης μέσω ιστοσελίδων μπορεί να πλοηγηθεί σε RDF συνδέσμους. Ο χρήστης πρέπει να χρησιμοποιήσει λέξεις κλειδιά για να προβεί σε αναζήτηση που αυτή με τη σειρά της γίνεται από μηχανές αναζήτησης οι οποίες περνούν από τον ιστό δεδομένων σε συνδέσμους RDF, επιπλέον ενοποιούν και φιλτράρουν τα δεδομένα που βρίσκουν με τη χρήση των τάξεων. Γνωστοί linked data browsers είναι οι Tabulator, Disco, Ontology-browser, Falcons Explorer.
SPARQL endpoint: Μοιάζει πολύ με τη σχεσιακή SQL. Ουσιαστικά είναι κάποια σημεία στον Ιστό που παρέχουν πρόσβαση σε RDF δεδομένα μέσω του SPARQL Protocol and RDF Query Language (SPARQL) το οποίο είναι η γλώσσα ερωτήσεων των παραπάνω δεδομένων.
Download: αφορά τη λήψη των δεδομένων σε τοπικό αρχείο.
Ειδικές Εφαρμογές: Είναι εφαρμογές ειδικού σκοπού και προσφέρουν στο χρήστη συγκεκριμένα δεδομένα ενός πεδίου με συγκεκριμένο τρόπο. Για παράδειγμα μία τέτοια εφαρμογή είναι το Diseasome, η οποία αφορά την εξερεύνηση των ανθρώπινων ανωμαλιών κι ασθενειών.

Πίνακας 12: Τρόποι πρόσβαση στα Συνδεδεμένα Δεδομένα

3.9 Τα οφέλη των Συνδεδεμένων Δεδομένων για τις βιβλιοθήκες

Η τεχνολογία των συνδεδεμένων δεδομένων εμφανίζει ποικίλα πλεονεκτήματα, τόσο για τον ιδιοκτήτη δεδομένων που αποφασίζει να τα δημοσιεύσει σε μορφή συνδεδεμένων δεδομένων όσο και για τον τελικό χρήστη ο οποίος τα καταναλώνει και τα εκμεταλλεύεται.

Για παράδειγμα όσον αφορά τον ιδιοκτήτη δεδομένων, μπορεί να πετύχει με μικρό κόστος την ολοκλήρωση και σύνδεση των δεδομένων του με εξωτερικά ετερογενή σύνολα δεδομένων που μπορεί να είναι δομημένα, ημιδομημένα και αδόμητα, ενώ η υιοθέτηση του RDF μοντέλου για την περιγραφή των δεδομένων μιας εφαρμογής παρέχει ευέλικτη μοντελοποίηση ενός συγκεκριμένου τομέα ενδιαφέροντος και εύκολη ενημέρωση του σχήματος οργάνωσης των δεδομένων της εφαρμογής.

Ενώ όσον αφορά στον χρήστη-καταναλωτή συνδεδεμένων δεδομένων εκμεταλλεύεται τη χρήση URI για τον μονοσήμαντο προσδιορισμό οντοτήτων και την ύπαρξη συνδέσμων μεταξύ σχετικών δεδομένων και αυτό έχει σαν αποτέλεσμα την αποφυγή σε μεγάλο βαθμό χρονοβόρων και επίπονων διαδικασιών εύρεσης αντιστοιχιών μεταξύ διαφορετικών συνόλων δεδομένων καθώς και ανακάλυψης νέων συνόλων δεδομένων.

Τα συνδεδεμένα δεδομένα εκτός από τα παραπάνω πλεονεκτήματα τους, κρίνονται επωφελή και για τις βιβλιοθήκες κάτι το οποίο έχει αναγνωριστεί διεθνώς τα τελευταία χρόνια. Τα οφέλη που παρέχουν τα συνδεδεμένα δεδομένα αφορούν όχι μόνο στις βιβλιοθήκες ως οργανισμούς αλλά και σε μεμονωμένα άτομα όπως βιβλιοθηκονόμους, προγραμματιστές, ερευνητές και φοιτητές. Συγκεκριμένα, η αποτύπωση των μεταδεδομένων των Βιβλιοθηκών σε μορφή Συνδεδεμένων Δεδομένων καθιστά ευκολότερη την ανακάλυψη των τεκμηρίων της από ενδιαφερόμενους χρήστες.

Ο σημασιολογικός ιστός αποτελεί μία επέκταση του παγκόσμιου ιστού , όπως αναφέρθηκε και παραπάνω, όπου τα δεδομένα αποκτούν δομημένη περιγραφή με αυστηρά καθορισμένο νόημα. Τα εν λόγω δομημένα δεδομένα, που συχνά προστίθενται σε υπάρχοντα έγγραφα ιστού μέσω γλωσσών σημασιολογικής επισημείωσης, όπως η RDFa και τα μικροδεδομένα, λαμβάνονται συνήθως υπόψη από τους αλγόριθμους υπολογισμού συνάφειας των μηχανών αναζήτησης και βελτιώνουν τη θέση στην οποία της ορατότητας των τεκμηρίων των Βιβλιοθηκών που δημοσιεύουν τους πόρους τους με τη μορφή συνδεδεμένων δεδομένων, ενισχύοντας την παρουσία τους στον παγκόσμιο ιστό και προσελκύοντας νέους χρήστες.

Ακόμα, για τους ίδιους τους τελικούς χρήστες η ύπαρξη συνδεδεμένων δεδομένων προσφέρει προχωρημένες δυνατότητες πλοήγησης μέσω σύνδεσης των τεκμηρίων των Βιβλιοθηκών με εξωτερικές υπηρεσίες όπως η Wikipedia και ανακάλυψης των πόρων εκείνων που καλύπτουν

τις πληροφοριακές ανάγκες τους. Παράλληλα, η εύκολη διασύνδεση δεδομένων από διαφορετικές πηγές προσφέρει στις βιβλιοθήκες τη δυνατότητα να βελτιώσουν σημαντικά την ποιότητα των μεταδεδομένων τους. Ένας δεδομένος πόρος μπορεί να περιγραφεί σε συνεργασία με άλλες Βιβλιοθήκες και να συνδεθεί με δεδομένα που προέρχονται από άλλες κοινότητες ή ακόμα και άτομα.

Με αυτόν τον τρόπο, αναδεικνύεται για τις Βιβλιοθήκες η ευκαιρία να προσφέρουν στους χρήστες τους περιγραφές καλύτερης ποιότητας και λεπτομέρειας, αλλάζοντας τον παραδοσιακό τρόπο καταλογογράφησης, ο οποίος παράγει αυτόνομες μεμονωμένες περιγραφές τεκμηρίων, σε ένα συμμετοχικό μοντέλο επαναχρησιμοποίησης πληροφορίας που έχει δημοσιευθεί από εξωτερικούς οργανισμούς και άτομα. Αυτό το νέο μοντέλο καταλογογράφησης θα αποφέρει σημαντικά οφέλη και στο σύνολο των βιβλιοθηκονόμων, οι οποίοι θα είναι ελεύθεροι πλέον να εστιάζουν στον τομέα στον οποίο ειδικεύονται, χωρίς να χρειάζεται να τεκμηριώνουν εξ αρχής πόρους που έχουν ήδη τεκμηριωθεί επιτυχημένα από τρίτους.

Ακόμα τα συνδεδεμένα δεδομένα ανοίγουν νέους δρόμους για την ανάπτυξη καινοτόμων εφαρμογών τα οποία συνδυάζουν τα βιβλιοθηκονομικά δεδομένα με άλλα ανοικτά σύνολα δεδομένων και προσφέρουν μία διαφορετική οπτική στα βιβλιοθηκονομικά δεδομένα, τα οποία μέχρι σήμερα ήταν διαθέσιμα μόνο μέσω των ολοκληρωμένων βιβλιοθηκονομικών συστημάτων στα οποία είναι αποθηκευμένα.

Σε γενικές γραμμές οι βιβλιοθήκες ανοίγοντας τα δεδομένα τους και προσφέροντάς τα σε μία εύκολα επεξεργάσιμη και μηχαναγνώσιμη μορφή, ενσωματώνουν τα δεδομένα τους στον παγκόσμιο ιστό, μέσα από τον οποίο μπορούν να ξαναχρησιμοποιηθούν από εφαρμογές και να αξιοποιηθούν με μη αναμενόμενους τρόπους. Για παράδειγμα, ο συνδυασμός των δεδομένων δανεισμού μιας Βιβλιοθήκης με γεωχωρικά δεδομένα και δεδομένα δρομολογίων αστικών συγκοινωνιών θα μπορούσε να χρησιμοποιηθεί από μια εφαρμογή που προτείνει σε φοιτητές τη συντομότερη διαδρομή για την απόκτηση ενός συγκεκριμένου βιβλίου. Τέτοιες εφαρμογές αυξάνουν την ορατότητα των τεκμηρίων των Βιβλιοθηκών και συντελούν στην αύξηση της αναγνωσιμότητάς τους.

Τα συνδεδεμένα δεδομένα επίσης προσφέρουν πολλά και αξιόλογα πλεονεκτήματα χρησιμοποιώντας τις τρέχουσες πρακτικές για τη δημιουργία της σύνδεσης των δεδομένων των βιβλιοθηκών και παρέχοντας μία φυσική επέκταση των συνεργατικών μοντέλων επιμερισμού, που ιστορικά διαχειρίζονται οι βιβλιοθήκες. Τα πλεονεκτήματα της διασύνδεσης των δεδομένων μιας βιβλιοθήκης επηρεάζουν κατά πολύ τις βασικές λειτουργίες των βιβλιοθηκών αφού είναι άμεσα συνυφασμένες με το ύφος και τη δομή των υπηρεσιών των βιβλιοθηκονόμων, των πληροφορικών και των προμηθευτών των συστημάτων τους. Τα συνδεδεμένα δεδομένα

και ειδικά τα ανοιχτά συνδεδεμένα δεδομένα, είναι διαμοιραζόμενα, επεκτάσιμα και επαναχρησιμοποιήσιμα. Επιπλέον, παρέχουν υποστήριξη στην πολύγλωσση λειτουργία των δεδομένων και των υπηρεσιών των χρηστών, όπως την ταυτοποίηση του προσδιορισμού των εννοιών μέσω των "αναγνώσιμων γλωσσών" URIs. Επιπλέον, διευκολύνουν την επαναχρησιμοποίηση του συνόλου των δεδομένων σε όλο το φάσμα της πολιτιστικής κληρονομιάς εμπλουτίζοντας την περιγραφή τους με προερχόμενες πληροφορίες από δεδομένα κοινοτήτων εκτός βιβλιοθηκών ή προσώπων αλλά και από σχετικές εμπειρίες χρηστών πάνω σε συγκεκριμένους τομείς.

Η χρήση των παγκόσμιων αναγνωριστικών εξασφαλίζει τη δυνατότητα μιας συνεργατικής περιγραφής των πόρων με δυνατότητα πολλών περιγραφών στο ίδιο περιεχόμενο και παρέχει ανεξαρτησία από τα Ενοποιημένα Συστήματα Βιβλιοθηκών. Οι πολλές διασυνδέσεις που προκύπτουν εμπεριέχουν περαιτέρω δεδομένα από αξιόπιστες πηγές διευρύνοντας μια συλλογή βιβλιοθήκης πέρα από το τοπικό σύνολο των πηγών της επαυξάνοντας με αυτό τον τρόπο την αξία των πόρων. Η χρήση των URIs διευκολύνει την προσβασιμότητα των μεταδεδομένων με την αναφορά των πόρων τους σε ένα ευρύτερο φάσμα πηγών δεδομένων περιγραφής έργων, τόπων, ανθρώπων, γεγονότων και θεμάτων. Η περιοχή των Διαδικτυακών Διευθύνσεων (Internet's Domain Name) παρέχει σταθερότητα και εμπιστοσύνη τοποθετώντας αυτά τα αναγνωριστικά μέσα σε ένα πλαίσιο μίας συστηματοποιημένης και καλά κατανοητής ιδιοκτησίας σαν αυτής των βιβλιοθηκών που χαρακτηρίζεται από την ιδιαιτερότητα της παροχής αξιόπιστων μεταδεδομένων ως δεδομένων στον Ιστό, για πόρους μακροπρόθεσμης πολιτισμικής σημασίας.

Η επαναχρησιμοποίηση των URIs επιτρέπει στους παραγωγούς δεδομένων να συνεισφέρουν τμήματα των δεδομένων τους ως δηλώσεις (Statements), για έναν πόρο. Στο σύστημα των εγγράφων η ανταλλαγή των δεδομένων πραγματοποιείται με τη μορφή ολόκληρων εγγραφών η κάθε μία εκ των οποίων θεωρείται ως πλήρης περιγραφή. Αντίθετα, σε ένα οικοσύστημα γράφων, σαν αυτό του Σημασιολογικού Ιστού, οι μεμονωμένες δηλώσεις ενός πόρου μπορούν να συγκεντρώνονται σε ένα παγκόσμιο γράφημα.

3.10 Η γλώσσα ερωτημάτων SPARQL

Για τη διαχείριση των δεδομένων του Σημασιολογικού Ιστού, απαιτείται κάποιος μηχανισμός που να είναι εύκολα αντιληπτός από τους χρήστες. Για τον σκοπό αυτό, χρησιμοποιούνται γλώσσες ερωτημάτων, όπως συμβαίνει και στον κλασικό Παγκόσμιο Ιστό. Η πιο γνωστή γλώσσα του Σημασιολογικού Ιστού αποτελεί η SPARQL, η οποία δίνει τη δυνατότητα δημιουργίας, ανεύρεσης, ενημέρωσης και διαγραφής RDF δεδομένων και βασίζεται στην αντιστοίχιση γράφων (graph-matching) μέσω μεταβλητών. Για τη σύνταξή της,

χρησιμοποιούνται προκαθορισμένες λέξεις (όροι) εύκολα αντιληπτές στον άνθρωπο, μερικές από τις οποίες είναι οι εξής:

- PREFIX : Χρησιμοποιείται για την ανάθεση προθέματος συντόμευσης για τους χώρους ονοματοδοσίας.
- SELECT : Καθορίζει τις μεταβλητές μέσω των οποίων θα εμφανιστούν τα αποτελέσματα ύστερα από την εκτέλεση κάποιου ερωτήματος.
- FROM : Καθορίζει την πηγή προέλευσης των δεδομένων. Η πηγή μπορεί να είναι ένας ονομαστικός γράφος ή ο προεπιλεγμένος γράφος που μπορεί να περιλαμβάνει έναν ή περισσότερους ονομαστικούς.
- WHERE : Καθορίζει τη λογική του υπολογισμού των αποτελεσμάτων κάποιου ερωτήματος.
- LIMIT : Καθορίζει τον αριθμό των αποτελεσμάτων που θα επιστραφούν.
- GROUP BY : Χρησιμοποιείται για την ομαδοποίηση αποτελεσμάτων.
- ORDER BY : Χρησιμοποιείται για να καθοριστεί η ταξινόμηση των αποτελεσμάτων.
- DISTINCT : Συνδυάζεται με το SELECT ώστε να επιστραφούν μοναδικά αποτελέσματα.
- OPTIONAL : Καθορίζει ότι κάποιο μοτίβο γράφου (graph pattern) είναι προαιρετικό για τον υπολογισμό των αποτελεσμάτων.
- FILTER : Επιτρέπει το φιλτράρισμα των αποτελεσμάτων βάσει κριτηρίου.
- UNION : Επιτρέπει τον συνδυασμό αποτελεσμάτων από διαφορετικά μοτίβα γράφων

Πίνακας 13: Όροι Σύνταξης της SPARQL

Η SPARQL δεν είναι case sensitive γλώσσα, με αποτέλεσμα, στους όρους της να μπορούν να χρησιμοποιηθούν και κεφαλαία και μικρά γράμματα. Μια μεταβλητή της SPARQL αποτελεί μια συμβολοσειρά, η οποία ξεκινά με το σύμβολο του αγγλικού ερωτηματικού "?" και περιλαμβάνει ένα η περισσότερα γράμματα όπως για παράδειγμα ?subject. Μεταβλητή μπορεί να χρησιμοποιηθεί για κάθε στοιχείο μιας RDF δήλωσης, αναφέροντας σε συγκεκριμένο τμήμα γράφου. Παρακάτω δίνεται παράδειγμα απλού SPARQL ερωτήματος:

```
SELECT
?s ?p ?o
 WHERE
{ ?s ?p ?o .
}
```

Παράδειγμα απλού SPARQL ερωτήματος. Στο συγκεκριμένο παράδειγμα γίνεται ανεύρεση και επιλογή όλων των RDF δηλώσεων από έναν προεπιλεγμένο γράφο. Η 1η γραμμή, περιέχει τον όρο SELECT που καθορίζει τις μεταβλητές μέσω των οποίων θα εμφανιστούν τα

αποτελέσματα του ερωτήματος και η 2η περιέχει τις μεταβλητές αυτές. Η 3η , περιέχει τον όρο WHERE βάσει του οποίου γίνεται η ανεύρεση δεδομένων με χρήση των μεταβλητών της 4ης γραμμής. Αγκύλες όπως αυτές της 3ης και της 5ης γραμμής, χρησιμοποιούνται για να καθορίσουν την αρχή και το τέλος του περιεχομένου, που αναφέρεται στον όρο που υπάρχει πριν από την αρχική αγκύλη (ο όρος WHERE για την συγκεκριμένη περίπτωση). Οι μεταβλητές της 2ης γραμμής, εναλλακτικά θα μπορούσαν να αντικατασταθούν από το σύμβολο του πολλαπλασιασμού "*" (που σημαίνει επιλογή όλων) για απλότητα, καθώς αποτελούν όλες τις μεταβλητές που περιλαμβάνονται στην 4η γραμμή.

Αν θα χρειαζόταν τα δεδομένα του ερωτήματος να προέρχονται από συγκεκριμένο γράφο, θα συμπεριλαμβάναμε μεταξύ των γραμμών 2 και 3 τον όρο FROM ακολουθούμενο από το URI του γράφου, ως εξής:

```
FROM <graph – uri>
```

Όπου το graph-uri αποτελεί το URI του γράφου (ονομαστικού ή προεπιλεγμένου).

Ο αριθμός των αποτελεσμάτων που θα επιστραφούν από το ερώτημα, μπορεί να καθοριστεί με χρήση του όρου LIMIT ακολουθούμενο από έναν ακέραιο αριθμό.

Για παράδειγμα, για την επιστροφή 100 αποτελεσμάτων, θα τοποθετούσαμε την έκφραση LIMIT 100, στο τέλος του ερωτήματος.

Τα αποτελέσματα που θα επιστραφούν, μπορούν να ταξινομηθούν βάσει ενός ή περισσότερων μεταβλητών. Για τον σκοπό αυτό, χρησιμοποιείται ο όρος ORDER BY ακολουθούμενο από μια ή παραπάνω μεταβλητές κατά σειρά προτεραιότητας. Για παράδειγμα για να ταξινομηθούν τα αποτελέσματα, πρώτα βάσει της μεταβλητής ?s και ύστερα βάσει της μεταβλητής ?p, θα χρησιμοποιούσαμε την έκφραση ORDER BY ?s ?p, μετά το περιεχόμενο του όρου WHERE.

Σε μια έκφραση ταξινόμησης, μπορεί να οριστεί ανάστροφη σειρά ταξινόμησης για μια ή περισσότερες μεταβλητές. Αυτό επιτυγχάνεται με χρήση του όρου DESC ακολουθούμενο από τη μεταβλητή μέσα σε παρενθέσεις. Για παράδειγμα, για να οριστεί ότι η μεταβλητή ?p θα έχει κανονική σειρά ταξινόμησης, ενώ η ?p, θα έχει ανάστροφη, θα χρησιμοποιούσαμε την έκφραση ORDER BY ?s DESC(?p). Ο όρος DISTINCT, μπορεί να χρησιμοποιηθεί για να διατηρηθούν μόνο οι μοναδικές τιμές εκ των αποτελεσμάτων ενός ερωτήματος και ορίζεται μετά τον όρο SELECT (πριν τις μεταβλητές). Για παράδειγμα, η έκφραση SELECT DISTINCT ?s, θα επέστρεφε μόνο τις μοναδικές τιμές για την μεταβλητή ?s, ενώ η έκφραση SELECT DISTINCT ?s ?p, θα επέστρεφε μόνο τις μοναδικές τιμές του συνδυασμού ?s ?p.

Υπάρχουν περιπτώσεις όπου απαιτείται η ομαδοποίηση αποτελεσμάτων βάσει μιας ή περισσότερων μεταβλητών, για χρήση συναρτήσεων που εφαρμόζονται σε ομαδοποιημένα δεδομένα. Συναρτήσεις αυτής της κατηγορίας αποτελούν εκείνη του μέσου όρου, του ελαχίστου, του μεγίστου, κ.ά. Η ομαδοποίηση πραγματοποιείται μέσω του όρου GROUP BY ακολουθούμενο από μια ή παραπάνω μεταβλητές κατά σειρά προτεραιότητας. Για παράδειγμα, η χρήση της έκφρασης GROUP BY ?s, μετά το περιεχόμενο του όρου WHERE, θα είχε ως αποτέλεσμα την ομαδοποίηση των αποτελεσμάτων που θα επιστρεφόταν μέσω της μεταβλητής ?s

Στα ερωτήματα SPARQL, εκτός από τις μεταβλητές, μπορούν να χρησιμοποιηθούν και URI για κάθε στοιχείο μιας RDF δήλωσης. Τα URI, είτε περικλείονται σε γωνιακές αγκύλες όταν συντάσσονται στην πλήρη τους μορφή, είτε συντομεύονται με χρήση προθέματος συντόμευσης. Για τον ορισμό ενός προθέματος συντόμευσης, χρησιμοποιείται στην αρχή του ερωτήματος, ο όρος PREFIX με την εξής σύνταξη:

PREFIX prefix-name:

Όπου το prefix-name αποτελεί μια συμβολοσειρά η οποία αντιπροσωπεύει το πρόθεμα συντόμευσης και το uri αποτελεί το URI για το οποίο χρησιμοποιείται το πρόθεμα. Παρακάτω δίνεται παράδειγμα SPARQL ερωτήματος, το οποίο επιστρέφει το URI της ιστοσελίδας του Tim Berners-Lee, χρησιμοποιώντας ως πηγή RDF δηλώσεων:

```
PREFIX dc:
PREFIX foaf:
SELECT
?homepage
WHERE {
?subject dc:title "CERN experience." .
?subject dc:creator ?creator .
 ?creator foaf:homepage ?homepage .
}
```

Οι δύο πρώτες γραμμές του ερωτήματος, περιέχουν τα προθέματα συντόμευσης που θα χρησιμοποιηθούν. Η 5η , περιέχει τη μεταβλητή μέσω της οποίας θα επιστραφεί το URI της ιστοσελίδας, ύστερα από τον υπολογισμό του περιεχομένου των γραμμών 7,8 και 9. Από τις

γραμμές αυτές, η 7η , απομονώνει από την πηγή RDF δηλώσεων, μόνο εκείνες τις δηλώσεις που διαθέτουν ως τιμή για την ιδιότητα dc:title, την CERN experience. Η 8 η , εξειδικεύει περισσότερο αυτές τις δηλώσεις, απομονώνοντας μόνο εκείνες που διαθέτουν ταυτόχρονα και την ιδιότητα dc:creator και η 9η , ορίζει ότι η τιμή της ιδιότητας dc:creator θα πρέπει να συνδέεται μέσω της ιδιότητας foaf:homepage σε κάποια τιμή.

Η SPARQL διαθέτει διάφορες συναρτήσεις μέσω των οποίων διευκολύνονται διάφορες διαδικασίες. Παρακάτω ακολουθούν κάποιες από αυτές και οι περιγραφές τους.

• str - Μπορεί να μετατρέψει ένα URI σε απλή συμβολοσειρά.

• strlen - Επιστρέφει το πλήθος χαρακτήρων μιας συμβολοσειράς.

• strstarts - Συγκρίνει δύο συμβολοσειρές και επιστρέφει δυαδική τιμή ως ένδειξη ομοιότητας. Το αποτέλεσμα της σύγκρισης είναι αληθές, όταν τα αρχικά γράμματα των συμβολοσειρών είναι ίδια.

• strends - Λειτουργεί όπως και η προηγούμενη, όμως το αποτέλεσμα της σύγκρισης είναι αληθές, όταν τα τελικά γράμματα των συμβολοσειρών είναι ίδια.

• contains - Εξετάζει αν μια συμβολοσειρά περιέχεται σε μια άλλη και επιστρέφει ανάλογη δυαδική τιμή.

• regex - Αναζητά ένα παρεχόμενο μοτίβο μέσα σε μια συμβολοσειρά και επιστρέφει δυαδική τιμή ανάλογα με το αποτέλεσμα της αναζήτησης.

• concat - Πραγματοποιεί συνένωση δύο συμβολοσειρών. • round - Στρογγυλοποιεί έναν πραγματικό αριθμό στον πλησιέστερο ακέραιο.

• abs - Επιστρέφει την απόλυτη τιμή ενός αριθμού.

3.11 DBpedia

Η DBpedia, είναι ένα έργο το οποίο ξεκίνησε το 2007 βασισμένο στην κοινότητα και αποτελεί πλέον έναν από τους μεγαλύτερους παρόχους δεδομένων του Σημασιολογικού Ιστού. Τα περιεχόμενά της, προέρχονται από την εξαγωγή δομημένων δεδομένων (όπως είναι τα πλαίσια πληροφορικών, οι λίστες, τα δεδομένα κατηγοριοποίησης) που περιλαμβάνονται στα άρθρα της εγκυκλοπαίδειας Wikipedia. Δεδομένα αυτής της μορφής εξάγονται από 125 διαφορετικές γλωσσικές εκδόσεις της Wikipedia, εκ των οποίων η αγγλική προσφέρει τα περισσότερα κατ' αναλογία. Το πλαίσιο πληροφορίας (infobox), είναι ένας πίνακας που βρίσκεται συνήθως στο

δεξί μέρος των ιστοσελίδων της Wikipedia και παρέχει σύντομες πληροφορίες. Η σταθερή δομή του, καθιστά εύκολη την αυτοματοποιημένη εξαγωγή και αξιοποίηση των περιεχομένων του. Η DBpedia, αποσπά τα δεδομένα ανάλογων πλαισίων χαρακτηρίζοντάς τα σημασιολογικά με χρήση οντολογίας που περιλαμβάνει 320 κλάσεις και 1.650 ιδιότητες.

Η ίδια οντολογία χρησιμοποιείται για τα δεδομένα που προέρχονται από 27 γλωσσικές εκδόσεις της Wikipedia περιλαμβάνοντας την αγγλική, βουλγάρικη, ελληνική, ισπανική κ.ά.. Σε αντίθεση με την Wikipedia στην οποία επιτρέπονται μόνο αναζητήσεις σε μορφή κειμένου, η DBpedia επιτρέπει σύνθετες αναζητήσεις βασισμένη σε τεχνικές του Σημασιολογικού Ιστού. Αποτελεί έναν από τους κεντρικούς κόμβους του νέφους των ΑΔΔ, και μπορεί να διασυνδεθεί με RDF σύνολα δεδομένων προερχόμενα από διαφορετικούς τομείς, λόγω της ποικιλίας των θεμάτων που καλύπτει. Υπάρχουν διάφορες εφαρμογές που αξιοποιούν τα περιεχόμενά της, παρέχοντας σύνθετες δυνατότητες. Παράδειγμα ανάλογης εφαρμογής, αποτελεί η DBpedia Relationship Finder, η οποία επιτρέπει την ανεύρεση σχέσεων μεταξύ οντοτήτων που περιέχονται σε άρθρα της Wikipedia.

Η DBpedia, παρέχει ελεύθερη πρόσβαση στα μεταδεδομένα της, μέσω τριών μηχανισμών που είναι οι εξής:

1. Τελικό σημείο SPARQL.

2. RDF σύνολο δεδομένων.

3. HTTP URI

Το τελικό σημείο που διαθέτει, στηρίζεται σε triplestore OpenLink Virtuoso. Για την προστασία υπερφόρτωσής του, επιτρέπει την εκτέλεση μόνο των ερωτημάτων που δεν επιφέρουν μεγάλο κόστος και περιορίζει το πλήθος των αποτελεσμάτων στις 10.000 εγγραφές. Το RDF σύνολο δεδομένων της, διατίθεται για κατέβασμα και χρήση μέσω των σειριοποιήσεων N-Triples, N-Quads και Turtle, ενώ παρέχει και HTTP URIs για την πρόσβαση σε μεταδεδομένα.

3.12 Geonames

Η βάση δεδομένων Geonames είναι ένα μεγάλο, δωρεάν σύνολο δεδομένων που περιέχει πληροφορίες για γεωγραφικές τοποθεσίες, είναι δηλαδή μία γεωγραφική βάση δεδομένων. Τα δεδομένα, που κυκλοφόρησαν με την άδεια της Creative Commons(*Creative Commons Attribution cc-by license*), είναι διαθέσιμα μέσω της υπηρεσίας web ή ως ημερήσια λήψη

αρχείου κειμένου διαχωρισμένου με καρτέλες για εμπορική και μη εμπορική χρήση. Περιέχει πάνω από 11 εκατομμύρια εγγραφές. Η GeoNames συγκεντρώνει δεδομένα από πολλές πηγές, η σημαντικότερη από τις οποίες είναι η Εθνική Υπηρεσία Γεωχωρικών Πληροφοριών των Ηνωμένων Πολιτειών. Άλλες σημαντικές πηγές είναι η υπηρεσία εθνικής χαρτογράφησης ή υπηρεσίες εθνικής στατιστικής όλων των χωρών κάθε φορά που δημοσιεύουν δεδομένα συμβατά με την cc-by άδεια.

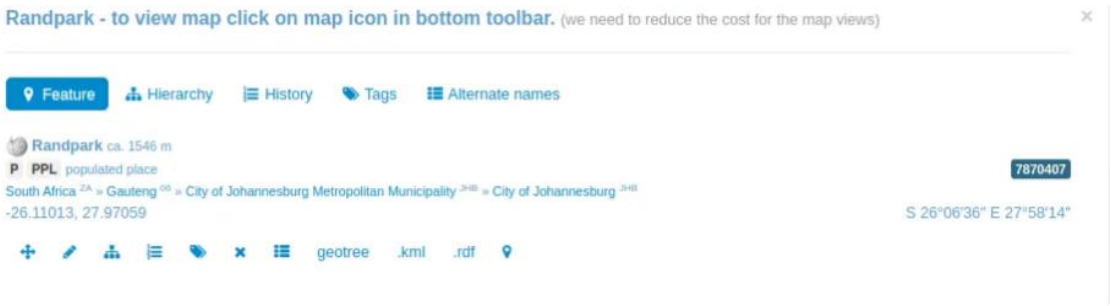
Τα δεδομένα στη βάση δεδομένων Geonames προέρχονται από πολλαπλά επίσημα σύνολα δεδομένων προτού η ομάδα της τα μετατρέψει σε τυπική μορφή. Όπως συμβαίνει με όλα τα έργα ανοιχτού κώδικα, οι χρήστες μπορούν να ενημερώσουν τα δεδομένα μέσω μιας διεπαφής ιστού εάν βρουν τυχόν αποκλίσεις. Ο ιστότοπος Geonames παρέχει επίσης έναν τρόπο περιήγησης στα δεδομένα και τα εκθέτει μέσω ενός API . Μια βιβλιοθήκη πελάτη, που αναπτύχθηκε για πολλές γλώσσες, μειώνει το βάρος της ανάπτυξης έναντι αυτού του API. Η ομάδα Geonames παρέχει επίσης μια premium υπηρεσία δεδομένων που προσφέρει καθαρά δεδομένα και υποστήριξη για τον πελάτη.

Όπως αναφέρθηκε παραπάνω, τα δεδομένα είναι διαθέσιμα ως διάφορα αρχεία με δυνατότητα λήψης, διαχωρισμένα σε καρτέλες. Το κύριο αρχείο, γενικά γνωστό ως Geonames, αντιπροσωπεύει ένα πλήρες, αποκανονικοποιημένο σύνολο δεδομένων. Ωστόσο, για να αξιοποιήσετε στο έπακρο τα δεδομένα, συνιστάται η χρήση των διαθέσιμων βοηθητικών αρχείων για τη βελτίωση της χρηστικότητας των δεδομένων. Για παράδειγμα, το αρχείο Geonames περιέχει εναλλακτικά ονόματα για τη δεδομένη τοποθεσία ως λίστα διαχωρισμένη με κόμματα στο ίδιο πεδίο ονόματος. Ωστόσο, χρησιμοποιώντας το αρχείο εναλλακτικών ονομάτων, παρέχονται περαιτέρω λεπτομέρειες του εναλλακτικού τίτλου, συμπεριλαμβανομένης της γλώσσας και της χρήσης. Μπορούμε να το δούμε αυτό κοιτάζοντας μια πραγματική τοποθεσία στη βάση δεδομένων. Για παράδειγμα, χρησιμοποιώντας τη λειτουργία αναζήτησης στον ιστότοπο, αναζήτησα το προάστιο Randpark στο Γιοχάνεσμπουργκ της Νότιας Αφρικής.

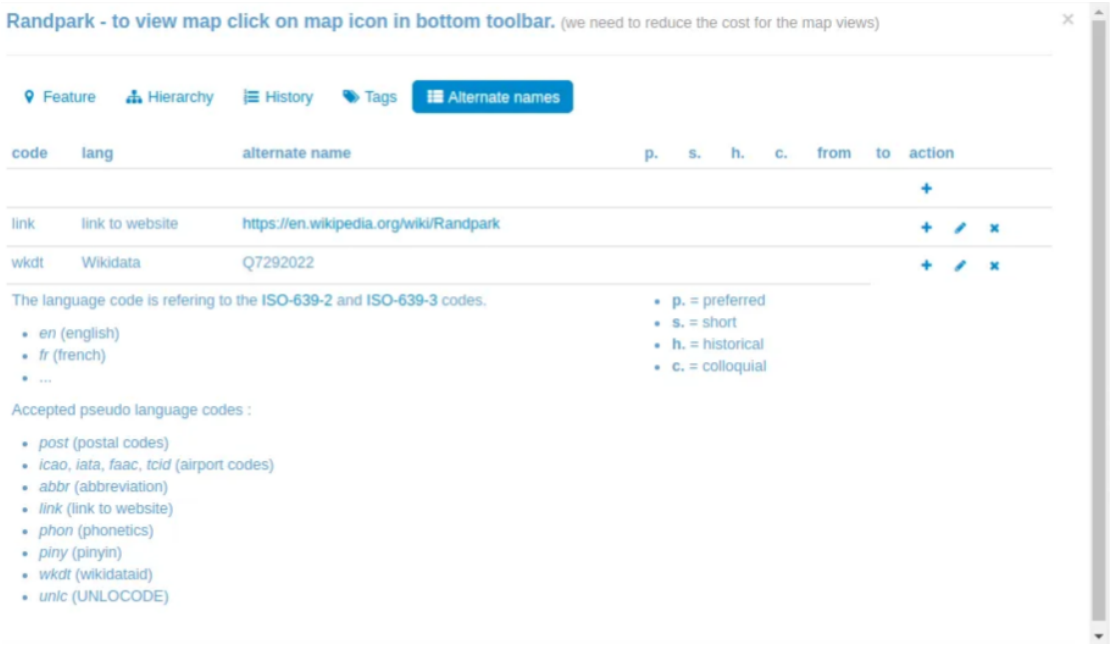
Η καρτέλα χαρακτηριστικών του αποτελέσματος αναζήτησης είναι η πρώτη οθόνη που βλέπουμε όταν ανοίγουμε την τοποθεσία. Το IT μας δείχνει ότι το Randpark είναι ένα «κατοικημένο μέρος» και μας δίνει λεπτομέρειες για το πού μπορούμε να το βρούμε, συμπεριλαμβανομένου του γεωγραφικού πλάτους και του γεωγραφικού μήκους.

Από τις υπόλοιπες καρτέλες, η καρτέλα Hierarchy-Ιεραρχία εμφανίζει τις διοικητικές λεπτομέρειες για την τοποθεσία, δηλαδή επαρχία, δήμο ή/και περιφέρεια στην περίπτωση της Νότιας Αφρικής. Η καρτέλα History-Ιστορικό εμφανίζει το ιστορικό επεξεργασίας για το Γεωόνομα και η καρτέλα Tags-Ετικέτες εμφανίζει τυχόν καρτέλες που έχουν αντιστοιχιστεί στην ενεργή εγγραφή.

Μεγαλύτερο ενδιαφέρον παρουσιάζει η καρτέλα Alternative Name-Εναλακτικό Όνομα που φαίνεται παρακάτω.



Μεγαλύτερο ενδιαφέρον παρουσιάζει η καρτέλα Alternate Names που φαίνεται παρακάτω.



Εικόνα 7: Η γεωχωρική βάση δεδομένων geonames κατά την αναζήτηση του Randpark

Πηγή: https://www.geonames.org/

Στην αρχή του, το πεδίο alternative name χρησίμευε παλαιότερα για την αποθήκευση του ονόματος της τοποθεσίας σε διαφορετικές γλώσσες. Ωστόσο, τώρα έχει φτάσει να αντιπροσωπεύει έναν τρόπο προσθήκης δεδομένων σε ένα Geonname. Στο αρχείο alternative name , ένα πεδίο που ονομάζεται lang είναι ένας περιγραφέας των δεδομένων που περιέχονται στην εγγραφή. Αυτό το πεδίο μπορεί να περιέχει έναν κωδικό γλώσσας iso δύο γραμμάτων (π.χ. en, fr, st) ή μπορεί να έχει μια περιγραφική τιμή των δεδομένων. Οι τιμές μπορεί να είναι οποιαδήποτε από τις παρακάτω:

- post (postal codes)
- icao, iata, faac, tcid (airport codes)
- abbr (abbreviation)

- link (link to website)
- phon (phonetics)
- piny (pinyin)
- wkdt (wikidataid)
- unlc (UNLOCODE)
-

Ο κύριος πίνακας "geonname" έχει τα ακόλουθα πεδία:

geonameid : ακέραιο αναγνωριστικό (id) της εγγραφής στη βάση δεδομένων geonames
name : όνομα γεωγραφικού σημείου (utf8), varchar(200)
asciiname : όνομα γεωγραφικού σημείου με απλούς χαρακτήρες ascii, varchar(200)
alternatenames : εναλλακτικά ονόματα, διαχωρισμένα με κόμματα, ονόματα ascii που μεταγράφονται αυτόματα, χαρακτηριστικό convenience από τον πίνακα, alternatename table,varchar(10000)
latitude : γεωγραφικό πλάτος σε δεκαδικούς βαθμούς (wgs84)
longitude : γεωγραφικό μήκος: γεωγραφικό μήκος σε δεκαδικούς βαθμούς (wgs84)
feature class : βλέπε http://www.geonames.org/export/codes.html, char(1)
feature code : βλέπε http://www.geonames.org/export/codes.html, varchar(10)
country code : ISO-3166 Κωδικός χώρας 2 γραμμάτων, 2 χαρακτήρες
cc2 : εναλλακτικοί κωδικοί χωρών, διαχωρισμένοι με κόμματα, κωδικός χώρας ISO-3166 2 γραμμάτων, 200 χαρακτήρες
admin2 code : κωδικός για το δεύτερο διοικητικό τμήμα, μια κομητεία στις ΗΠΑ, βλέπε αρχείο admin2Codes.txt. varchar(80)
admin3 code : κωδικός για διοικητική διαίρεση τρίτου επιπέδου, varchar(20)
admin4 code : κωδικός για διοικητική διαίρεση τέταρτου επιπέδου, varchar(20)
population : bigint (8 byte int)
elevation : υψόμετρο ,bigint (8 byte int)
dem : ψηφιακό υψομετρικό μοντέλο, srtm3 ή gtopo30, μέσο υψόμετρο {3x3}(περίπου 90mx90m) ή {30x30}(περίπου 900mx900m) επιφάνεια σε μέτρα, ακέραιος αριθμός. srtm επεξεργασία από cgiar/ciat.
timezone : το αναγνωριστικό ζώνης ώρας iana (δείτε αρχείο timeZone.txt) varchar(40)
modification date : ημερομηνία τελευταίας τροποποίησης σε μορφή εεεε-ΜΜ-ηη

Πίνακας 14: Πεδία του πίνακα "geonname

4. Queries

Παρακάτω γίνεται συλλογή των queries και παρουσίαση τους, ώστε να απεικονιστεί γραφικά η σημασία του Σημασιολογικού Ιστού μέσω της χρήσης της γεωγραφικής βάσης Geonames.Συγκεκριμένα, έγινε προσπάθεια δημιουργίας και μελέτης queries σε διαφορετικές γλώσσες προγραμματισμού (json,python,sql) προκειμένου να αναδειχτεί η ικανότητα χρήσης

της βάσης από διάφορες γλώσσες προκειμένου να συλλεχθούν τα γεωγραφικά δεδομένα που χρειαζόμαστε κάθε φορά.

4.1 Queries in sparql

```
PREFIX gn: <http://www.geonames.org/ontology#>

PREFIX wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>

PREFIX dcterms: <http://purl.org/dc/terms/>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?country ?population ?lattitude ?longitude ?creationDate
?modificationDate

WHERE {

?country gn:name ?country_name ;

        gn:population ?population ;

        wgs84_pos:lat ?latitude ;

        wgs84_pos:long ?longitude ;

        rdfs:isDefinedBy ?countryDocument .

?countryDocument dcterms:created ?creationDate ;

                dcterms:modified ?modificationDate .}
```

Παράδειγμα για τις United States:

- https://sws.geonames.org/6252001/ (concept, i.e., the actual country)
- https://sws.geonames.org/6252001/about.rdf (GeoNames document about this concept)

Αυτά τα URIs αναφέρονται μεταξύ τους μέσω του rdfs:isDefinedBy/foaf:primaryTopic.
Η ιδιότητα dcterms:created δίνει την ημερομηνία δημιουργίας του εγγράφου και όχι την ημερομηνία δημιουργίας της χώρας. Το ίδιο ισχύει και για το dcterms:modified.
Τρέχοντας το παραπάνω query για συγκεκριμένη χώρα παίρνουμε ως αποτέλεσμα έναν πίνακα 5x5 που για τη συγκεκριμένη χώρα μας δίνει δεδομένα που αφορούν το όνομα της χώρας, τον πληθυσμό της και τις γεωγραφικές συντεταγμένες της (latitude, longitude).

Εάν δεν χρειάζεται να ανατρέξουμε στο "έγγραφο σχετικά με το concept" στα αποτελέσματα, μπορούμε να χρησιμοποιήσουμε συγκεκριμένα paths για να λάβουμε τις τιμές:

```
PREFIX gn: <http://www.geonames.org/ontology#>

PREFIX wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>

PREFIX dcterms: <http://purl.org/dc/terms/>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?country ?population ?lattitude ?longitude ?creationDate
?modificationDate

WHERE {

?country gn:name ?country_name ;

        gn:population ?population ;

        wgs84_pos:lat ?lattitude ;

        wgs84_pos:long ?longitude ;

        rdfs:isDefinedBy/dcterms:created ?creationDate ;

        rdfs:isDefinedBy/dcterms:modified ?modificationDate .}
```

Διαφορετικά μπορούμε να χρησιμοποιήσουμε μια άλλη μεταβλητή για αυτό:

```
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?country ?population ?lattitude ?longitude ?creationDate
?modificationDate
WHERE {

?country gn:name ?country_name ;
        gn:population ?population ;
        wgs84_pos:lat ?lattitude ;
        wgs84_pos:long ?longitude ;
        rdfs:isDefinedBy ?countryDocument .

?countryDocument dcterms:created ?creationDate ;
                dcterms:modified ?modificationDate .
}
```

Εάν θέλουμε να εστιάσουμε εκτός από την χώρα-περιοχή σε συγκεκριμένα αξιοθέατα λόγου χάρη {Μουσεία} μπορούμε να τρέξουμε το παρακάτω script.

Χρησιμοποιούμε το παρακάτω query για να εντοπίσουμε από την dbpedia γεωγραφική βάση τα Μουσεία που υπάρχουν και τρέχοντάς τα εντοπίζουμε.

Η dbpedia ένα project για την εξαγωγή, διασύνδεση και επαναχρησιμοποίηση δομημένης πληροφορίας διαμέσου του Web από την Wikipedia. Τα δεδομένα που αντλούνται μπορούν να είναι αντικείμενο επεξεργασίας από λογισμικό και να διασυνδεθούν με οποιοδήποτε τρόπο, προσφέροντας δυνατότητες για διάφορες άλλες εφαρμογές.

```
SELECT DISTINCT ?Museum
  (SAMPLE(?name) as ?name)
  (SAMPLE(?abstract) as ?abstract)
  (SAMPLE(?thumbnail) as ?thumbnail)
  (MAX(?latitude) as ?latitude)
  (MAX(?longitude) as ?longitude)
  (SAMPLE(?photoCollection) as ?photoCollection)
  (SAMPLE(?website) as ?website)
  (SAMPLE(?homepage) as ?homepage)
  (SAMPLE(?wikilink) as ?wikilink)
WHERE {
  ?Museum a dbpedia-owl:Museum ;
        dbpprop:name ?name ;
        dbpedia-owl:abstract ?abstract ;
        dbpedia-owl:thumbnail ?thumbnail ;
        geo:lat ?latitude ;
        geo:long ?longitude ;
        dbpprop:hasPhotoCollection ?photoCollection ;
        dbpprop:website ?website ;
        foaf:homepage ?homepage ;
        foaf:isPrimaryTopicOf ?wikilink .
  FILTER(langMatches(lang(?abstract),"EN"))
  FILTER (langMatches(lang(?name),"EN"))
}
GROUP BY ?Museum
LIMIT 20
```

Τα queries που κατασκευάσαμε μπορούμε να τα τρέξουμε σε έναν online editor, όπως: http://geosparql.org/.

Αν θέλουμε να εμπλουτίσουμε το query, χρησιμοποιώντας ως clause το NEARBY, εστιάζουμε ως εξής:

```
PREFIX spatial:<http://jena.apache.org/spatial#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX geo:<http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX gn:<http://www.geonames.org/ontology#>

Select *
WHERE{
?object spatial:nearby(40.74 -73.989 1 'mi').
?object rdfs:label ?label
}LIMIT 10
```

Αν έχουμε μια εκτενή λίστα με αναγνωριστικά Geonames IDs για τα οποία θέλουμε να βρούμε τα αντίστοιχα αναγνωριστικά Wikidata, μπορούμε να χρησιμοποιήσουμε το Pywikibot που είναι μία βιβλιοθήκη της Python.

Συγκεκριμένα,το query σε SPARQL για μοναδικά Geonames ID είναι το παρακάτω:

```
SELECT DISTINCT ?item ?itemLabel WHERE {
  SERVICE wikibase:label { bd:serviceParam wikibase:language "de". }
  {
    SELECT DISTINCT ?item WHERE {
      ?item p:P1566 ?statement0.
      ?statement0 (ps:P1566) "2867714".
    }
  }
}
```

Για παράδειγμα για το Μόναχο- Munich το Geonames ID είναι 2867714 και τρέχοντας το query μας επιστρέφει το σωστό Wikidata ID, το οποίο είναι το **wikidata:Q32664319**:

```
import pywikibot
from pywikibot import pagegenerators as pg

# read query file

with open('C:\\Users\\p70076654\\Downloads\\SPARQL_mapGeonamesID.rq', 'r') as
query_file:
    QUERY = query_file.read()
    #print(QUERY)

# create generator based on query
# returns an iterator that produces a sequence of values when iterated over
# useful when creating large sequences of values

wikidata_site = pywikibot.Site("wikidata", "wikidata")
generator = pg.WikidataSPARQLPageGenerator(QUERY, site=wikidata_site)

print(generator)

# OUTPUT: <generator object WikidataSPARQLPageGenerator.<locals>.<genexpr> at
0x00000169FAF3FD10>

# iterate over generator

for item in generator:
    print(item)
```

Η βάση GeoNames, όπως αναφέραμε παρέχει πληροφορίες για πάνω από οκτώ εκατομμύρια γεωχωρικά χαρακτηριστικά. Τα δεδομένα εκτίθενται μέσω ενός RDF webservice που παρέχει πληροφορίες για κάθε ένα δεδομένο σύμφωνα με τους πόρους που διαθέτει στη βάση του. Ωστόσο, δεν υπάρχει δυνατότητα εκτέλεσης οποιουδήποτε τύπου query σε SPARQL για την ανάκτηση δεδομένων.

Μια άλλη σημαντική πηγή γεωχωρικών δεδομένων πέρα από τη GeoNames είναι η βάση DBPedia. Πολλές από αυτές τις εξαγόμενες οντότητες είναι γεωχωρικού χαρακτήρα (πόλεις, κομητείες, χώρες, ορόσημα, κ.λπ.) και πολλές από αυτές τις οντότητες περιέχουν ήδη κάποιες πληροφορίες γεωχωρικής θέσης. Στην πραγματικότητα, η DBPedia αντλεί αρκετά δεδομένα από τη GeoNames και φτάνει στο σημείο να περιλαμβάνει τις πληροφορίες γεωγραφικού πλάτους και μήκους για πολλές οντότητες. Παρέχει επίσης owl:sameAs συνδέσεις μεταξύ των πόρων του DBPedia και του GeoNames.

Η κοινότητα των συνδεδεμένων δεδομένων έχει κυκλοφορήσει το σύνολο συνδεδεμένων Γεωδεδομένων. Αυτό το σύνολο δεδομένων είναι μια βάση αιχμής χωρικής γνώσης, που προέρχεται από το Open Street Map και συνδέεται με πόρους από τις δύο βάσεις , DBpedia και GeoNames. Περιέχει πάνω από 200 εκατομμύρια τριάδες που περιγράφουν τους κόμβους και τα μονοπάτια από το OpenStreetMap. Το σύνολο δεδομένων είναι προσβάσιμο μέσω των

τερματικών σημείων SPARQL που εκτελούνται στην πλατφόρμα OpenLink Virtuoso με όλα τα πλεονεκτήματα και τους περιορισμούς που αυτή η πλατφόρμα παρέχει.

Μια άλλη πηγή γεωχωρικών δεδομένων είναι η United States Geological Survey (USGS). Η USGS έχει εκμισθώσει ένα SPARQL endpoint για τριπλέτες δεδομένων που προέρχονται από τον Εθνικό Χάρτη(National Map), η οποία είναι μια συλλογική προσπάθεια για χρήσιμες τοπογραφικές πληροφορίες για τις Ηνωμένες Πολιτείες. Αυτό το σύνολο δεδομένων είναι πολύ πιο συγκεκριμένο και πιο ειδικό από τα δεδομένα που παρέχονται από τη DBPedia και το GeoNames. Περιλαμβάνει γεωγραφικά ονόματα, υδρογραφία, όρια, μέσα μεταφοράς, δομές και τον τρόπο με τον οποίο καλύπτεται εκεί η γη , όπως λόγου χάρη τα όρη που περιλαμβάνει.

Λόγω έλλειψης διαθέσιμων τριπλών καταστημάτων GeoSPARQL, το δημοσιευμένο σύνολο δεδομένων περιλαμβάνει έναν προ-υπολογισμό όλων των τοπολογικών σχέσεων μεταξύ οντοτήτων. Ένα ερώτημα όπως « Η Εμφάνιση όλων των σιδηροδρομικών γραμμών που διασχίζουν τα ποτάμια» είναι μάλιστα δυνατό να απαντηθεί κοιτάζοντας τα τρέχοντα προυπολογισμένα δεδομένα. Ωστόσο, αυτό σημαίνει ότι εάν προστεθεί μια νέα οντότητα, η βάση γνώσεων πρέπει να υπολογίσει όλα όσα σχετίζονται με την οντότητα και να ενημερώσει και αυτές τις οντότητες. Αν αυτά τα δεδομένα δεν ήταν προυπολογισμένα, ο μόνος τρόπος να απαντηθεί το ερώτημα είναι μέσω της ευρετηρίασης των δεδομένων και της αναζήτησης των σχέσεων με ένα κατηγόρημα σχέσης.

Το GeoSPARQL παρέχει τα μέσα για τη σύνδεση όλων των γεωχωρικών δεδομένων και εξάγει περισσότερα αποτελέσματα από το σχέσεις που αναπτύσσονται μεταξύ των δεδομένων. Καθώς δεν είναι δυνατό να προ-υπολογιστούν όλες οι σχέσεις μεταξύ όλων των διαθέσιμων γεωχωρικών δεδομένων, εμπλουτίζοντας τα υπάρχοντα σύνολα δεδομένων με αναπαραστάσεις GeoSPARQL και μέσω της δημιουργίας ευρετηρίων για τα δεδομένα δημιουργείται έτσι ένας τρόπος με τον οποίο οι πάροχοι δεδομένων μπορούν να μοιράζονται διαρκώς δεδομένα παρέχοντας έτσι παράλληλα πρόσβαση σε γεωχωρικά δεδομένα.

Για τα παραδείγματα που ακολουθούν, έχουμε επεξεργαστεί ένα υποσύνολο του συνόλου δεδομένων GeoNames RDF24 και τα δεδομένα του USGS GeoSPARQL για την Ατλάντα. Αυτά τα δεδομένα είναι προσβάσιμα μέσω ενός τελικού σημείου SPARQL του Εθνικού Χάρτη με έναν χωρικό δείκτη GeoSPARQL25. Ο χωρικός αυτός δείκτης GeoSPARQL υποστηρίζει δεδομένα ευρετηρίασης σύμφωνα με το WKT και σειριοποίηση GML. Τα υποστηριζόμενα GeoSPARQL queries περιλαμβάνουν σχέσεις Simple Features, Egen hofer και RCC8, μη τοπολογικές συναρτήσεις και κοινές τοπολογικές συναρτήσεις των queries.

Καθώς τόσα πολλά δεδομένα αντιπροσωπεύονται ως μεμονωμένα δεδομένα γεωγραφικού πλάτους και μήκους χρησιμοποιώντας το W3C Basic Geo λεξιλόγιο, συμπεριλαμβανομένου

αυτού που παρέχεται από το GeoNames, είναι αναγκαίο να μπορούμε να το μετατρέψουμε εύκολα σε GeoSPARQL.

Προσπαθώντας να απαντήσουμε στην ερώτηση:

«Ποια μνημεία περιέχονται μέσα σε πάρκο;» Κατασκευάσαμε τα παρακάτω queries.

Queries σε Geosparql

```
SELECT ?m ?p WHERE { ?m a ex:Monument ; geo:hasGeometry ?mgeo . ?p a ex:Park ;
geo:hasGeometry ?pgeo . ?mgeo geo:within ?pgeo . }
```

Αυτό το ερώτημα απαιτεί μια τοπολογική σύγκριση μεταξύ των συντεταγμένων των μνημείων και των συντεταγμένων των πάρκων. Το δείχνουμε στο παραπάνω query χρησιμοποιώντας την ιδιότητα της δυαδικής τοπολογίας geo:within. Οι δύο οντότητες έχουν δηλώσεις τύπου, geo:hasGeometry, ιδιότητες για να τα δέσουν με τις συντεταγμένες τους, και μετά το geo:within συνάρτησης για να τα συνδέσουμε μεταξύ τους.

Μπορούμε να ξαναγράψουμε και να τρέξουμε το παραπάνω query ακόμα πιο απλά χρησιμοποιώντας μια τοπολογική σχέση χαρακτηριστικού προς χαρακτηριστικό feature-to-feature. Αυτή η μέθοδος παρουσιάζεται στο επόμενο query.

```
SELECT ?m ?p WHERE {
?p a ex:Park . ?m a ex:Monument ; geo:within ?p . }
```

Οι χωρικές συντεταγμένες που πρέπει ο χρήστης συχνά να αναζητά οντότητες συγκεκριμένου τύπου που εμπίπτουν σε ένα πλαίσιο. Για παράδειγμα αν θέλουμε να βρούμε :

Τι αξιοθέατα βρίσκονται μέσα στο πλαίσιο οριοθέτησης που ορίζεται από (-77.089005, 38.913574) αυτές τις συντεταγμένες;»

```
SELECT ?m ?p WHERE { ?p a ex:Park .
?m a ex:Monument ; geo:within ?p . }
```

Για να εντοπίσει η βάση το σημείο που ενδιαφέρει το χρήστη «αντιμετωπίζει» το σημείο αναζήτησης ως πολύγωνο και έτσι δίνει το αποτέλεσμα.

```
SELECT ?a WHERE { ?a a ex:Attraction; geo:hasGeometry ?ageo . FILTER(geof:within(?ageo, "POLYGON(( -77.089005 38.913574, -77.029953 38.913574, -77.029953 38.886321, -77.089005 38.886321, -77.089005 38.913574 ))"^^geo-sf:WKTLiteral)) }
```

Με την ίδια λογική βρίσκουμε για παράδειγμα την απάντηση στο παρακάτω ερώτημα: «Ποια πάρκα είναι εντός των 3 χιλιόμετρων από το μνημείο της Ουάσιγκτον;»

Το αποτέλεσμα δίνεται αφού ανακτούμε τις συντεταγμένες και χρησιμοποιήσουμε τη συνάρτηση geof:distance για τον υπολογισμό της απόστασης μεταξύ τους. Στη συνέχεια εφαρμόζεται μια τυπική συνάρτηση SPARQL μικρότερη από αυτή που χρησιμοποιήθηκε. Αυτή η λογική παρουσιάζεται ακριβώς στο παρακάτω query:

```
SELECT ?p WHERE { ?p a ex:Park ; geo:hasGeometry ?pgeo . ex:WashingtonMonument geo:hasGeometry ?wgeo . FILTER(geof:distance(?pgeo, ?wgeo, units:m) < 3000) }
```

4.2 Query in Python

Η δημιουργία ενός script για όλες τις διοικητικές διαιρέσεις – ονόματα και κωδικούς και πώς φωλιάζουν τα τμήματα μεταξύ τους – για όλες τις χώρες της γαλλόφωνης Καραϊβικής (χώρες που είναι αυτή τη στιγμή, ή παλαιότερα, υπερπόντια εδάφη της Γαλλίας) σε γλώσσα Python απεικονίζεται στο παρακάτω script:

```
import requests,csv
from time import strftime
ccodes=['BL','DM','GD','GF','GP','HT','KN','LC','MF','MQ','VC']
fclass='A'
lang='fr'
uname='REQUEST FROM GEONAMES'
#Columns to keep
fields=['countryId','countryName','countryCode','geonameId','name','asciiName',
       'alternateNames','fcode','fcodeName','adminName1','adminCode1',
       'adminName2','adminCode2','adminName3','adminCode3','adminName4','adminCode4',
       'adminName5','adminCode5','lng','lat']
fcode=fields.index('fcode')
#Divisions to keep
divisions=['ADM1','ADM2','ADM3','ADM4','ADM5','PCLD','PCLF','PCLI','PCLIX','PCLS']
base_url='http://api.geonames.org/searchJSON?'
def altnames(names,lang):
    "Given a dict of names, extract preferred names for a given language"
    aname=''
    for entry in names:
        if 'isPreferredName' in entry.keys() and entry['lang']==lang:
            aname=entry.get('name')
        else:
            pass
    return aname
places=[]
tossed=[]
for country in ccodes:
 data_url =f'{base_url}?name=*&country={country}&featureClass={fclass}&lang={lang}&style=full&username={uname}'
    response=requests.get(data_url)
```

```
  data=response.json() #total retrieved and results in list of dicts
  gnames=response.json()['geonames'] #create list of dicts only
  gnames.sort(key=lambda e: (e.get('countryCode',''),e.get('fcode',''),
                        e.get('adminCode1',''),e.get('adminCode2',''),
                        e.get('adminCode3',''),e.get('adminCode4',''),
                        e.get('adminCode5','')))
  for record in gnames:
     r=[]
     for f in fields:
        item=record.get(f,'')
        if f=='alternateNames' and f !='':
           aname=altnames(item,'en')
           r.append(aname)
        else:
           r.append(item)
     if r[fcode] in divisions: #keep certain admin divs, toss others
        places.append(r)
     else:
        tossed.append(r)

filetoday=strftime('%Y_%m_%d')
outfile='geonames_fwi_adm_'+filetoday+'.csv'

writefile=open(outfile,'w', newline='', encoding='utf8')
writer=csv.writer(writefile, delimiter=',', quotechar='',quoting=csv.QUOTE_NONNUMERIC)
writer.writerow(fields) #header row
writer.writerows(places)
writefile.close()
print(len(places),'records written to file',outfile)
```

Αρχικά, προσδιορίζουμε όλες τις μεταβλητές που χρειάζόμαστε: τους κωδικούς ISO δύο γραμμάτων των χωρών, μια λίστα με τα χαρακτηριστικά Geonames που θέλουμε να διατηρήσουμε, τον κωδικό γλώσσας δύο γραμμάτων και τον συγκεκριμένο τύπο χαρακτηριστικού που με ενδιαφέρει. Υπάρχουν διαφορετικοί κωδικοί χαρακτηριστικών που ταξινομούνται με ένα μόνο γράμμα και ένας αριθμός υποτύπων κάτω από αυτό. Η κλάση δυνατοτήτων A είναι για εγγραφές που αντιπροσωπεύουν διοικητικές υποδιαιρέσεις και σε αυτήν την κατηγορία χρειαζόμαστε εγγραφές που αντιπροσώπευαν τη χώρα ως σύνολο (κωδικοί PCL) και τις υποδιαιρέσεις της (κωδικοί ADM). Υπάρχουν πολλές διαφορετικές μεταβλητές τοπωνυμίων που περιλαμβάνουν επίσημα ονόματα, σύντομες φόρμες και μια φόρμα ASCII που περιλαμβάνει μόνο χαρακτήρες που βρίσκονται στο λατινικό αλφάβητο που χρησιμοποιείται στα αγγλικά. Ο κωδικός γλώσσας που μεταβιβάζετε στη διεύθυνση url θα αλλάξει αυτά τα αποτελέσματα.

Χρησιμοποιούμε την αναζήτηση πλήρους κειμένου Geonames, όπου πραγματοποιείτε αναζήτηση για χαρακτηριστικά με το όνομα (τα ξεχωριστά API για εργασία με ιεραρχίες για γονικές και θυγατρικές οντότητες είναι μια άλλη επιλογή). Χρησιμοποίησαμε έναν αστερίσκο ως μπαλαντέρ για να ανακτήσουμε όλα τα ονόματα και τις άλλες παραμέτρους για να φιλτράρω για συγκεκριμένες χώρες και κατηγορίες χαρακτηριστικών. Στο τέλος του βασικού url πρόσθεσαμε JSON για την αναζήτηση. Εάν το αφήσουμε εκτός λειτουργίας, οι εγγραφές επιστρέφονται ως XML.

Ο κύριος βρόχος for βρίσκεται σε κάθε χώρα και μεταβιβάζει τις παραμέτρους στη διεύθυνση url δεδομένων για να ανακτήσει τα δεδομένα για αυτήν τη χώρα: μεταβιβάζουμε τον κωδικό χώρας, την κατηγορία χαρακτηριστικών A και τα γαλλικά ως γλώσσα για τα τοπωνύμια. Επίσης, προσθέτουμε style=full για να ανακτήσουμε όλες τις πιθανές πληροφορίες που είναι διαθέσιμες για μια δεδομένη εγγραφή. Η προεπιλογή είναι να καταγράψουμε ένα υποσύνολο βασικών πληροφοριών, το οποίο δεν είχε τους κωδικούς διαχειριστή που χρειαζόμασταν.

Χρησιμοποιούμε τη λειτουργική μονάδα Python Requests για να αλληλεπιδράσουμε με το API. Η Geonames επιστρέφει δύο αντικείμενα στο JSON: έναν ακέραιο αριθμό των συνολικών εγγραφών που ανακτήθηκαν και ένα άλλο αντικείμενο JSON που ουσιαστικά αντιπροσωπεύει μια λίστα λεξικών python, όπου κάθε λεξικό περιέχει όλα τα χαρακτηριστικά μιας εγγραφής ως μια σειρά από ζεύγη κλειδιών και τιμών όπου το κλειδί είναι το όνομα του χαρακτηριστικού.

Δημιουργούμε μια νέα μεταβλητή gnames για να απομονώσουμε μόνο αυτήν τη λίστα και, στη συνέχεια, ταξινομούμε τη λίστα με βάση το πώς θέλουμε να εμφανίζεται η τελική έξοδος ανά χώρα και κατά κωδικούς διαχειριστή, έτσι ώστε παρόμοια επίπεδα κωδικών διαχειριστή να ομαδοποιούνται μαζί. Το κόλπο της χρήσης lamba για την ταξινόμηση ένθετων λιστών ή λεξικών είναι καλά τεκμηριωμένο, αλλά μια παραλλαγή ήταν να χρησιμοποιήσουμε τη μέθοδο λήψης λεξικού. Ορισμένες λειτουργίες ενδέχεται να μην έχουν πέντε επίπεδα κωδικών διαχειριστή. Εάν δεν το κάνουν, τότε δεν υπάρχει κλειδί για αυτό το χαρακτηριστικό και χρησιμοποιώντας την απλή προσέγγιση dict[key] επιστρέφει ένα σφάλμα σε αυτές τις περιπτώσεις. Η χρήση του dict.get(key") επιτρέπει να μεταβιβάσουμε μια προεπιλεγμένη τιμή εάν δεν υπάρχει κλειδί. Παρέχουμε μια κενή συμβολοσειρά ως σύμβολο κράτησης θέσης, καθώς τελικά θέλουμε κάθε εγγραφή να έχει τον ίδιο αριθμό στηλών στην έξοδο και να ευθυγραμμιστεί σωστά.

gnames - List (8 elements)

nde	Type	Size	Value
0	dict	27	{'timezone':{'gmtOffset':0, 'timeZoneId':'Europe/Londo…
1	dict	30	{'timezone':{'gmtOffset':-4, 'timeZoneId':'America/St_…
2	dict	30	{'timezone':{'gmtOffset':-4, 'timeZoneId':'America/St_…
3	dict	30	{'timezone':{'gmtOffset':-4, 'timeZoneId':'America/St_…
4	dict	30	{'timezone':{'gmtOffset':-4, 'timeZoneId':'America/St_…
5	dict	30	{'timezone':{'gmtOffset':-4, 'timeZoneId':'America/St_…
6	dict	30	{'timezone':{'gmtOffset':-4, 'timeZoneId':'America/St_…
7	dict	27	{'timezone':{'gmtOffset':-4, 'timeZoneId':'America/St_…

Εικόνα 8: Εγγραφές που επιστράφηκαν από τη Geonames ως λίστα, όπου κάθε στοιχείο λίστας είναι ένα λεξικό ζευγών κλειδιών/τιμών για ένα δεδομένο μέρος

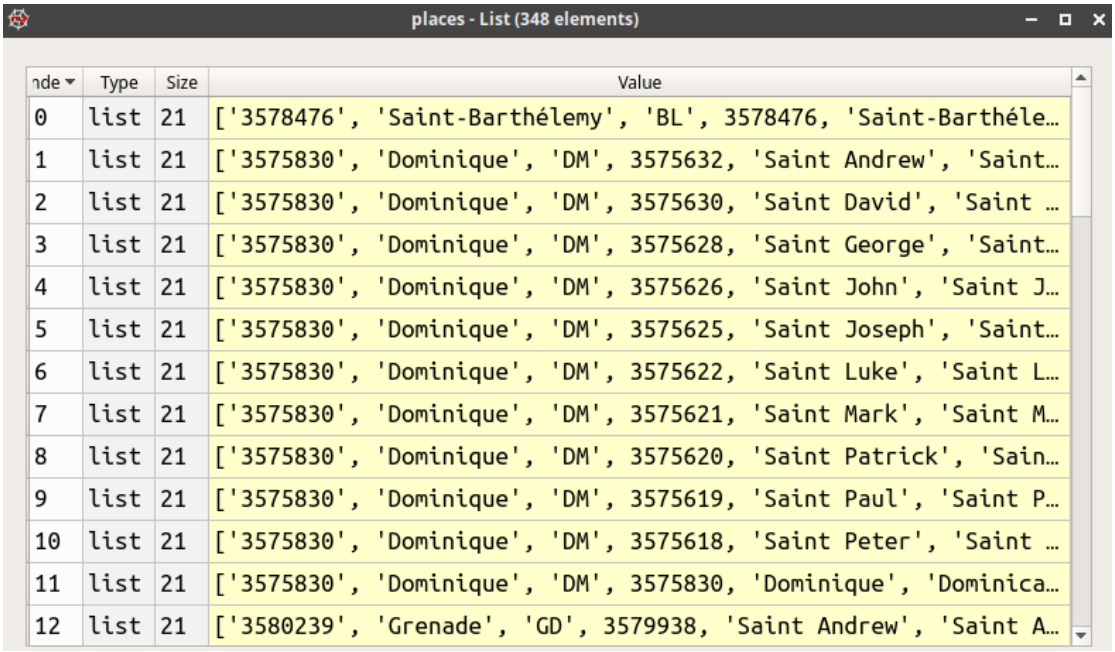
Πηγή: https://www.geonames.org/

Key	Type	Size	Value
asciiName	str	1	Parish of Charlotte
astergdem	int	1	264
bbox	dict	5	{'east':-61.123931443999936, 'south':13.…
continentCode	str	1	NA
countryCode	str	1	VC
countryId	str	1	3577815
countryName	str	1	Saint-Vincent-et-les-Grenadines
fcl	str	1	A
fclName	str	1	country, state, region,...
fcode	str	1	ADM1
fcodeName	str	1	first-order administrative division
geonameId	int	1	3577934
lat	str	1	13.25064
lng	str	1	-61.15196
name	str	1	Charlotte
population	int	1	38000

Εικόνα 9: Παράδειγμα μεμονωμένου στοιχείου λίστας, λεξικό ζευγών κλειδιών/τιμών για την Ενορία της Σάρλοτ, ένα τμήμα διαχείρισης 1ης τάξης του Saint-Vincent-et-les-Grenadines. Τα ονόματα των μεταβλητών είναι κλειδιά.

Πηγή: https://www.geonames.org/

Μόλις έχουμε εγγραφές για την πρώτη χώρα, τις κάνουμε κύκλο και επιλέγουμε μόνο τα χαρακτηριστικά που θέλουμε από τη λίστα πεδίων. Το όνομα του χαρακτηριστικού είναι το κλειδί, παίρνουμε τη συσχετισμένη τιμή, αλλά αν αυτό το κλειδί δεν υπάρχει, εισάγουμε μια κενή συμβολοσειρά. Στις περισσότερες περιπτώσεις, η τιμή που συσχετίζεται με ένα κλειδί είναι μια συμβολοσειρά ή ακέραιος αριθμός, αλλά σε μερικές περιπτώσεις είναι ένα άλλο κοντέινερ, όπως στην περίπτωση των εναλλακτικών ονομάτων που είναι μια άλλη λίστα λεξικών. Εάν υπάρχουν εναλλακτικά ονόματα, θέλουμε να βγάλουμε ένα προτιμώμενο όνομα στα αγγλικά, εάν υπάρχει. Το χειριζόμαστε αυτό με μια συνάρτηση, ώστε ο βρόχος να φαίνεται λιγότερο ακατάστατος. Τέλος, εάν αυτή η εγγραφή αντιπροσωπεύει ένα τμήμα διαχειριστή ή είναι μια εγγραφή σε επίπεδο χώρας, τότε τη διατηρούμε, διαφορετικά την προσαρτούμε σε μια γρήγορη λίστα που θα επιθεωρήσουμε αργότερα.



places - List (348 elements)

nde	Type	Size	Value
0	list	21	['3578476', 'Saint-Barthélemy', 'BL', 3578476, 'Saint-Barthéle…
1	list	21	['3575830', 'Dominique', 'DM', 3575632, 'Saint Andrew', 'Saint…
2	list	21	['3575830', 'Dominique', 'DM', 3575630, 'Saint David', 'Saint …
3	list	21	['3575830', 'Dominique', 'DM', 3575628, 'Saint George', 'Saint…
4	list	21	['3575830', 'Dominique', 'DM', 3575626, 'Saint John', 'Saint J…
5	list	21	['3575830', 'Dominique', 'DM', 3575625, 'Saint Joseph', 'Saint…
6	list	21	['3575830', 'Dominique', 'DM', 3575622, 'Saint Luke', 'Saint L…
7	list	21	['3575830', 'Dominique', 'DM', 3575621, 'Saint Mark', 'Saint M…
8	list	21	['3575830', 'Dominique', 'DM', 3575620, 'Saint Patrick', 'Sain…
9	list	21	['3575830', 'Dominique', 'DM', 3575619, 'Saint Paul', 'Saint P…
10	list	21	['3575830', 'Dominique', 'DM', 3575618, 'Saint Peter', 'Saint …
11	list	21	['3575830', 'Dominique', 'DM', 3575830, 'Dominique', 'Dominica…
12	list	21	['3580239', 'Grenade', 'GD', 3579938, 'Saint Andrew', 'Saint A…

Εικόνα 10: Τελική λίστα που περιέχει εγγραφές για όλα τα τμήματα διαχειριστή για συγκεκριμένες χώρες και κατηγορίες χαρακτηριστικών, όπου τα στοιχεία είναι υπολίστες που αντιπροσωπεύουν κάθε μέρος

Πηγή: https://www.geonames.org/

Συνολικά η προσέγγιση αυτή λειτούργησε καλά, αλλά υπάρχουν κάποιες μικρές επιφυλάξεις. Ορισμένες από τις χώρες που μελετάμε δεν είναι ανεξάρτητες, αλλά εξαρτώνται από τη Γαλλία. Για τις εξαρτημένες χώρες, οι κωδικοί υποδιαίρεσης 1ου και μερικές φορές ακόμη και 2ου επιπέδου εμφανίζονται πανομοιότυποι με τον κωδικό χώρας ανώτατου επιπέδου, καθώς αντιπροσωπεύουν μια υποδιαίρεση μιας ανεξάρτητης χώρας (πολλά υπερπόντια εδάφη είναι διαμερίσματα της Γαλλίας). Εάν χρειαστεί να εναρμονίσουμε αυτούς τους κωδικούς μεταξύ των χωρών, ίσως χρειαστεί να προσαρμόσουμε τις εξαρτήσεις. Τα εναλλακτικά αγγλικά ονόματα μερών εμφανίζονται πάντα για την εγγραφή σε επίπεδο χώρας, αλλά συνήθως όχι κάτω από αυτό. Νομίζω ότι θα έπρεπε να κάνουμε κάποιες πρόσθετες τροποποιήσεις ή ακόμη και να εκτελέσουμε ένα δεύτερο σύνολο αιτημάτων στα Αγγλικά, αν θέλαμε όλα τα αγγλικά ορθογραφικά. Για παράδειγμα στα γαλλικά πολλά σύνθετα τοπωνύμια όπως το Saint-Paul χωρίζονται με παύλα, αλλά στα αγγλικά χωρίζονται με κενό. Δεν είναι κάτι σπουδαίο καθώς μας ενδιέφεραν πρωτίστως η εναλλακτική ορθογραφία των χωρών, δηλαδή η Γουιάνα έναντι της Γαλλικής Γουιάνας.

Το τελικό αποτέλεσμα παρακάτω για τη Γουιάνα είναι το εξής:

	A	B	C	D	E	F	G	H	I
1	countryId	countryName	countryCode	geonameId	name	asciiName	alternateNames	fcode	fcodeName
22	3381670	Guyane	GF	6690605	Guyane	Guyane	Guyane	ADM1	first-order administrative division
23	3381670	Guyane	GF	6690606	Guyane	Guyane		ADM2	second-order administrative division
24	3381670	Guyane	GF	3382159	Arrondissement de Cayenne	Arrondissement de Cayenne		ADM3	third-order administrative division
25	3381670	Guyane	GF	3380386	Arrondissement de Saint-Laurent-du-Maroni	Arrondissement de Saint-Laurent-du-Maroni		ADM3	third-order administrative division
26	3381670	Guyane	GF	6690700	Régina	Regina		ADM4	fourth-order administrative division
27	3381670	Guyane	GF	6690689	Cayenne	Cayenne		ADM4	fourth-order administrative division
28	3381670	Guyane	GF	6690691	Iracoubo	Iracoubo		ADM4	fourth-order administrative division
29	3381670	Guyane	GF	6690692	Kourou	Kourou		ADM4	fourth-order administrative division
30	3381670	Guyane	GF	6690693	Macouria	Macouria		ADM4	fourth-order administrative division
31	3381670	Guyane	GF	6690696	Matoury	Matoury		ADM4	fourth-order administrative division
32	3381670	Guyane	GF	6690704	Saint-Georges	Saint-Georges		ADM4	fourth-order administrative division
33	3381670	Guyane	GF	6690701	Remire-Montjoly	Remire-Montjoly		ADM4	fourth-order administrative division
34	3381670	Guyane	GF	6690702	Roura	Roura		ADM4	fourth-order administrative division
35	3381670	Guyane	GF	6690707	Sinnamary	Sinnamary		ADM4	fourth-order administrative division
36	3381670	Guyane	GF	6690697	Montsinéry-Tonnegrande	Montsinery-Tonnegrande		ADM4	fourth-order administrative division
37	3381670	Guyane	GF	6690698	Ouanary	Ouanary		ADM4	fourth-order administrative division
38	3381670	Guyane	GF	6690688	Camopi	Camopi		ADM4	fourth-order administrative division
39	3381670	Guyane	GF	6690703	Saint-Élie	Saint-Elie		ADM4	fourth-order administrative division
40	3381670	Guyane	GF	6690694	Mana	Mana		ADM4	fourth-order administrative division
41	3381670	Guyane	GF	6690705	Saint-Laurent-du-Maroni	Saint-Laurent-du-Maroni		ADM4	fourth-order administrative division
42	3381670	Guyane	GF	6690706	Saül	Sauel		ADM4	fourth-order administrative division
43	3381670	Guyane	GF	6690695	Maripasoula	Maripasoula		ADM4	fourth-order administrative division
44	3381670	Guyane	GF	6690690	Grand-Santi	Grand-Santi		ADM4	fourth-order administrative division
45	3381670	Guyane	GF	6690686	Apatou	Apatou		ADM4	fourth-order administrative division
46	3381670	Guyane	GF	6690687	Awala-Yalimapo	Awala-Yalimapo		ADM4	fourth-order administrative division
47	3381670	Guyane	GF	6690699	Papaichton	Papaichton		ADM4	fourth-order administrative division
48	3381670	Guyane	GF	3381670	Guyane	Guyane	French Guiana	PCLD	dependent political entity

Εικόνα 11: Το 1ο μισό του αρχείου CSV εισήχθη σε υπολογιστικό φύλλο, εγγραφές που δείχνουν τμήματα διαχειριστή της Γουιάνας / Γαλλικής Γουιάνας

Πηγή: https://www.geonames.org/

	J	K	L	M	N	O	P	Q	R	S	T	U
1	adminName1	adminCode1	adminName2	adminCode2	adminName3	adminCode3	adminName4	adminCode4	adminName5	adminCode5	lng	lat
22	Guyane	GF									-52.99856	3.99929
23	Guyane	GF	Guyane	973							-52.99994	3.99886
24	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731					-52.5	4
25	Guyane	GF	Guyane	973	Arrondissement de Saint-Laurent-du-Maroni	9732					-53.75	4
26	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Régina	97301			-52.1292	4.3136
27	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Cayenne	97302			-52.335	4.9386
28	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Iracoubo	97303			-53.2056	5.48
29	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Kourou	97304			-52.6428	5.1583
30	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Macouria	97305			-52.4739	5.0136
31	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Matoury	97307			-52.3311	4.8506
32	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Saint-Georges	97308			-51.8011	3.8894
33	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Remire-Montjoly	97309			-52.2767	4.905
34	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Roura	97310			-52.3242	4.7283
35	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Sinnamary	97312			-52.9586	5.3775
36	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Montsinéry-Tonnegrande	97313			-52.4928	4.8928
37	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Ouanary	97314			-51.6717	4.2092
38	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Camopi	97356			-52.3436	3.1703
39	Guyane	GF	Guyane	973	Arrondissement de Cayenne	9731	Saint-Élie	97358			-53.2881	4.8256
40	Guyane	GF	Guyane	973	Arrondissement de Saint-Laurent-du-Maroni	9732	Mana	97306			-53.7769	5.6672
41	Guyane	GF	Guyane	973	Arrondissement de Saint-Laurent-du-Maroni	9732	Saint-Laurent-du-Maroni	97311			-54.0289	5.5039
42	Guyane	GF	Guyane	973	Arrondissement de Saint-Laurent-du-Maroni	9732	Saül	97352			-53.2083	3.6272
43	Guyane	GF	Guyane	973	Arrondissement de Saint-Laurent-du-Maroni	9732	Maripasoula	97353			-54.0269	3.6428
44	Guyane	GF	Guyane	973	Arrondissement de Saint-Laurent-du-Maroni	9732	Grand-Santi	97357			-54.38	4.2728
45	Guyane	GF	Guyane	973	Arrondissement de Saint-Laurent-du-Maroni	9732	Apatou	97360			-54.3431	5.1567
46	Guyane	GF	Guyane	973	Arrondissement de Saint-Laurent-du-Maroni	9732	Awala-Yalimapo	97361			-53.9081	5.7361
47	Guyane	GF	Guyane	973	Arrondissement de Saint-Laurent-du-Maroni	9732	Papaichton	97362			-54.1722	3.8086
48		00									-53	4

Εικόνα 12: Αυτοί οι κωδικοί υποδιαίρεσης προέρχονται από το INSEE COG, οι οποίοι είναι οι επίσημοι κωδικοί που χρησιμοποιούνται από τη γαλλική κυβέρνηση για τον προσδιορισμό όλων των γεωγραφικών περιοχών τόσο για τη μητροπολιτική Γαλλία ,όσο και για τα υπερπόντια διαμερίσματα και συλλογικότητες

Πηγή: https://www.geonames.org/

---

5. Συμπεράσματα

5.1 Σύγκριση των γεωχωρικών βάσεων δεδομένων

Η DBpedia και η GeoNames είναι δύο διαφορετικά συστήματα που παρέχουν γεωγραφικά δεδομένα, αλλά έχουν διαφορετικές προσεγγίσεις και παρέχουν διαφορετικού είδους πληροφορίες. Ας δούμε πώς μπορούμε να συγκρίνουμε τα δύο αυτά συστήματα:

1. **Περιεχόμενο:** Και η DBpedia και το GeoNames περιέχουν γεωγραφικά δεδομένα, αλλά με διαφορετικές πτυχές. Η DBpedia είναι ένα σύστημα που εξάγει δεδομένα από τη Wikipedia και τα οργανώνει σε μια δομή βάσης δεδομένων, περιλαμβάνοντας γεωγραφικές πληροφορίες όπως τοποθεσίες, ιστορικά γεγονότα κλπ. Το GeoNames, από την άλλη πλευρά, είναι ένα σύστημα που παρέχει κυρίως γεωγραφικά δεδομένα όπως γεωγραφικές τοποθεσίες, γεωγραφικές συντεταγμένες, ονόματα τοποθεσιών κλπ.
2. **Κάλυψη:** Η DBpedia καλύπτει μια ευρεία ποικιλία θεμάτων, συμπεριλαμβανομένων των γεωγραφικών πληροφοριών, αλλά δεν είναι αποκλειστικά αφιερωμένη σε αυτά. Το GeoNames είναι εξειδικευμένο στην παροχή γεωγραφικών πληροφοριών και έχει εκτεταμένη κάλυψη σε όλο τον κόσμο.
3. **Διαθεσιμότητα και άδειες:** Τόσο η DBpedia όσο και το GeoNames παρέχουν τα δεδομένα τους υπό ανοικτές άδειες, που επιτρέπουν την ελεύθερη χρήση και επαναχρησιμοποίηση τους από το ευρύ κοινό.
4. **Ακρίβεια και αξιοπιστία:** Η ακρίβεια και η αξιοπιστία των δεδομένων μπορεί να διαφέρει μεταξύ των δύο συστημάτων, ανάλογα με την πηγή και τη μέθοδο συλλογής τους.

Συνολικά, η σύγκριση της DBpedia και του GeoNames απαιτεί να ληφθούν υπόψη η φύση των δεδομένων που παρέχουν, η κάλυψη, η διαθεσιμότητα και άλλοι παράγοντες, ανάλογα με τις ανάγκες του κάθε χρήστη ή εφαρμογής.

Ο παρακάτω πίνακας συνοψίζει για διαφορετικούς παρόχους τον αριθμό των διαθέσιμων γεωχωρικών δεδομένων καθώς και τον τρόπο πρόσβασης σε αυτά τα δεδομένα:

Provider	#Geodata	Data access
DBpedia	727 232 triples	SPARQL endpoint
Geonames	5 240 032 (feature).	API
LinkedGeoData	60 356 364 triples	SPARQL endpoint, Snorql
Foursquare	n/a	API
Freebase	8,5MB	RDF Freebase Service
Ordnance Survey(Cities)	6 295 triples	Talis API
GeoLinkedData.es	101 018 triples	SPARQL endpoint
Google Places	n/a	Google API
GADM project data	682 605 triples	Web Service
NUTS project data	316 238 triples	Web Service
IGN experimental	629 716 triples	SPARQL endpoint

Εικόνα 13: Γεωχωρικά δεδομένα ανά πάροχο και διαφορετικό τύπο πρόσβασής τους

Πηγή: Comparing Vocabularies for Representing Geographical Features and Their Geometry Ghislain Auguste Atemezing, Raphael Troncy

Συνεπώς, όταν αξιολογούμε τις αλλαγές που έχουν γίνει σε ένα μοντέλο δεδομένων είναι απαραίτητο να το συγκρίνουμε με την υπάρχουσα έκδοση για να δούμε τις διαφορές. Για παράδειγμα, μπορεί να θέλουμε να δούμε πόσα πεδία έχουν προστεθεί στο νέο σχήμα για να προσδιορίσουμε εάν είναι απαραίτητη ή όχι μια ενημέρωση της βάσης.

Το εργαλείο σύγκρισης δεδομένων – Data Comparison Tool επιτρέπει να συγκρίνουμε μια υπάρχουσα βάση δεδομένων με μια ενημερωμένη έκδοση του μοντέλου δεδομένων. Χρησιμοποιώντας αυτό το εργαλείο, μπορούμε να παρακολουθούμε τις διαφορές μεταξύ διαφόρων πτυχών της βάσης δεδομένων, συμπεριλαμβανομένων του σχήματος, της γεωμετρίας, των χαρακτηριστικών και της χωρικής αναφοράς. Μπορούμε επίσης να επιλέξουμε να συγκρίνουμε τις βάσεις δεδομένων με βάση όλες αυτές τις πτυχές ή μια συγκεκριμένη πτυχή.

Ακόμη, μια προσέγγιση είναι η χρήση γλωσσών προγραμματισμού όπως η Python μαζί με σχετικές βιβλιοθήκες για χειρισμό και σύγκριση δεδομένων. Για να υλοποιήσουμε αυτή τη σύγκριση πρέπει να ακολουθήσουμε το παρακάτω μοτίβο:

Ανάκτηση δεδομένων (*Data Retrieval*) : Λήψη δεδομένων τόσο από τη GeoNames όσο και από τη DBpedia. Μπορούμε να χρησιμοποιήσουμετα API που παρέχονται από αυτές τις υπηρεσίες ή να κατεβάσουμε απευθείας σύνολα δεδομένων αν είναι διαθέσιμα.

Προεπεξεργασία δεδομένων (*Data Preprocessing*): Προεπεξεργασία των δεδομένων που περιλαμβάνει την τυποποίηση μορφών,τον χειρισμό τιμών που λείπουν και την αφαίρεση διπλοτύπων.

Σύγκριση δεδομένων (*Data Comparison*): Χρήση αλγορίθμων και βιβλιοθηκών π.χ. GeoPandas για να συγκρίνουμε γεωγραφικές συντεταγμένες και γεωμετρίες.

Παρακάτω είναι ένα παράδειγμα Python που χρησιμοποιεί GeoPandas για χειρισμό και σύγκριση δεδομένων:

```
import pandas as pd

# Load data from GeoNames and DBpedia (assuming they are in CSV format)

geonames_data = pd.read_csv('geonames_data.csv')

dbpedia_data = pd.read_csv('dbpedia_data.csv')

# Example: Compare place names

merged_data = pd.merge(geonames_data, dbpedia_data, on='place_name', how='inner')

print("Common place names between GeoNames and DBpedia:")

print(merged_data['place_name'])

# Example: Compare geographical coordinates

geonames_coords = geonames_data[['latitude', 'longitude']]

dbpedia_coords = dbpedia_data[['latitude', 'longitude']]

# Calculate distance between coordinates (using Euclidean distance as an example)

distance = ((geonames_coords - dbpedia_coords) ** 2).sum(axis=1) ** 0.5

merged_data['distance'] = distance

print("Mean distance between coordinates:")

print(distance.mean())
```

Η ανάπτυξη του τομέα της οντολογίας μπορεί να εξελιχθεί και να βοηθήσει εξαιρετικά στη δημιουργία ενός κοινού πλαισίου ανταλλαγής και επεξεργασίας της γνώσης. Οι οντολογίες που δημιουργούνται πλέον είναι από πολλά επιστημονικά πεδία και μας δίνεται η δυνατότητα να λειτουργήσουμε συνεργατικά στη δημιουργία συνδυασμών που πριν δεν ήταν εφικτοί. Ωστόσο υπάρχουν σημαντικά προβλήματα μέχρι να γίνει ένα τέτοιο πλάνο πραγματικότητα. Οι οντολογίες υψηλού πεδίου σε πολλές περιπτώσεις δεν επιτρέπουν τον συνδυασμό όλων των οντολογιών, καθώς χρησιμοποιούν διαφορετική νοηματοδότηση. Ακόμα υπάρχουν πολλές οντολογίες που έχουν δημιουργηθεί σε διαφορετική γλώσσα αναπαράστασης. Σε κάθε

περίπτωση όμως η κατηγοριοποίηση αυτή της γνώσης και η δόμηση της σε κοινή παγκόσμια γλώσσα μπορεί να αποφέρει πολλούς σημαντικούς καρπούς στο μέλλον. Ένα άλλο είδος προβλήματος που δημιουργείται είναι η δυσκολία απόδοσης σύνθετων εννοιών με αντικειμενικό και σαφή τρόπο. Επίσης είναι αρκετά δύσκολη η περιγραφή των σχέσεων μεταξύ των εννοιών. Παρά τα δομικά προβλήματα των οντολογιών, έχουν γίνει εξαιρετικά επιτυχημένες προσπάθειες ιδιαίτερα στους τομείς των θετικών επιστήμων, της ιατρικής και της οικονομική επιστήμης. Διαφαίνεται τελικώς, ότι πολλοί φορείς αναμένεται να ασχοληθούν εκτενώς με τον τομέα, και το ακαδημαϊκό ενδιαφέρον αυξάνεται συνεχώς. Έχουν γίνει ήδη μεγάλες βιβλιοθήκες εννοιών και οντολογιών που θα μπορούσαν να χρησιμοποιηθούν ως βάση για την περαιτέρω ανάπτυξη του κλάδου.

Υπογραμμίζεται ότι η ρίζα *(root node*) μιας οντολογίας πεδίου καλό είναι να συσχετιστεί (ή να οριστεί) με όρους που εμπεριέχονται σε μια οντολογία υψηλού επιπέδου, προκειμένου να καταστεί αποδοτικότερη η διαχείριση της επιστημονικής πληροφορίας που εμπεριέχεται στη συγκεκριμένη οντολογία. Η εν λόγω διαδικασία θα διασφαλίσει ότι η οντολογία είναι δομημένη χρησιμοποιώντας την αρχιτεκτονική μιας οντολογίας υψηλού επιπέδου την οποία μοιράζεται με άλλες οντολογίες (πολλές άλλες οντολογίες που έχουν πια κοινή αρχιτεκτονική) (*Arp & Spear 2015*).

Οι οντολογίες υψηλού επιπέδου λειτουργούν ως σημασιολογικές γέφυρες για τη διευκόλυνση της σημασιολογικής ολοκλήρωσης (επίλυση προβλημάτων σημασιολογικής ετερογένειας) οντολογιών πεδίου και για την καθοδήγηση της ανάπτυξης νέων οντολογιών. Για τον λόγο αυτό, επικεντρώνονται στην τυποποίηση γενικών και αφηρημένων εννοιών που δεν σχετίζονται με κάποιο συγκεκριμένο τομέα, αλλά αντίθετα ισχύουν για όλους τους τομείς και επομένως μπορούν να επαναχρησιμοποιηθούν εύκολα από αυτούς. Δηλαδή, παρέχουν ένα κοινό οντολογικό υπόβαθρο για οντολογίες πεδίου (Hoehndorf 2010, Schmidt και άλλοι 2016). Επιπλέον, μια οντολογία υψηλού επιπέδου παρέχει περιορισμούς στις κατηγορίες που εμπεριέχει (*με τη μορφή αξιωμάτων*) οι οποίοι κληρονομούνται από τις οντολογίες πεδίου που συσχετίζονται με τη συγκεκριμένη οντολογία υψηλού επιπέδου.

Συνεπώς, οι οντολογίες υψηλού επιπέδου αποτελούν ένα μέσο επαλήθευσης των οντολογιών πεδίου. Κάτι τέτοιο είναι ιδιαίτερα χρήσιμο κατά τη διαδικασία δημιουργίας μιας νέας οντολογίας για την οποία επιδιώκεται σημασιολογική διαλειτουργικότητα με μια άλλη, υφιστάμενη οντολογία. Κατά τη διαδικασία της δημιουργίας/ανάπτυξης μιας νέας οντολογίας, οι οντολογίες υψηλού επιπέδου συμβάλλουν στην επαλήθευση βασικών οντολογικών περιορισμών. Μπορούν επίσης να χρησιμοποιηθούν για την επαλήθευση της συμβατότητας μιας νέας οντολογίας με άλλες οντολογίες που βασίζονται στην ίδια οντολογία υψηλού επιπέδου.

Κατά συνέπεια, παρέχουν υψηλής ποιότητας έλεγχο συμβατότητας και αξιοπιστίας για οντολογίες πεδίου και τη σημασιολογική ολοκλήρωσή τους (*Hoehndorf 2010*). Επιπλέον, αν σκεφτεί κανείς το πλήθος των οντολογιών πεδίου που υπάρχουν, η συσχέτιση αυτών με μια οντολογία υψηλού επιπέδου είναι εξαιρετικά χρήσιμη, καθώς είναι σημαντικό να επαναχρησιμοποιηθεί η καλά τεκμηριωμένη γνώση που υπάρχει στις οντολογίες υψηλού επιπέδου μαζί με τις οντολογίες πεδίου, προκειμένου να μειωθεί ο χρόνος μοντελοποίησης, το πρόβλημα ετερογένειας της αναπαράστασης γνώσης και η πολυπλοκότητα της μοντελοποίησης με οντολογίες (*Schmidt και άλλοι 2016*).

6.Βιβλιογραφία

Aurona Gerber, Alta van der Merwe, Andries Barnard – A Functional Semantic Web Architecture, 2008, [http://www.eswc2008.org/final-pdfs-for-web-site/fisr-1.pdf]

Ian Horrocks, Bijan Parsia, Peter-Patel-Schneider, James Hendler – Semantic Web Architecture: Stack or Two Towers?, 2005,

[http://www.cs.manchester.ac.uk/~ horrocks/Publications/download/2005/HPPH05.pdf]

Βασιλειάδης Νικόλαος – Σημειώσεις μαθήματος 2ου εξαμήνου ΔΠΜΣ Πληροφορικής και Διοίκησης: Διαχείριση Γνώσης, 2009, [http://lpis.csd.auth.gr/mtpx/km/slides1.pdf]

Natalya F. Noy, Deborah L. McGuinness – Ontology Development 101: A Guide to Creating Your First Ontology, [http://www-ksl.stanford.edu/people/dlm/papers/ontologytutorial-noy-mcguinness.pdf]

Grigoris Antoniou, Enrico Franconi, Frank van Harmelen – Introduction to Semantic Web Ontology Languages, 2005, Springer Verlag

Oscar Corcho, Mariano Fernández-López, Asunción Gómez-Pérez – Methodologies, tools and languages for building ontologies. Where is their meeting point?, 2002, Madrid, Spain

York Sure – Methodology, Tools & Case Studies for Ontology Based knowledge Management, May 2003, Fridericiana University, Karlsruhe

Oscar Corcho, Mariano Fernández-López, Asunción Gómez-Pérez, Angel López-Cima – Building legal ontologies with METHONTOLOGY and WebODE, Work supported by the project Esperonto (IST-2001-34373)

Geonames Wordnet(geown):extracting wordnets from GeoNames,Francis Bond,Arthur Bond

Assessment of the accuracy of GeoNames gazetteer data, Dirk Ahlers

Enhancing and Validating GeoNames Data with Digital Nautical Charts Data: A Case Study in the Mapping of Freeform Map Labels Dakotah D. Maguire* a , Jason C. Kaufmana , Alexandre Sorokinea , Robert Stewarta

GeoNames website. https://www.geonames.org/. Accessed June 15, 2022

“About GeoNames.” https://www.geonames.org/about.html. Accessed June 15, 2022

International Hydrographic Organization. 2000. IHO Transfer Standard for Digital Hydrographic Data. Monaco: International Hydrographic Bureau. https://iho.int/uploads/user/pubs/standards/s-57/31Main.pdf

International Hydrographic Organization. 2018. S-100 - Universal Hydrographic Data Model, Edition 4.0.0. Monaco: International Hydrographic Organization. https://iho.int/uploads/user/pubs/standards/s-100/S100_Ed%204.0.0_Clean_17122018.pdf

National Geospatial-Intelligence Agency. Digital Nautical Chart. https://dnc.nga.mil/dncp/home.php

Accessed June 15th, 2022 National Oceanic and Atmospheric Administration. Nautical Cartography: The Making of a NOAA Nautical Chart

https://nauticalcharts.noaa.gov/learn/nauticalcartography.html Accessed June 15th, 2022

Kim, Junchul, Maria Vasardani, and Stephan Winter. 2017. “Similarity matching for integrating spatial information extracted from place descriptions.” International Journal of Geographical Information Science. doi:https://doi.org/10.1080/13658816.2016.1188930

Šimbera, Jan, Dusan Drbohlav, and Přemysl Štych. 2021. “Geocoding Freeform Placenames: An Example of Deciphering the Czech National Immigration Database.” International Journal of Geo-Information 335–351. doi: https://doi.org/10.3390/ijgi10050335

Dunn, Jonathan. 2020. “Mapping Languages: The Corpus of Global Language Use.” Language Resources and Evaluation vol. 54: 999–1018. doi: https://doi.org/10.1007/s10579-020-09489-2. NGA.

2019. “U.S. Chart No. 1: Symbols, Abbreviations and Terms used on Paper and Electronic Navigational Charts.” https://nauticalcharts.noaa.gov/publications/docs/us-chart-1/ChartNo1.pdf

Ο Σημασιολογικός Ιστός και τα Ανοιχτά Διασυνδεδεμένα Δεδομένα στην Εκπαίδευση Ανάπτυξη θεματικής πύλης δυναμικής παρουσίασης ανοικτών και διασυνδεδεμένων εκπαιδευτικών δεδομένων, Αικατερίνη Τσαρτσάλη

Σημασιολογικός Ιστός και σχετικές τεχνολογίες,πιθανές χρήσεις στην ακαδημαική κοινότητα,Παναγιώτης Ντούσικος,Παναγιώτης Τσιώλης

Αντωνίου, Γ., van Harmelen, F., (2009), Εισαγωγή στο σημασιολογικό ιστό, Δεύτερη Αμερικάνικη Έκδοση, Εκδόσεις Κλειδάριθμος

Αρναούτη, Ε., (2007), ‘ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΖΗΤΗΣΗ: ΑΡΧΙΤΕΚΤΟΝΙΚΕΣ ΣΗΜΑΣΙΟΛΟΓΙΚΩΝ ΥΠΗΡΕΣΙΩΝ ΙΣΤΟΥ.’, Πτυχιακή εργασία, ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ, ΤΜΗΜΑ ΕΦΑΡΜΟΓΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗ ΔΙΟΙΚΗΣΗ ΚΑΙ ΟΙΚΟΝΟΜΙΑ, ΜΕΣΟΛΟΓΓΙ

Baader, F., Calvanese, D., McGuiness, D., Nardi, D. and Patel-Schneider, P.,F., (2003), The Description Logic Handbook: Theory, Implementation, and Applications., Cambridge: Cambridge University Press

BernersLee, T., Fischetti, M., (1999), Weaving the Web. HarperSanFrancisco, chapter 12, ISBN 9780062515872 Berners-Lee, T., Hendler, J., & Lassila, O., (2001), The Semantic Web: A new form of Web content that is meaningful to computers would unleash a revolution of new possibilities, Scientific American

Γκατσώνη, Γ., Βαασιλειάδης, Ν., (2009), Εφαρμογή Τεχνολογιών του Σημασιολογικού Ιστού στη Διαχείριση Γνώσης στα Πλαίσια της ηλεκτρονικής Τραπεζικής, Τμήμα Πληροφορικής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης Collins, A.,M., Quillian, M.,R., (1969), «Retrieval time from semantic memory», Journal of verbal learning and verbal behavior 8 (2): 240–247

Collins, A.,M., Quillian, M.,R., (1970), «Does category size affect categorization time?», Journal of verbal learning and verbal behavior 9 (4): 432–438

Collins, A.,M., Loftus, E.,F., (1975), «A spreading activation theory of semantic processing», Psychological Review 82 (6): 407–428

Dean, M., Schreiber, G., Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.,L., Patel-Schneider, P.,F. and Stein, L.,A., (2004), OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004. Retrieved December 2016, from http://www.w3.org/TR/owl-ref/.

Καρακατσούλης, Δ., (2011), Ηλεκτρονικό Εμπόριο & Σημασιολογικός Ιστός: Υλοποίηση του Ηλεκτρονικού Καταστήματος YourBooks, Thesis, Πανεπιστήμιο Πατρών, Πολυτεχνική Σχολή, Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών & Πληροφορικής, Πάτρα

Κορμπάκης, Π., (2012), Πρακτική Εφαρμογή των Οντολογιών ως Εργαλεία Αναπαράστασης και Διαχείρισης Γνώσης στην Ηλεκτρονική Υγεία., THESIS., ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ, ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ, ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ, ΚΟΖΑΝΗ

Lukasiewicza, T., Stracciab, U., (2008) Managing uncertainty and vagueness in description logics for the Semantic Web, Web Semantics: Science, Services and Agents on the World Wide Web Volume 6, Issue 4, November 2008, Pages 291–308 Semantic Web Challenge 2006/2007

Σημασιολογικός Ιστός και Συνδεδεμένα Δεδομένα: Εφαρμογή στην ηλεκτρονική διακυβέρνηση» Ελένη Μπουλογεώργου-Νημά

Παγκόσμιος Ιστός http://users.forthnet.gr/ath/skonstan/site_1/articles/history_files/Web.html

Linked Data: Principles and State of the Art http://www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdf

Future of the Web, Tim Berners-Lee http://net.educause.edu/ir/library/pdf/EPO0719.pdf

Technical Report on Linked Data Applications - The Genesis and the Challenges of Using Linked Data on the Web http://linkeddata.deri.ie/sites/linkeddata.deri.ie/files/lod-app-tr-2009-07-26_0.pdf

Exploiting Linked Data for Building Web Applications, Michael Hausenblas http://sw-app.org/pub/exploit-lod-webapps-IEEEIC-preprint.pdf

Linked Data – The story so far http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

How to publish Link Data on the Web (tutorial) http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

Cool URIs for the Semantic Web http://www.dfki.uni-kl.de/~sauermann/2006/11/cooluris/

Tony Segaran, Jeff Hammerbacher, Beautiful Data, O'Reilly, August 2009 (page 335)

Interlinking Open Data on the Web http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkingOpenData.pdf

A Practical Introduction by Michael Hausenblas http://www.deri.ie/fileadmin/documents/teaching/tutorials/LinkedData_MichaelHa usenblas_2009-03-05.pdf

«Ανάπτυξη οντολογίας και συναφών δεδομένων (triples) για την περιγραφή της λειτουργίας του Ανοικτού Πανεπιστημίου» Παναγιώτης Τριανταφύλλου»

Gruber, T. 1993. A translation approach to portable ontologies. Knowledge Acquisition 5(2): 199-220

Kovács, L. και Micsik, A. 2005. An Ontology-Based Model of Digital Libraries. In the 8th International Conference on Asian Digital Libraries (ICADL 2005), LNCS 3815, 38-43

Mizoguchi, R., και Bourdeau, J. 2000. Using Ontological Engineering to Overcome Common AI-ED Problems. International Journal of Artificial Intelligence in Education 11: 107–121

W3C OWL Web Ontology Language Overview, Διαθέσιμο από http://www.w3.org/TR/owl-features/

W3C OWL Web Ontology Language Reference, Διαθέσιμο από http://www.w3.org/TR/owl-ref/

W3C Web-Ontology (Web-Onto) Working Group, Διαθέσιμο από http://www.w3.org/2001/sw/WebOnt/

ΑΝΑΠΤΥΞΗ ΟΝΤΟΛΟΓΙΑΣ ΣΤΟΝ ΤΟΜΕΑ ΠΑΡΟΧΗΣ ΥΠΗΡΕΣΙΩΝ ΥΓΕΙΑΣ ΣΕ ΠΕΡΙΒΑΛΛΟΝ WEBPROTEGE ΛΟΛΟΤΣΗ ΧΡΙΣΤΙΝΑ

Σχεδιασμός Οντολογίας για τη Σημασιολογική Διερεύνηση των Μικρών και Μεσαίων Έξυπνων Πόλεων στον Μεσογειακό Χώρο,Παναγιωτοπούλου