



**ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**ΕΞΑΓΩΓΗ ΣΗΜΑΣΙΟΛΟΓΙΚΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΜΕΣΩ ΤΩΝ ΑΝΟΙΧΤΩΝ ΔΙΑΣΥΝΔΕΜΕΝΩΝ
ΔΕΔΟΜΕΝΩΝ**

| Semantic Information Extraction using Linked Open Data |

ΣΩΚΡΑΤΗΣ ΑΘΑΝΑΣΙΑΔΗΣ

ΑΕΜ: 2547

Κατεύθυνση: Πληροφοριακά Συστήματα

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:

N. Βασιλειάδης, Καθηγητής Τμήματος Πληροφορικής ΑΠΘ

ΘΕΣΣΑΛΟΝΙΚΗ 2020

ΠΕΡΙΛΗΨΗ

Αντικείμενο της παρούσας εργασίας είναι η εξαγωγή σημασιολογικής πληροφορίας μέσω Ανοιχτών Διασυνδεδεμένων Δεδομένων

Τα Ανοιχτά Διασυνδεδεμένα Δεδομένα (Linked Open Data) αποτελούν πλέον αναπόσπαστο κομμάτι των περισσότερων συστημάτων τα οποία κάνουν οποιαδήποτε μορφή εξαγωγής γνώσης από το Διαδίκτυο. Η δομημένη σε RDF μορφή αναπαράστασης τους, καθώς και τα εξειδικευμένα εργαλεία, για την εξαγωγή πληροφορίας όπως η SPARQL, καθιστούν πιο εύκολη για την μηχανή την εξαγωγή και κατανόηση της πληροφορίας. Η πληροφορία που εξάγεται μέσω των Ανοιχτών Διασυνδεδεμένων Δεδομένων πλέον στηρίζει πάρα πολλά συστήματα Μηχανικής Μάθησης (Machine Learning) και Ρομποτικής.

Στην εργασία αυτή στόχος είναι να αναπτυχθεί ένα εργαλείο το οποίο θα χρησιμοποιεί την πληροφορία που βρίσκεται μέσα σε οντολογίες όπως η DBpedia και η ConceptNet και να ανακαλεί σημασιολογική πληροφορία. Συγκεκριμένα ως είσοδος θα δίδεται μια λίστα από οντότητες (π.χ. sugar, milk, mug, water) και ως έξοδος θα επιστρέφεται μια λίστα από οντότητες που είναι πιο σημασιολογικά δεμένες. Το εργαλείο αναπτύσσεται με την διαδικαστική γλώσσα προγραμματισμού Python.

ABSTRACT

The subject of the present dissertation is the extraction of Semantic Information through Linked Open Data.

Linked Open Data is now an integral part of most systems that perform any form of knowledge retrieval from the Internet. Their RDF-structured representation format, as well as specialized information extraction tools such as SPARQL, make it easier for the machine to retrieve and understand information from the (Semantic) Web. The extracted information through Linked Open Data can now support many Machine Learning and Robotics-related tasks.

The purpose of this work is to develop a tool that will use data found in ontologies, such as Dbpedia and ConceptNet, and retrieve semantic information from them. Specifically, an entity (e.g. sugar, milk, mug, water) will be given as an input and a list of entities most related to the entity given will be returned as an output. This tool is developed with a procedural programming language, Python.

ΕΥΧΑΡΙΣΤΙΕΣ

Πριν την παρουσίαση των αποτελεσμάτων της παρούσας εργασίας, αισθάνομαι την υποχρέωση να ευχαριστήσω ορισμένους από τους ανθρώπους που γνώρισα, συνεργάστηκα μαζί τους και έπαιξαν πολύ σημαντικό ρόλο στην πραγματοποίησή της, καθώς και τους ανθρώπους που αποτέλεσαν σημαντικό ψυχικό εφόδιο για την ολοκλήρωση των σπουδών μου.

Εκφράζω τις ευχαριστίες μου στον καθηγητή κ. Νικόλαο Βασιλειάδη που μου εμπιστεύτηκε την υλοποίηση αυτής της πτυχιακής. Ενώ εκφράζω και τις ευχαριστίες μου και στον συνεπιβλέποντα της πτυχιακής Αλέξανδρο Βασιλειάδη, ο οποίος με τις συμβουλές του και την καθοδήγηση του αποτέλεσε καθοριστικό ρόλο στην ανάπτυξη της εργασίας, καθώς επίσης του είμαι ευγνώμων για όλη την υπομονή που έδειξε καθ' όλη την διάρκεια της συνεργασίας μας.

Θα ήθελα να ευχαριστήσω τον συμφοιτητή μου Κώστα για τις άπειρες φορές που συνεργαστήκαμε στη διάρκεια των σπουδών μου και για τις ώρες που επενδύσαμε πάνω από έναν υπολογιστή προγραμματίζοντας.

Ευχαριστώ πολύ τους γονείς μου που με στήριζαν σε όλη την διάρκεια των σπουδών μου και που ήταν δίπλα μου στις δύσκολες στιγμές.

Ευχαριστώ τον αδερφό μου Θοδωρή που η μόνη απάντηση που μου έδινε όταν του έκανα μια ερώτηση σχετικά με την Python ήταν η «διάβασες το documentation;» και με άφηνε να παιδεύομαι.

Δεν θα είχα καταφέρει τίποτα και δεν θα ήμουν αυτός που είμαι σήμερα, χωρίς τους φίλους μου Θάνο και Άγγελο.

Ευχαριστώ επίσης τους φίλους μου Έλενα και Μιχάλη.

Τέλος θα ήθελα να ευχαριστήσω την Δανάη που ήταν η πρώτη που άκουσε τις πρόβες που έκανα για την παρουσίαση της πτυχιακής.

Θεσσαλονίκη, 9/10/2020

Σωκράτης Αθανασιάδης

ΠΕΡΙΕΧΟΜΕΝΑ

1.1 ΓΕΝΙΚΗ ΙΔΕΑ	XX
1.2 ΕΝΟΤΗΤΕΣ ΕΡΓΑΣΙΑΣ	XX
2.1 ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΙΣΤΟΣ (SEMANTIC WEB).....	25
2.1.1 ΣΗΜΑΣΙΟΛΟΓΙΚΟ ΔΙΚΤΥΟ.....	26
2.1.2 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ ΣΗΜΑΣΙΟΛΟΓΙΚΩΝ ΔΙΚΤΥΩΝ	26
2.1.3 ΤΕΧΝΟΛΟΓΙΑ ΓΙΑ ΤΟΝ ΣΗΜΑΣΙΟΛΟΓΙΚΟ ΙΣΤΟ.....	28
2.1.4 ΑΠΟ ΤΑ ΔΕΔΟΜΕΝΑ ΣΤΗΝ ΓΝΩΣΗ	29
2.2 LAYER APPROACH OF SEMANTIC WEB	29
2.3 ΑΝΟΙΧΤΑ ΔΙΑΣΥΝΔΕΔΕΜΕΝΑ ΔΕΔΟΜΕΝΑ (LINKED OPEN DATA)..	31
2.4 ΔΙΑΦΟΡΑ ΣΗΜΑΣΙΟΛΟΓΙΚΟΥ ΙΣΤΟΥ ΜΕ ΤΗΝ ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ	33
2.5 Ο ΡΟΛΟΣ ΤΟΥ ΣΗΜΑΣΙΟΛΟΓΙΚΟΥ ΙΣΤΟΥ ΣΤΗΝ ΚΑΘΗΜΕΡΙΝΗ ΖΩΗ.....	34
3.1 HTML ΚΑΙ RDF.....	39
3.2 RESOURCE DESCRIPTION FRAMEWORK (RDF)	40
3.3 SPARQL	42
3.3.1 ΒΑΣΙΚΑ SPARQL ΕΡΩΤΗΜΑΤΑ	42
3.3.2 ΕΝΤΟΛΗ SELECT ΚΑΙ WHERE	43
3.3.3 ΕΝΤΟΛΗ LIMIT	45
3.3.4 ΕΝΤΟΛΗ FILTER.....	46
3.3.5 ΕΝΤΟΛΗ PREFIX.....	46
3.3.6 ΜΕΡΙΚΑ ΑΚΟΜΗ ΠΙΟ ΣΥΝΘΕΤΑ ΠΑΡΑΔΕΙΓΜΑΤΑ	46
3.4 ΟΝΤΟΛΟΓΙΑ (ONTOLOGY).....	49
3.5 ΓΛΩΣΣΑ ΟΝΤΟΛΟΓΙΑΣ ΙΣΤΟΥ (WEB ONTOLOGY LANGUAGE - OWL)	50
4.1 CONCEPTNET	55
4.1.1 OPEN MIND COMMON SENSE	55
4.2 DBPEDIA.....	61
4.2.1 DBPEDIA ΚΑΙ ΑΝΟΙΧΤΑ ΔΙΑΣΥΝΔΕΔΕΜΕΝΑ ΔΕΔΟΜΕΝΑ	62
4.2.2 DBPEDIA ENDPOINT	66
4.3 WORDNET	67

4.3.1 ΣΧΕΣΗ ΥΠΕΡΩΝΥΜΙΑΣ (HYPERONYMY) ΚΑΙ ΜΕΡΩΝΥΜΙΑΣ (MERONYMY)	68
4.3.2 ΡΗΜΑΤΑ	69
4.3.3 ΕΠΙΘΕΤΑ	69
5.1 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ ΚΑΙ ΠΑΡΑΔΟΣΙΑΚΑ ΣΥΣΤΗΜΑΤΑ	
ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΑΠΟ ΤΟ WEB	73
5.1.1 ΟΜΟΙΟΤΗΤΑ ΣΥΝΗΜΙΤΟΝΩΝ (COSINE SIMILARITY)	74
5.1.2 ΜΟΝΤΕΛΟ TF-IDF	76
5.2 ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ	77
5.3 Η ΕΞΕΛΙΞΗ ΤΩΝ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ	78
5.4 ΔΙΚΤΥΟ ΣΗΜΑΣΙΟΛΟΓΙΚΗΣ ΟΜΟΙΟΤΗΤΑΣ (SEMANTIC	
SIMILARITY NETWORK)	81
5.5 ΑΝΑΓΚΗ ΓΙΑ ΣΗΜΑΣΙΟΛΟΓΙΚΑ ΔΕΔΟΜΕΝΑ	82
6 ΥΛΟΠΟΙΗΣΗ	85
6.1 CONCEPTNET.PY	85
6.2 DBPEDIA.PY-TFIDF.PY	88
6.2.1 DBPEDIA.PY	88
6.2.2 Η ΜΕΘΟΔΟΣ QUERY ΤΗΣ DBPEDIA.PY	90
6.2.3 TFIDF.PY	91
7 ΑΠΟΤΕΛΕΣΜΑΤΑ	95
7.1 ΣΧΕΤΙΚΑ ΜΕ ΤΗΝ ΣΥΝΟΛΙΚΗ ΜΕΤΡΙΚΗ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ... ..	95
7.1.1 ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΟΝ ΟΡΟ ΑΝΑΖΗΤΗΣΗΣ «TIME»	96
7.1.2 ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΟΝ ΟΡΟ ΑΝΑΖΗΤΗΣΗΣ «SUGAR»	98
7.1.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΟΝ ΟΡΟ ΑΝΑΖΗΤΗΣΗΣ «RED»	100
7.1.4 ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΟΝ ΟΡΟ ΑΝΑΖΗΤΗΣΗΣ «KITCHEN»	101
7.1.5 ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΟΝ ΟΡΟ ΑΝΑΖΗΤΗΣΗΣ «COMPUTER»	102
8 ΕΠΙΛΟΓΟΣ	106

ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ

ΕΙΚΟΝΑ 1 ΔΙΚΤΥΟ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ	25
ΕΙΚΟΝΑ 2 ΣΗΜΑΣΙΟΛΟΓΙΚΟ ΔΙΚΤΥΟ	26
ΕΙΚΟΝΑ 3 ΠΟΡΦΥΡΙΚΟ ΔΕΝΤΡΟ	27
ΕΙΚΟΝΑ 4 ΜΟΝΤΕΛΟ RDF ΚΑΙ ΚΛΗΡΟΝΟΜΙΚΟΤΗΤΑ	29
ΕΙΚΟΝΑ 5 ΕΠΙΠΕΔΑ ΣΗΜΑΣΙΟΛΟΓΙΚΟΥ ΙΣΤΟΥ	30
ΕΙΚΟΝΑ 6 ΓΡΑΦΗΜΑ ΑΝΟΙΧΤΩΝ ΔΙΑΣΥΝΔΕΔΕΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ	32
ΕΙΚΟΝΑ 7 ΑΝΑΓΝΩΡΙΣΗ ΑΝΤΙΚΕΙΜΕΝΩΝ ΑΠΟ ΡΟΜΠΟΤ	33
ΕΙΚΟΝΑ 8 ΓΡΑΦΟΣ ΣΗΜΑΣΙΟΛΟΓΙΚΗΣ ΠΛΗΡΟΦΟΡΙΑΣ	35
ΕΙΚΟΝΑ 9 ΣΧΕΣΗ RDF	40
ΕΙΚΟΝΑ 10 ΤΡΙΠΛΕΤΑ RDF	40
ΕΙΚΟΝΑ 11 ΠΑΡΑΔΕΙΓΜΑ RDF ΤΡΙΠΛΕΤΑΣ	41
ΕΙΚΟΝΑ 12 SPARQL ENDPOINT	43
ΕΙΚΟΝΑ 13 ΑΠΟΤΕΛΕΣΜΑΤΑ SPARQL ΕΡΩΤΗΜΑΤΟΣ ΓΙΑ ΤΟΝ ΟΡΟ «ΤΕΑ»	43
ΕΙΚΟΝΑ 14 THUMBNAIL ΑΠΟ ΕΡΩΤΗΜΑ SPARQL ΓΙΑ ΤΟΝ ΟΡΟ «ΤΕΑ»	45
ΕΙΚΟΝΑ 15 ΠΑΡΑΔΕΙΓΜΑ ΟΝΤΟΛΟΓΙΩΝ	49
ΕΙΚΟΝΑ 16 OWL2 ΟΝΤΟΛΟΓΙΑ	51
ΕΙΚΟΝΑ 17 ΤΡΙΠΛΕΤΕΣ RDF	56
ΕΙΚΟΝΑ 18 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΖΗΤΗΣΗΣ ΑΠΟ ΤΟ CONCEPTNET ΓΙΑ ΤΟΝ ΟΡΟ «COFFEE»	57
ΕΙΚΟΝΑ 19 ΔΕΔΟΜΕΝΑ ΑΠΟ ΤΟ CONCEPTNET	58
ΕΙΚΟΝΑ 20 ΔΕΔΟΜΕΝΑ ΑΠΟ CONCEPTNET	59
ΕΙΚΟΝΑ 21 ΔΕΔΟΜΕΝΑ ΑΠΟ CONCEPTNET	59
ΕΙΚΟΝΑ 22 ΔΕΔΟΜΕΝΑ ΑΠΟ CONCEPTNET	60
ΕΙΚΟΝΑ 23 ΔΕΔΟΜΕΝΑ ΑΠΟ CONCEPTNET	60
ΕΙΚΟΝΑ 24 ΤΡΙΠΛΕΤΑ ΑΠΟ DBPEDIA	62
ΕΙΚΟΝΑ 25 ΝΕΦΟΣ ΑΝΟΙΧΤΩΝ ΔΙΑΣΥΝΔΕΔΕΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ	63
ΕΙΚΟΝΑ 26 ΛΕΞΙΛΟΓΙΚΟ ΠΑΡΑΔΕΙΓΜΑ ΑΠΟ WORDNET	67
ΕΙΚΟΝΑ 27 ENTITY ΤΟΥ WORDNET	68
ΕΙΚΟΝΑ 28 ΛΕΞΙΛΟΓΙΚΗ ΑΝΑΛΥΣΗ ΤΟΥ ΟΡΟΥ «BOOK» ΑΠΟ WORDNET	69
ΕΙΚΟΝΑ 29 ΠΑΡΑΔΕΙΓΜΑ ΜΙΚΡΗΣ ΣΥΛΛΟΓΗΣ ΕΓΓΡΑΦΩΝ	73
ΕΙΚΟΝΑ 30 ΑΝΤΕΣΤΡΑΜΜΕΝΟΣ ΚΑΤΑΛΟΓΟΣ ΓΙΑ ΤΗΝ ΣΥΛΛΟΓΗ ΕΓΓΡΑΦΩΝ	74
ΕΙΚΟΝΑ 31 ΔΙΑΝΥΣΜΑΤΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ ΤΩΝ ΕΓΓΡΑΦΩΝ	75
ΕΙΚΟΝΑ 32 ΔΙΑΝΥΣΜΑΤΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ ΤΩΝ ΕΓΓΡΑΦΩΝ ΣΤΟΝ ΔΙΣΔΙΑΣΤΑΤΟ ΧΩΡΟ 75	
ΕΙΚΟΝΑ 33 ΔΟΜΗ ΜΗΧΑΝΗΣ ΑΝΑΖΗΤΗΣΗΣ ΤΟΥ WWW	78
ΕΙΚΟΝΑ 34 ΜΕΓΕΘΟΣ ΤΟΥ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ	80
ΕΙΚΟΝΑ 35 ΜΟΝΤΕΛΟ SUPPORT VECTOR MACHINE	81
ΕΙΚΟΝΑ 36 ΣΗΜΑΣΙΟΛΟΓΙΚΟ ΜΟΝΤΕΛΟ, ΑΝΑΠΑΡΑΣΤΑΣΗ ΠΡΑΓΜΑΤΙΚΟΥ ΚΟΣΜΟΥ ΜΕ ΔΕΔΟΜΕΝΑ	83
ΕΙΚΟΝΑ 37 ΤΡΙΠΛΕΤΕΣ RDF	86
ΕΙΚΟΝΑ 38 ΤΡΙΠΛΕΤΑ RDF ΓΙΑ ΤΟΝ ΟΡΟ ΑΝΑΖΗΤΗΣΗΣ	87
ΕΙΚΟΝΑ 39 ΑΠΟΤΕΛΕΣΜΑΤΑ CONCEPTNET.PY	87
ΕΙΚΟΝΑ 40 COMMENTBOX ΤΟΥ ΟΡΟΥ «ΤΕΑ» ΣΤΟ DBPEDIA	88
ΕΙΚΟΝΑ 41 ΤΕΛΙΚΑ ΒΑΡΗ ΣΤΙΣ ΟΝΤΟΤΗΤΕΣ	95

ΕΙΚΟΝΑ 42 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΖΗΤΗΣΗΣ ΓΙΑ ΤΟΝ ΟΡΟ «TIME»	97
ΕΙΚΟΝΑ 43 ΤΡΙΠΛΕΤΕΣ RDF ΓΙΑ ΤΟΝ ΟΡΟ «TIME»	98
ΕΙΚΟΝΑ 44 ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΗΣ ΣΧΕΣΗΣ «ISA»	98
ΕΙΚΟΝΑ 45 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΖΗΤΗΣΗΣ ΓΙΑ ΤΟΝ ΟΡΟ «SUGAR».....	99
ΕΙΚΟΝΑ 46 ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΗΝ ΣΧΕΣΗ «ISA»	100
ΕΙΚΟΝΑ 47 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΖΗΤΗΣΗΣ ΓΙΑ ΤΟΝ ΟΡΟ «RED»	100
ΕΙΚΟΝΑ 48 ΑΝΑΓΝΩΡΙΣΗ ΟΝΤΟΤΗΤΩΝ ΑΠΟ ΚΑΠΟΙΟ ΡΟΜΠΟΤ	101
ΕΙΚΟΝΑ 49 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΖΗΤΗΣΗΣ ΓΙΑ ΤΟΝ ΟΡΟ «KITCHEN»	102
ΕΙΚΟΝΑ 50 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΖΗΤΗΣΗΣ ΓΙΑ ΤΟΝ ΟΡΟ «COMPUTER»	103
ΕΙΚΟΝΑ 51 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΖΗΤΗΣΗΣ ΓΙΑ ΤΟΝ ΟΡΟ «COMPUTER»	103
ΕΙΚΟΝΑ 52 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΖΗΤΗΣΗΣ ΓΙΑ ΤΟΝ ΟΡΟ «COMPUTER», ΙΔΙΟΤΗΤΑ «USEDFor»	104
ΕΙΚΟΝΑ 53 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΖΗΤΗΣΗΣ ΓΙΑ ΤΟΝ ΟΡΟ «COMPUTER», ΙΔΙΟΤΗΤΑ «PARTOf»	104
ΕΙΚΟΝΑ 54 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΖΗΤΗΣΗΣ ΓΙΑ ΤΟΝ ΟΡΟ «COMPUTER», ΙΔΙΟΤΗΤΑ «ISA»	104

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

ΠΙΝΑΚΑΣ 1 ΣΥΝΤΑΞΕΙΣ ΟΝΤΟΛΟΓΙΩΝ.....	51
ΠΙΝΑΚΑΣ 2 ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ ΑΠΟ ΤΟ DBPEDIA	64
ΠΙΝΑΚΑΣ 3 ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ.....	80

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

ΕΙΣΑΓΩΓΗ

1.1 ΓΕΝΙΚΗ ΙΔΕΑ

Η συνεχής αύξηση του όγκου των δεδομένων που συνυπάρχουν στον Παγκόσμιο Ιστό συμβαδίζει με την απαίτηση για την ανάπτυξη εργαλείων που έχουν ως στόχο τον συνδυασμό ετερογενών αυτών δεδομένων, αφού αυτό είναι που θα φέρει την εξέλιξη του Παγκόσμιου Ιστού ενώ θα αποτελέσει και θεμέλιο ανάπτυξης του Σημασιολογικού Ιστού. Τα Σημασιολογικά Δεδομένα (Semantic Data) που υπάρχουν δεν είναι αρκετά ενώ ταυτόχρονα οι ιστοσελίδες που έχουν ενσωματωμένη κάποια μορφή Σημασιολογικής Πληροφορίας είναι κατά πολλές τάξεις μεγέθους μικρότερες από τις παραδοσιακές HTML ιστοσελίδες που στερούνται Σημασιολογικής Πληροφορίας. Η λύση όμως σε αυτό το πρόβλημα δεν είναι σίγουρα η αντικατάσταση των παραδοσιακών ιστοσελίδων με νέες που θα έχουν Σημασιολογική Πληροφορία, αφού κάτι τέτοιο θα ήταν αδύνατον αν σκεφτεί κανείς το μέγεθος του Ιστού που στερείται Σημασιολογικής Πληροφορίας, όμως η ανάπτυξη εργαλείων για την ενοποίηση των δεδομένων που υπάρχουν στο διαδίκτυο και την δημιουργία ενός κοινού συστήματος προσπέλασης ετερογενών δεδομένων είναι κάτι υλοποιήσιμο.

Η ανάπτυξη τέτοιων εργαλείων απαιτεί πρωτίστως την ανακάλυψη Σημασιολογικής Πληροφορίας μεταξύ οντοτήτων που συνυπάρχουν στο Διαδίκτυο. Τέτοιες οντότητες έχουν οριστεί μέσω των δικτύων που ορίζουν τα Ανοιχτά Διασυνδεδεμένα Δεδομένα (Linked Open Data), και η ανακάλυψη Σημασιολογικών Σχέσεων μεταξύ τους αποτελεί το βασικό αντικείμενο διατριβής της παρούσας πτυχιακής εργασίας. Φυσικά τα Ανοιχτά Διασυνδεδεμένα Δεδομένα θα ήταν αδύνατον να απουσιάζουν από την ανάπτυξη ενός τέτοιου εργαλείου αφού αυτά είναι που υποστηρίζουν οποιαδήποτε μορφή εξαγωγής γνώσης από το Διαδίκτυο. Ενώ άλλες τεχνολογίες όπως η RDF και η SPARQL μας δίνουν την δυνατότητα δόμησης τέτοιων δεδομένων ενώ μας προσφέρουν και κάποιες βασικές μορφές προσπέλασης σε αποθήκες Ανοιχτών Διασυνδεδεμένων Δεδομένων.

Πιο συγκεκριμένα το εργαλείο που αναπτύχθηκε χρησιμοποιεί τα Ανοιχτά Διασυνδεδεμένα Δεδομένα που υπάρχουν σε πόρους όπως το ConceptNet και το DBpedia και ανακαλεί Σημασιολογική Πληροφορία. Στο εργαλείο θα δίνεται ως είσοδος μια οντότητα, όπως για παράδειγμα το «Τσάι» και θα επιστρέφεται ένα σύνολο από άλλες οντότητες που συνδέονται σημασιολογικά με αυτήν, όπως για παράδειγμα «Ζάχαρη, Νερό, Καφές, Βότανο...».

1.2 ΕΝΟΤΗΤΕΣ ΕΡΓΑΣΙΑΣ

Η εργασία αναλύεται στα εξής κεφάλαια:

- Στο Κεφάλαιο 2 γίνεται αναφορά στις βασικές έννοιες που χρειάζονται για την κατανόηση της πτυχιακής. Αυτές είναι ο Σημασιολογικός Ιστός μαζί με τις τεχνολογίες που υπάρχουν γύρω από αυτόν, τα Ανοιχτά Διασυνδεδεμένα Δεδομένα που όπως αναφέρθηκε είναι απαραίτητα για την εκτέλεση του

εργαλείου. Ενώ τέλος τονίζονται κάποιες διαφορές που υπάρχουν στις έννοιες με άλλες αντίστοιχες του τομέα της Τεχνητής Νοημοσύνης.

- Στο Κεφάλαιο 3 θα μιλήσουμε πιο συγκεκριμένα για τις τεχνολογίες RDF, SPARQL και OWL που είναι οι βασικότερες τεχνολογίες που υποστηρίζουν την ανάπτυξη του Σημασιολογικού Ιστού. Επίσης στο κεφάλαιο αυτό γίνεται αναφορά στο τι ακριβώς ορίζουμε ως «οντολογία».
- Στο Κεφάλαιο 4 θα περιγράψουμε ακριβώς τους πόρους ConceptNet και DBpedia από τους οποίους εξάγουμε τα δεδομένα που είναι απαραίτητα για την λειτουργία του εργαλείου και πιο συγκεκριμένα τα δεδομένα που μας είναι απαραίτητα για την ανάκληση της Σημασιολογικής Πληροφορίας. Θα γίνει επίσης μια σύντομη περιγραφή σχετικά με το πώς ξεκίνησε η ανάπτυξη αυτών των πόρων. Επίσης γίνεται και σαφής αναφορά στο λεξικό της WordNet το οποίο επίσης χρησιμοποιήθηκε με παρόμοιο τρόπο όπως το ConceptNet και το DBpedia, για να αντλήσουμε δηλαδή κάποια πληροφορία την οποία χρησιμοποιούμε για την ανάκληση της Σημασιολογικής Πληροφορίας.
- Στο Κεφάλαιο 5 θα αναφερθούμε σε κάποια προβλήματα τα οποία παρουσιάζουν τα παραδοσιακά συστήματα ανάκτησης πληροφορίας από το διαδίκτυο, και για ποιο λόγο χρειαζόμαστε τόσο πιο σύγχρονα εργαλεία όσο και εργαλεία τα οποία λαμβάνουν υπόψιν τους περισσότερες παραμέτρους από τα παραδοσιακά συστήματα ανάκτησης πληροφορίας από το διαδίκτυο. Ενώ θα γίνει μια σύντομη ιστορική αναφορά στην ανάπτυξη των Μηχανών Αναζήτησης και στο πώς σιγά σιγά οι Μηχανές Αναζήτησης έχουν την ανάγκη να συγχρονιστούν με τις απαιτήσεις που υπάρχουν σήμερα, που δεν είναι άλλες από την ένταξη της Σημασιολογικής Πληροφορίας ως βασική παράμετρο στις Μηχανές Αναζήτησης.
- Στο Κεφάλαιο 6 θα αναφερθούμε σε πρακτικά θέματα υλοποίησης, και στον κώδικα που αναπτύχθηκε στα πλαίσια αυτής της πτυχιακής εργασίας .
- Στο Κεφάλαιο 7 θα αναφέρουμε κάποια παραδείγματα εκτέλεσης του εργαλείου και θα αναλυθούν ακριβώς τα αποτελέσματα τα οποία προκύπτουν από το εργαλείο. Ενώ επίσης θα αναλυθούν θέματα που ορίζουν το πώς σχηματίζεται το τελικό αποτέλεσμα και ποιοι παράγοντες λαμβάνονται υπόψιν για την τελική αναπαράσταση των αποτελεσμάτων.
- Τέλος, στο Κεφάλαιο 8 παρουσιάζονται τα συμπεράσματα που εξήχθησαν από την εργασία, και διατυπώνονται και ορισμένες προτάσεις για μελλοντική εργασία και μελέτη.

ΚΕΦΑΛΑΙΟ 2: ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΙΣΤΟΣ

ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΙΣΤΟΣ

2.1 ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΙΣΤΟΣ (SEMANTIC WEB)

Το Διαδίκτυο είναι ένα ανοιχτό σύστημα διασυνδεδεμένων πληροφοριών/εγγράφων πολυμεσικού περιεχομένου, που επιτρέπει στους Χρήστες του Διαδικτύου να αναζητήσουν πληροφορίες μεταβαίνοντας από το ένα έγγραφο στο άλλο.

- **Ανοιχτό σύστημα διασυνδεδεμένων πληροφοριών**, σημαίνει δημοσιοποίηση δεδομένων/πληροφοριών ώστε να είναι αλληλένδετα και να γίνουν πιο χρήσιμα στον αναγνώστη.
- **Πολυμεσικό περιεχόμενο** χαρακτηρίζουμε ένα σύνολο μέσων: κείμενο, εικόνες, υπερσυνδέσμων, βίντεο, τα οποία χρησιμοποιούνται για να αναπαραστήσουν με πολυδιάστατο τρόπο την πληροφορία που παρουσιάζεται σε ένα έγγραφο.
- Τέλος οι Χρήστες του Διαδικτύου **αναζητούν πληροφορίες** μέσω μιας μηχανής αναζήτησης η οποία έχει την δυνατότητα να ανακτά Έγγραφα του Διαδικτύου τα οποία σχετίζονται με την αναζήτηση του Χρήστη.

Μπορούμε να αναπαραστήσουμε το σημερινό Διαδίκτυο σαν ένα διασυνδεδεμένο δίκτυο υπολογιστών, σαν ένα πλέγμα υπολογιστών όπως φαίνεται στην εικόνα. Το περιεχόμενο που διαμοιράζεται στο δίκτυο είναι ουσιαστικά όχι δομημένο ή ημιδομημένο. Για παράδειγμα όταν κάποιος επισκέπτεται την σελίδα του πανεπιστημίου μας “<https://www.csd.auth.gr/>” βλέπει ένα σύνολο μέσων: κείμενο, εικόνες, σύνδεσμοι, τα οποία δεν έχουν κάποια ουσιαστική δομή και απλά προβάλλονται στον περιηγητή (browser).



Εικόνα 1 Δίκτυο Παγκόσμιου Ιστού

Η απουσία ξεκάθαρης δομής στο περιεχόμενο των πληροφοριών περιορίζει την χρήση του Ιστού από τις μηχανές/υπολογιστές σε έναν απλό μεσάζοντα μεταξύ των εγγράφων και του Χρήστη. Η νοητική συμπεριφορά που έχει αναπτυχθεί στα υπολογιστικά συστήματα με την εξέλιξη της τεχνολογίας είναι ουσιαστικά σχεδόν ανύπαρκτη στον χώρο του Διαδικτύου και η εδραίωση του είναι ένα από τα

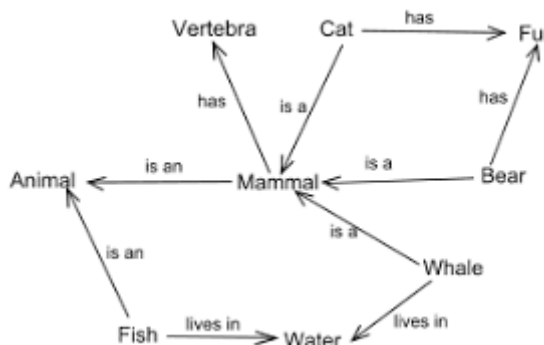
σημαντικότερα κίνητρα στην δημιουργία ενός Ιστού που θα έχει πιο αξιοποιήσιμο περιεχόμενο από τις μηχανές.

Ο *Σημασιολογικός Ιστός (ΣΙ)* η αλλιώς *Web 3.0*, είναι μια επέκταση του σημερινού Ιστού, που θα φέρει δομή στο σημασιολογικό περιεχόμενο των ιστοσελίδων. Δομή στο ουσιαστικό περιεχόμενο των σελίδων σημαίνει ότι θα περιέχει κάποια *μεταδεδομένα* (meta-data) τα οποία θα εκφράζουν σημασιολογικά τα δεδομένα που είναι κατανοητά από τις μηχανές. Ο ΣΙ είναι δηλαδή μια επέκταση του σημερινού Διαδικτύου που προσφέρει σημασιολογική δομή στο περιεχόμενο των ιστοσελίδων του σημερινού Ιστού.

Είναι σαφές ότι θα πρέπει να οριστούν κάποια τεχνολογικά πρότυπα στα οποία θα βασιστεί το μοντέλο του ΣΙ. Κάποια από αυτά είναι το μοντέλο *Resource Description Framework (RDF)*, το πρωτόκολλο *SPARQL* που αποτελεί ένα πρωτόκολλο για την ανάκτηση RDF σχημάτων, και το μοντέλο *Web Ontology Language (OWL)* που χρησιμοποιείται για την γενικότερη αναπαράσταση γνώσης σε σχήματα Οντολογιών.

2.1.1 Σημασιολογικό Δίκτυο

Το Σημασιολογικό Δίκτυο, είναι μια βάση γνώσης που αντιπροσωπεύει σημασιολογικές σχέσεις μεταξύ εννοιών στο δίκτυο αυτό. Είναι δηλαδή μια μορφή αναπαράστασης της γνώσης. Πρόκειται για ένα γράφημα του οποίου οι κορυφές είναι κάποιες οντολογίες/έννοιες (όπως για παράδειγμα “Animal” ή “Water”), και οι ακμές ορίζουν τις σημασιολογικές σχέσεις μεταξύ των κορυφών. Χρησιμοποιούνται σε εφαρμογές επεξεργασίας της φυσικής γλώσσας, με σκοπό την σημασιολογική ανάλυση και την αποσαφήνιση της λογικής. Ο λόγος που η γνώση αναπαρίσταται με γράφο είναι επειδή η γνώση που προβάλλεται κατανοείται καλύτερα αν υλοποιείται με ένα σύνολο εννοιών/κορυφών που σχετίζονται μεταξύ τους. Συνήθως στα Σημασιολογικά Δίκτυα υπάρχει και ιεραρχία.

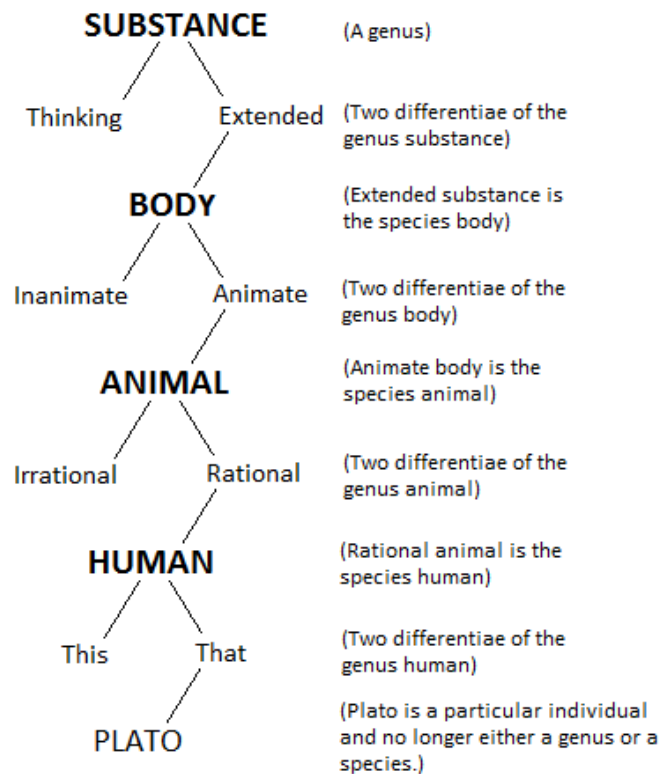


Εικόνα 2 Σημασιολογικό Δίκτυο

2.1.2 Ιστορική αναδρομή Σημασιολογικών Δικτύων

Τέτοια γραφήματα χρονολογούνται από τον 3^ο αιώνα π.Χ. όταν ο νεοπλατωνικός φιλόσοφος Πορφύριος παρουσίασε την ταξινόμηση των κατηγοριών

του Αριστοτέλη, ως προς την κλίμακα της ύπαρξης με τρόπο που αργότερα υιοθετήθηκε σε δέντρα-διαγράμματα, όπως φαίνεται στο Πορφυρικό Δέντρο της εικόνας.



Εικόνα 3 Πορφυρικό Δέντρο

Πολύ αργότερα τα Σημασιολογικά Δίκτυα χρησιμοποιήθηκαν από τον **Richard Hook Richens** ερευνητή του πανεπιστημίου Cambridge ως εργαλείο για την αυτόματη μετάφραση των φυσικών γλωσσών [1], ενώ το 1960 ο Robert F. Simmons και ο Sheldon Klein, ανεξάρτητα ο ένας από τον άλλον δημιούργησαν Σημασιολογικά Δίκτυα τα οποία χρησιμοποιήθηκαν αργότερα για να δημοσιευτούν αλγόριθμοι που με την χρήση αυτών των δικτύων σχηματίζουν προτάσεις που είναι συντακτικώς ορθές αλλά στερούν νοήματος [2] .

Το 1980, δύο ολλανδικά πανεπιστήμια, του Groningen και του Twente, άρχισαν από κοινού ένα έργο που ονομάζεται «Γράφημα Γνώσεων» που είχε το πρόσθετο χαρακτηριστικό ότι οι ακμές του προέρχονται από ένα πεπερασμένο σύνολο πιθανών σχέσεων και υποστηρίζουν βασικές αλγεβρικές πράξεις. Ενώ πολύ αργότερα ο Hermann Helbig περιγράφει το μοντέλο MultiNet [3] το οποίο αποτελεί ένα από τα πιο ολοκληρωμένα μοντέλα αναπαράσταση γνώσης για την ερμηνεία φυσικής γλώσσας. Τα επόμενα χρόνια δόθηκε έμφαση σε Σημασιολογικά Δίκτυα κοινωνικού περιεχομένου, και σημαντικό σημείο αναφοράς αποτελεί η διδακτορική διατριβή του Fawsy Bendeck το 2008 [4], που δημοσίευσε το Δίκτυο Σημασιολογικής Ομοιότητας (Semantic Similarity Network) που περιέχει εξειδικευμένες σχέσεις και αλγόριθμους για την αναπαράσταση και τον υπολογισμό της σημασιολογικής ομοιότητας.

2.1.3 Τεχνολογία για τον Σημασιολογικό Ιστό

Για να δημιουργηθεί ουσιαστική σημασιολογική δομή στο περιεχόμενο των ιστοσελίδων του Παγκοσμίου Ιστού (ΠΙ), θα πρέπει πρώτα να αναγνωριστούν όλα τα διαφορετικά αντικείμενα (objects) που υπάρχουν στον ΠΙ στα οποία θέλουμε να αποδώσουμε Σημασιολογική Πληροφορία (ΣΠ). Θα πρέπει δηλαδή να εδραιωθεί ένα μοντέλο το οποίο θα ξεχωρίζει τα διαφορετικά *Data Items* που υπάρχουν στον Ιστό. Αυτό το πετυχαίνουμε με την χρήση *URI (Uniform Resource Identifier)*, αποδίδουμε δηλαδή σε κάθε Data Item ένα διαφορετικό URI.

Για την αναγνώριση και την διατύπωση των σχέσεων μεταξύ των αντικειμένων στο καθένα από τα οποία έχουμε αποδώσει ένα ξεχωριστό αναγνωριστικό URI, χρησιμοποιείται ένα απλοϊκό μοντέλο διασυνδεδεμένου Γράφου. Όπου οι Κορυφές είναι τα ξεχωριστά αντικείμενα (Data Item) στα οποία έχει αποδοθεί ένα ξεχωριστό URI για την αναγνώριση τους, και οι Ακμές του Γράφου θα προσδιορίζουν τις Σχέσεις μεταξύ αυτών των Αντικειμένων. Ακριβώς αυτό το μοντέλο είναι η τεχνολογία RDF.

Αξίζει στο σημείο αυτό να κατανοήσουμε πως ακριβώς θα δημιουργηθεί αυτός ο Γράφος. Ένα από τα σημαντικότερα στοιχεία που οδήγησαν στην παγκόσμια αναγνώριση του σημερινού Ιστού είναι το γεγονός ότι η πληροφορία δημιουργείται από τους Χρήστες, για τους Χρήστες, αφού “Ο καθένας μπορεί να πει οτιδήποτε για οτιδήποτε”. Ένας παρόμοιος μηχανισμός υποβοηθά και στην δημιουργία του ΣΙ. Όμως στην περίπτωση του ΣΙ είναι κατανοητό ότι θα πρέπει να υπάρχει μεγαλύτερος έλεγχος στο περιεχόμενο που δημοσιοποιείται, τουλάχιστον στην αρχική φάση της δημιουργίας του.

Για αυτό τον λόγο ορίστηκαν κάποια πρωτόκολλα όπως:

1. Θα πρέπει να υπάρχει μια πρότυπη σύνταξη για τα δεδομένα και τα μεταδεδομένα που πρόκειται να δημοσιοποιηθούν.
2. Θα πρέπει να υπάρχει ένα Λεξιλόγιο (Vocabulary of Terms) το οποίο θα περιέχει την σημασιολογική πληροφορία που θα αποδοθεί στα Δεδομένα που προαναφέραμε.
3. Θα πρέπει οι Χρήστες να δημοσιοποιήσουν πολλά Δεδομένα του κανόνα ένα, με την χρήση των λεξιλογίων αυτών που αναφέραμε στον κανόνα 2.
4. Τα δεδομένα θα πρέπει να είναι άμεσα προσβάσιμα και επαναχρησιμοποιήσιμα.

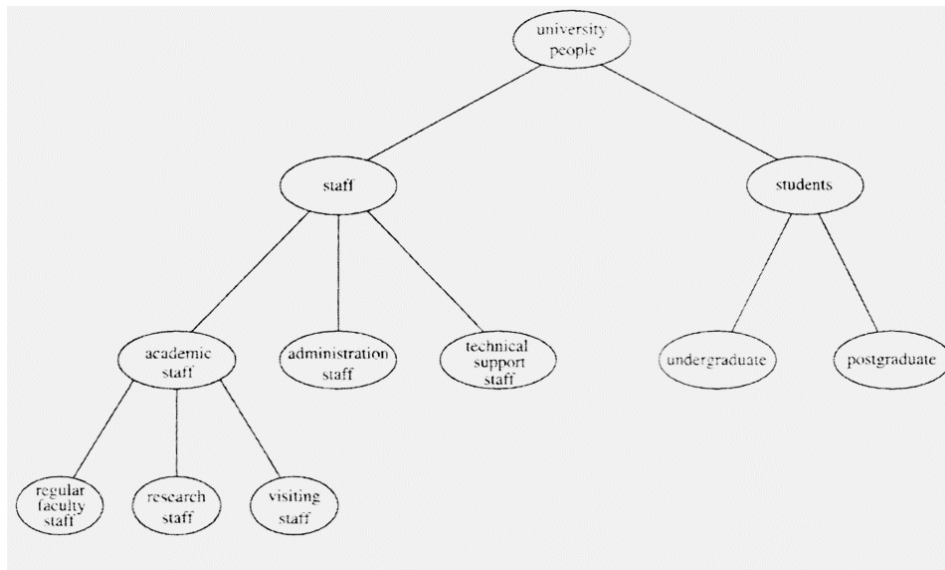
Πλέον είναι ευρέως αποδεκτό το μοντέλο που αναλύθηκε και έχουν ήδη δημοσιοποιηθεί πολλές εκατοντάδες δεδομένα και λεξιλόγια με τον τρόπο αυτό. Σημειώνεται ότι τέτοιες τεχνολογίες συναντούμε στην Ιστοσελίδα όπου συναντούμε το δίκτυο ConceptNet¹ το οποίο θα αναλυθεί σε επόμενη ενότητα.

¹ <http://conceptnet.io/>

2.1.4 Από τα Δεδομένα στην Γνώση

Αναφέρθηκε πως στόχος του ΣΙ είναι να βοηθήσει τα υπολογιστικά συστήματα να διασυνδέουν και να ενσωματώνουν δεδομένα τα οποία είναι καταναμημένα σε διάφορα σημεία στο ΠΙ. Το περιεχόμενο του ΠΙ, είναι κατά βάση κείμενο, επομένως μια τέτοια διασύνδεση δεδομένων θα ήταν αποτελεσματική αν βρεθούν συσχετίσεις μεταξύ των λέξεων. Η βασική τεχνολογία του ΣΙ που προσδιορίζει αυτές τις συσχετίσεις μπορεί να θεωρηθεί σαν ένας γράφος με ετικέτες (labeled graph), όπου οι κορυφές του θα προσδιορίζουν τα αντικείμενα και οι ακμές τις σχέσεις μεταξύ των αντικειμένων.

Το μοντέλο RDF (όπως και το μοντέλο OWL) δεν είναι απλά ένα σύστημα το οποίο περιγράφει με αποδοτικό τρόπο τα δεδομένα και τα μεταδεδομένα, αλλά εμπεριέχει σε ένα αξιοποιήσιμο βαθμό και κάποια γνώση. Επομένως δεν είναι απλά μια data-description language, αλλά και μια knowledge representation language. Η γνώση που εμπεριέχεται στα μοντέλα αυτά μπορεί να γίνει αξιοποιήσιμη με την χρήση μηχανισμών εξαγωγής συμπερασμάτων χρησιμοποιώντας όχι μόνο τις σχέσεις μεταξύ των αντικειμένων που αποδίδει το μοντέλο RDF, αλλά και την υποστήριξη της κληρονομικότητας μεταξύ των αντικειμένων σε ένα σχήμα RDF.



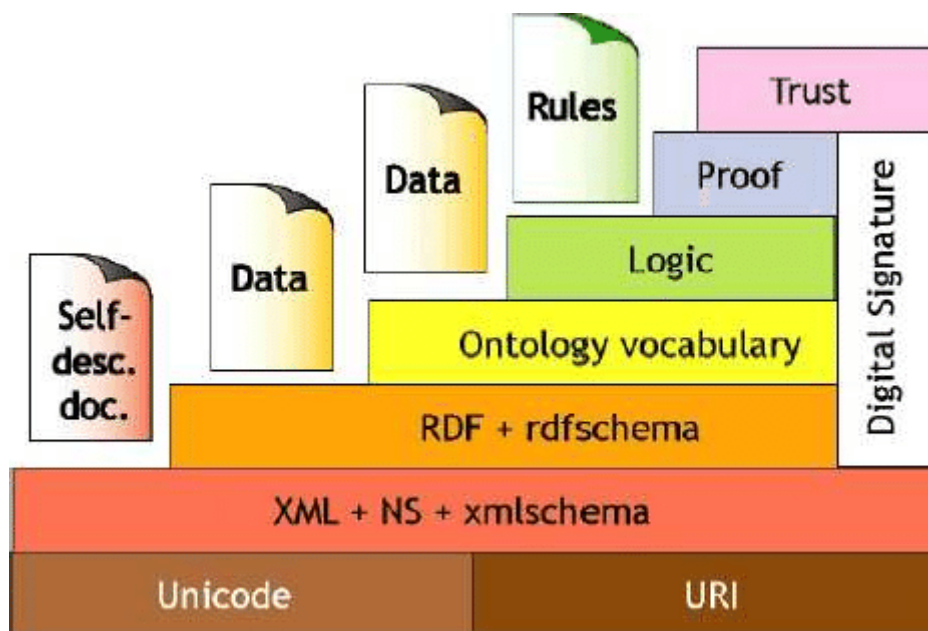
Εικόνα 4 Μοντέλο RDF και Κληρονομικότητα

2.2 LAYER APPROACH OF SEMANTIC WEB

Η ανάπτυξη του Παγκόσμιου Ιστού προχωράει με βήματα που το καθένα αποτελεί θεμέλιο για το επόμενο του. Όπως συνηθίζεται στην ανάπτυξη των τεχνολογιών στον τομέα της Πληροφορικής, υπάρχει πληθώρα από ερευνητικά κέντρα που το καθένα κινείται στην κατεύθυνση που θεωρεί την περισσότερο κρίσιμη ανάλογα με τους στόχους του και το εργαλείο που αναπτύσσει, φυσικά το εργαλείο ανάπτυξης μπορεί να είναι κοινό, και στην περίπτωση μας να είναι ο ΣΙ, αυτή η πρακτική όμως, των πολλών ερευνητών που κινούνται σε διαφορετικές κατευθύνσεις γύρω από ένα

αντικείμενο ενδιαφέροντος είναι ο λόγος που η επιστήμη της Πληροφορικής βιώνει τόσο ραγδαία ανάπτυξη τα τελευταία χρόνια. Παρόλα αυτά οι ερευνητές συμφωνούν ως προς τα θεμέλια γύρω από την ανάπτυξη του ΣΙ.

Ως προς τα θεμέλια του ΣΙ οι ερευνητές δίνουν το μοντέλο των επιπέδων.



Εικόνα 5 Επίπεδα Σημασιολογικού Ιστού

Στο κατώτερο επίπεδο βρίσκεται η XML, που αποτελεί την γλώσσα που επιτρέπει στην συγγραφή δομημένων ιστοσελίδων με λεξιλόγια που ορίζει ο Χρήστης. Επιπλέον τα URIs που χρησιμοποιούνται στην XML μπορούν να ομαδοποιηθούν από το namespace τους.

Στο αμέσως επόμενο επίπεδο υπάρχει η RDF. Η RDF είναι ένα βασικό μοντέλο με το οποίο μπορούν να γραφτούν απλές δηλώσεις που αφορούν τα αντικείμενα που υπάρχουν στον ΠΙ. Το RDF μοντέλο δεν βασίζεται στην XML αλλά η σύνταξη των RDF δηλώσεων είναι σε μορφή XML και για αυτό τον λόγο βρίσκεται πάνω από το XML επίπεδο. Μαζί με το RDF υπάρχει το RDF schema το οποίο μας επιτρέπει την ιεράρχηση των αντικειμένων του Web. Το RDF schema είναι ένα από τα βασικότερα μοντέλα με το οποίο μπορούμε να αναπαραστήσουμε τις οντολογίες, όμως δεν είναι αρκετά ισχυρό και δεν εμπεριέχει πολύπλοκες σχέσεις μεταξύ των αντικειμένων του Web.

Για αυτόν τον λόγο υπάρχει το Logic επίπεδο που επιτρέπει την δήλωση γνώσεων σε συγκεκριμένες εφαρμογές ως ένα πιο ενισχυμένο μοντέλο που αναπαριστά οντολογίες.

Το Proof Layer είναι το επίπεδο που γίνεται η συμπερασματική διαδικασία και η αναπαράσταση των proofs στις γλώσσες του ΠΙ, και η επικύρωση της απόδειξης (proof validation).

Τέλος, το Trust layer προκύπτει μέσω της χρήσης ψηφιακών υπογραφών και άλλων ειδών γνώσεων, που βασίζονται σε συστάσεις αξιόπιστων πρακτόρων ή σε οργανισμούς αξιολόγησης και πιστοποίησης. Όταν θέλουμε να αναφερθούμε σε αυτήν την κοινότητα συνήθως χρησιμοποιούμε την ορολογία “Web of Trust”. Ο λόγος που το Trust layer αποτελεί το υψηλότερο επίπεδο στην ιεραρχία έχει πολύ μεγάλη σημασία και υπονοεί ότι ο ΣΙ θα πετύχει πλήρως μόνο όταν οι χρήστες αρχίσουν να το εμπιστεύονται σε θέματα ασφαλείας και ποιότητας των πληροφοριών που παρέχονται.

Υπάρχουν δυο πολύ βασικές αρχές που πρέπει να τηρούνται στην Ανάλυση των επιπέδων του ΣΙ.

1. Θα πρέπει να υπάρχει συμβατότητα προς τα κάτω, και αυτό σημαίνει πως πράκτορες που γνωρίζουν ένα επίπεδο, θα πρέπει να είναι σε θέση να ερμηνεύσουν και όλα τα κατώτερα επίπεδα. Για παράδειγμα πράκτορες που κατανοούν την Σημασιολογία των OWL, θα πρέπει να είναι σε θέση να κατανοούν και την πληροφορία που δίνεται σε RDF ή RDF Schema.
2. Θα πρέπει να υπάρχει μερική συμβατότητα προς τα πάνω, δηλαδή η ανάπτυξη πρέπει να είναι τέτοια ώστε πράκτορες που είναι σε θέση να ερμηνεύσουν πλήρως ένα επίπεδο, θα πρέπει να είναι σε θέση να κατανοήσουν μερικώς ένα επίπεδο που βρίσκεται πάνω από αυτό. Για παράδειγμα ένας πράκτορας που είναι σε θέση να κατανοήσει πλήρως τις αναπαραστάσεις σε RDF και RDF Schema θα είναι σε θέση να κατανοήσει μερικώς και την γνώση που δίνεσαι σε μορφή OWL.

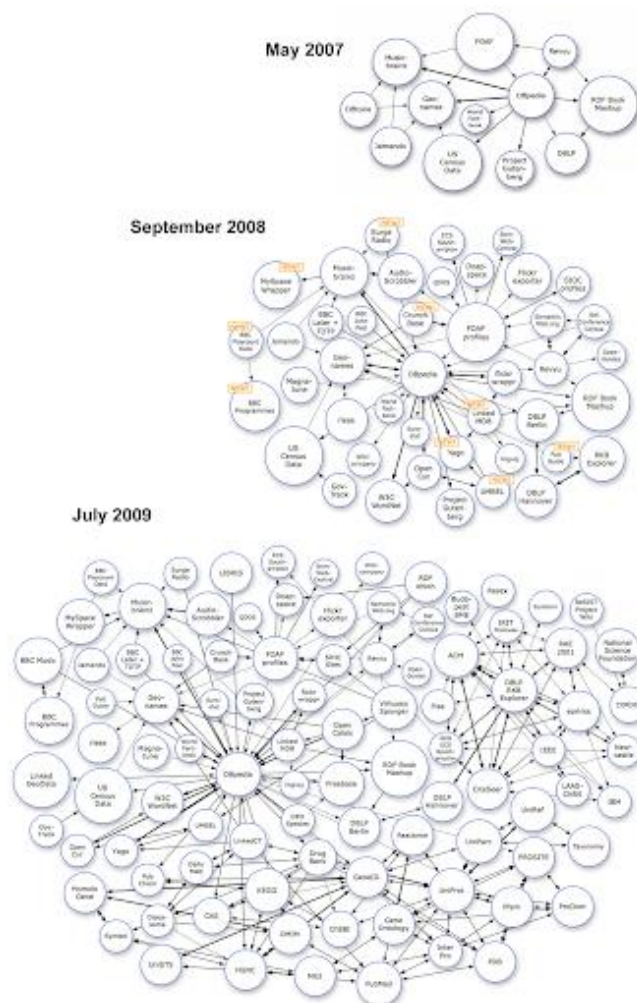
2.3 ΑΝΟΙΧΤΑ ΔΙΑΣΥΝΔΕΔΕΜΕΝΑ ΔΕΔΟΜΕΝΑ (LINKED OPEN DATA)

Ο όρος Διασυνδεδεμένα Δεδομένα, περιγράφει μια μέθοδο δημοσιοποίησης δομημένων δεδομένων ώστε να είναι αλληλένδετα και να γίνουν πιο χρήσιμα. Η μέθοδος στηρίζεται στις τεχνολογίες HTTP και URIs του ιστού. Το HTTP είναι ένα πρωτόκολλο επικοινωνίας για μεταφορά κειμένου, και μεταφέρει δεδομένα ανάμεσα σε έναν διακομιστή (server) και έναν πελάτη (client). Ενώ το URI αποτελεί ένα αναγνωριστικό ενός πόρου που βρίσκεται στο διαδίκτυο. Όμως στα Ανοιχτά Διασυνδεδεμένα Δεδομένα (ΑΔΔ) αντί να χρησιμοποιούνται οι τεχνολογίες αυτές για να εξυπηρετούν τους ανθρώπινους αναγνώστες, τις επεκτείνονται με τρόπο που να μπορούν να διαβαστούν αυτόματα από υπολογιστές, έτσι πολλά δεδομένα από διαφορετικές πηγές συνδέονται και είναι αναζητήσιμα από υπολογιστικά συστήματα.

Ο όρος επινοήθηκε από τον Tim Berners Li, αν και στην πραγματικότητα η κοινότητα της πληροφορικής είχε συλλάβει πολύ νωρίτερα την έννοια, και πιο συγκεκριμένα ο Tim Berners Lee ανέφερε τέσσερις αρχές σχετικά με τα ΑΔΔ [5]:

1. Την χρήση των URIs ώστε να προσδιοριστούν οι πόροι του διαδικτύου.

2. Την χρήση των HTTP URIs έτσι ώστε να μπορούν να αναφερθούν σε αυτούς τους πόρους και να μπορούν να αναζητηθούν από ανθρώπους και από πράκτορες.
3. Την δημοσιοποίηση πληροφοριών σχετικά με τον πόρο στον οποίο αναφερόμαστε χρησιμοποιώντας τυποποιημένους τύπους αναπαράστασης όπως RDF
4. Την ενσωμάτωση συνδέσεων με άλλα URIs με στόχο να είναι πιο πιθανή η ανακάλυψη τους από άλλες σχετικές πληροφορίες στον Ιστό και να γίνει η πληροφορία που προσφέρουν τα δεδομένα πιο πλούσια.



Εικόνα 6 Γράφημα Ανοιχτών Διασυνδεδεμένων Δεδομένων

Στόχος της παγκόσμιας κοινότητας της επιστήμης της πληροφορικής είναι να επεκταθεί ο ΠΙ δημοσιεύοντας διάφορα σύνολα δεδομένων βάση των παραπάνω αρχών. Τον Οκτώβριο του 2007 τα σύνολα δεδομένων αποτελούνταν από παραπάνω από δύο δισεκατομμύρια RDF τριάδες, οι οποίες ήταν αλληλένδετες με πάνω από δύο εκατομμύρια RDF συνδέσεις. Τον Σεπτέμβριο του 2011 αυτό το μέγεθος αυξήθηκε σε 31 δισεκατομμύρια RDF τριάδες, αλληλένδετες με περίπου 504 εκατομμύρια RDF συνδέσεις. Σήμερα η ιστοσελίδα <https://lod-cloud.net/> αναπαριστά αυτό το Project.

2.4 ΔΙΑΦΟΡΑ ΣΗΜΑΣΙΟΛΟΓΙΚΟΥ ΙΣΤΟΥ ΜΕ ΤΗΝ ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Οι περισσότερες τεχνολογίες που χρησιμοποιούνται ως βάση στον ΣΙ έχουν αναπτυχθεί στα πλαίσια του κλάδου της Τεχνητής Νοημοσύνης (ΤΝ), όμως μία από τις διαφορές είναι ότι στην ΤΝ ένας πράκτορας, δηλαδή ένα υπολογιστικό σύστημα δρα ως ένα αυτόνομο σύστημα το οποίο έχει κάποιους στόχους και κάποιες προκαθορισμένες μεθόδους και στρατηγικές για την επίτευξη των στόχων του, οι οποίες συναγωνίζονται, αλλά και σε πολλές περιπτώσεις ξεπερνούν, την ανθρώπινη ευφυΐα. Στον ΣΙ όμως δεν υπάρχει κάποια ευφυΐα ή κάποια τεχνική που πρέπει να ξεπεραστεί, αλλά στόχος είναι να δημιουργηθεί ένα περιβάλλον το οποίο θα είναι πιο αξιοποιήσιμο από τα υπολογιστικά συστήματα έτσι ώστε να είναι σε θέση να βοηθήσουν στον χρήστη να κάνει πιο εύκολα τις καθημερινές του online δραστηριότητες.

Φυσικά είναι αποδεκτό ότι με την ύπαρξη του ΣΙ θα υποβοηθηθούν και συστήματα ΤΝ. Για παράδειγμα στην περίπτωση που έχουμε ως στόχο να αναπτύξουμε έναν πράκτορα ο οποίος θα αναγνωρίζει οπτικά τα αντικείμενα που υπάρχουν σε ένα δωμάτιο. Έστω ότι ο πράκτορας μας βρίσκεται στην κουζίνα και βλέπει το εξής:



Εικόνα 7 Αναγνώριση Αντικειμένων από Ρομπότ

Ο πράκτορας έχει προγραμματιστεί να αναγνωρίζει τα αντικείμενα τα οποία υπάρχουν μέσα στην κουζίνα. Επομένως μέχρι στιγμής γνωρίζει επιτυχώς:

1. Ότι το δωμάτιο στο οποίο βρίσκεται είναι η κουζίνα
2. Ότι αναγνωρίζει 3 διαφορετικά βάζα και ξέρει πως το περιεχόμενο των 2 από αυτών είναι στιγμιαίος καφές, και καφές εσπρέσσο.

3. Αναγνωρίζει επίσης την κούπα που υπάρχει δίπλα στον νεροχύτη της κουζίνας

Και το ζητούμενο του είναι να βρει τι υπάρχει στο τρίτο βάζο. Με την κατάλληλη εξόρυξη γνώσης από τον ΣΙ μπορεί να ανακαλύψει ότι η παρουσία ζάχαρης στο τρίτο βάζο είναι μάλλον το περισσότερο πιθανό σενάριο.

2.5 Ο ΡΟΛΟΣ ΤΟΥ ΣΗΜΑΣΙΟΛΟΓΙΚΟΥ ΙΣΤΟΥ ΣΤΗΝ ΚΑΘΗΜΕΡΙΝΗ ΖΩΗ

Στον σημερινό Ιστό η αναζήτηση του περιεχομένου, έχει πολύ υψηλή ανάκληση, και η ακρίβεια και η ποιότητα των αποτελεσμάτων δεν έχει φθάσει ακόμη στο βέλτιστο επίπεδο. Η αναζήτηση στον ΠΙ έχει ως αποτέλεσμα ένα σύνολο Ιστοσελίδων, και ένα πολύ σημαντικό τμήμα του περιεχομένου του ΠΙ δεν είναι δομημένο σωστά έτσι ώστε να ανακληθεί σε ερωτήματα που σχετίζονται σε αυτό. Στον σημερινό ΠΙ οι περισσότερες ιστοσελίδες είναι γραμμένες σε γλώσσα HTML, και η HTML περιγράφει το συντακτικό περιεχόμενο και όχι το σημασιολογικό. Αν όμως καταφέρουμε να δώσουμε Σημασιολογική Δομή στο περιεχόμενο των ιστοσελίδων τότε τα υπολογιστικά συστήματα θα είναι σε θέση να κατανοήσουν την σημασιολογία των ερωτημάτων που τους θέτουμε, και δεν θα δρουν σαν απλοί διαμεσολαβητές του περιεχομένου του ΠΙ και του χρήστη.

Έστω για παράδειγμα ότι επιθυμούμε να δούμε μια ταινία, και να φάμε κάτι, και έστω ότι επιθυμούμε να δούμε μια ταινία δράσης, και να φάμε Ιταλικό φαγητό. Στον σημερινό Ιστό αναζητούμε τα ερωτήματα «ταινίες σήμερα» και «ιταλικά εστιατόρια». Αυτά είναι δύο τελείως διαφορετικά ερωτήματα και δεν υπάρχει καμία απολύτως συσχέτιση μεταξύ τους. Όμως στον ΣΙ θα θέταμε το ερώτημα με διαφορετικό τρόπο, για παράδειγμα «θέλω να δω μια ταινία δράσης και να φάω ιταλικό φαγητό», και το αποτέλεσμα θα ήταν να μας επιστρέψει ένα σύνολο κινηματογράφων της πόλης μας, τα οποία είναι κοντά σε κάποιο Ιταλικό εστιατόριο. Στο πρώτο σενάριο, θα έπρεπε εμείς να εξαντλήσουμε όλες τις επιλογές που θα είχαμε σχετικά με το ερώτημα «σε ποιο σινεμά θα δω την ταινία δράσης που παίζει σήμερα» με γνώμονα το ποιο από αυτά τα σινεμά είναι κοντά σε κάποιο Ιταλικό εστιατόριο.



Εικόνα 8 Γράφος Σημασιολογικής Πληροφορίας

Ο σημερινός ιστός αποτελείται από ένα σύνολο δεδομένων στα οποία έχει γίνει κάποια προεπεξεργασία για να είναι πιο εύκολα αναζητήσιμα. Ο ΣΙ όμως, δεν έχει να κάνει μόνο με ένα σύνολο δεδομένων τα οποία επιθυμούμε να ανακτήσουμε. Σχετίζεται με αντικείμενα και μπορεί να αναγνωρίσει ανθρώπους, μέρη, γεγονότα, εταιρείες, προϊόντα, ταινίες, και γενικότερα μπορεί να αναπαραστήσει όλον τον πραγματικό κόσμο. Ο ΣΙ ιστός αναγνωρίζει αυτά τα διαφορετικά αντικείμενα, τα ιεραρχεί και τους αποδίδει τον ορισμό «Οντολογίες».

ΚΕΦΑΛΑΙΟ 3: ΤΕΧΝΟΛΟΓΙΕΣ RDF, SPARQL, OWL

ΤΕΧΝΟΛΟΓΙΕΣ RDF, SPARQL, OWL

Η επιτυχία του ΠΙ προκύπτει εν μέρη από την δύναμη ύπαρξης ξεκάθαρων μηχανισμών που υπάρχουν για να υποστηρίζεται η σύνθεση και η ανταλλαγή της πληροφορίας που δημοσιοποιείται σε αυτόν. Η γλώσσα HTML είναι η βασική γλώσσα με την οποία γράφονται οι ιστοσελίδες του ΠΙ. Η HTML επιτρέπει σε οποιονδήποτε να γράψει μια ιστοσελίδα και να την δημοσιοποιήσει στον ΠΙ. Φυσικά η HTML εγγυάται ότι η ιστοσελίδα θα προβληθεί σωστά από οποιονδήποτε περιηγητή ιστού. Είναι προφανές ότι θα πρέπει να αναπτυχθούν τεχνολογίες που θα έχουν ανάλογη δύναμη και να αποτελέσουν θεμέλια για την δημιουργία του ΣΙ. Έχουν αναπτυχθεί πολλές τέτοιες τεχνολογίες από μέλη της παγκόσμιας κοινότητας της πληροφορικής και μερικές από αυτές αναλύονται στο παρόν κεφάλαιο.

3.1 HTML ΚΑΙ RDF

Οι σελίδες στον ΠΙ γράφονται σε γλώσσα HTML. Η γλώσσα HTML απαρτίζεται από 3 πολύ σημαντικά συστατικά στοιχεία:

1. Το συντακτικό της
2. Ένα Μοντέλο Δεδομένων (Data Model)
3. Το σημασιολογικό περιεχόμενο που προσφέρει

Το σημασιολογικό περιεχόμενο της HTML όμως δίνει πληροφορίες μόνο για το πως πρέπει να ερμηνεύσει αποδοτικά ένας περιηγητής τα συστατικά τμήματα τα οποία βρίσκονται μέσα σε μια HTML ιστοσελίδα. Έτσι ο περιηγητής, χρησιμοποιώντας αυτό το σημασιολογικό περιεχόμενο παρουσιάζει την ιστοσελίδα με τέτοιο τρόπο έτσι ώστε να μπορεί να την κατανοήσει ο αναγνώστης.

Για παράδειγμα στην παρακάτω HTML ιστοσελίδα:

```
<html>
  <head>
    <title>Apartments for Rent</title>
  </head>
  <body>
    <ol>
      <li> Studio in Thessaloniki
      <li> 3 bedroom Apartment in Thessaloniki
    </ol>
  </body>
</html>
```

Ο περιηγητής αναγνωρίζει την ΣΠ που υπάρχει στην παραπάνω HTML Ιστοσελίδα, ότι για παράδειγμα το συστατικό <title>..<</title> αφορά έναν τίτλο και θα το παρουσιάσει με ειδική γραμματοσειρά και ότι τα συστατικά αφορούν εγγραφές μιας ταξινομημένης λίστας και θα παρουσιαστούν μια ειδική μορφολογία. Όμως η HTML δεν μας προσφέρει κάποια ΣΠ η οποία μπορεί να χρησιμοποιηθεί από υπολογιστικά

συστήματα ή εφαρμογές που θέλουν να εξάγουν πιο χρήσιμη γνώση ή να εφαρμόσουν τεχνικές που απαιτούν δεδομένα με πλούσια ΣΠ.

Στον ΣΙ υπάρχει η ανάγκη η ΣΠ που υπάρχει σε μια ιστοσελίδα να είναι κάτι πιο πλούσιο, καθώς υπάρχει η απαίτηση, όχι μόνο να γίνεται η σωστή αναπαράσταση των εγγράφων για τους χρήστες, αλλά να μπορεί να γίνει αξιοποίηση της πληροφορίας που υπάρχει στην HTML ιστοσελίδα από διάφορων τύπων εφαρμογές και υπολογιστικά συστήματα. Το μοντέλο RDF χρησιμοποιείται για να ξεπεραστεί αυτό το φράγμα στην ΣΠ την οποία εμπεριέχουν οι ιστοσελίδες που είναι γραμμένες σε HTML γλώσσα.

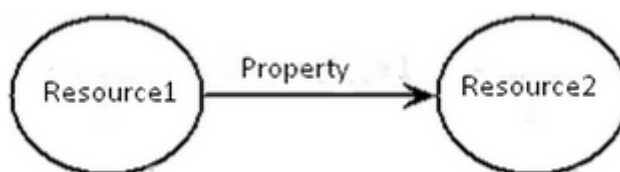
3.2 RESOURCE DESCRIPTION FRAMEWORK (RDF)

Το μοντέλο **RDF**, χρησιμοποιεί 4 βασικές έννοιες:

1. Resources
2. Properties
3. Statements
4. Graphs

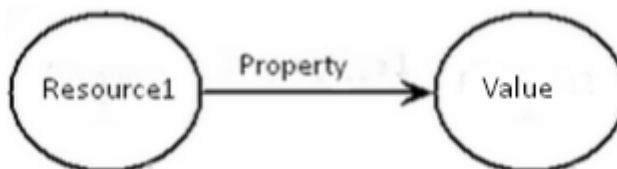
Έγινε αναφορά ότι κάθε διαφορετικό Data Item δηλαδή κάθε διαφορετικό Resource στον Ιστό, θα προσδιορίζεται από ένα μοναδικό URI. Η χρήση αυτών των URI μας επιτρέπει να δημιουργήσουμε ένα καθολικό naming scheme.

Τα Properties Περιγράφουν σχέσεις ανάμεσα σε δύο Resources.



Εικόνα 9 Σχέση RDF

Τα Properties αναγνωρίζονται επίσης με την απόδοση ενός μοναδικού URI σε κάθε property, ενώ αυτή η ανάθεση που γίνεται στο παραπάνω σχήμα ονομάζεται Statement. Στην γενική του μορφή ένα Statement είναι μια τριπλέτα της μορφής:



Εικόνα 10 Τριπλέτα RDF

Ενώ φυσικά ένα Value μπορεί να είναι κι αυτό Resource. Παράδειγμα τριπλέτας:



Εικόνα 11 Παράδειγμα RDF Τριπλέτας

Δηλώνει ότι η Ακρόπολη βρίσκεται στην Αθήνα.

Τέλος ένας γράφος κάνει αναπαράσταση και διασυνδέει με Properties πολλές τέτοιες τριπλέτες μεταξύ τους. Έτσι μπορεί πολύ εύκολα να αναπαρασταθεί μια οντολογία και να χρησιμοποιηθεί για τον ορισμό Semantics σε μια Ιστοσελίδα (Σχήμα σελίδας 28). Επομένως με την χρήση όσων μας προσφέρει το μοντέλο RDF, εισάγουμε ΣΠ στο ήδη υπάρχον περιεχόμενο μιας Ιστοσελίδας.

Αρχικά η Ιστοσελίδα μας θα ήταν σε αυτή την μορφή:

```
<div>
  <h1>Avatar</h1>
  <span>Director: James Cameron (born August 16, 1954)</span>
  <span>Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html">Trailer</a>
</div>
```

Από όσα έχουμε ήδη πει είναι κατανοητό πως η προβολή και η ερμηνεία των συστατικών θα γίνει σωστά από έναν περιηγητή, και ο χρήστης θα μπορέσει αποδοτικά να αναγνωρίσει ότι η ιστοσελίδα αφορά μια ταινία με τίτλο "Avatar" ότι ο σκηνοθέτης είναι ο James Cameron, ότι το είδος της ταινίας είναι "Science fiction" και ότι υπάρχει ένας υπερσύνδεσμος που οδηγεί στο trailer της ταινίας. Όμως μια μηχανή δεν μπορεί να αναγνωρίσει τίποτα από τα παραπάνω αφού δεν υπάρχει η απαραίτητη ΣΠ.

Έστω ότι το URI το οποίο θα χρησιμοποιηθεί ως αναγνωριστικό του Data Item ταινία είναι το <http://schema.org/Movie>. Τότε θα εισάγουμε το URI αυτό στο <div>...</div> συστατικό της ιστοσελίδας και κατευθείαν έχουμε προσθέσει σημασιολογικό περιεχόμενο στο συγκεκριμένο συστατικό, το οποίο είναι αναγνωρίσιμο από μια μηχανή.

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1>Avatar</h1>
  <span>Director: James Cameron (born August 16, 1954)</span>
  <span>Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html">Trailer</a>
</div>
```

Με τον ίδιο τρόπο προσθέτουμε σημασιολογικό περιεχόμενο σε όλα τα Components που έχουμε στο συγκεκριμένο παράδειγμα

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <div itemprop="director" itemscope itemtype="http://schema.org/Person">
    Director: <span itemprop="name">James Cameron</span> (born <span itemprop="b
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="../../movies/avatar-theatrical-trailer.html" itemprop="trailer">Traile
</div>
```

Είναι προφανές ότι τα URI που χρησιμοποιήθηκαν για την αναγνώριση των διαφορετικών Resources (<http://schema.org/Person>) και τα Properties (`itemprop="birthDate"`, `itemprop="name"`, `itemprop="director"`, `itemprop="trailer"`), προσθέτουν το απαραίτητο ΣΠ στην HTML ιστοσελίδα. Πλέον η ιστοσελίδα μας είναι κατανοητή όχι μόνο από έναν άνθρωπο, αλλά και από μια εφαρμογή.

3.3 SPARQL

Η SPARQL είναι μια Γλώσσα Ερωτημάτων (Query Language QL) για δομές RDF. Σε μια Βάση Δεδομένων (ΒΔ) με SQL μπορούν να ανακτηθούν δεδομένα που περιέχονται στην βάση αυτή. Με παρόμοιο τρόπο, σε μια ΒΔ που περιέχει δεδομένα με ΣΠ μπορούν να ανακτηθούν με την χρήση SPARQL ερωτημάτων. Πρωτοεμφανίστηκε το 2008 και χρησιμοποιείται και σήμερα για την πρόσβαση την εισαγωγή και την ανάκτηση ΣΠ σε ΒΔ.

Για να εκτελέσει κανείς ένα SPARQL ερώτημα χρειάζεται το ανάλογο λογισμικό, που στην περίπτωση της SPARQL γλώσσας ονομάζεται *triple store*. Επομένως ένα *triple store* είναι μια ΒΔ για δομές RDF. Η σύνδεση της SPARQL με το *triple store* γίνεται μέσω ενός *endpoint* το οποίο διατίθεται από το *triple store*. Ένα τέτοιου είδους *endpoint* προσφέρει η DBpedia² έτσι ώστε να είναι διαθέσιμο στους χρήστες της γλώσσας SPARQL να θέσουν ερωτήματα (queries) στην αντίστοιχη DBpedia, που στην περίπτωση της αυτή είναι ένα σύνολο RDF σχημάτων που έχει δημιουργηθεί από δεδομένα του Wikipedia. Υπάρχει μια πλήρης λίστα για όλα τα SPARQL endpoints στην ιστοσελίδα³.

3.3.1 Βασικά SPARQL ερωτήματα

Έστω ότι θέλουμε να ανακτήσουμε πληροφορίες σχετικά με τον όρο «Tea» μέσω SPARQL ερωτήματος στο endpoint της DBpedia. Αρχικά, μπορούμε να συνδεθούμε με κάποιον browser στην διεύθυνση: <https://dbpedia.org/sparql> όπου θα εμφανιστεί το παρακάτω περιβάλλον διάδρασης

² dbpedia.org/sparql

³ CKAN.org

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)
http://dbpedia.org

Query Text

(Security restrictions of this server do not allow you to retrieve remote RDF data, see [details](#).)

Results Format: JSON

Execution timeout: 30000 milliseconds (values less than 1000 are ignored)

Options:

- ☒ Strict checking of void variables
- ☐ Log debug info at the end of output (has no effect on some queries and output formats)
- ☐ Generate SPARQL compilation report (instead of executing the query)

(The result can only be sent back to browser, not saved on the server, see [details](#))

Run Query Reset

Copyright © 2020 [OpenLink Software](#)
Virtuoso version 07.20.3232 on Linux (x86_64-generic-linux-glibc25), Single Server Edition

Εικόνα 12 SPARQL Endpoint

3.3.2 Εντολή *SELECT* και *WHERE*

Στην εφαρμογή, υπάρχει ένα πεδίο όπου θα τοποθετηθεί το SPARQL ερώτημα, ενώ έχουμε την επιλογή στο πεδίο «Results Format» να επιλέξουμε τον τρόπο με τον οποίο επιθυμούμε να μας επιστραφούν τα αποτελέσματα του ερωτήματος μας. Δυο από τις βασικότερες πράξεις σε όλες τις γλώσσες ερωτημάτων, είναι η *SELECT* και η *WHERE*. Έστω ότι θέλουμε να επιλέξουμε για τον όρο «Tea» το πεδίο *abstract*, το οποίο φαίνεται στην εικόνα που ακολουθεί:

About: Tea

An Entity of Type : [Τεα](#), from Named Graph : <http://dbpedia.org>, within Data Space : dbpedia.org

(This article is about the beverage. For other uses, see [Tea \(disambiguation\)](#).)("Cup of tea" redirects here. For other uses, see [Cup of Tea](#).) Tea is an aromatic beverage commonly prepared by pouring hot or boiling water over cured leaves of the *Camellia sinensis*, an evergreen shrub native to Asia. After water, it is the most widely consumed drink in the world. There are many different types of tea; some teas, like Darjeeling and Chinese greens, have a cooling, slightly bitter, and astringent flavour, while others have vastly different profiles that include sweet, nutty, floral or grassy notes.

Property	Value
dbpedia:abstract	<ul style="list-style-type: none"> (This article is about the beverage. For other uses, see Tea (disambiguation).)("Cup of tea" redirects here. For other uses, see Cup of Tea.) Tea is an aromatic beverage commonly prepared by pouring hot or boiling water over cured leaves of the <i>Camellia sinensis</i>, an evergreen shrub native to Asia. After water, it is the most widely consumed drink in the world. There are many different types of tea; some teas, like Darjeeling and Chinese greens, have a cooling, slightly bitter, and astringent flavour, while others have vastly different profiles that include sweet, nutty, floral or grassy notes. Tea originated in southwestern China, where it was used as a medicinal drink. It was popularized as a recreational drink during the Chinese Tang dynasty, and tea drinking spread to other East Asian countries. Portuguese priests and merchants introduced it to the West during the 16th century. During the 17th century, drinking tea became fashionable among Britons, who started large-scale production and commercialization of the plant in India to bypass a Chinese monopoly at that time. The phrase herbal tea usually refers to infusions of fruit or herbs made without the tea plant, such as steeps of rosehip, chamomile, or rooibos. These are also known as tisanes or herbal infusions to distinguish them from "tea" as it is commonly understood. ^(en)
dbpedia:thumbnail	<ul style="list-style-type: none"> wiki-commons:Special:FilePath/Tea_leaves_steeping_in_a_zhong_cai_05.jpg?width=300
dbpedia:wikiPageExternalLink	<ul style="list-style-type: none"> https://books.google.com/books?id=DJ2I_bX6WTUC&pg=PA8#v=onepage&q&f=false https://books.google.com/books?id=_TR_PQAACAAJ

Εικόνα 13 Αποτελέσματα SPARQL Ερωτήματος για τον όρο «Tea»

Το πεδίο «abstract» αποτελεί έναν ορισμό, για το τι είναι «Tea». Για να ανακτήσουμε το συγκεκριμένο πεδίο, το SPARQL ερώτημα που πρέπει να γράψουμε στο endpoint της DBpedia είναι το εξής:

```
SELECT ?abstract
WHERE {<http://dbpedia.org/resource/Tea> dbo:abstract ?abstract }
```

Το ερώτημα εμπεριέχει την εντολή SELECT, που σημαίνει επέλεξε και την μεταβλητή ?abstract, δηλαδή σε φυσική γλώσσα το ερώτημα σημαίνει: «επέλεξε το abstract από το <http://dbpedia.org/resource/Tea>. Έχουμε επιλέξει στο endpoint της DBpedia να μας επιστρέψει το αποτέλεσμα σε JSON μορφή, και το αποτέλεσμα είναι το εξής:

```
"(This article is about the beverage. For other uses, see Tea
(disambiguation).) ("Cup of tea" redirects here. For other uses, see
Cup of Tea.)
```

```
Tea is an aromatic beverage commonly prepared by pouring hot or
boiling water over cured leaves of the Camellia sinensis, an
evergreen shrub native to Asia. After water, it is the most widely
consumed drink in the world. [...]."@en
```

```
"El té es una infusión de las hojas y brotes de la planta del té
(Camellia sinensis). La popularidad de esta bebida es solamente
sobrepasada por el agua. Su sabor es fresco, ligeramente amargo y
astringente; este gusto es agradable para mucha gente. Se argumenta
que el consumo de té (especialmente verde) es benéfico para la
salud por contener antioxidantes, flavanoles, flavonoides,
catequinos y polifenoles. [...]."@es
```

```
"Le thé est une boisson aromatique préparée en infusant des
feuilles et des bourgeons de théier, un arbuste à feuilles
persistantes originaire d'Asie, et pouvant être bue chaude ou
froide. [...]."@fr
```

Μας έχει επιστρέψει το πεδίο «abstract» για την λέξη «Tea» σε τρεις διαφορετικές γλώσσες, στα αγγλικά, στα ισπανικά, και στα γαλλικά. Φυσικά με τον τρόπο αυτόν, μπορούμε αν πάρουμε οποιοδήποτε από τα πεδία θέλουμε. Για παράδειγμα μπορούμε να ζητήσουμε το πεδίο με την εικόνα (thumbnail) του όρου «Tea» ως εξής:

```
SELECT ?thumbnail
WHERE {<http://dbpedia.org/resource/Tea>dbo:thumbnail?thumbnail }
```

Με το αποτέλεσμα:



Εικόνα 14 Thumbnail από ερώτημα SPARQL για τον όρο «Tea»

3.3.3 Εντολή *LIMIT*

Αν δεν επιθυμούμε να διατηρήσουμε τόσα πολλά αποτελέσματα, έστω ότι πχ θέλουμε μόνο τα 2 πρώτα, τότε μπορούμε να χρησιμοποιήσουμε την εντολή *LIMIT* για να κρατήσουμε μόνο τα πρώτα δυο αποτελέσματα ως εξής:

```
SELECT ?abstract
WHERE { <http://dbpedia.org/resource/Tea> dbo:abstract ?abstract }
LIMIT 2
```

Αποτελέσματα:

"(This article is about the beverage. For other uses, see Tea (disambiguation).) ("Cup of tea" redirects here. For other uses, see Cup of Tea.)

Tea is an aromatic beverage commonly prepared by pouring hot or boiling water over cured leaves of the *Camellia sinensis*, an evergreen shrub native to Asia. After water, it is the most widely consumed drink in the world. [...]"@en

"El té es una infusión de las hojas y brotes de la planta del té (*Camellia sinensis*). La popularidad de esta bebida es solamente sobrepasada por el agua. Su sabor es fresco, ligeramente amargo y astringente; este gusto es agradable para mucha gente. Se argumenta que el consumo de té (especialmente verde) es benéfico para la salud por contener antioxidantes, flavanoles, flavonoides, catequinos y polifenoles. [...]"@es

Όπως φαίνεται στα αποτελέσματα έχουμε μόνο αυτά της αγγλικής και της ισπανικής γλώσσας.

3.3.4 Εντολή *FILTER*

Με την εντολή *FILTER* , μπορούμε να φιλτράρουμε τα αποτελέσματα που επιθυμούμε. Για παράδειγμα μπορούμε να κρατήσουμε μόνο τα αποτελέσματα τα οποία είναι στην ισπανική γλώσσα ως εξής:

```
SELECT ?abstract
WHERE { <http://dbpedia.org/resource/Tea> dbo:abstract ?abstract
       FILTER (lang(?abstract) = 'es') }
```

"El té es una infusión de las hojas y brotes de la planta del té (Camellia sinensis). La popularidad de esta bebida es solamente sobrepasada por el agua. Su sabor es fresco, ligeramente amargo y astringente; este gusto es agradable para mucha gente. Se argumenta que el consumo de té (especialmente verde) es benéfico para la salud por contener antioxidantes, flavanoles, flavonoides, catequinos y polifenoles. [...]."@es

3.3.5 Εντολή *PREFIX*

Μπορούμε να χρησιμοποιήσουμε την εντολή *PREFIX* για να ορίσουμε ένα πρόθεμα που θα χρησιμοποιούμε στο πρόγραμμα. Για παράδειγμα αν θέλουμε να χρησιμοποιήσουμε το πρόθεμα *dbpedia* το οποίο θα αναφέρεται στο <http://dbpedia.org/resource/>

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
SELECT ?abstract
WHERE { dbpedia:Tea dbo:abstract ?abstract
        FILTER (lang(?abstract) = 'en') }
```

Και στη συνέχεια του κώδικα για να αναφερθούμε σε αυτό χρησιμοποιούμε μόνο το πρόθεμα και όχι ολόκληρο το URI.

3.3.6 Μερικά ακόμη πιο σύνθετα παραδείγματα

1. Έστω ότι ψάχνουμε να βρούμε τους ανθρώπους που γεννήθηκαν στην Ελλάδα πριν το 1900. Χρειαζόμαστε τα πεδία, name, birth, death και person.

SPARQL-ερώτημα:

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX : <http://dbpedia.org/resource/>
```

```

SELECT ?name ?birth ?death ?person WHERE { ?person dbo:birthPlace
:Greece . ?person dbo:birthDate ?birth . ?person foaf:name ?name .
?person dbo:deathDate ?death . FILTER (?birth < "1900-01-
01"^^xsd:date) . } ORDER BY ?name

```

Η καινούργια εντολή που χρησιμοποιείται σε αυτό το παράδειγμα είναι η ORDER BY, η οποία στο συγκεκριμένο παράδειγμα ταξινομεί τα αποτελέσματα ως προς το όνομα. Προφανώς θα μπορούσαμε να επιλέξουμε ταξινόμηση ως προς οποιοδήποτε πεδίο επιθυμούμε να λάβουμε ταξινομημένα τα αποτελέσματα μας.

Αποτελέσματα:

"(Αλέξανδρος Ζαΐμης)"@en	1855- 11-09	1936-09-15	http://dbpedia.org/resource/Alexandros_Zaimis
"(Ελευθέριος Βενιζέλος)"@e n	1864- 08-23	1936-03-18	http://dbpedia.org/resource/Eleftherios_Venizelos
"(Στέφανος Στεφανόπουλος)"@en	1898- 07-03	"1982-10- 4"^^<http://www.w3.org/ 2001/XMLSchema#date>	http://dbpedia.org/resource/Stefanos_Stefanopoulos
"(Χαρίλαος Τρικούπης)"@e n	1832- 06-11	"1896-3- 30"^^<http://www.w3.org /2001/XMLSchema#date>	http://dbpedia.org/resource/Charilaos_Trikoupis

- Εστω ότι ψάχνουμε να βρούμε τους ποδοσφαιριστές, οι οποίοι έχουν γεννηθεί σε μια χώρα με περισσότερους από δέκα εκατομμύρια κατοίκους, που έπαιξαν ως τερματοφύλακες σε μια ομάδα που παίζει σε γήπεδο με περισσότερες από τριάντα χιλιάδες θέσεις και η χώρα του συλλόγου είναι διαφορετική από την χώρα γέννησης του παίχτη. Τότε το ερώτημα που θα θέταμε στο endpoint μας θα ήταν το εξής:

```

PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT distinct ?soccerplayer ?countryOfBirth ?team ?countryOfTeam
?stadiumcapacity
{
?soccerplayer a dbo:SoccerPlayer ;
    dbo:position|dbp:position
<http://dbpedia.org/resource/Goalkeeper_(association_football)> ;
    dbo:birthPlace/dbo:country* ?countryOfBirth ;
    #dbo:number 13 ;
    dbo:team ?team .
    ?team dbo:capacity ?stadiumcapacity ; dbo:ground ?countryOfTeam.
    ?countryOfBirth a dbo:Country ; dbo:populationTotal ?population.
}

```

```

?countryOfTeam a dbo:Country .
FILTER (?countryOfTeam != ?countryOfBirth)
FILTER (?stadiumcapacity > 30000)
FILTER (?population > 10000000)
} order by ?soccerplayer

```

Αποτελέσματα:

http://dbpedia.org/resource/Abdellah_Benabdelah	http://dbpedia.org/resource/Algeria	http://dbpedia.org/resource/Wydad_Casablanca	http://dbpedia.org/resource/Morocco	"67000"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger>
---	---	---	---	--

http://dbpedia.org/resource/Airton_Moraes_Michellon	http://dbpedia.org/resource/Brazil	http://dbpedia.org/resource/FC_Red_Bull_Salzburg	http://dbpedia.org/resource/Austria	"31000"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger>
---	---	---	---	--

http://dbpedia.org/resource/Alain_Gouaméné	http://dbpedia.org/resource/Ivory_Coast	http://dbpedia.org/resource/Raja_Casablanca	http://dbpedia.org/resource/Morocco	"67000"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger>
---	---	---	---	--

http://dbpedia.org/resource/Allan_McGregor	http://dbpedia.org/resource/United_Kingdom	http://dbpedia.org/resource/Beşiktaş_J.K.	http://dbpedia.org/resource/Turkey	"41903"^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger>
---	---	---	---	--

...

...

...

...

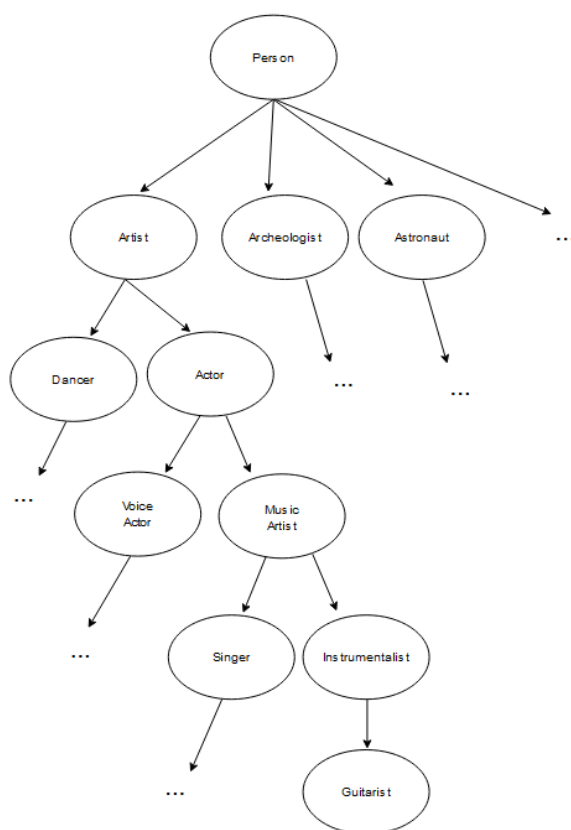
...

Σε αυτό το σημείο αξίζει να αναφέρουμε ότι είναι ξεκάθαρη η ισχύς της ΣΠ και ότι σύνθετα ερωτήματα που αφορούν οντότητες του πραγματικού κόσμου, μπορούν να απαντηθούν από μηχανές αν διασυνδέσουμε τα δεδομένα με τέτοιο τρόπο έτσι ώστε να είναι αναζητήσιμα και επεξεργάσιμα από τα υπολογιστικά συστήματα.

3.4 ΟΝΤΟΛΟΓΙΑ (ONTOLOGY)

Σύμφωνα με τη Wikipedia⁴, «Οντολογία είναι ένας τυπικός και σαφής ορισμός μιας κοινής και συμφωνημένης μορφοποίησης που αφορά σε ένα πεδίο ενδιαφέροντος». Με άλλα λόγια οι οντολογίες είναι δομημένα πλαίσια οργάνωσης της πληροφορίας, είναι πλαίσια στα οποία κατηγοριοποιούμε οποιονδήποτε ορισμό αντικειμένου, προσώπου, μέρους, επιστήμης κλπ. Ενώ οι οντολογίες είναι συνδεδεμένες μεταξύ τους μέσα από μια ιεραρχία οντολογικών επιπέδων. Χρησιμοποιούνται στους τομείς της ΤΝ, στον ΣΙ, στην Βιοπληροφορική, ανάμεσα σε άλλους τομείς. Σε γενικότερο πλαίσιο οι Οντολογίες ορίζουν μια μορφή αναπαράστασης γνώσης του κόσμου στον οποίο υπάρχουμε.

Παράδειγμα Οντολογιών:



Εικόνα 15 Παράδειγμα Οντολογιών

Στο παράδειγμα φαίνεται τόσο η Ιεραρχία των κλάσεων όσο και το ότι όλες αυτές οι ξεχωριστές κλάσεις συνθέτουν μια ενιαία οντολογία, με ανώτερη κλάση την «Person». Η γνώση που μας προσφέρει η αναπαράσταση του γνωστού κόσμου σε

⁴ <https://en.wikipedia.org/wiki/Ontology>

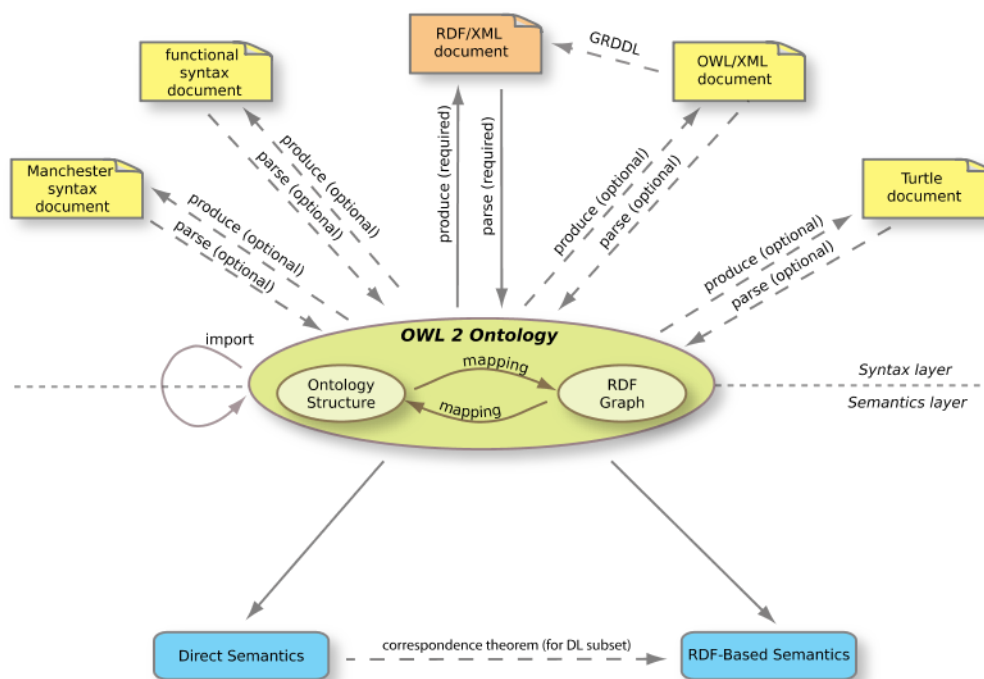
οντολογίες είναι άμεσα αξιοποιήσιμη από τις μηχανές ενώ επίσης αξιοποιήσιμο είναι το σημασιολογικό περιεχόμενο που προσφέρουν.

3.5 ΓΛΩΣΣΑ ΟΝΤΟΛΟΓΙΑΣ ΙΣΤΟΥ (WEB ONTOLOGY LANGUAGE -OWL)

Η Γλώσσα Οντολογίας Ιστού (OWL) είναι μια οικογένεια γλωσσών αναπαράστασης γνώσης σχετικά με τις οντολογίες. Χαρακτηρίζονται από επίσημη σημασιολογία (*formal semantics*), είναι βασισμένη στο πρότυπο XML και αφορά αντικείμενα που είναι σε μορφή RDF. Οι OWL, όπως και η τεχνολογία RDF, έχουν προσελκύσει σημαντικό ακαδημαϊκό ενδιαφέρον. Επίσημα οι γλώσσες OWL εξελίχθηκαν σε εμπορική ιδέα από το W3C που ανακοίνωσε την επίσημη έκδοση της OWL στις 27 Οκτωβρίου του 2009. Αυτή η έκδοση ονομάζεται OWL2, και παρέχει κλάσεις, ιδιότητες, αναγνωριστικά, και δεδομένα τα οποία αποθηκεύονται ως έγγραφα του Σημασιολογικού Ιστού. Στον ιστότοπο: <https://www.w3.org/TR/2012/REC-owl2-overview-20121211/> υπάρχει αναρτημένη η επίσημη έκδοση της OWL 2 από την W3C.

Ο λόγος που υπάρχουν οι Γλώσσες Οντολογιών, είναι για να διατυπώσουμε με ακρίβεια και με πλούσιο τρόπο την γνώση που έχουμε γύρω από μια οντολογία να φτιάξουμε δηλαδή ένα μοντέλο γνώσης. Τα μοντέλα γνώσης που μπορούμε να τυποποιήσουμε μόνο με RDF υστερούν πολύ συγκριτικά με αυτά της OWL2 αφού τα OWL2 είναι αυτά που εισάγουν άμεσα τα σημασιολογικά δεδομένα. Αν ρίξουμε μια προσεκτική ματιά στην Εικόνα μπορούμε να θεωρήσουμε ότι ο λόγος που το επίπεδο «Ontology vocabulary» βρίσκεται πάνω από το RDF/RDFS είναι επειδή το OWL2 έρχεται και προσθέτει αυτά τα σημασιολογικά δεδομένα, που αφορούν τον πλουσιότερο προσδιορισμό των κλάσεων, τον ιεραρχιών και των ιδιοτήτων που έχουν αποδοθεί από τα RDF/RDFS. Για παράδειγμα, αν ξέρουμε από RDF ότι «A isMarriedTo B», τότε αυτό υποδηλώνει ότι «B isMarriedTo A». Η αν ξέρουμε ότι «C isAnAncestorOfD» και «D isAnAncestorOf E» τότε ξέρουμε και ότι «C isAnAncestorOf E». Επίσης με τις OWL γλώσσες μπορούμε να καθορίσουμε ότι μια οντολογία είναι ίδια με μια άλλη. Για παράδειγμα ο «Elvis Presley» στη Wikipedia, είναι ακριβώς ο ίδιος με τον «Elvis Presley» στο BBC. Η παραδοχή αυτή από έναν άνθρωπο προφανώς είναι κάτι το αυτονόητο (Common Sence) , μια μηχανή όμως έχει την ανάγκη από τον πλούσιο καθορισμό της Σημασιολογίας των Οντολογιών, και αυτό αποσκοπούν να πετύχουν οι Γλώσσες Οντολογιών.

Σκοπός της OWL2 είναι να διευκολύνει την ανάπτυξη οντολογιών, και την κοινή χρήση τους μέσω του Διαδικτύου, με απώτερο στόχο να κάνει το περιεχόμενο του Ιστού προσβάσιμο σε μηχανήματα. Στο σχήμα της εικόνας που ακολουθεί, υπάρχει μια επισκόπηση της γλώσσας OWL2, που δείχνει τα κύρια δομικά στοιχεία της, και πως σχετίζονται μεταξύ τους.



Εικόνα 16 OWL2 Οντολογία

Η έλλειψη στο κέντρο αντιπροσωπεύει την αφηρημένη έννοια μιας οντολογίας, η οποία μπορεί να θεωρηθεί είτε ως μια αφηρημένη δομή είτε ως ένα γράφημα RDF αφού οποιαδήποτε οντολογία OWL2 μπορεί να θεωρηθεί ως ένα γράφημα RDF.

Στην κορυφή υπάρχουν οι εναλλακτικές συντάξεις της οντολογίας που μπορούν να χρησιμοποιηθούν για την σειριοποίηση και την ανταλλαγή της. Η κυριότερη από αυτές τις συντάξεις, είναι η *RDF/XML* και υποστηρίζεται από την πλειοψηφία των εργαλείων OWL2 και επομένως τους παρέχει την απαραίτητη διαλειτουργικότητα. Φυσικά υπάρχουν και εναλλακτικές συντάξεις όπως της *Turtle* και μια πιο ευανάγνωστη σύνταξη την *Manchester syntax*. Η κάθε σύνταξη έχει τον δικό της διαφορετικό σκοπό, όπως φαίνεται από τον πίνακα που ακολουθεί.

Πίνακας 1 Συντάξεις Οντολογιών

Όνομα σύνταξης	Σκοπός
RDF / XML	Διαλειτουργικότητα (μπορεί να γραφτεί και να διαβαστεί από όλο το συμβατό λογισμικό OWL 2)
OWL / XML	Ευκολότερη επεξεργασία με εργαλεία XML
Λειτουργική σύνταξη	Επίσημη δομή των οντολογιών
Σύνταξη Μάντσεστερ	Ευκολότερο να διαβαστεί

Ενώ στο κάτω μέρος του σχήματος βρίσκονται οι δύο κυριότερες σημασιολογικές προδιαγραφές που καθορίζουν την οντολογία. Υπάρχουν δυο εναλλακτικοί τρόποι στην εκχώρηση σημασιολογίας στις OWL2 οντολογίες. Η *Άμεση Σημασιολογία (Direct Semantics)* αποδίδει νόημα απευθείας σε δομές οντολογίας, με αποτέλεσμα μια πιο ευανάγνωστη σημασιολογία, και οι οντολογίες που τους έχουν αποδοθεί σημασιολογικά δεδομένα με αυτόν τον τρόπο, ονομάζονται *OWL2 DL*. Η *Σημασιολογία βασισμένη σε RDF (RDF Based Semantics)* εκχωρεί το νόημα άμεσα σε γραφήματα RDF και έμμεσα στις δομές οντολογίας. Είναι πλήρως συμβατή με την τεχνολογία RDF και μπορεί φυσικά να εφαρμοστεί σε οποιαδήποτε οντολογία OWL2. Τέτοιες οντολογίες ονομάζονται *OWL2 Full* και χρησιμοποιείται ανεπίσημα για να αναφερθούμε σε όλα τα γραφήματα RDF τα οποία θεωρούνται οντολογίες.

ΚΕΦΑΛΑΙΟ 4: CONCEPTNET – DBPEDIA - WORDNET

CONCEPTNET - DBPEDIA - WORDNET

Στην εργασία χρησιμοποιήθηκε πληροφορία και δεδομένα που εξορύχθηκαν από τρεις ιστοσελίδες οι οποίες υποστηρίζουν η κάθε μία ένα δίκτυο με σημασιολογική πληροφορία σε ημιδομημένη μορφή. Τα δίκτυα αυτά είναι: το ConceptNet⁵ που αναπτύχθηκε από το ερευνητικό κέντρο του MIT, η DBpedia⁶ το οποίο κατά κάποιο τρόπο αποτελεί προϊόν της Wikipedia⁷, και το Wordnet⁸ το οποίο είναι ένα λεξικό που αναπτύχθηκε από το πανεπιστήμιο Princeton.

Το κάθε δίκτυο προσφέρει ένα διαφορετικό είδος πληροφορίας και δεδομένων. Αρχικά γίνεται εξόρυξη πρωτόγονης πληροφορίας από το ConceptNet που σχετίζεται με ένα είδος γνώσης που στον κόσμο της πληροφορικής ονομάζουμε «Common Sense» (Κοινή Λογική). Στη συνέχεια από τη DBpedia προσφέρεται ένα είδος πληροφορίας που ερμηνεύει μια λέξη η γενικότερα μια έννοια. Τέλος από το Wordnet γίνεται εξόρυξη λεξιλογικής πληροφορίας. Αυτές οι τρεις συνιστώσες αποτελούν βασικά συστατικά στην ΣΠ που συνθέτει η εφαρμογή.

4.1 CONCEPTNET

Το **Concept Net** είναι ένα Δίκτυο ΣΠ, σχεδιασμένο έτσι ώστε να βοηθάει τα υπολογιστικά συστήματα και τις εφαρμογές να «κατανοούν» τη σημασιολογική πλευρά των λέξεων που χρησιμοποιεί ο άνθρωπος καθημερινά. Δηλαδή να έχουν Κοινή Λογική (Common Sense). Ο Όρος «Κοινή Λογική» εκφράζει την αίσθηση που έχουμε για κάτι χωρίς πολλή σκέψη αλλά με γνώμονα κυρίως την εμπειρία και την γνώση, και από αυτά τα δύο, η γνώση είναι κάτι το οποίο μπορούμε να προσφέρουμε στα υπολογιστικά συστήματα που αναπτύσσουμε. Φυσικά στους ανθρώπους η Κοινή Λογική είναι κάτι που αναπτύσσουν σταδιακά, αφού όχι μόνο αποκτούν γνώσεις και εμπειρίες με την πάροδο του χρόνου, αλλά η Κοινή Λογική τους διαμορφώνεται από τους νόμους, την κουλτούρα και τον πολιτισμό μέσα στον οποίο βιώνει ο εκάστοτε άνθρωπος και αποκτάει συγκεκριμένες συνήθειες και τρόπους ζωής. Στα υπολογιστικά συστήματα ωστόσο ο όρος «Κοινή Λογική» αφορά την ικανότητα που έχουν τα υπολογιστικά συστήματα να κατανοούν και να ερμηνεύουν καταστάσεις του περιβάλλοντος στο οποίο ορίζονται. Για παράδειγμα ο συλλογισμός που κάνει ο πράκτορας μας στο παράδειγμα της προηγούμενης ενότητας που προσπαθεί να καταλάβει τι έχει το 3^ο βάζο μέσα είναι «Κοινή Λογική».

4.1.1 Open Mind Common Sense

Το Σημασιολογικό Δίκτυο ConceptNet αποτελεί το δεύτερο στάδιο της ανάπτυξης ενός project που ξεκίνησε το 1999 στο MIT Media Lab το «Open Mind

⁵ <https://conceptnet.io/>

⁶ <https://wiki.dbpedia.org/>

⁷ <https://en.wikipedia.org/wiki/DBpedia>

⁸ <https://wordnet.princeton.edu/>

Common Sense» [6]. Το Open Mind Common Sense αναπτύχθηκε με σκοπό να βοηθήσει τα υπολογιστικά συστήματα και τις εφαρμογές να κατανοήσουν τον καθημερινό κόσμο των ανθρώπων και να αποκτήσουν μια συλλογιστική ικανότητα που θα είναι ακριβώς σαν την «Κοινή Λογική» που αναφέραμε. Το εγχείρημα αποτελούνταν από ένα δίκτυο το οποίο αποσκοπούσε στο να κάνει πιο εύκολη και πιο διασκεδαστική την θέληση των χρηστών να δουλέψουν μαζί και να προσφέρουν ο καθένας το δικό του μικρό κομμάτι Σημασιολογικής Γνώσης (ΣΓ) στο δίκτυο με τέτοιο τρόπο έτσι ώστε συλλογικά να δημιουργηθεί ένα δίκτυο «Κοινής Λογικής». Τέτοια κομμάτια ΣΓ μπορεί να είναι κάτι πολύ απλό όπως για παράδειγμα η γνώση ότι «μια κούπα περιέχει μέσα καφέ», ή «η ζάχαρη συνδυάζεται με τον καφέ», τα οποία μπορούν να εκφραστούν και με μια τριπλέτα RDF:



Εικόνα 17 Τριπλέτες RDF

Η λογική του ConceptNet είναι ότι αν μάθουμε στους Υπολογιστές πως να κατανοούν το ΣΠ τέτοιων μικρών κομματιών γνώσης, τότε αν τους προσφέρουμε μεγάλη ποσότητα αυτής της μορφής ΣΠ θα αποκτήσουν «Κοινή Λογική» και επομένως θα είναι σε θέση να κάνουν λογικούς συλλογισμούς για τον κόσμο. Η επίπτωση που θα έχει αυτό το εγχείρημα στον κόσμο του Διαδικτύου, σημαίνει ότι θα έχουμε δημιουργήσει την απαιτούμενη τεχνοτροπία για να πάμε το Διαδίκτυο στο επόμενο του επίπεδο, δηλαδή τον ΣΙ.

Το Open Mind Common Sense εξελίχθηκε μέσα από την συνεισφορά του κοινού στο ConceptNet. Στόχος του ConceptNet είναι να δημιουργήσει για κάθε λέξη μια τριπλέτα που θα περιέχει την ΣΠ, και φυσικά αυτό σημαίνει ότι θα συνδέεται σημασιολογικά με άλλες λέξεις. Επομένως το ConceptNet μπορούμε να πούμε ότι είναι ένας Γράφος που συνδέει λέξεις με labeled ακμές ώστε να είναι αλληλένδετα και να γίνουν πιο χρήσιμα (RDF graph).

Η γνώση που υπάρχει στο ConceptNet είναι crowd-sourced, αυτό σημαίνει ότι ο καθένας μπορεί να γράψει οτιδήποτε για οτιδήποτε και να το συνδυάσει με οτιδήποτε θεωρεί η «Κοινή Λογική». Ακριβώς με τον ίδιο τρόπο με τον οποίο ξεκίνησε η ανάπτυξη του ΠΙ. Ενώ τα δεδομένα που υπάρχουν στο ConceptNet είναι διαθέσιμα μέσω ενός JSON-LD API, όπου το LD σημαίνει Linked Data. Εκτός από το JSON-LD API όμως είναι διαθέσιμη για το κοινό και μια διεπαφή όπου μπορούμε να αναζητήσουμε και να περιηγηθούμε στα δεδομένα του ConceptNet.

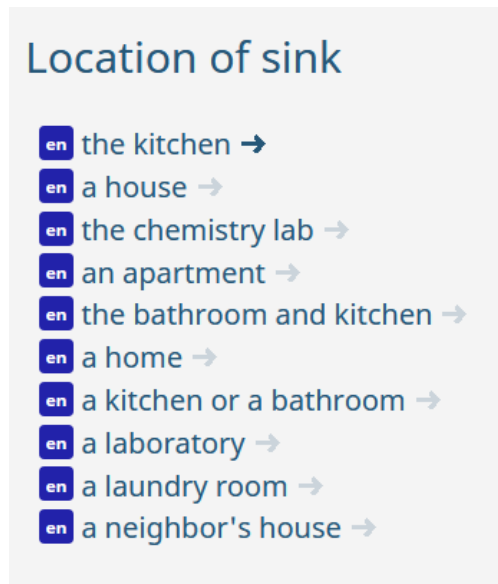
Παράδειγμα (ConceptNet): Αν ψάξουμε στο ConceptNet τον όρο «Coffee», έχουμε το παρακάτω αποτέλεσμα.

Εικόνα 18 Αποτελέσματα αναζήτησης από το ConceptNet για τον όρο «Coffee»

Αρχικά βλέπουμε τα sources της πληροφορίας που παίρνουμε κάτω από τον όρο μας, ενώ λαμβάνουμε ως επιστροφή ένα σύνολο άλλων όρων που συνδέονται σημασιολογικά με τον όρο της αναζήτησης. Στην συγκεκριμένη περίπτωση λαμβάνουμε τα πεδία:

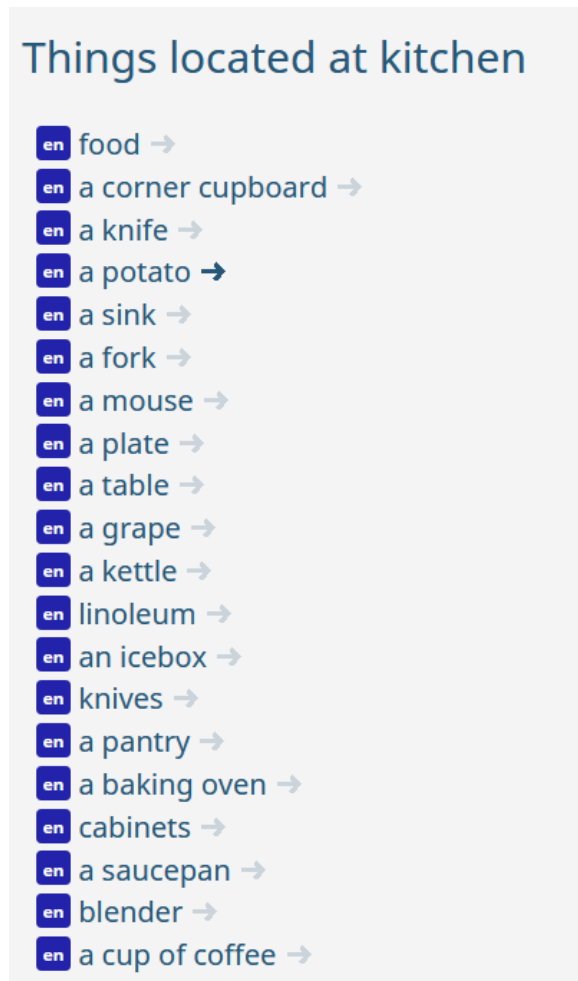
- Synonyms
- Types of coffee
- Coffee is a type of...
- Related terms
- Location of coffee
- Derived terms
- Etymologically derived terms
- Coffee has...
- Properties of coffee
- Coffee can be...
- Word forms
- Things used for coffee
- Things located at coffee

Ας επανέλθουμε στο παράδειγμα της προηγούμενης ενότητας όπου ένας πράκτορας προσπαθεί να ανιχνεύσει το περιεχόμενο του βάζου. Ο πράκτορας αρχικά έχει προγραμματιστεί να ανιχνεύει σε ποιο δωμάτιο βρίσκεται, και αναγνωρίζει έναν νιπτήρα. Σύμφωνα με το ConceptNet η αναζήτηση του όρου «Sink» στο πεδίο «Location of sink» μας επιστρέφει:



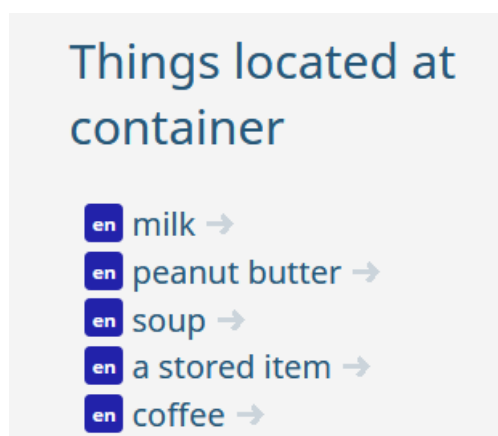
Εικόνα 19 Δεδομένα από το ConceptNet

Βλέπουμε ότι μια από τις επιλογές είναι η κουζίνα. Στη συνέχεια η αναζήτηση του όρου “Kitchen” στην εφαρμογή του ConceptNet.



Εικόνα 20 Δεδομένα από ConceptNet

Με την πληροφορία που έχει στην βάση γνώσης του ο πράκτορας σχετικά με την γεωμετρία των αντικειμένων καταφέρνει να αναγνωρίσει επιτυχώς την κούπα (Cup) και τα δύο δοχεία (Container) που βρίσκονται πάνω στην επιφάνεια. Στη συνέχεια με την εξόρυξη του πεδίου «Things located at container»:



Εικόνα 21 Δεδομένα από ConceptNet

Ο πράκτορας ανιχνεύει ότι το περιεχόμενο του ενός δοχείου πιθανότατα θα είναι καφές, βάση του παρακάτω συλλογισμού.

→ Sink

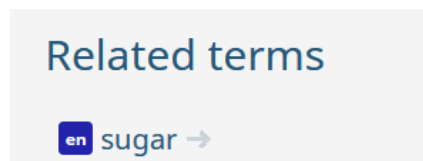
→ Location of Sink → Kitchen

→ Things located at Kitchen → a cup of coffee

→ Container (ανιχνεύτηκε από το γεωμετρικό του σχήμα)

→ Things located at container → Coffee

Και με την αναζήτηση του όρου “Coffee” παίρνουμε:



Εικόνα 22 Δεδομένα από ConceptNet

Ενώ βάση της ίδιας λογικής στη συνέχεια έχουμε την αναζήτηση του όρου «Sugar»



Εικόνα 23 Δεδομένα από ConceptNet

Όπου είναι ξεκάθαρο ότι συνήθως η ζάχαρη βρίσκεται μέσα σε ένα “Container”. Επομένως βάση του συλλογισμού.

→ Coffee → Sugar

→ Sugar → Container

Ο πράκτορας μας είναι σε θέση να βγάλει το συμπέρασμα το οποίο προέκυψε μέσα από λογικούς συλλογισμούς ότι πιθανότατα τα δοχεία περιέχουν Καφέ και Ζάχαρη, ενώ η διαδικασία αυτή ορίζει ακριβώς αυτό που ονομάσαμε «Κοινή Λογική».

4.2 DBPEDIA

Η *DBpedia*, είναι έργο το οποίο έχει δημιουργηθεί από δεδομένα που παρέχει η κοινότητα (*crowd-sourced community*) με μεγάλο όγκο πολυγλωσσικά δεδομένα τα οποία αναπαριστούν γνώση η οποία έχει ληφθεί από εξόρυξη γνώσης από μέσα της *Wikipedia*. Επομένως η *DBpedia* χρησιμοποιείται για να γίνει εξόρυξη δεδομένων από διάφορα *Wikimedia Projects*. Πιο συγκεκριμένα ορίζεται ένας Γράφος Ανοιχτής Γνώσης (*Open Knowledge Graph - OKG*) ο οποίος περιέχει δεδομένα σε μορφή Ανοιχτών Δεδομένων (*Linked Data*) και είναι διαθέσιμος για όλους.

Ο λόγος για τον οποίο υπάρχει η *DBpedia* είναι παρόμοιος με αυτόν του *ConceptNet*, επειδή υπάρχει η απαίτηση πρόσβασης της πληροφορίας της *Wikipedia*⁹ από εφαρμογές και υπολογιστικά συστήματα που αναζητούν πληροφορία με σημασιολογικό περιεχόμενο. Η πρόσβαση των συστημάτων στην τεχνολογία αυτή γίνεται με την βοήθεια *SPARQL* ερωτημάτων όπως ακριβώς αναφέρθηκε στην πρώτη ενότητα.

Η ποσότητα πληροφορίας που είναι διαθέσιμη στο *DBpedia* στην Αγγλική γλώσσα ανέρχεται στις 4.58 εκατομμύρια εγγραφές. Από αυτές ένα σημαντικό ποσοστό (4.22 εκατομμύρια) αποτελούν κομμάτι μιας Οντολογίας. Πιο συγκεκριμένα υπάρχουν εγγραφές για:

- 1.445.000 Άνθρωποι
- 735.000 Μέρη
- 411.000 Δημιουργικά έργα (όπως μουσικά έργα, ταινίες, βιντεοπαιχνίδια)
- 241.000 Οργανισμοί
- 251.000 Ζωικά είδη
- 6.000 Ασθένειες

Η *DBpedia* επίσης συνδέεται με *Linked Datasets* που δημιουργούν έναν αριθμό περίπου 50 εκατομμυρίων *RDF triples* σε όλες τις γλώσσες. Όπως δείχνουν τα νούμερα το εγχείρημα έχει κάνει διαθέσιμο έναν τρομερά μεγάλο όγκο πληροφοριών στους υπολογιστές και σε εφαρμογές που χρειάζονται γνώση ενώ οι διασυνδέσεις όλων των παραπάνω σε *Linked Data* προσφέρουν ένα ανεκτίμητης αξίας σημασιολογικό περιεχόμενο. Η *DBpedia* κατηγοριοποιεί όλα τα δεδομένα σε σαφώς ορισμένες κλάσεις οντολογιών¹⁰. Αν αναζητήσουμε τον όρο «Coffee» στη *DBpedia* παίρνουμε ένα σύνολο από τριπλέτες:

is *dbo:ingredient of*

- *dbp:Tiramisu*
- *dbp:Hopje*
- *dbp:Opera_cake*
- *dbp:Café_con_leche*
- *dbp:Brown_bread*

⁹ <https://en.wikipedia.org/wiki/DBpedia>

¹⁰ <http://mappings.dbpedia.org/server/ontology/classes/>

[rdf:type](#)

- owl:Thing
- dul:FunctionalSubstance
- wikidata:Q2095
- dbo:Beverage
- dbo:Food

Εικόνα 24 Τριπλέτα από DBpedia

Είναι προφανές ότι το ConceptNet και η DBpedia είναι δυο διαφορετικά *project* που έχουν κάποιους κοινούς προσανατολισμούς, οι οποίοι είναι το να κάνουν περισσότερο προσβάσιμη στις μηχανές την ΣΠ που υπάρχει στον ΠΙ.

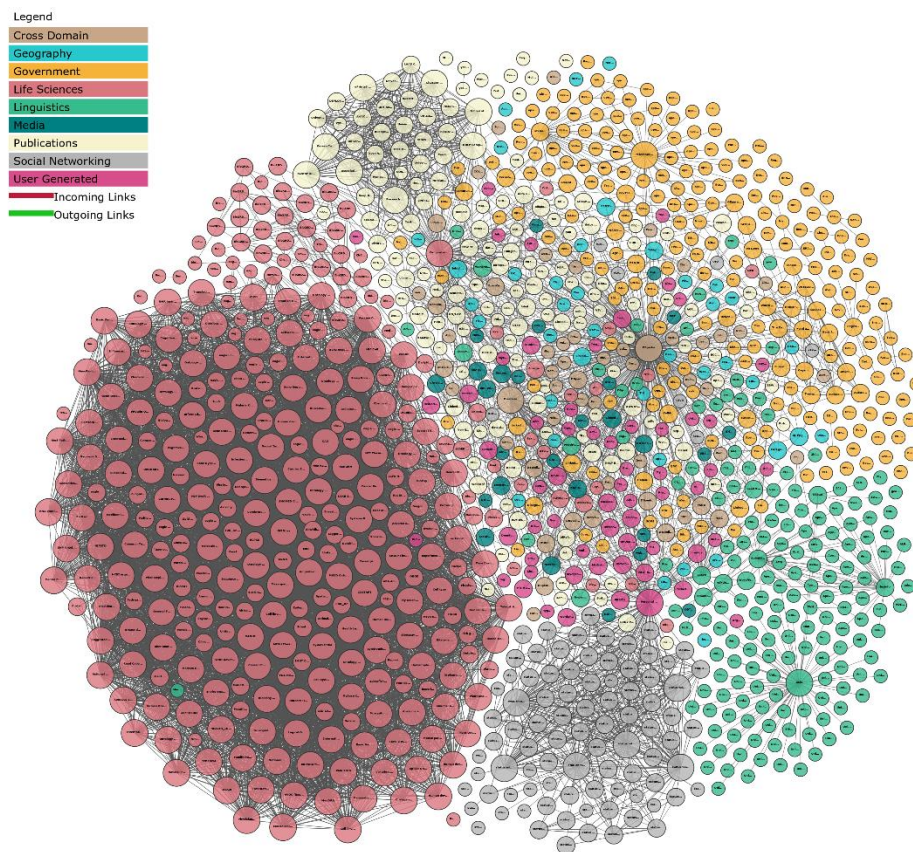
Το έργο DBpedia περιλαμβάνει τρεις κύριους τομείς, έναν βασικό τομέα για εξαγωγή και μετασχηματισμό δομημένων δεδομένων, που εξάγει οντότητες, τύπους σχέσεων μεταξύ των οντοτήτων και ένα σύνολο σχέσεων μεταξύ των οντοτήτων που έχει προέλθει από εξόρυξη γνώσεων από έγγραφα της Wikimedia. Περιλαμβάνει επίσης τον τομέα ανάπτυξης ΑΔΔ, που καθιστά τις σχέσεις οντοτήτων διαθέσιμες στον Ιστό με τον παραδοσιακό τρόπο δημοσιοποίησης Διασυνδεδεμένων Δεδομένων όπως αναλύθηκε στην ενότητα 1.5. Αυτό το σύνολο των ΑΔΔ χρησιμοποιούνται για την δημιουργία γραφημάτων σχέσεων οντοτήτων (γράφοι RDF).

Το τρίτο βασικό κομμάτι είναι μια Υπηρεσία Διαδικτυακών Ερωτημάτων (SPARQL endpoint), που παρέχει πρόσβαση σε δομημένα δεδομένα, όπου τα αποτελέσματα των ερωτημάτων παραδίδονται ως σχεσιακοί πίνακες ή ως γραφήματα οντοτήτων ή ως μια ποικιλία τύπων και μορφών εγγράφων που αναλύονται στην ενότητα 3.2.2 DBpedia endpoint.

4.2.1 DBpedia και Ανοιχτά Διασυνδεδεμένα Δεδομένα

Τα δεδομένα που είναι προσβάσιμα από τη DBpedia έχουν δημοσιευτεί αυστηρά με τις αρχές δημοσίευσης των Ανοιχτών Διασυνδεδεμένων Δεδομένων που περιγράφηκαν στην Ενότητα 1.3. Δηλαδή οι οντότητες αναγνωρίζονται χρησιμοποιώντας υπερσυνδέσμους (HTTP URIs), και οι οντότητες περιγράφονται με την χρήση προτάσεων και δηλώσεων που βασίζονται στη γλώσσα RDF όπου το κάθε μέρος της τριπλέτας αναγνωρίζεται επίσης από ένα HTTP URI είτε από ένα Literal. Οι περιγραφές των οντοτήτων αυτές, δημοσιεύονται σε δίκτυα HTTP (για παράδειγμα στον Παγκόσμιο Ιστό) χρησιμοποιώντας έγγραφα RDF, των οποίων το περιεχόμενο έχει σειριοποιηθεί χρησιμοποιώντας κάποια από τις γνωστές μορφές σειριοποίησης που περιγράφηκαν στην ενότητα 2.4 (HTML, JSON-LD, RDF, Turtle, RDF/XML, και άλλα).

Όπως είναι αναμενόμενο, το σύνολο των δεδομένων της DBpedia συνδέεται με διάφορες άλλες πηγές δεδομένων. Το παρακάτω διάγραμμα παρέχει την επισκόπηση ορισμένων από αυτές τις πηγές δεδομένων.



Εικόνα 25 Νέφος Ανοιχτών Διασυνδεδεμένων Δεδομένων

Ενώ ο πίνακας που ακολουθεί εμπεριέχει τα *Σύνολα Δεδομένων (Data Sets)* που χρησιμοποιήθηκαν, τον αριθμό *υπερσυνδέσεων* του κάθε *Συνόλου Δεδομένων* καθώς και μια περιγραφή του κάθε *Συνόλου Δεδομένων*.

Πίνακας 2 ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ ΑΠΟ ΤΟ DBPEDIA

Σύνολο Δεδομένων	Περιγραφή	Αριθμός Συνδέσμων
Amsterdam Museum	Πληροφορίες για αντικείμενα πολιτιστικής κληρονομιάς που σχετίζονται με την πόλη του Άμστερνταμ	630
BBC Wildlife Finder	Πληροφορίες σχετικά με τους άγριους βιότοπους, τους οικοτόπους, και τις οικολογικές ζώνες.	450
Book Mashup	Παρέχει πληροφορίες σχετικά με βιβλία	9,000
Bricklink	Ανεπίσημη αγορά Lego	10,000
CORDIS	Πληροφορίες για όλα τα προγράμματα και έργα της ΕΕ.	300
Dailymed	Παρέχει πληροφορίες σχετικά με φάρμακα	900
DBLP Bibliography	Παρέχει πληροφορίες σχετικά με επιστημονικές δημοσιεύσεις	200
DBTune	Παρέχει ελεύθερα διαθέσιμα δεδομένα σχετικά με τη μουσική	840
Diseasome	Παρέχει πληροφορίες σχετικά με ασθένειες και γονίδια	2,300
Drugbank	Παρέχει πληροφορίες σχετικά με φάρμακα και γονίδια	4,800
EUNIS	Πληροφορίες για τα είδη, τύπους οικοτόπων και τοποθεσίες	11,000
Eurostat (Linked Statistics)	Καλύπτει έναν αριθμό τομέων από την οικονομία σε σχέση με τα δημογραφικά στοιχεία έως το εμπόριο και τα δεδομένα μεταφοράς	250
Eurostat (WBSG)	Παρέχει πληροφορίες σχετικά με τις ευρωπαϊκές χώρες και περιοχές	140
CIA World Factbook	Παρέχει πληροφορίες για χώρες	550
flickr wrappr	Προσπαθεί να δημιουργήσει μια συλλογή φωτογραφιών για κάθε οντότητα που υπάρχει στο DBpedia	4,000,000

Freebase	Μια ανοιχτή βάση δεδομένων για εκατομμύρια πράγματα από διάφορους τομείς	3,900,000
GADM	Δεδομένα τοποθεσίας των διοικητικών περιοχών του κόσμου	39,000
GeoNames	Παρέχει πληροφορίες σχετικά με γεωγραφικά χαρακτηριστικά	425,000
GeoSpecies	Πληροφορίες σχετικά με βιολογικές αρχές, οικογένειες ειδών, καθώς και δεδομένα εμφάνισης ειδών και σχετικά δεδομένα	16,000
Global Health Observatory	Παρέχει πρόσβαση σε στατιστικά δεδομένα σχετικά με προβλήματα υγείας	200
Project Gutenberg	Παρέχει πληροφορίες σχετικά με συγγραφείς και προσφέρει ανοιχτή πρόσβαση στο έργο τους	2,500
Italian Public Schools	Παρέχει πληροφορίες για δημόσια σχολεία στην Ιταλία	5,800
LinkedGeoData	Βάση γνώσεων	104,000
LinkedMDB	Παρέχει πληροφορίες για ταινίες	14,000
MusicBrainz	Παρέχει πληροφορίες για καλλιτέχνες και μουσική	23,000
New York Times	Σύνδεσμοι μεταξύ επικεφαλίδων θέματος NYT και οντοτήτων της DBpedia	9,700
OpenCyc	Δεδομένα σχετικά με τις οντότητες του Cyc	27,000
OpenEI (Open Energy Info)	Παρέχει πληροφορίες σχετικά με την ενέργεια	680
Revyu	Παγκόσμιες αξιολογήσεις	6
Sider	Παρέχει πληροφορίες σχετικά με τις παρενέργειες των φαρμάκων	2000
TCMGeneDIT	Πληροφορίες για την παραδοσιακή κινεζική ιατρική, τα γονίδια και τις ασθένειες	900
UMBEL	Μια ελαφριά, θεματική δομή αναφοράς εννοιών που προέρχεται από το Cyc	900,000
US Census	Παρέχει δεδομένα απογραφής από τις ΗΠΑ	12,600

WikiCompany	Παρέχει πληροφορίες για εταιρείες	8,300
Wikidata	δομημένα δεδομένα που σχετίζονται με στοιχεία της Wikipedia	5,200,000
WordNet	RDF / OWL αναπαράσταση της οντολογίας του WordNet	470,000
YAGO	Βάση γνώσεων μεταξύ τομέων του YAGO	2,900,000 instance links, 41,000,000 type statements

4.2.2 DBpedia endpoint

Τα συνδεδεμένα δεδομένα είναι μια μέθοδος δημοσίευσης δεδομένων στον Ιστό και διασύνδεσης δεδομένων μεταξύ διαφορετικών πηγών δεδομένων. Μπορεί κανείς να έχει πρόσβαση σε συνδεδεμένα δεδομένα χρησιμοποιώντας κάποια λογισμικά περιήγησης ΣΙ, με την διαφορά από τον παραδοσιακή μέθοδο πλοήγησης μεταξύ εγγράφων του Ιστού, ότι στην περίπτωση των ΑΔΔ, αντί να ακολουθούνται οι υπερσύνδεσμοι που υπάρχουν μεταξύ των εγγράφων, η πλοήγηση γίνεται με την χρήση των συνδέσμων RDF που υπάρχουν μεταξύ των διαφορετικών πηγών δεδομένων. Οι σύνδεσμοι αυτοί RDF, μπορούν επίσης να ακολουθούνται από ρομπότ και από μηχανές αναζήτησης του ΣΙ.

Το DBpedia προσφέρει επίσης το δικό του SPARQL endpoint, το «*Virtuoso SPARQL Query Editor*», Βέβαια, μέχρι στιγμής, το DBpedia endpoint δεν διαθέτει στο κοινό όλα τα δεδομένα τα οποία έχουν συμπεριληφθεί στο γενικότερο έργο του DBpedia. Τα Σύνολα Δεδομένων που έχουν συμπεριληφθεί στο DBpedia endpoint είναι διαθέσιμα στους παρακάτω υπερσυνδέσμους:

- <http://downloads.dbpedia.org/2016-10/links/>
- <http://downloads.dbpedia.org/2016-10/core/>
- http://downloads.dbpedia.org/2016-10/core-i18n/en/instance_types_lhd_dbo_en.ttl.bz2
- http://downloads.dbpedia.org/2016-10/core-i18n/en/instance_types_lhd_ext_en.ttl.bz2

Το SPARQL endpoint του DBpedia μας δίνει τη δυνατότητα να ζητήσουμε τα δεδομένα σε όποιον τύπο μας βολεύει, ανάλογα με τις ανάγκες του λογισμικού που αναπτύσσουμε. Πιο συγκεκριμένα μας δίνει την δυνατότητα να εξάγουμε δεδομένα σε μορφή HTML, Spreadsheet, XML, JSON, Javascript, Turtle, RDF/XML, N-Triples, CSV, TSV, CXML.

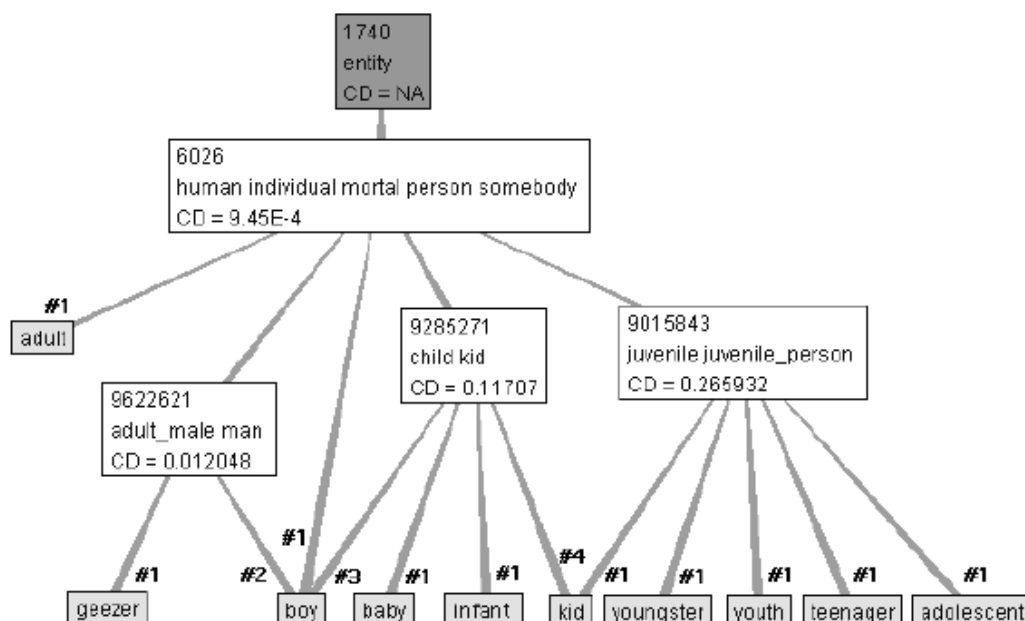
Αν και το DBpedia είναι ένα από τα σημαντικότερα SPARQL endpoints που έχουν αναπτυχθεί, αξίζει να αναφέρουμε πως δεν είναι το μοναδικό. Η παγκόσμια κοινότητα της πληροφορικής, έχει αναπτύξει με την πάροδο του χρόνου μια μεγάλη γκάμα SPARQL endpoints. Τα endpoints που έχουν αναπτυχθεί υπάρχουν στην ιστοσελίδα <https://www.w3.org/wiki/SparqlEndpoints>.

4.3 WORDNET

Το WordNet είναι μια μεγάλη λεξική βάση δεδομένων αγγλικής γλώσσας. Τα ουσιαστικά τα ρήματα, τα επίθετα και τα επιρρήματα ομαδοποιούνται σε σύνολα γνωστικών συνωνύμων (synsets) και το καθένα εκφράζει μια ξεχωριστή έννοια. Τα σύνολα αλληλοσυνδέονται μέσω εννοιολογικών, σημασιολογικών και λεξικών σχέσεων, έτσι προκύπτει ένα δίκτυο λέξεων και εννοιών στο οποίο μπορεί κανείς να περιηγηθεί¹¹. Το WordNet είναι ένα χρήσιμο εργαλείο για λογισμικά που σχετίζονται με την λεξιλογική πλευρά του κειμένου, και με την φυσική επεξεργασία της γλώσσας.

Το WordNet ομαδοποιεί τις λέξεις με βάση τι έννοιες που αυτές αντιπροσωπεύουν. Οι λέξεις διασυνδέονται όχι μόνο επειδή μοιάζουν λεξικογραφικά, αλλά και επειδή αντιπροσωπεύουν κοινές έννοιες. Άρα λέξεις που βρίσκονται σε κοντινή απόσταση μεταξύ τους στο δίκτυο του WordNet είναι σημασιολογικά διαφορούμενες. Η κύρια συσχέτιση μεταξύ των λέξεων στο WordNet είναι το συνώνυμο (synonym), δηλαδή λέξεις που υποδηλώνουν την ίδια έννοια, όπως «Car» και «Vehicle» ή «Warm» και «Hot». Οι λέξεις που είναι συνώνυμες μεταξύ τους, ομαδοποιούνται σε μη ταξινομημένα σύνολα και σχηματίζονται έτσι συνολικά 117.000 σύνολα, τα οποία συνδέονται μεταξύ τους μέσω εννοιολογικών σχέσεων. Για κάθε σύνολο υπάρχει ένας ορισμός, ενώ λέξεις που έχουν πολλές διαφορετικές έννοιες, μπορούν να βρίσκονται σε παραπάνω από ένα σύνολα.

Υπάρχει πλήρης ανάλυση της βιβλιοθήκης WordNet σε Python στον σύνδεσμο: <https://www.nltk.org/howto/wordnet.html>

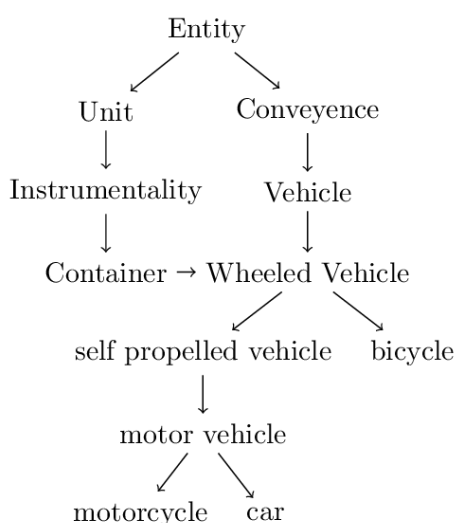


Εικόνα 26 Λεξιλογικό Παράδειγμα από WordNet

¹¹ <http://wordnetweb.princeton.edu/perl/webwn>

4.3.1 Σχέση Υπερωνυμίας (*hyperonymy*) και Μερωνυμίας (*meronymy*)

Η συνηθέστερη σχέση μεταξύ των συνόλων είναι η σχέση υπερωνυμίας (ISA). Συνδέει γενικά σύνολα όπως {furniture, piece_of_furniture} με πιο συγκεκριμένα όπως {bed} και {bunkbed}. Με τον τρόπο αυτό δηλώνεται ότι η κατηγορία των «επίπλων» περιλαμβάνει το «κρεβάτι», το οποίο με την σειρά του, περιλαμβάνει την «κουκέτα». Ενώ αντίστροφα, έννοιες όπως «κρεβάτι» και «κουκέτα» είναι ένα είδος «επίπλου». Δημιουργείται έτσι ένας γράφος, με όλες τις έννοιες τελικά να ανεβαίνουν στον κόμβο ρίζα που είναι το {entity} που σημαίνει «οντότητα». Προφανώς η σχέση που αναλύθηκε είναι μεταβατική, αυτό σημαίνει πως αν για παράδειγμα μια «πολυθρόνα» είναι ένα είδος «καρέκλας» και μια «καρέκλα» είναι ένα είδος «επίπλου», τότε μια «πολυθρόνα» είναι ένα είδος «επίπλου».



Εικόνα 27 Entity του WordNet

Η σχέση *Meronymy*, η αλλιώς σχέση μερικής όψης, αναφέρεται σε κληρονομικότητα μεταξύ των συνόλων. Για παράδειγμα αν μια «καρέκλα» έχει «πόδια» τότε μια «πολυθρόνα» που είναι «καρέκλα», έχει «πόδια». Φυσικά τέτοια είδη κληρονομικότητας δεν είναι δυνατόν να κληρονομούνται «προς τα πάνω», καθώς μπορεί να αναφέρονται σε χαρακτηριστικά μόνο συγκεκριμένων ειδών και όχι της τάξης στο σύνολο της. Για παράδειγμα οι «καρέκλες» έχουν «πόδια», αλλά δεν έχουν όλα τα είδη των «επίπλων» πόδια.

ΚΕΦΑΛΑΙΟ 5: PROBLEM STATEMENT – USE CASE

PROBLEM STATEMENT – USE CASE

5.1 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ ΚΑΙ ΠΑΡΑΔΟΣΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΑΠΟ ΤΟ WEB

Ένα Σύστημα Ανάκτησης Πληροφορίας (ΣΑΠ) έχει δύο βασικούς στόχους. Ο πρώτος έχει να κάνει με την ποιότητα και την επάρκεια των αποτελεσμάτων που επιστρέφει και ο δεύτερος σχετίζεται με την ταχύτητα ανάκτησης της ζητούμενης πληροφορίας, δηλαδή με την απόδοση του συστήματος. Και στα δύο κομμάτια, η παγκόσμια κοινότητα της πληροφορικής έχει κάνει γιγαντιαία βήματα βελτίωσης ΣΑΠ. Ένα από τα βασικότερα συστατικά στοιχεία ενός ΣΑΠ είναι ο *Αντεστραμμένος Κατάλογος (AK) (Inverted Index)*. Ο ΑΚ είναι μια μορφή καταλόγου για την οργάνωση των όρων μιας συλλογής εγγράφων. Η χρήση του καταλόγου στοχεύει στην αποδοτική επεξεργασία των ερωτημάτων που κάνουμε στο ΣΑΠ, και είναι απαραίτητος ώστε να αποφευχθεί η εξέταση όλων των εγγράφων της συλλογής, που στην περίπτωση μας είναι όλες οι ιστοσελίδες που είναι καταχωρημένες στον ΠΙ. Ο αντεστραμμένος κατάλογος αποτελείται από δύο βασικά τμήματα, το *λεξικό (lexicon)*, και τις *λίστες εμφανίσεων (posting lists)*.

Έστω ότι στην διάθεση μας έχουμε την παρακάτω συλλογή εγγράφων:

d_1 :	Ο κομήτης του Χάλλεϋ μας επισκέπτεται περίπου κάθε εβδομήντα έξι χρόνια.
d_2 :	Ο κομήτης του Χάλλεϋ ανακαλύφθηκε από τον αστρονόμο Έντμοντ Χάλλεϋ.
d_3 :	Ένας κομήτης διαγράφει ελλειπτική τροχιά.
d_4 :	Ο πλανήτης Αρης έχει δύο φυσικούς δορυφόρους, το Δείμο και το Φόβο.
d_5 :	Ο πλανήτης Δίας έχει εξήντα τρεις γνωστούς φυσικούς δορυφόρους.
d_6 :	Ο Ήλιος είναι ένας αστέρας.
d_7 :	Ο Αρης είναι ένας πλανήτης του ηλιακού μας συστήματος.

Εικόνα 29 Παράδειγμα μικρής συλλογής εγγράφων

Ο ΑΚ για την παραπάνω συλλογή εγγράφων, φαίνεται στο σχήμα που ακολουθεί και περιέχει στο μέρος του λεξικού όλες τις λέξεις που εμφανίζονται στο σύνολο των εγγράφων που έχουμε, και στο τμήμα των λιστών εμφανίσεων, περιέχει το πόσες φορές εμφανίζεται ο κάθε όρος συνολικά στην συλλογή, και το σε ποια έγγραφα εμφανίζεται ο όρος. Πάνω στον ΑΚ στηρίζεται το μοντέλο tf-idf που αναλύεται παρακάτω.

λεξιικό	λίστες εμφανίσεων
ο	$[5: d_1, d_2, d_4, d_5, d_6, d_7]$
κομήτης	$[3: d_1, d_2, d_3]$
του	$[3: d_1, d_2, d_7]$
Χάλεϋ	$[2: d_1, d_2]$
μας	$[2: d_1, d_7]$
επισκέπτεται	$[1: d_1]$
περίπου	$[1: d_1]$
κάθε	$[1: d_1]$
εβδομήντα	$[1: d_1]$
έξι	$[1: d_1]$
χρόνια	$[1: d_1]$
ανακαλύφθηκε	$[1: d_2]$
από	$[1: d_2]$
τον	$[1: d_2]$
αστρονόμο	$[1: d_2]$
Έντμοντ	$[1: d_2]$
ένας	$[3: d_2, d_6, d_7]$
διαγράφει	$[1: d_3]$
ελλειπτική	$[1: d_3]$
τροχιά	$[1: d_3]$
πλανήτη	$[3: d_4, d_5, d_7]$
Άρης	$[2: d_4, d_7]$
έχει	$[3: d_4, d_5]$
δύο	$[1: d_4]$
φυσικούς	$[2: d_4, d_5]$
δορυφόρους	$[2: d_4, d_5]$
το	$[1: d_4]$
Δείμο	$[1: d_4]$
και	$[1: d_4]$
Φόβο	$[1: d_4]$
Δίας	$[1: d_5]$
εξήντα	$[1: d_5]$
τρεις	$[1: d_5]$
γνωστούς	$[1: d_5]$
Ήλιος	$[1: d_6]$
είναι	$[2: d_6, d_7]$
αστέρας	$[1: d_6]$
ηλιακού	$[1: d_6]$
συστήματος	$[1: d_7]$

Εικόνα 30 Αντεστραμμένος Κατάλογος για την συλλογή εγγράφων

5.1.1 Ομοιότητα Συνημιτόνων (Cosine Similarity)

Στόχος του μοντέλου, είναι ότι αφού αναπαρασταθούν τα δύο Comment Boxes, τα οποία εξάγονται από το DBpedia όπως αναφέρθηκε στην ενότητα 3, με δυο μη μηδενικά διανύσματα, ως Ομοιότητα Συνημιτόνου, ορίζεται το εσωτερικό γινόμενο των δύο διανυσμάτων και προσδιορίζεται το πόσο πολύ μοιάζουν τα δύο Comment Boxes με μέτρο σύγκρισης μόνο το πόσες κοινές λέξεις έχουν μεταξύ τους. Πιο συγκεκριμένα, αν έχουμε τα Comment Boxes d1 έως d7:

- d_1 : Ο κομήτης του Χάλλεϋ μας επισκέπτεται περίπου κάθε εβδομήντα έξι χρόνια.
 d_2 : Ο κομήτης του Χάλλεϋ ανακαλύφθηκε από τον αστρονόμο Έντμοντ Χάλλεϋ.
 d_3 : Ένας κομήτης διαγράφει ελλειπτική τροχιά.
 d_4 : Ο πλανήτης Άρης έχει δύο φυσικούς δορυφόρους, το Δείμο και το Φόβο.
 d_5 : Ο πλανήτης Δίας έχει εξήντα τρεις γνωστούς φυσικούς δορυφόρους.
 d_6 : Ο Ήλιος είναι ένας αστέρας.
 d_7 : Ο Άρης είναι ένας πλανήτης του ηλιακού μας συστήματος.

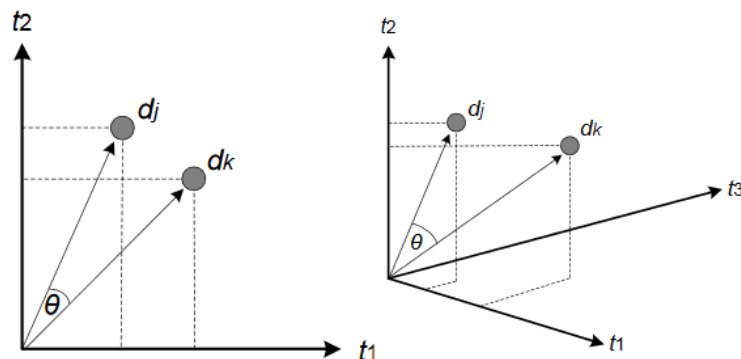
Τότε μπορούμε να σχηματίσουμε έναν πίνακα όρων-εγγράφων με δυαδικά βάρη ανάλογα με το αν ένας όρος εμπεριέχεται σε ένα d-Comment Box η όχι.

όρος	d_1	d_2	d_3	d_4	d_5	d_6	d_7
κομήτης	1	1	1	0	0	0	0
πλανήτης	0	0	0	1	1	0	1
Χάλλεϋ	1	1	0	0	0	0	0
Άρης	0	0	0	1	0	0	1
Δίας	0	0	0	0	1	0	0
τροχιά	0	0	1	0	0	0	0

Εικόνα 31 Διανυσματική αναπαράσταση των εγγράφων

Πλέον όπως φαίνεται και από το σχήμα το κάθε d-Comment Box έχει αναπαρασταθεί από ένα διάνυσμα. Αν πάρουμε το εσωτερικό γινόμενο των διανυσμάτων αυτών, θα προσδιορίσουμε ποια Comment Boxes μοιάζουν περισσότερο βάση των όρων που περιέχουν.

Το εσωτερικό γινόμενο δύο διανυσμάτων, υπολογίζει την γωνία που σχηματίζουν τα διανύσματα στον χώρο. Αν αναπαραστήσουμε τα Comment Boxes στις δυο και στις τρεις διαστάσεις που είναι δυνατή η γραφική αναπαράσταση των διανυσμάτων, τότε δύο Comment Boxes θα είχαν εσωτερικό γινόμενο 1, αν τα διανύσματα τους βρίσκονταν στην ίδια ευθεία, και αυτό θα σήμαινε ότι είναι όμοια.



Εικόνα 32 Διανυσματική αναπαράσταση των εγγράφων στον δισδιάστατο χώρο

5.1.2 Μοντέλο *tf-idf*

Το μοντέλο *tf-idf* είναι ένα εργαλείο το οποίο προσδιορίζει έναν εναλλακτικό τρόπο προσδιορισμού των βαρών των διανυσμάτων που χρησιμοποιούνται για τον υπολογισμό της ομοιότητας συνημιτόνου. Με το μοντέλο αυτό δεν χρησιμοποιούνται δυαδικά βάρη, όπως στην υλοποίηση που αναλύθηκε στην προηγούμενη παράγραφο, αλλά χρησιμοποιείται κάτι πιο πλούσιο σε πληροφορία για τον προσδιορισμό των βαρών των όρων στο κάθε Comment Box.

Έστω t ένας όρος και d ένα Comment Box που έχουμε ανακτήσει από τη DBpedia. Η συχνότητα (frequency) εμφάνισης του όρου t στο d συμβολίζεται με $f_{t,d}$ (term frequency) και προσδιορίζει τον αριθμό των εμφανίσεων του όρου στο συγκεκριμένο έγγραφο. Η συχνότητα εμφάνισης του όρου στο έγγραφο μπορεί να χρησιμοποιηθεί για να δηλώσει τη σημαντικότητα (βάρος) του όρου για το έγγραφο. Επομένως, μία πρώτη προσέγγιση για τον προσδιορισμό του βάρους w είναι να χρησιμοποιηθεί ο τύπος:

$$w_{t,d} = f_{t,d}$$

Με την χρήση του τύπου αυτού για τον υπολογισμό του βάρους w_t στο έγγραφο d , οι όροι που εμφανίζονται σε μεγάλα έγγραφα ενδεχομένως να έχουν και μεγαλύτερο βάρος, διότι αυξάνεται η πιθανότητα ύπαρξής τους στο έγγραφο. Για τον λόγο αυτό, και για να μη γίνεται διάκριση μεταξύ μικρών και μεγάλων εγγράφων (Comment Box) μπορεί να χρησιμοποιηθεί η κανονικοποιημένη συχνότητα εμφάνισης (normalized frequency) που συμβολίζεται με $nf_{t,d}$ και ορίζεται ως εξής:

$$nf_{t,d} = \frac{f_{t,d}}{\max\{f_{x,d}\}}$$

Το πλήθος των εμφανίσεων ενός όρου σε ένα έγγραφο δηλώνει τη σημαντικότητα του όρου για το έγγραφο αυτό. Ωστόσο, θα πρέπει να παρατηρήσουμε ότι όροι που εμφανίζονται σε πολλά έγγραφα έχουν μικρή διακριτική ικανότητα. Αυτό σημαίνει, ότι αν και οι όροι αυτοί μπορεί να εμφανίζονται σε πολλά από αυτά μειώνει τη σημαντικότητά τους. Για παράδειγμα, σε μία συλλογή Comment Boxes που έχουν προκύψει από την αναζήτηση του όρου «tea» στο πρόγραμμα, είναι λογικό τα Comment Boxes να περιέχουν πολλές φορές τον όρο «tea». Όμως, είναι επίσης λογικό ο όρος «tea» να εμφανίζεται πολλές φορές στο ίδιο έγγραφο. Επομένως είναι αναγκαίος ένας νέος παράγοντας στην χρήση για τον υπολογισμό των βαρών. Ο παράγοντας αυτός καλείται *αντίστροφη συχνότητα εγγράφων* (*inverse document frequency*) και συμβολίζεται με idf_t . Αν συμβολίσουμε με N το πλήθος των Comment Boxes που έχουμε διαθέσιμα και με n_t το πλήθος των Comment Boxes που περιέχουν τον όρο t , τότε ο παράγοντας αυτός υπολογίζεται για κάθε όρο ξεχωριστά ως εξής:

$$idf_t = \ln\left(\frac{N}{n_t}\right)$$

Χρησιμοποιώντας την κανονικοποιημένη συχνότητα εμφάνισης και την αντίστροφη συχνότητα εγγράφων, προκύπτει ένας νέος τρόπος υπολογισμού των βαρών $w_{t,d}$ που είναι:

$$w_{t,d} = n f_{t,d} \cdot idf_t = \frac{f_{t,d}}{\max_x(f_{x,d})} \cdot \ln\left(\frac{N}{n_t}\right)$$

Όμως ο παράγοντας idf_t δεν είναι κανονικοποιημένος. Η κανονικοποίηση του παράγοντα αυτού μπορεί να πραγματοποιηθεί διαιρώντας με το λογάριθμο του πλήθους των εγγράφων. Με τον τρόπο αυτό προκύπτει η *κανονικοποιημένη αντίστροφη συχνότητα εγγράφων* (*normalized inverse document frequency*) η οποία υπολογίζεται ως εξής:

$$nidf_t = \frac{idf_t}{\ln(N)} = \frac{\ln\left(\frac{N}{n_t}\right)}{\ln(N)}$$

Χρησιμοποιώντας τους ορισμούς για τους παράγοντες $n f_{t,d}$ και $nidf_t$ προκύπτει ο ακόλουθος τρόπος υπολογισμού των βαρών:

$$w_{t,d} = n f_{t,d} \cdot nidf_t = \frac{f_{t,d}}{\max_x(f_{x,d})} \cdot \frac{\ln\frac{N}{n_t}}{\ln(N)}$$

Φυσικά στον χώρο της Πληροφορικής έχουν χρησιμοποιηθεί και προταθεί διάφορες παραλλαγές του τρόπου προσδιορισμού των βαρών χρησιμοποιώντας ως βάση το σχήμα $tf-idf$, ανάλογα με τις απαιτήσεις και τις ανάγκες του εκάστοτε προβλήματος.

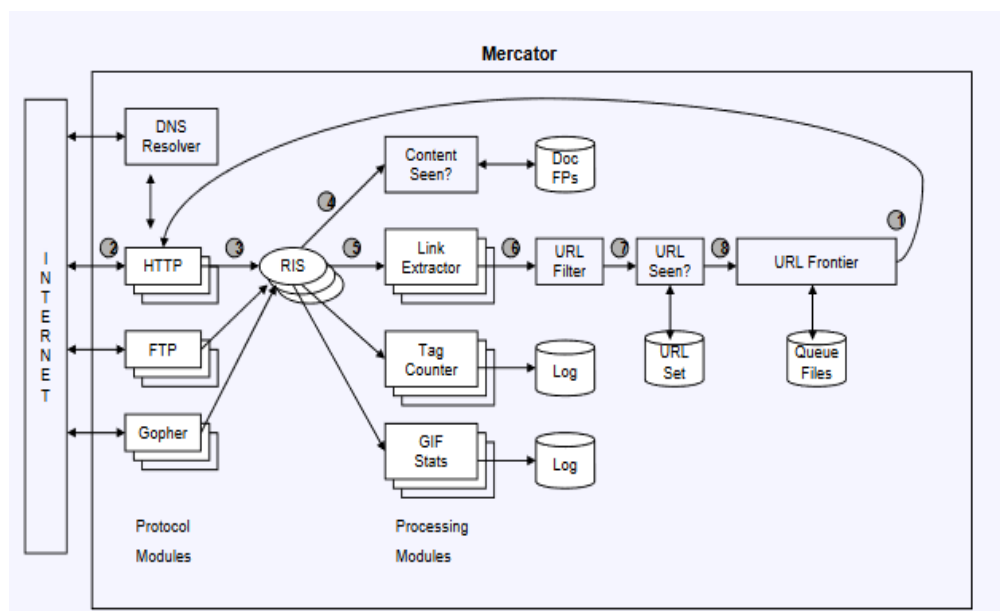
5.2 ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Η μηχανή αναζήτησης είναι μια εφαρμογή που επιτρέπει την αναζήτηση κειμένων και αρχείων στο Διαδίκτυο. Αποτελείται από ένα πρόγραμμα υπολογιστή που βρίσκεται σε έναν ή περισσότερους υπολογιστές στους οποίους δημιουργείται μια βάση δεδομένων με τις πληροφορίες που συλλέγονται από το διαδίκτυο, και το διαδραστικό περιβάλλον που εμφανίζεται στον τελικό χρήστη ο οποίος χρησιμοποιεί την εφαρμογή από άλλον υπολογιστή ο οποίος είναι διασυνδεδεμένος στο διαδίκτυο. Οι μηχανές αναζήτησης αποτελούνται από 3 είδη λογισμικού, τον επεξεργαστή ερωτημάτων, τον κατάλογο και τον διαχειριστή καταλόγου, και τον Crawler.

Ο Web Crawler, είναι ένα πρόγραμμα που διαβάζει τις ιστοσελίδες που υπάρχουν στο Web. Επίσης καλείται και Spider bot, η WebBot. Ο Crawler δεν είναι κάποιο είδος πράκτορα, τρέχει σε κάποιον server και απλά παράγει HTTP αιτήσεις για να «κατεβάσει» τις σελίδες. Το ίδιο πράγμα που κάνει και ένας χρήστης με την βοήθεια ενός Web Browser, με την διαφορά ότι ο crawler είναι πιο συστηματικός και φυσικά πιο γρήγορος. Ενδεικτικά, ο Metacore¹² είναι ένας crawler υψηλών επιδόσεων που δημιουργήθηκε από τους Allan Heydon, Marc Njork, Raymie Stata και συνεργάτες στο

¹² Heydon, Allan, and Marc Najork. "Mercator: A scalable, extensible web crawler." *World Wide Web* 2.4 (1999): 219-229.

Comaq Systems Research Center (αποτελεί συνέχιση της δουλειάς της Alta Vista). Ενώ υπάρχει και ο Heritrix¹³. Ο οποίος είναι ένας ανοικτού κώδικα crawler υψηλών επιδόσεων που υλοποιήθηκε από τον Raymie Stata και συνεργάτες στο Internet Archive.



Εικόνα 33 Δομή Μηχανής Αναζήτησης του WWW

Ο Κατάλογος (Indexer) μιας μηχανής αναζήτησης είναι συνήθως παραλλαγές του αντεστραμμένου καταλόγου, και η διαχείριση του απαιτεί τεράστια υπολογιστική ισχύ.

5.3 Η ΕΞΕΛΙΞΗ ΤΩΝ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ

Το 1990 εμφανίζεται η πρώτη μηχανή αναζήτησης περιεχομένου η *Archie Search Engine*. Η εφαρμογή υλοποιεί μια πολύ απλή μορφή αναζήτησης αρχείων (FTP) μεταξύ ενός local server και του χρήστη. Λέγεται πως είναι η πρώτη μηχανή αναζήτησης και πρόκειται για μια πολύ απλή αναζήτηση σε αρχεία τα οποία είναι πολύ περιορισμένα σε αριθμό.

Το 1993 εμφανίζεται επίσημα η πρώτη μηχανή αναζήτησης η *W3Catalog*¹⁴ η οποία δεν υλοποιείται με *crawler* και *indexer* αλλά πρόκειται για μια εφαρμογή που εκμεταλλεύεται το γεγονός ότι υπήρχαν πολλές ποιοτικές λίστες που περιέχουν διαδικτυακούς πόρους. Όμως η ανάγκη για συνεχή πρόσβαση σε κάθε σελίδα εκατοντάδες φορές μέσα στην μέρα από το *bot* που υλοποιούσε την αναζήτηση αποτέλεσε βασικό περιορισμό για την εξέλιξη της μηχανής. Την ίδια χρονιά κάνει την

¹³ <http://crawler.archive.org/team-list.html>

¹⁴ <https://www.w3catalog.com/>

εμφάνιση της και η *Aliweb*¹⁵ η οποία αντίθετα με την *W3Catalog* δεν χρησιμοποιεί κάποιο “web robot” αλλά βασίζεται σε ένα ευρετήριο το οποίο είχε συγκεκριμένη μορφή, και η ύπαρξη του κάθε ιστοτόπου στο ευρετήριο αυτό εξαρτόνταν από τους διαχειριστές του εκάστοτε ιστότοπου. Το σημαντικό πλεονέκτημα έναντι της *W3Catalog* ήταν ότι η απουσία του web robot (spider) βελτιώνει σημαντικά το εύρος ζώνης που χρησιμοποιείται, όμως οι περισσότεροι διαχειριστές ιστοτόπων δεν γνώριζαν την ανάγκη υποβολής της σελίδας τους. Αργότερα την ίδια χρονιά η μηχανή *JumpStation*¹⁶ αποτελεί την πρώτη μηχανή αναζήτησης που χρησιμοποιεί τα τρία βασικά χαρακτηριστικά μιας μηχανής αναζήτησης (*crawler*, *indexing* και *αναζήτηση*).

Αργότερα έκαναν την εμφάνιση τους οι πρώτοι κατάλογοι Ιστού που περιέχουν λίστες με ιστοτόπους του Web. Πιο συγκεκριμένα το 1994 ιδρύεται από τους Jerry Yang και David Filo ένας από τους σημαντικότερους καταλόγους ιστού ο *Yahoo! Directory*¹⁷, ο οποίος γίνεται και ο δημοφιλέστερος. Ενώ την ίδια χρονιά από τον Brian Pinkerton στο πανεπιστήμιο της Ουάσιγκτον δημιουργείται η πρώτη μηχανή αναζήτησης που επιτρέπει στους χρήστες να αναζητούν οποιαδήποτε λέξη σε οποιαδήποτε ιστοσελίδα η *WebCrawler*¹⁸, η οποία αποτέλεσε σημαντικό πρότυπο για όλες τις μηχανές αναζήτησης που έχουν βγει από τότε. Τον επόμενο μόλις χρόνο βγαίνουν δυο σημαντικές μηχανές αναζήτησης η *Lycos*¹⁹ και η *Yahoo!*²⁰ ενώ επίσης κυκλοφορεί και ο κατάλογος ιστού *LookSmart*²¹ που αποτελεί τον κύριο ανταγωνιστή του καταλόγου της *Yahoo!*. Το 1995 κυκλοφορεί το *Altavista*²² που είναι και η πρώτη μηχανή αναζήτησης που υποστηρίζει ερωτήματα φυσικής γλώσσας.

Εκτός από τους καταλόγους ιστών και τις μηχανές αναζήτησης κάνουν την εμφάνιση τους το 1996 και οι αλγόριθμοι βαθμολογίας Ιστοτόπων και πιο συγκεκριμένα κυκλοφορεί ο *RankDex* από τον Robin Li. Ο αλγόριθμος κατατάσει τα αποτελέσματα των μηχανών αναζήτησης χρησιμοποιώντας υπερσυνδέσμους για την μέτρηση της ποιότητας των ιστοσελίδων που ευρετηρίαζε. Εξέλιξη του αποτελεί ο αλγόριθμος *PageRank* της Google. Ενώ την ίδια χρονιά ο Larry Page και ο Sergey Brin εργάζονται στο *BackRub* που είναι ο προκάτοχος της *Google*²³ η οποία έγινε διαθέσιμη στο κοινό το 1997. Μέσα στα επόμενα χρόνια μέχρι και της αρχής του 2010 κάνουν την εμφάνιση τους μηχανές αναζήτησης όπως οι *Yandex*, το *MSN* που πολύ αργότερα εξελίχθηκε στο *Bing*, η *AlltheWeb*, η *Baidu*, η *Blekk*, η *DuckDuckGo*, η *Cuil* οι οποίες βασίζονται όλες σε τεχνολογίες που αναπτύχθηκαν προγενέστερα.

Το 2011 ξεκινά η ύπαρξη ΣΠ στο Web με τις Google, Yahoo! και Microsoft να ανακοινώνουν το *Schema.org* που αποτελεί μια πλούσια γκάμα ετικετών που μπορούν

¹⁵ <http://www.aliweb.com/>

¹⁶ <https://en.wikipedia.org/wiki/JumpStation>

¹⁷ https://en.wikipedia.org/wiki/Yahoo!_Directory

¹⁸ <https://www.webcrawler.com/>

¹⁹ <https://www.lycos.com/>

²⁰ <https://us.yahoo.com/>

²¹ <https://www.looksmart.com/>

²² <https://en.wikipedia.org/wiki/AltaVista>

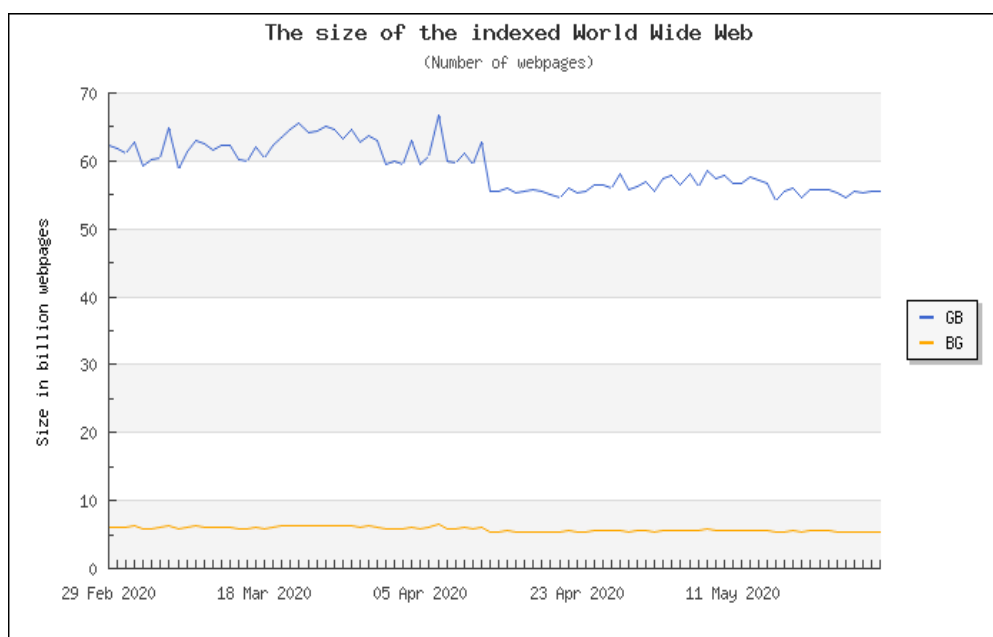
²³ https://en.wikipedia.org/wiki/History_of_Google

Πίνακας 3 Μηχανές Αναζήτησης

Yandex	https://yandex.com/
Msn	https://www.msn.com/el-gr/
Bing	https://www.bing.com/
AlltheWeb	https://en.wikipedia.org/wiki/AlltheWeb
Baidu	http://www.baidu.com/
Blekko	https://www.blekko.com
DuckDuckGo	https://duckduckgo.com/
Cuil	https://en.wikipedia.org/wiki/Cuil
Google	https://www.google.com/

να χρησιμοποιήσουν οι ιστότοποι για να μεταφέρουν καλύτερες πληροφορίες. Ενώ ένα χρόνο αργότερα το 2012 η Google να κυκλοφορεί το *Γράφημα Γνώσεων (Knowledge Graph)* το οποίο χρησιμοποιείται από την Google για την αποθήκευση σημασιολογικών σχέσεων μεταξύ αντικειμένων.

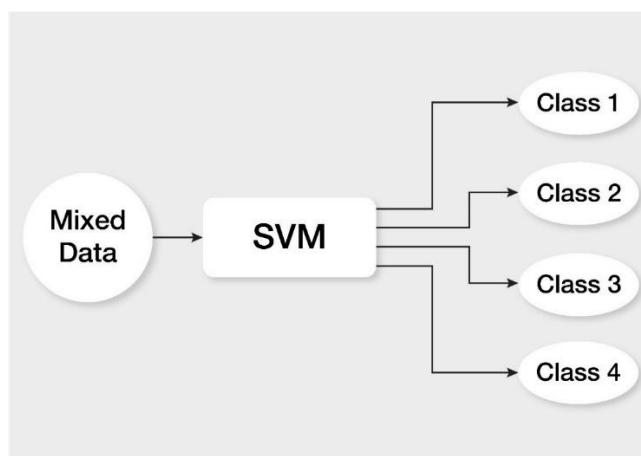
Το Web χαρακτηρίζεται ως η εφαρμογή «killer» για την ανάκτηση πληροφορίας. Υπάρχουν μεγάλες ποσότητες δεδομένων που πρέπει να ανακτηθούν που είναι καταναμημένες σε εκατομμύρια ιστοσελίδες που αλλάζουν συνεχώς και ενώ πολλά από αυτά τα δεδομένα επαναλαμβάνονται. Επίσης γίνεται μεγάλος λόγος για το πόσο ποιοτικά είναι τα δεδομένα που ανακτούμε από τον παγκόσμιο ιστό. Εκτός όμως από την μεγάλη ποσότητα των δεδομένων που υπάρχει για να ανακτηθεί, γίνεται λόγος και για το πόσο πολυμορφικά είναι τα δεδομένα στο Web. Υπάρχουν αδόμητα, ημιδομημένα και πλήρως δομημένα δεδομένα, ενώ παρουσιάζουν επίσης και σημαντική ετερογένεια.



Εικόνα 34 Μέγεθος του Παγκόσμιου Ιστού

5.4 ΔΙΚΤΥΟ ΣΗΜΑΣΙΟΛΟΓΙΚΗΣ ΟΜΟΙΟΤΗΤΑΣ (SEMANTIC SIMILARITY NETWORK)

Ο υπολογισμός της Σημασιολογικής Ομοιότητας μεταξύ δύο συνόλων λέξεων που περιγράφουν δυο οντότητες είναι ένα από τα σημαντικότερα προβλήματα στον τομέα του ΣΙ. Φυσικά είναι αναγκαίο εργαλείο και σε εφαρμογές που κάνουν οποιαδήποτε μορφή εξόρυξης γνώσης από το ΠΙ όπως αναζητήσεις, συστήματα συστάσεων (recommendation systems), και στον τομέα της στοχευμένης διαφήμισης. Οι παραδοσιακές τεχνικές ανάκτησης πληροφορίας από κείμενα που έχουν χρησιμοποιηθεί μέχρι και σήμερα για την υλοποίηση πολλών μηχανών αναζήτησης στο ΠΙ, αν και έχουν επεκταθεί αρκετά έτσι ώστε εκτός από απλή συσχέτιση μεταξύ των λέξεων να κάνουν συσχετίσεις μεταξύ των εννοιών (concepts) που αντιπροσωπεύουν οι λέξεις συνήθως αδυνατούν να αντικατοπτρίζουν την Σημασιολογική Γνώση που κρύβουν οι συσχετίσεις μεταξύ των εννοιών. Αυτό οφείλεται αφενός στο γεγονός ότι τα παραδοσιακά συστήματα δεν λαμβάνουν υπόψιν τους σχεδόν καθόλου τις Οντολογίες, και αφετέρου στο ότι εμπιστευόμαστε τυφλά κάποια μαθηματικά μοντέλα όπως το *Support Vector Machine (SVM)* που χρησιμοποιείται στα παραδοσιακά συστήματα ανάκτησης πληροφορίας κειμένου για να μας μεταφέρει από τον χώρο των λέξεων στον χώρο των εννοιών, χωρίς όμως να υπάρχει σαφής προσδιορισμός των εννοιών που αυτές εκπροσωπούν.



Εικόνα 35 Μοντέλο Support Vector Machine

Επομένως είναι κατανοητό ότι στα πλαίσια του Ιστού παρόλο που χρησιμοποιείται το μοντέλο *Term Vector Similarity matching* για τον λόγο ότι είναι αρκετά ικανοποιητικό και αρκετά απλό με μια λογική ευστοχία και ακρίβεια, είναι γνωστό πως η εξέταση μόνο των όρων οδηγεί σε προβλήματα που οφείλονται στην έλλειψη σημασιολογίας. Τα προβλήματα αυτά συνήθως οφείλονται στην πολυσημία και στην συνωνυμία.

Στόχος του ΠΙ είναι να αποτελέσει την επιτομή της νοημοσύνης των υπολογιστών με την ανθρώπινη γνώση σχετικά με συγκεκριμένους τομείς που αναπαρίστανται με την μορφή οντολογιών. Οι οντολογίες αυτές περιέχουν σημασιολογικές σχέσεις μεταξύ εννοιών που μπορούν να χρησιμοποιηθούν για την ανακάλυψη σχέσεων μεταξύ των οντοτήτων. Ένας από τους σημαντικότερους παράγοντες, όπως αναφέρθηκε σε προηγούμενη ενότητα, που εμποδίζουν στην επιτομή

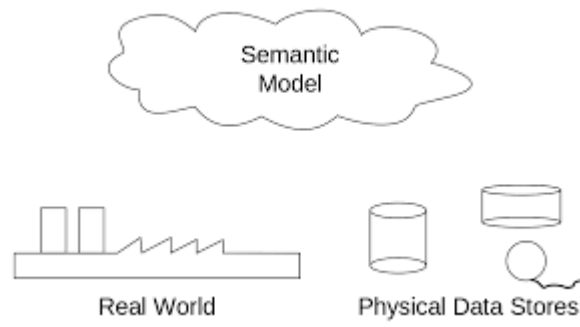
αυτή είναι η έλλειψη μηχανικά αναγνώσιμης γνώσης. Και η ζήτηση για νέες τεχνικές συσχέτισης πληροφορίας αποτελεί την βασικότερη τεχνολογία που πρέπει να μοντελοποιηθεί για την υλοποίηση του ΣΙ.

Στην παρούσα φάση ανάπτυξης του ΣΙ, το πρόβλημα είναι ότι δεν υπάρχει αρκετό υλικό σε Σημασιολογική Μορφή για να δημιουργηθεί ένα σημασιολογικό δίκτυο που θα μπορεί να αποτελέσει μια τεχνολογία πάνω στην οποία θα βασιστεί μια Μηχανή Αναζήτησης. Από την ιστορική αναδρομή της εξέλιξης των τεχνολογιών που αφορούν τις Μηχανές Αναζήτησης φαίνεται πώς το σημασιολογικό δίκτυο θα αποτελέσει μάλλον το πιο κατάλληλο εργαλείο στο οποίο θα βασιστούν οι μηχανές αναζήτησης της επόμενης γενιάς. Η σημασιολογική πληροφορία που υπάρχει πίσω από τις λέξεις δεν έχει αξιοποιηθεί σε κανένα από τα εγχειρήματα που συνόδευαν μέχρι τώρα την λειτουργία των μηχανών αναζήτησης, γιατί η συγκεκριμένη τεχνολογία δεν έχει φτάσει ακόμα στο ικανοποιητικό επίπεδο που ζητούμε, ενώ η παγκόσμια κοινότητα του Διαδικτύου, φαίνεται να είναι ανοιχτή πως οποιοδήποτε τεχνολογικό εγχείρημα βασίζεται πάνω στις έννοιες της Σημασιολογίας και του Σημασιολογικού Ιστού.

Στα πλαίσια αυτής της εργασίας, γίνεται εξαγωγή σημασιολογικής πληροφορίας μέσω των ΑΔΔ. Τα Ανοιχτά Διασυνδεδεμένα Δεδομένα αποτελούν πλέον αναπόσπαστο κομμάτι των περισσότερων συστημάτων τα οποία κάνουν οποιαδήποτε μορφής εξαγωγής γνώσης από το Διαδίκτυο. Η δομημένη σε RDF μορφή αναπαράστασης τους, καθώς και τα εξειδικευμένα εργαλεία για την εξαγωγή πληροφορίας, όπως η SPARQL, καθιστούν πιο εύκολη για την μηχανή την εξαγωγή και κατανόηση της πληροφορίας. Η πληροφορία που εξάγεται μέσω των ΑΔΔ πλέον στηρίζει πάρα πολλά συστήματα που δεν περιορίζονται μόνο στον τομέα του ΠΙ, άλλα στηρίζουν και πολλά συστήματα Μηχανικής Μάθησης (Machine Learning) και Ρομποτικής. Στόχος αυτής της εργασίας είναι να αναπτυχθεί ένα εργαλείο το οποίο θα χρησιμοποιεί την πληροφορία που βρίσκεται μέσα σε οντολογίες όπως η DBpedia και η ConceptNet και να ανακαλεί σημασιολογική πληροφορία. Το εργαλείο αυτό θα δέχεται μια λίστα από οντότητες και θα επιστρέφει μια λίστα από οντότητες που είναι σημασιολογικά «κοντινότερα» σε αυτές που δόθηκαν ως είσοδο. Τέτοια εργαλεία στα πλαίσια του ΣΙ, έχουν ως στόχο να φέρουν τις μηχανές ένα βήμα πιο κοντά στο να κατανοούν την σημασιολογία των λέξεων.

5.5 ΑΝΑΓΚΗ ΓΙΑ ΣΗΜΑΣΙΟΛΟΓΙΚΑ ΔΕΔΟΜΕΝΑ

Η δομή των δεδομένων ενός Συστήματος Ανάκτησης Πληροφορίας (ΣΑΠ), προφανώς δεν μπορεί να ικανοποιήσει πλήρως τις απαιτήσεις για έναν εννοιολογικό ορισμό των δεδομένων. Επομένως δεν μπορεί να αποτελέσει το απαραίτητο εργαλείο πάνω στο οποίο θα βασιστεί η ανάπτυξη του ΣΙ. Υπάρχει η ανάγκη καθορισμού δεδομένων από μια εννοιολογική άποψη, η οποία οδηγεί στην ανάπτυξη ολοένα και περισσότερων τεχνικών μοντελοποίησης σημασιολογικών δεδομένων, για τον καθορισμό της έννοιας των δεδομένων και των σχέσεων τους με άλλα δεδομένα. Σε γενικό επίπεδο, ο γενικός στόχος των σημασιολογικών μοντέλων δεδομένων είναι να συλλάβει περισσότερο το νόημα των δεδομένων, με ενσωμάτωση εννοιών που είναι ήδη γνωστές από το πεδίο της ΤΝ, και να αναπαρασταθεί ο φυσικός κόσμος και οι πραγματικές καταστάσεις με τρόπο που θα είναι κατανοητός από τις μηχανές.



Εικόνα 36 Σημασιολογικό Μοντέλο, Αναπαράσταση πραγματικού κόσμου με δεδομένα

ΚΕΦΑΛΑΙΟ 6: ΥΛΟΠΟΙΗΣΗ

ΥΛΟΠΟΙΗΣΗ

6 ΥΛΟΠΟΙΗΣΗ

Για την εξαγωγή ΣΠ από ΑΔΔ, αναπτύχθηκαν 5 διαδικασίες-κλάσεις σε γλώσσα Python:

- ConceptNet.py
- DBpedia.py
- WordNet.py
- Tfidf.py
- MainSemantic.py

Ο κώδικας αναπτύχθηκε με την χρήση του Ολοκληρωμένου Περιβάλλοντος Ανάπτυξης (Integrated Development Environment – IDE) Spyder 4.0.1.

Ως είσοδο στο πρόγραμμα δίνεται μια λέξη (για παράδειγμα ‘tea’) και η διαδικασία επιστρέφει μια λίστα λέξεων, που είναι σημασιολογικά κοντινότερα στην λέξη που δόθηκε. Υποστηρίζονται μόνο λέξεις της Αγγλικής γλώσσας.

6.1 CONCEPTNET.PY

Στο πρώτο κομμάτι του κώδικα χρησιμοποιήθηκε η βιβλιοθήκη «requests» για να λάβουμε μέσω του Web API του ConceptNet τις απαραίτητες πληροφορίες σε JSON μορφή.

```
response=requests.get('http://api.conceptnet.io/c/en/'+self.term+'?
offset=0&limit=1000')
obj=response.json()
```

Ως «self.term» συμβολίζουμε τον όρο-λέξη ο οποίος δόθηκε ως είσοδο από τον χρήστη. Το παραπάνω «response» αποτελείται από ένα σύνολο όρων που σύμφωνα με το ConceptNet είναι σημασιολογικά κοντινότερα στον όρο της αναζήτησης. Φυσικά ο κάθε όρος που επιστρέφεται συνδέεται σημασιολογικά με τον όρο της αναζήτησης μέσω μιας σχέσης (Relation), αυτό μπορούμε να το αναπαραστήσουμε με μια τριπλέτα RDF.

Στη συνέχεια επεξεργαζόμαστε την πληροφορία που έχουμε λάβει σε json μορφή, έτσι ώστε να την φέρουμε στην μορφή μιας δομής λεξικού (Dictionary)

```
'''Convert json text Data to a Dictionary'''
for resurl in obj['edges']:
    #take the source
    source = resurl['sources'][0]['@id']
    #take the id
    resurl1 = resurl['@id']
```

```

resurl1 = resurl1.replace(", ", " ")
tempWords = resurl1.split()
#filtering the sources of results
if 'wordnet' in source or 'verbosity' in source:
    relation = tempWords[0][7:][-1]
    #keep only the triples with the given entity
    if '/c/en/'+self.term+'/' in tempWords[1]:
        second=tempWords[2].split("/") [3]
    if '/c/en/'+self.term+'/' in tempWords[2]:
        second=tempWords[1].split("/") [3]
    weight = resurl['weight']
    #add on dictionary
    if relation in self.data:
        self.data[relation].append([second, weight])
    else:
        self.data[relation] = []
        self.data[relation].append([second, weight])

```

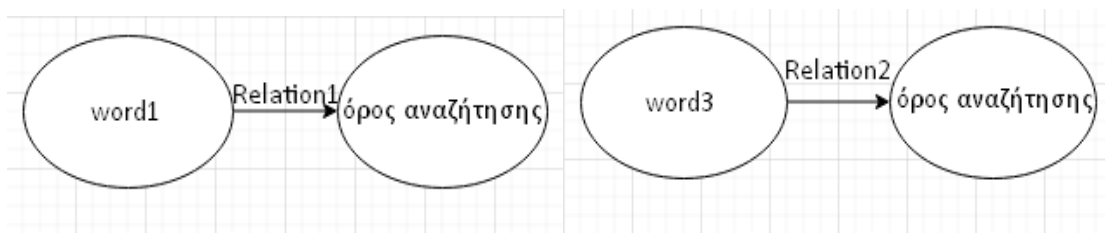
Η δομή λεξικού είναι στην μορφή

```

[[Relation1: [word1 , weight1],[word2 , weight2]],
 [Relation2: [word3 , weight3],[word4 , weight4]],
 ...
 [RelationN: [wordN , weightN],[wordN , weightN]]]

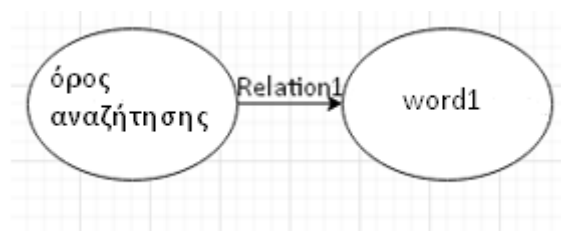
```

Όπου οι όροι «word1» και «word2» συνδέονται με τον όρο της αναζήτησης μέσω της σχέσης «Relation1» και οι όροι «word3» και «word4» συνδέονται με τον όρο της αναζήτησης μέσω της σχέσης «Relation2». Αυτό σε RDF τριπλέτες θα μπορούσε να αναπαρασταθεί ως εξής:



Εικόνα 37 Τριπλέτες RDF

Φυσικά οι όροι «word1» και «όρος αναζήτησης» ενδέχεται να είναι ανάποδα στην τριπλέτα.



Εικόνα 38 Τριπλέτα RDF για τον όρο αναζήτησης

Ενώ τα «weight1», «weight2», «weight3» και «weight4» είναι ένα επιπλέον δεδομένο που μας προσφέρει το ConceptNet ως προς το πόσο πολύ μοιάζουν δύο όροι, το οποίο ορίζεται από την κοινότητα του ConcceryNet. Για παράδειγμα αν αναζητήσουμε τον όρο «cherry» στο ConceptNet μέσω του API της εφαρμογής, θα πάρουμε το αποτέλεσμα:

```
RelatedTo
['fruit', 2.150074024348682]
['red', 1.399359801889989]
['small', 0.670909090909091]
['small_fruit', 0.603644453919114]
['sundae', 0.5470051023260155]
```

Εικόνα 39 Αποτελέσματα ConceptNet.py

Είναι προφανές ότι η λέξη «fruit», σημασιολογικά μοιάζει περισσότερο με τον όρο της αναζήτησης από την λέξη «sundae» η οποία έχει πολύ μικρή σημασιολογική ομοιότητα με τον όρο της αναζήτησης. Αυτό φαίνεται ξεκάθαρα άμα συγκρίνουμε τα «weight» των δύο λέξεων στα αποτελέσματα.

Τέλος, θεωρήθηκε στα πλαίσια της εργασίας ότι δεν είναι απαραίτητη η χρήση όλων των ιδιοτήτων (properties) που επιστρέφονται από το API του ConceptNet, για αυτό τον λόγο στο παρακάτω τμήμα του κώδικα αφαιρούνται όσα δεν θεωρούνται σημαντικά.

```
propertiesToRem = ['Synonym', 'Antonym', 'NotHasProperty',
'ExternalURL', 'EtymologicallyDerivedFrom', 'FormOf', 'HasSubevent',
'HasFirstSubevent', 'HasLastSubevent', 'DistinctFrom', 'SymbolOf',
'DefinedAs']

for remKey in propertiesToRem:
    self.data.pop(remKey, None)

return self.data
```

6.2 DBPEDIA.PY-TFIDF.PY

6.2.1 DBpedia.py

Σκοπός της κλάσης DBpedia.py είναι η διασύνδεση και η εξαγωγή δομημένης πληροφορίας από την Wikipedia. Πιο συγκεκριμένα τα δεδομένα τα οποία αντλούνται είναι ένα πεδίο τύπου Comment Box, το οποίο αποτελεί μια αφηρημένη περιγραφή της έννοιας που θέλουμε. Για παράδειγμα στην εικόνα που ακολουθεί το ζητούμενο πεδίο είναι το «abstract» του όρου «tea».

About: Tea

An Entity of Type : <http://dbpedia.org>, from Named Graph : <http://dbpedia.org>, within Data Space : dbpedia.org

(This article is about the beverage. For other uses, see Tea (disambiguation).)("Cup of tea" redirects here. For other uses, see Cup of Tea.) Tea is an aromatic beverage commonly prepared by pouring hot or boiling water over cured leaves of the *Camellia sinensis*, an evergreen shrub native to Asia. After water, it is the most widely consumed drink in the world. There are many different types of tea; some teas, like Darjeeling and Chinese greens, have a cooling, slightly bitter, and astringent flavour, while others have vastly different profiles that include sweet, nutty, floral or grassy notes.

Property	Value
dbpedia:abstract	<div><ul style="list-style-type: none">(This article is about the beverage. For other uses, see Tea (disambiguation).)("Cup of tea" redirects here. For other uses, see Cup of Tea.) Tea is an aromatic beverage commonly prepared by pouring hot or boiling water over cured leaves of the <i>Camellia sinensis</i>, an evergreen shrub native to Asia. After water, it is the most widely consumed drink in the world. There are many different types of tea; some teas, like Darjeeling and Chinese greens, have a cooling, slightly bitter, and astringent flavour, while others have vastly different profiles that include sweet, nutty, floral or grassy notes. Tea originated in southwestern China, where it was used as a medicinal drink. It was popularized as a recreational drink during the Chinese Tang dynasty, and tea drinking spread to other East Asian countries. Portuguese priests and merchants introduced it to the West during the 16th century. During the 17th century, drinking tea became fashionable among Britons, who started large-scale production and commercialization of the plant in India to bypass a Chinese monopoly at that time. The phrase herbal tea usually refers to infusions of fruit or herbs made without the tea plant, such as steeps of rosehip, chamomile, or rooibos. These are also known as tisanes or herbal infusions to distinguish them from "tea" as it is commonly understood. ^(en)</div>

Εικόνα 40 CommentBox του όρου «Tea» στο DBpedia

Η κλάση δέχεται ως είσοδο δύο ορίσματα, το πρώτο (data) είναι το λεξικό που δημιουργήθηκε στην κλάση ConceptNet.py, και το δεύτερο (term) είναι ο όρος που αναζητήθηκε στην αρχή της εκτέλεσης του προγράμματος από τον χρήστη.

```
def __init__(self, data, term):  
    self.data = data  
    self.term = term  
    self.DBdata = dict()
```

Στόχος είναι να δημιουργήσουμε αρχικά ένα λεξικό της μορφής:

```
[ [Relation1: [word1 , CommentBox1], [word2, CommentBox2] ],  
  [Relation2: [word3 , CommentBox3], [word4, CommentBox4] ],  
  ...
```



```
[RelationN: [wordN , CommentBoxN],[wordN,CommentBoxN]]
```

Στη συνέχεια του κώδικα καλούμε την μέθοδο «Query» για όλες τις λέξεις που βρίσκονται μέσα στο λεξικό «data». (‘σημείο-1’ του κώδικα)

```
def getData(self):
    for relation in self.data.keys():
        for word in self.data[relation]:
            ΣΗΜΕΙΟ-1 results = DBpedia.Query(word[0].capitalize())
            if not results["results"]["bindings"]:
                ΣΗΜΕΙΟ-2 results = DBpedia.Query(word[0].capitalize()+"_"+self.term+"")

            if not results["results"]["bindings"]:
                ΣΗΜΕΙΟ-3 wikiTerms = wikipedia.search(word[0].capitalize())
                wikiWord = wikiTerms[0].replace(" ", "_")
                results = DBpedia.Query(wikiWord)
                if not results["results"]["bindings"]:
                    if relation in self.DBdata:
                        self.DBdata[relation].append([word[0], None])
                    else:
                        self.DBdata[relation] = []
                        self.DBdata[relation].append([word[0], None])
                for result in results["results"]["bindings"]:
                    if relation in self.DBdata:
                        self.DBdata[relation].append([word[0],
result["abstract"]["value"]])
                    else:
                        self.DBdata[relation] = []
                        self.DBdata[relation].append([word[0],
result["abstract"]["value"]])
```

Σημειώνεται πως υπάρχουν κάποιες υποπεριπτώσεις σχετικά με την αναζήτηση που είναι αναγκαίο να προσδιοριστούν στον σχολιασμό του κώδικα που αναπτύχθηκε έτσι ώστε να ξεκαθαριστεί ο τρόπος με τον οποίο υλοποιήθηκε η αναζήτηση. Έστω ότι για παράδειγμα, θέλουμε να ανακτήσουμε το Comment Box του όρου «Bass» (fish). Στη DBpedia δεν υπάρχει ο όρος Bass σκέτος. Αυτό συμβαίνει επειδή εξαιτίας της πολυσημίας της λέξης, δεν είναι ξεκάθαρο το σε ποια οντότητα αναφερόμαστε, κι αυτό επειδή η λέξη «Bass» μπορεί να αναφέρεται στον Μπάσο Ήχο, η μπορεί να αναφέρεται στο Ψάρι. Στην περίπτωση αυτή λοιπόν, αν αναζητήσουμε σκέτο τον όρο «Bass» δεν θα λάβουμε αποτέλεσμα από την μέθοδο «Query». Επομένως όπως φαίνεται στο ‘σημείο-2’ του κώδικα προσθέτουμε στο τέλος της λέξης τον απαιτούμενο προσδιορισμό για να αντιμετωπίσουμε το πρόβλημα της πολυσημίας, στην περίπτωση του παραδείγματος που αναλύουμε, προσθέτουμε στο τέλος της συμβολοσειράς «Bass» την συμβολοσειρά «_(fish)». Δηλαδή στην γενική περίπτωση προσθέτουμε την συμβολοσειρά «_(term)», που όπου term είναι ο όρος που δόθηκε από τον χρήστη στην αρχή της εκτέλεσης του προγράμματος. Αυτή η υποπερίπτωση βελτίωσε τα αποτελέσματα της κλάσης DBpedia αρκετά, αφού πιάνει υποπεριπτώσεις που θα χάναμε αν δεν την υλοποιούσαμε με αυτόν τον τρόπο.

Όμως υπάρχει πάλι η περίπτωση στο DBpedia να μην υπάρχει καταχωρημένη η λέξη που αναζητήσαμε ακόμα και με την συνένωση της με την συμβολοσειρά «_(term)». Σε αυτήν την περίπτωση, στο ‘σημείο-3’ του κώδικα αναζητούμε την λέξη πρώτα στο Wikipedia. Ο λόγος που γίνεται αυτό είναι για να αντιμετωπίσουμε

προβλήματα με λέξεις που αναφέρονται στην ίδια έννοια αλλά μόνο η μία από όλες αυτές είναι καταχωρημένη στην DBpedia. Για παράδειγμα η λέξεις «Schrod» και «Scord» αναφέρονται στην ίδια οντότητα, όμως η οντότητα αυτή είναι καταχωρημένη στη DBpedia μόνο ως Scrod. Επομένως με την αναζήτηση που κάνουμε στη Wikipedia μετατρέπουμε την λέξη «Schrod» σε «Scrod» και στη συνέχεια του κώδικα καλώντας την μέθοδο «Query» με την λέξη «Scrod» ως είσοδο ανακτούμε τα δεδομένα που επιθυμούμε. Ο λόγος που είναι ασφαλές να κάνουμε αυτήν την μετατροπή είναι επειδή η δομημένη πληροφορία που βρίσκεται στη DBpedia έχει προέλθει από την Wikipedia.

Όπως και στην κλάση ConceptNet στόχος της κλάσης DBpedia είναι να δημιουργηθεί ένα λεξικό που θα είναι της μορφής:

```
[ [Relation1: [word1 , weight1], [word2 , weight2]],  
  [Relation2: [word3 , weight3], [word4 , weight4]],  
  ...  
  [RelationN: [wordN , weightN], [wordN , weightN]] ]
```

Όμως στην περίπτωση της κλάσης DBpedia το βάρος (weight) δεν θα είναι μια μετρική που θα αναφέρει πόσο σημασιολογικά όμοιες είναι οι δύο λέξεις, αλλά ορίζεται ως το Cosine Similarity των δύο CommentBox που αντιστοιχούν στον όρο της αναζήτησης, και στον όρο που υπάρχει στο λεξικό (word1, word2,...) αντίστοιχα.

Όπως φαίνεται στο κομμάτι του κώδικα που ακολουθεί, πρώτα υπολογίζεται το Cosine Similarity των κειμένων, και στο τέλος από την κλάση DBpedia επιστρέφεται το λεξικό που αναφέραμε.

```
TfIdf_CosineSimilarityData = TfIdf(self.DBdata, self.term).getData()  
return TfIdf_CosineSimilarityData
```

6.2.2 Η μέθοδος *Query της DBpedia.py*

Στην μέθοδο που ακολουθεί, υλοποιήθηκε με την χρήση της SPARQL γλώσσας το ερώτημα (query) που χρειαζόταν για να ανακτηθεί το κατάλληλο πεδίο ‘abstract’ που αναφέρθηκε προηγουμένως.

```
def Query(word) :  
    sparql = SPARQLWrapper("http://dbpedia.org/sparql")  
    sparql.setQuery("""  
        SELECT ?abstract  
        WHERE {  
<http://dbpedia.org/resource/"""+word+"""> dbo:abstract ?abstract  
        FILTER (lang(?abstract) = 'en') }  
        """)  
    sparql.setReturnFormat(JSON)  
    results = sparql.query().convert()  
return results
```

Η πράξη που κάνουμε με το SPARQL ερώτημα είναι η ‘SELECT’, με την οποία επιλέγουμε το πεδίο ‘abstract’. Στην συνθήκη ‘WHERE’ έχουμε ορίσει τον σύνδεσμο στον οποίο υλοποιούμε το ερώτημα, και ως ‘FILTER’ ορίζουμε το αποτέλεσμα να επιστρέφεται μόνο αν η γλώσσα είναι η αγγλική (‘en’). Τέλος, το αποτέλεσμα του ερωτήματος μετατρέπεται σε μορφή JSON για να είναι ευκολότερη η επεξεργασία του από την κλάση DBpedia.py.

6.2.3 *TfIdf.py*

Στόχος της κλάσης είναι να υπολογίσει την ομοιότητα μεταξύ του κάθε Comment Box που ανακτήσαμε από τη DBpedia με το Comment Box του όρου που έδωσε ο χρήστης στην είσοδο. Για τον υπολογισμό της ομοιότητας χρησιμοποιήθηκαν η ομοιότητα συνημιτόνου (cosine similarity) σε διανύσματα που δημιουργήθηκαν με την χρήση του μοντέλου tf-idf.

Η κλάση δέχεται ως είσοδο δύο ορίσματα το data που είναι τα δεδομένα που έχουμε ανακτήσει από τη DBpedia από μεθόδους που αναλύθηκαν προηγουμένως και τον όρο term που είναι ο όρος που δόθηκε από τον χρήστη ως είσοδο στο πρόγραμμα.

```
def __init__(self, data, term):
    self.term = term
    self.data = data
    self.corpus = []
    self.weights = []
    self.TfIdfData = dict()
```

Στο όρισμα «corpus» θα αποθηκευτεί το αποτέλεσμα από την ομοιότητα συνημιτόνου. Ενώ το όρισμα εξόδου «TfIdfData» είναι ένα λεξικό που θα πάρει τη μορφή:

```
[[Relation1: [word1 , CosSim1],[word2 , CosSim2]],
 [Relation2: [word3 , CosSim3],[word4 , CosSim4]],
 ...
 [RelationN: [wordN , CosSimN],[wordN , CosSimN]]]
```

Όπου «CosSimn» συμβολίζουμε το Cosine Similarity του CommentBox της λέξης «wordn» με το CommentBox του όρου που δόθηκε από τον χρήστη. Θυμίζουμε πως και τα δύο έχουν ανακτηθεί από την κλάση DBpedia.py.

Στην αρχή της κλάσης με την μέθοδο «getData» ανακτούμε από το DBpedia το CommentBox του όρου «term» που είναι η λέξη που δόθηκε ως είσοδος από τον χρήστη.

```
word = self.term.capitalize()
sparql = SPARQLWrapper("http://dbpedia.org/sparql")
sparql.setQuery("""
```

```

        SELECT ?abstract
        WHERE {
<http://dbpedia.org/resource/""+word+""> dbo:abstract ?abstract
        FILTER (lang(?abstract) = 'en')}
        """)
    sparql.setReturnFormat(JSON)
    result = sparql.query().convert()
    if not result["results"]["bindings"]:
        '''3rd try || word schrod doesn't exist, but Scrod
(same word) exists
        uses wikipedia to correct the word '''
        wikiTerms = wikipedia.search(word[0].capitalize())
        #replace " " with "_" because the words in DBpedia are
seperated by this char "_"
        wikiWord = wikiTerms[0].replace(" ", "_")
        result = TfIdf.Query(wikiWord)
    termCB =
result["results"]["bindings"][0]["abstract"]["value"] #get
CommentBox for term
    self.corpus.append(termCB)

```

Στη συνέχεια προσθέτουμε (append) στη δομή «corpus» όλα τα CommentBoxes που έχουμε ανακτήσει από το DBpedia, και με αυτόν τον τρόπο δημιουργούμε το σύνολο των εγγράφων στα οποία θα εφαρμόσουμε tf-idf score όπως αναλύθηκε στην ενότητα 4.

```

for key in self.data.keys():
    for triple in range(len(self.data[key])):
        if(self.data[key][triple][1] == None):
            self.corpus.append("WeHaveNoDataFromDBpedia")
        else:
            self.corpus.append(self.data[key][triple][1])

```

Έπειτα με τη βοήθεια σχετικών βιβλιοθηκών δημιουργούμε τα διανύσματα πάνω στα οποία θα εφαρμόσουμε την ομοιότητα συνημιτόνου.

```

vectorizer = TfidfVectorizer()
vectors = vectorizer.fit_transform(self.corpus)
dense = vectors.todense()
denselist = dense.tolist()
for vector in denselist:
    self.weights.append(Tfidf.getCosSimilarity(denselist[0],
vector))

```

Στο τέλος της κλάσης δημιουργούμε το λεξικό που είναι στη μορφή που αναφέραμε στην αρχή αυτής της υποενότητας, και το επιστρέφουμε ως αποτέλεσμα της μεθόδου.

```

wpos = 1
for relation in self.data.keys():

```

```
        for triple in range(len(self.data[relation])):
            if relation in self.TfIdfData:

self.TfIdfData[relation].append([self.data[relation][triple][0],
self.weights[wpos]])
                else:
                    self.TfIdfData[relation] = []

self.TfIdfData[relation].append([self.data[relation][triple][0],
self.weights[wpos]])
                    wpos = wpos + 1
        return self.TfIdfData
```

Υπάρχει επίσης η μέθοδος «getCosineSimilarity» που δέχεται δύο ορίσματα «vector1» και «vector2» που είναι δύο διανύσματα, και υπολογίζει την ομοιότητα συνημιτόνου.

ΚΕΦΑΛΑΙΟ 7: ΑΠΟΤΕΛΕΣΜΑΤΑ

ΑΠΟΤΕΛΕΣΜΑΤΑ

7 ΑΠΟΤΕΛΕΣΜΑΤΑ

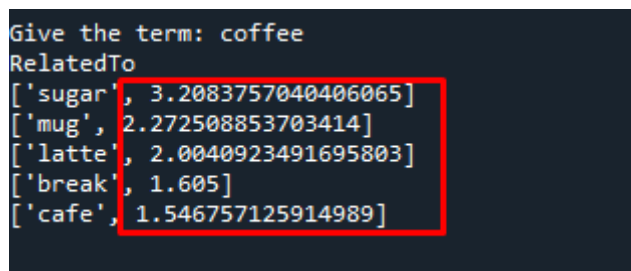
Στο παρόν κεφάλαιο μελετάμε τα αποτελέσματα κάποιων ορών που δίνονται ως είσοδο. Σημειώνουμε ότι επειδή δεν υπάρχει ένα data model δεν μπορούμε να κάνουμε αξιολόγηση του λογισμικού πάνω σε κάποιο σύνολο δεδομένων. Για αυτόν τον λόγο δίνουμε ενδεικτικά κάποια αποτελέσματα έτσι ώστε να γίνει κατανοητή η εκτέλεση του λογισμικού που αναπτύχθηκε και να ελεγχθεί η ποιότητα των αποτελεσμάτων του λογισμικού.

7.1 ΣΧΕΤΙΚΑ ΜΕ ΤΗΝ ΣΥΝΟΛΙΚΗ ΜΕΤΡΙΚΗ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ

Η τελική μετρική που υπολογίζεται στην κλάση MainSemantic.py είναι η εξής:

$$weight_w = \frac{1}{WordNetDistance_w} + cosineSim_{w,t} + ConceptNetWeight_w$$

Με $weight_w$ συμβολίζουμε το τελικό βάρος που έχει μια λέξη ανάλογα με το πόσο μοιάζει με τον όρο (term) που δόθηκε ως είσοδο από τον χρήστη. Για παράδειγμα, αν δώσουμε ως είσοδο τον όρο «coffee» το αποτέλεσμα που παίρνουμε είναι το εξής:



```
Give the term: coffee
RelatedTo
['sugar', 3.2083757040406065]
['mug', 2.272508853703414]
['latte', 2.0040923491695803]
['break', 1.605]
['cafe', 1.546757125914989]
```

Εικόνα 41 Τελικά βάρη στις οντότητες

Το $weight_w$ σημειώνεται με κόκκινο πλαίσιο, ενώ τα αποτελέσματα είναι ταξινομημένα με βάση αυτό.

Η μετρική αυτή συνυπολογίζεται με τη συνεισφορά των τριών βαρών που υπολογίσαμε από τις τρεις βασικές κλάσεις που υπάρχουν στο λογισμικό. Αυτές είναι οι: ConceptNet.py από όπου παίρνουμε την μετρική $ConceptNetWeight_w$, DBpedia.py/Tfidf.py από όπου παίρνουμε τη μετρική $cosineSim_{w,t}$ και την WordNet.py όπου υπολογίζεται η μετρική $WordNetDistance_w$.

Αρχικά, η μετρική $ConceptNetWeight_w$ αποτελεί το βάρος που ορίζεται μέσα στο δίκτυο του ConceptNet. Σε αυτό δεν έχει γίνει καμία απολύτως αλλαγή από το λογισμικό μας και χρησιμοποιείται αυτούσιο.

Η μετρική $cosineSim_{w,t}$ θυμίζουμε ότι είναι η ομοιότητα συνημιτόνου μεταξύ του Comment Box που έχουμε εξορύξει από το DBpedia για μια λέξη «w» και του Comment Box του όρου που δόθηκε ως είσοδο από τον χρήστη «t».

Τέλος η μετρική $WordNetDistance_w$ ορίζει την απόσταση που έχει στο δίκτυο του wordnet η λέξη «w» από τον όρο «t» που δόθηκε ως είσοδος από τον χρήστη. Επειδή αυτή η μετρική αναπαριστά μια απόσταση σε ένα δέντρο, πήραμε τον λόγο $\frac{1}{WordNetDistance_w}$ έτσι ώστε όσο πιο κοντά βρίσκεται μια λέξη «w» στον όρο «t» τόσο πιο κοντά στη μονάδα θα είναι το πηλίκο του λόγου αυτού. Σημειώνουμε επίσης ότι αν μια λέξη δεν βρεθεί μέσα στο δίκτυο του WordNet τότε η μετρική $WordNetDistance_w$ παίρνει την τιμή 1000 έτσι ώστε ο λόγος να είναι κοντά στο 0.

7.1.1 Αποτελέσματα για τον όρο αναζήτησης «time»

Στις εικόνες που ακολουθούν φαίνονται τα αποτελέσματα αναζήτησης για τον όρο «time». Αρχικά, στην εικόνα που ακολουθεί φαίνεται το πεδίο «RelatedTo» (σχετίζεται με) που περιλαμβάνει το σύνολο εκείνων των λέξεων με τις οποίες σχετίζεται περισσότερο σημασιολογικά ο όρος «time».


```

Give the term: time
RelatedTo
['minute', 10.099120879776951]
['hour', 9.427018317855124]
['clock', 9.211774166556175]
['watch', 7.727845922148497]
['morning', 7.024847036987094]
['evening', 6.937927140888179]
['hours', 6.444574719717019]
['minutes', 6.438498678010946]
['year', 6.204632217616342]
['month', 5.797691887567057]
['seconds', 5.695593525808177]
['always', 5.629497912787117]
['day', 5.460611279606702]
['century', 5.451047192374548]
['measurement', 4.9601375927551645]
['week', 4.568541733572539]
['watch', 4.534522883424791]
['clocks', 4.4172407935671565]
['age', 3.9221477155010507]
['hour', 3.7540345446157315]
['measure', 3.665262179231507]
['ago', 3.6528478478161857]
['moment', 3.4921399772913047]
['now', 3.3869438693108216]
['minute', 3.2001992995906123]
['dimension', 3.0298160650525805]
['party', 2.916501035467129]
['past', 2.869059969376905]
['present', 2.7179309784559207]
['passing', 2.6661314501522395]
['hours_minutes', 2.6448118592658965]
['break', 2.539978165834775]
['clock', 2.52592360371548]
['night', 2.519579976253114]
['future', 2.5148725549144166]
['recent', 2.372279081517542]
['stitch', 2.3664922316412014]
['alarm_clock', 2.32154689907146]
['temporal', 2.29128384422406]
['measures', 2.2319368814807126]
['chronological', 2.215886279836239]
['rest', 2.1011111111111114]
['clock_measures', 2.010099075302716]
['when', 1.9629990009990008]
['seconds_minutes', 1.8769990009990014]
['tells', 1.8729230769230771]
['calendar', 1.7222922180057127]
['sleep', 1.655177908106796]
['minutes_hours', 1.6514791050384934]
['second', 1.6139984425679577]
['fourth_dimension', 1.613909090909091]
['fourth', 1.613909090909091]
['hour_minute', 1.6028743834690429]
['late', 1.593999000999001]
['days', 1.5811245853761418]
['birthday', 1.566530679569641]
['story', 1.5091111111111113]
['watches', 1.5031511973517144]
['minutes_seconds', 1.4589990009990013]
['current', 1.4434285714285713]
['after', 1.425999000999001]
['play', 1.4229090909090907]

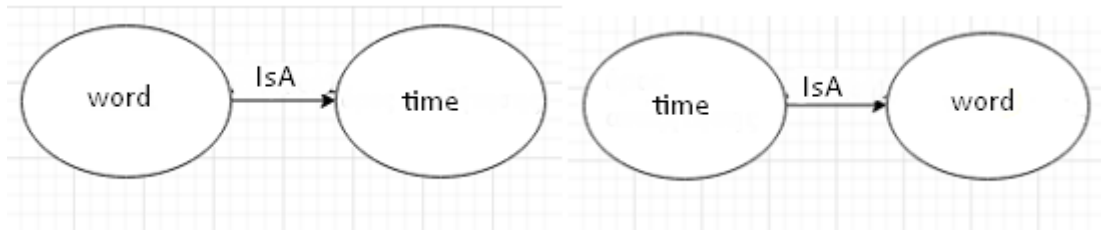
```

Εικόνα 42 Αποτελέσματα αναζήτησης για τον όρο «Time»

Ενδεικτικά αναφέρουμε ότι η λέξεις που σχετίζονται περισσότερο σημασιολογικά με τον όρο αναζήτησης είναι οι minute, hour, clock, watch, morning, evening, hours, minutes, year, month, seconds, always, day. Κάτι τέτοιο είναι λογικό αφού οι περισσότερες από αυτές τις λέξεις είτε αναφέρονται για να περιγράψουν μια χρονική στιγμή (time) μέσα στη μέρα, είτε είναι αυτές που μας ενημερώνουν για τον χρόνο (day, month, year, clock, watch).

Ενώ φυσικά υπάρχουν και λέξεις που σχετίζονται λιγότερο με τον όρο. Κάποιες από αυτές είναι οι: fourth_dimension που σχετίζεται με τον όρο «time» επειδή ο χρόνος αποτελεί την τέταρτη διάσταση, ενώ φυσικά σχετίζεται και με σκέτη την λέξη «fourth» για αυτόν τον λόγο. Επίσης, σχετίζεται με την λέξη «tells» και «measures» αφού ο όρος «time» χρησιμοποιείται πολύ συχνά όταν θέλουμε να ρωτήσουμε «τι ώρα είναι» και αντιστοίχως όταν θέλουμε να ενημερώσουμε κάποιον για το «τι ώρα είναι». Ενώ επίσης ο χρόνος αποτελεί μια μετρική «measures» της ώρας που έχει περάσει.

Επίσης έχει εξορυχθεί και ένα άλλο πεδίο τύπου «IsA» που ορίζει μια τριπλέτα της μορφής:



Εικόνα 43 τριπλέτες RDF για τον όρο «Time»

Στην εικόνα 44 φαίνονται ενδεικτικά κάποια αποτελέσματα της σχέσης «IsA».

```
IsA
['future', 2.6108725549144167]
['day', 2.566124585376142]
['geological_time', 2.557285136625069]
['present', 2.5133902099510603]
['moment', 2.511536011649441]
['moment', 2.511536011649441]
['case', 2.5]
['time_period', 2.488876656340896]
['greenwich_mean_time', 2.4859388567370324]
['past', 2.4841071396546903]
['experience', 2.429969662726194]
['eternity', 2.4116289370261152]
['civil_time', 2.377134423026222]
['while', 2.3291708201734265]
['incarnation', 2.3219311604144264]
['daylight_saving_time', 2.3083720323604955]
['space_age', 2.303961776229816]
['wee', 2.2960294013042173]
['biological_time', 2.295040601330114]
['occasion', 2.25]
['musical_time', 2.1942269181498792]
['ephemera', 2.1703307826139437]
['continuum', 2.111111111111111]
['dead', 2.111111111111111]
['cosmic_time', 2.111111111111111]
['attribute', 2.1]
['duration', 2.1]
['high_time', 2.090909090909091]
['hard_times', 2.090909090909091]
```

Εικόνα 44 Αποτελέσματα της σχέσης «IsA»

7.1.2 Αποτελέσματα για τον όρο αναζήτησης «sugar»

Στην εικόνα που ακολουθεί φαίνονται τα αποτελέσματα αναζήτησης σημασιολογικής πληροφορίας για τον όρο «sugar».

```

Give the term: sugar
RelatedTo
['sweet', 8.617034445694014]
['white', 6.225406211815076]
['sweetener', 5.632997628976413]
['salt', 3.7708408157230617]
['cane', 3.7281182247921736]
['powder', 3.347768464653846]
['granules', 3.2707998145547994]
['crystals', 3.194699941665205]
['food', 3.144346562253094]
['coffee', 3.1093809432206374]
['grains', 2.965709546716773]
['sucrose', 2.737657159172084]
['cake', 2.7336915938412827]
['ingredient', 2.664274348062939]
['spice', 2.5908580755596873]
['salt', 2.517244015784319]
['condiment', 2.4599194916733764]
['sweetness', 2.304463696900606]
['tea', 1.9209858153761579]
['candy', 1.7137824040090155]
['taste', 1.5975544916463025]
['glucose', 1.5908916989265898]
['granular', 1.5704096841701172]
['sweet_substance', 1.5684636969006063]
['baking', 1.5003842675021428]
['white_powder', 1.3863530674964646]
['sweet_powder', 1.377466103216974]
['substance', 1.3001111111111112]
['white_granules', 1.2877969999381043]
['sweet_granules', 1.2095756962846327]
['white_grains', 1.1442972903982853]

```

Εικόνα 45 Αποτελέσματα αναζήτησης για τον όρο «sugar»

Από την εικόνα βγάζουμε το συμπέρασμα ότι περισσότερο η λέξη «sugar» σχετίζεται σημασιολογικά με τη λέξη «sweet» αφού η ζάχαρη είναι γλυκιά, με τον όρο «white» αφού η ζάχαρη είναι άσπρη.

Σχετίζεται όμως και σε μεγάλο βαθμό με την λέξη «salt». Αυτό οφείλεται αφενός στο γεγονός ότι και τα δύο είναι άσπροι κόκκοι και μοιάζουν μεταξύ τους, και αφετέρου στο γεγονός ότι τα συναντούμε σε κοινά μέρη (για παράδειγμα στην κουζίνα ενός σπιτιού, σε φαγητά, σε βάζα).

Επίσης είναι αξιοσημείωτο ότι στα αποτελέσματα με λιγότερη σημασιολογική ομοιότητα εμφανίζονται και οι λέξεις «coffee» και «tea» όπου αντικατοπτρίζεται το χαρακτηριστικό του πραγματικού κόσμου, ότι συνήθως βάζουμε ζάχαρη στον καφέ μας και στο τσάι μας.

Όπως και στο προηγούμενο παράδειγμα έτσι και εδώ στο πεδίο «IsA» έχουμε αντίστοιχα αποτελέσματα

```
IsA
['sweetening', 2.878464695901605]
['brown_sugar', 2.801614188447414]
['lump_sugar', 2.6148571700316126]
['sugarloaf', 2.5630334919999043]
['cane_sugar', 2.506962200196294]
['granulated_sugar', 2.5]
['beet_sugar', 2.46648704910757]
['caramel', 2.3211718726128003]
['corn_sugar', 2.0588235294117645]
```

Εικόνα 46 αποτελέσματα για την σχέση «IsA»

Συναντάμε λέξεις όπως «brown sugar» και «corn sugar» που και στις δύο περιπτώσεις έχουμε ένα είδος ζάχαρης.

7.1.3 Αποτελέσματα για τον όρο αναζήτησης «red»

Στην εικόνα 47 φαίνονται τα αποτελέσματα αναζήτησης ΣΠ για τον όρο «red».

```
Give the term: red
RelatedTo
['apple', 9.618761997252793]
['color', 8.352962315312173]
['blood', 7.554477576396906]
['color', 6.286996275001902]
['stop', 4.964726577070207]
['wine', 4.301985644714426]
['colour', 4.278883653030679]
['fire', 4.0997156600761215]
['blue', 3.7578658634725595]
['colour', 3.462502894203068]
['squirrel', 3.0707869231065454]
['tomato', 2.7671513086985415]
['fox', 2.393720402500796]
['blush', 2.3624724292292]
['scarlet', 2.347]
['lip', 1.9423143646398509]
['raspberry', 1.8003113546076133]
['farm', 1.7202365446092538]
['primary_colour', 1.6904445943262418]
['primary', 1.5017125970646288]
['maroon', 1.4612013509456794]
['blood', 1.446857844526732]
['cherry', 1.393559457029873]
['sunburn', 1.3349813401639676]

PartOf
['louisiana', 2.4111746259372486]
['texas', 2.3983029844049044]
['oklahoma', 2.3544237558008048]

IsA
['dark_red', 2.8112782740226026]
['crimson', 2.8109517342808026]
['purplish_red', 2.774727129305675]
['chromatic_color', 2.755617737188375]
['turkey_red', 2.7528881243095435]
['sanguine', 2.7246219623721792]
['chrome_red', 2.601862809521325]
['cerise', 2.5]
['scarlet', 2.5]
['cardinal', 2.0833333333333335]
```

Εικόνα 47 Αποτελέσματα αναζήτησης για τον όρο «red»

Αρχικά στο πεδίο «RelatedTo» συναντάμε κάποια πράγματα τα οποία είναι κόκκινα, όπως το μήλο, το αίμα, την πινακίδα «stop», την φωτιά, το κρασί, την αλεπού, την τομάτα. Ενώ επίσης υπάρχουν και κάποια άλλα πράγματα τα οποία δεν είναι κόκκινα αλλά έχουν ισχυρή σημασιολογική ομοιότητα με την λέξη «κόκκινο». Κάποια από αυτά είναι οι λέξεις «χρώμα», το «μπλε» γιατί πολύ συχνά το μπλε ορίζεται ως το

αντίθετο χρώμα του κόκκινου σύμφωνα με την κοινή γνώμη, ενώ υπάρχει και η λέξη «primary colour» που εννοεί ότι το χρώμα «κόκκινο» είναι ένα από τα βασικά χρώματα.

Στο πεδίο «PartOf» είναι αξιοσημείωτο το ότι έχουν εμφανιστεί όροι όπως το «texas» πιθανόν από την formula «Texas red»²⁴ ενώ εμφανίζεται και η λέξη «Oklahoma» που ετυμολογικά προέρχεται από τις λέξεις «okla» + «humma» που σημαίνει «κόκκινοι άνθρωποι».

Τέλος στο πεδίο «IsA» όπως και στα αντίστοιχα πεδία των προηγούμενων παραδειγμάτων υπάρχουν οντότητες και πράγματα τα οποία είναι κόκκινα, η το κόκκινο είναι κάτι από αυτά.

7.1.4 Αποτελέσματα για τον όρο αναζήτησης «kitchen»

Όπως φαίνεται στην εικόνα 48 (η οποία είναι μέρος της εργασίας [29]) υπάρχουν τα αποτελέσματα αναζήτησης ΣΠ για τον όρο «kitchen». Σύμφωνα με τα αποτελέσματα της εικόνας, η λέξη «kitchen» σχετίζεται σημασιολογικά με τις λέξεις τραπέζι, πιάτο, φούρνος, νεροχύτης, ψυγείο, καρέκλα, μίξερ, ποτήρι, σπάτουλα, σπίτι, φαγητό, αλάτι, κουταλάκι, βούτυρο. Αυτές οι λέξεις αποτελούν πράγματα τα οποία βρίσκονται μέσα σε μία κουζίνα.

Αν επιστρέψουμε στο παράδειγμα μας στην ενότητα 1.4 το οποίο αφορά τον συλλογισμό που κάνει ένας πράκτορας ο οποίος προσπαθεί να ανιχνεύσει ότι βρίσκεται μέσα σε μια κουζίνα. Ο συλλογισμός αυτός θα μπορούσε να είναι βασισμένος σε λογισμικά όπως αυτό που έχει αναπτυχθεί στα πλαίσια αυτής της εργασίας. Φυσικά μπορούμε να θεωρήσουμε ότι ο συλλογισμός όχι μόνο είναι σωστός, αλλά μπορεί να γίνει «on the fly» από τον πράκτορα μας αν αυτός είναι σε θέση να εξάγει ΣΠ με τον τρόπο που δουλεύει το λογισμικό που αναπτύχθηκε, και επομένως δεν απαιτείται ειδικός σχεδιασμός συλλογισμού για κάθε διαφορετικό δωμάτιο στο οποίο πιθανόν να βρίσκεται ο πράκτορας.



Εικόνα 48 Αναγνώριση οντοτήτων από κάποιο ρομπότ

²⁴ https://en.wikipedia.org/wiki/Texas_Red

```
Give the term: kitchen
RelatedTo
['table', 3.84343069231875]
['plate', 3.3887871966006884]
['room', 2.3142259157147684]
['oven', 2.050058100388644]
['toaster', 1.5339321291066186]
['cabinet', 1.3601111111111108]
['sink', 1.143063078735083]
['refrigerator', 1.0526075421676853]
['chair', 1.0323063365454341]
['range', 0.9016666666666666]
['mixer', 0.8595000000000002]
['spatula', 0.8409545461749977]
['glass', 0.7873461834395777]
['bowl', 0.715853623852752]
['napkin', 0.6552824287007023]
['tray', 0.6373012296252065]
['house', 0.6180935291738688]
['food', 0.5765186175669085]
['kettle', 0.5749259550849192]
['servant', 0.5560989244763573]
['phone', 0.5067791127488512]
['bacteria', 0.496729517505068]
['chicken', 0.493092755091747]
['salt', 0.48871085078758036]
['sponge', 0.4722509653142333]
['tablespoon', 0.47165037144846766]
['butter', 0.4453291704475036]
['pot', 0.4431111111111096]
['steel', 0.38002621792678487]
['platter', 0.36703058437800806]
['cook', 0.3481111111111112]
['pan', 0.34390909090909105]
['cleaver', 0.3435170635431623]
```

```
IsA
['kitchenette', 2.773695303770603]
['room', 2.5692259157147683]
['galley', 2.300992775902893]
```

Εικόνα 49 Αποτελέσματα αναζήτησης για τον όρο «kitchen»

7.1.5 Αποτελέσματα για τον όρο αναζήτησης «computer»

Στα αποτελέσματα για εξόρυξη ΣΠ σχετικά με τον όρο «computer» λαμβάνουμε πέντε πεδία. Το πρώτο είναι το «RelatedTo» δηλαδή το «Σχετίζεται με». Αυτό περιλαμβάνει λέξεις όπως «apple» και «dell» από τους αντίστοιχους κατασκευαστές υπολογιστών apple και dell, «desk» που σχετίζεται με υπολογιστές γραφείου, «programmer» που είναι ο προγραμματιστής υπολογιστών, και «bit».


```

Give the term: computer
RelatedTo
['apple', 4.4676587342335115]
['desk', 4.1023084244261385]
['print', 3.5306167518166682]
['dell', 2.8263384325544036]
['worker', 2.28070929146014]
['programmer', 2.2563860678485925]
['data', 2.243805192008484]
['work', 2.066666666666667]
['memory', 1.6930121138152758]
['bit', 1.440727520437393]

```

Εικόνα 50 Αποτελέσματα αναζήτησης για τον όρο «computer»

Το δεύτερο πεδίο που λαμβάνουμε είναι το «HasContext» που ορίζει ένα σύνολο λέξεων που αφορούν τον υπολογιστή (computer) ως πλαίσιο. Σε αυτές συμπεριλαμβάνονται λέξεις όπως «computer science», «visual display unit», «data structure», «digital communication», «interconnection» και άλλες που φαίνονται αναλυτικά με το βάρος Σηματολογικής Ομοιότητας στην εικόνα 51.

```

['computer_science', 2.388585416154097]
['visual_display_unit', 2.362152854647076]
['plotter', 2.3180797637593544]
['throughput', 2.307052117388072]
['data_structure', 2.2759211874008827]
['scratchpad', 2.25]
['printout', 2.2481590324874166]
['digital_communication', 2.2452939104858833]
['interconnection', 2.220206179719311]
['alpha_test', 2.211219617805458]
['beta_test', 2.211219617805458]
['read_out', 2.2036276357129454]
['faceplate', 2.125]
['console', 2.125]
['back_up', 2.0858820620344583]
['slot', 2.076923076923077]
['outage', 2.076923076923077]
['module', 2.0714285714285716]
['format', 2.0714285714285716]
['pass', 2.0526315789473686]
['up', 2.000999000999001]
['incompatible', 2.000999000999001]
['compatible', 2.000999000999001]

```

Εικόνα 51 Αποτελέσματα αναζήτησης για τον όρο «computer»

Το τρίτο πεδίο είναι το «UsedFor» που περιλαμβάνει ένα σύνολο λέξεων τα οποία σχετίζονται με το πώς χρησιμοποιείται ο υπολογιστής. Πιο συγκεκριμένα παίρνουμε την πληροφορία ότι ο υπολογιστής χρησιμοποιείται για υπολογισμούς και για εκτέλεση διεργασιών (process).

```
UsedFor
['calculate', 2.242189755130076]
['work', 2.066666666666667]
['cybernate', 2.0351067297398737]
['process', 2.000999000999001]
```

Εικόνα 52 Αποτελέσματα αναζήτησης για τον όρο «computer», ιδιότητα «UsedFor»

Στο τέταρτο πεδίο «PartOf» είναι ένα σύνολο που περιέχει λέξεις οι οποίες αποτελούν λειτουργικό συστατικό ενός υπολογιστή, η αποτελεί ο υπολογιστής συστατικό αυτών.

```
PartOf
['central_processing_unit', 2.6003938550692363]
['peripheral', 2.5968172645069396]
['computer_circuit', 2.4122705874997545]
['busbar', 2.362101167272172]
['memory', 2.358012113815276]
['cathode_ray_tube', 2.2885305633786173]
['diskette', 2.27639428660938]
['keyboard', 2.25]
['computer_accessory', 2.2219533266786726]
['data_converter', 2.2]
['hardware', 2.1666666666666665]
['disk_cache', 2.142857142857143]
['platform', 2.125]
['monitor', 2.1]
['chip', 2.0833333333333335]
```

Εικόνα 53 Αποτελέσματα αναζήτησης για τον όρο «computer», ιδιότητα «PartOf»

Είναι γεγονός πως ο υπολογιστής είναι μια κεντρική μονάδα επεξεργασίας, ή είναι μέρος ενός κυκλώματος υπολογιστών. Ενώ επίσης είναι γνωστό πως αποτελείται από chips και διάφορα άλλα είδη hardware, και επομένως λέξεις όπως οι «central processing unit», «computer circuit», «memory», «chip», «hardware» και άλλες που φαίνονται αναλυτικά στην εικόνα 53, δεν απουσιάζουν από το σύνολο «PartOf».

Τέλος έχουμε και ένα πέμπτο πεδίο το «IsA», το οποίο συναντήθηκε και σε παραδείγματα που αναλύσαμε προηγουμένως. Η εικόνα 54 δείχνει αναλυτικά τα περιεχόμενα του.

```
IsA
['digital_computer', 2.85582879741693]
['analog_computer', 2.8385898034667223]
['home_computer', 2.826992629449951]
['turing_machine', 2.7181734881064323]
['machine', 2.6988874289788227]
['web_site', 2.6840781607656723]
['predictor', 2.2913079017424094]
['pari_mutuel_machine', 2.1486302449105557]
['number_cruncher', 2.111111111111111]
['server', 2.1]
['node', 2.0714285714285716]
```

Εικόνα 54 Αποτελέσματα αναζήτησης για τον όρο «computer», ιδιότητα «IsA»

ΚΕΦΑΛΑΙΟ 8: ΕΠΙΛΟΓΟΣ

8 ΕΠΙΛΟΓΟΣ

Από όσα εκθέσαμε παραπάνω, συνάγεται το συμπέρασμα ότι οι μηχανές δεν είναι ακόμα σε θέση να κατανοήσουν με βεβαιότητα το περιεχόμενο του Παγκόσμιου Ιστού. Αξιολογώντας τις παραπάνω απόψεις γίνεται, λοιπόν, αντιληπτό ότι για την αντιμετώπιση του προβλήματος αυτού απαιτείται η ένταξη της Σημασιολογικής Πληροφορίας στον σημερινό Ιστό.

Αρχικά καταλήξαμε στο συμπέρασμα ότι τα παραδοσιακά Συστήματα Ανάκτησης Πληροφορίας που υπάρχουν για την δημιουργία καταλόγων του διαδικτύου και την δημιουργία crawler δεν επαρκούν για τη δημιουργία και την εξαγωγή της Σημασιολογικής Πληροφορίας στον χώρο του διαδικτύου, κάτι το οποίο είναι και λογικό αφού δεν έχουν σχεδιαστεί για αυτόν τον σκοπό.

Επίσης έγινε κατανοητό πως τα εργαλεία και οι τεχνολογίες που χρειάζονται για την δημιουργία Σημασιολογικής Πληροφορίας στις ιστοσελίδες του διαδικτύου υπάρχουν ήδη, και το μόνο που χρειάζεται είναι να συνδυαστούν κατάλληλα. Αυτά τα εργαλεία είναι το RDF, η SPARQL, η OWL2, ενώ υπάρχουν διάφορες οντολογίες που θα παίζουν καθοριστικό ρόλο στη μετάβαση από τον Κοινωνικό Ιστό στον Σημασιολογικό Ιστό, όπως η DBpedia, ο ιστότοπος ConceptNet, το λεξικό WordNet.

Η εργασία αυτή, αποσκοπεί να αποτελέσει ένα ακόμα εργαλείο που ίσως έχει να προσφέρει κάτι σε όλη αυτή την εξέλιξη του διαδικτύου, και αυτό το «κάτι» είναι η Σημασιολογική Πληροφορία που εξάγεται από την κατάλληλη χρήση των εργαλείων που αναφέρθηκαν παραπάνω και την ερμηνεία της πληροφορίας που παρέχουν αυτά. Τέλος αξίζει να θυμηθούμε ότι η εξέλιξη του διαδικτύου στηρίχτηκε και στηρίζεται σε τέτοιες συνεισφορές που γίνονται από την παγκόσμια κοινότητα της πληροφορικής. Επομένως στην ίδια φιλοσοφία είναι δυνατόν να στηριχθούμε και για την εξέλιξη του Σημασιολογικού Ιστού.

Το εργαλείο θα μπορούσε να χρησιμοποιηθεί και να δοκιμαστεί στο Quora/Kaggle Challenge “Duplicate question pairs”²⁵. Καθώς το συγκεκριμένο εργαλείο με την χρήση της Σημασιολογικής Πληροφορίας μπορεί να συγκρίνει δύο ερωτήσεις και να καταλάβει αν εκφράζουν το ίδιο πράγμα. Υπήρξαν αρκετές περιπτώσεις που μία μικρή αλλαγή λέξης σε δύο προτάσεις τις κάνει να έχουν εντελώς διαφορετικό νόημα. Για παράδειγμα «Ποια είναι η τιμή αγοράς της ελιάς στην Ελλάδα;» και «Ποια είναι η τιμή αγοράς της ελιάς στην Γερμανία;», οι δύο ερωτήσεις μοιάζουν αρκετά, όμως είναι σαφές ότι οι απαντήσεις τους είναι δύο τελείως ανεξάρτητα γεγονότα. Ενώ υπήρχαν προτάσεις που δεν υπήρξε ούτε μια κοινή λέξη αλλά είχαν την ίδια απάντηση. Για παράδειγμα «Πότε κυκλοφόρησε η ταινία “Το Χόμπιτ”;» και «Ποια μέρα ήταν η προβολή στο σινεμά του πρώτου βιβλίου του Τόλκιν;». Είναι φανερό ότι με την χρήση εργαλείων που υποστηρίζουν τα παραδοσιακά ΣΑΠ θα έβγαινε λανθασμένο συμπέρασμα ότι το πρώτο ζευγάρι είναι πολύ πιθανόν να

²⁵ <https://www.kaggle.com/c/quora-question-pairs>

είναι η ίδια ερώτηση, ενώ στην περίπτωση του δεύτερου ζεύγους μάλλον οι ερωτήσεις μας δεν μοιάζουν και πολύ. Με την χρήση του εργαλείου που αναπτύχθηκε όμως θα είχαμε μια πιο ισχυρή γνώση για να συγκρίνουμε το κάθε ζεύγος ερωτήσεων, δηλαδή την Σημασιολογική Πληροφορία.

Επίσης ένα άλλο σενάριο χρήσης του εργαλείου που αναπτύχθηκε θα μπορούσε να είναι η υποστήριξη Βάσεων Δεδομένων που αποτελούνται από ετερογενή γραφήματα και υποστηρίζουν noSQL τύπους ερωτημάτων. Ένα παράδειγμα αυτής της περίπτωσης είναι το Neo4j. Το Neo4j υποστηρίζει την ανάπτυξη τέτοιων γραφημάτων, επομένως θα ήταν πολύ πρακτικό να δημιουργηθεί ένα δίκτυο Σημασιολογικής Πληροφορίας που θα αποτελείται από ένα μεγάλο πλήθος RDF τριπλετών που συνδέονται με πολλούς διαφορετικούς τρόπους. Σε ένα τέτοιο γράφημα η απόσταση δύο Κόμβων θα μπορούσε να ερμηνευτεί ως μια μετρική Σημασιολογικής «απόστασης».

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] LEHMANN, Fritz. *Semantic networks in artificial intelligence*. Elsevier Science Inc., 1992.
- [2] Simmons, Robert F. *Synthetic language behavior*. System Development Corporation, 1963.
- [3] Helbig, Hermann. *Die semantische Struktur natürlicher Sprache: Wissensrepräsentation mit MultiNet*. Springer-Verlag, 2013.
- [4] Bendeck, Fawsy. *WSM-P Workflow Semantic Matching Platform*. Verlag Dr. Hut, 2008.
- [5] Bizer, Christian, Tom Heath, and Tim Berners-Lee. "Linked data: Principles and state of the art." *World wide web conference*. Vol. 1. 2008.
- [6] Singh, Push, et al. "Open mind common sense: Knowledge acquisition from the general public." *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, Berlin, Heidelberg, 2002.
- [7] Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web." *Scientific american* 284.5 (2001): 34-43.
- [8] Antoniou, Grigoris, and Frank Van Harmelen. *A semantic web primer*. MIT press, 2004.
- [9] Maedche, Alexander, and Steffen Staab. "Ontology learning for the semantic web." *IEEE Intelligent systems* 16.2 (2001): 72-79.
- [10] Broekstra, Jeen, Arjohn Kampman, and Frank Van Harmelen. "Sesame: A generic architecture for storing and querying rdf and rdf schema." *International semantic web conference*. Springer, Berlin, Heidelberg, 2002.
- [11] Παπαδόπουλος, Απόστολος, Ιωάννης Μανωλόπουλος, and Κωνσταντίνος Τσίχλας. "Εισαγωγή στην Ανάκτηση Πληροφορίας." (2015).
- [12] Παπαδόπουλος, Απόστολος, Ιωάννης Μανωλόπουλος, and Κωνσταντίνος Τσίχλας. "Αποτίμηση Αποτελεσματικότητας." (2015).
- [13] Quilitz, Bastian, and Ulf Leser. "Querying distributed RDF data sources with SPARQL." *European semantic web conference*. Springer, Berlin, Heidelberg, 2008.
- [14] Sowa, John F. "Semantic networks." (1987).
- [15] Lehmann, Fritz. *Semantic networks in artificial intelligence*. Elsevier Science Inc., 1992.
- [16] Bizer, Christian, Tom Heath, and Tim Berners-Lee. "Linked data: The story so far." *Semantic services, interoperability and web applications: emerging concepts*. IGI Global, 2011. 205-227.

- [17] Bauer, Florian, and Martin Kaltenböck. "Linked open data: The essentials." *Edition mono/monochrom*, Vienna 710 (2011).
- [18] Maedche, Alexander, and Steffen Staab. "Ontology learning for the semantic web." *IEEE Intelligent systems* 16.2 (2001): 72-79.
- [19] Antoniou, Grigoris, Enrico Franconi, and Frank Van Harmelen. "Introduction to semantic web ontology languages." *Reasoning web*. Springer, Berlin, Heidelberg, 2005. 1-21.
- [20] Hitzler, Pascal, et al. "OWL 2 web ontology language primer." *W3C recommendation* 27.1 (2009): 123.
- [21] Singh, Push, et al. "Open mind common sense: Knowledge acquisition from the general public." *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, Berlin, Heidelberg, 2002.
- [22] Liu, Hugo, and Push Singh. "ConceptNet—a practical commonsense reasoning tool-kit." *BT technology journal* 22.4 (2004): 211-226.
- [23] Bizer, Christian, et al. "DBpedia-A crystallization point for the Web of Data." *Journal of web semantics* 7.3 (2009): 154-165.
- [24] Auer, Sören, et al. "Dbpedia: A nucleus for a web of open data." *The semantic web*. Springer, Berlin, Heidelberg, 2007. 722-735.
- [25] Miller, George A., et al. "Introduction to WordNet: An on-line lexical database." *International journal of lexicography* 3.4 (1990): 235-244.
- [26] Li, Baoli, and Liping Han. "Distance weighted cosine similarity measure for text classification." *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, Berlin, Heidelberg, 2013.
- [27] Seymour, Tom, Dean Frantsvog, and Satheesh Kumar. "History of search engines." *International Journal of Management & Information Systems (IJMIS)* 15.4 (2011): 47-58.
- [28] Heydon, Allan, and Marc Najork. "Mercator: A scalable, extensible web crawler." *World Wide Web* 2.4 (1999): 219-229.
- [29] Young, Jay, et al. "Towards lifelong object learning by integrating situated robot perception and semantic web mining." *22nd European Conference on Artificial Intelligence, ECAI 2016*. Vol. 285. IOS Press, 2016.
- [30] Chernova, Sonia, et al. "Situated bayesian reasoning framework for robots operating in diverse everyday environments." *Robotics Research*. Springer, Cham, 2020. 353-369.
- [31] Icarte, Rodrigo Toro, et al. "How a general-purpose commonsense ontology can improve performance of learning-based image retrieval." *arXiv preprint arXiv:1705.08844* (2017).
- [32] Wang, Peng, et al. "Fvqa: Fact-based visual question answering." *IEEE transactions on pattern analysis and machine intelligence* 40.10 (2018): 2413-2427.
- [33] Presutti, Valentina, et al. "Extracting core knowledge from linked data." *Proceedings of the Second International Conference on Consuming Linked Data-Volume 782*. CEUR-WS. org, 2011.

- [34] Mendes, Pablo N., Max Jakob, and Christian Bizer. "DBpedia: A Multilingual Cross-domain Knowledge Base." *LREC*. 2012.
- [35] Bizer, Christian, et al. "DBpedia-A crystallization point for the Web of Data." *Journal of web semantics* 7.3 (2009): 154-165.
- [36] Speer, Robyn, Joshua Chin, and Catherine Havasi. "Conceptnet 5.5: An open multilingual graph of general knowledge." *arXiv preprint arXiv:1612.03975* (2016).