



ΜΕΤΑΠΤΥΧΙΑΚΟ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΙΟΝΙΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ
ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΙΣ ΑΝΘΡΩΠΙΣΤΙΚΕΣ ΚΑΙ
ΚΟΙΝΩΝΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παραγωγή πρωτότυπων σημασιολογικών δεδομένων βιβλιογραφικού
περιεχομένου σε μορφή RDF και ανάπτυξη εργαλείων χρήσης τους

Πουλημένου Χριστίνα
Α.Μ. ΠΜ201711

Επιβλέπων Καθηγητής

Μιχαήλ Στεφανιδάκης, Επίκουρος Καθηγητής, Ιόνιο Πανεπιστήμιο

Τριμελής Επιτροπή

Μιχαήλ Στεφανιδάκης, Επίκουρος Καθηγητής, Ιόνιο Πανεπιστήμιο
Θεόδωρος Ανδρόνικος, Αναπληρωτής Καθηγητής, Ιόνιο Πανεπιστήμιο
Δημήτριος Τσουμάκος, Αναπληρωτής Καθηγητής, Ιόνιο Πανεπιστήμιο

Φεβρουάριος 2019

Contents

1	Εισαγωγή	7
1.1	Αντικείμενο της παρούσης εργασίας	8
1.2	Διάρθρωση της παρούσης εργασίας	8
2	Θεωρητικό υπόβαθρο - Βιβλιογραφική ανασκόπηση	10
2.1	Σημασιολογικός ιστός	10
2.2	Δεδομένα και μεταδεδομένα	12
2.3	Ανοιχτά Συνδεδεμένα Δεδομένα	13
2.4	Τεχνολογίες του σημασιολογικού ιστού	13
2.4.1	URIs	13
2.4.2	Οντολογίες-Λεξιλόγια	14
2.4.3	Το μοντέλο RDF	15
2.4.4	Αποθετήρια τριάδων (Triplestores)	16
2.4.5	Γλώσσες ερωτημάτων σε RDF δεδομένα	19
3	Ανάλυση Προβλήματος	21
3.1	Διατύπωση προβλήματος	21
3.2	Κατασκευή οντολογίας	21
3.3	Λεξιλόγια που χρησιμοποιήθηκαν	25
3.3.1	RDF - RDFs	25
3.3.2	Dublin Core	28
3.3.3	Bibliographic Ontology	28
3.3.4	FOAF	29
4	Περιγραφή Υλοποίησης	31
4.1	Εργαλεία που χρησιμοποιήθηκαν	31
4.1.1	Python3	31
4.1.2	Open Link Virtuoso	32
4.1.3	SPARQL	33
4.2	Αναζήτηση και συλλογή των βιβλιογραφικών δεδομένων σε μορφή bibtex.	34
4.3	Απόδοση URIs στις οντότητες.	34
4.4	Μετατροπή των bibtex δεδομένων σε RDF μέσω Python.	35
4.5	Έλεγχος εγκυρότητας του RDF αρχείου.	36
4.6	Φόρτωση του αρχείου στο Open Link Virtuoso και διατύπωση ερωτημάτων με τη γλώσσα SPARQL.	37
5	Συμπεράσματα και μελλοντικές εργασίες	42
	Βιβλιογραφία	43

List of Figures

1	Η στοίβα του σημασιολογικού ιστού (Semantic Web Stack).	11
2	Αναπαράσταση μιας τριάδας RDF	15
3	Αναπαράσταση τριάδων RDF	16
4	Η Αρχιτεκτονική μιας Triplestore.	17
5	Δείγμα βιβλιογραφικής εγγραφής σε μορφή bibtex.	22
6	Σχηματική αναπαράσταση της οντολογίας.	23
7	Στιγμιότυπα των κλάσεων 'Publication' και 'Author'.	24
8	Η οντολογία με τη χρήση όρων γνωστών λεξιλογίων.	26
9	Μια εγγραφή σε μορφή Bibtex.	35
10	Οι τριάδες που προκύπτουν από το Bibtex μιας δημοσίευσης.	36
11	Η καρτέλα Quad Store Upload του Virtuoso.	37
12	Η καρτέλα SPARQL του Virtuoso.	37
13	Τα αποτελέσματα ενός ερωτήματος στον Virtuoso.	38
14	Αποτελέσματα ερωτήματος 1.	39
15	Αποτελέσματα ερωτήματος 2.	40
16	Αποτελέσματα ερωτήματος 3.	41

List of Tables

1	Δομές μοντελοποίησης της γλώσσας RDFS	18
2	Τα χαρακτηριστικά της οντότητας Publication	22
3	Τα χαρακτηριστικά της οντότητας Author	23
4	Οι ιδιότητες της κλάσης Publication	26
5	Οι ιδιότητες της κλάσης Author	26
6	Κλάσεις που χρησιμοποιούν τα λεξιλόγια της RDF	27
7	Ιδιότητες που χρησιμοποιούν τα λεξιλόγια RDF	27

Περίληψη

Οι τελευταίες εξελίξεις στις τεχνολογίες του σημασιολογικού ιστού έχουν σαν αποτέλεσμα να περνάμε από μια παραδοσιακή παρουσίαση των εγγράφων στον ιστό, σε μια άλλη μορφή με αυξημένη σημασιολογία. Αυτό δεν θα μπορούσε να αφήσει ανεπηρέαστο τον χώρο των βιβλιογραφικών δεδομένων. Καθώς οι επιστημονικές δημοσιεύσεις είναι στοιχεία καθοριστικής σημασίας για την πρόοδο του επιστημονικού έργου, οι τεχνικές δημοσίευσης που προσφέρουν διασύνδεση με πολλές πηγές δεδομένων, ολοκλήρωση και ενοποίηση των δεδομένων από διαφορετικές πηγές και σημασιολογική περιγραφή των βιβλιογραφικών εγγράφων, είναι πλέον απαραίτητες και προκαλούν το ενδιαφέρον πολλών ερευνητών που κινούνται στο χώρο του σημασιολογικού ιστού. Τέτοιες δυνατότητες προσφέρει το μοντέλο δεδομένων RDF, τα διάφορα λεξιλόγια-οντολογίες που αναπαριστούν το πεδίο των βιβλιογραφικών εγγράφων καθώς και διάφορα εργαλεία δημοσίευσης και διαχείρισης διασυνδεδεμένων δεδομένων.

Στην παρούσα εργασία επιχειρείται να εφαρμοσθούν οι τεχνολογίες του σημασιολογικού ιστού στο χώρο των επιστημονικών δημοσιεύσεων των καθηγητών του τμήματος πληροφορικής του Ιονίου Πανεπιστημίου. Τα εργαλεία που χρησιμοποιούνται σ' αυτήν την προσπάθεια είναι κάποια από τα υπάρχοντα λεξιλόγια για την κατασκευή της οντολογίας, η γλώσσα προγραμματισμού Python 3 για την μετατροπή των δεδομένων σε RDF και ο Virtuoso Universal Server για τη διαχείριση των δεδομένων.

Abstract

The latest developments in semantic web technologies have resulted in transition from a traditional presentation of documents in the web, to another form with enhanced semantics. This could not leave the field of the bibliographic data unaffected. As scientific publications are key to the advancement of scientific work, publishing techniques that offer interlinking with many data sources, integration of data from different sources, and semantic description of bibliographic records, are now necessary and attract the interest of many researchers in the field of the semantic web. Such features include the RDF data model, various ontology vocabularies that represent the field of bibliographic records, and various tools for publishing and managing linked data.

This thesis attempts to apply the semantic web technologies in the field of scientific publications of the professors of the Ionian University Department of Informatics. The tools used in this effort are some of the existing and well known vocabularies for building the ontology, the Python 3 programming language to convert the available data to RDF model and Virtuoso Universal Server for data management.

1 Εισαγωγή

Η δυνατότητα αποτελεσματικής πρόσβασης στα αποτελέσματα των επιστημονικών ερευνών είναι θέμα ζωτικής σημασίας για την αποτελεσματικότητα των μελλοντικών επιστημονικών ερευνών σε όλη τη διάρκεια του κύκλου ζωής τους, από την διατύπωση υποθέσεων και το πείραμα μέχρι την παρουσίαση του πονήματος και τη δημοσίευσή του, καθώς τα ευρήματα του ενός επιστημονικού εγχειρήματος θέτουν τις βάσεις για την ανάπτυξη του επόμενου. Το κύριο μέσο για τη διάδοση της επιστημονικής γνώσης είναι τα επιστημονικά άρθρα που δημοσιεύονται και τα οποία είναι δομικά στοιχεία στη ροή εργασιών της επιστημονικής κοινότητας. [1]

Τις τελευταίες δεκαετίες η αλματώδης ανάπτυξη του Παγκόσμιου Ιστού έφερε μια επανάσταση στον τρόπο διάδοσης της γνώσης και ο κόσμος των επιστημονικών δημοσιεύσεων δεν θα μπορούσε να μείνει ανεπηρέαστος. Από την αναζήτηση των δημοσιεύσεων, είτε με φυσική παρουσία στις βιβλιοθήκες είτε με άλλους παραδοσιακούς τρόπους, περάσαμε στην αναζήτηση της ψηφιακής μορφής τους μέσα από τις σελίδες του ιστού με ότι αυτό συνεπάγεται σε χρόνο, χρήμα και ευκολία χρήσης.

Μέχρι τώρα οι ηλεκτρονικές εκδόσεις των επιστημονικών άρθρων παρουσιάζονται συμβατικά είτε σαν ιστοσελίδες σε μορφή HTML είτε σαν έγγραφα PDF, με τον ιστό να χρησιμοποιείται κυρίως ως βολικό μέσο διανομής για το παραδοσιακό κείμενο. Αυτός ο τρόπος παρουσίασης είναι φιλικός προς τον άνθρωπο όσον αφορά το διάβασμα του εγγράφου αλλά προσφέρει πολύ μικρές δυνατότητες αλληλεπίδρασης με το χρήστη και σχεδόν καμία δυνατότητα επεξεργασίας από υπολογιστή εμποδίζοντας έτσι την ανάπτυξη εφαρμογών που θα μπορούσαν να συνδέσουν αυτόματα τις πληροφορίες μεταξύ των διαφόρων άρθρων. [2]

Αυτό έρχεται σε αντίθεση με το όραμα του δημιουργού του παγκόσμιου ιστού, Tim Berners Lee, για έναν "ιστό των δεδομένων" (web of data) ή αλλιώς έναν "σημασιολογικό ιστό" (semantic web) δομημένο με τέτοιο τρόπο ώστε οι υπολογιστές να μπορούν να βοηθούν τον άνθρωπο στο όλο και διογκούμενο έργο της αναζήτησης της πληροφορίας, της ενοποίησης των δεδομένων από κατανεμημένες πηγές και της αναπαράστασης της γνώσης. [1]

Έτσι ένα μεγάλο κομμάτι της έρευνας έχει στραφεί στην ανάπτυξη των τεχνολογιών του σημασιολογικού ιστού. Η περιγραφή του πεδίου που αφορά τους ερευνητές και τις ερευνητικές κοινότητες με τις δημοσιεύσεις τους, τις δραστηριότητές τους και τις αμοιβαίες σχέσεις τους είναι από τις πρώτες εφαρμογές που υλοποιήθηκαν με τη διαχείριση κατανεμημένων, διασυνδεδεμένων και καλά δομημένων πληροφοριών στο πνεύμα του σημασιολογικού ιστού. [3]

Οι τελευταίες εξελίξεις σ' αυτόν τον τομέα μπορούν να χρησιμοποιηθούν για την σημασιολογική βελτίωση των επιστημονικών άρθρων επιτρέποντας καλύτερη διασύνδεση με άλλες πηγές, προσθέτοντας περιγραφικά μεταδεδομένα τα οποία επιτρέπουν την καλύτερη εξερεύνηση του άρθρου, προσδιορίζοντας τις έννοιες και τους όρους του άρθρου και επιτρέποντας στον χρήστη να έχει πρόσβαση στο περιεχόμενο με διαδραστικό τρόπο, σε επαναταξινομούμενες λίστες αναφορών, σε περιλήψεις εγγράφων και σε σύνολα αριθμητικών δεδομένων που μπορούν να αποθηκευτούν και να είναι προσβάσιμα και εύκολα στην επεξεργασία. [2]

Διάφορες τεχνολογίες του σημασιολογικού ιστού όπως το μοντέλο RDF και τα λεξιλόγια των οντολογιών που περιγράφουν συγκεκριμένα πεδία γνώσης καθώς και εργαλεία δημοσίευσης και διαχείρισης συνδεδεμένων δεδομένων υποστηρίζουν την προσπάθεια περιγραφής των βιβλιογραφικών δεδομένων σαν σημασιολογικά βιβλιογραφικά δεδομένα.

1.1 Αντικείμενο της παρούσης εργασίας

Οι τεχνικές διαχείρισης της πληροφορίας στο σύγχρονο παγκόσμιο ιστό περιλαμβάνουν πρότυπα όπως οι γλώσσες RDF και OWL 2 για να κωδικοποιήσουν την πληροφορία σε τέτοια μορφή ώστε να μπορούν οι εφαρμογές να απευθύνουν ερωτήματα στα μεταδεδομένα και να ενοποιούν τις πληροφορίες που προέρχονται από διάφορες κατακευκτωμένες πηγές, με αυτοματοποιημένο τρόπο. Από τη στιγμή που η επικοινωνία μεταξύ των διαφόρων επιστημονικών πηγών είναι πρωταρχικής σημασίας στις επιστημονικές πρακτικές, είναι βασικό, οι εκδότες των άρθρων να υιοθετούν αυτά τα πρότυπα ώστε να μπορούν οι αναγνώστες τους να οδηγούνται εύκολα στην ανακάλυψη της γνώσης που εκπροσωπείται στα περιοδικά, τα βιβλία και τα συνέδρια. Αυτό απαιτεί να υπάρχουν διαθέσιμες οι κατάλληλες οντολογίες που ικανοποιούν τις απαιτήσεις των συγγραφέων, των εκδοτών και των αναγνωστών, εκφρασμένες με τις κατάλληλες γλώσσες. [4]

Ο σκοπός αυτής της εργασίας είναι να προτείνει ένα τρόπο να εφαρμοσθούν οι τεχνολογίες του σημασιολογικού ιστού στο χώρο των δημοσιεύσεων των καθηγητών του τμήματος πληροφορικής του Ιονίου Πανεπιστημίου. Η βασική ιδέα ήταν αυτό να γίνει σε δύο στάδια. Στο πρώτο έπρεπε αφού μελετηθεί το συγκεκριμένο γνωστικό πεδίο και οι απαιτήσεις του, να κατασκευαστεί μια οντολογία που να το αντιπροσωπεύει. Στο δεύτερο έπρεπε να υλοποιηθεί αυτή η οντολογία σε γλώσσα RDF και να διατυπωθούν κάποια ερωτήματα πάνω στα δεδομένα, για να δειχθεί ο εύκολος και αποτελεσματικός τρόπος χρήσης τους.

1.2 Διάρθρωση της παρούσης εργασίας

Η διάρθρωση της εργασίας έχει ως εξής:

Στην ενότητα 1 γίνεται μια εισαγωγή στο γνωστικό πεδίο με το οποίο ασχολείται η εργασία.

Στην ενότητα 2 γίνεται μια βιβλιογραφική ανασκόπηση του θέματος. Συγκεκριμένα γίνεται αναφορά στον Σημασιολογικό Ιστό, στην έννοια των μεταδεδομένων και των ανοιχτών συνδεδεμένων δεδομένων, στις τεχνολογίες του σημασιολογικού ιστού, όπως στην απόδοση URIs στις οντότητες, στις οντολογίες και την αναπαράστασή τους με λεξιλόγια, και στο μοντέλο RDF, γίνεται μια περιγραφή του τι είναι ένα αποθετήριο τριάδων RDF, και τέλος μια αναφορά στις γλώσσες δόμησης των ερωτημάτων που απευθύνονται σε δεδομένα RDF.

Στην ενότητα 3 γίνεται η ανάλυση του προβλήματος. Αφού διατυπωθεί το πρόβλημα περιγράφεται λεπτομερώς ο τρόπος με τον οποίο κατασκευάστηκε η οντολογία καθώς και τα λεξιλόγια που χρησιμοποιήθηκαν για να περιγράψουν τις οντότητες, τις μεταξύ τους σχέσεις και τις ιδιότητές τους.

Η ενότητα 4 περιλαμβάνει την υλοποίηση της οντολογίας που αντιπροσωπεύει τον χώρο των βιβλιογραφικών δεδομένων που αφορούν την εργασία. Κατ' αρχήν περιγράφονται τα εργαλεία που χρησιμοποιήθηκαν και συγκεκριμένα η γλώσσα προγραμματισμού Python 3 που χρησιμοποιήθηκε για την μετατροπή των βιβλιογραφικών δεδομένων από μια συμβατική μορφή σε RDF, ο Virtuoso Universal Server που χρησιμεύει στη δημοσίευση δεδομένων και η γλώσσα SPARQL που είναι μια γλώσσα ερωτημάτων πάνω σε δεδομένα RDF. Στη συνέχεια περιγράφεται το πως έγιναν η αναζήτηση και η συλλογή των βιβλιογραφικών δεδομένων σε μορφή Bibtext, η απόδοση URIs στις οντότητες, η μετατροπή των δεδομένων σε RDF και σε μορφή σειριοποίησης N-Triples, ο έλεγχος εγκυρότητας του RDF αρχείου που δημιουργήθηκε, η φόρτωση των δεδομένων στον Virtuoso και η διατύπωση κάποιων ερωτημάτων μέσω του SPARQL endpoint του Virtuoso.

Τέλος στην ενότητα 5 διατυπώνονται κάποια συμπεράσματα και γίνονται προτάσεις για μελλοντικές εργασίες.

2 Θεωρητικό υπόβαθρο - Βιβλιογραφική ανασκόπηση

2.1 Σημασιολογικός ιστός

Μέχρι πρόσφατα ο Παγκόσμιος Ιστός έδινε αρκετά αποτελεσματικά την ευκαιρία στο χρήστη είτε να δημοσιεύει είτε να αποκτά πρόσβαση σε έγγραφα και με την απεριόριστη δυνατότητα σύνδεσης αυτών των εγγράφων με οποιοδήποτε άλλο έγγραφο του δικτύου να αποκτά ένα μεγάλο όγκο πληροφορίας. Το μειονέκτημα ήταν ότι αυτή η πληροφορία δεν μπορούσε να είναι διαχειρίσιμη από τον υπολογιστή απ'ευθείας σαν δεδομένα παρά μόνο μετά από αρκετή δουλειά που έπρεπε να γίνει από το χρήστη [5].

Αυτό ακριβώς φιλοδοξεί να διορθώσει η εξέλιξη του Παγκόσμιου Ιστού, που ονομάζουμε Σημασιολογικό Ιστό, έτσι ώστε να μπορούμε να τον αντιμετωπίζουμε μάλλον σαν μια τεράστια βάση από δεδομένα παρά απλά σαν ένα σύνολο διασυνδεδεμένων εγγράφων. Έτσι με τα δεδομένα να έχουν αποκτήσει ένα καλά ορισμένο νόημα οι υπολογιστές θα μπορούν όχι μόνο να τα διαβάσουν αλλά και να τα επεξεργαστούν με αποτέλεσμα την όλο και περισσότερο αυτοματοποιημένη λήψη αποφάσεων και την καλύτερη συνεργασία ανθρώπου-υπολογιστή [6].

Ο Σημασιολογικός Ιστός αποτελείται από στρώματα (layers) κάθε ένα από τα οποία υλοποιεί μια λειτουργικότητα (functionality) και τα οποία αποτελούν μια στοίβα, τη λεγόμενη στοίβα του σημασιολογικού ιστού (Semantic Web Stack). Η λειτουργικότητα των επιπέδων επεκτείνεται όσο ανεβαίνουμε από τα χαμηλότερα προς τα ανώτερα επίπεδα. Έτσι λοιπόν στα χαμηλά επίπεδα υλοποιούνται λειτουργίες οι οποίες είναι πολύ κοντά στον Παγκόσμιο Ιστό και στις μηχανές ενώ όσο πηγαίνουμε προς τα ανώτερα επίπεδα υλοποιούνται λειτουργικότητες αναπαράστασης γνώσης, πολύπλοκης συλλογιστικής και εμπιστοσύνης που πλησιάζουν την ανθρώπινη σκέψη. Παρακάτω περιγράφονται σύντομα τα επίπεδα του σημασιολογικού ιστού που φαίνονται στην εικόνα Figure 1:[7]

- Unicode (Χαρακτήρες)

Ο Σημασιολογικός Ιστός βασίζεται στο πρότυπο Unicode, που είναι ένα πλήρες σύνολο χαρακτήρων διεθνώς αποδεκτό για την αντιστοίχιση των δυαδικών ψηφίων σε χαρακτήρες κειμένου.

- URIs (Αναγνωριστικά)

Ο Σημασιολογικός Ιστός χρησιμοποιεί μοναδικά αναγνωριστικά ταυτοποίησης πόρων. Για τον λόγο αυτό χρησιμοποιούνται τα Uniform Resource Identifiers (URIs), τα οποία χρησιμοποιούνται ήδη στο διαδίκτυο για τον εντοπισμό των εγγράφων.

- XML , XML Namespaces, XML Query, XML Schema

Σε αυτό το επίπεδο βρίσκουμε τη γλώσσα XML που είναι μια γλώσσα σήμανσης (Markup Language) και τις τεχνολογίες που τη συνοδεύουν.

- RDF model

Σε αυτό το επίπεδο βρίσκουμε τα μεταδεδομένα ορισμένα σε RDF.

- Ontology

Στη συνέχεια συναντάμε τις γλώσσες οντολογιών που περιγράφουν τα μεταδεδομένα, όπως την OWL και την RDF Schema.

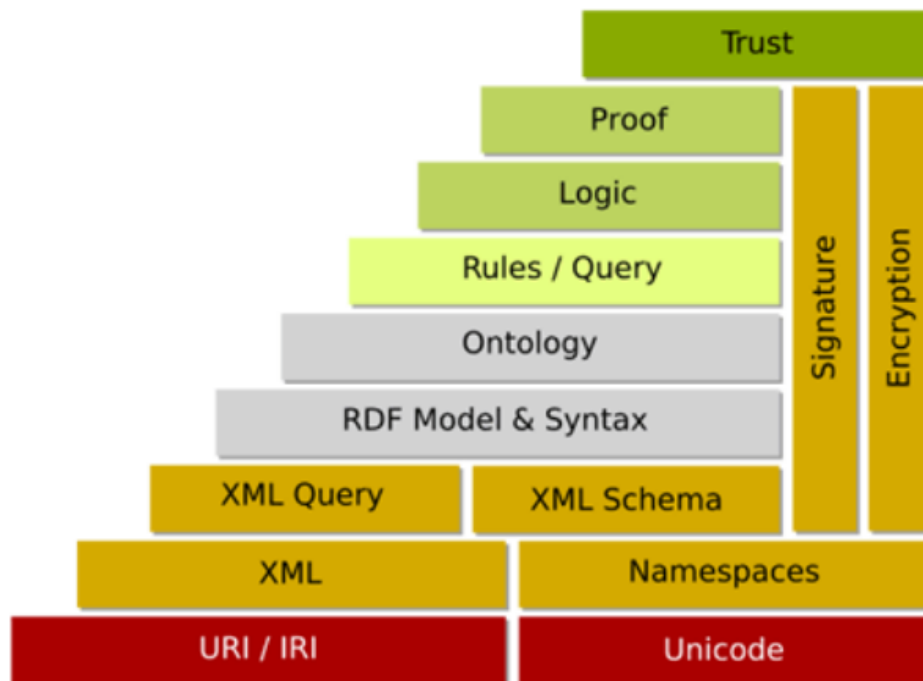


Figure 1: Η στοίβα του σημασιολογικού ιστού (Semantic Web Stack).

- Κανόνες και ερωτήματα

Για την παροχή κανόνων εκτός από τις δομές που διατίθενται από τις γλώσσες του Σημασιολογικού Ιστού, δομήθηκαν και γλώσσες κανόνων. Επίσης, για την ανάκτηση των δεδομένων δομήθηκαν πρότυπα ερωτημάτων όπως η γλώσσα SPARQL.

- Λογική

Σε αυτό το επίπεδο χρησιμοποιούνται οι τεχνολογίες των από κάτω επιπέδων έτσι ώστε να συνδυαστούν τα ερωτήματα και οι κανόνες και να έχουν μια ενιαία δράση πάνω στις οντολογίες που περιγράφονται από τα δεδομένα RDF με σκοπό την εξαγωγή συμπερασμάτων.

- Απόδειξη

Δεδομένου ότι τα δεδομένα RDF προέρχονται από διάφορες και κατακευκαλισμένες πηγές και επεξεργάζονται από πράκτορες λογισμικού, είναι λογικό αυτό το λογισμικό να παρέχει στον πελάτη τη δυνατότητα ελέγχου και της διαδικασίας και της εγκυρότητας των αποτελεσμάτων.

- Εμπιστοσύνη

Άμεσα συνδεδεμένο με τα δύο προηγούμενα επίπεδα αυτό προσφέρει ασφάλεια τόσο στον πελάτη, όσο και στον εξυπηρετητή, χρησιμοποιώντας διάφορους μηχανισμούς ελέγχου πρόσβασης στα δεδομένα από τους διάφορους πράκτορες λογισμικού.

- Κρυπτογράφηση

Τα παραπάνω εξασφαλίζονται μέχρι ενός σημείου με την κρυπτογράφηση και τις ψηφιακές υπογραφές. Το επίπεδο αυτό βρίσκεται στο πλάι της στοίβας υποδηλώνοντας ότι αν και σημαντικό, κατά κάποιο τρόπο είναι εφάπτομενο του πυρήνα των τεχνολογιών του σημασιολογικού ιστού.

2.2 Δεδομένα και μεταδεδομένα

Τα μεταδεδομένα έχουν τεράστια σημασία τόσο για τα μεμονωμένα άτομα όσο και για τους οργανισμούς και τις επιχειρήσεις που όλο και περισσότερο στρέφονται προς τον Παγκόσμιο Ιστό για την απόκτηση και την διάδοση κάθε είδους πληροφορίας. Όταν πρωτοεμφανίστηκαν σαν ιδέα εφαρμόζονταν κυρίως στο περιβάλλον μιας βιβλιοθήκης, φυσικής ή ψηφιακής. Σήμερα όμως παρατηρείται μια πρωτοφανής δραστηριότητα γύρω από την παραγωγή μεταδεδομένων σε μια προσπάθεια να αντεπεξέλθουμε στο εκθετικά αυξανόμενο μέγεθος του Παγκόσμιου Ιστού με αποτέλεσμα αυτά να υποστηρίζουν πολλά ακόμα πεδία όπως εταιρικά Intranets, αποθήκες δεδομένων, ηλεκτρονικό εμπόριο, εκπαιδευτικά περιβάλλοντα, ψηφιακά μουσεία και άλλους δημόσιους ιστότοπους.

Ο πλέον συνήθης και σύντομος ορισμός για τα μεταδεδομένα είναι: "δεδομένα σχετικά με άλλα δεδομένα". Ένας όμως πιο αναλυτικός και πιο σαφής ορισμός σύμφωνα με την Greenberg J. (2003) θα μπορούσε να είναι ο εξής: "Τα μεταδεδομένα είναι ένα σύνολο από καλά δομημένα δεδομένα που αφορούν ένα αντικείμενο και τις λειτουργίες που σχετίζονται με αυτό το αντικείμενο", όπου ως αντικείμενο μπορούμε να θεωρήσουμε από μια ψηφιακή πηγή δεδομένων μέχρι πρόσωπα, γεγονότα, συναλλαγές, οδηγίες χρήσης, προγραμματιστικές εφαρμογές και πολλά άλλα [8].

2.3 Ανοιχτά Συνδεδεμένα Δεδομένα

Ο όρος Ανοιχτά Συνδεδεμένα Δεδομένα αναφέρεται στη χρήση του Web με τη δημιουργία συνδέσεων μεταξύ δεδομένων που προέρχονται από διαφορετικές πηγές και στα οποία μπορεί ο κάθε χρήστης να έχει ελεύθερη πρόσβαση. Από τεχνική άποψη, τα Συνδεδεμένα Δεδομένα αφορούν δεδομένα που δημοσιεύονται στον Ιστό με τέτοιο τρόπο ώστε να είναι αναγνώσιμα από μηχανές, η σημασία τους ορίζεται ρητά, συνδέονται με άλλα εξωτερικά σύνολα δεδομένων και με τη σειρά τους μπορούν να αποτελέσουν το σημείο που δείχνει μια σύνδεση από ένα άλλο σύνολο δεδομένων.

Ο Berners-Lee περιέγραψε ένα σύνολο «κανόνων» για τη δημοσίευση δεδομένων στον Ιστό με τέτοιο τρόπο ώστε όλα τα δημοσιευμένα δεδομένα να γίνουν μέρος ενός ενιαίου συνολικού χώρου δεδομένων:

1. Χρησιμοποιήστε URI ως ονόματα για πράγματα
2. Χρησιμοποιήστε HTTP URIs έτσι ώστε οι χρήστες να μπορούν να αναζητούν αυτά τα ονόματα
3. Όταν κάποιος ψάχνει ένα URI, δώστε χρήσιμες πληροφορίες χρησιμοποιώντας τα πρότυπα (RDF, SPARQL)
4. Συμπεριλάβετε συνδέσμους με άλλα URIs, ώστε οι χρήστες να μπορούν να ανακαλύψουν περισσότερα πράγματα

Αυτές έχουν γίνει γνωστές ως «αρχές συνδεδεμένων δεδομένων» και παρέχουν μια βασική συνταγή για τη δημοσίευση και τη σύνδεση δεδομένων χρησιμοποιώντας την υποδομή του Ιστού, τηρώντας παράλληλα την αρχιτεκτονική και τα πρότυπα του.[9].

2.4 Τεχνολογίες του σημασιολογικού ιστού

2.4.1 URIs

Στην ορολογία του Παγκόσμιου Ιστού όλα τα αντικείμενα ενός πεδίου ενδιαφέροντος ονομάζονται πόροι. Διακρίνουμε δύο τύπους πόρων, τους πληροφοριακούς και τους μη πληροφοριακούς. Πληροφοριακούς πόρους ονομάζουμε ότι βρίσκεται στον Ιστό των εγγράφων δηλαδή αρχεία που περιέχουν κείμενα, εικόνες, σχέδια κ.λ.π. ενώ οτιδήποτε δεν είναι έγγραφο αλλά αντιπροσωπεύει απτά αντικείμενα του πραγματικού κόσμου ή ακόμα και έννοιες ονομάζονται μη πληροφοριακοί πόροι. Μη πληροφοριακοί πόροι θα μπορούσαν παραδείγματος χάριν να είναι άνθρωποι, όπως οι εργαζόμενοι μιας επιχείρησης, αντικείμενα όπως τα εμπορεύματα ενός καταστήματος, αφηρημένες έννοιες όπως έννοιες της ψυχολογίας ή τα φιλοσοφικά ρεύματα μιας εποχής.

Στην ορολογία του σημασιολογικού Ιστού ένας άλλος όρος που χρησιμοποιείται για να καλύψει τόσο τις άυλες όσο και τις απτές έννοιες είναι ο όρος οντότητα. Για να μπορούμε να κάνουμε δηλώσεις για τις οντότητες ενός πεδίου ενδιαφέροντος και να είναι εντελώς ξεκάθαρο σε ποιαν οντότητα αναφερόμαστε θα πρέπει σε κάθε μία να αναθέσουμε ένα όνομα το οποίο να την αντιπροσωπεύει μονοσήμαντα, δηλαδή να αποτελεί μίαν αναφορά (reference) γι' αυτήν. Οι ελάχιστες προϋποθέσεις για αποτελεσματική ονοματοδοσία στις οντότητες είναι η αποφυγή της αμφισημίας, μια κεντρική απονομή ονομάτων που θα ισχύει σε παγκόσμιο επίπεδο καθώς και η δυνατότητα να αποθηκεύονται αυτά τα αναγνωριστικά ονόματα σε αρχεία του παγκόσμιου ιστού και να μεταφέρονται με τα πρωτόκολλα του παγκόσμιου ιστού. Σαν τέτοια ονόματα οντοτήτων

που καλύπτουν απόλυτα αυτές τις προϋποθέσεις χρησιμοποιούμε τα URIs τα οποία είναι ένα σύνολο χαρακτήρων (string) με σαφείς κανόνες σύνταξης σύμφωνα με κάποιο διεθνές πρότυπο.

Η πιο γνωστή μορφή των URIs μοιάζει με διεύθυνση εγγράφου στο παγκόσμιο ιστό (Uniform Resource Locator - URL). Μια URL μπορεί να χρησιμοποιηθεί σαν αναγνωριστικό όνομα χωρίς να υπάρχει η αντίστοιχη πληροφορία για την οντότητα. Σ' αυτήν την περίπτωση, που είναι και η πιο συνηθισμένη, μιλάμε για τα HTTP URIs. Έχουν προταθεί και άλλες μορφές όπως τα URNs τα οποία λειτουργούν χωρίς να εμπλέκουν θέση ή μηχανισμό πρόσβασης ή τα info URIs που είχαν υποστηριχθεί από μεγάλα ιδρύματα βιβλιοθηκών αλλά και τα δύο παραγκωνίζονται όλο και περισσότερο από τα HTTP URIs[10].

Πολλές φορές παρατηρούμε μέσα στον όγκο των πληροφοριών του Σημασιολογικού Ιστού το φαινόμενο των "ταυτόσημων URIs". Αυτό συμβαίνει όταν διαφορετικοί χρήστες αναφέρονται στην ίδια οντότητα χωρίς να γνωρίζουν ο ένας για τη δουλειά του άλλου, οπότε δίνουν διαφορετικά URIs για το ίδιο αντικείμενο. Αυτό είναι απόλυτα θεμιτό και μάλλον αποτελεί και μια σημαντική κοινωνική προσφορά στο χώρο της Πληροφορίας μιας και επιτρέπει την έκφραση πολλών και διαφορετικών απόψεων για ένα θέμα. Μια καλή πρακτική που πρέπει όμως να ακολουθεί κανείς πάνω σε αυτό είναι όταν γνωρίζει άλλα URIs που αναφέρονται στις οντότητες που διαχειρίζεται, να θέτει συνδέσμους του τύπου owl:sameAs που δείχνουν προς τα ταυτόσημα URIs.

2.4.2 Οντολογίες-Λεξιλόγια

Σήμερα χρησιμοποιούνται όλο και περισσότερο οι οντολογίες σαν βασικό εργαλείο των ερευνητών που ασχολούνται με την ανάπτυξη εφαρμογών στο Σημασιολογικό Ιστό. Ο ορισμός του όρου "οντολογία" που χρησιμοποιείται συχνά σε σχετικά επιστημονικά άρθρα είναι αυτός που εισήγαγε ο Gruber: "η οντολογία είναι η ρητή προδιαγραφή μιας εννοιολογικής σύλληψης [11]. Αυτό σημαίνει ότι μπορεί να αναπαραστήσει με ακρίβεια ένα πεδίο ενδιαφέροντος και αυτό επιτυγχάνεται με τη χρήση "οντοτήτων" που αντιστοιχούν σε κάθε από αντικείμενο ή έννοια που εμπλέκεται σε αυτό το πεδίο καθώς και "ιδιοτήτων" που αντιστοιχούν σε σχέσεις που υφίστανται μεταξύ των οντοτήτων. Με αυτόν τον τρόπο καταφέρνει να επιτύχει διαμοιρασμό της γνώσης ανάμεσα σε κατανεμημένα συστήματα, δηλαδή προάγει τη διαλειτουργικότητα ανάμεσα σε προγράμματα εφαρμογών τα οποία δεν είναι εξ αρχής φτιαγμένα για να συνεργάζονται. Έτσι η χρήση οντολογιών καθίσταται το πιο σημαντικό δομικό συστατικό των εφαρμογών που στόχο έχουν να λύσουν τα προβλήματα ανομοιογένειας των σημασιολογικών δεδομένων και ως εκ τούτου βρίσκουν εφαρμογή σε πολλούς τομείς όπως η Τεχνητή Νοημοσύνη, το πεδίο διαχείρισης της γνώσης, το ηλεκτρονικό εμπόριο, η επεξεργασία φυσικών γλωσσών κ.ά [12].

Μια οντολογία αποτελείται από τέσσερεις τύπους δομικών συστατικών οι οποίοι είναι: οι έννοιες ή κλάσεις, τα στιγμιότυπα, οι σχέσεις και τα αξιώματα. Παρακάτω ορίζεται τι είναι το κάθε ένα από αυτά:

1. Η έννοια ή αλλιώς κλάση είναι ένα σύνολο αντικειμένων που αποτελούν τα θεμελιώδη στοιχεία του πεδίου ενδιαφέροντος και του οποίου τα μέλη έχουν κοινές ιδιότητες. Η βασική έννοια αντιπροσωπεύεται από μια υπερκλάση (super-class) οποία μπορεί να έχει πολλές υποκλάσεις (subclasses) που αντιπροσωπεύουν άλλες έννοιες που είναι υποσύνολα της βασικής. Για παράδειγμα ένα πανεπιστήμιο μπορεί να είναι μια υπερκλάση που έχει σαν υποκλάσεις τις σχολές, τις βιβλιοθήκες, τα εργαστήρια κ.λ.π.

2. Το στιγμιότυπο είναι ένα συγκεκριμένο αντικείμενο μιας κλάσης π.χ. η "Ελλάδα" είναι

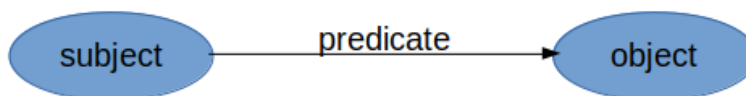


Figure 2: Αναπαράσταση μιας τριάδας RDF

ένα στιγμιότυπο της κλάσης "Χώρες".

3. Η σχέση αντιπροσωπεύει έναν τρόπο σύνδεσης μίας έννοιας με μία άλλη. Συγκεκριμένα μία έννοια από ένα πεδίο ορισμού συνδέεται μέσω μίας σχέσης με μία άλλη έννοια από ένα πεδίο τιμών. Παραδείγματος χάριν η "Ελλάδα" που ανήκει στο πεδίο ορισμού "Χώρες" συνδέεται με την σχέση "βρίσκεται σε" με την "Ευρώπη" που ανήκει στο πεδίο τιμών "Ήπειροι".

4. Τα αξιώματα χρησιμοποιούνται για να θέτουν περιορισμούς στις τιμές των κλάσεων και των στιγμιότυπων τους και μέσω αυτών να ελέγχεται η συνέπεια της οντολογίας[12].

Τα λεξιλόγια (vocabularies) είναι ένα σύνολο από όρους που χρησιμοποιεί μια γλώσσα αναπαράστασης οντολογιών για να περιγράψει ένα συγκεκριμένο γνωστικό πεδίο. Οι όροι αυτοί διατίθενται και ελέγχονται από μία αρμόδια αρχή και γι' αυτό λέγονται ελεγχόμενα λεξιλόγια. Με αυτόν τον τρόπο διασφαλίζονται κάποιες προϋποθέσεις που πρέπει να τηρούνται σχετικά με την εγκυρότητα των όρων όπως να αποφεύγεται η αμφισημία δηλαδή να χρησιμοποιείται ο ίδιος όρος για δύο διαφορετικές έννοιες ή να διασαφηνίζονται περιπτώσεις συνωνυμίας. Η ιδέα για αυτόν τον όρο "ελεγχόμενα λεξιλόγια" προέρχεται από τον χώρο των βιβλιοθηκών και των υπηρεσιών πληροφόρησης κατά τη δεκαετία του 1950. Αν και δεν υπάρχει σαφής διαχωρισμός μεταξύ οντολογίας και λεξιλογίου θα μπορούσαμε να πούμε ότι η αναπαράσταση μιας οντολογίας γίνεται με τη χρήση ενός λεξιλογίου και μιας γλώσσας αναπαράστασης οντολογιών όπως η RDF και RDFS, η OIL, η DAML+OIL και η εξέλιξη αυτών η OWL.[13].

2.4.3 Το μοντέλο RDF

Το μοντέλο RDF (Resource Description Framework) εκδόθηκε τον Φεβρουάριο του 1999 σαν οδηγία από τον οργανισμό W3C (World Wide Web Consortium) που ορίζει πλήρως τον τρόπο μοντελοποίησης των δεδομένων καθώς και τη σύνταξή του βασισμένη στην XML.

Το μοντέλο RDF θυμίζει κάπως ένα αντικειμενοστραφές μοντέλο δεδομένων (object-oriented data model). Αποτελείται από οντότητες που προσδιορίζονται από μοναδικά αναγνωριστικά και δυαδικές σχέσεις ή δηλώσεις μεταξύ αυτών των οντοτήτων. Σε μία γραφική αναπαράσταση μιας δήλωσης RDF, η πηγή της σχέσης ονομάζεται υποκείμενο (subject), το βέλος με την ετικέτα που το συνοδεύει ονομάζεται κατηγορήμα (ή ιδιότητα) (predicate) και ο προορισμός της σχέσης είναι το αντικείμενο (object). Συντακτικά οι προτάσεις αυτές δηλώνονται από μια διατεταγμένη τριάδα της μορφής, s p o (subject, predicate, object). Ένα σύνολο από τριάδες RDF μπορούμε να το αντιληφθούμε και ως έναν γράφο. Σε αυτόν τον γράφο τα υποκείμενα και τα αντικείμενα παίζουν το ρόλο των κόμβων ενώ οι ιδιότητες παίζουν το ρόλο των ακμών που τους συνδέουν, όπως φαίνεται στην εικόνα Figure 2.

Στο RDF μοντέλο γίνεται διαχωρισμός ανάμεσα στις οντότητες (entities) που προσδιορίζονται πάντα με URIs σαν αναγνωριστικά και στις απλές τιμές (literals) που είναι απλά συμβολοσειρές (strings). Το υποκείμενο και το κατηγορήμα μιας δήλωσης είναι πάντα οντότητες

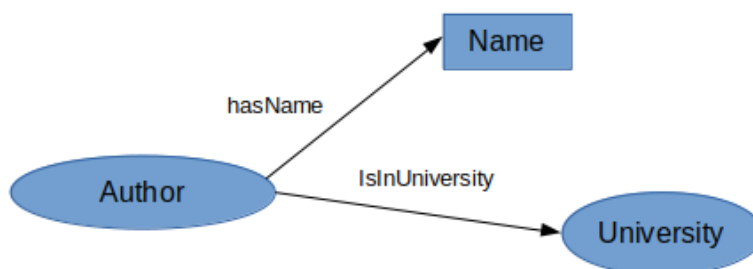


Figure 3: Αναπαράσταση τριάδων RDF

ενώ το αντικείμενο μπορεί να είναι είτε οντότητα είτε απλή τιμή. Όπως φαίνεται και στην εικόνα Figure 3 στις γραφικές απεικονίσεις η οντότητα αναπαριστάται με έλλειψη ενώ η απλή τιμή με τετράγωνο [14].

Οι βασικές δομές μοντελοποίησης της γλώσσας RDFS φαίνονται στον πίνακα Table 1 [13].

2.4.4 Αποθετήρια τριάδων (Triplestores)

Τα τελευταία χρόνια η διακίνηση μεγάλου όγκου δεδομένων έχει γίνει απαραίτητη σε πάρα πολλά πεδία όπως στην αγορά, στον τομέα της βιομηχανικής παραγωγής, στον τομέα του πολιτισμού (βιβλιοθήκες, μουσεία, πανεπιστήμια), στην έρευνα και αλλού. Η ανάγκη για ευέλικτα σχήματα δεδομένων που να είναι διαχειρίσιμα από ανομοιογενή ή κατακευματισμένα συστήματα επεξεργασίας δεδομένων γίνεται όλο και μεγαλύτερη, με αποτέλεσμα να προτιμάται από τους προμηθευτές η παροχή τους σε μορφή RDF. Πολλές φορές η μετατροπή των δεδομένων σε τέτοια μορφή γίνεται, από διάφορες εφαρμογές, on-the-fly τη στιγμή που ζητούνται. Ο μεγάλος όγκος όμως των δεδομένων και οι απαιτήσεις επεξεργασίας συνήθως απαιτούν έναν μόνιμο τρόπο αποθήκευσης. Σ' αυτή την περίπτωση μιλάμε για αποθετήρια τριάδων RDF (triplestores) και συνεπώς και για εφαρμογές που κατασκευάζουν και διαχειρίζονται αυτά τα αποθετήρια. Καθώς λοιπόν οι σημασιολογικές εφαρμογές καταλαμβάνουν όλο και μεγαλύτερο χώρο στη διαχείριση δεδομένων, τίθενται μεγάλες προκλήσεις στους ερευνητές και τους τεχνικούς με τα ζητήματα αποτελεσματικής αποθήκευσης και ανάκτησης των δεδομένων που ανακύπτουν [15] [16].

Η αρχιτεκτονική μιας τέτοιας αποθήκης είναι αυτή που φαίνεται στην εικόνα Figure 4 [15].

Τα κύρια συστατικά αυτής της αρχιτεκτονικής εξηγούνται παρακάτω:

α) Το αποθετήριο (repository): ένα σύνολο από αρχεία ή εναλλακτικά μία βάση δεδομένων
 β) Το ενδιάμεσο λογισμικό (middleware) που εξασφαλίζει την σταθερή επικοινωνία με την υποκείμενη βάση δεδομένων. Το σύνολο αυτού του λογισμικού περιλαμβάνει :

1) τον προμηθευτή των δεδομένων (storage provider) που είναι σε απ'ευθείας σύνδεση με το αποθετήριο

2) τη μηχανή ερωτημάτων (Query Engine) που απευθύνει δομημένα ερωτήματα σύμφωνα με τη σύνταξη μιας γλώσσας ερωτημάτων (π.χ. SPARQL) με σκοπό την λήψη πληροφορίας.

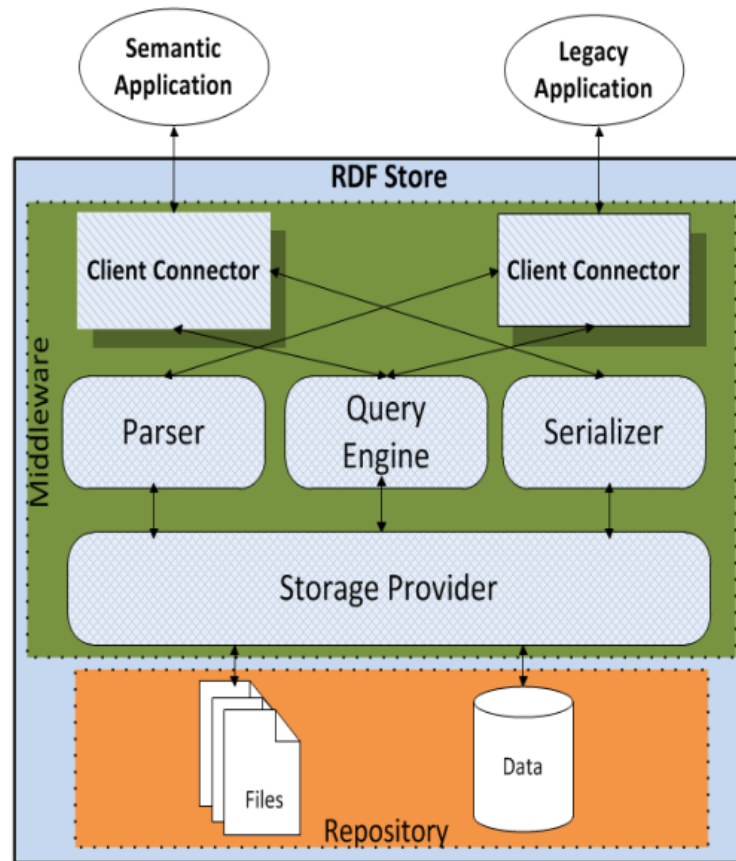


Figure 4: Η Αρχιτεκτονική μιας Triplestore.

Δομή	Σύνταξη τριάδας	Περιγραφή
Class (μια κλάση)	C rdf:type rdfs:Class	C (ένας πόρος) είναι μια κλάση RDF
Property (μια κλάση)	P rdf:type rdf:Property	P (ένας πόρος) είναι μια ιδιότητα RDF
type (μια ιδιότητα)	I rdf:type C	I (ένας πόρος) είναι ένα στιγμιότυπο της C
subClassOf (μια ιδιότητα)	C1 rdfs:subClassOf C2	C1 (μια κλάση) είναι υποκλάση της C2
subPropertyOf (ιδιότητα)	P1 rdfs:subPropertyOf P2	P1 (μια ιδιότητα) είναι υποιδιότητα της P2
domain (μια ιδιότητα)	P rdfs:domain C	Το πεδίο ορισμού P (μιας ιδιότητας) είναι C (μια κλάση)
range (μια ιδιότητα)	P rdfs:range C	Το σύνολο τιμών P (μιας ιδιότητας) είναι C (μια κλάση)

Table 1: Δομές μοντελοποίησης της γλώσσας RDFS

3) τον αναλυτή/σειριοποιητή (Parser/Serializer). Ο αναλυτής διαβάζει κείμενο που είναι γραμμένο σε μία γλώσσα σειριοποίησης RDF μοντέλων (π.χ. XML) και το μετατρέπει σε τριάδες RDF. Ο σειριοποιητής κάνει το αντίστροφο δηλαδή παίρνει ένα σύνολο από τριάδες RDF και δημιουργεί ένα αρχείο που περιέχει την πληροφορία εκφρασμένη σε μια σειριοποιημένη μορφή.

4) τους συνδέσμους για τον πελάτη (client connector) που είναι διαθέσιμοι σε κάποια γλώσσα προγραμματισμού για προγραμματιστική πρόσβαση και διαχείριση του περιεχομένου του αποθετηρίου.

Υπάρχουν αρκετά κριτήρια για να αξιολογηθεί η αποτελεσματικότητα μιας αποθήκης RDF τριάδων. Μερικά από αυτά είναι:[15]

1) Η πρόσβαση και η ανάκτηση των δεδομένων. Το πιο σημαντικό κριτήριο για την αποτελεσματικότητα μιας αποθήκης rdf τριάδων είναι η εύκολη και γρήγορη πρόσβαση στα δεδομένα καθώς και η ανάκτησή τους.

2) Η ασφάλεια (security). Καθώς οι χρήστες μπορούν να έχουν πρόσβαση και δυνατότητα επεξεργασίας στα δεδομένα πρέπει να υπάρχουν οι απαραίτητοι κανόνες ασφάλειας και να δίνονται δικαιώματα στους χρήστες με τέτοιο τρόπο ώστε να διασφαλίζεται η ακεραιότητα (Integrity), η συνέπεια (Consistency) και η ιδιαιτερότητα (Privacy) των δεδομένων.

3) Ο έλεγχος της έκδοσης (versioning). Πρέπει να ελέγχεται συχνά η βάση για αλλαγές στα δεδομένα ώστε να διασφαλίζεται η συνέπεια των δεδομένων.

4) Η διαχείριση δεδομένων σε δυαδική μορφή (handling binary data). Η διαχείριση δεδομένων μεγάλου όγκου γίνεται καλύτερα όταν τα δεδομένα είναι αποθηκευμένα σε δυαδική μορφή διότι έχουμε καλύτερη αναπαράσταση της πληροφορίας. Συγκεκριμένα τα πλεονεκτήματα είναι:

- α) Τυποποιημένη πρόσβαση στα δεδομένα και τα μεταδεδομένα.
- β) Μεγαλύτερη ευκολία στο να κρατάμε τα δεδομένα και τα μεταδεδομένα συγχρονισμένα.
- γ) Μεγαλύτερη ευκολία στο να ελέγχουμε τη βάση για καινούργια έκδοση.
- 5) Η διαχείριση της ροής των δεδομένων (handling streaming data). Σήμερα ο όγκος των δεδομένων που συλλέγονται από εξαιρετικά καταναεμημένα συστήματα συσκευών, όπως για παράδειγμα οι αισθητήρες, αυξάνεται εκθετικά, καθώς υπάρχει συνεχόμενη ροή δεδομένων σε πολύ μεγάλες ταχύτητες. Παρατηρείται λοιπόν μια δυσκολία στη διαχείρισή τους, ειδικά στα συστήματα πραγματικού χρόνου με αποτέλεσμα να παρατηρείται το φαινόμενο "υπερβολικά πολλά δεδομένα - λίγη χρήσιμη πληροφορία". Με στόχο να καλυφθεί αυτό το κενό είναι εξαιρετικά ενδιαφέρον να απαντηθεί εάν οι σημασιολογικές εφαρμογές άρα και οι αποθήκες RDF τριάδων, είναι ώριμες να διαχειριστούν αποτελεσματικά δεδομένα σε ροή.

Μερικές από τις εφαρμογές που διαχειρίζονται αποθετήρια RDF δεδομένων είναι οι: AllegroGraph, Bigdata, Marmotta, Openlink Virtuoso, Oracle 12c, Stardog, κ.ά.[16][17]

2.4.5 Γλώσσες ερωτημάτων σε RDF δεδομένα

Όπως αναφέρθηκε παραπάνω τα αποθετήρια δεδομένων σε μορφή RDF διαχειρίζονται έναν μεγάλο όγκο πληροφοριών, πράγμα το οποίο βελτιώνει την έρευνα και την δύναμη ανάλυσης των οργανισμών. Κάτι πολύ σημαντικό είναι να είμαστε σε θέση να μπορούμε να εξάγουμε πληροφορίες που υπονοούνται μέσα από ερωτήματα που απευθύνουμε στις βάσεις δεδομένων. Εντοπίζοντας τις σχέσεις που συνδέουν τα αρχικά δεδομένα με την βοήθεια μιας σημασιολογικής βάσης γράφων μετατρέπουμε την πληροφορία σε γνώση. Αυτό επιτρέπει στον χρήστη να ανακαλύψει την πληροφορία που βρίσκεται στα ήδη υπάρχοντα δεδομένα του, και κατ' επέκταση να βοηθήσει στην δημιουργία πιο έξυπνων λύσεων που θα βελτιώσουν το τρόπο διαχείρισης των προβλημάτων που καλείται να λύσει με τη βοήθεια των πληροφοριακών συστημάτων. Το παραπάνω υλοποιείται με την βοήθεια των γλωσσών ερωτημάτων (query languages) οι οποίες στην ουσία είναι λογισμικό που τηρεί ένα σύνολο από προδιαγραφές με βάση τις οποίες μπορεί να διαβάσει, να γράφει και να τροποποιεί τη βάση δεδομένων.

Κάποιες σημαντικές ιδιότητες που πρέπει να έχει μια γλώσσα ερωτημάτων για να είναι αποτελεσματική είναι :[18] [19] [20]

1) Εκφραστικότητα (expressiveness). Η εκφραστικότητα δείχνει πόσο ισχυρά ερωτήματα μπορεί να δομήσει μία γλώσσα, γι' αυτό πρέπει να διαθέτει κατ' ελάχιστο κάποια εκφραστικά μέσα όπως για παράδειγμα αυτά που προσφέρει η σχεσιακή άλγεβρα.

2) Closure. Αυτή η ιδιότητα σημαίνει ότι το αποτέλεσμα που επιστρέφει το ερώτημα πρέπει να είναι στην ίδια μορφή με το μοντέλο των δεδομένων έτσι ώστε να μπορεί με τη σειρά του να αποτελέσει είσοδο σε καινούργιο ερώτημα.

3) Επάρκεια (Adequacy), που σημαίνει ότι θα πρέπει να καλύπτει όλες τις έννοιες του υποκείμενου μοντέλου και μόνο αυτές.

4) Ορθογωνικότητα (Orthogonality). Η ορθογωνικότητα απαιτεί όλες οι λειτουργίες της γλώσσας να εφαρμόζονται το ίδιο αποτελεσματικά ανεξάρτητα από το περιεχόμενο της βάσης δεδομένων.

5) Ασφάλεια (safety). Μία γλώσσα ερωτημάτων θεωρείται ότι παρέχει ασφάλεια αν κάθε σωστά διατυπωμένο ερώτημα επιστρέφει ένα καλά ορισμένο σετ αποτελεσμάτων. Γενικά μία

γλώσσα θεωρείται μη ασφαλής όταν εκτελεί λειτουργίες αναδρομής (recursion), άρνησης (negation) ή περιλαμβάνει ενσωματωμένες (built-in) λειτουργίες.

Μερικές από τις πιο γνωστές γλώσσες ερωτημάτων είναι οι SPARQL, RDQL, Triple, SeRQL, Versa, N3, RQL.

3 Ανάλυση Προβλήματος

3.1 Διατύπωση προβλήματος

Στόχος αυτής της διπλωματικής εργασίας είναι να δημιουργηθεί ένα σύνολο από βιβλιογραφικά δεδομένα σε μορφή RDF και στη συνέχεια να παρουσιαστούν κάποια εργαλεία με τα οποία μπορούμε να κάνουμε διαχείριση αυτών των δεδομένων. Επιλέχθηκε αυτά τα βιβλιογραφικά δεδομένα να αφορούν τις δημοσιεύσεις των μελών του διδακτικού προσωπικού του τμήματος Πληροφορικής του Ιονίου Πανεπιστημίου. Στην παρούσα ενότητα παρουσιάζεται η θεωρητική μελέτη που έγινε προκειμένου να δημιουργηθεί η οντολογία που αναπαριστά το συγκεκριμένο γνωστικό πεδίο.

3.2 Κατασκευή οντολογίας

Αφού βρέθηκε το ζητούμενο σύνολο από βιβλιογραφικά δεδομένα σε μορφή Bibtex, με τον τρόπο που περιγράφεται αναλυτικά στην ενότητα 4, πραγματοποιήθηκε μια λεπτομερής εξέταση και ανάλυση των δεδομένων με σκοπό να εξαχθεί το μέγιστο της πληροφορίας που περικλείουν καθώς και η σημασιολογική τους αξία με σκοπό να παραχθεί η οντολογία που αναπαριστά το συγκεκριμένο γνωστικό πεδίο.

Παρακάτω αναλύεται μια εγγραφή από τα βιβλιογραφικά δεδομένα που βρέθηκαν σε μορφή Bibtex με όσα πεδία κρίθηκε αναγκαίο να χρησιμοποιηθούν για τις ανάγκες της παρούσας εργασίας. Η εγγραφή αυτή φαίνεται στην εικόνα Figure 5.

Όπως φαίνεται στην εικόνα η εγγραφή ξεκινάει με το σύμβολο '@' και μία λέξη που δείχνει αν η συγκεκριμένη εγγραφή είναι 'άρθρο' (article), δημοσίευση σε συνέδριο (inproceedings) ή βιβλίο (book) και στη συνέχεια ανοίγει η αγκύλη που περικλείει τα βιβλιογραφικά δεδομένα. Το πρώτο στοιχείο είναι ένας μοναδικός κωδικός που ταυτοποιεί την εγγραφή. Τα υπόλοιπα είναι τα πεδία που περιγράφουν την βιβλιογραφική εγγραφή και οι τιμές τους, που αντιστοιχούνται μεταξύ τους με το σύμβολο '='. Οι πληροφορίες που δίνονται στη συγκεκριμένη εγγραφή είναι οι εξής:

- Η δημοσίευση είναι ένα άρθρο με κωδικό 'DBLP:journals/monet/FlorosAP08'.
- Οι συγγραφείς (author) του άρθρου είναι οι Andreas Floros και Markos Avlonitis και Panayiotis M. Vlamos.
- Ο τίτλος (title) του άρθρου είναι 'Frequency-Domain Stochastic Error Concealment for Wireless Audio Applications'.
- Το περιοδικό (journal) στο οποίο δημοσιεύτηκε είναι το 'MONET'.
- Το έτος (year) δημοσίευσης είναι '2008'.
- Το URL (url) του άρθρου είναι 'https://doi.org/10.1007/s11036-008-0059-z'.

Μετά από τη μελέτη αυτών των πληροφοριών ορίστηκαν οι οντότητες που περιγράφουν το πεδίο των βιβλιογραφικών εγγραφών καθώς και τα χαρακτηριστικά της κάθε μιας και οι σχέσεις μεταξύ τους. Η βασική οντότητα είναι η 'Publication' (Δημοσίευση). Οι υπόλοιπες είναι οι 'Author' (Συγγραφέας), 'Publisher' (Εκδότης) και 'PublishedIn' (Μέσο Δημοσίευσης). Στους πίνακες Table 2 και Table 3 φαίνονται τα χαρακτηριστικά των οντοτήτων 'Publication' και 'Author'.

Στην εικόνα Figure 6 φαίνεται η σχηματική αναπαράσταση της οντολογίας όπου οι ελλείψεις

```
@article{DBLP:journals/monet/FlorosAP08,
  author = {Andreas Floros and
            Markos Avlonitis and
            Panayiotis M. Vlamos},
  title = {Frequency-Domain Stochastic Error Concealment
            for Wireless Audio Applications},
  journal = {{MONET}},
  year = {2008},
  url = {https://doi.org/10.1007/s11036-008-0059-z}|
}
```

Figure 5: Δείγμα βιβλιογραφικής εγγραφής σε μορφή bibtex.

ΟΝΤΟΤΗΤΑ PUBLICATION	
ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ	ΤΥΠΟΣ
Title	Literal
Author	Entity
TypeOfPublication	Literal (Journal, Conference ή Book)
PublishedIn	Entity
YearOfPublication	Literal
Publisher	Entity
URL	Literal

Table 2: Τα χαρακτηριστικά της οντότητας Publication

αντιστοιχούν σε οντότητες, τα παραλληλόγραμμα σε λεκτικές τιμές ενώ τα βέλη αντιστοιχούν στα χαρακτηριστικά των οντοτήτων 'Publication' και 'Author'.

Στην εικόνα Figure 7 φαίνεται ένα στιγμιότυπο της κλάσης 'Publication' και ένα της κλάσης 'Author'.

Στη συνέχεια έπρεπε να αναζητηθούν τα κατάλληλα λεξιλόγια έτσι ώστε να αναπαρασταθεί η οντολογία με τον τρόπο που απαιτούν οι βασικές αρχές του σημασιολογικού ιστού. Η έρευνα έγινε στη πλατφόρμα LOV ¹ η οποία περιλαμβάνει όλα τα λεξιλόγια που έχουν δημιουργηθεί για την αναπαράσταση οντολογιών στο σημασιολογικό ιστό. Μετά από προσεκτική μελέτη των λεξιλογίων που είναι σχετικά με το γνωστικό αντικείμενο που ενδιαφέρει την παρούσα μελέτη και λαμβάνοντας υπ' όψιν τα χαρακτηριστικά της οντολογίας κρίθηκαν κατάλληλα τα λεξιλόγια Rdf, RDFs, Dublin Core (DC) ², Bibliographic Ontology (BIBO) ³ και FOAF ⁴.

Οι όροι των λεξιλογίων που χρησιμοποιήθηκαν για να περιγράψουν τα χαρακτηριστικά των

¹<https://lov.linkeddata.es/dataset/lov/>

²<http://dublincore.org/documents/dces/>

³<http://bibliographic-ontology.org/bibo/bibo.php>

⁴<http://xmlns.com/foaf/spec/>

ΟΝΤΟΤΗΤΑ AUTHOR	
ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ	ΤΥΠΟΣ
Name	Literal
URL	Entity
MailAddress	Entity
University	Entity

Table 3: Τα χαρακτηριστικά της οντότητας Author

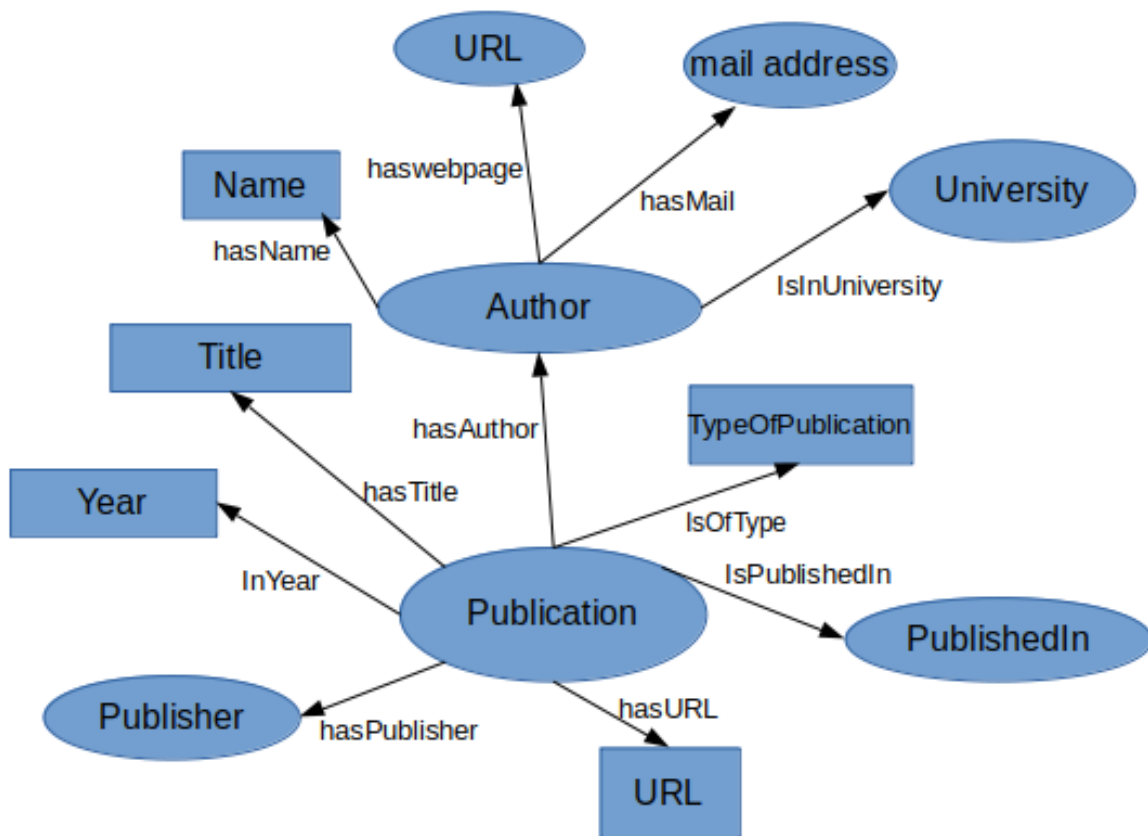


Figure 6: Σχηματική αναπαράσταση της οντολογίας.

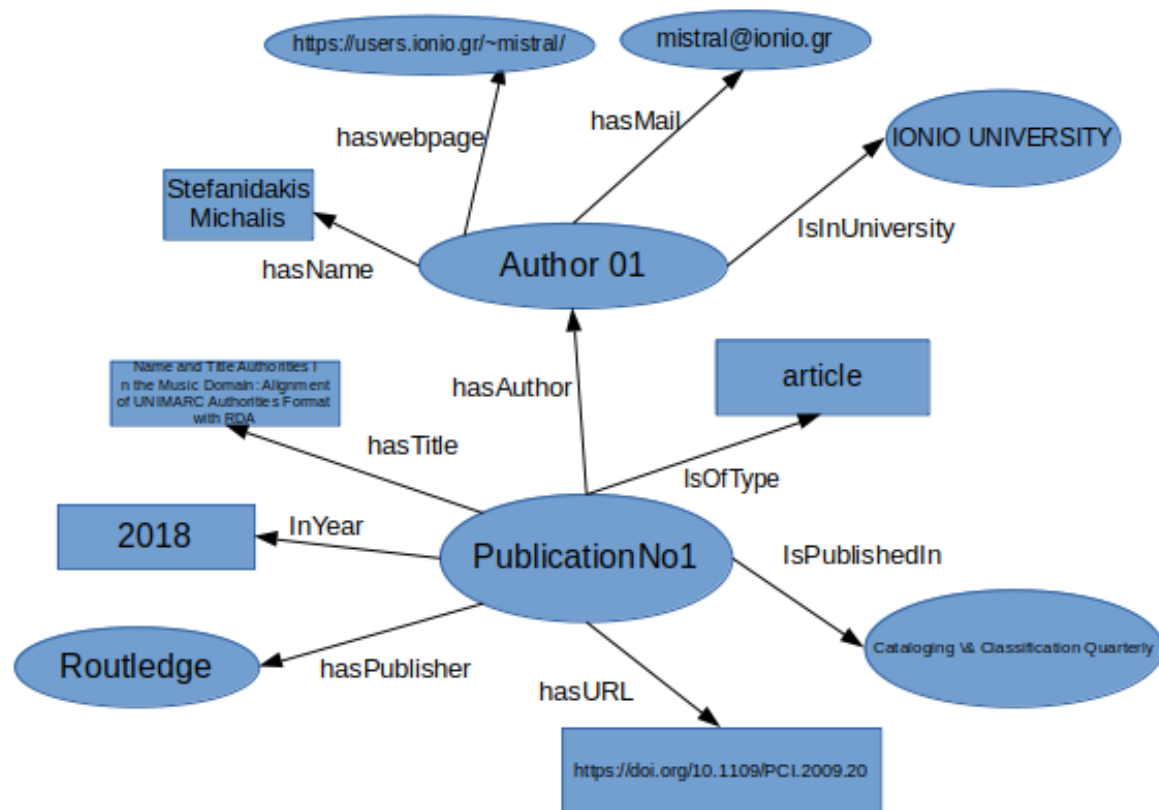


Figure 7: Στιγμιότυπα των κλάσεων 'Publication' και 'Author'.

οντοτήτων είναι:

- Από το λεξιλόγιο Dublin Core και από το χώρο ονομάτων (namespace) <http://purl.org/dc/elements/1.1/> (πρόθεμα dc:) χρησιμοποιήθηκαν οι όροι:

created, που δηλώνει την ημερομηνία δημιουργίας του πόρου,

creator, που δηλώνει τον δημιουργό του πόρου,

type, που δηλώνει τον τύπο του πόρου δηλαδή άρθρο, ιστοσελίδα, εικόνα κ.λ.π.,

publisher, που δηλώνει τον εκδότη του πόρου δηλαδή ένας οργανισμός, πρόσωπο ή υπηρεσία που ευθύνεται για τη διάθεση του πόρου,

title, που δηλώνει ένα όνομα που έχει δοθεί στον πόρο.

- Από το λεξιλόγιο Bibliographic Ontology (BIBO) και από το χώρο ονομάτων

<http://purl.org/ontology/bibo/> (πρόθεμα bibo:) χρησιμοποιήθηκαν οι όροι:

presentedAt, που συνδέει έναν πόρο με ένα γεγονός, για παράδειγμα ένα μια δημοσιευμένη εργασία με ένα συνέδριο,

uri, που δηλώνει το ενιαίο αναγνωριστικό ταυτοποίησης ενός πόρου.

- Από το λεξιλόγιο FOAF και από το χώρο ονομάτων <http://xmlns.com/foaf/0.1/> (πρόθεμα foaf:) χρησιμοποιήθηκαν οι όροι:

name, που δηλώνει το όνομα κάποιου προσώπου ή πράγματος,

homepage, που συνδέει κάποιον ή κάτι με μια ιστοσελίδα που περιέχει πληροφορίες σχετικά με αυτό,

mailbox, που συνδέει κάποιον με μια διεύθυνση ηλεκτρονικού ταχυδρομείου (σύμφωνα με το σχήμα <mailto:>),

workplaceHomepage, που συνδέει το υποκείμενο (που πρέπει να ανήκει στην κλάση Person) με μια ιστοσελίδα που περιέχει πληροφορίες για τον τόπο εργασίας του.

Εκτός από τα παραπάνω λεξιλόγια χρησιμοποιήθηκαν και τα λεξιλόγια `rdf` και `rdfs:Class` που περιγράφουν με πιο γενικό τρόπο τις κλάσεις και τις ιδιότητές τους, όπως για παράδειγμα δηλώνουν ότι ένας πόρος ανήκει σε μία κλάση ή ότι μια κλάση είναι υποκλάση μιας άλλης

Στους πίνακες Table 4 και Table 5 φαίνονται πως αντιστοιχούνται τα χαρακτηριστικά των οντοτήτων 'Publication' και 'Author' με τους όρους των παραπάνω λεξιλογίων. Κάθε χαρακτηριστικό είναι μια ιδιότητα που περιλαμβάνεται σε κάποιο λεξιλόγιο. Η τελευταία στήλη του κάθε πίνακα δείχνει το πεδίο τιμών της ιδιότητας δηλαδή σε ποιά κλάση ανήκουν οι τιμές που παίρνει.[13]

Στην εικόνα Figure 8 φαίνεται η οντολογία μετά την επιλογή των κατάλληλων όρων από τα λεξιλόγια.

3.3 Λεξιλόγια που χρησιμοποιήθηκαν

3.3.1 RDF - RDFs

Η γλώσσα RDF χρησιμοποιείται για την παραγωγή δηλώσεων που περιγράφουν τους πόρους ενός πεδίου γνώσης καθώς και τις μεταξύ τους σχέσεις οι οποίες με τη σειρά τους περιγράφονται από ιδιότητες που έχουν συγκεκριμένες τιμές. Για να υπάρχει η απαραίτητη τυποποίηση σ' αυτήν την περιγραφή χρησιμοποιούνται λεξιλόγια όρων όπως τα RDF και RDF Schema (RDFS). Τα

ONTOTHTA PUBLICATION				
ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ	ΤΥΠΟΣ	Λεξιλογο	Property Ιδιότητα	Range πεδίο τιμών
Title	Literal	Dublin Core	dc:title	rdfs:Literal
Author	Entity	Dublin Core	dc:creator	dcterms:Agent
TypeOfPublication	Literal	Dublin Core	dc:type	rdfs:Literal
PublishedIn	Entity	BIBO	bibo:presentedAt	bibo:event
YearOfPublication	Literal	Dublin Core	dc:created	rdfs:Literal
Publisher	Entity	Dublin Core	dc:publisher	dcterms:Agent
URL	Literal	BIBO	bibo:uri	rdfs:Literal

Table 4: Οι ιδιότητες της κλάσης Publication

ONTOTHTA AUTHOR				
ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ	ΤΥΠΟΣ	Λεξιλογο	Property Ιδιότητα	Range πεδίο τιμών
Name	Literal	FOAF	foaf:name	rdfs:Literal
URL	Entity	FOAF	foaf:homepage	foaf:Document
MailAddress	Entity	FOAF	foaf:mbox	owl:Thing
University	Entity	FOAF	foaf:workplaceHomepage	foaf:Document

Table 5: Οι ιδιότητες της κλάσης Author

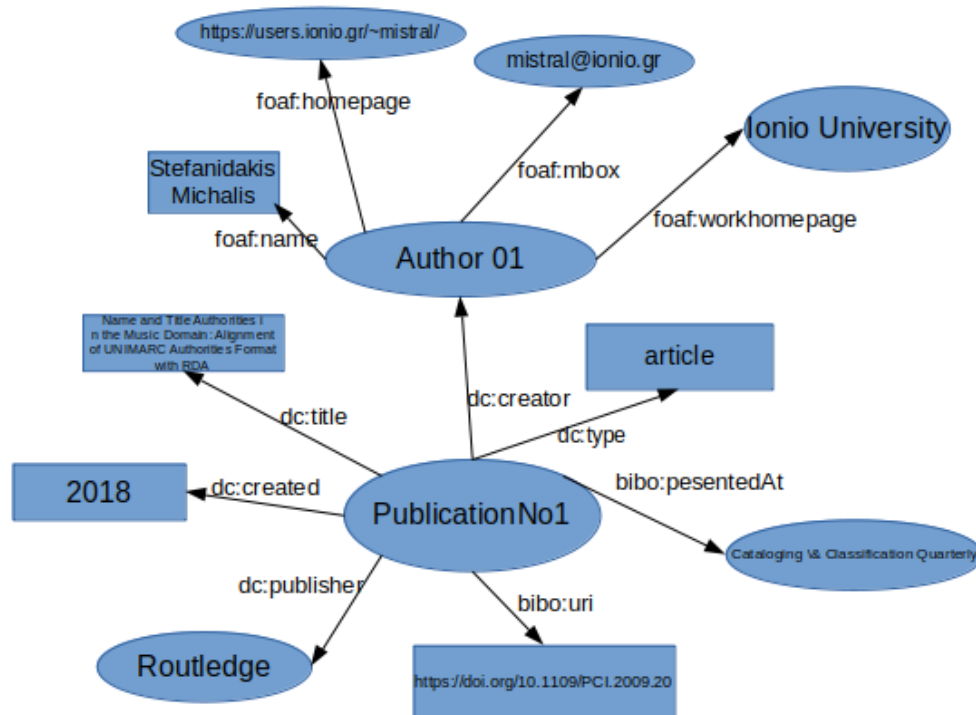


Figure 8: Η οντολογία με τη χρήση όρων γνωστών λεξιλογίων.

Κλάση	Περιγραφή
rdfs:Resource	Η κλάση όλων των πόρων.
rdfs:Literal	Η κλάση όλων των λεκτικών τιμών.
rdfs:Class	Η κλάση όλων των κλάσεων.
rdf:Property	Η κλάση όλων των ιδιοτήτων.
rdfs:Datatype	Η κλάση όλων των τύπων δεδομένων.

Table 6: Κλάσεις που χρησιμοποιούν τα λεξιλόγια της RDF

Ιδιότητα	Περιγραφή	Πεδίο ορισμού (domain)	Πεδίο τιμών (range)
rdf:type	Το υποκείμενο είναι στιγμιότυπο μίας κλάσης.	rdfs:Resource	rdfs:Class
rdfs:subClassOf	Το υποκείμενο είναι υποκλάση μίας κλάσης.	rdfs:Class	rdfs:Class
rdfs:domain	Το πεδίο ορισμού μιας ιδιότητας.	rdf:Property	rdfs:Class
rdfs:range	Το πεδίο τιμών μιας ιδιότητας.	rdf:Property	rdfs:Class
rdfs:label	Ένα όνομα για το υποκείμενο κατανοητό από τον άνθρωπο.	rdfs:Resource	rdfs:Literal
rdfs:comment	Ένα σχόλιο για το υποκείμενο.	rdfs:Resource	rdfs:Resource
rdfs:seeAlso	Επί πλέον πληροφορίες για το υποκείμενο.	rdfs:Resource	rdfs:resource

Table 7: Ιδιότητες που χρησιμοποιούν τα λεξιλόγια RDF

λεξιλόγια της RDF δεν περιλαμβάνουν όρους που περιγράφουν τις κλάσεις και τις ιδιότητές των κλάσεων συγκεκριμένων εφαρμογών, αλλά παρέχουν τον τρόπο για να περιγράψουμε αυτές τις κλάσεις και τις ιδιότητές τους. Μπορεί για παράδειγμα να προσδιορίσει την ιεραρχία των κλάσεων μιας οντολογίας ή να δηλώσει ότι ένας πόρος ανήκει σε μία κλάση. Για να περιγράψουμε όμως τις ιδιότητες ενός πόρου, όπως παραδείγματος χάριν τον τίτλο ενός βιβλίου, χρησιμοποιούμε άλλα πιο εξειδικευμένα λεξιλόγια.

Οι πόροι στα λεξιλόγια RDF διαθέτουν δικά τους URIs και συγκεκριμένα το λεξιλόγιο RDF το URI "http://www.w3.org/1999/02/22-rdf-syntax-ns" που σχετίζεται με το πρόθεμα rdf: και το RDF Schema το URI "http://www.w3.org/2000/01/rdf-schema" που σχετίζεται με το πρόθεμα rdfs:.

Στον πίνακα Table 6 φαίνονται μερικές από τις κλάσεις που χρησιμοποιούν τα λεξιλόγια της RDF.

Στον πίνακα Table 7 φαίνονται κάποιες από τις ευρέως χρησιμοποιούμενες ιδιότητες που περιγράφονται από τα λεξιλόγια της RDF.

3.3.2 Dublin Core

Το Dublin Core (DC), εδραιώθηκε το 1995 σε ένα συνέδριο που οργανώθηκε στις Ηνωμένες Πολιτείες και συγκεκριμένα στο Dublin του Ohio από τον οργανισμό OCLC (Online Computer Library Centre, Inc.) και το Εθνικό Κέντρο για Εφαρμογές Υπερυπολογιστών (National Centre for Supercomputing Applications).ref www.*****

Οι στόχοι του εγχειρήματος του Dublin Core ήταν:[21]

- Απλότητα στην δημιουργία και στην συντήρηση του συνόλου των όρων.
- Κατανοητή σημασιολογία.
- Συμμόρφωση με άλλα υπάρχοντα ή αναδυόμενα πρότυπα.
- Εφαρμογή σε παγκόσμιο επίπεδο.
- Επεκτασιμότητα.
- Διαλειτουργικότητα μεταξύ διαφόρων συλλογών ή συστημάτων ευρετηριοποίησης.

Χρησιμοποιείται για να περιγράφει πόρους που βρίσκονται στον παγκόσμιο ιστό όπως για παράδειγμα εικόνες, ιστοσελίδες, πολυμέσα και φυσικούς πόρους όπως για παράδειγμα βιβλία, άρθρα ή έργα τέχνης.

Το Dublin core βασίζεται στην περιγραφή με βάση την γλώσσα RDF και περιλαμβάνει δύο χώρους ονομάτων (namespaces) με URIs <http://purl.org/dc/terms/> και <http://purl.org/dc/elements/1.1>.

Οι κλάσεις που χρησιμοποιεί το Dublin Core είναι:

Agent, AgentClass, BibliographicResource, FileFormat, Frequency, Jurisdiction, LicenseDocument, LinguisticSystem, Location, LocationPeriodOrJurisdiction, MediaType, MediaType-OrExtent, MethodOfAccrual, MethodOfInstruction, PeriodOfTime, PhysicalMedium, PhysicalResource, Policy, ProvenanceStatement, RightsStatement, SizeOrDuration, Standard.

Οι ιδιότητες στο χώρο ονομάτων /terms/ είναι οι:

abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, contributor, coverage, created, creator, date, dateAccepted, dateCopyrighted, dateSubmitted, description, educationLevel, extent, format, hasFormat, hasPart, hasVersion, identifier, instructionalMethod, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, language, license, mediator, medium, modified, provenance, publisher, references, relation, replaces, requires, rights, rightsHolder, source, spatial, subject, tableOfContents, temporal, title, type και valid. ενώ οι ιδιότητες στο χώρο ονομάτων /elements/1.1 είναι οι:

contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title και type.

3.3.3 Bibliographic Ontology

Το λεξιλόγιο Bibliographic Ontology (BIBO) ξεκίνησε από τους Frédéric Giasson και Bruce D'Arcus σε μια προσπάθεια να περιγράψουν και να συσχετίσουν βιβλιογραφικές αναφορές. Η οντολογία περιγράφει βιβλιογραφικά αντικείμενα στο σημασιολογικό ιστό σε γλώσσα RDF. Μπορεί να χρησιμοποιηθεί σε διάφορους ρόλους όπως η ταξινόμηση εγγράφων, η περιγραφή βιβλιογραφικών αναφορών και παραπομπών και γενικότερα στην περιγραφή οποιοδήποτε έγγραφου σε γλώσσα RDF.

Το λεξιλόγιο αυτό είναι βασισμένο στη γλώσσα RDF διότι οι απαιτήσεις για την περιγραφή ενός τέτοιου χώρου όπως τα βιβλιογραφικά δεδομένα είναι τόσο μεγάλες που ένα αυτόνομο πρότυπο θα ήταν πολύ δύσκολο να τις καλύψει. Με τη χρήση της RDF το λεξιλόγιο αποκτά δυνατότητα επεκτασιμότητας και διασύνδεσης με οποιοδήποτε άλλο προϋπάρχουσα οντολογία RDF όπως η Dulin Core και η FOAF.

Ο χώρος ονομάτων της οντολογίας BIBO είναι το URI "<http://purl.org/ontology/bibo/>". Οι κλάσεις που χρησιμοποιεί η οντολογία BIBO είναι:

Κλάσεις: Book, Document, Article, AudioVisualDocument, Thesis, Series, AudioDocument, Image, Manuscript, Map, Conference, Collection, Collection, Slideshow, AcademicArticle, Proceedings, DocumentStatus, MultiVolumeBook, Patent, Website, Periodical, Periodic Statuteal, BookSection, Report, Event, Film, Issue, LegalDecision, Slide, Bill, Chapter, Code, Journal, DocumentPart, LegalDocument, Legislation, Letter, Magazine, Magazine, CourtReporter, CourtReporter, LegalCaseDocument, Quote, Standard, Workshop, Excerpt, Interview, Manual, CollectedDocument, CollectedDocument, CollectedDocument, PersonalCommunication-Document, ReferenceSource, ThesisDegree, PersonalCommunication.

Οι ιδιότητες που χρησιμοποιεί η συγκεκριμένη οντολογία είναι:

Ιδιότητες: Issue, presentedAt, editorList, uri, contributorList, isbn10, pageEnd, pageStart, suffixName, degree, isbn, authorList, volume, edition, abstract, issn, identifier, editor, eissn, lccn, number, pages, annotates, content, interviewee, locator, producer, asin, chapter, coden, affirmedBy, citedBy, doi, eanucc13, numPages, pmid, presents, translationOf, upc, court, distributor, interviewer, issuer, performer, prefixName, reversedBy, shortTitle, transcriptOf, translator, cites, gtin14, director, handle, numVolumes, organizer, owner, reproducedIn, section, shortDescription, sici, subsequentLegalDecision.

3.3.4 FOAF

Το όνομα της FOAF προκύπτει από το ακρωνύμιο των λέξεων Friend Of A Friend. Είναι μια οντολογία που ξεκίνησε να αναπτύσσεται στα μέσα του 2000 από τους Libby Miller και Dan Brickley και χρησιμοποιείται για να περιγράψει ανθρώπους, τις δραστηριότητές τους και τις σχέσεις μεταξύ τους. Η τεχνολογία κατασκευής της χρησιμοποιεί τις βασικές αρχές τριών διαφορετικών ειδών δικτύων. Τα κοινωνικά συνεργατικά δίκτυα, τα αντιπροσωπευτικά δίκτυα που αποδίδουν μια απλοποιημένη μορφή ενός πολύπλοκου συστήματος και τα δίκτυα πληροφορίας που χρησιμοποιούν την τεχνολογία του Παγκόσμιου Ιστού για να παρουσιάζουν πληροφορίες από διάφορες ανεξάρτητες πηγές.

Χωρίς να κάνει την ίδια δουλειά με τα κοινωνικά δίκτυα, μπορεί να παρέχει έναν τρόπο περιγραφής τους χρησιμοποιώντας τη γλώσσα RDF, και την OWL. Οι εφαρμογές μπορούν χρησιμοποιώντας τα προφίλ της FOAF να βρουν και να ομαδοποιήσουν άτομα που πληρούν ορισμένες συνθήκες όπως για παράδειγμα τα πρόσωπα που γνωρίζει κάποιος ή τα άτομα που δουλεύουν σε συγκεκριμένη επιχείρηση. Αυτό επιτυγχάνεται αποδίδοντας ένα ID σε κάθε προφίλ και μετά με τον ορισμό των σχέσεων μεταξύ των διαφόρων προσώπων.

Μερικές από τις κλάσεις της FOAF είναι:

Agent, Document, Group, Image, LabelProperty, OnlineAccount, OnlineChatAccount, OnlineEcommerceAccount, OnlineGamingAccount, Organization, Person, PersonalProfileDocument, Project.

Μερικές από τις ιδιότητες της FOAF είναι:

account, accountName, accountServiceHomepage, age, aimChatID, basednear, birthday, currentProject, depiction, depicts, dnaChecksum, familyName, familyname, firstName, focus, fundedBy, geekcode, gender, givenName, givenname, holdsAccount, homepage, icqChatID, img, interest, isPrimaryTopicOf, jabberID, knows, lastName, logo, made, maker, mbox, mboxsha1sum, member, membershipClass, msnChatID, myersBriggs, name, nick, openid, page, pastProject, phone, plan, primaryTopic, publications, schoolHomepage, sha1, skypeID, status, surname, theme, thumbnail, tipjar, title, topic, topicinterest, weblog, workInfoHomepage, workplaceHomepage, yahooChatID.

4 Περιγραφή Υλοποίησης

4.1 Εργαλεία που χρησιμοποιήθηκαν

Τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση της οντολογίας είναι: η γλώσσα προγραμματισμού Python 3 με την οποία δημιουργήθηκε ο κώδικας που είχε σαν τελικό αποτέλεσμα το αρχείο με τις τριάδες RDF, το Open Link Virtuoso στο οποίο δημοσιεύθηκαν τα RDF δεδομένα, και η γλώσσα SPARQL με τη βοήθεια της οποίας δημιουργήθηκαν τα ερωτήματα. Παρακάτω παρατίθεται μια μικρή περιγραφή αυτών των εργαλείων.

4.1.1 Python3

Η Python είναι μια διερμηνευόμενη γλώσσα υψηλού επιπέδου με ανεπτυγμένη σημασιολογία (Semantics). Η δημιουργία της ξεκίνησε το 1989 από τον Ολλανδό Guido van Rossum στο ερευνητικό κέντρο Centrum Wiskunde and Informatica. Θεωρείται διάδοχος της γλώσσας προγραμματισμού ABC και αρχικά χρησιμοποιήθηκε ως γλώσσα σεναρίων για το κατανεμημένο λειτουργικό σύστημα Amoeba. Η πρώτη έκδοση κυκλοφόρησε το 1991. Η ονομασία της οφείλεται στη δημοφιλή κωμική σειρά Monty Python's Flying Circus του BBC της Μεγάλης Βρετανίας (δεκαετία του '70).

Τα κυριότερα χαρακτηριστικά της είναι:

1) η ευκολία στην εκμάθηση, την αναγνωσιμότητα και τη συντήρηση του κώδικα.

Θεμελιώδης ιδέα από τα πρώτα βήματα της ανάπτυξης της γλώσσας ήταν το να είναι εύκολη στην εκμάθηση, πράγμα που για να γίνει χρειάζεται να έχει απλή σύνταξη, το οποίο οδηγεί στην εύκολη αναγνωσιμότητα του κώδικα και επομένως στην εύκολη αποσφαλμάτωση και συντήρησή του.

2) η εύκολη ανάπτυξη εφαρμογών.

Τα προηγούμενα χαρακτηριστικά που αναφέρθηκαν οδηγούν στη γρήγορη ανάπτυξη εφαρμογών, ειδικά σε σχέση με τα προγράμματα που είναι γραμμένα σε C, C++ ή Java. Βέβαια αυτό γίνεται ως ένα βαθμό σε βάρος της ταχύτητας εκτέλεσης των προγραμμάτων όμως τελευταία υπάρχει η τάση να προτιμάται η αποδοτικότητα του προγραμματιστή σε σχέση με την αποτελεσματικότητα του προγράμματος.

3) οι υψηλού επιπέδου δομές δεδομένων.

Η Python βασίζεται σε μερικές πολύ αποδοτικές δομές δεδομένων όπως οι λίστες, οι πλειάδες (tuples), τα σύνολα (sets) και τα λεξικά (dictionaries). Αυτό έχει σαν συνέπεια να κάνει μια αποτελεσματική προσέγγιση στον αντικειμενοστραφή προγραμματισμό χωρίς να αποκλείει και άλλες μορφές προγραμματισμού όπως τον διαδικαστικό ή το συναρτησιακό.

4) η ανοικτότητα του κώδικα.

Ο κώδικάς της είναι λογισμικό ανοικτού κώδικα, δηλαδή μπορεί ο καθένας να έχει πρόσβαση σε αυτόν, να τον αλλάζει, να τον αντιγράφει και να τον διανέμει χωρίς κόστος. Αναπτύσσεται με βάση τις προδιαγραφές της κοινότητας των προγραμματιστών της και ο συντονισμός και η διαχείριση γίνεται από τον μη κερδοσκοπικό οργανισμό Python Software Foundation (PSF) (www.python.org/psf).

5) η επεκτασιμότητα.

Το μοντέλο της Python σχεδιάστηκε με τέτοιο τρόπο ώστε να είναι ευέλικτα επεκτάσιμο, να παρέχει ενσωματωμένα στοιχεία (εντολές, τύπους αντικειμένων, κ.λπ.) αλλά και να δίνει

τη δυνατότητα στους προγραμματιστές να προσθέτουν τα δικά τους στοιχεία ανάλογα με τις ανάγκες τους και το σύστημα που χρησιμοποιούν. Αποτελείται από ένα βασικό πυρήνα ο οποίος πλαισιώνεται από αρθρώματα (modules) και πακέτα (packages). Διαθέτει πολλές έτοιμες βιβλιοθήκες που είναι πολύ εύκολες στη χρήση και οι οποίες μπορεί πολύ εύκολα να επεκταθούν ακόμα και σε γλώσσα C ή C++. Όλα αυτά σε συνδυασμό με την ανοικτότητά της εξασφαλίζουν το ότι εξελίσσεται και παρακολουθεί όλες τις αλλαγές και τις ανάγκες που παρουσιάζονται στο χώρο του προγραμματισμού.

6) η ωριμότητα.

Η εμφάνισή της έγινε στα τέλη της δεκαετίας του 1980 οπότε μέσα στο πέρασμα των χρόνων λύθηκαν πάρα πολλά προβλήματα που παρουσίαζε και εμπλουτίστηκε με πολλές βιβλιοθήκες.

7) η λειτουργία σε οποιοδήποτε περιβάλλον.

Λειτουργεί σε όλα τα λειτουργικά όπως Linux, Windows, OS/2, MAC, Amiga και συνεργάζεται με άλλες γλώσσες προγραμματισμού, χρησιμοποιώντας τις βιβλιοθήκες τους, όπως τη JAVA ή τη C/C++ με την οποία συνδέεται μέσω αρθρωμάτων.

8) η αποφυγή των segmentation faults.

Σε αντίθεση με άλλες γλώσσες προγραμματισμού ο Interpreter της Python ενημερώνει με μία εξαίρεση που πετάει (throw) για το σε ποια γραμμή είναι το σφάλμα ώστε να διορθωθεί άμεσα.

9) η αυτόματη διαχείριση μνήμης.

Αυτό σημαίνει ότι ο προγραμματιστής δεν ασχολείται με τη διαχείριση της μνήμης που δεσμεύει για τις μεταβλητές και τα διάφορα αντικείμενα που δημιουργεί. Όταν πάψει να χρησιμοποιείται μια μεταβλητή η Python αποδεσμεύει μόνη της το μέρος της μνήμης που χρειάστηκε. Επίσης όταν ένα αντικείμενο χρησιμοποιείται πάνω από μία φορά, το αναγνωρίζει και δεν το αποθηκεύει πάλι στη μνήμη (reference counting).

Τα παραπάνω χαρακτηριστικά την έχουν κάνει μια πολύ δημοφιλή γλώσσα μεταξύ τόσο των αρχάριων όσο και των έμπειρων προγραμματιστών και χρησιμοποιείται τόσο για εκπαιδευτικούς σκοπούς όσο και για την ανάπτυξη ολοκληρωμένων εφαρμογών.[22] [23]

4.1.2 Open Link Virtuoso

Ο Virtuoso ξεκίνησε σαν ένας εξυπηρετητής βασισμένος σε πολλά πρωτόκολλα που παρείχε ODBC/JDBC πρόσβαση σε σχεσιακά δεδομένα αποθηκευμένα είτε μέσω του Virtuoso είτε μέσω σύνδεσης με κάποια εξωτερική σχεσιακή βάση δεδομένων. Εκτός από την εξυπηρέτηση πελατών που απευθύνουν SQL ερωτήματα, ο Virtuoso έχει ενσωματωμένο HTTP εξυπηρετητή που παρέχει αποθετήριο DAV, endpoints που βασίζονται σε SOAP και WS πρωτόκολλα και δυναμικές ιστοσελίδες χρησιμοποιώντας μια ποικιλία γλωσσών σεναρίων (script languages).

Αρχικά τα RDF δεδομένα αποθηκεύονταν σε σχεσιακή βάση δεδομένων, όμως η ταχύτατη ανάπτυξη του σημασιολογικού ιστού δημιούργησε την ανάγκη να συμπεριληφθούν στον Virtuoso λειτουργίες που εφαρμόζονται απευθείας πάνω στα RDF δεδομένα. Γι' αυτό το λόγο προμηθευτές λογισμικού όπως η Oracle ή η OpenLink Software πρόσθεσαν λειτουργίες για δομένα απευθείας σε RDF (native RDF).[24]

Έτσι ο Virtuoso Universal Server απέκτησε μια υβριδική αρχιτεκτονική που βοηθάει στην αποθήκευση διαφόρων μοντέλων δεδομένων όπως σχεσιακά, RDF, XML και έγγραφα κειμένου. Μπορεί να τρέχει σαν ένας εξυπηρετητής για έγγραφα και εφαρμογές του ιστού αλλά και

για συνδεδεμένα δεδομένα. Η ανοιχτή έκδοση του Virtuoso περιλαμβάνεται στο σύνολο των τεχνολογιών που προτείνει το LOD2 για το χώρο των ανοιχτών συνδεδεμένων δεδομένων και πολλοί χρήστες τον χρησιμοποιούν για να φορτώνουν τα δεδομένα τους στο "Quad store" του. Επίσης περιλαμβάνει ένα SPARQL endpoint και ένα μηχανισμό, τον URL-Rewriter, με τον οποίο φτιάχνει επισκέψιμα URIs (dereferenceable URIs) [25].

Σήμερα οι μεγαλύτερες εφαρμογές του Virtuoso είναι στο χώρο των RDF δεδομένων για τη διαχείριση terrabytes RDF τριάδων που συνήθως δεν χωράνε στη RAM. Έτσι στο ερευνητικό πεδίο υπάρχει μεγάλο ενδιαφέρον για αποτελεσματικές τεχνικές συμπίεσης των δεδομένων αλλά και για ευέλικτες τεχνικές διαχείρισης που καλύπτουν τόσο τα RDF όσο και τα σχεσιακά δεδομένα. [26]

4.1.3 SPARQL

Η SPARQL είναι μία από τις πλέον διαδεδομένες τεχνολογίες των συνδεδεμένων δεδομένων και αποτελεί ένα σύνολο από προδιαγραφές για ανάκτηση και διαχείριση δεδομένων που είναι αποθηκευμένα σε μορφή RDF. Το όνομά της προέρχεται από το ακρωνύμιο του SPARQL πρωτοκόλλου και της RDF Query Language (Γλώσσα Ερωτημάτων RDF). Η τελευταία έκδοση που κυκλοφορεί είναι η SPARQL 1.1 που εκδόθηκε το 2013 σαν οδηγία του οργανισμού W3C (World Wide Web Consortium). Το πρότυπο SPARQL 1.1 περιλαμβάνει τις ακόλουθες προδιαγραφές:

1. Γλώσσα ερωτημάτων SPARQL 1.1 (SPARQL 1.1 Query Language)
Μια γλώσσα ερωτημάτων για δεδομένα RDF.
2. Μορφότυπα αποτελεσμάτων αναζήτησης σε SPARQL 1.1 (SPARQL 1.1 Query Results JSON Format και SPARQL 1.1 Query Results CSV and TSV Formats)
Επιτρέπει την έκφραση των αποτελεσμάτων σε: JavaScript Object Notation - JSON), comma-separated values - CSV και tab-separated values - TSV.
3. Ομόσπονδα ερωτήματα SPARQL 1.1 (SPARQL 1.1 Federated Query)
Επιτρέπει την εκτέλεση ερωτημάτων σε κατανεμημένα SPARQL endpoints.
4. Συστήματα συμπερασματολογίας SPARQL 1.1 (SPARQL 1.1 Entailment Regimes)
Προδιαγραφές που ορίζουν τη σημασιολογία ερωτημάτων SPARQL υπό το πρίσμα συστημάτων συμπερασματολογίας, όπως τα RDF Schema, OWL, Rule Interchange Format - RIF.
5. Γλώσσα ενημερώσεων SPARQL 1.1 (SPARQL 1.1 Update Language)
Μια γλώσσα ενημερώσεων (update) για γράφους RDF.
6. Πρωτόκολλο SPARQL 1.1 για RDF (SPARQL 1.1 Protocol for RDF)
Ένα πρωτόκολλο που ορίζει τρόπους μετάδοσης ερωτημάτων και ενημερώσεων SPARQL σε μια υπηρεσία SPARQL.
7. Περιγραφή υπηρεσιών SPARQL 1.1 (SPARQL 1.1 Service Description)
Προδιαγραφές που ορίζουν μια μέθοδο για ανακάλυψη και ένα λεξιλόγιο για περιγραφή υπηρεσιών SPARQL.
8. Πρωτόκολλο HTTP για αποθετήρια γράφου SPARQL 1.1 (SPARQL 1.1 Graph Store HTTP Protocol)
Οι προδιαγραφές αυτές ορίζουν ένα τρόπο χειρισμού δεδομένων σε ένα αποθετήριο τριάδων RDF μέσω λειτουργιών HTTP.
9. Σουίτα ελέγχων SPARQL 1.1 (SPARQL 1.1 Test Cases)
Μια σουίτα ελέγχων για το αν ένα σύστημα πληροί τις προδιαγραφές SPARQL 1.1.[13]

Παρακάτω αναπτύσσονται συνοπτικά τα κυριότερα χαρακτηριστικά της γλώσσας ερωτημάτων SPARQL.

Ουσιαστικά η SPARQL είναι μια γλώσσα ερωτημάτων που λειτουργεί με ταιριάσματα πάνω σε ένα γράφο. Δεδομένης μίας πηγής δεδομένων D, ένα ερώτημα αποτελείται από ένα σχέδιο που αντιστοιχείται στην D και οι τιμές που λαμβάνονται από αυτή την αντιστοίχιση επιδέχονται μιας επεξεργασίας για να δοθεί η απάντηση. Τα βασικά δομικά συστατικά ενός SPARQL ερωτήματος είναι: οι δηλώσεις προθεμάτων που χρησιμεύουν στη συντομογραφία των χρησιμοποιούμενων URIs, το σχέδιο ταιριάσματος που περιλαμβάνει επιλογές όπως ένωση μοτίβων (union), επιλεκτικά ταιριάσματα (optional), εφαρμογή φίλτρων στις επιστρεφόμενες τιμές και δυνατότητα επιλογής της πηγής δεδομένων και οι μετατροπές των αποτελεσμάτων που λειτουργούν, εφαρμόζοντας πάνω στις τιμές που επιστρέφονται σαν αποτέλεσμα, κλασσικούς τελεστές όπως "projection", "distinct", "order", "limit" και "offset". Τέλος η έξοδος ενός SPARQL ερωτήματος μπορεί να είναι διαφόρων τύπων όπως: ναι/οχι, επιλογή τιμών που ταιριάζουν στο μοτίβο, διαγραφή ή κατασκευή νέων τριάδων και περιγραφές πόρων.[27] [13]

4.2 Αναζήτηση και συλλογή των βιβλιογραφικών δεδομένων σε μορφή bibtex.

Μετά την επιλογή των εργαλείων που επρόκειτο να χρησιμοποιηθούν έπρεπε να βρεθεί το σύνολο των δεδομένων πάνω στο οποίο θα κατασκευάζονταν η οντολογία σύμφωνα με τη μεθοδολογία που προέκυψε από τη θεωρητική μελέτη.

Μετά από σχετική έρευνα στο διαδίκτυο εντοπίστηκε ένα μεγάλο σύνολο βιβλιογραφικών δεδομένων που αφορά τις δημοσιεύσεις των μελών του τμήματος Πληροφορικής του Ιονίου Πανεπιστημίου. Τα δεδομένα αυτά εντοπίστηκαν στον εξυπηρετητή github και συγκεκριμένα στη διεύθυνση https://github.com/ioniodi/scholari/tree/master/_bibtex. Η μορφή των δεδομένων ήταν σύμφωνα με το μορφότυπο Bibtex και προτιμήθηκε γιατί είναι ένας ανοικτός μορφότυπος περιγραφής βιβλιογραφικών αναφορών που χρησιμοποιείται ευρέως από την επιστημονική κοινότητα για την αποθήκευση στοιχείων βιβλιογραφίας. Επί πλέον διαθέτει μια πολύ συγκεκριμένη δομή που καθιστά το διάβασμα του bibtex αρχείου και την οργάνωση των δεδομένων σε άλλη μορφή, από μια εφαρμογή γραμμένη σε μία γλώσσα προγραμματισμού όπως για παράδειγμα η Python που χρησιμοποιήθηκε στην παρούσα εργασία, διαδικασίες αρκετά γρήγορες.

Στο github τα δεδομένα αυτά ήταν γραμμένα σε ξεχωριστό αρχείο για κάθε καθηγητή οπότε έγινε αντιγραφή όλων των δημοσιεύσεων σε ένα αρχείο με όνομα papers.bib. Τα πεδία που υπήρχαν στο συγκεκριμένο bibtex αρχείο ήταν τα "author", "title", "journal" ή "booktitle", "publisher", "volume", "number", "pages", "year", "url", "doi", "timestamp", "biburl" και "bibsourc" από τα οποία για τις ανάγκες της οντολογίας χρησιμοποιήθηκαν τα "author", "title", "journal" ή "booktitle", "publisher", "year" και "url". Στην εικόνα Figure 9 φαίνεται μία εγγραφή του αρχείου papers.bib.

4.3 Απόδοση URIs στις οντότητες.

Από τα παραπάνω πεδία που χρησιμοποιήθηκαν στην οντολογία, αυτά που αντιστοιχούν σε οντότητες είναι τα "Author", "journal", "booktitle" και "publisher" και όπως αναφέρθηκε

```
@article{DBLP:journals/nc/GiannakisA17,
  author    = {Konstantinos Giannakis and
               Theodore Andronikos},
  title     = {Membrane automata for modeling biomolecular processes},
  journal   = {Natural Computing},
  volume    = {16},
  number    = {1},
  pages     = {151--163},
  year      = {2017},
  url       = {https://doi.org/10.1007/s11047-015-9518-1},
  doi       = {10.1007/s11047-015-9518-1},
  timestamp = {Thu, 08 Jun 2017 01:00:00 +0200},
  biburl    = {http://dblp.org/rec/bib/journals/nc/GiannakisA17},
  bibsource = {dblp computer science bibliography, http://dblp.org}
}
```

Figure 9: Μια εγγραφή σε μορφή Bibtex.

παραπάνω έπρεπε να τους αποδοθούν μοναδικά αναγνωριστικά ταυτοποίησης, δηλαδή URIs.

Για να γίνει αυτό ακολουθήθηκε η εξής διαδικασία: Με κώδικα σε Python δημιουργήθηκε ένα set για κάθε οντότητα το οποίο περιελάμβανε τα στιγμιότυπα της οντότητας και στη συνέχεια αποδόθηκε σε κάθε στοιχείο του set ένας μοναδικός αριθμητικός κωδικός. Στη συνέχεια αποδόθηκαν τα παρακάτω URIs:

Στα στιγμιότυπα της οντότητας "Author" αποδόθηκαν URIs του τύπου "http://sw2.lodbook.org/authors/0004".

Στα στιγμιότυπα της οντότητας "PublishedIn" αποδόθηκαν URIs του τύπου "http://sw2.lodbook.org/journals/0001" για τα περιοδικά και "http://sw2.lodbook.org/conf/0005" για τα συνέδρια.

Στα στιγμιότυπα της οντότητας "Publisher" αποδόθηκαν URIs του τύπου "http://sw2.lodbook.org/publishers/0006".

Μ'αυτόν τον τρόπο κάθε στιγμιότυπο των οντοτήτων της οντολογίας απέκτησε ένα μοναδικό αναγνωριστικό ταυτοποίησης, όπως ορίζουν οι βασικές αρχές δόμησης των ανοιχτών συνδεδεμένων δεδομένων στο σημασιολογικό ιστό.

4.4 Μετατροπή των bibtex δεδομένων σε RDF μέσω Python.

Στη συνέχεια έπρεπε να μετατραπούν τα δεδομένα του papers.bib τα οποία ήταν σε μορφή bibtex σε μορφή RDF.

Η μορφή σύνταξης των τριάδων RDF που χρησιμοποιήθηκε ήταν η N-Triples σύμφωνα με την οποία οι τριάδες αποθηκεύονται σε ένα αρχείο με επέκταση .nt. Η επιλογή της N-Triples έγινε διότι, παρ' όλο που δημιουργείται ένας μεγάλος όγκος από τριάδες, σε μορφή όχι τόσο εύκολα αναγνώσιμη από τον άνθρωπο όσο στην Turtle για παράδειγμα που διαθέτει συντομογραφίες (prefix) και ομαδοποιεί σε μπλοκ τριάδες με το ίδιο υποκείμενο, έχει μια σειρά από πλεονεκτήματα. Έχει σχετικά απλή δομή με κάθε γραμμή να αποτελείται από μία τριάδα "υποκείμενο-κατηγορημα-αντικείμενο" τα οποία χωρίζονται με κενό και τελειώνει με μία τελεία. Αυτό την κάνει πολύ εύκολη στην ανάλυση των δεδομένων από εφαρμογές και υποστηρίζεται από όλες τις βιβλιοθήκες RDF και όλα τα αποθετήρια RDF. Επίσης καθώς η προσπέλαση

```

<http://sw2.lodbook.org/journal/00076> <http://www.w3.org/2000/01/rdf-schema#label> "CoRR".
<http://sw2.lodbook.org/authors/00126> <http://xmlns.com/foaf/0.1/name> "Gerth Stolling Brodal".
<http://sw2.lodbook.org/authors/00340> <http://xmlns.com/foaf/0.1/name> "Alexis C. Kaporis".
<http://sw2.lodbook.org/authors/00171> <http://xmlns.com/foaf/0.1/name> "Apostolos N. Papadopoulos".
<http://sw2.lodbook.org/authors/00060> <http://xmlns.com/foaf/0.1/name> "Spyros Sioutas".
<http://sw2.lodbook.org/authors/00182> <http://xmlns.com/foaf/0.1/name> "Konstantinos Tsakalidis".
<http://sw2.lodbook.org/authors/00335> <http://xmlns.com/foaf/0.1/name> "Kostas Tsichlas".
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/dc/elements/1.1/created> "2012".
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/dc/elements/1.1/title> "Dynamic 3-sided Planar Range Queries with
Logarithmic Time".
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/dc/elements/1.1/type> "article".
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/ontology/bibo/uri> <http://arxiv.org/abs/1201.2702>.
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/ontology/bibo/presentedAt> <http://sw2.lodbook.org/journal/00076>.
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/dc/elements/1.1/creator> <http://sw2.lodbook.org/authors/00126>.
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/dc/elements/1.1/creator> <http://sw2.lodbook.org/authors/00340>.
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/dc/elements/1.1/creator> <http://sw2.lodbook.org/authors/00171>.
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/dc/elements/1.1/creator> <http://sw2.lodbook.org/authors/00060>.
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/dc/elements/1.1/creator> <http://sw2.lodbook.org/authors/00182>.
<http://sw2.lodbook.org/Papers/00327> <http://purl.org/dc/elements/1.1/creator> <http://sw2.lodbook.org/authors/00335>.

```

Figure 10: Οι τριάδες που προκύπτουν από το Bibtex μιας δημοσίευσης.

(streaming) γίνεται ανά γραμμή, αν χαθεί μια γραμμή από το αρχείο, χάνεται μόνο μια τριάδα ενώ στις μορφές σύνταξης τύπου turtle, αν χαθεί ένα μπλοκ, μπορεί να χαθούν ακόμα και χιλιάδες τριάδων. Ακόμη δεν έχει καμία σημασία η σειρά των τριάδων, οπότε ακόμη και αν με διάφορες αλλαγές πειραχτεί αυτή η σειρά, ο γράφος που αναπαριστούν οι τριάδες μένει ανεπηρέαστος ενώ σε μια άλλη μορφή το αρχείο θα έχανε την εγκυρότητά του. Τέλος η συμπαγής και ισχυρή δομή της την καθιστά ιδιαίτερα φιλική σε αποδοτικές τεχνικές συμπίεσης.

Η δημιουργία του .nt αρχείου έγινε με πρόγραμμα που κατασκευάστηκε με τον κώδικα της γλώσσας προγραμματισμού Python. Αρχικά προσπελάστηκε το αρχείο papers.bib με τη χρήση των εντολών της βιβλιοθήκης bibtexparser, που υποστηρίζει την ανάλυση bibtex δεδομένων και τα δεδομένα του γράφηκαν σε ένα άλλο αρχείο με όνομα ionio.bib, με τη μορφή μιας λίστα από λεξικά, όπου κάθε λεξικό περιείχε μία δημοσίευση. Τα κλειδιά (keys) του λεξικού αντιστοιχίστηκαν στα πεδία του αρχείου papers.bib και οι τιμές τους (values) στις τιμές αυτών των πεδίων. Στη συνέχεια διαβάστηκαν τα περιεχόμενα του ionio.bib και τα περιεχόμενά του γράφηκαν, με τρόπο ώστε να υπακούουν στη δομή N-Triples, στο αρχείο triplets.nt.

Στην εικόνα Figure 10 φαίνονται όλες οι τριάδες που σχετίζονται με την δημοσίευση που έχει URI "http://sw2.lodbook.org/Papers/00327"

4.5 Έλεγχος εγκυρότητας του RDF αρχείου.

Στη συνέχεια με το 'riot' ο οποίος είναι ένα εργαλείο που ελέγχει την εγκυρότητα σύνταξης των διάφορων μορφοτύπων σειριοποίησης δεδομένων RDF ελέγχθηκε αν το triplets.nt περιείχε μη έγκυρους χαρακτήρες και αν η μορφή όλων των τριάδων που περιείχε πληρούσαν τις προδιαγραφές της μορφής N-Triples. Μετά από τις απαραίτητες διορθώσεις, όπως για παράδειγμα την απαλοιφή του χαρακτήρα 'escape sequence' που περιείχαν αρκετά ονόματα συγγραφέων στο bibtex αρχείο, το triplets.nt πήρε την έγκυρη μορφή που ήταν απαραίτητη για το επόμενο βήμα.

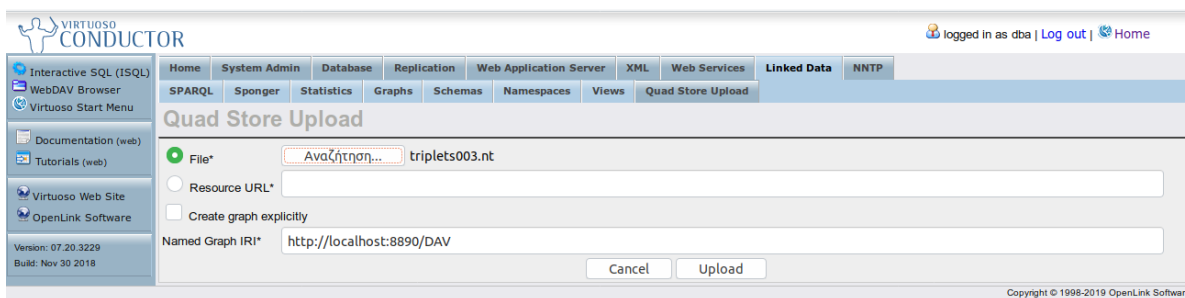


Figure 11: Η καρτέλα Quad Store Upload του Virtuoso.

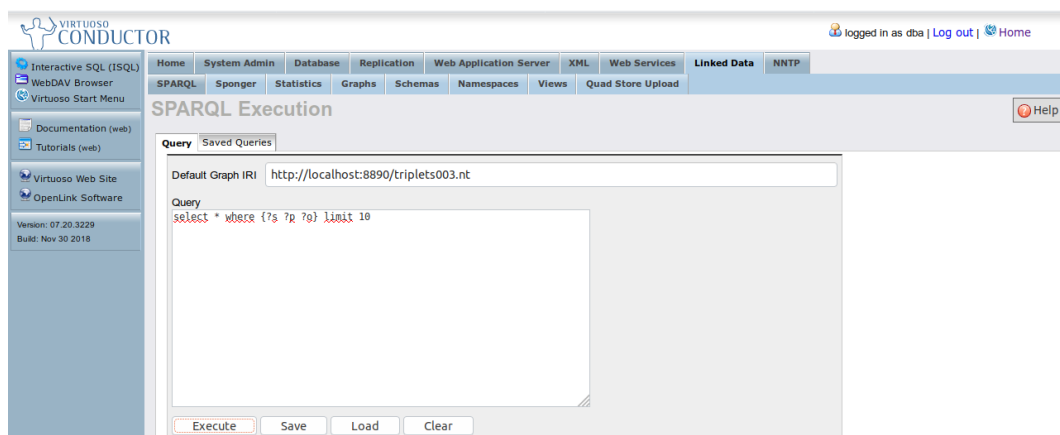


Figure 12: Η καρτέλα SPARQL του Virtuoso.

4.6 Φόρτωση του αρχείου στο Open Link Virtuoso και διατύπωση ερωτημάτων με τη γλώσσα SPARQL.

Αφού έγινε και ο έλεγχος της εγκυρότητας του αρχείου triplets.nt χρησιμοποιήθηκε η εφαρμογή OpenLink Virtuoso που όπως αναφέρθηκε παραπάνω διαθέτει μεταξύ άλλων και τελικό σημείο SPARQL. Η σύνδεση έγινε στον Virtuoso conductor με όνομα χρήστη dba σαν superuser. Για να φορτωθούν τα δεδομένα, από την καρτέλα "Linked Data" επιλέχθηκε η καρτέλα "Quad Store Upload" η οποία φαίνεται στην εικόνα Figure 11 όπου δόθηκε στο πεδίο "File" το path και το όνομα του αρχείου triplets.nt και έγινε "Upload".

Στη συνέχεια έγινε μετάβαση στην καρτέλα "SPARQL". Στο πλαίσιο "Default Graph IRI" δόθηκε το IRI του γράφου, δηλαδή "http://localhost:8890/triplets.nt". Στο πλαίσιο "Query" διατυπώθηκαν τα ερωτήματα σε σύνταξη SPARQL και πατώντας το κουμπί "Execute" έγινε η εμφάνιση των αποτελεσμάτων. Στην εικόνα Figure 12 φαίνεται η καρτέλα "SPARQL" με το όνομα του γράφου και ένα ερώτημα, ενώ στην εικόνα Figure 13 φαίνονται τα αποτελέσματα του συγκεκριμένου ερωτήματος όπως τα εμφανίζει ο Virtuoso.

Μερικά ενδεικτικά ερωτήματα που μπορεί να απευθύνει κανείς στο συγκεκριμένο γράφο, με τη σύνταξη σε SPARQL και τα αποτελέσματά τους (Figure 14, Figure 15, Figure 16) φαίνονται παρακάτω:

s	p	o
DBLP:journals/eaai/GiannakisTPAV16	http://purl.org/dc/terms/created	"2016"
DBLP:conf/algocloud/2015	http://purl.org/dc/terms/created	"2016"
DBLP:conf/algocloud/GiannakopoulosK16	http://purl.org/dc/terms/created	"2016"
DBLP:conf/algocloud/KarydisSAMK16	http://purl.org/dc/terms/created	"2016"
DBLP:conf/algocloud/KokotinisKNRSTT16	http://purl.org/dc/terms/created	"2016"
DBLP:conf/algocloud/KorasidisGDTK16	http://purl.org/dc/terms/created	"2016"
DBLP:conf/algocloud/NodarakisRSTTTP16	http://purl.org/dc/terms/created	"2016"
DBLP:conf/algocloud/SioutasMPGVKSK16	http://purl.org/dc/terms/created	"2016"
DBLP:conf/bigdataconf/DokaPGTK16	http://purl.org/dc/terms/created	"2016"
DBLP:conf/bigdataconf/GiannakourisPTK16	http://purl.org/dc/terms/created	"2016"

Figure 13: Τα αποτελέσματα ενός ερωτήματος στον Virtuoso.

onoma	titlos
"Lazaros S. Iliadis"	"Artificial Intelligence Applications and Innovations - {AIAI} 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece,
"Harris Papadopoulos"	"Artificial Intelligence Applications and Innovations - {AIAI} 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece,
"Lazaros S. Iliadis"	"Artificial Intelligence Applications and Innovations - {AIAI} 2012 International Workshops: AIAB, AIEIA, CISE, COPA, IIVC,
"Harris Papadopoulos"	"Artificial Intelligence Applications and Innovations - {AIAI} 2012 International Workshops: AIAB, AIEIA, CISE, COPA, IIVC,

Figure 14: Αποτελέσματα ερωτήματος 1.

Ερώτημα 1: "Να εμφανίσεις τις δημοσιεύσεις των συγγραφέων Λάζαρου Ηλιάδη και Χάρη Παπαδόπουλου, που έχουν δημοσιευθεί στο "Springer".

```

PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
PREFIX foaf: http://xmlns.com/foaf/0.1/
PREFIX bibo: http://purl.org/ontology/bibo/
PREFIX dc: http://purl.org/dc/terms/

```

```

select ?onoma ?titlos
where
{
  ?paper dc:creator ?auth.
  ?paper dc:title ?titlos.
  ?auth foaf:name ?onoma.
  FILTER (?onoma = "Lazaros S. Iliadis" OR ?onoma = "Harris Papadopoulos")
  ?paper dc:publisher ?pub.
  ?pub rdfs:label "Springer".
}

```

Ερώτημα 2: "Να εμφανίσεις τις δημοσιεύσεις του καθηγητή Μ. Στεφανιδάκη, που έχουν δημοσιευθεί ως άρθρα στο περιοδικό Journal in Computer Virology.

```

PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
PREFIX foaf: http://xmlns.com/foaf/0.1/
PREFIX bibo: http://purl.org/ontology/bibo/
PREFIX dc: http://purl.org/dc/terms/

```

```

select ?titlos
where
{
  ?paper dc:creator ?auth.
  ?paper dc:title ?titlos.
  ?auth foaf:name ?onoma.
  FILTER (regex (?onoma, "Michalis Stefanidakis"))
  ?paper bibo:presentedAt ?journ.
  ?journ rdfs:label ?journal.
  FILTER (regex(?journal, "Journal in Computer Virology"))
}

```

titlos
"A spatial stochastic model for worm propagation: scale effects"
"Treating scalability and modelling human countermeasures against local preference worms via gradient models"

Figure 15: Αποτελέσματα ερωτήματος 2.

}

Ερώτημα 3: "Να εμφανίσεις τον τίτλο των δημοσιεύσεων και το όνομα των συγγραφέων που δημοσίευσαν άρθρα σε περιοδικά το "2006".

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

PREFIX foaf: <http://xmlns.com/foaf/0.1/>

PREFIX bibo: <http://purl.org/ontology/bibo/>

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

```

select ?onoma ?titlos
where
{
  ?paper dc:creator ?auth.
  ?auth foaf:name ?onoma.
  ?paper dc:title ?titlos.
  ?paper dc:type "article".
  ?paper dc:created ?year.
  FILTER (?year = "2006")
}
```

onoma	titlos
"Rosa Mavropodi"	"Secure distributed intelligent networks"
"George M. Giaglis"	"Wireless value-added services in exhibition shows"
"Pavlos A. Vlachos"	"Drawing Emerging Business Models for the Mobile Music Industry"
"Vassilios Chrissikopoulos"	"Secure distributed intelligent networks"
"Panayiotis Kotzanikolaou"	"Secure distributed intelligent networks"
"Christos Douligeris"	"Secure distributed intelligent networks"
"Diomidis Spinellis"	"Wireless value-added services in exhibition shows"
"Adam P. Vrechopoulos"	"Drawing Emerging Business Models for the Mobile Music Industry"
"Adamantia G. Pateli"	"Wireless value-added services in exhibition shows"
"Adamantia G. Pateli"	"Drawing Emerging Business Models for the Mobile Music Industry"

Figure 16: Αποτελέσματα ερωτήματος 3.

5 Συμπεράσματα και μελλοντικές εργασίες

Η εργασία αυτή είχε σαν στόχο να παρουσιάσει μια τεχνική παραγωγής βιβλιογραφικών δεδομένων σε μορφή RDF και κάποια εργαλεία χρήσης τους. Παίρνοντας υπ' όψιν την αναγκαιότητα της διασύνδεσης των δεδομένων στο χώρο των επιστημονικών δημοσιεύσεων, επιλέχθηκε το σύνολο των δεδομένων να προέρχεται από τις δημοσιεύσεις του διδακτικού προσωπικού του τμήματος πληροφορικής του Ιονίου Πανεπιστημίου.

Το συμπέρασμα που συνάγεται από αυτή την προσπάθεια μετατροπής των βιβλιογραφικών δεδομένων του Ιονίου Πανεπιστημίου σε RDF είναι ότι από τη στιγμή που αυτά διατίθενται σε μια καλά δομημένη μορφή όπως η bibtex, η διαδικασία της μετατροπής τους είναι αρκετά γρήγορη. Η δουλειά που έχει γίνει από τους ερευνητές που ασχολούνται με τις τεχνολογίες του σημασιολογικού ιστού έχει οδηγήσει σε αρκετά και αξιόπιστα λεξιλόγια για την περιγραφή των βιβλιογραφικών εγγραφών και με εργαλεία σαν την γλώσσα προγραμματισμού Python 3 που είναι εύχρηστη και ευέλικτη και τον Virtuoso Server που είναι αξιόπιστος στη δημοσίευση και διαχείριση των δεδομένων, το εγχείρημα μετατροπής του συνόλου των εγγραφών των επιστημονικών δημοσιεύσεων του Ιονίου Πανεπιστημίου σε μορφή σύμφωνη με τα πρότυπα του σημασιολογικού ιστού είναι κάτι που στο μέλλον είναι καλό να γίνει.

Σαν μελλοντική εργασία θα μπορούσε να γίνει μια πιο συστηματική μελέτη των λεξιλογίων για μια πιο ολοκληρωμένη οντολογία αναπαράστασης των βιβλιογραφικών εγγραφών με σκοπό τη δημοσίευσή τους στο σημασιολογικό ιστό. Με αυτές τις τεχνικές της ολοκλήρωσης και ενοποίησης των δεδομένων το έργο της επιστημονικής κοινότητας του πανεπιστημίου αποκτά προστιθέμενη αξία μιας και εφοδιάζεται με αυξημένη σημασιολογία, διασύνδεση με άλλες πηγές δεδομένων, εύκολη προσβασιμότητα και αύξηση των ετεροαναφορών από άλλους ερευνητές.

References

- [1] D. SHOTTON, “Semantic publishing: the coming revolution in scientific journal publishing,” *Learned Publishing*, vol. 22, no. 2, pp. 85–94.
- [2] D. Shotton, K. Portwin, G. Klyne, and A. Miles, “Adventures in semantic publishing: Exemplar semantic enhancements of a research article,” *PLOS Computational Biology*, vol. 5, pp. 1–17, 04 2009.
- [3] Y. Sure, S. Bloehdorn, P. Haase, J. Hartmann, and D. Oberle, “The swrc ontology – semantic web for research communities,” in *Progress in Artificial Intelligence* (C. Bento, A. Cardoso, and G. Dias, eds.), (Berlin, Heidelberg), pp. 218–231, Springer Berlin Heidelberg, 2005.
- [4] S. Peroni and D. Shotton, “Fabio and cito: Ontologies for describing bibliographic resources and citations,” *Journal of Web Semantics*, vol. 17, pp. 33 – 43, 2012.
- [5] T. Berners-Lee, “The world wide web - past, present and future,” *Journal of Digital Information*, vol. 1, no. 1, 2006.
- [6] T. Berners-Lee, J. Hendler, and O. Lassila, “The semantic web” in scientific american,” *Scientific American Magazine*, vol. 284, 05 2001.
- [7] A. Hogan, “Linked data and the semantic web standards,” 2014. Διαθέσιμο στο: http://aidanhogan.com/docs/ldmngmt_local.chapter.pdf.
- [8] J. Greenberg, “Metadata and the world wide web,” in *Encyclopedia of Library and Information Science*, pp. 1876–1888, Marcel Dekker, 2003.
- [9] C. Bizer, T. Heath, and T. Berners-Lee, “Linked data - the story so far,” *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.
- [10] L. Sauermann, R. Cyganiak, and M. Völkel, “Cool uris for the semantic web,” Technical Memo TM-07-01, DFKI GmbH, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, February 2007. Written by 29.11.2006.
- [11] T. R. Gruber, “Toward principles for the design of ontologies used for knowledge sharing,” *Int. J. Hum.-Comput. Stud.*, vol. 43, pp. 907–928, Dec. 1995.
- [12] M. M. Taye, “Understanding semantic web and ontologies: Theory and applications,” *CoRR*, vol. abs/1006.4567, 2010.
- [13] S. M., P. I., and A. T., *ΑΝΟΙΚΤΑ ΣΥΝΔΕΔΕΜΕΝΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ Μία Πρακτική Προσέγγιση στον Σημασιολογικό Ιστό*. Kallipos, 2015.
- [14] S. Decker, S. Melnik, and P. Mitra, “Framework for the semantic web: An rdf tutorial,” *IEEE Internet Computing*, vol. 4, pp. 68–73, 11 2000.
- [15] M. G., T. W., and S. M., “A survey of rdf store solutions,” in *2014 International Conference on Engineering, Technology and Innovation (ICE)*, IEEE, 2014.

- [16] B. Haslhofer, E. M. Roochi, B. Schandl, and S. Zander, “Europeana rdf store report,” technical report, University of Vienna, Vienna, March 2011.
- [17] N. K. S. I., “Survey of rdf storage managers,” in *The Sixth International Conference on Advances in Databases, Knowledge, and Data Applications*, DBKDA, 2014.
- [18] P. Haase, J. Broekstra, A. Eberhart, and R. Volz, “A comparison of rdf query languages,” in *The Semantic Web – ISWC 2004* (S. A. McIlraith, D. Plexousakis, and F. van Harmelen, eds.), (Berlin, Heidelberg), pp. 502–517, Springer Berlin Heidelberg, 2004.
- [19] K. Hutt, “A comparison of rdf query languages,” in *21st Computer Science Seminar*.
- [20] J. Broekstra and A. Kampman, *An RDF Query and Transformation Language*, pp. 23–39. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
- [21] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf, “Dublin core metadata for resource discovery,” 1998.
- [22] G. Manis, *Εισαγωγή στον Προγραμματισμό με αρωγό τη γλώσσα Python*. Kallipos, 2015.
- [23] G. Van Rossum and F. L. Drake, *An Introduction to Python*. Network Theory Ltd., 2011.
- [24] O. Erling and I. Mikhailov, *Virtuoso: RDF Support in a Native RDBMS*, pp. 501–519. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.
- [25] A.-D. Mezaour, B. Van Nuffelen, and C. Blaschke, *Building Enterprise Ready Applications Using Linked Open Data*, pp. 155–174. Cham: Springer International Publishing, 2014.
- [26] O. Erling, “Virtuoso, a hybrid rdbms/graph column store.,” *IEEE Data Eng. Bull.*, vol. 35, no. 1, pp. 3–8, 2012.
- [27] J. Pérez, M. Arenas, and C. Gutierrez, “Semantics and complexity of sparql,” *ACM Trans. Database Syst.*, vol. 34, pp. 16:1–16:45, Sept. 2009.