ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Διαχείριση Πολιτιστικών Θησαυρών με Χρήση Τεχνολογιών Σημασιολογικού Ιστού

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΙΩΑΝΝΗ ΓΚΑΛΑΝΑΚΗ

Επιβλέπων : Στέφανος Κόλλιας,
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2011

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Διαχείριση Πολιτιστικών Θησαυρών με Χρήση Τεχνολογιών Σημασιολογικού Ιστού

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΙΩΑΝΝΗ ΓΚΑΛΑΝΑΚΗ

Επιβλέπων : Στέφανος Κόλλιας,
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Ιουλίου 2011.

.....	.....	.....
Στέφανος Κόλλιας	Γεώργιος Στάμου,	Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π	Λέκτορας Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

.

Αθήνα, Ιούλιος 2011

.....

ΙΩΑΝΝΗΣ ΓΚΑΛΑΝΑΚΗΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στην προσπάθεια εξέλιξης του Παγκόσμιου Ιστού αναπτύχθηκαν νέες τεχνολογίες που θα μας βοηθήσουν να μεταβούμε στον Σημασιολογικό Ιστό. Προκειμένου να αναπτυχθούν έξυπνες και αυτόματες εφαρμογές που θα αξιοποιούν την σημασιολογία των δεδομένων μας, πρέπει να τα δημοσιεύσουμε ακολουθώντας τις αρχές των Συνδεδεμένων Δεδομένων. Ανεπηρέαστος, από αυτήν την αλλαγή, δεν έμεινε ούτε ο χώρος των πολιτιστικών θησαυρών. Σκοπός της παρούσας διπλωματικής είναι να παρουσιάσουμε την μεθοδολογία μεταφοράς των θησαυρών, από την υπάρχουσα μορφή τους σε μια νέα που θα χρησιμοποιεί τις τεχνολογίες του Σημασιολογικού Ιστού.

Την διαδικασία την χωρίσαμε σε δύο διαφορετικά τμήματα ώστε να επιτύχουμε τον στόχο μας. Επιπλέον, παρουσιάσαμε τα προβλήματα που δημιουργούνται, μέσα από τις διαφορετικές περιπτώσεις που εξετάσαμε, αλλά και τον τρόπο, με τον οποίο τα αντιμετωπίσαμε. Αναλύθηκε η διαδικασία της μετατροπή των θησαυρών, ώστε να δημοσιευθούν σύμφωνα με το σύστημα οργάνωσης γνώσης του SKOS. Επίσης, χρησιμοποιήσαμε το εργαλείο Amalgame, για να συνδέσουμε τους θησαυρούς μας με εξωτερικά σύνολα. Η αντιστοίχιση έγινε με βάση έναν απλό τρόπο σύνδεσης όμοιων πόρων, που περιλαμβάνει το εργαλείο, έχοντας όμως ως αποτέλεσμα ποιοτικές συνδέσεις, μεταξύ θησαυρών με μεγάλο αριθμό όρων. Με αυτό τον τρόπο οι θησαυροί μας, συνδέθηκαν με τον σύννεφο των Συνδεδεμένων Δεδομένων.

Λέξεις Κλειδιά: Σημασιολογικός Ιστός, Συνδεδεμένα Δεδομένα, Θησαυροί, SKOS, Amalgame

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

In order to extend the World Wide Web to the Semantic Web, new technologies were developed. Intelligent and automated applications that exploit the semantics of our data require them to be published with the principles of Linked Data. The cultural thesaurus could not be unaffected by these changes. The main scope of this thesis is to present the methodology with which a thesaurus can be transferred from the existing form to a new one, using the technologies of the Semantic Web.

To achieve our objective, we separated the process into two different parts. Furthermore, examining different use cases, we presented some problems that arose and the ways that we solved them. Specifically, we analyzed the process of conversion of thesaurus and how to be published using the knowledge organization system, the SKOS. We have also used the tool Amalgame, to connect our treasures with external target datasets. The alignment was based on the tool's way to connect similar resources, resulting quality semantic links between large thesauruses. Finally, the concepts of our thesaurus are included in the Linked Data cloud.

Keywords: Semantic Web, Linked Data, Thesaurus, SKOS, Amalgame

Ευχαριστίες

Θέλω να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Στέφανο Κόλλια που μου έδωσε την ευκαιρία να ασχοληθώ με ένα τόσο πρωτοποριακό επιστημονικά αντικείμενο στον τομέα του Σημασιολογικού Ιστού, αλλά και τον ερευνητή Βασίλη Τζουβάρα, υπό την καθοδήγησή του οποίου έγινε η παρούσα διπλωματική.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου και όσους με στήριξαν, υλικά και πνευματικά, κατά την διάρκεια των φοιτητικών μου χρόνων.

Πίνακας περιεχομένων

Κατάλογος εικόνων

Κατάλογος πινάκων

Η σελίδα αυτή είναι σκόπιμα λευκή.

1

Εισαγωγή

1.1 Πολιτιστικοί θησαυροί και Σημασιολογικός Ιστός

Ο Παγκόσμιος Ιστός έφερε επανάσταση στον τρόπο επικοινωνίας των ανθρώπων και στην πρόσβαση που έχουμε στο χώρο των πληροφοριών, αφού μεγάλος όγκος δεδομένων είναι πλέον συνεχώς διαθέσιμος και αποθηκευμένος στον Ιστό. Ωστόσο, τα τελευταία χρόνια, ο Ιστός έχει λάβει μία νέα μορφή (το Web 2.0), που δίνει έμφαση στην προσωπική επικοινωνία αλλά και στην εξατομίκευση των παρεχόμενων υπηρεσιών. Το γεγονός αυτό, μας οδήγησε στην δημιουργία έξυπνων και αυτόματων πρακτόρων, που θα διαχωρίζουν και θα διαχειρίζονται την πληροφορία που ενδιαφέρει κάθε χρήστη. Η εξέλιξη αυτή έφερε αντιμέτωπη την κοινότητα του Ιστού με ένα σημαντικό πρόβλημα. Σε αντίθεση με ένα ανθρώπινο χειριστή που μπορεί να πλοηγηθεί εύκολα μεταξύ των δεδομένων, για τις μηχανές δεν ισχύει το ίδιο γιατί το μεγαλύτερο μέρος της πληροφορίας, που είναι διαθέσιμη στον Ιστό, δεν χρησιμοποιεί κάποια συγκεκριμένη δομή και δεν έχει σημασιολογία. Για το λόγο αυτό, παρουσιάστηκε από τον Tim Berners-Lee μια επέκταση του σημερινού Ιστού, ο Σημασιολογικός Ιστός (Semantic Web).

Για την παροχή αυτοματοποιημένων εφαρμογών, ο Σημασιολογικός Ιστός έθεσε ως στόχο την μετατροπή της υπάρχουσας πληροφορίας σε μία δομημένη μορφή, όπου με την χρήση των μεταδεδομένων, θα είναι εύκολα προσβάσιμη από ανθρώπους και από μηχανές. Επιπλέον, εισήχθηκε η έννοια των Συνδεδεμένων Δεδομένων (Linked Data), όπου τα διαφορετικά σύνολα πληροφοριών θα είναι συνδεμένα μεταξύ τους με σημασιολογικούς δεσμούς. Παρουσιάστηκαν επίσης, κάποιες βασικές αρχές για την

επιτυχή μεταφορά τους. Όμως, ο όγκος δεδομένων που υπάρχουν στον Ιστό είναι πολύ μεγάλος και δεν υπάρχει κάποια ομοιομορφία στην δομή τους. Έτσι, η δημιουργία μιας καθολικής και αυτοματοποιημένης διαδικασίας συναντά αρκετές δυσκολίες και οι διαχειριστές των δεδομένων αναζητούν λύσεις για την δική τους εξιδανικευμένη περίπτωση.

Οι πολιτιστικοί θησαυροί αποτελούν βασικά σύνολα δεδομένων στον χώρο του Ιστού. Ιδρύματα και πανεπιστήμια από όλο τον κόσμο προσπαθούν να δημοσιεύσουν τους θησαυρούς που διαθέτουν στον χώρο των Συνδεδεμένων Δεδομένων. Η σωστή μεταφορά τους στον Σημασιολογικό Ιστό αλλά και η χρησιμοποίηση κοινού λεξιλογίου από τους διάφορους φορείς είναι μία από τις προκλήσεις που πρέπει να αντιμετωπιστούν. Ακόμα, αφού μετατραπεί σωστά ένας θησαυρός πρέπει να επιλεχθούν, από το σύννεφο των Συνδεδεμένων Δεδομένων, οι σωστοί στόχοι που θα αντιστοιχηθεί. Τέλος, οι συνδέσεις που θα δημιουργηθούν οφείλουν να είναι σωστές σημασιολογικά και να περιγράφουν την πραγματική ομοιότητα μεταξύ των όρων. Επειδή το μέγεθος των θησαυρών είναι αρκετά μεγάλο η δημιουργία και ο έλεγχος των συνδέσεων αυτών, όπως προαναφέρθηκε, δεν είναι εύκολο να γίνει από μια αυτόματη εφαρμογή.

1.2 Αντικείμενο διπλωματικής

Στη διπλωματική αυτή εξετάσαμε αρχικά τις συνθήκες που επικρατούν στον χώρο των πολιτιστικών θησαυρών αλλά και τις τεχνολογίες που χρησιμοποιεί ο Σημασιολογικός Ιστός. Αντικείμενο της εργασίας μας ήταν η παρουσίαση και η επίλυση των προβλημάτων που προκύπτουν κατά την διαδικασία μεταφοράς των θησαυρών, ώστε να είναι έτοιμοι για την δημοσίευση τους ως Συνδεδεμένα Δεδομένα.

Όπως αναφέραμε, μεγάλος όγκος πληροφορίας βρίσκεται υπό την μορφή θησαυρών που έχουν δημιουργηθεί και διατηρούνται από διαφορετικούς φορείς. Η μεταφορά τους στο Σημασιολογικό Ιστό πρέπει να γίνει με βάση τους κανόνες που απαιτούνται ώστε να επιτευχθεί και η πλήρης αξιοποίηση της γνώσης που περιέχουν (λεπτομερής παρουσίαση των κανόνων γίνεται σε επόμενο κεφάλαιο).

Πρώτο εμπόδιο στην προσπάθεια αυτή είναι η διαφορετικότητα και η ποικιλία που συναντάμε στην δομή των θησαυρών. Το γεγονός αυτό, εμποδίζει την ανάπτυξη μιας εφαρμογής που θα μετατρέπει οποιοδήποτε σύνολο σε μία νέα μορφή. Επόμενο θέμα

είναι η υιοθέτηση ενός κοινού λεξιλογίου, που θα έχει εφαρμογή σε όλους τους θησαυρούς και θα επέτρεπε την δημιουργία νέων όρων από τους σχεδιαστές. Στην εργασία μας, επιλέξαμε το σύστημα οργάνωσης γνώσης SKOS και παρουσιάσαμε μεθοδολογίες μετατροπής θησαυρών στο σύστημα αυτό. Κάθε διαχειριστής θησαυρών μπορεί να προχωρήσει στην διαδικασία της μετατροπής, αφού αναλύσει τις απαιτήσεις που έχει και το αποτέλεσμα που επιθυμεί. Εμείς εξετάζουμε τρία σενάρια όπου το καθένα από αυτά παρουσιάζει και αντιμετωπίζει συγκεκριμένα προβλήματα που εμφανίζονται. Το πρώτο αφορά αυτόματη λήψη δεδομένων από την εξωτερική βάση του Geonames, το δεύτερο την υιοθέτηση μοναδικών αναγνωριστικών URI από το IPTC, κρατώντας τοπικά στοιχεία του θησαυρού μας και τέλος, γίνεται μια ολοκληρωμένη μετατροπή θησαυρών, συγκεκριμένης δομής σε SKOS.

Έχοντας τους θησαυρούς στην επιθυμητή μορφή, επόμενη απαίτηση ήταν η δημιουργία σημασιολογικών συνδέσεων μεταξύ τους αλλά και με άλλες γνωστές πηγές. Οι αντιστοιχίσεις μεταξύ των όρων των θησαυρών δεν είναι απλή διαδικασία γιατί τα σύνολα αυτά περιλαμβάνουν συνήθως πολλούς όρους με πιθανώς πολλές ετικέτες ο καθένας (αν είναι για παράδειγμα πολυγλωσσικός). Οι συγκρίσεις που απαιτούνται να γίνουν είναι αρκετά πολύπλοκες και οι συνδέσεις είναι πολύ σημαντικό να περιγράφουν την ακριβή εννοιολογική ομοιότητα που έχουν οι δύο όροι. Για τις αντιστοιχίσεις αυτές εξετάσαμε ξανά μια σειρά από διαφορετικές περιπτώσεις, κάνοντας χρήση του εργαλείου Amalgame, του οποίου ακολουθεί και αναλυτική παρουσίαση.

1.2.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήσαμε τον τρόπο λειτουργίας των Συνδεδεμένων Δεδομένων και τις σημερινές εξελίξεις στο χώρο των πολιτιστικών θησαυρών.
2. Παρουσιάσαμε την βασική μεθοδολογία για την μεταφορά και την δημοσίευση θησαυρών στο Σημασιολογικό Ιστό.
3. Προσδιορίσαμε τα προβλήματα που αντιμετωπίζουμε κατά την μετατροπή θησαυρών σε μορφή SKOS και τον τρόπο που αντιμετωπίζονται μέσα από την υλοποίηση τριών διαφορετικών περιπτώσεων.

4. Υλοποιήσαμε μια εφαρμογή που κάνει αυτόματη λήψη μεγάλου όγκου δεδομένων από την βάση του Geonames, ανάλογα με τα κριτήρια που θέτουμε.
5. Υλοποιήσαμε μια εφαρμογή για την αυτόματη μετατροπή θησαυρών XML σε SKOS, που να διατηρεί την αρχική ιεραρχία του εγγράφου αλλά και να δημιουργεί για κάθε πόρο ένα μοναδικό URI.
6. Εξετάσαμε τις δυνατότητες του νέου εργαλείου Amalgame σε πραγματικές συνθήκες, για αντιστοιχίσεις μεταξύ θησαυρών και δημιουργήσαμε ένα μικρό σύνολο Συνδεδεμένων Δεδομένων.
7. Συζητήσαμε τις αδυναμίες και τα πλεονεκτήματα του εργαλείου Amalgame στην δημιουργία αντιστοιχίσεων μεταξύ λεξιλογίων, αλλά και εξετάσαμε αναλυτικά τον τρόπο λειτουργίας του.

1.3 Οργάνωση κειμένου

Η παρούσα διπλωματική περιέχει 10 Κεφάλαια. Το Κεφάλαιο 1 είναι μια εισαγωγή στο αντικείμενο που θα μας απασχολήσει. Το σχετικό θεωρητικό υπόβαθρο με το αντικείμενο της διπλωματικής, παρουσιάζεται στο Κεφάλαιο 2. Στο Κεφάλαιο 3 αναφερόμαστε στο χώρο των πολιτιστικών θησαυρών, παρουσιάζουμε κάποια γνωστά σύνολα δεδομένων και αναλύουμε τη μορφή SKOS. Το Κεφάλαιο 4 περιλαμβάνει την διαδικασία για την δημοσίευση θησαυρών στον Σημασιολογικό Ιστό, από την μετατροπή τους, στην κατάλληλη μορφή, έως την σύνδεση τους με άλλες πηγές Δεδομένων. Στο Κεφάλαιο 5 περιγράφουμε την διαδικασία μετατροπής σε SKOS, θησαυρών, εξετάζοντας τρείς περιπτώσεις. Στο Κεφάλαιο 6 αναλύουμε τον τρόπο δημιουργίας συνδέσεων μεταξύ των θησαυρών κάνοντας χρήσης της εφαρμογής του Amalgame. Στο Κεφάλαιο 7 παρουσιάζουμε τα εργαλεία που χρησιμοποιήθηκαν στην παρούσα εργασία. Τέλος, ακολουθεί το Κεφάλαιο 8 με τα συμπεράσματα και τις μελλοντικές επεκτάσεις της διπλωματικής, το Κεφάλαιο 9 που αναφέρει την βιβλιογραφία που μελετήθηκε και το Κεφάλαιο 10 που περιλαμβάνει ένα παράρτημα με συμπληρωματικές πληροφορίες.

2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζεται το θεωρητικό υπόβαθρο, πάνω στο οποίο στηρίζεται η παρούσα διπλωματική και η κατανόηση του οποίου είναι απαραίτητη για τον αναγνώστη. Ο αναγνώστης, μέσα από αυτό, θα μπορέσει να λάβει τις απαραίτητες γνώσεις σχετικά με τη θεματική περιοχή της διπλωματικής.

2.1 Σημασιολογικός Ιστός (Semantic Web)

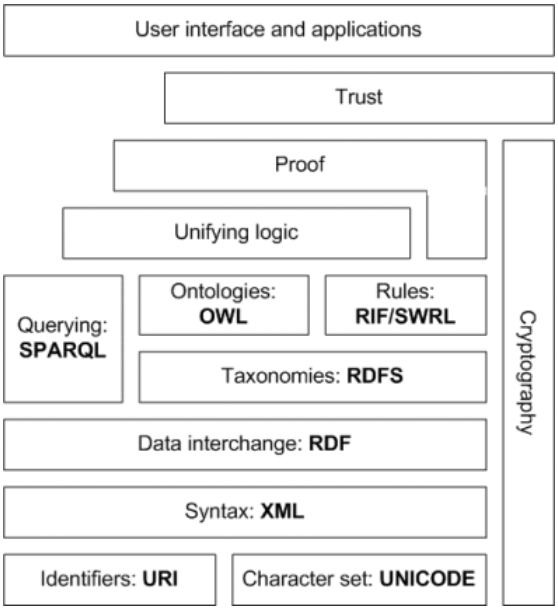
Το 1989 ο βρετανός Tim Berners-Lee δημιούργησε τη τεχνολογία του Ιστού. Η εφεύρεσή αυτή ονομάστηκε από τον ίδιο, Παγκόσμιος Ιστός (World Wide Web) όρος γνωστός από το "www". Ο Lee οραματιζόταν ένα κόσμο όπου ο καθένας θα μπορούσε να ανταλλάσσει πληροφορίες και ιδέες άμεσα προσβάσιμες από τους υπολοίπους. Από τότε ο ιστός έχει αλλάξει ριζικά και οι περισσότερες σελίδες χρησιμοποιούν για την αποθήκευση τους την γλώσσα της HTML. Αν και η HTML περιέχει μεταδεδομένα αυτά χρησιμοποιούνται για την περιγραφή της σελίδας, όπως τις λέξεις κλειδιά που την περιγράφουν. Με αυτό τον τρόπο έχουμε καταλήξει στο παράδοξο ο Παγκόσμιος Ιστός να έχει κατασκευαστεί από υπολογιστές αλλά να χρησιμοποιείτε από ανθρώπους. Οι σελίδες χρησιμοποιούν την φυσική γλώσσα, τις εικόνες και σχεδιάζονται ώστε η πληροφορία να παρουσιάζεται ώστε να την αντιλαμβάνεται εύκολα ο χρήστης. Οι υπολογιστές αντίθετα, ενώ είναι το κέντρο στη δημιουργία και στη διατήρηση του Ιστού, οι ίδιοι δεν βγάζουν κάποιο νόημα με όλη αυτή την πληροφορία. Δεν μπορούν ούτε να διαβάσουν ούτε να δουν συσχετίσεις όπως θα κάνει ένας χρήστης.

Ο Lee οραματίστηκε το 1999 μια νέα μορφή ιστού η οποία θα λειτουργούσε ως επέκταση στην ήδη υπάρχουσα. Ο Σημασιολογικός Ιστός, όπως ονομάστηκε (1), είναι ένας Ιστός από δεδομένα που θα δίνει στις μηχανές τη δυνατότητα, αφού επεξεργαστούν, να αντιλαμβάνονται το νόημα της πληροφορίας που υπάρχει σ' αυτόν και να εξάγουν συμπεράσματα. Αποτέλεσμα της διαδικασίας αυτής θα είναι μια νέα γενιά υπολογιστών που θα εκτελούν πολύπλοκες λειτουργίες για τις οποίες, σήμερα, είναι απαραίτητη η ανθρώπινη συμμετοχή. Με αυτό τον τρόπο, ευφυείς πράκτορες θα μπορούν να αποκτήσουν πρόσβαση στον Ιστό ποιο έξυπνα και να εκτελούν λειτουργίες εκ μέρους του χρήστη.

2.1.1 Σήμερα

Σήμερα βρισκόμαστε στο Web 2.0. Όπως είδαμε, η δεύτερη γενιά του Παγκόσμιου Ιστού έχει εστιαστεί στην δυνατότητα για τους ανθρώπους να συνεργαστούν και να ανταλλάξουν πληροφορία. Ενώ ο απλός Ιστός περιέχει στατικές HTML σελίδες, το Web 2.0 είναι δυναμικό . Προσφέρει εφαρμογές στους χρήστες και δίνει έμφαση στις διαδικτυακές κοινότητες. Μιας και το Web 2.0 δίνει έμφαση στους ανθρώπους και την επικοινωνία περιλαμβάνει ένα μεγάλο αριθμό από τεχνολογίες όπως, AJAX, Ruby, XHTML. Όμως η τεχνολογία είναι λιγότερη σημαντική από τον άνθρωπο αφού το μόνο που θέλουμε είναι το αποτέλεσμα, το οποίο είναι η κοινωνική αλληλεπίδραση σε μια ελκυστική και εύκολη στη χρήση εφαρμογή.

Η σημερινή μορφή του Ιστού απέχει πολύ από το «όραμα» του Σημασιολογικού Ιστού. Είναι φανερό ότι μεταξύ των δύο αυτών κόσμων υπάρχει ένα χάσμα. Τα δεδομένα τις πιο πολλές φορές περιλαμβάνονται σε αρχεία μη δομημένα και μη τυποποιημένα. Η μετάβαση λοιπόν δεν είναι ούτε σύντομη ούτε απλή υπόθεση. Παρόλα αυτά έχουν οριστεί τα θεμέλια πάνω στα οποία θα στηριχθεί ο Σημασιολογικός Ιστός και έχουν ξεκινήσει να γίνονται σημαντικά βήματα προς αυτή την κατεύθυνση. Νέες τεχνολογίες και εργαλεία έχουν αρχίσει να δημιουργούνται και να εξελίσσονται τις οποίες τις βλέπουμε στο παρακάτω σχήμα. Όσες από αυτές χρησιμοποιήθηκαν για την παρούσα διπλωματική θα παρουσιαστούν αναλυτικά στη συνέχεια.



Εικόνα 2.1.1.1 Τεχνολογίες Σημασιολογικού Ιστού (2)

2.2 XML (xtensible Markup Language)

Πριν προχωρήσουμε στις τεχνολογίες που εξελίχθηκαν με τον Σημασιολογικό Ιστό, θα αναλύσουμε σύντομα κάποια βασικά στοιχεία μίας γλώσσας σήμανσης , που περιέχει ένα σύνολο κανόνων για την ηλεκτρονική κωδικοποίηση κειμένων. Η XML σχεδιάστηκε δίνοντας έμφαση στην απλότητα, τη γενικότητα και τη χρησιμότητα στο Διαδίκτυο (3). Είναι μία μορφοποίηση δεδομένων κειμένου, με ισχυρή υποστήριξη Unicode για όλες τις γλώσσες του κόσμου. Αν και η σχεδίαση της XML εστιάζει στα κείμενα, χρησιμοποιείται ευρέως για την αναπαράσταση αυθαίρετων δομών δεδομένων, που προκύπτουν για παράδειγμα στις υπηρεσίες ιστού. Υπάρχει μία ποικιλία διεπαφών προγραμματισμού εφαρμογών, που μπορούν να χρησιμοποιούν οι προγραμματιστές, για να προσπελαύνουν δεδομένα XML, αλλά και διάφορα συστήματα σχημάτων XML, τα οποία είναι σχεδιασμένα για να βοηθούν στον ορισμό γλωσσών, που προκύπτουν από την XML.

Οι περισσότεροι θησαυροί με τους οποίους ασχολείται η παρούσα διπλωματική βρίσκονται σε μορφή XML. Προκειμένου τα δεδομένα των θησαυρών αυτών να είναι έτοιμα για χρήση στον Σημασιολογικό Ιστό, πρέπει να μετατραπούν με σωστό τρόπο σε άλλες μορφές (π.χ. RDF που αναλύεται στην συνέχεια). Η σύντομη παρουσίαση κάποιων βασικών δομικών στοιχείων της XML είναι απαραίτητη για την κατανόηση της μεθόδου μετατροπής που θα δούμε στα επόμενα κεφάλαια.

2.2.1 Δομικά στοιχεία XML

Αρχικά όλα τα XML αρχεία μπορούν να αρχίζουν με μια XML δήλωση. Αυτή δείχνει την ελάχιστη έκδοση της XML που χρησιμοποιείται στο κείμενο. Ακόμα διευκρινίζει την κωδικοποίηση των χαρακτήρων του αρχείου αυτού. Τυπικά τα δεδομένα στο Σημασιολογικό Ιστό κωδικοποιούνται με το UTF-8.

```
<?xml version="1.0" encoding="UTF-8"?>
```

Το βασικό δομικό στοιχείο ενός XML εγγράφου με το οποίο περιγράφονται τα δεδομένα είναι τα στοιχεία (element). Ένα στοιχείο αρχίζει με μια ετικέτα ανοίγματος(opening tag), ακολουθεί το περιεχόμενο του και τελειώνει με άλλη ετικέτα(closing tag). Η ονομασία του στοιχείου γίνεται ελεύθερα από τον χρήστη και συνήθως δηλώνει συνοπτικά και κατανοητά το περιεχόμενο του. Υπάρχουν και μερικοί περιορισμοί όμως όπως ότι ο πρώτος χαρακτήρας της ετικέτας δεν μπορεί να είναι γράμμα, κάτω παύλα ή ελληνικό ερωτηματικό. Επίσης η ετικέτα δεν μπορεί να ξεκινάει με το λεκτικό xml σε οποιοδήποτε συνδυασμό πεζών/μικρών γραμμάτων. Δύο παραδείγματα είναι:

```
<person>Tony Blair</person>
```

Το στοιχείο έχει ετικέτα ‘person’ που δηλώνει ότι στο περιεχόμενου του θα μιλήσουμε για όνομα ενός ατόμου . Το περιεχόμενο του στοιχείου μπορεί να είναι είτε κενό, είτε απλό κείμενο είτε άλλα στοιχεία.

Το ακόλουθο στοιχείο είναι παράδειγμα με κενό περιεχόμενο

```
<line-break/>
```

Επόμενο βασικό στοιχείο της γλώσσας είναι τα χαρακτηριστικά (attributes) που μπορεί να έχει ένα στοιχείο. Αυτά μπορεί να είναι περισσότερα από ένα και αποτελούνται από ένα ζευγάρι όνομα/τιμή, το οποίο υπάρχει μέσα σε μία ετικέτα-αρχής ή σε μία ετικέτα-χωρίς-περιεχόμενο. Η χρήση τους μοιάζει και μπορεί να αντικατασταθεί από ενθυλακωμένα στοιχεία. Η επιλογή ανάμεσα σε ενθυλακωμένα στοιχεία ή γνωρίσματα εξαρτάται από την οργάνωση που θέλουμε στο έγγραφο.

Ένα παράδειγμα φαίνεται παρακάτω:

```
<person name=”Tony” lastname=”Blair”>
<country> Great Britain </country>
< function > Prime Minister </ function >
</ person >
```

Το οποίο θα μπορούσε να γραφτεί μονάχα με στοιχεία ως εξής:

```
<person>
<name>Tony</name>
< lastname>Blair</lastname>
<function>Prime Minister</function>
<country>Great Britain</country>
</person>
```

Η XML έχει ένα σύνολο κανόνων ώστε το έγγραφο να θεωρείται έγκυρο και σε αντίθεση με την HTML απαιτεί μια αυστηρή δομή στη σύνταξη του εγγράφου. Έτσι κάθε στοιχείο, περιέχει μία αρχική ετικέτα και μια αντίστοιχη τελική ετικέτα. Οι ετικέτες δεν πρέπει να επικαλύπτονται και τα στοιχεία πρέπει να έχουν μοναδικά ονόματα. Τέλος ενώ η σειρά των χαρακτηριστικών δεν μας ενδιαφέρει, η σειρά των στοιχείων είναι εξαιρετικά σημαντική για την δομή του αρχείου μας.

2.2.2 Μειονεκτήματα της XML

Όπως είδαμε παραπάνω η XML είναι μια καθολική μετά-γλώσσα σήμανσης , η οποία παρέχει ένα ενιαίο πλαίσιο για διαχείριση των δεδομένων και των μεταδεδομένων μεταξύ των εφαρμογών. Με την βοήθεια της δημιουργούμε δομημένα έγγραφα στον Ιστό, κατάλληλα για την ανάγνωση τους από μηχανές. Όμως για την επίτευξη των στόχων του Σημασιολογικού Ιστού, τους οποίους είδαμε στην προηγούμενη ενότητα, η XML δεν είναι κατάλληλη με την παρούσα μορφή της αφού δεν παρέχει τα μέσα για να περιγράψουμε την σημασιολογία, δηλαδή το νόημα, των δεδομένων μας. Στο παρακάτω παράδειγμα βλέπουμε ότι με την XML δεν υπάρχει τυπικός τρόπος αντιστοίχισης νοήματος στην ένθεση ετικετών(nesting of tags) και η σωστή ερμηνεία αφήνεται στην εφαρμογή:

```
<car name>” Yaris”
<manufacture>Toyota </ manufacture></car name >
```

```
< manufacture name>” Toyota”
< car > Yaris </car ></manufacture >
```

```
<carjapan>< car >” Yaris” </car >
< manufacture > Toyota </manufacture ></carjapan >
```

Για την αντιμετώπιση των προβλημάτων σημασιολογίας που είδαμε ότι έχει η XML, αναπτύχθηκε ένα νέο πρότυπο, η RDF, η οποία, βασίστηκε σ’ αυτήν και προσαρμόστηκε στις νέες ανάγκες του Σημασιολογικού Ιστού.

2.3 RDF (Resource Description Framework)

Η Resource Description Framework (RDF) είναι ένα πρότυπο για ανταλλαγή δεδομένων στον Ιστό, που σχεδιάστηκε σαν ένα μοντέλο για μεταδομένα. Είναι μια γλώσσα που χρησιμοποιείτε στην αναπαράσταση της πληροφορίας σχετικής με πόρους στον Παγκόσμιο Ιστό (4). Ο τίτλος, ο συγγραφέας και η ημερομηνία τροποποίησης μίας ιστοσελίδας είναι μερικά παραδείγματα τέτοιων πόρων. Επίσης, γενικεύοντας τον όρο «πόρος», η RDF μπορεί να χρησιμοποιηθεί και για την περιγραφή πραγμάτων, τα οποία μπορούν να προσδιοριστούν στον Ιστό, αλλά δεν μπορούν να ανακτηθούν άμεσα μέσα σε αυτόν. Τέτοιες περιπτώσεις μεταδεδομένων είναι οι πληροφορίες για τα προϊόντα ενός on-line καταστήματος (ειδικά χαρακτηριστικά, τιμή και διαθεσιμότητα) αλλά και η περιγραφή των προτιμήσεων παράδοσης για ένα χρήστη.

Η RDF χρησιμεύει σε περιπτώσεις στις οποίες, η πληροφορία χρειάζεται να επεξεργαστεί από εφαρμογές παρά από ανθρώπους, προσφέρει δηλαδή ένα πλαίσιο (framework) έκφρασης της πληροφορίας, έτσι ώστε να μπορεί αυτή, να μεταφερθεί μεταξύ εφαρμογών χωρίς καμία απώλεια νοήματος. Επειδή το πλαίσιο αυτό είναι κοινό, η πληροφορία αυτή μπορεί να γίνει διαθέσιμη σε εφαρμογές διαφορετικές από εκείνες για τις οποίες είχε αρχικά δημιουργηθεί.

2.3.1 Βασικές έννοιες της RDF

Κυρίως στόχος της RDF είναι να παρέχει ένα τρόπο για να κάνουμε δηλώσεις για πόρους στον Ιστό. Για να συνεχίσουμε την εξήγηση, θα αναλύσουμε τις βασικές έννοιες της RDF: τους πόρους, τις ιδιότητες και τις προτάσεις (5).

- Οι πόροι είναι όλα εκείνα τα «αντικείμενα» τα οποία αφορά η δήλωσή μας. Σε κάθε πόρο αντιστοιχεί ένα μοναδικό αναγνωριστικό, το URI(Uniform Resource Identifier) του. Αυτό είναι μια σειρά από χαρακτήρες που χρησιμοποιείτε για να αναγνωριστεί το αντικείμενο σε ένα δίκτυο(συνήθως στον Παγκόσμιο Ιστό). Επειδή στην πράξη ένα συχνά το να γράφουμε ολόκληρο το URI ενός πόρου δεν είναι βολικό αφού καταλήγουμε σε πολύ μακριές γραμμές σε μια σελίδα, χρησιμοποιούμε για συντόμευση προθέματα (URI prefixes) τα οποία αναθέτουμε σε κάποια (ευρέως χρησιμοποιούμενα) URIs χώρων ονομάτων (namespace URIs). Έτσι στην αρχή δηλώνουμε για παράδειγμα ότι το πρόθεμα foaf αντικαθιστά το URI χώρου ονομάτων http://xmlns.com/foaf/0.1/,με την εντολή

xmlns:foaf="http://xmlns.com/foaf/0.1/".

Με αυτό τον τρόπο μπορούμε μετά αντί για http://xmlns.com/foaf/ 0.1/knows, να γράφουμε foaf:knows.

- Οι ιδιότητες είναι μία ειδική περίπτωση πόρων που περιγράφουν τις σχέσεις μεταξύ πόρων. Τέτοιες σχέσεις είναι οι : «δημιουργός», «ημερομηνία γέννησης», κ.α. Οι ιδιότητες στην RDF πρέπει να είναι πόροι . Έτσι καθορίζονται επίσης από URI κάτι που προσφέρει μία μέθοδο παγκόσμιας, μοναδικής ονομασίας. Ένα σημαντικό πρόβλημα για τον Σημασιολογικό Ιστό ήταν η ομωνυμία που «μαστίζει» την αναπαράσταση των κατανεμημένων δεδομένων μέχρι σήμερα. Όπως θα δούμε και στη συνέχεια για τη λύση του προβλήματος αυτού δημιουργήθηκαν επίσημες οντολογίες, στο πλαίσιο της RDF, όπως η FOAF που είδαμε .
- Η λογική στην RDF είναι ότι όλες οι προτάσεις συνθέτονται από τριάδες (triple) υποκειμένου-κατηγορήματος –αντικείμενου (subject - predicate - object). Έστω ότι θέλουμε να περιγράψουμε ότι ο κάποιος με το όνομα John Smith έφτιαξε μια συγκεκριμένη σελίδα στον Ιστό. Χρησιμοποιώντας την φυσική γλώσσα θα λέγαμε

http://www.example.org/index.html έχει **δημιουργό** του οποίου η τιμή είναι **John Smith**

μέρη της δήλωσης έχουν τονιστεί για να δώσουμε έμφαση ότι για την διατύπωση μιας πρότασης χρειάζεται:

1. Το αντικείμενο το οποίο αφορά η δήλωση (η σελίδα, στην περίπτωση αυτή)
2. Μια ειδική ιδιότητα (δημιουργός, στην περίπτωση αυτή) του αντικειμένου που περιγράφεται
3. Αυτό που είναι η τιμή για το αντικείμενο της πρότασης (ποιος είναι ο δημιουργός, στην περίπτωση αυτή)

Όπως είδαμε η σελίδα έχει ένα URL (Uniform Resource Locator) το οποίο χρησιμοποιείτε για να αναγνωριστεί. Επιπλέον η λέξη «δημιουργός χρησιμοποιείτε για την αναγνώριση της ιδιότητας και οι δυο λέξεις «John Smith» για να αναγνωριστεί το αντικείμενο(ένα άτομο) που είναι η τιμή της ιδιότητας.

Το παραπάνω παράδειγμα μας δείχνει πώς θα γινόταν μια δήλωση σε φυσική γλώσσα.

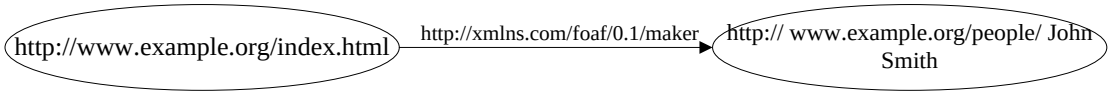
Αντίστοιχα στην RDF για μια δήλωση(rdf:Statement) έχουμε τις εξής ιδιότητες:

- Πόρος(rdf:subject)
- Ιδιότητα (rdf:predicate)
- Αντικείμενο (rdf:object) που μπορεί να είναι είτε ένας πόρος είτε ένα λεκτικό (literal)

Το αντίστοιχο παράδειγμα δηλαδή στην RDF θα γινόταν η παρακάτω τριάδα :

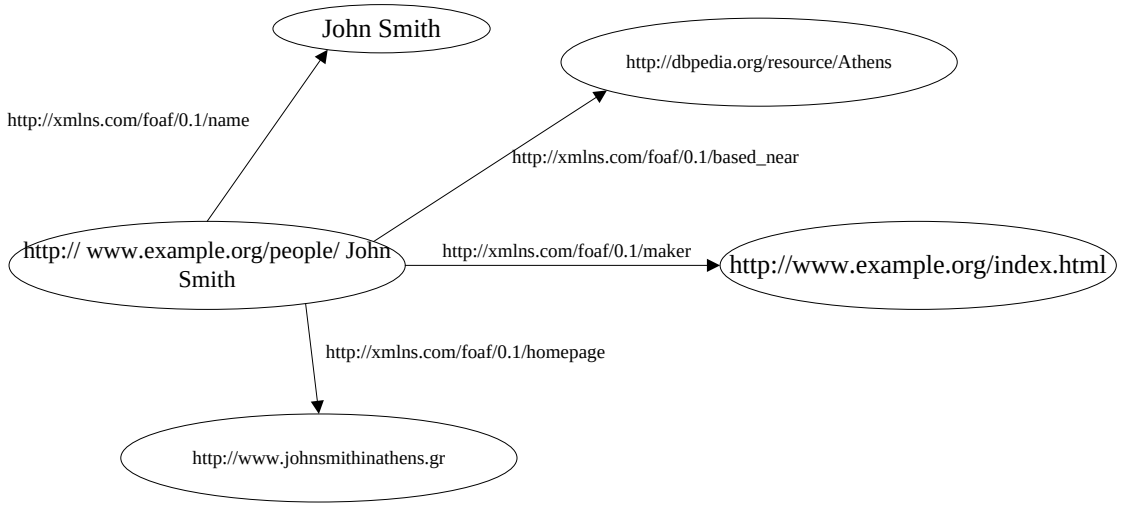
http://www.example.org/index.html, http://xmlns.com/foaf/0.1/maker, http:// www.example.org/people/ John Smith.

Οι προτάσεις μπορούν να αναπαρασταθούν και γραφικά. Ο πόρος και το αντικείμενο σχηματίζονται ως δύο κόμβοι που συνδέονται με μία κατευθυνόμενη ακμή που είναι η ιδιότητα. Για παράδειγμα, η παραπάνω πρόταση αναπαρίσταται γραφικά ως εξής:



Εικόνα 2.3.1.1 Γραφική αναπαράσταση RDF τριάδας

Στο σημασιολογικό δίκτυο η κάθε πρόταση δεν είναι απομονωμένη. Αντίθετα οι προτάσεις μπορεί να παριστάνουν ένα κατευθυνόμενο γράφο όπως ο παρακάτω:



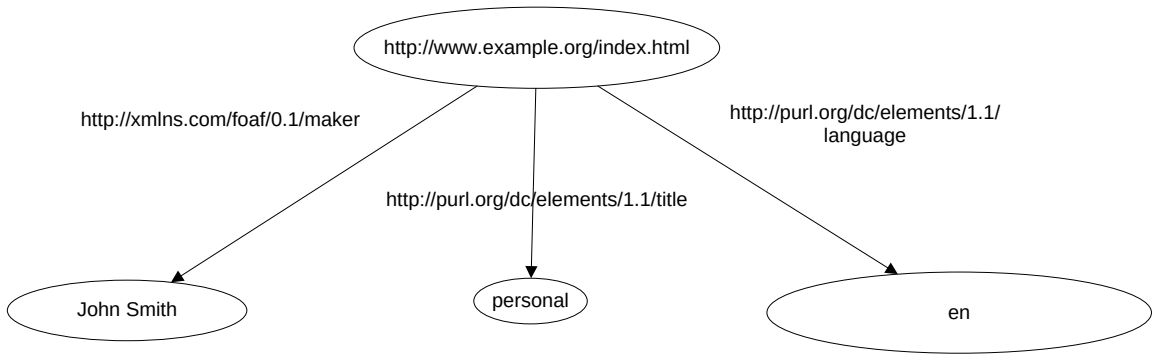
Εικόνα 2.3.1.2 Ένα σημασιολογικό δίκτυο

2.3.2 Μορφές σειριακής διάταξης της RDF (RDF serialization formats)

Όπως περιγράψαμε στο παραπάνω κεφάλαιο 2.2.1 το μοντέλο της RDF είναι ένας γράφος. Σε αρκετές περιπτώσεις είναι αναγκαίο η πληροφορία που βρίσκετε στο RDF αρχείο μας να αποθηκευτεί, να επεξεργαστεί και να μεταφερθεί σε μια άλλη εφαρμογή. Η μορφή του γράφου δεν είναι βολική για αυτές τις χρήσεις και γι' αυτό

αναπτύχθηκαν μορφές σειριακής διάταξης της RDF. Τα δύο ποιο κοινά πρότυπα είναι είτε ως σημειογραφία τριάδων(triples notation) είτε ως XML έγγραφο (RDF/XML). Παραδείγματα και περιγραφή αυτών των τρόπων αναπαράστασης ακολουθούν.

Η RDF/XML σύνταξη για την αναπαράσταση και ανταλλαγή RDF μεταδεδομένων είναι βασισμένη στην XML και οι αναλυτικές προδιαγραφές της παρουσιάζονται στην αντίστοιχη σύσταση της W3C (6). Το W3C μαζί με την σειριακή διάταξη της RDF σαν XML πρότεινε την σύνταξη Notation 3 (N3) (7). Σχεδιάστηκε με στόχο την ευκολότερη γραφή με το χέρι και την καλύτερη κατανόηση. Επειδή βασίζεται σε μία πινακοειδή σημειογραφία (tabular notation), προσφέρει ευκολότερη αναγνώριση των τριάδων που περιέχονται σε κάποιο RDF έγγραφο. Η Notation 3 είναι στενά συνδεδεμένη με άλλες δύο μορφές σειριακής διάταξης της RDF, οι τη Turtle (8) και η N-Triples (9). Η Turtle είναι υποσύνολο της Notation 3 και η μόνη διαφορά των δύο γλωσσών είναι ότι η Turtle δεν υποστηρίζει τη σύνταξη κανόνων RDF καθώς και τη σύνταξη μονοπατιών (path syntax). Η N-Triples είναι υποσύνολο της Turtle και αποτελεί την πιο απλή μορφή σειριακής διάταξης της RDF. Τέλος παρουσιάζουμε μια αναπαράσταση μιας RDF περιγραφής με γράφο σε όλους τους τύπους που αναφέρθηκαν.



Εικόνα 2.3.2.1 Αναπαράσταση RDF περιγραφής με γράφο

Αναπαράσταση σε RDF/XML :

```
1: <?xml version="1.0"?>
2: <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3: xmlns:dc="http://purl.org/dc/elements/1.1/"
4: xmlns:foaf="http://xmlns.com/foaf/0.1/">
5:    <rdf:Description rdf:about="http://www.example.org/index.html">
```

```
6:     <foaf:maker>John Smith</foaf:maker>
7:     <dc:title>personal</dc:title>
8:      <dc:language>en</dc:language>
9:    </rdf:Description>
10: </rdf:RDF>
```

Αναπαράσταση σε Notation 3 ή Turtle :

```
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
<http://www.example.org/index.html> dc:language "en";
        dc:title "personal";
        foaf:maker "John Smith".
```

Αναπαράσταση σε N-Triples :

```
<http://www.example.org/index.html><http://xmlns.com/foaf/0.1/maker> "John
Smith".
<http://www.example.org/index.html> <http://purl.org/dc/elements/1.1/title>
personal".
<http://www.example.org/index.html> <http://purl.org/dc/elements/1.1/language>
"en"
```

2.3.3 Δομημένες τιμές(structured values) με χρήση κενών κόμβων (blank nodes)

Στην RDF, ονομάζουμε κενό κόμβο αυτόν που δεν έχει κάποιο URIref ή μία σταθερή τιμή. Συχνά, τα αληθινά δεδομένα έχουν πολύπλοκες δομές, όπου ορισμένες έννοιες χρησιμοποιούνται για την συνάθροιση τιμών και δεν τυγχάνουν καμία αναφορά πέρα από τα πλαίσια ενός εγγράφου. Για παράδειγμα, έστω ότι θέλουμε να περιγράψουμε τη διεύθυνση ενός ατόμου. Ένας τρόπος είναι ο παρακάτω :

```
exstaff:85740 exterms:address   "1501 Grant Avenue, Bedford, Massachusetts
01730".
```

Όμως, με αυτό τον τρόπο μία εφαρμογή δεν μπορεί να αναγνωρίσει τα συστατικά στοιχεία της διεύθυνσης. Για να γίνει αυτό, πρέπει να εισάγουμε ένα βοηθητικό πόρο, ο οποίος θα αποτελέσει το συνδετικό κρίκο για να μετασχηματίσουμε τη διεύθυνση σε μία περισσότερο δομημένη μορφή. Έτσι, τα δεδομένα μας μετασχηματίζονται ως εξής :

```
exstaff:85740      exterms:address       exaddressid:85740 .
exaddressid:85740  exterms:street        "1501 Grant Avenue" .
exaddressid:85740  exterms:city          "Bedford" .
exaddressid:85740  exterms:state         "Massachusetts" .
exaddressid:85740  exterms:postalCode    "01730" .
```

Ο πόρος exaddressid:85740, όπως, βλέπουμε βοηθάει στη συνεκτικότητα και στη δομή της πληροφορίας μας, αλλά δεν είναι χρήσιμο να ταυτοποιηθεί με ένα μοναδικό URL. Κάθε κενός κόμβος πρέπει να είναι διακριτός σε ένα γράφο και να ταυτοποιείται με διαφορετικό όνομα. Στη σημασιολογία τριάδων, τέτοιοι κόμβοι, ταυτοποιούνται προσθέτοντας το όνομα τους με τους χαρακτήρες `_:name`. Η αντικατάσταση του πόρου exaddressid:85740 με έναν κενό κόμβο στο παράδειγμά μας, θα είχε σαν αποτέλεσμα οι τριάδες μας να μετασχηματιστούν ως εξής :

```
exstaff:85740   exterms:address       _:johnaddress .
_:johnaddress   exterms:street        "1501 Grant Avenue" .
_:johnaddress   exterms:city          "Bedford" .
_:johnaddress   exterms:state         "Massachusetts" .
_:johnaddress   exterms:postalCode    "01730" .
```

Οι κενοί κόμβοι, όπως βλέπουμε μπορούν να εμφανιστούν μόνο ως υποκείμενα ή αντικείμενα σε μια δήλωση και είναι χρήσιμοι μόνο στην αναπαράσταση n-αδικών σχέσεων, αφού στην RDF υποστηρίζονται άμεσα μόνο οι δυαδικές σχέσεις.

2.3.4 Τυποποιημένα λεκτικά (typed literals)

Ένα λεκτικό(literal) το οποίο εμφανίζεται ως το αντικείμενο μιας δήλωσης μπορεί να συσχετιστεί με κάποιο τύπο δεδομένων. Έστω η πρόταση:

```
exstaff:85740   exterms:age     27 .
```

Στην πρόταση αυτή, δεν υπάρχει καμία ένδειξη αν το «27» πρέπει να ερμηνευθεί ως αριθμός ή ως αλφαριθμητικό. Ακόμα, στη περίπτωση που είναι αριθμός, δεν ξέρουμε αν η αναπαράστασή του είναι δεκαδική ή οκταδική άρα και έχει την τιμή είκοσι τρία. Στην RDF ο τύπος ταυτοποιείται με ένα URIref. Το URIref του τύπου ακολουθεί το στοιχείο και χωρίζεται από αυτό κάνοντας χρήση της σημειογραφίας ^^ .Το RDF δεν έχει δικό του σύστημα τύπων και γι' αυτό οι τύποι δεδομένων είναι ορισμένοι εξωτερικά από την RDF, και προσδιορίζονται από τα URIs τους. Έτσι, το παραπάνω παράδειγμα μετασχηματίζεται ως εξής :

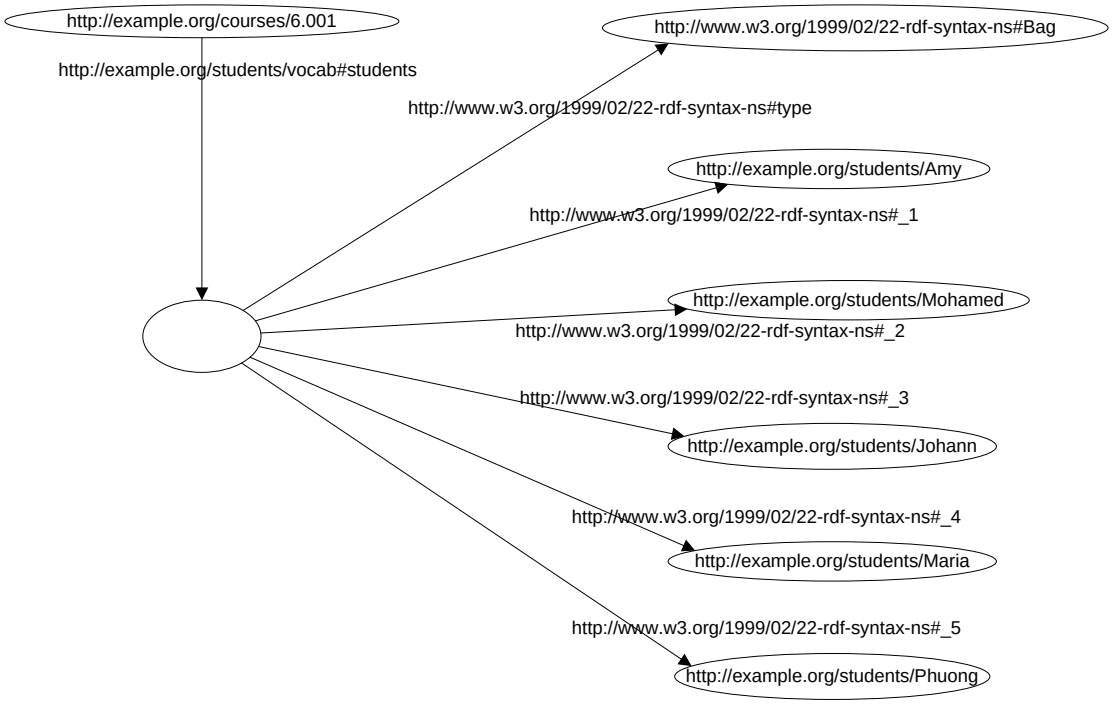
<http://www.example.org/staffid/85740><http://www.example.org/terms/age>
"27"^^http://www.w3.org/2001/XMLSchema#integer

2.3.5 Στοιχεία - υποδοχείς (Containers)

Συχνά υπάρχει η ανάγκη να περιγράψουμε ομάδες αντικειμένων, όπως για παράδειγμα τους συγγραφείς ενός βιβλίου ή τη λίστα των μαθητών που γράφτηκαν σε ένα μάθημα. Η RDF προσφέρει κάποιους προκαθορισμένους τύπους και ιδιότητες για την περιγραφή αυτών των ομάδων. Το κατηγόρημα rdf:type ορίζει τον τύπο ενός αντικειμένου. Υπάρχουν τρεις τύποι υποδοχέων στην RDF :

- rdf:Bag, η σειρά των μελών δεν είναι σημαντική και ενδεχομένως να περιέχει διπλότυπα μέλη
- rdf:Seq, η σειρά των μελών είναι σημαντική και ο υποδοχέας μπορεί να περιέχει πολλαπλές εμφανίσεις
- rdf:Alt, τα μέλη αποτελούν ένα σύνολο εναλλακτικών επιλογών

Ένας υποδοχέας συνδέεται με τα μέλη του μέσω ιδιοτήτων με ονόματα rdf:_1, rdf:_2, …, rdf:_n. Ο γράφος που αντιστοιχεί σε έναν rdf:Bag υποδοχέα που αναπαριστά τους φοιτητές ενός μαθήματος φαίνεται στο παρακάτω σχήμα:



Εικόνα 2.3.5.1 Ένας υποδοχέας τύπου rdf:Bag

2.3.6 Χαρακτηρισμός δηλώσεων RDF (Reification)

Μια RDF εφαρμογή μπορεί να χρειαστεί να περιγράψει δηλώσεις RDF με στόχο να καταγράφει πληροφορίες, όπως πότε έγινε μία πρόταση, από ποιον, κλπ. Η κύρια ιδέα είναι η εισαγωγή ενός βοηθητικού πόρου και η συσχέτισή του με καθένα από τα τρία τμήματα της αρχική πρότασης με τη βοήθεια κατάλληλων ιδιοτήτων. Η RDF προσφέρει ένα ενσωματωμένο λεξιλόγιο για την περιγραφή RDF δηλώσεων , το οποίο περιλαμβάνει το αντικείμενο rdf:Statement για το κατηγόρημα rdf:type και τις ιδιότητες rdf:subject, rdf:predicate και rdf:object. Για παράδειγμα έστω ότι έγινε η δήλωση:

```
exproducts:item10245   exterms:weight   "2.4"^^xsd:decimal
```

με χρήση του χαρακτηρισμού δηλώσεων γίνεται

```
exproducts:triple12345   rdf:type        rdf:Statement .
exproducts:triple12345   rdf:subject     exproducts:item10245 .
exproducts:triple12345   rdf:predicate   exterms:weight .
exproducts:triple12345   rdf:object      "2.4"^^xsd:decimal .
exproducts:triple12345   dc:creator      exstaff:85740 .
```

έχοντας το παραπάνω παράδειγμα, βλέπουμε ότι μια δήλωση συνήθως περιγράφεται με μία τετράδα προτάσεων (στο παράδειγμα οι τέσσερις πρώτες προτάσεις), η οποία αναφέρεται και ως «τετράδα του χαρακτηρισμού δηλώσεων» (reification quad) . Η δήλωση λοιπόν, μπορεί να συσχετιστεί με το περιεχόμενο της. Στη συνέχεια χρησιμοποιώντας το Statement URIref ως υποκείμενο, μπορούμε να προβούμε σε περεταίρω χαρακτηρισμούς της δήλωσης.

2.3.7 RDF Schema

Πολλές φορές οι κοινότητες της RDF χρειάζονται την δυνατότητα να ορίζουν το δικό τους λεξιλόγιο, το οποίο σκοπεύουν να χρησιμοποιήσουν στις δηλώσεις τους. Η RDF δεν παρέχει από μόνη της τρόπο να οριστούν κλάσεις και ιδιότητες ανάλογα με την εφαρμογή. Αντίθετα, τέτοιες κλάσεις και ιδιότητες, περιγράφονται ως ένα RDF Schema (10). Η RDFS, όπως αναφέρεται, είναι μία γλώσσα για την περιγραφή των RDF λεξιλογίων, η αλλιώς των οντολογιών. Οι χρήστες, μέσω της RDFS είναι υπεύθυνοι για τον ορισμό της δικής τους ορολογίας. Η RDFS παρέχει διευκολύνσεις για την περιγραφή ομάδων σχετικών πόρων, που ονομάζονται κλάσεις, και για την περιγραφή των σχέσεων μεταξύ των πόρων, που ονομάζονται ιδιότητες . Τέλος, πρέπει να διευκρινίζουμε ότι η RDFS είναι μια επέκταση της RDF,ένα εξειδικευμένο

και προκαθορισμένο σύνολο RDF πόρων. Στη συνέχεια παραθέτουμε τα βασικά στοιχεία μοντελοποίησης της RDFS.

2.3.7.1 Επεκτάσεις του RDF Schema που αφορούν κλάσεις

Πόροι :

rdfs:Class, η κλάση όλων των κλάσεων (ακόμα και του εαυτού της).

rdfs:Resource, η κλάση όλων των πόρων.

Ιδιότητες (κατηγορήματα):

rdfs:subClassOf, η ιδιότητα που συσχετίζει μία κλάση με τις υπερκλάσεις της. Σημειώνουμε ότι στην RDFS επιτρέπεται η πολλαπλή κληρονομικότητα (multiple inheritance). Τέλος είναι μεταβατική.

rdf:type, η ιδιότητα που συνδέει έναν πόρο με κάποια κλάση στην οποία ανήκει, ή ισοδύναμα συνδέει έναν πόρο με κάποια κλάση της οποίας είναι στιγμιότυπο.

2.3.7.2 Επεκτάσεις του RDF Schema που αφορούν ιδιότητες

Πόροι:

- rdfs:Datatype, χρησιμοποιείται ως αντικείμενο για το rdf:type και δηλώνει ότι ο πόρος είναι ένας τύπος δεδομένων.
- rdf:Property, χρησιμοποιείται ως αντικείμενο με το rdf:type για να δηλώσει νέες ιδιότητες.

Ιδιότητες:

- rdfs:domain, η ιδιότητα που συσχετίζει μία ιδιότητα με κάθε μία από τις κλάσεις του πεδίου ορισμού της και δηλώνει ότι όλοι οι πόροι που έχουν μία δεδομένη ιδιότητα είναι στιγμιότυπα των κλάσεων του πεδίου ορισμού της.
- rdfs:range, η ιδιότητα που συσχετίζει μία ιδιότητα με κάθε μία από τις κλάσεις του πεδίου τιμών της, δηλώνοντας ότι όλοι οι πόροι που είναι τιμές μίας δεδομένης ιδιότητας είναι στιγμιότυπα των κλάσεων του πεδίου τιμών της.
- rdfs:subPropertyOf, η ιδιότητα που συσχετίζει μία ιδιότητα με τις υπεριδιότητές της. Τέλος είναι μεταβατική.

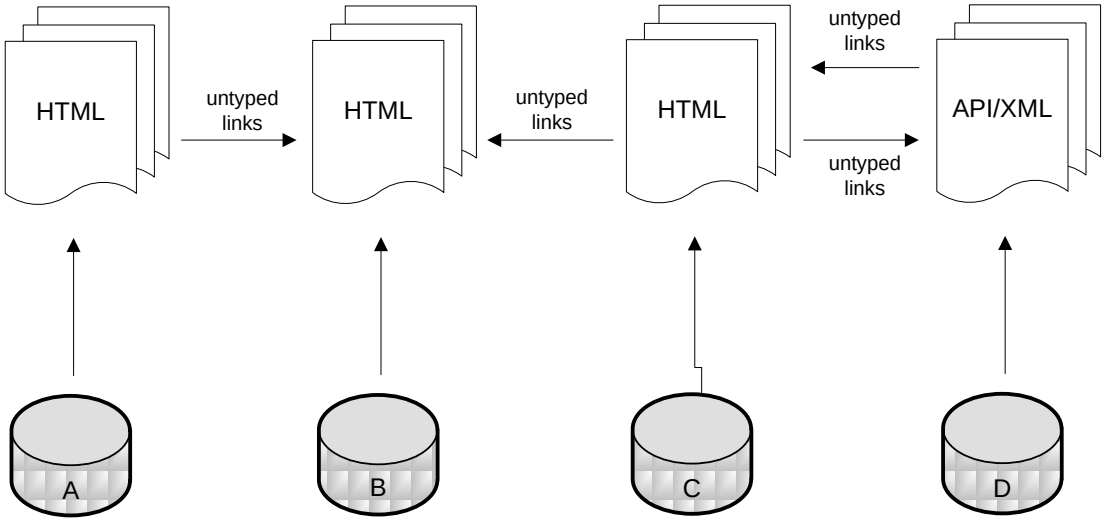
Άλλες διευκολύνσεις του RDF Schema

- rdfs:comment,εισάγει ένα σχόλιο για ανάγνωση από κάποιο χρήστη.
- rdfs:label,δίνει ένα όνομα για κάποιο πόρο, το οποίο μπορεί να διαβαστεί από κάποιο άνθρωπο.

- rdfs:seeAlso,χρησιμοποιείτε για να υποδείξει μια πηγή, η οποία μπορεί να παρέχει επιπλέον πληροφορία για το αντικείμενο.
- rdfs:isDefinedBy,είναι υποσύνολο της rdfs:seeAlso και χρησιμοποιείτε για να υποδείξει μια πηγή που την πηγή του πόρου.

2.4 Συνδεδεμένα Δεδομένα (Linked Data)

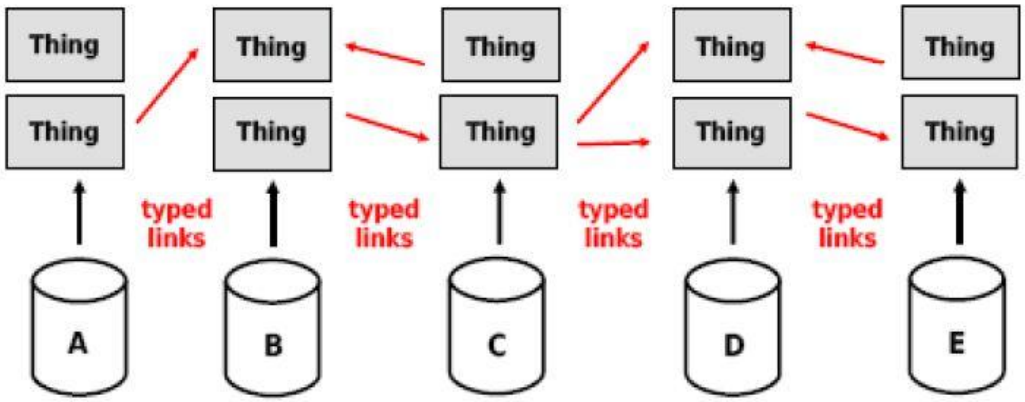
Στον Παγκόσμιο Ιστό μέχρι πρόσφατα τα δεδομένα ήταν διαθέσιμα σε ακατέργαστες βάσεις, με την μορφή CSV, XML ή σε πίνακες HTML, θυσιάζοντας έτσι την δομή τους αλλά και τα μεταδεδομένα. Ακόμα τα δεδομένα είναι συνήθως σε κλειστές και αποκομμένες μεταξύ τους βάσεις δεδομένων. Έτσι ο μοναδικός τρόπος εμφάνισης των δεδομένων στον Ιστό είναι μέσα από HTML σελίδες που απευθύνονται κυρίως σε ανθρώπινο χρήστη Για τους λόγους αυτούς, ο Παγκόσμιος Ιστός λέγεται και Ιστός των Εγγράφων (Web of Documents).



Εικόνα 2.3.7.1 Ο Ιστός των Εγγράφων (11)

Τα τελευταία χρόνια ο Ιστός έχει εξελιχθεί από ένα παγκόσμιο πληροφοριακό σύμπαν από διασυνδεδεμένα αρχεία, σε ένα όπου και τα δεδομένα και τα αρχεία είναι συνδεδεμένα. Πίσω από αυτήν την εξέλιξη βρίσκετε ο Tim Berners-Lee, ο οποίος πρότεινε μια σειρά από βέλτιστες πρακτικές για την δημοσίευση και την σύνδεση δομημένων δεδομένων στον Ιστό. Ο όρος που χρησιμοποιήθηκε για αυτά είναι Συνδεδεμένα Δεδομένα (Linked Data) (12). Η νέα μορφή του Ιστού των Εγγράφων δίνει δυνατότητα για νέους τύπους εφαρμογών, όπως οι ειδικοί φυλλομετρητές που θα επιτρέπουν στους χρήστες να πλοηγούνται από μία πηγή δεδομένων σε άλλες παρεμφερής πηγές μέσω συνδέσεων που θα υπάρχουν. Η δυνατότητα των νέων

εφαρμογών, να πλοηγούνται σε έναν αδέσμευτο παγκόσμιο χώρο δεδομένων θα τους δίνει την δυνατότητα να δίνουν ποίο ολοκληρωμένες απαντήσεις. Στο παρακάτω σχήμα παρουσιάζεται ο Ιστός των Συνδεδεμένων Δεδομένων:



Εικόνα 2.3.7.2 Ο Ιστός των Συνδεδεμένων Δεδομένων (11)

2.4.1 Οι κανόνες των Συνδεδεμένων Δεδομένων

Το 2006 ο Berners-Lee έγραψε ένα σημείωμα στο οποίο περιέγραφε ένα σύνολο από κανόνες για την δημοσίευση δεδομένων στον Ιστό. Για να μεταβεί ο Παγκόσμιος Ιστός από τον Ιστό των Εγγράφων στον ενιαίο Ιστό των Συνδεδεμένων Δεδομένων (13):

1. Χρησιμοποίηση URI για την ταυτοποίηση όλων των αντικειμένων..
2. Χρησιμοποίηση HTTP URIs, ώστε οι χρήστες να μπορούν να ζητήσουν πληροφορίες για τους πόρους που αυτά αντιπροσωπεύουν χρησιμοποιώντας το πρωτόκολλο HTTP.
3. Παροχή χρήσιμων πληροφοριών για ένα αντικείμενο, όταν κάποιος ζητάει κάποιο URI, κάνοντας χρήση των προτύπων (RDF/XML,SPARQL).
4. Πρέπει να περιλαμβάνονται συνδέσεις σε άλλα σχετικά URIs, ώστε ο χρήστης να μπορεί να ανακαλύψει περισσότερα πράγματα στον Ιστό.

Τα παραπάνω έχουν γίνει γνωστά ως «Αρχές των Συνδεδεμένων Δεδομένων» (Linked Data principles). Εδώ πρέπει να σημειώσουμε ότι το 2009, τροποποιήθηκαν από τον Berners-Lee σε τρείς πολύ απλούς κανόνες (14) :

1. Όλοι οι τύποι αντικειμένων, τώρα έχουν ονόματα που αρχίζουν με HTTP.
2. Παίρνω σημαντικές πληροφορίες πίσω. Σε κάποιο τυποποιημένο πρότυπο, θα λάβω πίσω δεδομένα. Αυτά θα είναι χρήσιμα δεδομένα για κάποιον που θα θέλει να μάθει για το αντικείμενο αυτό.

3. Θα παίρνω πίσω πληροφορίες που δεν θα περιορίζονται στο ύψος και στο βάρος κάποιου ή πότε γεννήθηκε αλλά θα περιλαμβάνουν και σχέσεις. Και όταν θα υπάρχουν αυτές οι σχέσεις μεταξύ των αντικειμένων, το σχετικό αντικείμενο θα δίνεται με όνομα και αυτό που αρχίζει με HTTP.

Να σημειώσουμε εδώ ότι παρόλο που στον δεύτερο κανόνα αναφέρεται η χρήση τυποποιημένων πρότυπων, αυτά δεν περιορίζονται σε κάποιο συγκεκριμένο πρότυπο, όπως το RDF/XML.

2.4.2 Αρχιτεκτονική Ιστού με Συνδεδεμένα Δεδομένα

Στη προηγούμενη ενότητα αναφερθήκαμε στους βασικούς κανόνες των Συνδεδεμένων Δεδομένων, όπως αυτοί παρουσιάστηκαν από τον εμπνευστή τους. Δύο βασικές τεχνολογίες που βασίζονται τα Συνδεδεμένα Δεδομένα είναι τα Ενιαία Αναγνωριστικά Πόρων (Uniform Resource Identifiers-URIs) και το Πρωτόκολλο Μεταφοράς Υπερκειμένου (Hypertext Transfer Protocol - HTTP) . Επίσης η RDF που αναλύσαμε στην ενότητα 2.3 είναι το βασικό μοντέλο, βασισμένο σε γράφο, που χρησιμοποιείται για την δομή και την σύνδεση των δεδομένων .Όπως είδαμε το RDF μοντέλο αναλύει τα δεδομένα σε τριάδες, όπου υποκείμενο και αντικείμενο έχουν URIs για την ταυτοποίηση του πόρου που περιγράφουν και το κατηγόρημα, που περιγράφει τη σχέση τους αναπαρίσταται και αυτό από ένα URI.

Η χρήση του μοντέλου RDF για τον Ιστό των Συνδεδεμένων Δεδομένων παρέχει αρκετά οφέλη. Ο πελάτης μπορεί να επισκεφθεί όλα τα URI ενός RDF γράφου με στόχο την ανάκτηση των επιπλέων πληροφοριών. Οι πληροφορίες από διαφορετικές πηγές συγχωνεύονται με φυσικό τρόπο αφού το μοντέλο αυτό επιτρέπει να θέσουμε RDF συνδέσμους από διαφορετικές πηγές. Όμως προκειμένου να κάνουμε εύκολο για τους πελάτες την συγχώνευση και την αναζήτηση δεδομένων πρέπει να προσέχουμε να μην χρησιμοποιούμε το πλήρες μοντέλο της RDF γιατί υπάρχουν περιπτώσεις που δημιουργεί προβλήματα. Για παράδειγμα η χρήση κενών κόμβων(blank nodes) είναι καλό να αποφεύγεται μια και είναι αδύνατο να θέσουμε εξωτερικούς RDF συνδέσμους σε έναν τέτοιο κόμβο. Η συγχώνευση δεδομένων γίνεται εξαιρετικά δύσκολη με την χρήση κενών κόμβων. Άλλωστε όλες οι πηγές πρέπει να αντιπροσωπεύονται από κάποιο URI. Αποθαρρύνεται επίσης η χρήση RDF reification επειδή η σημασιολογία της είναι ασαφής ενώ τα ερωτήματα με SPARQL γίνονται πιο πολύπλοκα. Μεταδεδομένα μπορούν να επικολληθούν στους πληροφοριακούς πόρους

ως εναλλακτική. Τέλος Οι RDF Containers και collections δεν λειτουργούν καλά με τη SPARQL. Οπότε θα πρέπει να χρησιμοποιηθούν μόνο όταν είναι αναγκαίο.

2.4.3 Πληροφοριακοί και μη-πληροφοριακοί πόροι

Στην ορολογία της Αρχιτεκτονικής του Ιστού όλα τα αντικείμενα που μας ενδιαφέρουν, ονομάζονται πόροι. Στον Ιστό των Δεδομένων οι πόροι χωρίζονται σε πληροφοριακούς (informational resources) και μη-πληροφοριακούς (non-informational resources) ή αλλιώς «άλλους» πόρους (15). Στον ιστό των Συνδεδεμένων Δεδομένων η διαφορά αυτή είναι αρκετά σημαντική και ορίστηκε στην αντίστοιχη σύσταση της W3C (16) .

Όλοι οι πόροι που βρίσκουμε στο παραδοσιακό Ιστό, όπως έγγραφα, εικόνες και άλλα αρχεία μέσων, είναι πληροφοριακοί πόροι. Οι πληροφοριακοί πόροι έχουν μία ή περισσότερες αναπαραστάσεις, για παράδειγμα, σε διαφορετικές μορφές, ποιότητες ευκρίνειας ή φυσικές γλώσσες. Μια αναπαράσταση είναι μια ροή bytes σε συγκεκριμένη μορφή, όπως HTML, RDF/XML ή JPEG. Η προσπέλαση αυτών γίνεται χρησιμοποιώντας το πρωτόκολλο HTTP. Τέλος, οι αναπαραστάσεις μπορούν να μεταδοθούν μέσα σε κάποιο μήνυμα. Όλοι οι άλλοι πόροι που μένουν αν αφαιρέσουμε τους πληροφοριακούς πόρους είναι οι μη-πληροφοριακοί πόροι. Σε αυτή την κατηγορία πόρων περιλαμβάνονται οι άνθρωποι, οι ιδέες, όλα τα φυσικά αντικείμενα του κόσμου μας και γενικά οτιδήποτε δεν περιλαμβάνεται στο χώρο του Ιστού. Οι πόροι αυτοί, σε αντίθεση με τους πληροφοριακούς, δεν έχουν αναπαράσταση.

2.4.4 Dereferencing των HTTP URIs

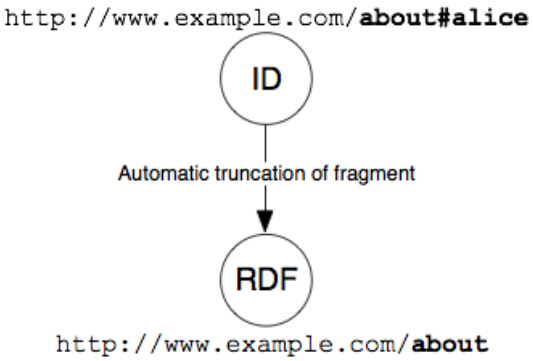
Η διαδικασία της «επίσκεψης» ενός URI στον Ιστό με σκοπό να ανακτήσουμε πληροφορίες σχετικά με τον αναφερόμενο πόρο λέγεται dereferencing HTTP URIs και χρησιμοποιεί έναν HTTP μηχανισμό που ονομάζεται διαπραγμάτευση περιεχομένου (content negotiation). Το προσχέδιο του W3C TAG εισήγαγε μια διάκριση στο πώς γίνεται dereferencing στα URIs που ταυτοποιούν πληροφοριακούς πόρους και μη-πληροφοριακούς πόρους (17):

- Πληροφοριακοί Πόροι: Όταν επισκεφθούμε ένα URI που ταυτοποιεί έναν πληροφοριακό πόρο, ο εξυπηρετητής του ιδιοκτήτη του URI συνήθως παράγει μια νέα αναπαράσταση, μια νέα στιγμιαία αναπαράσταση της τρέχουσας κατάστασης της πληροφοριακής πηγής και τη στέλνει πίσω στον πελάτη χρησιμοποιώντας τον HTTP κωδικό απόκρισης 200 OK.

- Μη-πληροφοριακοί πόροι: αυτούς δεν μπορούμε να τους επισκεφθούμε απευθείας. Επομένως, η αρχιτεκτονική του Ιστού χρησιμοποιεί ένα κόλπο για να επιτρέψει την επίσκεψη στα URIs που ταυτοποιούν μη-πληροφοριακούς πόρους: αντί να αποστέλλεται μια αναπαράσταση του πόρου, ο εξυπηρετητής στέλνει στον πελάτη το URI ενός πληροφοριακού πόρου που περιγράφει τον μη-πληροφοριακό πόρο, χρησιμοποιώντας τον HTTP κωδικό απόκρισης 303 See Other. Αυτό ονομάζεται 303 ανακατεύθυνση. Σε ένα δεύτερο βήμα, ο πελάτης επισκέπτεται αυτό το νέο URI και παίρνει μια αναπαράσταση που περιγράφει τον αρχικό μη-πληροφοριακό πόρο.

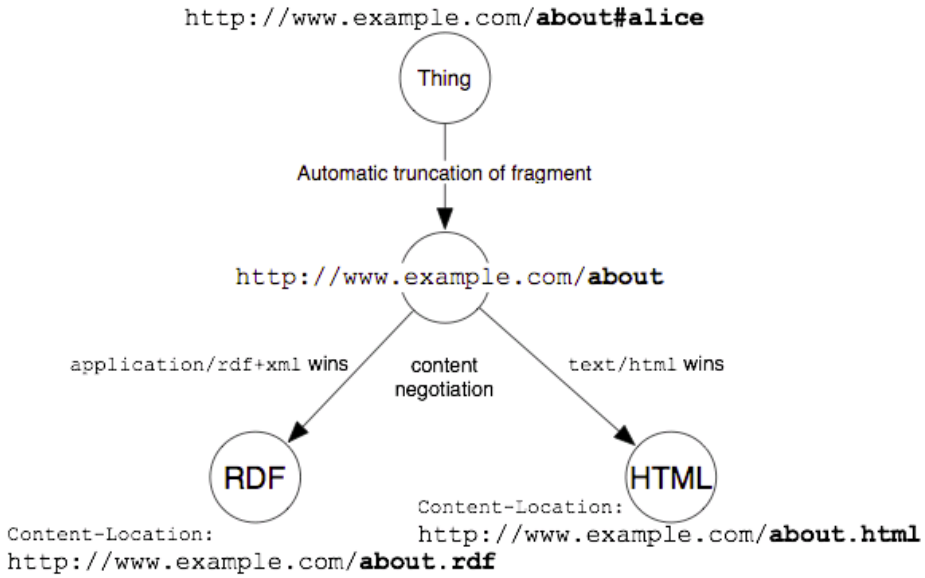
Υπάρχουν δύο προσεγγίσεις για το πώς μπορούν οι εκδότες δεδομένων να παρέχουν στους πελάτες URIs πληροφοριακών πόρων που να περιγράφουν μη-πληροφοριακούς πόρους: τα Hash (δίεση) URIs και τα 303 URIs (18).

Για πόρους, που δεν είναι έγγραφα, η λύση είναι τα URIs με δίεση (#), ή αλλιώς hash URIs . Το σύμβολο '#' χωρίζει σε δύο μέρη, το μέρος πριν το '#' και το μέρος μετά. Όταν ένας πελάτης αναζητά κάποιο hash URI, το πρωτόκολλο HTTP επιβάλλει την αφαίρεση του τμήματος μετά το '#' πριν την αποστολή της αίτησης στον εξυπηρετητή. Φυσικά για τον εξυπηρετητής το κομμάτι του hash URI πριν το '#' ,αντιστοιχεί σε κάποιον πληροφοριακό πόρο που περιγράφει όλους τους πόρους των οποίων το hash URI προέρχεται από το δικό του.



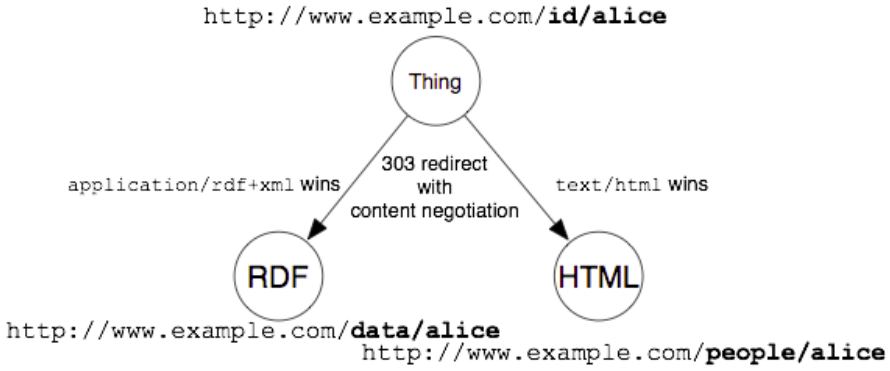
Εικόνα 2.4.4.1 Η λύση του hash URI χωρίς διαπραγμάτευση περιεχομένου

Μετά την αποστολή του κομματιού του hash URI πριν το '#' στον εξυπηρετητή ακολουθεί έμμεση <επίσκεψη> με διαπραγμάτευση περιεχομένου.



Εικόνα 2.4.4.2 Η λύση του hash URI με διαπραγμάτευση περιεχομένου

Η δεύτερη λύση είναι η χρήση του HTTP κωδικού 303 See Other ως ένδειξη ότι ο ζητούμενος πόρος δεν είναι έγγραφο και δε συνδέεται με κάποια αναπαράσταση . Ο κωδικός 303 ανακατευθύνει τον πελάτη σε κάποιο URI, το οποίο είτε αντιστοιχεί σε ένα έγγραφο που περιγράφει τον πόρο που αρχικά αναζήτησε ο πελάτης, μετά από διαπραγμάτευση περιεχομένου είτε αποτελεί την αφετηρία για να ανακατευθυνθεί σε ένα τέτοιο στη συνέχεια, μετά από διαπραγμάτευση περιεχομένου.



Εικόνα 2.4.4.3 303 Ανακατεύθυνση σε έγγραφο με διαπραγμάτευση περιεχομένου

Ο πίνακα που ακολουθεί συνοψίζει τις πιθανές απαντήσεις που μπορεί να λάβει ο πελάτης όταν ανατρέχει σε κάποιο URI:

Πίνακας 2-1 Πιθανές απαντήσεις μετά την προσπάθεια προσπέλασης ενός URI

HTTP Κωδικός Απάντησης	Αντικείμενο που επιστρέφεται	Συμπέρασμα
200 (success)	Μία αναπαράσταση του ζητούμενου πόρου	Ο πόρος είναι ένας πληροφοριακός πόρος και μία αναπαράστασή του έχει επιστραφεί.
303 (see other)	Ένα URI	Ο πόρος μπορεί να είναι είτε πληροφοριακός είτε μη-πληροφοριακός. Υπάρχει ένας σχετιζόμενος με αυτόν πόρος, του οποίου το URI έχει επιστραφεί. Ο σχετιζόμενος πόρος μπορεί να είναι είτε πληροφοριακός είτε μη-πληροφοριακός.
4XX ή 5XX (error)	Τίποτα	Δεν μπορούμε να συμπεράνουμε τίποτα για τη φύση του πόρου.

2.4.5 Διαπραγμάτευση Περιεχομένου (Content Negotiation)

Οι HTML φυλλομετρητές συνήθως εμφανίζουν τις RDF αναπαραστάσεις ως «γυμνό» κώδικα RDF ή απλά τις «κατεβάζουν» ως αρχεία RDF χωρίς να τις εμφανίζουν. Αυτό δεν είναι πολύ βοηθητικό για το μέσο χρήστη. Επομένως, η παρουσίαση μιας σωστής HTML αναπαράστασης σε συνδυασμό με την RDF αναπαράσταση ενός πόρου βοηθάει τον άνθρωπο να καταλάβει σε τί αναφέρεται ένα URI. Επομένως, διαπραγμάτευση περιεχομένου είναι η διαδικασία επιλογής της καλύτερης αναπαράστασης για μια δεδομένη απόκριση όταν υπάρχουν διαθέσιμες πολλαπλές αναπαραστάσεις. Οι πελάτες HTTP στέλνουν κεφαλίδες HTTP με κάθε αίτηση για να δηλώσουν τί είδος αναπαράστασης προτιμούν. Εάν οι κεφαλίδες δηλώνουν ότι ο πελάτης προτιμά HTML, τότε ο εξυπηρετητής μπορεί να παράγει μια HTML αναπαράσταση. Εάν ο πελάτης προτιμά RDF, τότε ο εξυπηρετητής μπορεί να παράγει RDF.

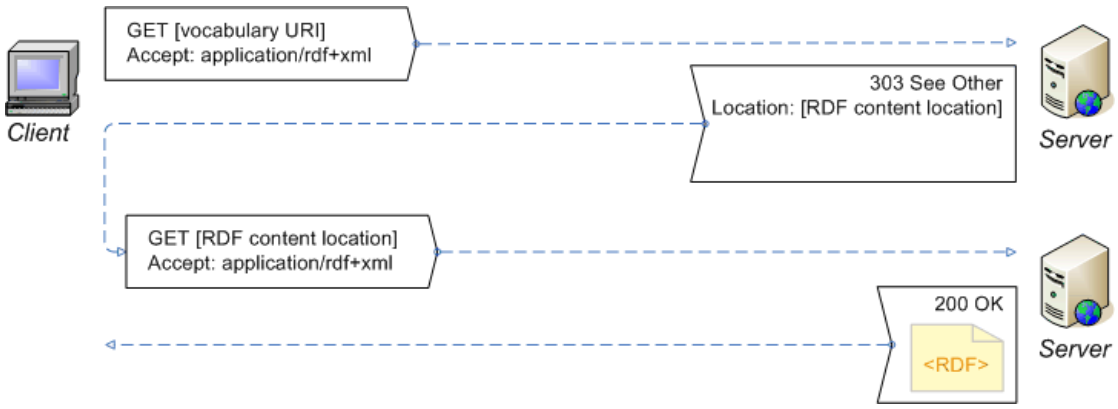
Η διαπραγμάτευση περιεχομένου για μη- πληροφοριακούς πόρους υλοποιείται συνήθως με τον ακόλουθο τρόπο. Όταν επισκεφτούμε ένα URI που ταυτοποιεί ένα μη- πληροφοριακό πόρο, ο εξυπηρετητής στέλνει μια ανακατεύθυνση 303 προς έναν πληροφοριακό πόρο κατάλληλο για τον πελάτη. Επομένως, μια πηγή δεδομένων συχνά εξυπηρετεί τρία URIs που σχετίζονται με μη- πληροφοριακούς πόρους, για

παράδειγμα:

http://www4.wiwiss.fu-berlin.de/factbook/resource/Russia (URI που ταυτοποιεί το μη- πληροφοριακό πόρο Ρωσία) http://www4.wiwiss.fu-berlin.de/factbook/data/Russia (πληροφοριακός πόρος με μια RDF/XML αναπαράσταση που περιγράφει τη Ρωσία) http://www4.wiwiss.fu-berlin.de/factbook/page/Russia (πληροφοριακός πόρος με μια HTML αναπαράσταση που περιγράφει τη Ρωσία)

Η εικόνα που ακολουθεί δείχνει μια επίσκεψη σε ένα HTTP URI το οποίο ταυτοποιεί ένα μη- πληροφοριακό πόρο και χρησιμοποιεί τη διαπραγμάτευση περιεχομένου:

1. Ο πελάτης πραγματοποιεί ένα HTTP GET αίτημα προς ένα URI το οποίο ταυτοποιεί ένα μη-πληροφοριακό πόρο. Στην περίπτωσή μας, ένα URI λεξιλογίου. Εάν ο πελάτης είναι φυλλομετρητής Συνδεδεμένων Δεδομένων και προτιμά μια RDF/XML αναπαράσταση του πόρου, στέλνει μια κεφαλίδα Accept:application/rdf+xml μαζί με την αίτηση. Οι HTML φυλλομετρητές θα έστελναν αντί γι' αυτό μια κεφαλίδα Accept: text/html.
2. Ο εξυπηρετητής αναγνωρίζει το URI που ταυτοποιεί ένα μη-πληροφοριακό πόρο. Καθώς ο εξυπηρετητής δεν μπορεί να επιστρέψει μια αναπαράσταση αυτού του πόρου, απαντά χρησιμοποιώντας τον HTTP κωδικό απόκρισης 303 See Other και στέλνει στον πελάτη το URI ενός πληροφοριακού πόρου που περιγράφει το μη- πληροφοριακό πόρο. Στην περίπτωση του RDF: RDF content location.
3. Ο πελάτης τώρα ζητά από τον εξυπηρετητή να λάβει (GET) μια αναπαράσταση του πληροφοριακού πόρου, ζητώντας ξανά application/rdf+xml.
4. Ο εξυπηρετητής στέλνει στον πελάτη ένα έγγραφο RDF/XML το οποίο περιέχει μια περιγραφή του αρχικού πόρου vocabulary URI.



Εικόνα 2.4.5.1 Διαπραγμάτευση περιεχομένου

2.4.6 Ταυτόσημα URIs (URI aliases)

Σε ένα ανοιχτό περιβάλλον όπως είναι ο Ιστός, συχνά συμβαίνει διαφορετικοί πάροχοι πληροφοριών να αναφέρονται στον ίδιο μη-πληροφοριακό πόρο, όπως π.χ. στην ίδια γεωγραφική τοποθεσία ή στο ίδιο διάσημο πρόσωπο. Καθώς δε γνωρίζουν ο ένας την ύπαρξη του άλλου εισάγουν διαφορετικά URIs για την ταυτοποίησή του ίδιου αντικειμένου. Για παράδειγμα, η DBpedia, μια πηγή δεδομένων που παρέχει πληροφορίες οι οποίες έχουν εξαχθεί από τη Wikipedia, χρησιμοποιεί το URI http://dbpedia.org/resource/Berlin για να ταυτοποιεί το Βερολίνο. Το Geonames, μια πηγή δεδομένων που παρέχει πληροφορίες για εκατομμύρια γεωγραφικές τοποθεσίες, χρησιμοποιεί το URI http://sws.Geonames.org/2950159/ για να ταυτοποιεί το Βερολίνο. Καθώς και τα δύο URIs αναφέρονται στον ίδιο μη- πληροφοριακό πόρο, καλούνται ταυτόσημα URI.

Τα ταυτόσημα URIs συνηθίζονται στον Ιστό των Δεδομένων καθώς δεν μπορεί να περιμένουμε ότι όλοι οι πάροχοι πληροφοριών θα συμφωνήσουν πάνω στο ίδιο URI για να ταυτοποιήσουν μια οντότητα. Τα ταυτόσημα URIs παρέχουν μια σημαντική κοινωνική λειτουργία στον Ιστό Δεδομένων καθώς αναφέρονται σε διαφορετικές περιγραφές του ίδιου μη- πληροφοριακού πόρου και έτσι επιτρέπουν την έκφραση διαφορετικών όψεων και απόψεων. Για να είμαστε σε θέση να εντοπίσουμε ότι οι διαφορετικοί πάροχοι πληροφοριών μιλούν για τον ίδιο μη-πληροφοριακό πόρο, είναι κοινή πρακτική, για τους παρόχους πληροφοριών, να θέτουν συνδέσμους owl:sameAs προς ταυτόσημα URI τα οποία γνωρίζουν.

2.4.7 Επιλέγοντας URIs

Ένα από τα σημαντικότερα θεμέλια για τα Συνδεδεμένα Δεδομένα είναι όλοι οι πόροι να έχουν ως όνομα ένα αναγνωριστικό URI. Τα URIs σύμφωνα με τον Tim Berners-

Lee πρέπει να είναι «καλά» (“cool”) (18). Οι σχεδιαστές των δεδομένων θα πρέπει να παρέχουν URIs με βάση την απλότητα, τη σταθερότητα και τη δυνατότητα διαχείρισής τους. Τα σύντομα, μνημονικά URIs δε θα «σπάσουν» τόσο εύκολα, όταν διακινούνται στον Ιστό και είναι γενικά ευκολότερα στην απομνημόνευση. Ακόμα, τα URIs θα πρέπει να διατηρούνται σταθερά στο χρόνο όσο το δυνατό περισσότερο, έχοντας στο νου τα επόμενα δέκα ή είκοσι χρόνια. Τα URIs πρέπει να δημιουργούνται με τρόπο που να μπορούμε να τα διαχειριστούμε. Είναι προτιμότερο ο προσδιορισμός των URIs να γίνεται σε ένα HTTP όνομα πεδίου το οποίο έχουμε υπό έλεγχο, όπου μπορούμε πραγματικά να τα κάνουμε dereferenceable. Συχνά καταλήγουμε με τρία URIs που σχετίζονται με έναν και μόνο μη-πληροφοριακό πόρο:

- ένα αναγνωριστικό για τον πόρο,
- ένα αναγνωριστικό για ένα σχετιζόμενο πληροφοριακό πόρο κατάλληλο για HTML φυλλομετρητές (με μια αναπαράσταση ιστοσελίδας),
- ένα αναγνωριστικό για ένα σχετιζόμενο πληροφοριακό πόρο κατάλληλο για RDF φυλλομετρητές (με μια RDF/XML αναπαράσταση).

Στη συνέχεια παραθέτουμε κάποιες προτάσεις για την επιλογή σχετιζόμενων URIs:

1. http://dbpedia.org/resource/Berlin
2. http://dbpedia.org/page/Berlin
3. http://dbpedia.org/data/Berlin

Τέλος δύο παραδείγματα καλών URIs:

http://dbpedia.org/resource/Boston,

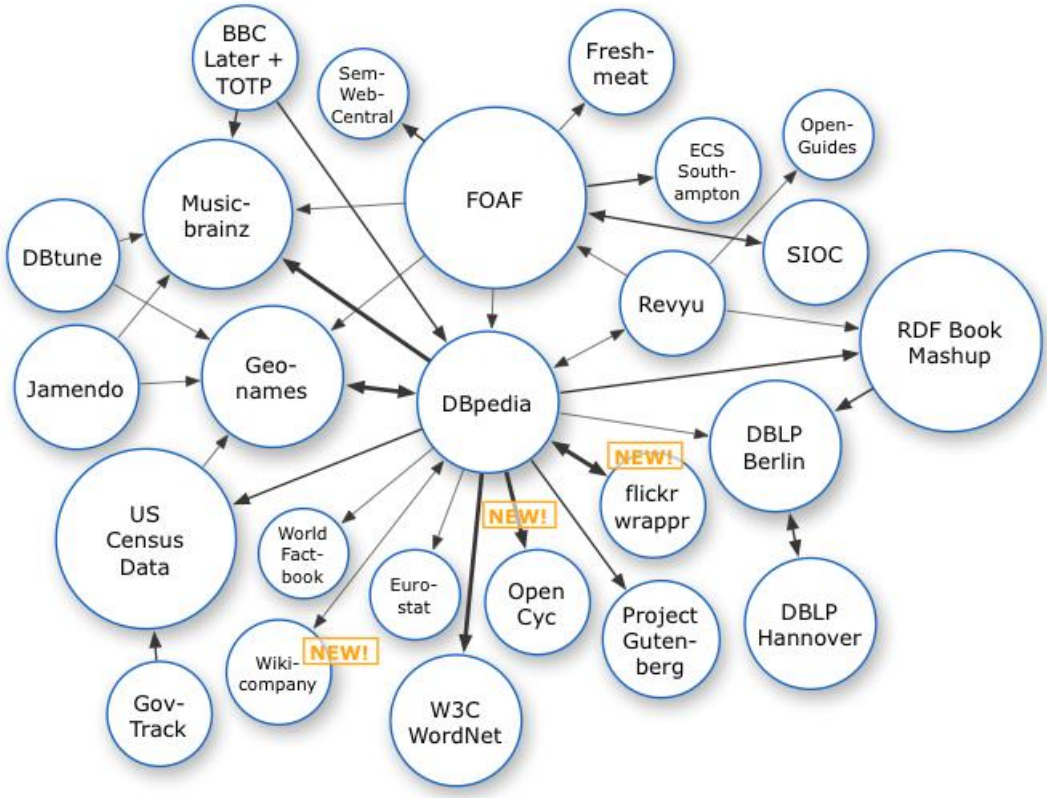
http://www4.wiwiss.fu-berlin.de/bookmashup/books/006251587X

2.4.8 Ανασκόπηση των κύριων ομάδων εργασίας των Συνδεδεμένων Δεδομένων

Στις παραπάνω ενότητες είδαμε κάποια τεχνικά στοιχεία για την υλοποίηση των Συνδεδεμένων Δεδομένων. Αυτά θα είναι χρήσιμα στον αναγνώστη στο επόμενο κεφάλαιο, όπου θα δούμε τα βασικά βήματα μιας μεθοδολογίας που πρέπει να ακολουθήσουμε προκειμένου να δημοσιευτεί ένα σύνολο δεδομένων μας ως Συνδεδεμένα Δεδομένα στον Ιστό. Σ’ αυτό το σημείο, κρίθηκε χρήσιμο, να αναφερθούμε στις σημαντικότερες ομάδες εργασίας πάνω στον χώρο αυτό και να τις παρουσιάσουμε συνοπτικά.

2.4.8.1 Linking Open Data project

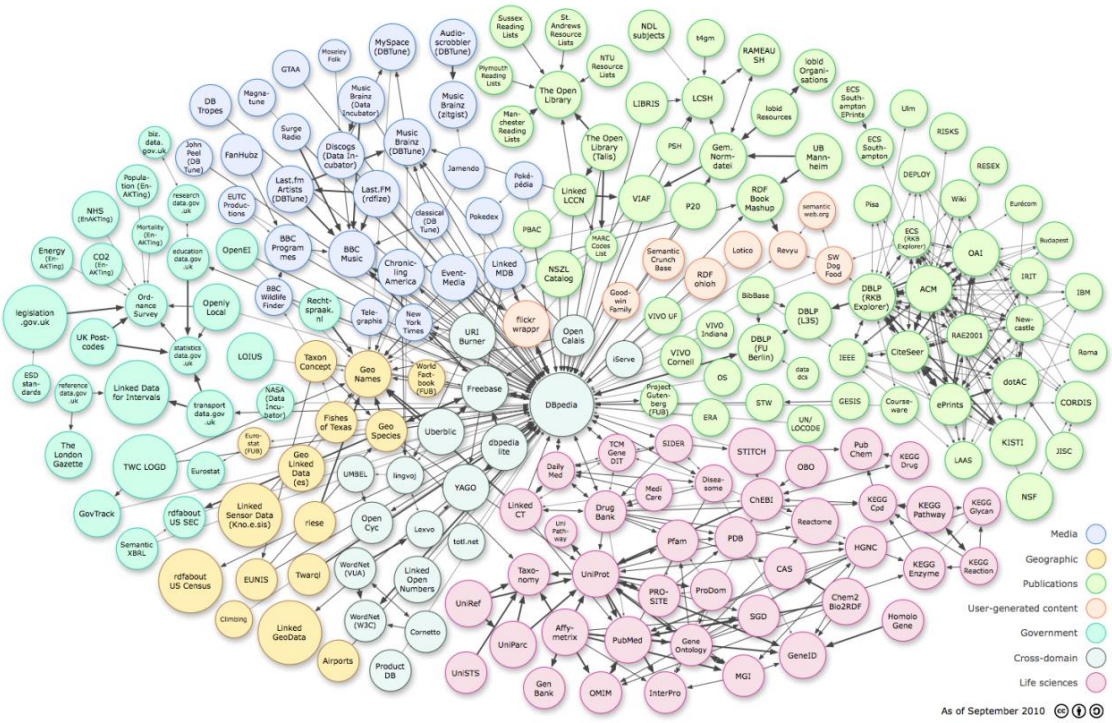
Πρόκειται για το ποιο σημαντικό project, αυτό της «Σύνδεση Ανοιχτών Δεδομένων» (Linking Open Data - LOD) (19). Τον Ιανουάριο του 2007 από τους Chris Bizer και Richard Cyganiak, δημιουργήθηκε ένα κοινωνικό project και υποστηρίχτηκε από το W3C Semantic Web Education και το Outreach Group. Στόχος του είναι να επεκτείνει τον Ιστό των Δεδομένων, με την έκδοση πολλών συνόλων δεδομένων σαν RDF στον Ιστό. Επιπλέον όμως, θέτει RDF συνδέσεις μεταξύ δεδομένων από διαφορετικές πηγές. Αρχικά, το σύννεφο αυτό από συνδεδεμένα δεδομένα αποτελούταν από 2 δισεκατομμύρια τριάδες RDF, οι οποίες δημιουργούσαν πάνω από δύο εκατομμύρια συνδέσεις (Οκτώβριος 2007).



Εικόνα 2.4.8.1 Σύννεφο Linking Open Data project, Οκτώβριος 2007

Από τότε το project, υποστηρίχτηκε αρχικά από ερευνητές σε πανεπιστήμια και μικρές εταιρίες και στη συνέχεια από μεγάλους οργανισμούς όπως το BBC, το Thomson Reuters και τη Βιβλιοθήκη του Κογκρέσου. Το project ήταν ανοιχτό και μπορούσε να συμμετάσχει ο καθένας, δημοσιεύοντας απλά ένα σύνολο δεδομένων σύμφωνα με τους κανόνες των Συνδεδεμένων Δεδομένων και διασυνδέοντάς το με τα υπάρχοντα σύνολα δεδομένων. Πολύ σύντομα το σύννεφο που είδαμε παραπάνω μεγάλωσε και το εύρος του το Σεπτέμβρη του 2010 είχε μεγαλώσει σε 25

δισεκατομμύρια τριάδες RDF, οι οποίες δημιουργούσαν πάνω από 397 εκατομμύρια συνδέσεις .



Εικόνα 2.4.8.2 Σύννεφο Linking Open Data project, 22 September 2010

Τα βέλη στο παραπάνω σχήμα υποδεικνύουν ότι υπάρχουν σύνδεσμοι μεταξύ αντικειμένων στα δύο συνδεόμενα σύνολα δεδομένων. Τα πιο «παχιά» βέλη αντιστοιχούν σε μεγαλύτερο αριθμό συνδέσμων μεταξύ δύο συνόλων δεδομένων, ενώ τα αμφίδρομα βέλη υποδεικνύουν ότι σε κάθε σύνολο δεδομένων υπάρχουν εξερχόμενοι σύνδεσμοι προς το άλλο. Το περιεχόμενο του σύννεφου έχει ποικίλη φύση, συνδυάζοντας δεδομένα από γεωγραφικές τοποθεσίες, εταιρίες, βιβλία, επιστημονικές δημοσιεύσεις, ταινίες, μουσική, τηλεοπτικά και ραδιοφωνικά προγράμματα, γονίδια, πρωτεΐνες, φάρμακα και κλινικές δοκιμές, διαδικτυακές κοινότητες, στατιστικά δεδομένα, δημογραφικά αποτελέσματα και αξιολογήσεις. Ακόμα υπάρχουν μερικά σύνολα δεδομένων που χρησιμεύουν ως συνδετικοί κόμβοι στον Ιστό των Δεδομένων. Το σύνολο δεδομένων της DBpedia αποτελείται από RDF τριάδες που έχουν αποσπασθεί από τα «infoboxes» που υπάρχουν συνήθως στη δεξιά πλευρά των άρθρων της Wikipedia, ενώ το Geonames παρέχει RDF περιγραφές εκατομμυρίων γεωγραφικών τοποθεσιών παγκοσμίως. Καθώς αυτά τα δύο σύνολα δεδομένων παρέχουν URIs και RDF περιγραφές για πολλές συνηθισμένες οντότητες ή έννοιες, συχνά αναφέρονται και σε άλλα πιο εξειδικευμένα σύνολα δεδομένων και

έτσι έχουν εξελιχθεί σε κόμβους στους οποίους συνδέονται όλο και περισσότερα σύνολα δεδομένων.

2.4.8.2 Linked Open Data Around the Clock (LATC) - EU project

Η ευρωπαϊκή ένωση προκειμένου να υποστηρίξει την δημοσίευση και κατανάλωση των Συνδεδεμένων Ανοιχτών Δεδομένων, παρέχει βοήθεια μέσω του 7th Framework Programme, το οποίο έχει στόχο την ενίσχυση της έρευνας και της καινοτομίας μέσω της χρηματοδότηση της έρευνας (20).

Στόχοι αυτής του project είναι :

- Βελτίωση μιας συνεχούς λειτουργίας εφαρμογής με στόχο να παρακολουθεί τη χρησιμοποίησης και να βελτιώνει την ποιότητα των Συνδεδεμένων Ανοιχτών Δεδομένων.
- Παροχή μιας εύκολης πρόσβασης για τους εκδότες δεδομένων και τους καταναλωτές τους.
- Δημιουργία μιας βιβλιοθήκης με ανοιχτού κώδικα εργαλεία επεξεργασίας των δεδομένων.
- Υποστήριξη της κοινότητας με συμβουλές για βέλτιστες τεχνικές και σεμινάρια.

2.4.8.3 PlanetData - EU project

Το PlanetData (21) είναι ένα έργο χρηματοδοτούμενο από την ευρωπαϊκή ένωση μέσω του «Network of Excellence» που ασχολείται με τη διαχείριση δεδομένων μεγάλης κλίμακας .Αυτό το project περιλαμβάνει δεδομένα του Σημασιολογικού Ιστού (RDF) που είναι δημοσιευμένα με βάση τις αρχές των Συνδεδεμένων Δεδομένων. Η κοινότητα που αποτελείται από ακαδημαϊκούς και βιομηχανικούς συνεργάτες υποστηρίζει την έρευνα στη διαχείριση των δεδομένων μεγάλης κλίμακας. Αποτέλεσμα θα είναι ένα ολοκληρωμένο πρόγραμμα κατάρτισης, διάδοσης, τυποποίησης καθώς και η μεταφορά των αποτελεσμάτων της έρευνας στη βιομηχανία.

2.4.8.4 Linking Open Data 2 - EU project

Μέσω του 7th Framework Programme, ευρωπαική επιτροπη χρηματοδότησε το LOD2 project (22), με στόχο την συνέχιση της εργασίας πάνω στο project Linking Open Data που είδαμε παραπάνω μέχρι το 2014. Αυτό το έργο επικεντρώνει την

προσπάθεια του στην «Δημιουργία γνώσης μέσα από το διασυνδεδεμένα δεδομένα» με την ανάπτυξη σε:

- Εργαλεία και μεθοδολογίες έτοιμα για επιχειρήσεις με στόχο την έκθεση και διαχείριση μεγάλων ποσοτήτων δομημένης πληροφορίας στον Ιστό των Δεδομένων.
- Δημιουργία ενός υψηλής ποιότητας δικτύου με μία πλατφόρμα δοκιμών, σε πολλούς τομείς, με πολύγλωσσες οντολογίες βασισμένο σε πηγές όπως η Wikipedia και το OpenStreetMap.
- Αλγόριθμοι που βασίζονται στη μηχανική μάθηση με στόχο την αυτόματη διασύνδεση δεδομένων.
- Παροχή πρότυπων και μεθόδων για αξιόπιστη παρακολούθηση προέλευσης, εξασφαλίζοντας την ιδιωτικότητα και την ασφάλεια των δεδομένων, καθώς και για την αξιολόγηση της ποιότητας των πληροφοριών.
- Εργαλεία που προσαρμόζονται για την αναζήτηση, περιήγηση και συγγραφή των Συνδεδεμένων δεδομένων.

3

Θησαυροί και οντολογίες στην πολιτισμική κληρονομιά

Στο κεφάλαιο αυτό εξετάζουμε την περιοχή των θησαυρών και των ελεγχόμενων λεξιλογίων. Παρουσιάζουμε μια σειρά από γνωστά σύνολα δεδομένων, ευρωπαϊκών προγραμμάτων διαχείρισης θησαυρών και τέλος το συστήματα οργάνωσης γνώσης SKOS για τον Σημασιολογικό Ιστό.

3.1 Λεξιλόγια, θησαυροί και θεματικές επικεφαλίδες

Τα ελεγχόμενα λεξιλόγια παρέχουν έναν τρόπο για την οργάνωση της γνώσης και αποτελούν ένα σύνολο τυποποιημένων λέξεων ή φράσεων, που χρησιμοποιούνται για την εύρεση και την ανάκτηση πληροφορίας. Χρησιμοποιούνται σε θεματικές επικεφαλίδες (Subject headings), θησαυρούς (thesauri) και ταξινομίες (taxonomies). Η λειτουργία τους βασίζεται στην χρήση προκαθορισμένων, εξουσιοδοτημένων όρων, οι οποίοι έχουν προεπιλεγεί από το δημιουργό του λεξιλογίου. Αυτό έρχεται σε αντίθεση με τα φυσικά λεξιλόγια γλώσσας, όπου δεν υπάρχει κανένας περιορισμός σχετικά με το λεξιλόγιο.

Στην επιστήμη βιβλιοθηκονομίας και πληροφορίας τα ελεγχόμενα λεξιλόγια είναι μια λίστα από προσεκτικά επιλεγμένες λέξεις και φράσεις, οι οποίες χρησιμοποιούνται σαν ετικέτες για μονάδες πληροφορίας, ώστε αυτές να μπορούν να ανακτηθούν πιο εύκολα. Στόχος τον παραπάνω είναι να αντιμετωπιστεί το πρόβλημα της φυσικής γλώσσας, όπου στην ίδια έννοια μπορεί να δοθεί διαφορετική ονομασία. Έτσι τα ελεγχόμενα λεξιλόγια λύνουν προβλήματα όπως τη χρήση συνωνύμων.

Παραδείγματα ελεγχόμενων λεξιλογίων αποτελούν τα θεματικά περιγραφικά πεδία, όπως του Library of Congress Subject Headings (LCSH) αλλά και οι θησαυροί όρων, όπως ο Art & Architecture Thesaurus (AAT) .

Στην περιοχή του Σημασιολογικού Ιστού, τα ελεγχόμενα λεξιλόγια διευκολύνει την πρόσβαση σε πληροφορία που παράγεται από πολλούς διαφορετικούς δημιουργούς, ενώ ταυτόχρονα βοηθάει τον χρήστη όταν αναζητά πληροφορία για ένα πόρο, να βρίσκει νέους σχετικούς πόρους πάνω στην ίδια θεματική ενότητα.

Ένας θησαυρός είναι ένα έργο αναφοράς που παραθέτει λέξεις ομαδοποιούνται ανάλογα με την ομοιότητα της έννοια (που περιέχει συνώνυμα και, ενίοτε, αντώνυμα, σε αντίθεση με το λεξικό , το οποίο περιέχει τους ορισμούς και προφορές . Στο χώρο της επιστήμης της πληροφορικής, εξειδικευμένοι θησαυροί είναι σχεδιασμένοι για την ανάκτηση πληροφορίας. Αποτελούν ένα συγκεκριμένο τύπο ελεγχόμενου λεξιλογίου που δομείται με συγκεκριμένη σειρά, στον οποίο οι σχέσεις ισοδυναμίας, ιεραρχίας και συσχέτισης ανάμεσα στους όρους εμφανίζονται με σαφήνεια και αναγνωρίζονται με τυποποιημένο τρόπο.

Διεθνή πρότυπα αναπτύχθηκαν από τον Διεθνή Οργανισμό για την Τυποποίηση (International Organisation for Standardisation) με στόχο την τυποποίηση της διαδικασίας ανάπτυξης και παρουσίασης ενός θησαυρού γνώσης. Το πρότυπο για την κατασκευή και ανάπτυξη μονόγλωσσων θησαυρών είναι το ISO2788 ενώ για την ανάπτυξη πολύγλωσσων θησαυρών, χρησιμοποιείται το πρότυπο ISO5964 (23). Από το 2008, και τα δύο πρότυπα έχουν αναθεωρηθεί, διευρυνθεί και συνδυαστεί σε ένα νέο πρότυπο το ISO 25964 – «Θησαυροί και δια λειτουργικότητα με άλλα λεξιλόγια».

Θησαυροί για ανάκτηση πληροφοριών έχουν το δική τους μοναδικό λεξιλόγιό με το οποίο καθορίζουν τα διαφορετικά είδη των όρων και των σχέσεων (24):

- Όροι είναι οι βασικές σημασιολογικές μονάδες για τη μεταφορά εννοιών. Οι όροι είναι συνήθως μονολεκτική ουσιαστικά, ενώ τα ρήματα μετατρέπονται σε ουσιαστικά, όπως αντί για "διαβάζω" σε "ανάγνωση" και τέλος τα επίθετα και τα επιρρήματα δεν χρησιμοποιούνται συνήθως για μεταφορά εννοιών. Όταν ένα όρος είναι ασαφής , ένα "επεξηγηματικό σημείωμα"(scope note) μπορεί να προστεθεί για να εξασφαλιστεί η συνοχή, και να δώσει λεπτομέρειες για την ερμηνεία του όρου. Δεν είναι απαραίτητη η χρήση επεξηγηματικού σημειώματος σε κάθε όρο, αλλά η παρουσία του είναι σημαντική βοήθεια για την σωστή χρήση και κατανόηση του συγκεκριμένου θησαυρού.

- Ουσιαστικά οι σχέσεις που εισάγουμε μεταξύ των όρων είναι οι συνδέσεις που βοηθάνε στην συσχέτιση των όρων. Αυτές οι σχέσεις μπορούν να χωριστούν σε τρείς τύπους :
    1. Ιεραρχικών σχέσεων: Οι σχέσεις αυτές, δείχνουν όρους που είναι στενότερου ή ευρύτερου πεδίου. Ευρύτερος Όρος (BT- Broader Term) και Στενότερος Όρος (NT- Narrower Term).
    2. Σχέσεων Ισοδυναμίας: Χρησιμοποιείται κυρίως για την σύνδεση συνώνυμων και σχεδόν συνώνυμων όρων. «Χρησιμοποιείται για» (UF - Used For Term).
    3. Σχέσεων Συσχέτισης: Για την σύνδεση δύο παρεμφερών όρων, των οποίων η σχέση δεν είναι ούτε ιεραρχική ούτε ισοδύναμη, Συναφής Όρος (RT - Related Term).

Με βάση τα παραπάνω, βλέπουμε ότι ένας θησαυρός όρων περιλαμβάνει ένα σύνολο είτε εννοιών είτε αντικειμένων του πραγματικού κόσμου και τους δίνει μια εξήγηση σχετική με τον τομέα χρήσης, τον οποίο σχεδιάστηκε να εξυπηρετεί. Για παράδειγμα, δίνεται δυνατότητα στο σχεδιαστή του θησαυρού να διαχωρίσει την λέξη «γραφείο» - το μέρος που εργάζεται κάποιος και της λέξης «γραφείο» - το έπιπλο.

Οι θεματικές επικεφαλίδες (Subject headings) (25) είναι ένα σύνολο από ελεγχόμενο λεξιλόγιο που χρησιμοποιείται για την ταξινόμηση των υλικών και βοηθάει στην βελτίωση της ακρίβειας και την αποτελεσματικότητας μιας αναζήτησης. Αντί να παρέχουν ένα μοναδικό όρο ή φράση για χρήση, όπως γίνεται στην περίπτωση των θησαυρών, οι θεματικές επικεφαλίδες επιτρέπουν στο χρήστη να συνδέσει ή να συντονίσει όρους για να παράγει συμβολοσειρές όρων. Για παράδειγμα, οι θεματικές επικεφαλίδες Library of Congress Subject Headings ενώνουν τις έννοιες «Art» και «War» για να σχηματίσουν την επικεφαλίδα «Art and war». Στη συνέχεια, κάποιος μπορεί να συντονίσει την επικεφαλίδα αυτή με επικεφαλίδες από συγκεκριμένες περιοχές π.χ «Ινδιάνοι της Βόρειας Αμερικής Αλάσκα». Επειδή συνήθως ο αριθμός των θεματικές επικεφαλίδων είναι μεγάλος, ο ευκολότερος τρόπος για να βρεις την ζητούμενη πληροφορία είναι να ξεκινήσεις από μια λέξη κλειδί, μετά να κοιτάζεις θεματικές επικεφαλίδες σχετικών όρων, ώστε να εντοπίσεις άλλο σχετικό υλικό. Τέλος, η γλώσσα SKOS (Simple Knowledge Organisation System), που θα αναλυθεί λεπτομερώς στην συνέχεια, παρέχει ένα τρόπο για την έκφραση όρων εύρεσης και

θεματικών επικεφαλίδων με την γλώσσα RDF, για χρήση του περιεχομένου στον Σημασιολογικό Ιστό.

3.2 Θησαυροί, ελεγχόμενα λεξιλόγια και βάσεις γνώσης.

Στο σημείο αυτό κρίθηκε απαραίτητη η ανάλυση κάποιων βασικών πολιτιστικών θησαυρών, ελεγχόμενων λεξιλογίων και βάσεων δεδομένων που είναι θεμελιώδης για τα Συνδεδεμένα Δεδομένα,. Η παρουσίαση τους, θα βοηθήσει τον αναγνώστη στην κατανόηση της μορφής τους αλλά και θα δώσει μια εικόνα της σημερινής κατάστασης του πολιτισμικού χώρου που προσπαθεί να εισέλθει στο χώρο της τεχνολογίας και των Συνδεδεμένων Δεδομένων.

3.2.1 GEMET

GEMET (GEneral Multilingual Environmental Thesaurus) (26) είναι ένας πολύγλωσσος περιβαλλοντικός θησαυρός, που έχει αναπτυχθεί από το Ευρωπαϊκό Κέντρο για κατάλογους από πηγές δεδομένων. Δημοσιεύεται και διαχειρίζεται από την Ευρωπαϊκή Υπηρεσία Περιβάλλοντος και δικτύου πληροφοριών. Ο GEMET είναι μια συλλογή από διάφορα πολυγλωσσικά λεξιλόγια, και έχει σχεδιαστεί ως ένας γενικός θησαυρός , με στόχο τον καθορισμό μιας βασικής γενικής ορολογίας για το περιβάλλον. Η τρέχουσα έκδοση είναι διαθέσιμη σε 27 γλώσσες και περιλαμβάνει πάνω από 6.000 περιγραφές. Είναι διαθέσιμος δημόσια, είτε με άμεση πρόσβαση από το διαδίκτυο μέσω εφαρμογής είτε με κατέβασμα του, σε μορφή HTML ή SKOS αρχείου.

3.2.2 Eurovoc

Eurovoc (27) είναι ένας πολύγλωσσος θησαυρός που τηρείται από το Γραφείο Επισήμων Εκδόσεων της Ευρωπαϊκής Ένωσης. Υπάρχει σε 22 επίσημες γλώσσες της Ευρωπαϊκής Ένωσης (βουλγαρικά, ισπανικά, τσεχικά, δανικά, γερμανικά, εσθονικά, ελληνικά, αγγλικά, γαλλικά, ιταλικά, λετονικά, λιθουανικά, ουγγρικά, μαλτέζικα, ολλανδικά, πολωνικά, πορτογαλικά, ρουμανικά, σλοβακικά, σλοβενικά , της Φινλανδίας και της Σουηδίας), καθώς και σε κροατικά. Επίσης, έχει μεταφραστεί στα αλβανικά, ρωσικά και ουκρανικά. Ο Eurovoc χρησιμοποιείται από το Ευρωπαϊκό Κοινοβούλιο , στην Υπηρεσία Επισήμων Εκδόσεων των Ευρωπαϊκών Κοινοτήτων, τα εθνικά και περιφερειακά κοινοβούλια της Ευρώπης, ορισμένες εθνικές διοικητικές υπηρεσίες, καθώς και από ευρωπαϊκές οργανώσεις.

3.2.3 Getty Thesaurus of Geographic Names

Η Getty (28) είναι ένας Θησαυρός Γεωγραφικών Ονομάτων (συντομογραφία TGN ή GTGN). Το TGN περιλαμβάνει τα ονόματα και τις συναφείς πληροφορίες σχετικά με περιοχές. Οι περιοχές του TGN περιλαμβάνουν διοικητικές πολιτικές οντότητες (π.χ., οι πόλεις, τα έθνη) και φυσικά χαρακτηριστικά (π.χ., βουνά, ποτάμια). Περιλαμβάνονται τρέχουσα και ιστορικά μέρη. Επίσης παρέχονται και πληροφορίες σχετικές με την ιστορία, τον πληθυσμό, τον πολιτισμό, την τέχνη και την αρχιτεκτονική. Η πηγή του θησαυρού είναι διαθέσιμη για μουσεία, βιβλιοθήκες τέχνης, βιβλιογραφικά προγράμματα μέσω ιδιωτικών αδειών ή στη διάθεση κοινού, δωρεάν στην ιστοσελίδα του λεξιλογίου Getty. Είναι συμβατός με τα πρότυπα κατασκευής θησαυρού ISO και NISO. Περιέχει 1,106,000 όρους, ονόματα, και άλλες πληροφορίες για ανθρώπους, τόπους, πράγματα και έννοιες που σχετίζονται με την τέχνη, την αρχιτεκτονική και τον υλικό πολιτισμό. Περιέχει σχέσεις ιεραρχικές, ισοδυναμίας και συσχετιστικές. Τέλος να τονίσουμε ότι ο TGN δεν είναι GIS (Geographic Information System), παρόλο ότι πολλές εγγραφές του περιέχουν και τις συντεταγμένες της τοποθεσίας, οι οποίες είναι πάντα κατά προσέγγιση και με μόνη πρόθεση την αναφορά.

3.2.4 AGROVOC

AGROVOC (29)αναπτύχθηκε αρχικά στη δεκαετία του 1980 ως ένα πολυγλωσσικό δομημένο θησαυρό για όλες τις θεματικές στα πεδία της γεωργίας , της δασοκομίας , της αλιείας , των τροφίμων και συναφείς τομείς (π.χ. περιβάλλον). Σκοπός του ήταν να τυποποιηθεί η διαδικασία ευρετηρίασης για την AGRIS βάση δεδομένων, προκειμένου να κάνει την αναζήτηση απλούστερη, πιο αποτελεσματική, και να καθοδηγεί τον χρήστη για σχετικούς πόρους. Ο AGROVOC χρησιμοποιείται σε όλο τον κόσμο από ερευνητές, βιβλιοθηκάριους, διαχειριστές πληροφοριών, και άλλων, για την εύρεση, την ανάκτηση και τη διοργάνωση των δεδομένων στον τομέα των γεωργικών συστημάτων πληροφοριών. Ο ρόλος του είναι να βοηθήσει και να τυποποιήσει τη σημασιολογική περιγραφή των αντικειμένων, προκειμένου να επιτευχθεί η ενοποίηση της πληροφορίας σε όλα τα συστήματα, και να παράσχει πρόσβαση σε σχετικούς πόρους. AGROVOC διατίθεται σε έξι επίσημες γλώσσες του FAO : αγγλικά, γαλλικά, ισπανικά, αραβικά, κινέζικα και ρωσικά. Επίσης, έχει μεταφραστεί σε Τσεχική, περσικά, Γερμανικά, Χίντι, Ουγγρικά, Ιταλικά, Ιαπωνικά, Κορεατικά, Λάος, πολωνική, πορτογαλική, σλοβακική, και της Ταϊλάνδης, και

μεταφράζεται σε άλλες γλώσσες, όπως της Μαλαισίας, Μολδαβίας, Τελούγκου, Τουρκικά, και της Ουκρανίας.

Ο AGROVOC αποτελείται από όρους, που αποτελούνται από μία ή περισσότερες λέξεις που αντιπροσωπεύουν μία και ίδια έννοια. Για κάθε όρο, ένα μπλοκ λέξης εμφανίζεται, δείχνοντας την ιεραρχική και τη μη ιεραρχική σχέση του με άλλους όρους: BT (ευρύτερος όρος), NT (στενότερη έννοια), η RT (σχετικός όρος), UF (μη περιγραφικός). Οι σημειώσεις και ορισμοί, χρησιμοποιούνται στην AGROVOC για να διευκρινιστεί την έννοια και το περιεχόμενο του όρου. Ο AGROVOC έχει πλέον μετατραπεί από σύστημα οργάνωσης της γνώσης με βάση τον όρο και με τις παραδοσιακές σχέσεις θησαυρού (BT, NT, RT, και UF) σε σύστημα βασισμένο στην έννοια, το AGROVOC Concept Scheme (CS). Η AGROVOC CS επιτρέπει την εκπροσώπηση ποιο πολλών σημασιολογιών, όπως τις ειδικές σχέσεις μεταξύ εννοιών καθώς και τις σχέσεις μεταξύ των πολύγλωσσων τους, π.χ. «έχει συνώνυμο», «έχει μετάφραση». Τέλος ο AGROVOC είναι προσβάσιμος μέσω δικτυακών υπηρεσιών. Επίσης μπορεί να κατεβαστεί ελεύθερα για μη εμπορική χρήση και είναι διαθέσιμο σε διάφορες μορφές όπως MySQL, RDF, OWL, SKOS κ.α. .

3.2.5 Art & Architecture Thesaurus

Ο θησαυρός AAT (30)αποτελεί ένα δομημένο λεξιλόγιο από περίπου 34.000 εννοιών, με 131.000 όρους και πληροφορία για την περιγραφή, τεκμηρίωση και ανάκτηση αντικειμένων από πολλούς τομείς, όπως καλές τέχνες (ζωγραφική, γλυπτική, κτλ), αρχιτεκτονική, διακοσμητικές τέχνες (έπιπλα, κοστούμια, εξοπλισμός), αρχειακό (έγγραφα, επιστολές, κτλ) και πολιτισμικό υλικό (π.χ. πολιτισμικές παραδόσεις) και για ένα χρονικό διάστημα που εκτείνεται από την αρχαιότητα ως τη σημερινή εποχή σε όλο τον κόσμο. Ο χρήστης μπορεί να πλοηγηθεί στον AAT μέσω μίας on-line έκδοσης, που ανανεώνεται σε τακτά χρονικά διαστήματα και είναι διαθέσιμος σε εκτυπωμένη και ηλεκτρονική μορφή.

Ο AAT δομείται τόσο ιεραρχικά βάσει διαφορετικών κατηγοριών (facets) όσο και αλφαβητικά, ανακλώντας την κοινή πρακτική από ακαδημαϊκούς και καταλογογράφους. Ειδικότερα, τη βάση της δομής του AAT εξυπηρετεί ένα πλαίσιο από επτά κατηγορίες, οι οποίες αποτελούν τα υψηλότερα επίπεδα στην ιεραρχική δομή του AAT. Κάθε κατηγορία υποδιαιρείται σε ιεραρχίες, οι οποίες επί του παρόντος είναι 33 στον αριθμό. Το επίκεντρο κάθε εγγραφής στον AAT είναι μία έννοια και στη βάση δεδομένων κάθε έννοια (ή εγγραφή) αναγνωρίζεται από ένα

μοναδικό αριθμητικό αναγνωριστικό (ID). Σε κάθε έννοια έχουν συνδεθεί όροι, συναφείς έννοιες, μία πατρική έννοια (σε σχέση με την ιεραρχία), πηγές για τα δεδομένα και σημειώσεις. Οι όροι για κάθε έννοια μπορεί να περιλαμβάνουν τον ενικό και τον πληθυντικό αριθμό, τη φυσική διάταξη, γραμματικές παραλλαγές, διάφορους τύπους φωνής και συνώνυμα που έχουν ποικίλες ετυμολογικές ρίζες.

Ο AAT μπορεί να χρησιμοποιηθεί με τρείς τρόπους, είτε ως ένα πρότυπο τιμών δεδομένων στην τεκμηρίωση (καταλογογράφηση, ευρετηρίαση και περιγραφή) πολιτισμικής πληροφορίας, δεδομένου ότι οι όροι που διαθέτει μπορούν να αποτελέσουν τιμές των πεδίων που χρησιμοποιούνται για την τεκμηρίωση, είτε ως βοήθημα αναζήτησης στις βάσεις δεδομένων, δημιουργώντας ένα σημασιολογικό δίκτυο που απεικονίζει συνδέσμους και μονοπάτια μεταξύ όρων. Οι χρήστες μπορούν να ακολουθήσουν αυτά τα μονοπάτια που συντίθενται από συνώνυμους, ευρύτερους/στενότερους και συναφείς όρους για να εκλεπτύνουν, να επεκτείνουν και να βελτιώσουν τις αναζητήσεις τους.

3.2.6 FOAF

Το FOAF (Friend Of a Friend) (31) project είναι ένα περιγραφικό λεξιλόγιο που εκφράζεται με χρήση της RDF και της OWL(Web Ontology Language). Ξεκίνησε από τους Dan Brickley και Libby Mille και συνέχισε σαν μία ανοιχτή κοινοτική πρωτοβουλία που οδεύει προς το Σημασιολογικό Ιστό. Είναι μια οντολογία μηχανής για την περιγραφή των μεταδεδομένων σχετικά με πρόσωπα, τα ενδιαφέροντά τους, τις σχέσεις και τις δραστηριότητές τους. Όλοι μπορούν να χρησιμοποιήσουν την FOAF για να περιγράψουν τους ίδιους. Κάθε προφίλ χρήστη έχει ένα μοναδικό αναγνωριστικό όπως την e-mail διεύθυνση ή ένα URI της ιστοσελίδας του ατόμου, το οποίο και χρησιμοποιείται για να οριστούν οι σχέσεις μεταξύ των ανθρώπων. Έτσι όπως τα HTML έγγραφα μπορούν να διασυνδεθούν μεταξύ τους, τα FOAF έγγραφα-προφίλ, συνδέονται μέσω του μοναδικού αναγνωριστικού. Με την χρήση του FOAF δίνεται η δυνατότητα για τη δημιουργία του αντίστοιχου της προσωπικής ιστοσελίδας στον κόσμο του Σημασιολογικού Ιστού.

3.2.7 Dublin Core

Το Σύνολο μεταδεδομένων DC (Dublin Core) (32) είναι ένα πρότυπο λεξιλόγιο το οποίο περιλαμβάνει ένα σύνολο από μικρές και βασικές λέξεις μέσω των οπόιων οι περισσότεροι πόροι μπορούν κα περιγραφούν και να κατηγοριοποιηθούν. Μπορεί να περιγράψει φυσικές πηγές όπως βιβλία, υλικό ψηφιακού περιεχομένου όπως video,

ήχος, εικόνα, κείμενο αλλά και σύνθετα πολυμέσα όπως οι ιστοσελίδες. Συνήθως οι υλοποιήσεις του είναι βασισμένες σε XML και RDF. Το Dublic Core είναι ορισμένο από τον ISO, στο ISO Standard 15836 και το NISO Standard Z.39.85-2007. Η Dublin Core Metadata Initiative (DCMI) είναι μία οργάνωση που παρέχει ένα ανοιχτό φόρουμ για την ανάπτυξη διαλειτουργικών προτύπων μεταδεδομένων για το διαδίκτυο, τα οποία υποστηρίζουν ένα εύρος στόχων και επιχειρηματικών μοντέλων. Η DCMI περιλαμβάνει ομάδες εργασίες που λειτουργούν πάνω σε ομοφωνία, παγκόσμια συνέδρια και workshops. Το πρότυπο Dublic Core περιλαμβάνει δύο επίπεδα: το Απλό και το Qualified. Το απλό Dublin Core αποτελείται από δεκαπέντε στοιχεία ενώ το Qualified Dublic Core περιέχει τρία στοιχεία επιπλέον: τα στοιχεία Audience, Provenance και RightsHolder, καθώς και ένα σύνολο από εκλεπτύνσεις στοιχείων που λέγονται Qualifiers και που εκλεπτύνουν τη σημασιολογία των στοιχείων με χρήσιμους για την ανακάλυψη των πόρων τρόπους.

3.2.8 Wordnet

Το WordNet (33) δημιουργήθηκε και διατηρείται στο γνωστικό εργαστήριο επιστήμης του Πανεπιστήμιου του Princeton και η ανάπτυξη του άρχισε το 1985. Το WordNet είναι ένα σημασιολογικό λεξικό για Αγγλική γλώσσα. Ομαδοποιεί τις αγγλικές λέξεις στα σύνολα συνωνύμων αποκαλούμενων synsets, παρέχει τους σύντομους γενικούς ορισμούς, και καταγράφει τις σημασιολογικός σχέσεις μεταξύ αυτών των συνωνύμων συνόλων. Ο σκοπός είναι διπλός: να παραγάγει έναν συνδυασμό λεξικού και θησαυρού που είναι διαισθητικά χρησιμοποιήσιμος, και για να υποστηρίξει αυτόματες εφαρμογές με αντικείμενο την ανάλυση κειμένων και της τεχνητής νοημοσύνης.

Η τελευταία έκδοση είναι η 3.0 και το 2006 η βάση δεδομένων του περιείχε 155.287 λέξεις που οργανώνονται σε 117.659 synsets για ένα σύνολο 206.941 ζεύγη λέξεων. Το WordNet διακρίνει ουσιαστικά , ρήματα , επίθετα και επιρρήματα διότι ακολουθούν διαφορετικούς γραμματικούς κανόνες, ενώ δεν περιλαμβάνονται προθέσεις κλπ. Κάθε synset περιέχει μια ομάδα συνώνυμων λέξεων ή παραθέσεις (η παράθεση είναι μια ακολουθία λέξεων που πηγαίνουν μαζί για να διαμορφώσουν μια συγκεκριμένη έννοια) οι λέξεις συμμετέχουν χαρακτηριστικά σε διάφορα synsets.

Στο πλαίσιο της παρούσας διπλωματικής δεν μας απασχολεί να αναφερθούμε με μεγαλύτερη λεπτομέρεια στη δομή του WordNet, αλλά μόνο στο γεγονός ότι έχουν γίνει κάποιες προσπάθειες ώστε να μεταφερθεί και να χρησιμοποιηθεί η γνώση που

περιέχεται σε αυτό, στο χώρο του Σημασιολογικού Ιστού όπως μια προσπάθεια παρουσίασης του σε RDF/OWL.

3.2.9 GEONAMES

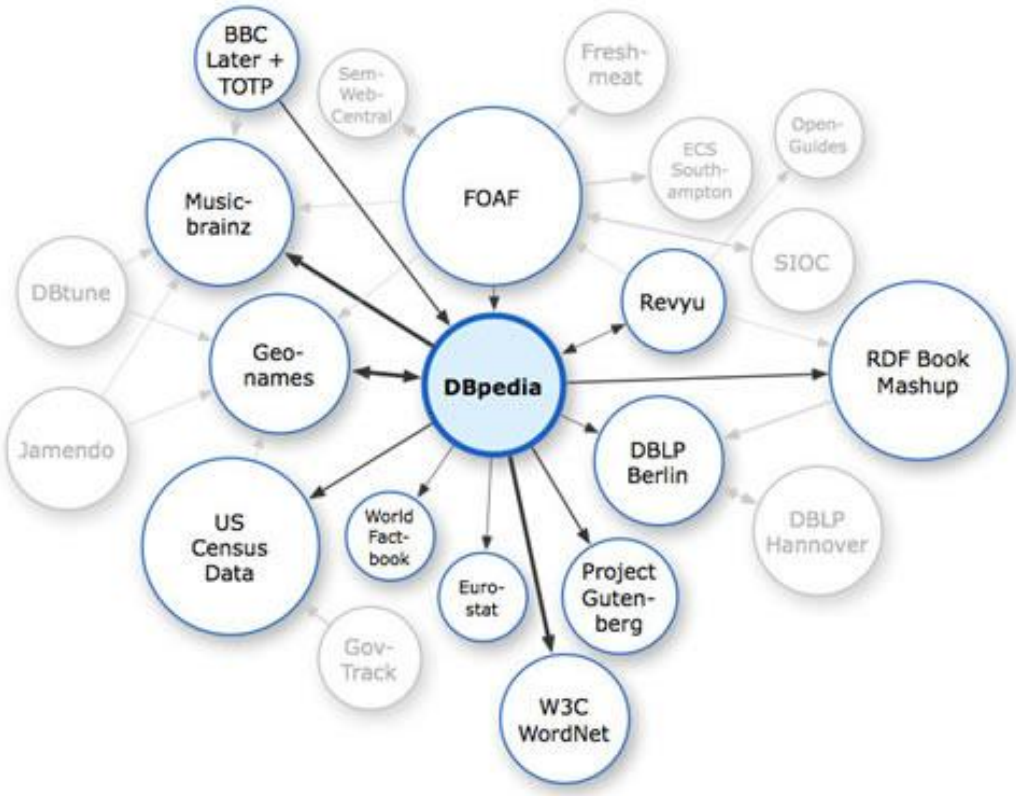
Geonames (34)είναι μια γεωγραφική βάση δεδομένων, διαθέσιμη και προσιτή μέσω διαφόρων υπηρεσιών Web και με την άδεια Creative Commons. Η βάση δεδομένων του Geonames περιέχει πάνω από 10.000.000 γεωγραφικές ονομασίες που αντιστοιχούν σε πάνω από 7.500.000 μοναδικά χαρακτηριστικά. Όλα τα χαρακτηριστικά κατηγοριοποιούνται σε μια από τις εννέα τάξεις και έπειτα υπό-κατηγοριοποιούνται με ένα από τους 645 κωδικούς. Πέρα από τα ονόματα των τόπων σε διάφορες γλώσσες, τα δεδομένα αποθηκεύονται μαζί με το γεωγραφικό πλάτος , μήκος , υψόμετρο, πληθυσμός, διοικητική υποδιαίρεση και τους ταχυδρομικούς κώδικες . Όλες οι συντεταγμένες χρησιμοποιούν το Παγκόσμιο Γεωδαιτικό Σύστημα 1984 (WGS84). Ένα σύνολο από δωρεάν εφαρμογές δίνουν πρόσβαση στη βάση του Geonames, η οποία διατίθεται και για κατέβασμα. Η εφαρμογή που είναι διαθέσιμη στο διαδίκτυο προσφέρει εύρεση περιοχών με αντίστροφη γεω κωδικοποίηση ,μέσω ταχυδρομικών κωδικών, με το να βρεθούν χώροι δίπλα σε ένα συγκεκριμένο τόπο, και την εξεύρεση άρθρων της Wikipedia σχετικά με τις γειτονικές περιοχές.

Κάθε αντικείμενου που περιέχει το Geonames αντιπροσωπεύεται ως πόρος του Ιστού και προσδιορίζεται από ένα σταθερό URI . Αυτό το URI παρέχει πρόσβαση είτε μέσω διαπραγμάτευσης περιεχομένου, είτε με τη σελίδα wiki που είναι σε HTML, ή με μια RDF περιγραφή του αντικειμένου, χρησιμοποιώντας στοιχεία της οντολογίας του Geonames. Η οντολογία (35) περιγράφει τις χαρακτηριστικές ιδιότητες των στοιχείων του Geonames χρησιμοποιώντας την OWL, ενώ οι κλάσεις και οι κωδικοί την SKOS γλώσσα. Τα δεδομένα του Geonames συνδέονται με τα δεδομένα της DBpedia και των άλλων RDF Linked Data, μέσω των άρθρων της Wikipedia, της οποίας έχουν τα URL, και συνδέονται με τις περιγραφές RDF. Η βάση δεδομένων της Geonames θα μας απασχολήσει εκτεταμένα παρακάτω, όπου προσπαθούμε να κάνουμε αυτόματη εξαγωγή από την βάση της και μετατροπή των εξαγόμενων στοιχείων στη γλώσσα της SKOS.

3.2.10 DBPEDIA

DBpedia (36) είναι μια προσπάθεια που αποσκοπεί στην εξαγωγή δομημένου περιεχόμενου από τις πληροφορίες που δημιουργήθηκαν στο πλαίσιο της εγκυκλοπαίδειας της Wikipedia. Αυτή η δομημένη πληροφορία στη συνέχεια

διατίθενται στον Παγκόσμιο Ιστό. Η DBpedia επιτρέπει στους χρήστες να αναζητούν σχέσεις και ιδιότητες που σχετίζονται με τους πόρους της Wikipedia, συμπεριλαμβανομένων των συνδέσεων με άλλες συναφείς βάσεις δεδομένων . Ο Tim Berners-Lee την περιέγραψε ως ένα από τα πιο διάσημα μέρη των Συνδεδεμένων Δεδομένων και δικαίως αφού όπως βλέπουμε και στην εικόνα είναι το κέντρο αλλά και η βάση με την οποία προσπαθούν να συνδεθούν όλα τα υπόλοιπα σύνολα δεδομένων.



Εικόνα 3.2.10.1 Η DBpedia στα Συνδεδεμένα Δεδομένα.

Το 2011 η DBpedia (37) περιλαμβάνει ένα σύνολο πάνω από 3.5 εκατομμύρια αντικείμενα εκ των οποίων 1.670.000 κατατάσσονται με σταθερή Οντολογία και συμπεριλαμβάνονται 364.000 άτομα, 462.000 θέσεων, 99.000 μουσικά άλμπουμ, ταινίες 54.000, 17.000 video games, 148.000 οργανώσεις , 169.000 είδη και 5.200 ασθένειες. Έχει 95 διαφορετικές γλώσσες, 850.000 συνδέσεις με εικόνες και 5.900.000 συνδέσεις με εξωτερικές ιστοσελίδες. Ακόμα, έχει 6.500.000 εξωτερικές συνδέσεις σε άλλες βάσεις δεδομένων RDF.

Το έργο DBpedia χρησιμοποιεί το Resource Description Framework (RDF) για την παρουσίαση των πληροφοριών που εξάγονται και αποτελείται πάνω από 672 εκατομμύρια κομμάτια πληροφορίας (RDF τριάδες) από τα οποία 286 εκατομμύρια

προήλθαν από την αγγλική έκδοση της Βικιπαίδειας και 386 εκατ. ευρώ προέρχονται από άλλες γλωσσικές εκδόσεις. Τέλος, βασικό πλεονέκτημα της DBpedia είναι ότι βασίζεται σε μια πολύ μεγάλη διαδικτυακή ανοικτή κοινότητα χρηστών για την ανάπτυξη και την εξέλιξη της.

3.2.11 IPTC

Ο διεθνής οργανισμός International Press Telecommunications Council (IPTC) (38) είναι μια κοινοπραξία που έχει στόχο την ανάπτυξη και διατήρηση τεχνικών πρότυπων για τη βελτίωση της ανταλλαγής ειδήσεων που χρησιμοποιούνται από σχεδόν κάθε σημαντικό οργανισμό ειδήσεων στον κόσμο. Οι περισσότερες από τις τρέχουσες εργασίες του IPTC αφορούν XML πρότυπα για την ανταλλαγή ειδήσεων και την ανάπτυξη προηγμένων μεταδεδομένων, με στόχο την περιγραφή και την ταξινόμηση των ειδήσεων, των φωτογραφιών, των βίντεο και άλλων μέσων ενημέρωσης.

Διάφορα πρότυπα έχουν αναπτυχθεί όπως το IPTC Information Interchange Model (IIM) (39) που είναι για μεταδεδομένα εικόνων ή το SportsML που είναι ένας βολικός τρόπος για να μοιράζονται τα αθλητικά στατιστικά με συνοπτικό και ευκρινή τρόπο. Βασική είναι επίσης η οικογένεια προτύπων IPTC G2 (NewsML-G2, EventsML-G2, SportsML-G2) που βασίζεται στην XML και σχεδιάστηκε για τον Σημασιολογικό Ιστό. Αυτή δίνει πολλές ευκαιρίες για την ενσωμάτωση μεταδεδομένων και επεκτείνει το πεδίο εφαρμογής της, πέρα από το περιεχόμενο ειδήσεων, ώστε να περιλαμβάνονται δεδομένα για ένα γεγονός αλλά και καλά οργανωμένες πληροφορίες σχετικά με τις οργανώσεις ατόμων, τα σημεία ενδιαφέροντος, τις γεωπολιτικές περιοχές ή κάποιες αφηρημένες έννοιες. Μεγάλο μέρος της διπλωματικής, βασίστηκε στη χρήση των IPTC NewsCodes. Το IPTC δεν παρέχει μόνο τις μορφές ανταλλαγής ειδήσεων για τη βιομηχανία ειδήσεων, αλλά παράλληλα δημιουργεί και συντηρεί ένα σύνολο από έννοιες που πρέπει να αποδοθούν ως μεταδεδομένα για αντικείμενα ειδήσεων όπως, φωτογραφίες κειμένου, ήχου, γραφικών, και αρχείων βίντεο. Αυτό επιτρέπει με την πάροδο του χρόνου μια ενιαία κωδικοποίηση των μεταδεδομένων για ειδήσεις .

Σήμερα οι NewsCodes χωρίζεται σε πολλά διαφορετικά σύνολα - ταξινομήσεις - με στόχο την καλύτερη διαχείριση μιας και τα θέματα αφορούν μια συγκεκριμένη περιοχή. Οι ταξινομίες των IPTC NewsCodes ομαδοποιούνται σε τέσσερεις βασικούς τομείς (40):

- Περιγραφικούς NewsCodes (Descriptive NewsCodes) Πρόκειται για μια ομάδα ταξινομιών για να περιγραφεί σωστά το περιεχόμενο των ειδήσεων.
- Διαχειριστικούς NewsCodes (Administrative NewsCodes) Πρόκειται για μια ομάδα ταξινομιών για την ορθή διαχείριση των ειδήσεων.
- NewsCodes για μετάδοση (Transmission NewsCodes) Αυτή είναι μια ομάδα ταξινομιών με ελεγχόμενες τιμές για τη διαδικασία της μετάδοσης.
- NewsCodes σε ανταλλάξιμες μορφές (Exchange Format NewsCodes) Αυτή είναι μια ομάδα ταξινομιών με τιμές που υποστηρίζουν συγκεκριμένες λειτουργίες των IPTC μορφών των προτύπων που αφορούν την ανταλλαγή ειδήσεων. Τα NewsML 1.x, NewsML-G2, EventsML-G2 και SportsML-G2 κάνουν χρήση αυτής της λειτουργίας.
- Το IPTC παρέχει το σύνολο των δεδομένων του ελεύθερα σε διάφορες μορφές όπως RDF/XML με SKOS ή NewsML-G2 αρχείο.

3.3 Συστήματα ευρωπαϊκής πολιτιστικής κληρονομιάς

Στην ενότητα αυτή παρουσιάζουμε μερικά συστήματα που χρησιμοποιούν τεχνολογίες Σημασιολογικού Ιστού για την διαχείριση των Πολιτιστικών Δεδομένων. Βλέπουμε αρχικά την Europeana, που είναι ένα έργο με στόχο τη διατήρηση της πολιτιστικής κληρονομιάς της Ευρώπης στον ψηφιακό χώρο και στη συνέχεια , το EUscreen με στόχο την διατήρηση της τηλεοπτικής κληρονομιάς .

3.3.1 Europeana

Η Europeana (41) είναι ένα έργο με στόχο να ολοκληρώσει τη πολιτιστική κληρονομιά ολόκληρης της Ευρώπης. Πρόκειται για μία διαδικτυακή πύλη που παρέχει πρόσβαση σε εκατομμύρια, πίνακες ζωγραφικής βιβλία, ταινίες, μουσειακά αντικείμενα και αρχειακά έγγραφα, που έχουν ψηφιοποιηθεί σε όλη την Ευρώπη. Στην Europeana συμμετείχαν περίπου 1500 ιδρύματα από όλη την Ευρώπη. Ο χρήστης μπορεί να πλοηγηθεί και να εξερεύνησει την πολιτιστική και επιστημονική κληρονομιά της Ευρώπης μέσα από μια κοινή συλλογή που αναπτύχθηκε.

Η Europeana παρέχει πρόσβαση σε διάφορους τύπους περιεχομένου από διαφορετικά είδη οργανισμών κληρονομιάς. Τα ψηφιακά αντικείμενα που οι χρήστες βρίσκουν στην Europeana δεν αποθηκεύονται σε έναν κεντρικό υπολογιστή, αλλά παραμένουν στο πολιτιστικό ίδρυμα και φιλοξενούνται στο δίκτυο τους. Η Europeana συγκεντρώνει συναφείς πληροφορίες - ή τα μεταδεδομένα - σχετικά με τα

αντικείμενα, μαζί με μια μικρή εικόνα. Όταν ο χρήστης βρει αυτό που τον ενδιαφέρει, μπορεί να έχει πρόσβαση στο πλήρες περιεχόμενο του , επισκεπτόμενος την αρχική τοποθεσία που κρατά το περιεχόμενο. Διαφορετικοί τύποι πολιτιστικών οργανισμών κληρονομιάς - βιβλιοθήκες, μουσεία, αρχεία και οπτικοακουστικών συλλογών – κατηγοριοποιούν το περιεχομένου τους με διαφορετικούς τρόπους και σε διαφορετικά πρότυπα. Οι προσεγγίσεις επίσης ποικίλλουν στις διάφορες χώρες. Για να καταστούν τα στοιχεία εύκολα αναζητήσιμα, πρέπει να αντιστοιχίζονται σε ένα μοναδικό και κοινό πρότυπο, γνωστό ως Europeana Semantic Elements. Η απόφαση σχετικά με το ποια αντικείμενα έχουν ψηφιοποιηθεί είναι του οργανισμού που κατέχει το υλικό και όχι της Europeana.

Στο σχέδιο στρατηγικής για τη Europeana που δημοσιεύθηκε τον Ιανουάριο του 2011 (42), η Europeana περιγράφει τέσσερις βασικές στρατηγικές στις οποίες θα βασιστεί η εξέλιξη της:

1. Συγκέντρωση - για την κατασκευή μιας ανοικτής και αξιόπιστης πηγής για την ευρωπαϊκή πολιτιστική και επιστημονική κληρονομιά περιεχομένου.
2. Διευκόλυνση - να στηρίξει τον τομέα πολιτιστικής και επιστημονικής κληρονομιάς, μέσω της μεταφοράς γνώσεων, της καινοτομίας και της υποστήριξης.
3. Διανομή – η κληρονομία να είναι διαθέσιμοι στους χρήστες όπου και αν βρίσκονται και όποτε την θέλουν.
4. Συνεργασία - να αναπτυχθούν νέοι τρόποι για τους χρήστες ώστε να συμμετάσχουν στην πολιτιστική και επιστημονική κληρονομιά τους.

Μέσω της Europeana υπάρχουν πολλά έργα τα οποία συμβάλλουν τεχνολογικά δίνοντας λύσεις αλλά και περιεχόμενο στη Europeana. Μερικά από τα σημαντικότερα είναι η ATHENA η οποία ελέγχει το περιεχόμενο των μουσείων και προωθεί τα πρότυπα για την ψηφιοποίηση μουσείων και των μεταδεδομένων, η Europeana Connect που προσθέτει ηχητικό υλικό στην Europeana και η Ευρωπαϊκή Βιβλιοθήκη που προσθέτει το περιεχόμενο των εθνικών βιβλιοθηκών.

3.3.2 EUSCREEN

Το έργο του EUscreen (43) ξεκίνησε τον Οκτώβριο του 2009 και συγκέντρωσε πάνω από 35.000 αντικείμενα από την τηλεοπτική κληρονομιά της Ευρώπης (βίντεο , φωτογραφίες , αντικείμενα). Το EUscreen άρχισε να παρέχει από το 2011 τυποποιημένη πρόσβαση στα αντικείμενα αυτά αλλά και συμπληρωματικές συναφείς

πληροφορίες μέσω μιας πολυγλωσικής πύλης ελεύθερης πρόσβαση. Η κοινοπραξία του EUscreen συντονίζεται από το Πανεπιστήμιο της Ουτρέχτης και αποτελείται από 27 συνεργάτες (οπτικοακουστικά αρχεία , τα ερευνητικά ιδρύματα , φορείς παροχής τεχνολογίας και της Europeana) από 19 ευρωπαϊκές χώρες.

Είναι μία από τις κύριες οπτικοακουστικές πηγές περιεχομένου για την Europeana, έτσι η συλλογή του είναι επίσης συνδεδεμένη απευθείας με την συλλογή που έχει εκατομμύρια ψηφιοποιημένα αντικείμενα από τα ευρωπαϊκά μουσεία, βιβλιοθήκες και αρχεία. Η επιλογή περιεχόμενου του EUscreen έχει χωριστεί σε τρία μέρη:

1. Ιστορικά Θέματα: Η συντριπτική πλειοψηφία του περιεχομένου EUscreen θα απαρτίζεται από κλιπ και προγράμματα που έχουν επιλεγεί για την εξερεύνηση 14 Ιστορικών Θεμάτων.
2. Παροχής περιεχομένου από Εικονικές Εκθέσεις: Στο πλαίσιο του έργου EUscreen, οι ραδιοτηλεοπτικοί φορείς θα αναπτύξουν την δική τους εικονική έκθεση , η οποία θα συνοδεύεται από μια σειρά μέσων, όπως φωτογραφίες, έγγραφα, ήχο και γραπτό κείμενο.
3. Συγκριτικές Εικονικές Εκθέσεις: Αυτά θα περιλαμβάνουν βίντεο κλιπ και προγράμματα από το σύνολο των EUscreen αρχείων των ραδιοτηλεοπτικών οργανισμών στόχο την εξερεύνηση τριών θεμάτων για να δείξει πώς τα γεγονότα έχουν παρατηρηθεί από διαφορετικές ευρωπαϊκές προοπτικές.

3.4 SKOS

Τα συστήματα οργάνωσης γνώσης (KOS – knowledge Organization Systems) χρησιμεύουν για σκοπούς ανάκτησης και διαχείρισης μίας συλλογής δεδομένων. Σε αυτά συμπεριλαμβάνονται διάφορα είδη λεξιλογίων όπως θησαυροί, σχήματα κατηγοριοποίησης, θεματικές λίστες και ταξινομήσεις. Διάφοροι οργανισμοί έχουν αναπτύξει και χρησιμοποιούν τα δικά τους συστήματα KOS, τα οποία όμως δεν συνδέονται μεταξύ τους και έτσι δεν υπάρχει ένα κοινό σημείο αναφοράς και η χρήση τους περιορίζεται τοπικά.

Το Simple Knowledge Organization System (SKOS) (44) είναι μια οικογένεια των τυπικών γλωσσών που έχει σχεδιαστεί για την εκπροσώπηση των θησαυρών , των συστημάτων ταξινόμησης, τις ταξινομίες ή οποιαδήποτε άλλη μορφή δομημένων ελεγχόμενου λεξιλογίου. Το SKOS είναι χτισμένο πάνω σε RDF και RDFS και ο κύριος του στόχος είναι να επιτρέπει εύκολη δημοσίευση των ελεγχόμενων δομημένων λεξιλογίων στον Σημασιολογικό Ιστό. Το SKOS σήμερα αναπτύσσεται

στο πλαίσιο του W3C. Οι αναπαραστάσεις των αντικειμένων σε SKOS είναι μηχανικά επεξεργάσιμες και μπορούν να ανταλλαχθούν μεταξύ εφαρμογών καθώς και να εκδοθούν στο διαδίκτυο. Όλες οι εργασίες ανάπτυξης του πραγματοποιούνται μέσω της λίστας που είναι εντελώς ανοιχτή και δημόσια. Στην ενότητα αυτή θα παρουσιαστεί το μοντέλο SKOS, τα στοιχεία από τα οποία αποτελείται, καθώς και το τρόπο χρήσης του μαζί με άλλες τεχνολογίες Σημασιολογικού Ιστού.

3.4.1 Βασικά στοιχεία του SKOS

Στη βασική χρήση της SKOS, το εννοιολογικό πόρων (έννοιες) μπορεί να ταυτοποιηθεί με URIs, τα οποία ονομάζονται μέσω ετικετών σε μία ή περισσότερες φυσικές γλώσσες, τεκμηριώνονται με διάφορες σημειώσεις και σημασιολογικά συνδέονται μεταξύ τους σε άτυπες ιεραρχίες και σε συνδεδεμένα δίκτυα και συγκεντρώνονται σε εννοιολογικά συστήματα.

Η έννοια (concept) είναι το κεντρικό στοιχείο όλων των SKOS λεξιλογίων. Μία έννοια μπορεί να αναπαριστά μία ιδέα, ένα αντικείμενο, ένα γεγονός ή οτιδήποτε άλλο στο πλαίσιο ενός γνωστικού πεδίου. Η αναγνώριση των εννοιών είναι τυπικά έργο κάποιου ειδικού πάνω στο γνωστικό πεδίο. Το SKOS εισάγει την κλάση skos:Concept για τον ορισμό μιας έννοιας. Αυτό γίνεται σε δύο στάδια:

- με τη δημιουργία (ή την επαναχρησιμοποίηση) ενός Uniform Resource Identifier (URI) για να προσδιορίσει μοναδικά η έννοια αυτή.
- προβάλλοντας στο RDF, με την χρήση της ιδιότητας rdf:type δείχνουμε ότι ο πόρος που προσδιορίζονται από αυτό το URI είναι τύπου skos:Concept .

Για παράδειγμα με:

```
<http://www.example.com/animals> rdf:type skos:Concept.
```

Χρησιμοποιώντας SKOS για την δημοσίευση εννοιολογικών συστημάτων, γίνεται εύκολη η αναφορά των εννοιών, σε περιγραφές πόρων στο Σημασιολογικό Ιστό. Οι ετικέτες μιας έννοιας αποτελούν το σύνολο των εκφράσεων τις οποίες χρησιμοποιούμε για να αναφερθούμε σε αυτήν την έννοια με φυσική γλώσσα. Το SKOS παρέχει τρεις ιδιότητες για την επισύναψη ετικετών σε έννοιες: skos:prefLabel, skos:altLabel και skos:hiddenLabel. Ο τύπος της ετικέτας δηλώνει και τη σχέση της ετικέτας με την έννοια και βοηθάει στην αναζήτησης της. Τα γνωρίσματα αυτά ορίζονται ως αμοιβαίως αποκλειόμενα οπότε είναι σφάλμα αν μία έννοια έχει το ίδιο αλφαριθμητικό για δύο τύπους ετικετών.

Οι προτιμώμενες λεκτικές ετικέτες(skos:prefLabel) αναθέτουν μία προτιμώμενη ετικέτα σε μία έννοια. Οι ετικέτες αυτές χρησιμοποιούνται ως περιγραφείς των εννοιών στα συστήματα εύρεσης. Τα αλφαριθμητικά των ετικετών μπορούν προαιρετικά να συνοδεύονται από ένα tag γλώσσας .π.χ.

```
ex:animals rdf:type skos:Concept;
  skos:prefLabel "animals"@en;
  skos:prefLabel "animaux"@fr.
```

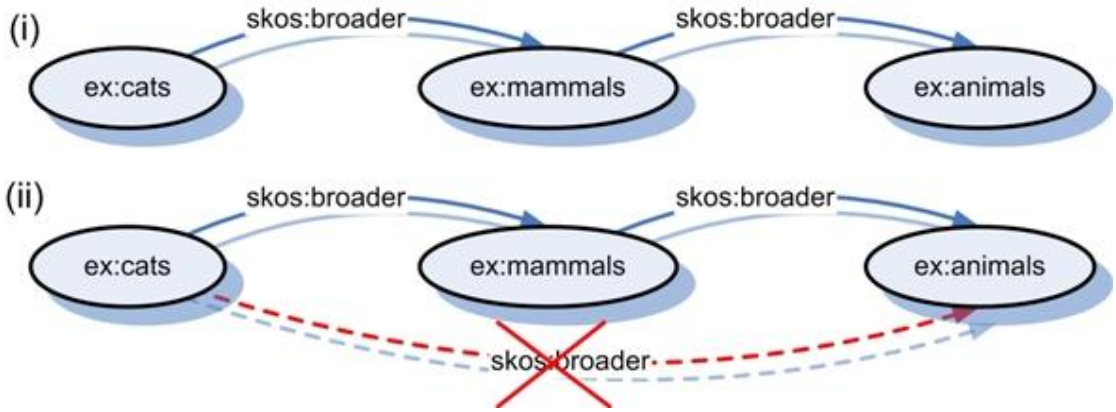
Οι σωστές πρακτικές υπαγορεύουν την χρήση μίας μόνο ετικέτας για κάθε γλώσσα ενώ μία ετικέτα δεν θα πρέπει να χρησιμοποιείται σε περισσότερες έννοιες από μία, εφόσον αυτές χρησιμοποιούνται για την αναπαράσταση της έννοιας. Οι εναλλακτικές λεκτικές ετικέτες (skos:altLabel) διευκολύνει όταν θέλουμε να εισάγουμε ετικέτες επιπλέον της επιθυμητής στο αντικείμενο που περιγράφεται. Επίσης μπορούμε αν και δεν προτείνεται, να εισάγουμε πάνω από μια εναλλακτική ετικέτα για την ίδια γλωσσά. Π.χ.

```
ex:rocks rdf:type skos:Concept;
  skos:prefLabel "rocks"@en;
  skos:altLabel "basalt"@en;
  skos:altLabel "granite"@en;
```

Τέλος οι κρυφές ετικέτες χρησιμοποιούνται όταν θέλουμε να είναι προσιτή μια ετικέτα από την εφαρμογή αναζήτησης αλλά να μην εμφανίζεται στον χρήστη.

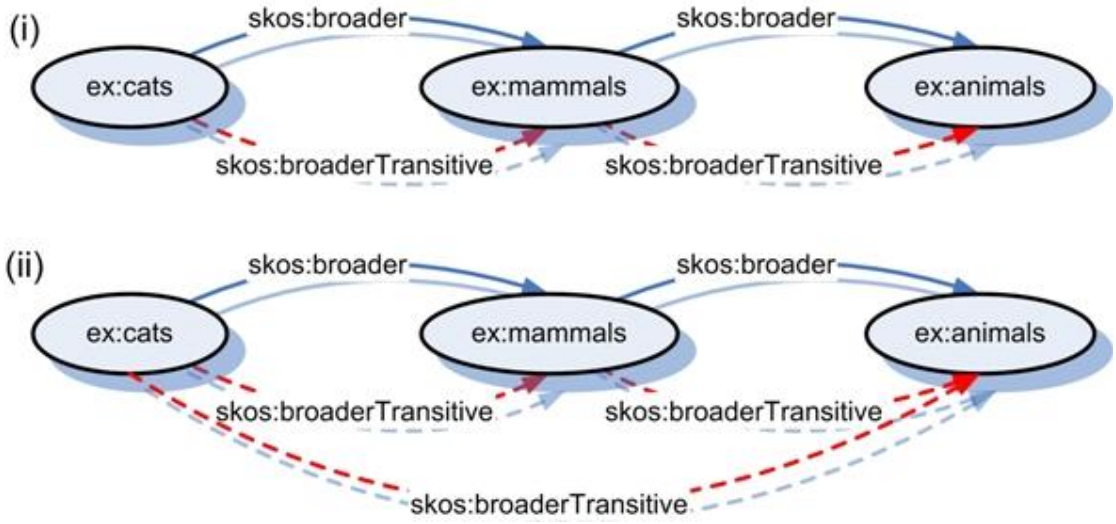
Οι σημασιολογικές σχέσεις είναι πολύ σημαντικές γιατί καθορίζουν το νόημα των εννοιών σε σχέση με άλλες έννοιες. Το SKOS παρέχει τρία γνωρίσματα: το skos:broader, το skos:narrower και το skos:related. Τα πρώτα δύο γνωρίσματα δίνουν τη δυνατότητα δημιουργίας ιεραρχικών σχημάτων εννοιών, ενώ η τρίτη επιτρέπει τις συνειρμικές και μη ιεραρχικές συνδέσεις.

Για να δείξουμε ότι μια έννοια έχει ευρύτερο περιεχόμενο (δηλαδή πιο γενικό) από ένα άλλο, χρησιμοποιείται η ιδιότητα skos:broader, ενώ η skos:narrower για να δηλώσουμε το αντίστροφο, δηλαδή όταν μια έννοια είναι πιο περιορισμένη από την άλλη. Να επισημάνουμε, όσον αφορά την κατεύθυνση της σχέσης, ότι η λέξη "broader" διαβάζεται "has broader concept". Ακόμα είναι σημαντικό ότι οι σχέσεις skos:broader/skos:narrower είναι αντίστροφη η κάθε μία στην άλλη και έτσι μπορούμε να συντομεύουμε το μέγεθος της πληροφορίας σε μία απεικόνιση SKOS. Στην SKOS οι δύο αυτές ιδιότητες δεν είναι μεταβατικές όπως φαίνεται στην εικόνα:



Εικόνα 3.4.1.1 skos:broader

Αντίθετα αν θέλουμε να δείξουμε ότι η έννοια μεταβιβάζεται, θα πρέπει η ιεραρχία να γίνεται με τις εντολές skos:broaderTransitive και skos:narrowerTransitive όπως φαίνεται:



Εικόνα 3.4.1.2 skos:broaderTransitive

Τέλος η ιδιότητα skos:related χρησιμοποιείται για να δείξει την σχέση που έχουν δύο έννοιες, είναι συμμετρική και δεν είναι μεταβατική.

Εκτός από τις σχέσεις που αφορούν την δομή και την ιεραρχία η SKOS παρέχει ένα σύνολο από ιδιότητες ώστε να ορίζονται χαρακτηριστικά των εννοιών που είναι για ανθρώπινη κατανάλωση. Το skos:scopeNote και το skos:definition είναι τα σημαντικότερα, αφού παρέχουν πρόσθετη πληροφορία σχετική με το νόημα που έχει η περιγραφόμενη έννοια.

Οι έννοιες μπορούν να δημιουργηθούν και να χρησιμοποιηθούν σαν ξεχωριστές οντότητες. Ωστόσο, συνήθως συλλέγονται σε λεξιλόγια ή σχήματα ταξινόμησης. Το SKOS παρέχει αυτή τη δυνατότητα με τη κλάση skos:ConceptScheme. Για παράδειγμα, για να ορίσουμε ένα σχήμα εννοιών σαν πόρο και να περιγράψουμε το τίτλο του και το δημιουργό του βάσει του Dublin Core μπορούμε να γράψουμε:

```
ex:animalThesaurus rdf:type skos:ConceptScheme;
  dct:title "Simple animal thesaurus";
  dct:creator ex:antoineIsaac.
```

Για την αποδοτική προσπέλαση των κορυφαίων εννοιών σε ένα σχήμα εννοιών, το SKOS ορίζει την ιδιότητα skos:hasTopConcept, με την οποία μπορούν να οριστούν οι πιο γενικές έννοιες που περιέχει ένα σχήμα εννοιών. Για το παράδειγμα από πάνω:

```
ex:animalThesaurus rdf:type skos:ConceptScheme;
skos:hasTopConcept ex:mammals;
skos:hasTopConcept ex:fish.
```

Τα σχήματα εννοιών έχουν σχεδιαστεί για την αναπαράσταση παραδοσιακών λεξιλογίων που βασίζονται στα πρότυπα οπότε οι σχεδιαστές θησαυρών θα πρέπει να ακολουθούν τις οδηγίες των προτύπων αυτών. Ωστόσο, πρέπει, να γίνει αντιληπτή η διάσταση του Σημασιολογικού Ιστού στο SKOS, σε αντίθεση με τα παραδοσιακά λεξιλόγια π.χ. μία έννοια μπορεί να ανήκει σε πολλά σχήματα εννοιών χρησιμοποιώντας την ιδιότητα: skos:inScheme.

3.4.2 Αντιστοίχιση σχημάτων εννοιών

Η αναπαράσταση ενός λεξιλογίου με το SKOS δεν εξυπηρετεί μόνο ως μηχανισμός δημοσίευσης, αλλά επιτρέπει να συμμετέχει σε ένα δίκτυο από σχήματα εννοιών. Στο Σημασιολογικό Ιστό οι πραγματικές δυνατότητες των δεδομένων εκτοξεύονται όταν διασυνδέονται. Καθώς οι έννοιες από διάφορα σχήματα εννοιών διασυνδέονται, αρχίζουν να διαμορφώνουν ένα κατανεμημένο, ετερογενές καθολικό σχήμα εννοιών. Ένας ιστός από σχήματα εννοιών μπορεί να εξυπηρετήσει ως η βάση για νέες εφαρμογές που θα επιτρέπουν τη σημασιολογική πλοήγηση μέσα στα λεξιλόγια.

Κάθε έννοια SKOS λαμβάνει ένα URI το οποίο αποτελεί αναμφίβολο αναγνωριστικό για την έννοια αυτή σε κάθε SKOS εφαρμογή. Αυτό γίνεται ιδιαίτερα χρήσιμο για την δημιουργία σημασιολογικών σχέσεων μεταξύ προϋπαρχόντων εννοιών. Αυτές οι αντιστοιχίσεις είναι κρίσιμης σημασίας για εφαρμογές όπως εργαλεία ανάκτησης

πληροφορίας που χρησιμοποιούν πολλαπλά λεξιλόγια με αλληλεπικαλυπτόμενα γνωστικά πεδία και που πρέπει να διασυνδεθούν σημασιολογικά.

Ένα σημαντικό χαρακτηριστικό της αντιστοίχισης είναι η δυνατότητα να δηλωθεί ότι δύο έννοιες σε διαφορετικά σχήματα έχουν συγκρίσιμα νοήματα και το να προσδιοριστεί η σχέση μεταξύ τους, ακόμα και αν έρχονται από διαφορετικά πλαίσια που χρησιμοποιούν διαφορετικές αρχές μοντελοποίησης. Το SKOS παρέχει μερικές ιδιότητες οι οποίες αντιστοιχούν έννοιες μεταξύ διαφορετικών σχημάτων εννοιών. Όταν δύο έννοιες έχουν παρόμοιο νόημα, αυτό μπορεί να δηλωθεί με τις ιδιότητες skos:exactMatch και skos:closeMatch. Επίσης δύο έννοιες από διαφορετικά σχήματα εννοιών μπορούν να αντιστοιχηθούν με ιδιότητες παράλληλες αυτών των σημασιολογικών σχέσεων που είδαμε πρωτύτερα: skos:broadMatch, skos:narrowMatch και skos:relatedMatch.

Η ιδιότητα skos:closeMatch δηλώνει ότι οι δύο έννοιες είναι επαρκώς παρόμοιες που μπορούν να χρησιμοποιηθούν η μία αντί για την άλλη στις εφαρμογές. Όμως η skos:closeMatch δεν είναι μεταβατική, κάτι που εμποδίζει να διαδοθεί η σχέση αυτή πέρα από τις δύο αυτές έννοιες. Η ιδιότητα skos:exactMatch δηλώνει μία ακόμα μεγαλύτερη σημασιολογική ομοιότητα. Πρέπει να σημειωθεί ωστόσο, ότι η τελευταία ιδιότητα δεν ορίζεται μέσω της ιδιότητας owl:sameAs στην γλώσσα OWL, η οποία συνδέει έννοιες που είναι ταυτόσημες και οι τριάδες που έχουν αυτούς τους πόρους γίνονται ένα.

3.4.3 Επέκταση του SKOS για ετικέτες (SKOS-XL)

Μια νεότερη επέκταση του SKOS μας απασχόλησε στα αρχικά στάδια της εργασίας και για αυτό η αναφορά σ' αυτήν κρίθηκε απαραίτητη. Κάποιες εφαρμογές απαιτούν τη δημιουργία άμεσων συνδέσμων μεταξύ των ετικετών που συνδέονται με τις έννοιες. Η χρήση του SKOS για τις ετικέτες, π.χ. skos:prefLabel , μας περιορίζει σε μία RDF τριάδα και έτσι, οι ετικέτες, δεν μπορούν να αποτελέσουν αντικείμενο μιας δήλωσης RDF, και άρα δεν μπορεί να επιτευχθεί σύνδεση μεταξύ τους.

Για να λυθεί αυτό το ζήτημα, δημιουργήθηκε μια επέκταση στο λεξιλόγιο SKOS για ετικέτες, η SKOS-XL (45). Η επέκταση αυτή εισάγει μια τάξη skosxl:Label που επιτρέπει στις ετικέτες να θεωρούνται πόροι RDF. Ότι θέλουμε να δηλώσουμε για την κλάση αυτή πρέπει πρώτα να επισυνάπτεται στο ίδιο RDF μέσω του skosxl:literalForm. Για παράδειγμα έστω ο «Οργανισμός Τροφίμων και Γεωργίας", ο

οποίος χαρακτηρίζεται τόσο από το επίσημο όνομα του, όσο και από τα αρχικά του ιδρύματος. Οι δύο ετικέτες μπορούν να γραφούν με τον ακόλουθο τρόπο:

```
ex:FAOlabel1 rdf:type skosxl:Label;
  skosxl:literalForm "Food and Agriculture Organization"@en.
ex:FAOlabel2 rdf:type skosxl:Label;
  skosxl:literalForm "FAO"@en.
```

Με την χρήση του skosxl:Label μπορούν οι ετικέτες να σχετιστούν με έννοιες χρησιμοποιώντας τις ιδιότητες(skosxl:prefLabel, skosxl:altLabel, skosxl:hiddenLabel) και να συνδεθούν μεταξύ τους με skosxl:labelRelation δηλώσεις:

```
ex:FAO rdf:type skos:Concept;
  skosxl:prefLabel ex:FAOlabel1;
  skosxl:altLabel ex:FAOlabel2.
ex:FAOlabel2 skosxl:labelRelation ex:FAOlabel1.
```

Τέλος να τονίσουμε ότι το SKOS-XL μοντέλο δεδομένων είναι συμβατό με την πρακτική του SKOS για τις ετικέτες. Εάν μια εντολή του skosxl:Label επισυνάπτεται σε μια έννοια, π.χ. η skosxl:altLabel , τότε σύμφωνα με το SKOS-XL, η skosxl:Label θα σχετίζεται με αυτή την έννοια σαν μία απλή ετικέτα skos:altLabel.

3.4.4 Πίνακες λεξιλογίου SKOS και SKOS-XL

Πίνακας 3-1 Λεξιλόγιο SKOS

skos:Collection	Συλλογή
skos:Concept	Έννοια
skos:ConceptScheme	Σχήμα εννοιών
skos:OrderedCollection	Διατεταγμένη συλλογή
skos:altLabel	Εναλλακτική ετικέτα
skos:broadMatch	Σύνδεση ευρύτερης σημασίας έννοιας
skos:broader	Ευρύτερος όρος
skos:broaderTransitive	Μεταβατικός ευρύτερος όρος
skos:changeNote	Τεκμηρίωση
skos:closeMatch	Σύνδεση κοντινών εννοιών
skos:definition	Ορισμός
skos:editorialNote	Σημείωση του editor
skos:exactMatch	Σύνδεση ακριβώς ίδιων εννοιών

skos:example	παράδειγμα
skos:hasTopConcept	Έχει ως κορυφαία έννοια
skos:hiddenLabel	Κρυφή ετικέττα
skos:historyNote	Ιστορική σημείωση
skos:inScheme	Ανήκει στο σχήμα
skos:member	Μέλος μιας συλλογής
skos:memberList	Μέλος μιας διατεταγμένη συλλογή
skos:narrowMatch	Σύνδεση στενότερης σημασίας έννοιας
skos:narrower	Όρος στενότερης σημασίας
skos:narrowerTransitive	Μεταβατικός όρος στενότερης σημασίας
skos:notation	Σημειογραφία
skos:note	Σημείωση
skos:prefLabel	Προτιμώμενη ετικέττα
skos:related	Σχετιζόμενος όρος
skos:relatedMatch	Σύνδεση σχετιζόμενου όρου
skos:scopeNote	Σημείωση εμβέλειας εννοιας
skos:topConceptOf	Είναι κορυφαία έννοια του

Πίνακας 3-2 Λεξιλόγιο SKOS-XL

skosxl:Label	Ετικέτα
skosxl:altLabel	Εναλλακτική ετικέτα
skosxl:hiddenLabel	Κρυφή ετικέτα
skosxl:labelRelation	Σχετική ετικέτα
skosxl:literalForm	Κυριολεκτική μορφή
skosxl:prefLabel	Προτιμόμενη ετικέτα

4

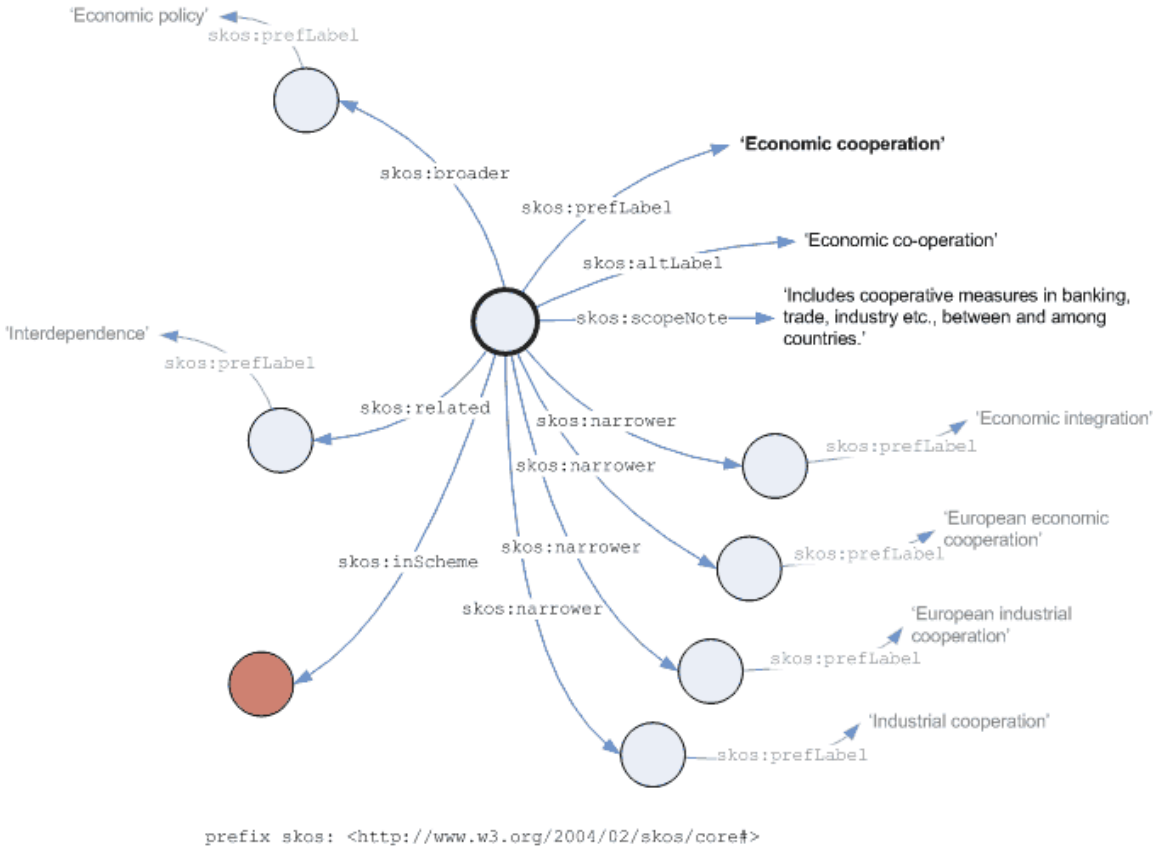
Μεθοδολογία για δημοσίευση θησαυρού στον Σημασιολογικό Ιστό

Ο Σημασιολογικός Ιστός, όπως είδαμε, παρέχει ένα κοινό πλαίσιο που επιτρέπει στα δεδομένα να μοιραστούν και να επαναχρησιμοποιηθούν από εφαρμογές. Η διαδικασία αυτή βασίζεται στην RDF, που παρέχει ένα απλό τρόπο παρουσίασης των δεδομένων ώστε να μιλάμε για πράγματα, τις ιδιότητές τους και τις μεταξύ τους σχέσεις. Χρησιμοποιώντας την RDF μπορούμε να συνδέσουμε είτε να συγχωνεύσουμε τα δεδομένα μας με άλλα δεδομένα RDF του Σημασιολογικού Ιστού. Στην πράξη, αυτό σημαίνει ότι οι πηγές των δεδομένων μπορούν να διανεμηθούν σε όλο τον Ιστό με αποκεντρωμένο τρόπο, αλλά ταυτόχρονα να ενσωματώνονται σε εφαρμογές, συχνά με νέους και απρόβλεπτους τρόπους. Για τους θησαυρούς χρησιμοποιούμε το SKOS το οποίο είναι ένα σύνολο ιδιοτήτων και κλάσεων που μπορούν να χρησιμοποιηθούν για να εκφράσουν το εννοιολογικό περιεχόμενο του, ως γράφημα RDF. Ακόμα για ένα θησαυρό πρέπει να διαθέτουμε URIs για κάθε έννοιά του, επιτρέπουμε έτσι, σε άλλους χρήστες να αναφέρονται στις έννοιες του, χωρίς να πρέπει να έχουν και το περιεχόμενο του. Τα URI με τα οποία συνδέουμε τις έννοιες μας πρέπει να είναι σταθερά σύμφωνα με τις αρχές για δημοσίευση καλών URIs (18) στο Σημασιολογικό Ιστό. Η RDF μπορεί επίσης, να χρησιμοποιηθεί για να εκφράσει τις μετά-ιδιότητες ενός θησαυρού, όπως ο τίτλος, η περιγραφή του, την ημερομηνία τροποποίησης και ούτω καθεξής. Τέλος, να αναφέρουμε ότι οι περισσότεροι θησαυροί διαχειρίζονται μέσω ενός συστήματος διαχείρισης θησαυρού. Το σύστημα αποθηκεύει το θησαυρό σε μια σχεσιακή βάση δεδομένων, ή σε XML ή σε άλλη

δομημένη μορφή αρχείου κειμένου. Θα δώσουμε ένα σύντομο παράδειγμα, μετατροπής ενός θησαυρού σε SKOS (46). Έστω ότι έχουμε το παρακάτω κομμάτι από τον θησαυρό UK Archival Thesaurus (UKAT):

Term: Economic cooperation
Used For:
Economic co-operation
Broader terms:
Economic policy
Narrower terms:
Economic integration
European economic cooperation
European industrial cooperation
Industrial cooperation
Related terms:
Interdependence
Scope Note:
Includes cooperative measures in banking, trade, industry etc., between and among countries.

Σαν γράφος RDF και κάνοντας χρήση του λεξιλογίου SKOS θα γίνει



Εικόνα 3.4.4.1 Γράφος RDF του UKAT

Η τελική μορφή του όταν τοποθετηθούν τα URIs των εννοιών του, αλλά και οι μετά-ιδιότητες του θησαυρού θα είναι:

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:skos="http://www.w3.org/2004/02/skos/core#"
   xmlns:dc="http://purl.org/dc/elements/1.1/">
   <skos:ConceptScheme rdf:about="http://www.ukat.org.uk/thesaurus">
      <dc:title>The UK Archival Thesaurus</dc:title>
      <dc:description>A subject thesaurus produced to support indexing in the UK
        archive sector.</dc:description>
      <dc:creator>UK Archival Thesaurus project</dc:creator>
      <dc:date>2004-08-22</dc:date>
      <dc:language>en</dc:language>
      <dc:rights>All rights reserved. Data in the UK Archival Thesaurus may be freely
```

```
        used and copied, without prior permission, for educational and other non-commercial purposes. These purposes include (but are not limited to) the incorporation of UKAT data into indexes, thesauri and finding aids created by organisations and projects in the archive sector and the wider heritage sector, in the UK and elsewhere. Under no circumstances may copies of UKAT data be sold without prior written permission fom the UKAT Project support@ukat.org.uk).</dc:rights>
        <skos:hasTopConcept rdf:resource="http://www.ukat.org.uk/thesaurus/field/1"/>
        <skos:hasTopConcept rdf:resource="http://www.ukat.org.uk/thesaurus/field/2"/>
        <skos:hasTopConcept rdf:resource="http://www.ukat.org.uk/thesaurus/field/3"/>
        <skos:hasTopConcept rdf:resource="http://www.ukat.org.uk/thesaurus/field/4"/>
        <skos:hasTopConcept rdf:resource="http://www.ukat.org.uk/thesaurus/field/5"/>
        <skos:hasTopConcept rdf:resource="http://www.ukat.org.uk/thesaurus/field/6"/>
        <skos:hasTopConcept rdf:resource="http://www.ukat.org.uk/thesaurus/field/8"/>
    </skos:ConceptScheme>
</rdf:RDF>
```

Στο παραπάνω παράδειγμα, χρησιμοποιούμε το skos:hasTopConcept ώστε να δείξουμε τις κορυφαίες έννοιες του UKAT. Η χρήση της ιδιότητας αυτής βοηθάει τις εφαρμογές, στον αποτελεσματικό εντοπισμό των δημοφιλέστερων εννοιών, για το δεδομένο σύστημα.

Έχοντας μετατρέψει τον θησαυρό μας, θα πρέπει να εκδώσουμε στον Ιστό τα δεδομένα μας (15). Ο απλούστερος τρόπος είναι η δημιουργία ενός ή περισσότερων RDF έγγραφων που περιέχουν τα δεδομένα και να τα δημοσιεύσουμε στο διαδίκτυο μέσω ενός διακομιστή HTTP. Υπάρχουν ειδικευμένοι διακομιστές για RDF όπως Joseki ή Sesame. Η δημοσίευση μέσω ενός εξυπηρετητή RDF επιτρέπει σε οποιονδήποτε να κάνει αναζήτηση στον θησαυρό ,στο διαδίκτυο, μέσω SPARQL, μιας γλώσσας επερωτήσεων για RDF. Στην συνεχεία του κεφαλαίου, θα δούμε ποιο αναλυτικά την διαδικασία μετατροπής θησαυρού σε SKOS.

4.1 Μετατροπή θησαυρών σε SKOS

Οι θησαυροί είναι ελεγχόμενα λεξιλόγια που δημιουργήθηκαν από ομάδες εργασίας με στόχο την είναι την ευρετηρίαση αλλά και την εύρεση δεδομένων από διάφορες πηγές. Οι περισσότεροι από αυτούς βρίσκονται σε μορφή XML αρχείων. Για να μπορεί να γίνει χρήση της γνώσης, που αυτοί περιέχουν, από τον Σημασιολογικό Ιστό

πρέπει να μετατραπούν σε RDF/OWL αναπαράσταση. Όμως, ο ίδιος θησαυρός μπορεί να καταλήξει να έχει διαφορετική δομή ανάλογα με την μετατροπή που έχει υποστεί. Η υιοθέτηση ενός κοινού πλαισίου για την αναπαράσταση του σε RDF/OWL θα έχει τα εξής οφέλη (47):

- μείωση του κόστους διαμοίρασης του θησαυρού.
- δυνατή η χρήση πολλαπλών θησαυρών στο πλαίσιο μιας εφαρμογής.
- δυνατή η ανάπτυξη πρότυπου λογισμικού για την επεξεργασία των θησαυρών.

Ωστόσο, οι θησαυροί διαφέρουν αρκετά ως προς τα χαρακτηριστικά τους. Η πρόκληση για το SKOS είναι να μπορέσει να συλλάβει όλα τα απαραίτητα χαρακτηριστικά των θησαυρών και να παρέχει επαρκή επεκτασιμότητα ώστε να είναι δυνατή η αναπαράσταση τοπικών χαρακτηριστικών.

Οι υπάρχουσες μέθοδοι μετατροπής θησαυρού σε SKOS είναι οι εξής:

1. Μέθοδοι μετατροπής για συγκεκριμένους θησαυρούς από την αρχική τους μορφή στην RDF/OWL. Το μειονέκτημα της είναι ότι δεν είναι ξεκάθαρο αν μπορούν να γενικευτούν και να εφαρμοστούν σε άλλους θησαυρούς εφόσον καλύπτονται μονάχα τα χαρακτηριστικά που εμφανίζονται στο συγκεκριμένο θησαυρό.
2. Μέθοδοι που μετατρέπουν τους θησαυρούς σε οντολογίες. Αυτή η μέθοδος έχει ι τρία βήματα: Πρώτα ορίζουμε το μετα-μοντέλο της οντολογίας. Έπειτα ορίζουμε τους κανόνες που θα χρησιμοποιηθούν για τη μετατροπή ενός παραδοσιακού θησαυρού στο μετα-μοντέλο και τέλος γίνεται χειρωνακτική διόρθωση.
3. Μέθοδοι που μετατρέπουν κάθε θησαυρό σε RDF/OWL χωρίς δημιουργία μιας οντολογίας. Αυτή η προσέγγιση απαιτεί τέσσερα βήματα:

1. Προετοιμασία
2. Συντακτική μετατροπή
3. Σημασιολογική μετατροπή
4. Προτυποποίηση

Στο πρώτο βήμα αναλύεται ο θησαυρός και η ψηφιακή μορφή του. Χρησιμοποιούμε την γνώση αυτή στο δεύτερο βήμα, για να γίνει η μετατροπή του σε βασική RDF, η οποία μετά μετατρέπεται σε ένα πιο κοινό μοντέλο που χρησιμοποιεί RDF και OWL. Στο τελικό βήμα, το RDF-OWL μετα-μοντέλο μετατρέπεται σε SKOS. Αυτή η μεθοδολογία βασίζεται σε δύο απαιτήσεις: τη διατήρηση της αρχικής σημασιολογίας του θησαυρού και την σταδιακή βελτίωση του μετά-μοντέλου RDF/OWL του θησαυρού.

Μια δεύτερη πρόταση που έχει προταθεί για να μετατραπεί ο θησαυρός σε μια πρώιμη μορφή SKOS έχει τρία βήματα (48):

1. Παραγωγή Κωδικοποίησης RDF.
2. Έλεγχος λαθών και επικύρωση της μορφής.
3. Δημοσιοποίηση κωδικοποίησης στο διαδίκτυο

Η προσέγγιση βασίζεται σε δύο απαιτήσεις: τη μετατροπή του θησαυρού στο SKOS μοντέλο, με στόχο την υποστήριξη διαλειτουργικότητας των θησαυρών και τη διατήρηση όλων των πληροφοριών που είναι κωδικοποιημένες στο θησαυρό. Το πρώτο βήμα διακρίνει τη μορφή του προς μετατροπή θησαυρού σε πρότυπο και μη πρότυπο. Οι πρότυποι βασίζονται στο ISO 2788 πρότυπο. Οι θησαυροί αυτοί μετατρέπονται σε στιγμιότυπα SKOS σχήματος χωρίς απώλεια πληροφορίας. Οι μη-πρότυποι θησαυροί είναι αυτοί που δεν έχουν δομικά χαρακτηριστικά και η δομή τους δεν ακολουθεί το ISO 2788. Η μέθοδος αυτή έχει στόχο την ανάπτυξη μίας επέκταση του σχήματος SKOS χρησιμοποιώντας τα rdfs:subClassOf και rdfs:subPropertyOf για την υποστήριξη των μη-πρότυπων χαρακτηριστικών. Το δεύτερο βήμα περιλαμβάνει επαλήθευση λαθών και επικύρωση της μορφής χρησιμοποιώντας τον RDF validator του W3C's, ενώ το τρίτο βήμα δεν έχει αναλυθεί περισσότερο.

Για την ανάπτυξη μιας μεθόδου μετατροπής θησαυρών, ο γενικός στόχος είναι η δημιουργία μιας μεθόδου που θα υποστηρίζει την διαλειτουργικότητα των θησαυρών που είναι κωδικοποιημένοι σε RDF/OWL. Η πρώτη απαίτηση που έχουμε από την μέθοδο αυτή είναι να παράγει πρόγραμμα μετατροπής που θα μετατρέπει την υπάρχουσα ψηφιακή απεικόνιση του θησαυρού σε SKOS παρέχοντας διαλειτουργικότητα. Επιπλέον θα πρέπει να παράγει ένα σωστό και έγκυρο SKOS RDF. Η δεύτερη απαίτηση είναι ο θησαυρός που θα προκύψει από την μετατροπή να έχει ολόκληρη την πληροφορία που υπάρχει στον αρχικό. Να τονίσουμε εδώ, ότι η διαλειτουργικότητα, σε περίπτωση σύγκρουσης, είναι ποιο σημαντική από την μεταφορά ολοκληρωμένης πληροφορίας.

Σε αρκετές περιπτώσεις δημιουργείται μια αναπαράσταση RDF του θησαυρού, μέσω αυτοματοποιημένης διαδικασίας (π.χ. XSLT μετασχηματισμός). Όμως, όταν χρησιμοποιούμε μια αυτοματοποιημένη διαδικασία, πρέπει να διασφαλιστεί ότι η έξοδος που παράγεται είναι λογική και σύμφωνη με την προτεινόμενη χρήση του SKOS Core Vocabulary. Για παράδειγμα, εάν ένα XML περιέχει ένα στοιχείο XML του οποίου το όνομα είναι «scopenote» δεν θα πρέπει να γίνεται αυτομάτως δεκτό ότι

θα πρέπει να μετατραπεί σε «skos:scopeNote». Ίσως το «scopenote» να περιέχει ορισμούς, οπότε θα πρέπει να χρησιμοποιηθεί η «skos:definition». Για αυτό το λόγο, είναι πολύ σημαντικό, ο δημιουργός του προγράμματος μετατροπής να έχει πολύ καλή γνώση του περιεχομένου του θησαυρού.

Ο τρόπος και η μεθοδολογία μετατροπής, που εφαρμόστηκε στην παρούσα διπλωματική, βασίστηκε σε όσα προαναφέρθηκαν. Η διαλειτουργικότητα ήταν η πρώτη απαίτηση που είχαμε και ανάλογα με την αρχική μορφή του κάθε θησαυρού, δημιουργήθηκε ένα πρόγραμμα για την σωστή και πλήρης μετατροπή του σε SKOS RDF. Τέλος, το περιεχόμενο, σε μερικές περιπτώσεις, δεν κρίθηκε απαραίτητο να μεταφερθεί ολόκληρο στην νέα μορφή του θησαυρού.

4.2 Μεθοδολογίες για Δημοσίευση Συνδεδεμένων Δεδομένων

Στην ενότητα αυτή θα δούμε τι ποιες είναι οι μεθοδολογίες για την δημοσίευση των δεδομένων μας στον Ιστό των Συνδεδεμένων Δεδομένων και τι χρειάζεται να προσέξουμε κατά την διαδικασία αυτή (15).

4.2.1 Επιλογή λεξιλογίου

Αφού επιλέξουμε σωστά URIs για την περιγραφή των εννοιών μας, όπως αναλύθηκε σε προηγούμενη ενότητα, το επόμενο βήμα είναι να επιλέξουμε ένα ή περισσότερα λεξιλόγια. Για να είναι όσο το δυνατόν ευκολότερο για τις εφαρμογές πελάτη , να επεξεργαστούν τα δεδομένα, θα πρέπει να επαναχρησιμοποιούνται τα γνωστά λεξιλόγια, όπου είναι δυνατόν. Ο διαχειριστής πρέπει να καθορίζει νέους όρους μόνο εάν δεν μπορεί να βρει τους απαιτούμενους όρους σε υπάρχοντα λεξιλόγια.

Στην κοινότητα του Σημασιολογικού Ιστού έχει αναπτυχθεί ένα σύνολο γνωστών λεξιλογίων και πριν τον ορισμό νέων όρων είναι καλό να γίνεται έλεγχος για το κατά πόσο τα δεδομένα μπορούν να αναπαρασταθούν χρησιμοποιώντας όρους από αυτά τα λεξιλόγια.

- Friend-of-a-Friend (FOAF), λεξιλόγιο για την περιγραφή ανθρώπων.
- Dublin Core (DC), προσδιορίζει γενικές ιδιότητες μετα-δεδομένων.
- Semantically-Interlinked Online Communities (SIOC), λεξιλόγιο για την αναπαράσταση διαδικτυακών κοινοτήτων.
- Description of a Project (DOAP), λεξιλόγιο για την περιγραφή projects.
- Simple Knowledge Organization System (SKOS), λεξιλόγιο για την αναπαράσταση ταξονομιών και χαλαρά δομημένης γνώσης.

- Music Ontology, παρέχει όρους για την περιγραφή καλλιτεχνών, άλμπουμ και μουσικών κομματιών.
- Review Vocabulary, λεξιλόγιο για την αναπαράσταση αξιολογήσεων.
- Creative Commons (CC), λεξιλόγιο για την περιγραφή όρων αδειών (license terms)

Είναι συνηθισμένη πρακτική η ανάμιξη όρων από διαφορετικά λεξιλόγια. Ποιο συγκεκριμένα, προτείνεται η χρήση των ιδιοτήτων rdfs:label και foaf:depiction όποτε αυτό είναι δυνατό καθώς οι όροι αυτοί υποστηρίζονται από τις εφαρμογές πελάτη. Εάν απαιτούνται URI αναφορές για γεωγραφικούς τόπους, περιοχές έρευνας, γενικά θέματα, καλλιτέχνες, βιβλία ή CDs προτείνεται η χρήση URIs από πηγές δεδομένων μέσα από το W3C SWEO Linking Open Data, για παράδειγμα Geonames, DBpedia, Musicbrainz, dbtune ή το RDF Book Mashup. Τα δύο κύρια οφέλη από τη χρησιμοποίηση URIs από αυτές τις πηγές δεδομένων είναι:

1. Τα URIs είναι dereferenceable, με την έννοια ότι η περιγραφή μιας έννοιας μπορεί να ανακτηθεί από τον Ιστό. Για παράδειγμα, η χρήση του URI http://dbpedia.org/page/Doom για την αναγνώριση του ηλεκτρονικού παιχνιδιού Doom μας δίνει μια εκτεταμένη περιγραφή του παιχνιδιού, συμπεριλαμβανομένων αποσπασμάτων σε 10 διαφορετικές γλώσσες και ποικίλες κατηγοριοποιήσεις.
2. Τα URIs είναι ήδη συνδεδεμένα με URIs από άλλες πηγές δεδομένων. Για παράδειγμα, μπορεί κανείς να πλοηγηθεί από το URI της DBpedia http://dbpedia.org/resource/Berlin στα δεδομένα για το Βερολίνο που παρέχονται από το Geonames και το Eurostat. Επομένως, χρησιμοποιώντας εννοιολογικά URIs από αυτά τα σύνολα δεδομένων, διασυνδέει κάποιος τα δεδομένα του με ένα πλούσιο και γρήγορα αναπτυσσόμενο δίκτυο άλλων πηγών δεδομένων.

4.2.2 Προσδιορισμός νέων όρων

Στην περίπτωση που δεν μπορούν να βρεθούν καλά υπάρχοντα λεξιλόγια που να καλύπτουν όλες τις απαιτούμενες κλάσεις και ιδιότητες, τότε απαιτείται ο προσδιορισμός νέων όρων. Ο προσδιορισμός νέων όρων δεν είναι δύσκολος. Οι RDF κλάσεις και ιδιότητες είναι και οι ίδιες πόροι που ταυτοποιούνται με URIs και δημοσιεύονται στον Ιστό, επομένως ότι έχει ειπωθεί για τη δημοσίευση Συνδεδεμένων Δεδομένων εφαρμόζεται και σε αυτά επίσης. Ο ορισμός λεξιλογίων

μπορεί να γίνει χρησιμοποιώντας την RDF Vocabulary Description Language 1.0: RDF Schema ή την Web Ontology Language (OWL).

Κάποιοι κανόνες που μπορούν να εφαρμοστούν από όσους είναι εξοικειωμένοι με τις γλώσσες αυτές:

1. Αντί για τον προσδιορισμό νέων λεξιλογίων από την αρχή, συστήνεται η συμπλήρωση των υπαρχόντων λεξιλογίων με επιπλέον όρους (στο όνομα πεδίου του δημιουργού κάθε φορά) για την αναπαράσταση των δεδομένων.
2. Να είναι κατανοητά και για τους ανθρώπους και για τις μηχανές. Στο στάδιο αυτό της ανάπτυξης του Ιστού των Δεδομένων, περισσότεροι άνθρωποι παρά μηχανές θα έρχονται αντιμέτωποι με τον κώδικα, αν και εν πρώτοις ο Ιστός Δεδομένων προοριζόταν για μηχανές. Χρήσιμος είναι και ο πεζός λόγος, για παράδειγμα rdfs:comments για κάθε όρο που εφευρίσκεται. Επίσης, για κάθε όρο πρέπει να παρέχεται και μια ετικέτα, κάνοντας χρήση της ιδιότητας rdfs:label.
3. Τα URIs όρων πρέπει να είναι dereferenceable. Είναι σημαντικό ώστε οι πελάτες να μπορούν να εντοπίσουν τον ορισμό ενός όρου.
4. Χρήση όρων άλλων ανθρώπων. Η χρησιμοποίηση όρων άλλων ανθρώπων ή η παροχή αντιστοιχίσεων(mappings) προς αυτούς βοηθά στην προώθηση της ανταλλαγής δεδομένων στον Ιστό Δεδομένων, με τον ίδιο τρόπο που οι σύνδεσμοι υπερκειμένου έχτισαν τον παραδοσιακό Ιστό των Εγγράφων. Συνηθισμένες ιδιότητες για την παροχή τέτοιων αντιστοιχίσεων είναι οι rdfs:subClassOf και rdfs:subPropertyOf.
5. Ρητή δήλωση όλων των σημαντικών πληροφοριών. Για παράδειγμα, δήλωση όλων των πεδίων και πεδίων τιμών ρητά. Οι άνθρωποι μπορούν να κάνουν υποθέσεις, οι μηχανές όμως όχι, επομένως δεν πρέπει να παραλείπονται σημαντικές πληροφορίες.
6. Όχι στη δημιουργία υπερβολικά περιορισμένων, εύθραυστων μοντέλων, χρειάζεται ευελιξία για την περίπτωση επέκτασης. Για παράδειγμα, αν κάποιος χρησιμοποιεί την full-featured OWL για τον ορισμό του λεξιλογίου του, είναι πιθανό να δηλώσει πράγματα που οδηγούν σε αθέλητες συνέπειες και αντιφάσεις όταν κάποιος άλλος αναφερθεί σε έναν όρο με ορισμό διαφορετικού λεξιλογίου. Επομένως, είναι προτιμότερη η χρήση RDF-Schema για τον προσδιορισμό λεξιλογίων.

Ο Ιστός των Δεδομένων επομένως, βασίζεται στο συνδυασμό χρησιμοποίησης συνηθισμένων λεξιλογίων μαζί με εξειδικευμένους όρους που συνδέονται με αντιστοιχήσεις όπως κρίνεται αναγκαίο. Ο καθένας είναι ελεύθερος να δημοσιεύει λεξιλόγια στον Ιστό Δεδομένων, τα οποία μπορούν με τη σειρά τους να συνδεθούν μ ε άλλα σχετικά λεξιλόγια.

4.3 Δημιουργία RDF Συνδέσμων προς άλλες Πηγές Δεδομένων

Το ποιο σημαντικό κομμάτι των Συνδεδεμένων Δεδομένων είναι ο τρόπος αλλά και η διαδικασία που ακολουθείται ώστε να δημιουργηθεί ο Ιστός από διάφορα σύνολα δεδομένων. Οι RDF σύνδεσμοι δίνουν τη δυνατότητα στους φυλλομετρητές Συνδεδεμένων Δεδομένων να περιηγούνται μεταξύ πηγών δεδομένων και να ανακαλύπτουν επιπλέον δεδομένα. Το πεδίο εφαρμογής είναι αυτό που θα καθορίσει ποιες RDF ιδιότητες θα χρησιμοποιηθούν ως κατηγορήματα. Για παράδειγμα, οι συνήθως χρησιμοποιούμενες συνδετικές ιδιότητες στο πεδίο των περιγραφών ανθρώπων είναι οι foaf:knows, foaf:based_near και foaf:topic_interest.

Είναι συνηθισμένη πρακτική να χρησιμοποιείται η ιδιότητα owl:sameAs για να δηλώσει ότι και μια ακόμα πηγή δεδομένων παρέχει επίσης πληροφορίες για ένα συγκεκριμένο μη-πληροφοριακό πόρο. Ένας σύνδεσμος owl:sameAs υποδηλώνει ότι δύο URI αναφορές αναφέρονται όντως στο ίδιο πράγμα. Επομένως, η owl:sameAs χρησιμοποιείται για να συνδέσει διαφορετικά ταυτόσημα URI.

Για την επίτευξη της σύνδεσης υπάρχουν δύο τρόποι, είτε οι RDF σύνδεσμοι να τεθούν χειρωνακτικά, όπως είναι η περίπτωση των προφίλ FOAF είτε να παραχθούν από αυτόματους αλγόριθμους σύνδεσης. Η τελευταία προσέγγιση χρησιμοποιείται κυρίως για να διασύνδεει μεγάλα σύνολα δεδομένων.

4.3.1 Χειρονακτική τοποθέτηση συνδέσμων

Προτού θέσουμε RDF συνδέσμους χειρωνακτικά, πρέπει να γνωρίζουμε τα σύνολα δεδομένων τα οποία θέλουμε να συνδέσουμε (μπορούμε να τα αναζητήσουμε στη λίστα συνόλων δεδομένων του Linking Open Data project). Αφού αναγνωρίσουμε συγκεκριμένα σύνολα ως κατάλληλους στόχους σύνδεσης, μπορούμε να ψάξουμε μέσα σε αυτά για URI αναφορές με τις οποίες θέλουμε να συνδεθούμε. Εάν μια πηγή δεδομένων δεν παρέχει διεπαφή αναζήτησης, για παράδειγμα ένα τερματικό σημείο SPARQL ή μια HTML ηλεκτρονική φόρμα, μπορούμε να χρησιμοποιήσουμε

φυλλομετρητές Συνδεδεμένων Δεδομένων όπως ο Tabulator ή ο Disco για την εξερεύνηση του συνόλου και την ανακάλυψη των σωστών URIs.

Μπορούμε ακόμα να χρησιμοποιήσουμε υπηρεσίες όπως η Uriqr και το Sindice για να βρούμε υπάρχοντα URIs και να διαλέξουμε το πιο γνωστό σε περίπτωση που υπάρχουν πολλά υποψήφια. Η Uriqr μας επιτρέπει να βρίσκουμε URIs ατόμων που γνωρίζουμε, απλά ψάχνοντάς τους με το όνομά τους. Τα αποτελέσματα ιεραρχούνται σύμφωνα με το πόσο συχνά ένα συγκεκριμένο URI αναφέρεται σε RDF έγγραφα στον Ιστό αλλά θα χρειαστεί και λίγη ανθρώπινη ευφυΐα για την επιλογή του πιο κατάλληλου URI προς χρήση. Το Sindice ευρετηριοποιεί το Σημασιολογικό Ιστό και μπορεί να πει ποιες πηγές αναφέρουν ένα συγκεκριμένο URI. Με τον τρόπο αυτό μια υπηρεσία μπορεί να βοηθήσει στην επιλογή του πιο δημοφιλούς URI για μία έννοια.

Πρέπει να έχουμε υπόψη μας ότι οι πηγές δεδομένων μπορεί να χρησιμοποιούν HTTP-303 ανακατευθύνσεις για να ανακατευθύνουν τους πελάτες από URIs που ταυτοποιούν μη-πληροφοριακούς πόρους προς URIs που ταυτοποιούν πληροφοριακούς πόρους οι οποίοι περιγράφουν τους μη-πληροφοριακούς. Στην περίπτωση αυτή, πρέπει να βεβαιωθούμε ότι συνδεόμαστε στο αναφορικό URI που ταυτοποιεί το μη-πληροφοριακό πόρο και όχι στο έγγραφο για αυτόν.

4.3.2 Αυτοματοποιημένη δημιουργία RDF συνδέσμων

Η προηγούμενη προσέγγιση δεν επιτρέπει την κλιμάκωση σε μεγάλα σύνολα δεδομένων, όπως η διασύνδεση 70.000 τοποθεσιών στη DBpedia με τις αντίστοιχες εγγραφές τους στο Geonames. Σε τέτοιες περιπτώσεις έχει νόημα η χρήση ενός αλγόριθμου αυτοματοποιημένης σύνδεσης εγγραφών για τη δημιουργία RDF συνδέσμων μεταξύ πηγών δεδομένων.

Η σύνδεση εγγραφών είναι γνωστό πρόβλημα στην κοινότητα των βάσεων δεδομένων. Το Project Σύνδεσης Ανοιχτών Δεδομένων συλλέγει υλικό σχετικό με τη χρήση αλγορίθμων σύνδεσης εγγραφών στο πλαίσιο των Συνδεδεμένων Δεδομένων στη wiki σελίδα Equivalence Mining. Ακόμα, υπάρχει έλλειψη καλών και εύκολων στη χρήση εργαλείων για την αυτόματη δημιουργία RDF συνδέσμων. Επομένως, είναι συνηθισμένη η υλοποίηση αλγορίθμων σύνδεσης εγγραφών συγκεκριμένων συνόλων δεδομένων για τη δημιουργία συνδέσμων μεταξύ πηγών δεδομένων. Στη συνέχεια θα περιγράψουμε δύο κατηγορίες τέτοιων αλγορίθμων:

4.3.2.1 Αλγόριθμοι βασισμένοι σε σχήματα (Pattern-based Algorithms)

Σε διάφορα πεδία, υπάρχουν γενικώς αποδεκτά σχήματα ονοματολογίας. Για παράδειγμα, στο χώρο των εκδόσεων υπάρχουν οι αριθμοί ISBN και ISSN, στο χρηματοοικονομικό τομέα υπάρχουν τα αναγνωριστικά ISIN. Εάν αυτά τα αναγνωριστικά χρησιμοποιούνται ως μέρος των HTTP URIs που ταυτοποιούν συγκεκριμένους πόρους, είναι δυνατόν να χρησιμοποιήσουμε απλούς, βασισμένους σε σχήματα αλγόριθμους, για την παραγωγή RDF συνδέσμων μεταξύ αυτών των πόρων.

Ένα παράδειγμα πηγής δεδομένων που χρησιμοποιεί ISBN αριθμούς ως μέρος των URI της είναι το RDF Book Mashup, το οποίο για παράδειγμα χρησιμοποιεί το URI http://www4.wiwiss.fu-berlin.de/bookmashup/books/0747581088

για να ταυτοποιήσει το βιβλίο «Ο Χάρυ Πότερ και ο Ημίαιμος Πρίγκιπας». Η παρουσία του αριθμού ISBN σε αυτά τα URIs κατέστησε εύκολη για τη DBpedia τη δημιουργία owl:sameAs συνδέσμων μεταξύ βιβλίων της και των αντίστοιχων βιβλίων του RDF Book Mashup. Η DBpedia χρησιμοποιεί τον παρακάτω αλγόριθμο βασισμένο σε σχήμα:

1. Επανέλαβε για όλα τα βιβλία της DBpedia που έχουν αριθμό ISBN.
2. Δημιούργησε ένα σύνδεσμο owl:sameAs μεταξύ του URI ενός βιβλίου στη DBpedia και του αντίστοιχου URI του Mashup βιβλίου χρησιμοποιώντας το ακόλουθο σχήμα για τα URIs των Mashup βιβλίων: http://www4.wiwiss.fuberlin.de/bookmashup/books/{ISBN number}.

Τρέχοντας τον αλγόριθμο αυτό σε όλα τα βιβλία στη DBpedia είχε ως αποτέλεσμα να παραχθούν 9000 RDF σύνδεσμοι που συγχωνεύτηκαν με το σύνολο δεδομένων της DBpedia. Για παράδειγμα, ο παραγόμενος σύνδεσμος για τον Χάρυ Πότερ είναι:

```
<http://dbpedia.org/resource/Harry_Potter_and_the_Half-Blood_Prince>
owl:sameAs<http://www4.wiwiss.fu-berlin.de/bookmashup/books/0747581088>
```

4.3.2.2 Σύνθετοι αλγόριθμοι βασισμένοι σε ιδιότητες

Σε περιπτώσεις όπου δεν υπάρχουν κοινά αναγνωριστικά μεταξύ συνόλων δεδομένων, είναι απαραίτητο να χρησιμοποιήσουμε πιο σύνθετους αλγόριθμους σύνδεσης βασισμένους σε ιδιότητες. Επισημαίνουμε δύο αλγόριθμους παρακάτω:

- Διασύνδεση μεταξύ DBpedia και Geonames. Πληροφορίες για γεωγραφικές τοποθεσίες εμφανίζονται τόσο στη βάση δεδομένων του Geonames καθώς και στη DBpedia. Για να αναγνωρίσουμε τα μέρη που εμφανίζονται και στα δύο

σύνολα δεδομένων, η ομάδα του Geonames χρησιμοποιεί μια βασισμένη σε ιδιότητες μέθοδο ευρετικής, η οποία βασίζεται στον τίτλο του άρθρου μαζί με σημασιολογικές πληροφορίες όπως το γεωγραφικό μήκος και πλάτος, αλλά και η χώρα, η διοικητική διαίρεση, ο τύπος του χαρακτηριστικού, ο πληθυσμός και οι κατηγορίες. Τρέχοντας αυτή την ευρετική μέθοδο και στις δύο πηγές δεδομένων είχε ως αποτέλεσμα 70.500 αντιστοιχίσεις οι οποίες συγχωνεύτηκαν ως σύνδεσμοι Geonames owl:sameAs και με το σύνολο της DBpedia αλλά και με αυτό του Geonames.

- Διασύνδεση Jamendo και MusicBrainz.Για να θέσουν RDF συνδέσμους εταξύ καλλιτεχνών στα σύνολα δεδομένων του Jamendo και του Musicbrainz, οι συγγραφείς χρησιμοποιούν μια μετρική ομοιότητας η οποία συγκρίνει τα ονόματα των καλλιτεχνών καθώς και τους τίτλους των άλμπουμ και των τραγουδιών τους.

4.4 Τι πρέπει να επιστρέφεται σαν RDF περιγραφή για ένα URI

Έστω ότι έχουμε μετατρέψει τα δεδομένα μας σε τριάδες RDF, πρέπει να ορίσουμε ποιες τριάδες πρέπει να επιστρέφονται με την RDF απεικόνιση (μετά από μια επαναδρομολόγηση 303) .

Η περιγραφή: Η παρουσίαση θα πρέπει να περιλαμβάνει όλες τις τριάδες από το σύνολο των δεδομένων σας που έχουν το URI του πόρου ως υποκείμενο (subject). Αυτή είναι η άμεση περιγραφή του πόρου.

Επιστροφή: Η παράσταση θα πρέπει να περιλαμβάνει επίσης όλες τις τριάδες από το σύνολο των δεδομένων π που έχουν το URI του πόρου ως αντικείμενο (subject). Αυτό μπορεί να είναι περιττό, καθώς αυτές οι τριάδες ίσως ήδη έχουν ανακτηθεί από το υποκείμενο URIs τους, αλλά επιτρέπει στους φυλλομετρητές και τα προγράμματα ανίχνευσης να διασχίσουν τους συνδέσμους προς κάθε κατεύθυνση.

Σχετικές περιγραφές: Μπορείτε να συμπεριλάβετε οποιαδήποτε περαιτέρω πληροφορίες σχετικά με τους πόρους που μπορεί να παρουσιάζει ενδιαφέρον για ένα τυπικό σενάριο χρήσης. Για παράδειγμα, μπορεί να θέλουμε να αποστείλουμε πληροφορίες για το δημιουργό, μαζί με πληροφορίες για ένα βιβλίο, γιατί πολλοί πελάτες που ενδιαφέρονται για το βιβλίο μπορεί επίσης να ενδιαφέρεται για τον συγγραφέα. Βέβαια δεν πρέπει να υπερβάλουμε στην επιστρεφόμενη πληροφορία η

επιστροφή ενός megabyte σε RDF θεωρείτε υπερβολικό στις περισσότερες περιπτώσεις.

Μεταδεδομένα: Η αναπαράσταση θα πρέπει να περιέχει μετά-δεδομένα που θέλουμε να επισυνάψουμε στα δημοσιευμένα μας δεδομένα, όπως το URI που προσδιορίζει την άδεια και έχει πληροφορίες για το συντάκτη. Για να έχουν οι πελάτες τη δυνατότητα να αξιολογούν την ποιότητα των δημοσιευμένων δεδομένων και να καθορίσουν εάν θέλουν να εμπιστευτούν τα δεδομένα, τα δεδομένα πρέπει να συνοδεύονται από μετά-πληροφορίες για το δημιουργό τους, την ημερομηνία δημιουργίας τους καθώς και τη μέθοδο δημιουργίας. Τέλος, κάθε έγγραφο RDF θα πρέπει να περιέχει μια άδεια βάσει της οποίας το περιεχόμενο μπορεί να χρησιμοποιηθεί(Creative Commons) ώστε οι καταναλωτές να ξέρουν, από νομική άποψη, πώς να χρησιμοποιήσουν τα δεδομένα του.

Σύνταξη: Υπάρχουν διάφοροι τρόποι για να κωδικοποιηθεί μια περιγραφή RDF . Θα πρέπει τουλάχιστον να παρέχετε η περιγραφή RDF ως RDF / XML, η μόνη επίσημη σύνταξη για RDF. Επειδή η RDF / XML δεν είναι εύκολα αναγνώσιμη από άνθρωπο, τα δεδομένα θα μπορούσαν να είναι σε επιπλέον μορφές όπως την Turtle.

4.5 Έλεγχος και Αποσφαλμάτωση Συνδεδεμένων Δεδομένων

Μετά τη δημοσίευση πληροφοριών ως Συνδεδεμένα Δεδομένα στον Ιστό, πρέπει να ελέγξουμε εάν οι πληροφορίες μας μπορούν να προσπελαστούν σωστά. Ένας εύκολος τρόπος ελέγχου είναι να βάλουμε μερικά από τα URIs μας στην Vapour Linked validation service, μια υπηρεσία που επαληθεύει ότι τα δημοσιευμένα δεδομένα συμμορφώνονται με τις αρχές των Συνδεδεμένων Δεδομένων και παράγει μια λεπτομερή αναφορά για το πώς τα URIs μας αντιδρούν σε διαφορετικά HTTP αιτήματα.

Εκτός από αυτό, είναι επίσης απαραίτητο να δούμε εάν οι πληροφορίες μας εμφανίζονται σωστά σε διαφορετικούς φυλλομετρητές Συνδεδεμένων Δεδομένων και εάν οι φυλλομετρητές μπορούν να ακολουθήσουν RDF συνδέσμους μέσα στα δεδομένα μας. Επομένως, μπορούμε να πάρουμε κάποια URIs από το σύνολο δεδομένων μας και να τα βάλουμε στην μπάρα πλοήγησης των παρακάτω φυλλομετητών Συνδεδεμένων Δεδομένων:

- Tabulator. Εάν ο Tabulator αργοπορεί κατά την εμφάνιση των πληροφοριών μας, τότε είναι ένδειξη ότι οι RDF γράφοι μας είναι πολύ μεγάλοι και πρέπει να διαχωριστούν. Ο Tabulator κάνει επίσης κάποιο βασικό συμπερασμό στα

δεδομένα Ιστού, χωρίς να κάνει ελέγχους συνέπειας. Επομένως, εάν ο Tabulator συμπεριφέρεται περίεργα, αυτό μπορεί να υποδεικνύει θέματα με τις δηλώσεις rdfs:subClassOf και rdfs:subPropertyOf μέσα σε RDFS και OWL σχήματα που χρησιμοποιούνται στα δεδομένα μας.

- Disco. Ο Disco φυλλομετρητής χρησιμοποιεί ένα διάλειμμα 2 δευτερολέπτων όταν ανακτά δεδομένα από τον Ιστό. Επομένως, ίσως είναι ένας δείκτης ότι ο εξυπηρετητής μας είναι πολύ αργός, εάν ο Disco δεν εμφανίζει σωστά τα δεδομένα μας.

4.6 Τρόποι για την παρουσίαση της πληροφορίας σαν Συνδεδεμένα Δεδομένα

Οι πληροφορίες που παρέχουμε, πρέπει να πληρούν τις παρακάτω ελάχιστες προϋποθέσεις ώστε να θεωρηθούν «δημοσιευμένες ως Συνδεδεμένα Δεδομένα στον Ιστό»:

- Τα αντικείμενα πρέπει να ταυτοποιούνται με dereferenceable HTTP URIs.
- Εάν επισκεφθούμε ένα τέτοιο URI ζητώντας τον MIME τύπο application/rdf+xml, μια πηγή δεδομένων πρέπει να επιστρέψει μια RDF/XML περιγραφή του ταυτοποιημένου πόρου.
- Τα URIs που ταυτοποιούν μη-πληροφοριακούς πόρους πρέπει να τεθούν με έναν από τους παρακάτω τρόπους: Είτε η πηγή των δεδομένων πρέπει να επιστρέφει μια HTTP απόκριση που να περιέχει μια HTTP 303 ανακατεύθυνση προς έναν πληροφοριακό πόρο που να περιγράφει το μη-πληροφοριακό πόρο, όπως περιγράφηκε νωρίτερα ή το URI για το μη-πληροφοριακό πόρο πρέπει να σχηματιστεί παίρνοντας το URI του σχετιζόμενου πληροφοριακού πόρου και προσαρτώντας σε αυτό ένα τμηματικό αναγνωριστικό (fragment identifier, π.χ. #foo).
- Εκτός από RDF συνδέσμους προς πόρους μέσα στην ίδια πηγή δεδομένων, οι RDF περιγραφές πρέπει να περιέχουν επίσης RDF συνδέσμους προς πόρους που παρέχονται από άλλες πηγές δεδομένων, έτσι ώστε οι πελάτες να μπορούν να περιηγούνται στον Ιστό των Δεδομένων ως όλο ακολουθώντας RDF συνδέσμους.

Μετά τη δημοσίευση των πληροφοριών μας ως Συνδεδεμένα Δεδομένα, πρέπει να σιγουρευτούμε ότι υπάρχουν εξωτερικοί RDF σύνδεσμοι που να δείχνουν σε URIs

του συνόλου δεδομένων μας, έτσι ώστε οι RDF φυλλομετρητές και οι crawlers να μπορούν να ανακαλύψουν τα δεδομένα μας. Υπάρχουν δύο βασικοί τρόποι για να το κάνουμε αυτό:

1. Να προσθέσουμε αρκετούς RDF συνδέσμους στο FOAF προφίλ μας που να δείχνουν σε URIs τα οποία αναγνωρίζουν κεντρικούς πόρους μέσα στο σύνολο των δεδομένων μας. Υποθέτοντας ότι κάποιος άλλος στον κόσμο μας γνωρίζει και επισκέπτεται το FOAF προφίλ, το νέο σύνολο δεδομένων μας είναι τώρα προσεγγίσιμο ακολουθώντας RDF συνδέσμους.
2. Να πείσουμε τους ιδιοκτήτες των σχετιζόμενων πηγών δεδομένων να παράγουν αυτόματα RDF συνδέσμους προς τα URIs του συνόλου μας. Ή, διευκολύνοντας τον ιδιοκτήτη του άλλου συνόλου, να δημιουργήσουμε τους RDF συνδέσμους οι ίδιοι και να τους στείλουμε σε αυτόν έτσι ώστε να τους συγχωνεύσει με το σύνολο των δεδομένων της. Ένα project το οποίο είναι εξαιρετικά ανοιχτό στην τοποθέτηση RDF συνδέσμων προς άλλες πηγές δεδομένων είναι το DBpedia community project. Αυτό μπορεί να γίνει ανακοινώνοντας απλά την πηγή δεδομένων στη DBpedia mailing list ή στέλνοντας ένα σύνολο RDF συνδέσμων στη λίστα.

5

Μετατροπή θησαυρών για χρήση στον Σημασιολογικό Ιστό σε μορφή SKOS

Στην ενότητα αυτή θα εφαρμόσουμε την παραπάνω γνώση σε πραγματικές συνθήκες. Στόχος μας είναι, η μετατροπή αλλά και η δημιουργία θησαυρών γνώσης από μια αρχική μορφή σε μορφή SKOS, ώστε να είναι διαθέσιμη για χρήση στον Σημασιολογικό Ιστό. Εξετάζουμε διαφορετικά σενάρια τα οποία απαιτούν και διαφορετικές αντιμετωπίσεις. Αρχικά θα δημιουργήσουμε ένα γεωγραφικό θησαυρό από το μηδέν μέσω του Geonames, στη συνέχεια θα δούμε πως μπορεί να γίνει χειρωνακτική μετατροπή ενός τοπικού θησαυρού σε SKOS κάνοντας χρήση δανικών URI μέσω του IPTC και τέλος εξετάζουμε την αυτοματοποιημένη μετατροπή ενός συγκεκριμένου τύπου XML θησαυρών σε SKOS.

5.1 Δημιουργία θησαυρού από το GEONAMES σε SKOS

Το Geonames, όπως αναλύσαμε είναι ένα σύνολο δεδομένων γεωγραφικού περιεχομένου με πάνω από 10,000,000 γεωγραφικά ονόματα. Κάθε στοιχείο της βάσης αυτής αναγνωρίζεται από ένα σταθερό και μοναδικό URI, το οποίο με τη σειρά του παρέχει πρόσβαση σε ένα σύνολο πληροφοριών. Έπειτα από διαπραγμάτευση περιεχομένου και μέσω διαφόρων παρεχόμενων εργαλείων, μπορούμε να εξάγουμε την πληροφορία, η οποία είναι διαθέσιμη με άδεια Creative Commons Attribution 3.0 License και σε διάφορες μορφές όπως, XML, JSON και RDF.

Ο γεωγραφικός θησαυρός μας θέλαμε να είναι ένα υποσύνολο του Geonames και να χρησιμοποιεί τα δικά του URI, τα οποία είναι σταθερά και dereferenceable. Έτσι δεν

χρειάζεται εμείς να τα παρέχουμε μέσω κάποιου διακομιστή. Η τεχνική χρησιμοποίησης γνωστών URIs για τον ορισμό των αντικειμένων μας, αντί για δημιουργία νέων, είναι συνηθισμένη για μικρές και τοπικές βάσεις και παρέχει το πλεονέκτημα ότι τα URIs αυτά, είναι ήδη συνδεδεμένα με URIs από άλλες πηγές δεδομένων. Για παράδειγμα, μέσω του Geonames, μπορεί κανείς να πλοηγηθεί σε άρθρα της wikipedia.

Για την λήψη δεδομένων το Geonames, παρέχει μια σειρά από εργαλεία. Εμείς χρησιμοποιήσαμε την απλή εντολή search, η οποία είναι η μόνη που παρείχε τα δεδομένα και σε RDF μορφή, γεγονός που θα διευκόλυνε την μετατροπή σε SKOS.

5.1.1 Περιγραφή του εργαλείου GEONAMES SEARCH

Το Geonames παρέχει στον χρήστη τη δυνατότητα να εξάγει από την βάση του τα δεδομένα που επιθυμεί. Με το εργαλείο «Geonames Search Webservice» μπορούμε επισκεπτόμενοι το Url:

api.Geonames.org/search?

να ρυθμίσουμε μια σειρά από παραμέτρους, ώστε να λάβουμε μια λίστα με ονόματα, που αντιστοιχούν στους περιορισμούς που θέσαμε. Στο παράρτημα Α επισυνάπτεται η λίστα με τις παραμέτρους του Geonames Search Web service. Για παράδειγμα, έστω ότι θέλουμε να πάρουμε πληροφορίες για το Λονδίνο θα πρέπει να εισάγουμε τα εξής URL ανάλογα με την μορφή που θέλουμε τα δεδομένα:

XML http://api.Geonames.org/search?q=london&maxRows=10&username=demo JSON http://api.Geonames.org/searchJSON?q=london&maxRows=10&username=demo RDF http://api.Geonames.org/search?q=london&maxRows=10&type=rdf&username=dem o

Όπως βλέπουμε ελέγχοντας τις παραμέτρους θα έχουμε και το αποτέλεσμα που θέλουμε. Δείγματα των μορφών XML και RDF (οι οποίες μας ενδιέφεραν άμεσα) παρουσιάζοντα παρακάτω.

XML	RDF
<geoname>	<gn:Feature rdf:about="http://sws.Geonames.org/2643741/">
<toponymName>	<rdfs:isDefinedBy>
City of London	http://sws.Geonames.org/2643741/about.rdf
</toponymName>	</rdfs:isDefinedBy>
<name>	<gn:name>
City of London	City of London
</name>	</gn:name>
<lat>	<gn:featureClass
51.51279	rdf:resource="http://www.Geonames.org/ontology#P"/>
</lat>	<gn:featureCode
<lng>	rdf:resource="http://www.Geonames.org/ontology#P.PPLX"/>
-0.09184	<gn:countryCode>
</lng>	GB
<geonameId>	</gn:countryCode>
2643741	<gn:population>
</geonameId>	7556900
<countryCode>	</gn:population>
GB	<wgs84_pos:lat>
</countryCode>	51.51279
<countryName>	</wgs84_pos:lat>
United Kingdom	<wgs84_pos:long>
</countryName>	-0.09184
<fcl>	</wgs84_pos:long>
P	<gn:parentCountry
</fcl>	rdf:resource="http://sws.Geonames.org/2635167/"/>
<fcode>	<gn:nearbyFeatures
PPLX	rdf:resource="http://sws.Geonames.org/2643741/nearby.rdf"/>
</fcode>	<gn:locationMap
</geoname>	rdf:resource="http://www.Geonames.org/2643741/city-of-london.html"/>
	</gn:Feature>

Ακόμα παρατηρήθηκε ότι στην RDF μορφή, υπήρχε η ονομασία της περιοχής σε σχεδόν 30 γλώσσες. Με βάση το λεξιλόγιο του Geonames, κάθε πόρος έχει από ένα επίσημο όνομα με την ετικέτα <gn:name> που είναι συνήθως στην αγγλική και ανάλογα με το πόσο γνωστό είναι ένα τοπωνύμιο, υπάρχουν μεταφράσεις του με την ετικέτα <gn:alternateName xml:lang="el">. Ακόμα, βλέπουμε ότι επιστρέφονται μια σειρά από στοιχεία για το μέρος, όπως γεωγραφικές συντεταγμένες, όνομα χώρας που ανήκει κλπ. Στην RDF υπήρχε το URI του πόρου αλλά και το URI της χώρας στην οποία ανήκει π.χ.

```
http://sws.Geonames.org/2643741/about.rdf
```

περιγράφει την πόλη του Λονδίνου η οποία ανήκει στο Ηνωμένο Βασίλειο που ανήκει στο http://sws.Geonames.org/2635167/ .

Αν τώρα επισκεφτούμε απευθείας το URL ,που αφορά τον πόρο με το όνομα «London», http://sws.Geonames.org/2643743/about.rdf

θα δούμε ένα RDF αρχείο με την λέξη «London» με 101 ετικέτες σε σχεδόν 90 γλώσσες, τον πληθυσμό της περιοχής, κοντινές περιοχές και διάφορα άλλα στοιχεία. Το σημαντικότερο όμως από όλα είναι οι ετικέτες με το όνομα «gn:wikipediaArticle». Αυτές περιέχουν συνδέσεις μεταξύ του πόρου και της wikipedia. Οι ετικέτες αυτές έχουν την εξής μορφή:

```
<gn:wikipediaArticle rdf:resource="http://an.wikipedia.org/wiki/Londres"/>
<gn:wikipediaArticle rdf:resource="http://ang.wikipedia.org/wiki/Lunden"/>
```

Ακόμα περιέχετε σύνδεση με την dbpedia μέσω του:

```
<owl:sameAs rdf:resource="http://dbpedia.org/resource/London"/>
```

Όπως βλέπουμε, δηλαδή, αν διαθέτουμε το URI του Geonames για κάθε περιοχή που θέλουμε, αυτόματα έχουμε συνδέσεις του θησαυρού μας με τον χώρο των Συνδεδεμένων Δεδομένων. Τέλος, το γεγονός ότι το λεξιλόγιο που δημιουργήθηκε για την μορφή RDF του Geonames, έχει κοινά χαρακτηριστικά με αυτό που χρησιμοποιούμε στην SKOS, βοήθησε στην εύκολη αντιστοίχιση όρων.

5.1.2 Λήψη δεδομένων, δημιουργία θησαυρού και μετατροπή σε SKOS

Αρχικά, στον θησαυρό που θα δημιουργούσαμε έπρεπε να επιλέξουμε τη πληροφορία, που θέλαμε να περιλαμβάνει. Σε πρώτη φάση, ζητήθηκε να περιοριστούμε στην περιοχή της Ευρώπης και να έχουμε τις μεγαλύτερες της πόλεις. Όμως, στη συνέχεια, αποφασίσαμε να μεγαλώσει η βάση μας και να έχουμε μια συνολική εικόνα από όλες τις χώρες του πλανήτη. Επειδή το Geonames έχει ένα

σύνολο τοπωνυμιών από πόλεις μέχρι ονόματα βουνών και αεροδρομίων, χρειαζόταν να περιορίσουμε την έρευνα μας. Μέσω της παραμέτρου Feature Codes δίνετε η δυνατότητα στον χρήστη να επιλέξει περιοχές που τον ενδιαφέρουν. Για να έχουμε έναν παγκόσμιο θησαυρό, χρειαζόμασταν τα ονόματα για τις ηπείρους, τις χώρες, τις πρωτεύουσες τους αλλά και τις πρωτεύουσες όλων των νομών κάθε χώρας. Δυστυχώς, δεν έχουν όλες οι χώρες την ίδια μορφή διοικητικής διαίρεσης με την Ελλάδα (Χώρα, Νομός, Πρωτεύουσα Νομού). Αντίθετα, υπάρχουν ποιο σύνθετες μορφές όπως οι ΗΠΑ με τις Πολιτείες της, αλλά και ανεξάρτητα κρατίδια όπως το Βατικανό. Για να θεωρηθεί ολοκληρωμένος ένας γεωγραφικός θησαυρός έπρεπε να περιείχε όλη αυτή την πληροφορία. Τα Geonames Feature Codes που χρησιμοποιήσαμε για να το πετύχουμε αυτό συνοψίζονται στον παρακάτω πίνακα:

Πίνακας 5-1 Geonames Feature Codes

PPLA	έδρα της πρώτης τάξης διοικητικής διαίρεσης
PPLA2	έδρα της δεύτερης τάξης διοικητικής διαίρεσης
PPLA3	έδρα της τρίτης για τη διοικητικής διαίρεσης
PPLC	πρωτεύουσα μιας πολιτικής οντότητας
PCLI	ανεξάρτητη πολιτική οντότητα
CONT	ήπειρος

Το σύνολο των δεδομένων σε πρώτη φάση, έπρεπε να αποθηκευτεί σε ένα αρχείο, το οποίο στη συνέχεια θα επεξεργαζόμασταν για να έρθει στην τελική του μορφή. Δημιουργήσαμε έναν κώδικα σε γλώσσα Ruby που κάνει όλη την διαδικασία από την αρχή μέχρι το τέλος. Αναλυτικά,

```
require 'open-uri'
1.step(7,1) { |x|
continent = case x
when 1 then "AF"
when 2 then "AS"
when 3 then "EU"
when 4 then "NA"
when 5 then "OC"
```

```
when 6 then "SA"
when 7 then "AN"
end
0.step(5000,1000) { |i|
open("http://api.Geonames.org/search?continentCode=" + continent.to_s() +
"&startRow=" + i.to_s() +
"&maxRows=1000&type=rdf&username=xalara&style=full&featureCode=PPLA&fe
atureCode=PPLA2&featureCode=PPLA3&featureCode=PPLC&featureCode=PCLI
") {|src|  open("world.rdf","ab") {|dst|     dst.write(src.read)  }}}}
```

Αφού καλέσουμε το RubyGem που καλεί και ανοίγει ένα URI, δημιουργούμε ένα διπλό βρόχο. Παρατηρήθηκε ότι το εργαλείο Geonames Search δεν δίνει στον απλό χρήστη πάνω από 1000 γραμμές σε κάθε κλίση του. Για αυτό το λόγο δημιουργήθηκαν βρόχοι ώστε κάθε φορά να περιορίζεται το επιλεγμένο δείγμα σε 1000 γραμμές και η επόμενη κλήση να αρχίζει από την γραμμή 1000+1. Σε κάθε κλήση, ο μεγάλος βρόχος δίνει μια ήπειρο, της οποίας την πληροφορία παίρνουμε αφού ολοκληρωθούν οι επαναλήψεις του ενσωματωμένου βρόχου. Ο μικρός βρόχος βλέπουμε ότι σε κάθε κλήση του δημιουργεί ένα URI, το οποίο κάνει χρήση του εργαλείου Search. Το URI χρησιμοποιεί τις παραμέτρους που είναι συνδεδεμένες με το ‘AND' :

- την ήπειρο που θα ψάξει (continentCode),
- την γραμμή από την οποία θα αρχίσουμε να καταγράφουμε(startRow),
- το μέγιστο αριθμό των γραμμών που θέλουμε να επιστρέψει(maxRows),
- τον τύπο δεδομένων που θα μας επιστρέψει(type=rdf) ,
- το όνομα του χρήστη που το καλεί (username),
- το γεγονός ότι θέλουμε την πλήρη πληροφορία για κάθε περιοχή και όχι κάποια συντόμευση της (style=full)
- όλες τις διοικητικές περιφέρειες που θέλουμε και αναλύθηκαν προηγουμένως (featureCode),

τέλος σε κάθε κλήση του μικρού βρόχου, τα δεδομένα αποθηκεύονται σε ένα αρχείο με το όνομα «world.rdf».

```
open("http://api.Geonames.org/search?type=rdf&username=xalara&style=full&featur
eCode=CONT") {|src|
  open("world.rdf","ab") {|dst|
```

```
    dst.write(src.read)  }}
```

με αυτήν την κλήση προσθέτουμε στο αρχείο «world.rdf» όλες τις Ηπείρους, αφού αυτό είναι απαραίτητο για την ιεραρχία που θα δημιουργήσουμε στην συνέχεια. Έτσι καταλήγουμε σε ένα αρχείο σε rdf που έχει στοιχεία της μορφής

```
<gn:Feature rdf:about="http://sws.Geonames.org/360630/">
<rdfs:isDefinedBy> http://sws.Geonames.org/360630/about.rdf </rdfs:isDefinedBy>
<gn:name> Cairo </gn:name>
<gn:alternateName xml:lang="af"> Kaïro </gn:alternateName>
<gn:alternateName xml:lang="als"> Kairo </gn:alternateName>
<gn:alternateName xml:lang="am"> ካይሮ </gn:alternateName>
<gn:featureClass rdf:resource="http://www.Geonames.org/ontology#P"/>
<gn:featureCode rdf:resource="http://www.Geonames.org/ontology#P.PPLC"/>
<gn:countryCode> EG </gn:countryCode>
<gn:population> 7734614 </gn:population>
<wgs84_pos:lat> 30.06263 </wgs84_pos:lat>
<wgs84_pos:long> 31.24967 </wgs84_pos:long>
<gn:parentCountry rdf:resource="http://sws.Geonames.org/357994/"/>
<gn:nearbyFeatures rdf:resource="http://sws.Geonames.org/360630/nearby.rdf"/>
<gn:locationMap rdf:resource="http://www.Geonames.org/360630/cairo.html"/>
</gn:Feature>
```

Έχοντας το θησαυρό μας σε ένα rdf αρχείο, που χρησιμοποιεί το λεξιλόγιο της Geonames, αυτό που μένει είναι να γίνει αντιστοίχιση των όρων μεταξύ Geonames και SKOS λεξιλογίου. Αρχικά, αποφασίστηκε η απαλοιφή των όρων: gn:featureCode, gn:countryCode, gn:population, wgs84_pos:lat, wgs84_pos:long, gn:nearbyFeatures, gn:locationMap. Η πληροφορία που περιλάμβαναν αυτοί οι όροι κρίθηκε περιττή, για τον τελικό θησαυρό μας, από την στιγμή που θα διαθέταμε το URI για κάθε τοπωνύμιο. Με απλή κλήση του URI του πόρου, θα έχουμε πρόσβαση στο σύνολο των πληροφοριών που έχει συλλέξει το Geonames για αυτόν. Αντίθετα, απαραίτητα στοιχεία για τον επιθυμητό θησαυρό ήταν το όνομα του πόρου, οι εναλλακτικές ετικέτες, ώστε να υποστηρίζετε άμεσα η πολυγλωσσικότητα και τέλος το στοιχείο gn:parentCountry, αφού είναι απαραίτητο στην SKOS να υπάρχει μια ιεραρχία μεταξύ των πόρων.

Έτσι δημιουργήσαμε την ακόλουθη χαρτογράφηση :

Πίνακας 5-2 Χαρτογράφηση Geonames

Geonames	SKOS
gn:alternateName	skos:prefLabel
gn:Feature	skos:Concept
rdfs:isDefinedBy	skos:exactMatch
gn:name	skos:prefLabel

Η διαδικασία τροποποίησης, επειδή η χαρτογράφηση ήταν αρκετά απλή, έγινε με αντικατάσταση:

```
text = File.read("world.rdf")
replace = text.gsub(/<rdf:RDF xmlns:cc="ht.*/, "")
replace = replace.gsub(/\<\?xml vers.*/, "")
replace = replace.gsub(/gn:alternateName/, "skos:prefLabel")
replace = replace.gsub(/gn:Feature/, "skos:Concept")
replace = replace.gsub(/rdfs:isDefinedBy/, "skos:exactMatch")
replace = replace.gsub(/gn:parentCountry/, "skos:broader")
replace = replace.gsub(/gn:name/, "skos:prefLabel")
replace = replace.gsub(/<gn:featureCode rdf.*/, "<skos:inScheme
rdf:resource=\"http://www.Geonames.org\"/>")
replace = replace.gsub(/<gn.*/, "")
replace = replace.gsub(/<wgs84.*/, "")
replace = replace.gsub(/<\/rdf:RDF>/, "")
File.open("world.rdf", "wb") {|file| file.puts replace}
```

Το αρχείο, που δημιουργήθηκε προηγουμένως, φορτώνεται στην μεταβλητή text έπειτα με μια σειρά από αντικαταστήσεις και διαγραφές φτάνουμε στο επιθυμητό αποτέλεσμα. Να σημειώσουμε εδώ, ότι σύμφωνα με τους κανόνες της SKOS, κάθε στοιχείο πρέπει να ανήκει σε ένα σχήμα, που στην περίπτωση μας είναι το:

```
skos:inScheme rdf:resource=http://www.Geonames.org/ .
```

Επιπλέον, όπως φαίνεται στην χαρτογράφηση, η ετικέτα gn:name και η gn:alternateName αντιστοιχήθηκε στην skos:prefLabel. Όπως είδαμε στην ανάλυση για SKOS ένα στοιχείο δεν μπορεί να έχει πάνω από ένα όνομα ανά γλώσσα. Εδώ, όμως δεν έχουμε πρόβλημα αφού η gn:alternateName έχει γλωσική ετικέτα ενώ το gn:name όχι. Επίσης, σε λιγότερους δημοφιλής όρους, το Geonames δεν παρέχει εναλλακτικές ετικέτες και έτσι το gn:name είναι η μόνη ονομασία που υπάρχει. Αφού

στον τελικό θησαυρό βάλουμε και τις αντίστοιχες επικεφαλίδες στην αρχή και στο τέλος του εγγράφου, έχουμε έναν θησαυρό σε έγκυρη μορφή SKOS. Για παράδειγμα το ένα δείγμα από το Κάιρο μετατράπηκε στην παρακάτω μορφή:

```
<skos:Concept rdf:about="http://sws.Geonames.org/360630/">
<skos:exactMatch>http://sws.Geonames.org/360630/about.rdf</skos:exactMatch>
<skos:prefLabel>Cairo</skos:prefLabel>
<skos:prefLabel xml:lang="tl">Lungsod ng Cairo</skos:prefLabel>
<skos:prefLabel xml:lang="th">ไคโร</skos:prefLabel>
<skos:prefLabel xml:lang="az">Qahirə</skos:prefLabel>
<skos:inScheme rdf:resource="http://www.Geonames.org"/>
<skos:broader rdf:resource="http://sws.Geonames.org/357994/"/>
</skos:Concept>
```

5.1.3 Τελικό αποτέλεσμα

Το τελικό αποτέλεσμα είναι ένας γεωγραφικός θησαυρός σε SKOS, που αποτελείται από 8040 τοπωνύμια από όλο τον κόσμο. Αναλυτικά έχουμε από

- Ευρώπη 3111
- Αφρική 776
- Ασία 3671
- Βόρεια Αμερική 173
- Ωκεανία 134
- Νότια Αμερική 166
- Ανταρκτική 2

Συνολικά 8033 περιοχές και 7 Ήπειροι.

Το πρόγραμμα της Ruby μπορεί εύκολα, να τροποποιηθεί, ανάλογα με τις ανάγκες μας, ώστε να εξαχτεί οτιδήποτε γεωγραφικός θησαυρός από το Geonames θελήσουμε. Ο τελικός θησαυρός ελέγχθηκε και πέρασε επιτυχώς από έναν SKOS validator, τα missing language tags είναι οι ετικέτες που αντιστοιχούν στο gn:name, το οποίο όπως είδαμε είναι αγγλικά αλλά δεν έχει κάποιο xml:lang και τα επτά loose concepts που αναφέρονται είναι οι επτά ήπειροι που θεωρητικά έχουν σαν ευρύτερο όρο την «ΓΗ» αλλά δεν αναφέρεται στο Geonames.

Εικόνα 5.1.3.1 Εικόνα του validator αφού γίνει ο έλεγχος στον γεωγραφικό θησαυρό

Αξίζει να υπενθυμίσουμε ότι ο θησαυρός μας, έμμεσα είναι συνδεδεμένος με το σύννεφο των Συνδεδεμένων δεδομένων. Επειδή τα URI του Geonames είναι έγκυρα κα σταθερά, ο νέος θησαυρός που φτιάξαμε είναι σαν ένας τροποποιημένος καθρέφτης και υποσύνολο του αρχικού. Μέσω ενός ειδικού φυλλομετριτή, για πλοήγηση στα Συνδεδεμένα δεδομένα, θα μπορούσαμε εύκολα να ξεκινήσουμε από τον τοπικό μας θησαυρό αναζητώντας ένα τοπωνύμιο, να μεταφερθούμε στο URI που ταυτίζεται με τον πόρο του π.χ. http://sws.Geonames.org/360630/about.rdf για το Κάιρο και από κει να μεταφερθούμε στο άρθρο της Wikipedia μέσω του

```
<gn:wikipediaArticle rdf:resource="http://en.wikipedia.org/wiki/Cairo"/>
```

ή στην dbpedia από το

```
<owl:sameAs rdf:resource="http://dbpedia.org/resource/Cairo"/>.
```

5.2 Μέθοδος χειρωνακτικής μετατροπής SKOS με URI του IPTC

Σήμερα, η ποιο συνηθισμένη μορφή ενός θησαυρού είναι αυτή της XML. Τα περισσότερα πολιτιστικά ιδρύματα, στην προσπάθεια να ψηφιοποιήσουν τις συλλογές τους, χρησιμοποίησαν την γλώσσα της XML γιατί επιτρέπει να ορίσουμε ένα αντικείμενο, μέσω της ετικέτας <Concept> , να του εισάγουμε επιπλέον πληροφορία αλλά και ένα σύνολο μεταδεδομένων. Στην αρχική μορφή του διαδικτύου, αυτός ο τύπος των δεδομένων ήταν αρκετός. Όπως όμως αναλύθηκε, η χρήση των θησαυρών

στον Σημασιολογικό Ιστό, απαιτεί να μετατραπούν σε RDF με βάση το λεξιλόγιο SKOS.

Σε αντίθεση με την προηγούμενη δοκιμή, τώρα έχουμε εξάγει, από μια βάση, έναν θησαυρό σε XML με στόχο να γίνει η μετατροπή του σε SKOS. Επιπλέον, υποθέτουμε ότι δεν έχουμε την δυνατότητα να παρέχουμε δικά μας dereferenceable URI στον Ιστό. Αρκετοί οργανισμοί, που διαθέτουν θησαυρούς σε τοπική XML μορφή, διστάζουν να κάνουν τις μετατροπές που χρειάζονται για την μετάβαση σε RDF. Ο λόγος είναι είτε έλλειψη τεχνογνωσίας, είτε κεφαλαίου που απαιτείτε για την δημιουργία και την σταθερή παροχή dereferenceable URIs. Όπως είχαμε αναλύσει, στον Σημασιολογικό Ιστό κάθε όρος πρέπει να ορίζετε από ένα μοναδικό αναγνωριστικό, που πρέπει να ακολουθεί τους κανόνες των «Cool» URIs. Για αυτό το λόγο, θα δούμε πως κάναμε χρήση μίας σχετικής βιβλιοθήκης, η οποία έχει ορίσει με URI, σχετικούς όρους με τους δικούς μας. Αναλύσαμε την μεθοδολογία σε απλά βήματα, για σχετικές περιπτώσεις:

1. Εξαγωγή του θησαυρού σε XML.
2. Ανάλυση του θησαυρού σε θεματολογίες. Πρέπει να χωρίσουμε τα <Concept>, με βάση το εννοιολογικό πεδίο που ανήκουν.
3. Ανάλυση των πεδίων κάθε <Concept> του θησαυρού. Βλέπουμε ποια στοιχεία μας ενδιαφέρουν να κρατήσουμε στην νέα μορφή του.
4. Δημιουργία χαρτογράφησης όρων, μεταξύ XML και SKOS μορφής.
5. Μετατροπή θησαυρού σε SKOS, χωρίς να υπάρχει ορισμός URI για τον πόρο, στο πεδίο <skos:Concept>
6. Εύρεση γνωστού και σχετικής θεματολογίας θησαυρού, που παρέχει τους όρους του με dereferenceable URI και κατά προτίμηση είναι συνδεδεμένος με τα Συνδεδεμένα Δεδομένα.
7. Χειροκίνητη αντιστοίχιση ταυτόσημων όρων μεταξύ του λεξιλογίου μας και του επιλεγμένου.
8. Τοποθέτηση εναπομεινάντων σε ξεχωριστή βάση, με στόχο την μελλοντική τους αντιστοίχιση .
9. Επανάληψη των βημάτων 6-8 για κάθε υποκατηγορία του θησαυρού μας.

5.2.1 Μετατροπή από XML σε SKOS.

Ο θησαυρός μας, που εξήχθηκε από την εφαρμογή ThesauriX, περιέχει διάφορες θεματολογίες. Όροι επιχειρηματικοί, τηλεοπτικοί, γεωγραφικοί και πνευματικών

δικαιωμάτων, συνθέτουν ένα ετερόκλιτο σε επίπεδο όρων σύνολο. Για την εφαρμογή μας, επιλέξαμε τον κλάδο που αφορά τηλεοπτικούς όρους με τίτλο «Θεματικό Σύστημα Αναφοράς». Ο κλάδος αυτός έχει πληροφορίες σχετικές με τους τύπους μέσων, τύπους αντικειμένων, τις ιδιότητες, τις θεματικές αναφορές και άλλα αναγνωριστικά για τηλεοπτικά προγράμματα.

Για την ανάλυση του περιεχομένου θα χρησιμοποιήσουμε την ιδιότητα των XML θησαυρών, όπου τα αντικείμενα ορίζονται από ένα <Concept> και τα πεδία μέσα σε αυτά ακολουθούν το ίδιο μοτίβο. Συγκεκριμένα στον δικό μας θησαυρό παρατηρούμε ότι για κάθε ετικέτα <Concept> υπάρχει η παρακάτω μορφή:

```
<Concept hierarchyId="5018" >
      <Translations>
       <Term termId="23014" lang="en" >
       <Name>Alternative Energy Issues</Name>
       <TermAttributes></TermAttributes>
       <URL>6001000</URL>
       <Proved>false</Proved>
       <GeneratedTranslation>false</GeneratedTranslation>
       </Term>
…
 </Translations>
<datingFrom>null</datingFrom>
<datingTill>null</datingTill>
<datingNotice>null</datingNotice>
</Concept>
```

Τοπικά κάθε όρος έχει το δικό του μοναδικό αναγνωριστικό που στη περίπτωση μας είναι το hierarchyId="5018" ακολουθούν για κάθε όρο μια σειρά από μεταφράσεις και τέλος λεπτομέρειες για την ημερομηνία δημιουργίας/τροποποίησης του όρου. Μέσα σε κάθε μετάφραση βλέπουμε ότι ακολουθείτε πάλι ένα μοτίβο. Ο δημιουργός του θησαυρού, θέλοντας να δώσει λεπτομέρειες για κάθε μετάφραση του όρου, αντιστοίχισε την μετάφραση στην ετικέτα <Term>. Όλοι οι όροι <Term> έχουν ένα μοναδικό αναγνωριστικό το termId="23014" και μια διευκρίνιση για το ποια γλώσσα περιέχεται. Στη συνέχεια, δίνουν το όνομα του όρου και κάποια άλλα στοιχεία για την μετάφραση αυτή, που έχουν τοπικό ενδιαφέρον μόνο.

Στην προσπάθεια μας να αντιστοιχίσουμε το λεξιλόγιο μεταξύ της XML και SKOS μορφής, βρεθήκαμε στο δίλλημα αν χρειαζόμαστε την πληροφορία που υπάρχει σε ένα <Term> ή αν μας αρκεί μόνο η μετάφραση του. Εφόσον η ενσωματωμένη πληροφορία κριθεί απαραίτητη, θα πρέπει να γίνει χρήση της επέκτασης του SKOS-XL αλλά και να δημιουργήσουμε δικό μας λεξιλόγιο για κάποιους όρους, που δεν υπάρχουν σε υπάρχον. Μέσω της SKOS-XL, έχουμε την δυνατότητα να ορίσουμε για κάθε μετάφραση που θα αφορά ένα <skos:Concept> ένα <skosxl:Label>. Έτσι κάθε μετάφραση θα έχει ένα μοναδικό αναγνωριστικό, με βάση το termId του, και θα ορίζεται ως πόρος. Το αποτέλεσμα για κάθε <Concept > θα είναι το εξής:

```
<skos:Concept rdf:about="http:/example.com/Concept/1816">
<skos:prefLabel>
        <skosxl:Label rdf:about="http:/example.com/Concept/1816#18806">
          <skosxl:literalForm xml:lang="hu">Zilah</skosxl:literalForm>
                        <my:TermAttributes></my:TermAttributes>
                        <my:Proved>false</my:Proved>
                        <my:GeneratedTranslation>false</my:GeneratedTranslation>
                        <skos:scopeNote></skos:scopeNote>
         </skosxl:Label>
<skos:prefLabel>
        <skosxl:Label rdf:about="http:/example.com/Concept/1816#18806">
          <skosxl:literalForm xml:lang="hu">Zilah</skosxl:literalForm>
                        <my:TermAttributes></my:TermAttributes>
                        <my:Proved>false</my:Proved>
                        <my:GeneratedTranslation>false</my:GeneratedTranslation>
                        <skos:scopeNote></skos:scopeNote>
         </skosxl:Label>
          <skos:broader rdf:resource="http:/example.com/Concept/1815"/>
             <my:datingFrom>null</my:datingFrom>
          <my:datingTill>null</my:datingTill>
             <my:datingNotice>null</my:datingNotice>
</skos:Concept>
```

Βλέπουμε ότι ανοίγουμε μια ετικέτα με το <skos:prefLabel> και μέσα αναλύουμε τον όρο με την πλήρη πληροφορία με το SKOS-XL. Όπου το λεξιλόγιο της SKOS ή

κάποιο άλλο δεν μας καλύπτει ορίζουμε δικό μας (my:Proved, my:GeneratedTranslation κτλ.), αρκεί φυσικά να ακολουθεί τους όρους δημιουργίας ενός ελεγχόμενου λεξιλογίου.

Στην περίπτωση που εξετάζουμε, αποφασίσαμε ότι μεγάλος μέρος της πληροφορίας, που έχει το αρχικό αρχείο, δεν είναι χρήσιμη για τον αναγνώστη του θησαυρού αλλά αφορά μόνο τον διαχειριστή του. Όμως, η αναφορά στην SKOS-XL μετατροπή, θεωρήσαμε ότι ήταν απαραίτητο να υπάρχει για να δείξουμε πως αυτή αναπτύσσεται και ποιο πρόβλημα μπορεί να λύσει.

Ο θησαυρός μας στην νέα του μορφή, θα χρειαζόταν μόνο τις μεταφράσεις του και τα xml attributes. Ειδικά τα δεύτερα, είναι σημαντικό να διατηρηθούν, αφού θα υπάρχει αντιστοιχία μεταξύ των παλιών όρων και των νέων, που θα δείχνουν σε εξωτερικά URI. Ένα άλλο ερώτημα στο οποίο έπρεπε να απαντήσουμε, κατά την διαδικασία της μετατροπής, ήταν τι ιεραρχία θα έχει το νέο SKOS αρχείο. Έπρεπε να κρατηθεί η ιεραρχία του τοπικού XML εγγράφου ή να ακολουθήσουμε την ιεραρχία που θα έχουν τα URI στον Ιστό? Η πρώτη περίπτωση είναι λάθος με την λογική του SKOS αφού το <skos:broader> δεν θα έχει κάποιο dereferenceable URI να δείξει αλλά θα δείχνει σε ένα τοπικό αρχείο με βάση το hierarchyId. Όμως και η δεύτερη δημιουργεί προβλήματα που θα τα συνετίσουμε στην συνέχεια.

Έτσι δημιουργήσαμε την ακόλουθη χαρτογράφηση :

Πίνακας 5-3 Χαρτογράφηση IPTC

XML	SKOS
Concept	skos:Concept
Term	skos:prefLabel
termId	xml:termId
lang	xml:lang
hierarchyId	xml:hierarchyId

Δημιουργήσαμε ένα σύντομο και απλό πρόγραμμα σε Ruby, που όπως και στο παράδειγμα των Geonames, εφαρμόζει την χαρτογράφηση και διαγράφει όσα στοιχεία δεν μας χρειάζονται. Επειδή, ο κώδικας είναι της ίδιας μορφής μ' αυτή του προηγούμενου παραδείγματος, δεν τον επανεισάγουμε και εδώ. Αποτέλεσμα είναι το Concept με hierarchyId =5018, που είδαμε να μετασχηματιστεί σε :

```
<skos:Concept rdf:about="" xml:hierarchyId="5018" >
```

```
<skos:prefLabel xml:termId="23014" xml:lang="en" >Alternative Energy Issues</skos:prefLabel>
<skos:prefLabel xml:termId="23015" xml:lang="nl" >Alternatieve energie-zaken</skos:prefLabel>
…
</skos:Concept>
```

5.2.2 Εύρεση στόχου και χειρωνακτική αντιστοίχιση όρων

Σε αυτό το στάδιο έχουμε ολόκληρο τον θησαυρό μας στην επιθυμητή μορφή SKOS. Όμως οι πόροι μας δεν είναι ορισμένοι με κάποιο URI, ούτε υπάρχει ιεραρχία στα δεδομένα μας. Αυτά τα δύο ζητήματα θα επιλυθούν αφού διαλέξουμε τον θησαυρό στόχο, με τον οποίο θα γίνει η σύνδεση.

Στον Ιστό υπάρχουν δημοσιευμένοι, από οργανισμούς και ιδρύματα, πολλοί θησαυροί γνώσης. Το επόμενο βήμα είναι να επιλέξουμε έναν σχετικό με τηλεοπτικούς όρους ώστε, να βρεθούν όσο το δυνατόν περισσότεροι όροι ίδιοι με τον δικό μας. Εδώ πρέπει να τονίσουμε ότι δεν επαρκεί οι δύο όροι να είναι απλά σχετικοί, αλλά πρέπει να είναι ταυτόσημοι εννοιολογικά. Από τον αρχικό θησαυρό μας, κρατήσαμε εκτός από την αρίθμηση του XML, την πολύγλωσση περιγραφή του όρου. Έτσι είναι πολύ σημαντικό η περιγραφή η δική μας για κάθε όρο να είναι ίδια με την περιγραφή που δίνεται στον στόχο.

Η μεγαλύτερη πηγή όρων για τηλεοπτική θεματολογία είναι το IPTC, το οποίο αναλύσαμε σε προηγούμενη ενότητα. Η κατηγορία Descriptive NewsCodes περιλαμβάνει όρους που χρησιμοποιούνται για την σωστή περιγραφή του περιεχομένου των ειδήσεων. Παρατηρήσαμε μεγάλη ομοιότητα των στοιχείων μας με την υποενότητα Subject Code. Το Subject Code περιέχει μια περιγραφή του περιεχομένου των Ειδήσεων σε υψηλό επίπεδο και έχει περίπου 1400 όρους.

Κάθε όρος του Subject Code είναι διαθέσιμος στην παρακάτω μορφή, που είναι αναγνωρίσιμη από άνθρωπο:

Concept Id: QCode = subj:01002000 URI = http://cv.iptc.org/newscodes/subjectcode/01002000	created: 2000-10-30T13:00:00+00:00	modified: 2000-10-30T01:00:00+00:00
Type (Qcode) = cpnat:abstract		
Name in en-GB is: architecture		
Definition in en-GB is: Designing of buildings, monuments and the spaces around them		
Broader concept:	http://cv.iptc.org/newscodes/subjectcode/01000000	

Αλλά και σε RDF:

```
<rdf:Description rdf:about="http://cv.iptc.org/newscodes/subjectcode/01002000">
<rdf:type rdf:resource="http://www.w3.org/TR/skos-reference/skos.html#Concept"/>
```

```
<skos:prefLabel xml:lang="en-GB">architecture</skos:prefLabel>
<skos:definition xml:lang="en-GB">Designing of buildings, monuments and the spaces around them</skos:definition>
<skos:broaderTransitive>
<rdf:Description rdf:about="http://cv.iptc.org/newscodes/subjectcode/01000000">
<rdf:type rdf:resource="http://www.w3.org/TR/skos-reference/skos.html#Concept"/>
</rdf:Description></skos:broaderTransitive></rdf:Description>
```

Βλέπουμε ότι κάθε όρος χαρακτηρίζεται από ένα μοναδικό URI, μια ετικέτα με το όνομα του, μια ετικέτα με την περιγραφή του και έναν όρο ευρύτερης έννοιας. Στην περίπτωση μας, επειδή θα δοκιμάσουμε την χειρωνακτική αντιστοίχιση όρων μας βολεύει η πρώτη μορφή που είδαμε.

Η διαδικασία είναι απλή αλλά και χρονοβόρα. Για κάθε όρο που έχουμε στον αρχικό μας θησαυρό προσπαθούμε να βρούμε έναν μέσα στο Subject Code, που είναι ο στόχος μας, και μετά αντιγράφουμε το URI του στόχου στο πεδίο που είχαμε αφήσει κενό <skos:Concept rdf:about="" xml:hierarchyId="5018" > .

Αν βρούμε ταυτόσημο όρο, τότε αυτός μεταφέρεται ολόκληρος στον νέο θησαυρό. Αν όμως, δεν βρεθεί αντιστοίχιση τότε τοποθετείται σε ένα άλλο αρχείο που το ονομάσαμε IPTCdump. Με αυτό τον τρόπο μπορούμε στο μέλλον, να επεξεργαστούμε τους υπολειπόμενους όρους.

Επόμενο βήμα είναι η δημιουργία ιεραρχίας. Στο νέο θησαυρό μας, θα πρέπει να κρατήσουμε την ιεραρχία που έχει ο στόχος και όχι η πηγή. Ο βασικός λόγος που δεν μπορούμε να κάνουμε το ανάποδο είναι γιατί δεν θα βρούμε αντιστοιχία σε όλους τους αρχικούς όρους. Έτσι, αν κρατάγαμε την ιεραρχία του XML, κατά την πλοήγηση μέσω του <skos:broader>, θα καταλήγαμε σε concept τα οποία δεν έχουν URI και θα είχαμε αρκετά blank nodes. Η τεχνική αυτή θα ήταν λανθασμένη και αντίθετη με τις αρχές των Συνδεδεμένων Δεδομένων. Έτσι, κρατώντας την ιεραρχία που υπάρχει στο IPTC, το επόμενο βήμα είναι να δημιουργήσουμε στον όρο μας το

```
<skos:broader rdf:resource="http://cv.iptc.org/newscodes/subjectcode/03000000"/>
```

Αν το URI στο οποίο αντιστοιχίσαμε τον όρο μας, δεν υπάρχει στον θησαυρό μας και επίσης, δεν βλέπουμε κάποιο σχετικό αντικείμενο που θα μπορούσε να είναι το <skos:broader>, τότε ο όρος αυτός θα μεταφερθεί στο IPTCdump. Ο λόγος που συμβαίνει αυτό είναι ότι στον νέο θησαυρό σε SKOS, δεν είναι σωστή τεχνική να

έχουμε πόρους χωρίς να ανήκουν σε κάποια ιεραρχία (loose Concepts). Τέλος προσθέτουμε το σχήμα που ανήκουν όλοι οι όροι μας :

```
<skos:inScheme
rdf:resource="http://www.iptc.org/site/NewsCodes/View_NewsCodes/"/>
```

Κάθε όρος θα έχει την παρακάτω μορφή

```
<skos:Concept rdf:about="http://cv.iptc.org/newscodes/subjectcode/10013000"
xml:hierarchyId="5333" >
<skos:inScheme
rdf:resource="http://www.iptc.org/site/NewsCodes/View_NewsCodes/"/>
<skos:prefLabel xml:termId="26141" xml:lang="en" >Fishing</skos:prefLabel>
<skos:prefLabel xml:termId="26142" xml:lang="nl" >Hengelsport</skos:prefLabel>
<skos:prefLabel xml:termId="26143" xml:lang="de" >Fischen</skos:prefLabel>
<skos:prefLabel xml:termId="26144" xml:lang="fr" >Pêche</skos:prefLabel>
<skos:prefLabel xml:termId="26145" xml:lang="da" >Lystfiskeri</skos:prefLabel>
<skos:prefLabel xml:termId="26146" xml:lang="it" >Fishing</skos:prefLabel>
<skos:prefLabel xml:termId="26147" xml:lang="el" >Ψάρεμα</skos:prefLabel>
<skos:prefLabel xml:termId="26148" xml:lang="hu" >horgászat</skos:prefLabel>
<skos:prefLabel xml:termId="26149" xml:lang="ca" >Pesca</skos:prefLabel>
<skos:prefLabel xml:termId="26150" xml:lang="sv" >Fiske</skos:prefLabel>
<skos:broader rdf:resource="http://cv.iptc.org/newscodes/subjectcode/10000000"/>
</skos:Concept>
```

Ο θησαυρός εγκρίθηκε από τον SKOS validator

Εικόνα 5.2.2.1 Εικόνα του validator αφού γίνει ο έλεγχος στον IPTC θησαυρό

5.2.3 Τελικό αποτέλεσμα

Είδαμε πως μπορούμε να δημιουργήσουμε έναν θησαυρό, βασιζόμενοι σε δικούς μας, τοπικούς όρους κάνοντας χειροκίνητη αντιστοίχιση με υπάρχοντες όρους στον Ιστό. Να βρούμε σε μια βάση όλους τους όρους είναι εξαιρετικά δύσκολο, οπότε συνιστάται η ξεχωριστή τοποθέτηση των εναπομεινάντων όρων σε ένα αρχείο (στην περίπτωση μας το IPTCdump). Σε επόμενη φάση θα μπορούσαμε να βρούμε άλλους στόχους ώστε να οριστούν και αυτοί οι όροι.

Κατά την μέθοδο αυτή, συναντήσαμε διλλήματα σχετικά με την ιεράρχηση, την διαθέσιμη πληροφορία αλλά και τον τρόπο αντιστοίχισης. Η χειροκίνητη μεθοδολογία έχει ως βασικό πλεονέκτημα ότι ο ανθρώπινος χρήστης έχει την κριτική ικανότητα να αντιλαμβάνεται, σε σύντομο χρόνο, πότε δύο όροι από διαφορετικές βιβλιοθήκες είναι ταυτόσημοι. Αντίθετα, ένα αυτόματο πρόγραμμα θα χρειαζόταν έναν σύνθετο αλγόριθμο για να πετύχει το ίδιο αποτέλεσμα. Ο λόγος είναι ότι απλοί αλγόριθμοι βασιζόμενοι στο ταίριασμα χαρακτήρων (string matching) θα είχαν πολλά σφάλματα αφού τα <Name> που είχε η πηγή σε σύγκριση με τα <skos:prefLabel> που είχε ο στόχος είχαν αρκετές διαφορές. Αντίθετα, με ένα πρόγραμμα η διαδικασία θα γινόταν πολύ πιο γρήγορα αφού η ανθρώπινη παρέμβαση σε 800 <Concept> απαιτεί αρκετό χρόνο.

Το αποτέλεσμα αυτής της μεθόδου είναι να έχουμε αντιστοίχιση σε 528 όρους, οι οποίοι χρησιμοποιούν τα URIs του IPTC, έχοντας τις ετικέτες ιεραρχίας και την πολυγλωσσικότητα του τοπικού XML αλλά και την ιεραρχία του IPTC σε SKOS.

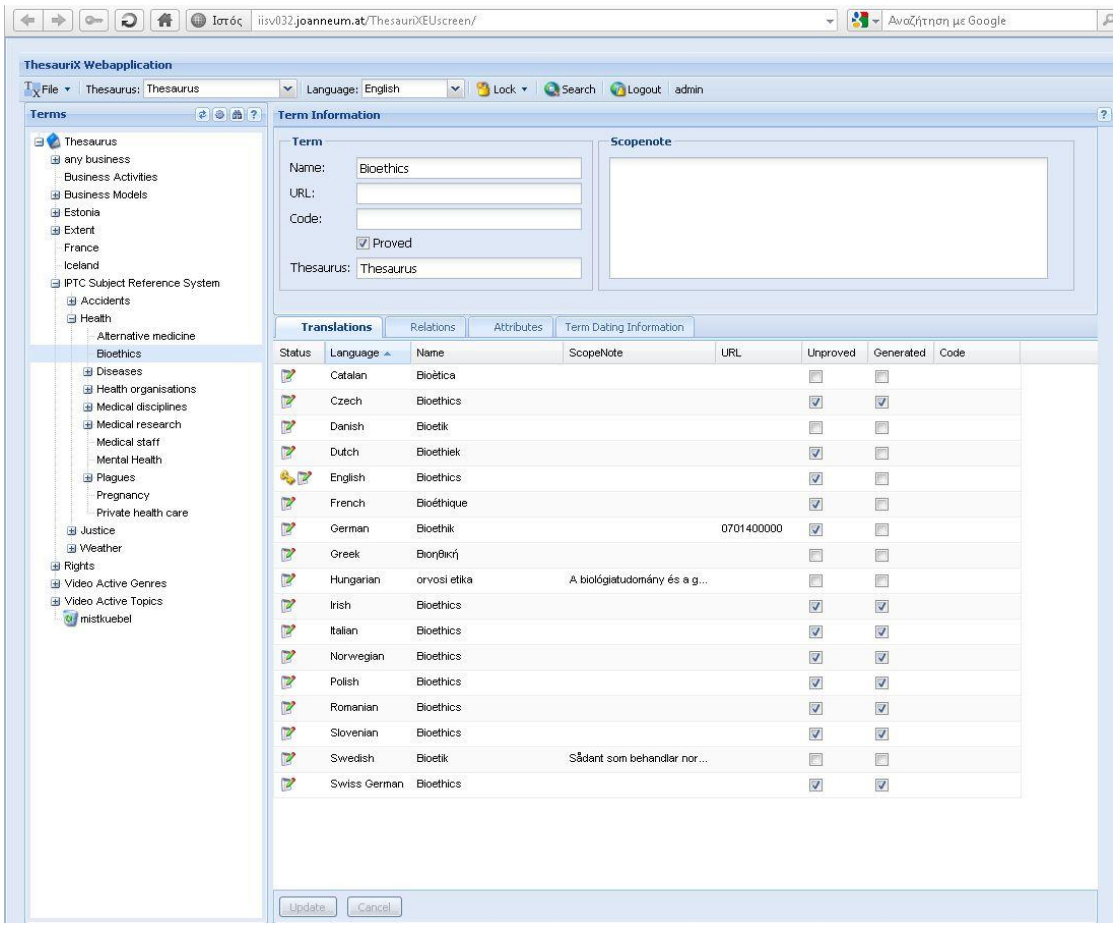
5.3 Μέθοδος αυτόματης μετατροπής σε SKOS, θησαυρών XML

Είδαμε ότι αρκετοί οργανισμοί έχουν τους θησαυρούς τους σε μορφή XML. Όταν το πλήθος αυτών είναι αρκετά μεγάλο, η χειροκίνητη μετατροπή δεν είναι ελκυστική λύση, αφού ο μεγάλος όγκος δεδομένων απαιτεί μια αυτοματοποιημένη μετατροπή αυτών. Οι αυτοματοποιημένες διαδικασίες είναι απαραίτητες και σε περιπτώσεις που δεν βρίσκουμε βιβλιοθήκη στον Ιστό με ταυτόσημους όρους ή απλά επιθυμούμε να παρέχουμε τα δικά μας αρχεία στον Ιστό. Μεθοδολογία ώστε να μετατρέπονται αυτόματα όλοι οι θησαυροί ανεξαρτήτως μορφής σε SKOS, δεν είναι εφικτό να γίνει γιατί δεν ακολουθούν την ίδια δομή. Αφού ορίσουμε τους επιθυμητούς στόχους, για την πληροφορία που θα περιέχεται στην SKOS μορφή, μένει να δημιουργήσουμε ένα

πρόγραμμα που θα μετατρέπει την πηγή σε ένα SKOS αρχείο με dereferenceable URI και ιεραρχία. Η εφαρμογή μας θα δημιουργηθεί με βάση την μορφή που έχουν οι εξαγόμενοι θησαυροί της εφαρμογής ThesauriX.

5.3.1 Ανάπτυξη εφαρμογής

Πριν τη δημιουργία της εφαρμογή μας, είναι εξαιρετικά σημαντικό να κατανοήσουμε την δομή της πληροφορίας στην αρχική της μορφής. Ο θησαυρός που επιλέχθηκε είναι γεωγραφικοί και έπειτα τηλεοπτικοί όροι που εξήχθηκαν από το ThesauriX. Είναι ένα εργαλείο που χρησιμοποιείται για την εισαγωγή στοιχείων, για την παρουσίαση και την διαχείριση των θησαυρών.



Εικόνα 5.3.1.1 ThesauriX

Όπως βλέπουμε οι θησαυροί αυτοί είναι πολύγλωσσοι και αποτελούνται από διάφορες θεματικές ενότητες, όπως και ότι είναι διαθέσιμο για κάθε στοιχείο, ένα σύνολο μεταδεδομένων που μπορούμε να έχουμε.

Η δομή είναι όμως η ίδια με αυτήν που αναλύθηκε στην προηγούμενη ενότητα (5.2.1) έτσι κάθε όρος <Concept> έχει την μορφή

```
<Concept hierarchyId="11881" >
        <Translations>
         <Term termId="63115" lang="en" >
          <Name>Valcea County</Name>
          <TermAttributes></TermAttributes>
          <Proved>false</Proved>
          <GeneratedTranslation>false</GeneratedTranslation>
         </Term>
</Translations>
<datingFrom>null</datingFrom>
<datingTill>null</datingTill>
<datingNotice>null</datingNotice>
       </Concept>
```

Συνοπτικά ξανά-αναφέρουμε ότι οι ετικέτες hierarchyId="" χρησιμεύουν στην ιεράρχηση των όρων μέσα στο XML αλλά ταυτόχρονα παρέχουν και ένα μοναδικό αναγνωριστικό σε κάθε <Concept>. Αντίστοιχα, οι ετικέτες termId="" παρέχουν μοναδικό αναγνωριστικό σε μια μετάφραση του όρου. Την μοναδικότητα αυτή που υπάρχει σε τοπικό επίπεδο, θα την εκμεταλλευτούμε για την δημιουργία δικών μας dereferenceable URI.

Για τους λόγους που αναλύθηκαν στην ενότητα στην ενότητα 5.2.1, οι ετικέτες <datingFrom>,<datingTill>,<datingNotice> αλλά και αυτές που είναι μέσα σε ένα <Term> μετάφρασης, εκτός του <Name>, μας είναι αδιάφορες στην νέα SKOS μορφή. Έτσι η χαρτογράφηση μας θα είναι ξανά

XML	SKOS
Concept	skos:Concept
Term	skos:prefLabel
termId	xml:termId
lang	xml:lang
hierarchyId	xml:hierarchyId

Εικόνα 5.3.1.2 Χαρτογράφηση Geography

Οι δύο βασικές απαιτήσεις που είχαμε από το πρόγραμμα είναι:

- Δημιουργία μοναδικών αναγνωριστικών URI της μορφής http://www.image.ntua.gr/geography/hierarchyId όπου θα γίνετε χρήση του μοναδικού αριθμού που υπάρχει ήδη για κάθε όρο.
- Διατήρηση της ιεραρχίας του XML θησαυρού. Στην XML μας, δυστυχώς δεν υπάρχει έτοιμο σε κάποιο πεδίο το hierarchyId του ευρύτερου αντικειμένου. Αντίθετα θα πρέπει να δημιουργήσουμε μια διαδικασία που για κάθε όρο να βρίσκει τον ευρύτερο του. Με λίγα λόγια, πρέπει να μετατρέψουμε την δενδρική ιεραρχία σε ένα επίπεδο SKOS αρχείο.

Για την δημιουργία του προγράμματος μετατροπής χρησιμοποιήθηκε η γλώσσα Ruby. Για την επεξεργασία του XML αρχείου, χρειαστήκαμε την βιβλιοθήκη της, REXML, η οποία είναι μια ολοκληρωμένη εφαρμογή του XPath.

Αρχικά εισάγουμε τις βιβλιοθήκες που θα κάνουμε χρήση και το αρχείο μας

```
require 'rexml/document'
include REXML
file = File.new("GEO_v1.xml")
doc = Document.new(file)
```

Στη συνέχεια, ορίζουμε μια δέσμη ενεργειών (function), δηλαδή ένα σετ εντολών που θα καλούνται αργότερα για επανεκτέλεση με πιο απλό τρόπο, χρησιμοποιώντας μόνο μια εντολή. Η get_path έχει σαν εισόδους ένα αρχείο σε xml και ένα μοναδικό αριθμό και δίνει σαν αποτέλεσμα τον πατέρα του τρέχοντος κόμβου.

```
def get_path(xml_doc, termId)
  node = REXML::XPath.first( xml_doc, termId )
    path = ['Concept']
  node.parent
 end
```

Έπειτα, για κάθε ετικέτα <Concept> λαμβάνουμε την τιμή που έχει το χαρακτηριστικό του hierarchyId και εισάγουμε ένα νέο χαρακτηριστικό (attribute) με όνομα rdf:about και τιμή ένα δημιουργημένο μοναδικό URI, βασισμένο στην μορφή που είδαμε στην πρώτη απαίτηση. Επίσης, εισάγουμε στην get_path το αρχείο μας και την τιμή hierarchyId για το τρέχον <Concept>. Αυτή επιστρέφει το <Concept> που είναι ο πατέρας του τρέχοντος κόμβου και μέσω της μεταβλητής father λαμβάνουμε την τιμή του hierarchyId για τον πατέρα. Τώρα είμαστε σε θέση να

προσθέσουμε ένα νέο στοιχείο με όνομα <skos:broader> και χαρακτηριστικό το rdf:resource που θα έχει τιμή:

```
http://www.image.ntua.gr/geography/+hierarchyIdFather,
```

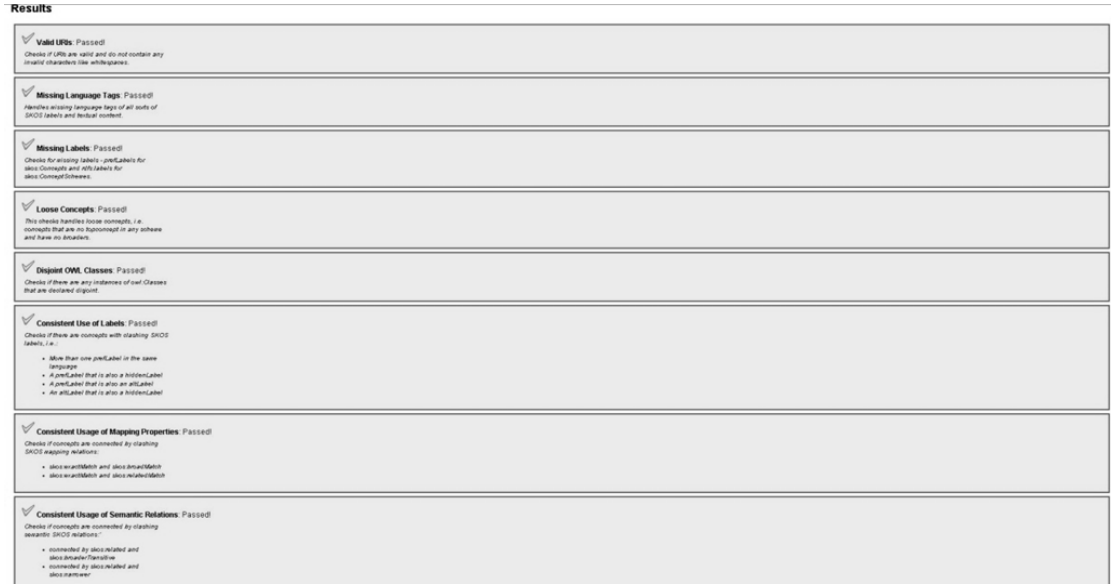
το οποίο το τοποθετούμε πριν από την ετικέτα < Translations>

```
doc.elements.each('//Concept') { |el|
id = el.attributes.get_attribute("hierarchyId").value
el.add_attribute("rdf:about", "http://www.image.ntua.gr/geography/"+ id.to_s() +"")
  path = get_path( doc, "//Concept[@hierarchyId='" + id.to_s() +"']")
father = path.attributes.get_attribute("hierarchyId").value
broad = el.add_element("skos:broader")
el.insert_before( "Translations", broad )
broad.add_attribute("rdf:resource", "http://www.image.ntua.gr/geography/"+
father.to_s() +"")
print "o " + id.to_s() + "exi patera ton " + father.to_s()}
File.open("res5.skos", "wb") {|file| file.puts doc}
```

Η διαδικασία αυτή όπως βλέπουμε θα γίνει για όλα τα <Concept> του αρχείου μας και το τελικό αποτέλεσμα κάθε φορά θα αποθηκεύεται σε ένα ενδιάμεσο αρχείο. Σε αυτό το αρχείο, εκμεταλλευόμενοι πάλι τις ομοιότητες μεταξύ XML και SKOS, με απλές εντολές εύρεσης και αντικατάστασης θα εφαρμόσουμε την χαρτογράφηση και θα σβήσουμε την πληροφορία που δεν μας χρειάζεται (παρουσιάσαμε παρόμοιο κώδικα στο 5.1.2). Τέλος μετά την διαδικασία αυτή, σβήσαμε τις κενές γραμμές του αρχείου ώστε να υπάρχει μια όμορφη εμφάνιση και εξάγαμε το τελικό αρχείο μας σε SKOS. Οι όροι μας είναι στην μορφή:

```
<skos:Concept rdf:about='http://www.image.ntua.gr/geography/11032'
xml:hierarchyId='11032'>
<skos:broader rdf:resource='http://www.image.ntua.gr/geography/11029'/>
<skos:inScheme rdf:resource="http://www.image.ntua.gr/geography/"/>
<skos:prefLabel xml:lang='en' xml:termId='52210'>La Rochelle</skos:prefLabel>
<skos:prefLabel xml:lang='nl' xml:termId='52213'>La Rochelle</skos:prefLabel>
…
</skos:Concept>
```

Και τέλος ο θησαυρός εγκρίθηκε από τον SKOS validator



Εικόνα 5.3.1.3 Εικόνα του validator αφού γίνει ο έλεγχος στον ολόκληρο Geography

Επόμενο στάδιο ήταν η εφαρμογή του κώδικα που αναπτύξαμε, σε θησαυρό που εξάχθηκε από το ίδιο εργαλείο. Για την δεύτερη δοκιμή πήραμε τον θησαυρό με τίτλο «Θεματικό Σύστημα Αναφοράς» που είδαμε και στο 5.2. Με μοναδική αλλαγή στον κώδικα το URI να έχει ως ρίζα το http://www.image.ntua.gr/iptc/ , φτιάξαμε ένα νέο SKOS με νέα μορφή του skos:Concept σε:

```
<skos:Concept rdf:about='http://www.image.ntua.gr/iptc/5430'
xml:hierarchyId='5430'>
<skos:broader rdf:resource='http://www.image.ntua.gr/iptc/5427'/>
<skos:inScheme rdf:resource="http://www.image.ntua.gr/iptc/"/>
<skos:prefLabel xml:lang='en' xml:termId='27109'>Veterans
Affairs</skos:prefLabel>
<skos:prefLabel xml:lang='nl'
xml:termId='27110'>Oorlogsveteranen</skos:prefLabel>
…
</skos:Concept>
```

Και η δεύτερη δοκιμή ήταν επιτυχής και εγκρίθηκε από τον Validator.

5.3.2 Τελικό αποτέλεσμα

Στην εφαρμογή αυτή είδαμε πως αναπτύσσουμε μια εφαρμογή, ώστε να επιτρέπει στο χρήστη την μετατροπή σε SKOS, ενός συνόλου θησαυρών, που έχουν ως χαρακτηριστικό την κοινή τους δομή. Αν το σύνολο αυτό είναι διαθέσιμο μέσω μιας

πύλης (π.χ. όπως είναι στην περίπτωση μας το ThesauriX) τότε με ενσωμάτωση του εργαλείου σ' αυτήν, θα μπορούσε να μετατρέπεται αυτόματα και να εξάγεται ο θησαυρός στη μορφή SKOS, έπειτα από εντολή του χρήστη. Να σημειώσουμε ότι τα dereferenceable URI που δημιουργήθηκαν στο πλαίσιο της εφαρμογής μας, δεν δημοσιεύτηκαν στον Ιστό, αφού η διαδικασία αυτή απαιτεί ειδικό εξοπλισμό και η υλοποίηση δεν ήταν στόχος της διπλωματικής αυτής.

Η πλήρης μετατροπή των θησαυρών μας, συνοδευόμενη με την δημοσίευση μοναδικών URI για κάθε όρο, μας δίνει αρκετή ανεξαρτησία. Δεν εξαρτόμαστε από εξωτερικές πηγές δεδομένων και έτσι μπορούμε να δημιουργήσουμε νέους όρους, τους οποίους, με τις διάφορες ετικέτες του SKOS, θα τις συνδέσουμε στον χώρο των Συνδεδεμένων Δεδομένων.

5.4 Συμπεράσματα

Στο κεφάλαιο αυτό, αντιμετωπίσαμε τρία σενάρια και είδαμε ότι η σωστή μετατροπή ενός θησαυρού με βάση το λεξιλόγιο του SKOS δεν είναι απλή διαδικασία και κάθε περίπτωση είναι διαφορετική.

Αν υπάρχει δυνατότητα από τεχνικής άποψης, προτείνεται η ανάπτυξη μιας αυτόματης εφαρμογής μετατροπής, όπως στην τρίτη περίπτωση. Ειδικά όταν διαθέτουμε μια βιβλιοθήκη με πολλούς θησαυρούς, η δομή των οποίων είναι κοινή, τότε η διαδικασία αυτή θα μας διευκολύνει πολύ, να εισέλθουμε στον χώρο του Σημασιολογικού Ιστού.

Μια αντιμετώπιση τέτοια δεν είναι εφικτή ούτε απαραίτητη πάντα. Για παράδειγμα, αν μας ενδιαφέρει να μετατρέψουμε ένα μικρό σύνολο όρων, χωρίς να μπούμε στην διαδικασία δημοσίευσης και δημιουργίας dereferenceable URIs, τότε ο δεύτερος τρόπος είναι ποιο εύκολος αν και χρονοβόρος. Τέλος, σε περίπτωση που δεν διαθέτουμε κάποιον θησαυρό ή απλά θέλουμε να εμπλουτίσουμε τα υπάρχοντα δεδομένα μας, ο δανεισμός όρων από δημοσιευμένο και ήδη συνδεδεμένο θησαυρό είναι η καλύτερη λύση.

Βλέπουμε δηλαδή ότι ο χειριστής που είναι υπεύθυνος για την μετατροπή, θα πρέπει να έχει πλήρη επίγνωση των απαιτήσεων, των αναγκών αλλά και των διαθέσιμων εργαλείων που έχει, ώστε να επιλέξει αντίστοιχη μεθοδολογία.

6

Δημιουργία συνδέσεων θησαυρών SKOS με τα Συνδεδεμένα Δεδομένα με χρήση του εργαλείου Amalgame

Στον Ιστό υπάρχει ένα σύνολο από θησαυρούς και βιβλιοθήκες με όρους. Αν οι όροι αυτοί είναι δημοσιευμένοι σε κάποιο RDF ή SKOS αρχείο, θα έχουν ένα μοναδικό αναγνωριστικό URI. Όμως, είναι συχνό φαινόμενο διαφορετικοί θησαυροί να περιέχουν τον ίδιο όρο. Για παράδειγμα σε έναν γεωγραφικό θησαυρό ο όρος με το όνομα «ΑΘΗΝΑ» δηλώνει την πρωτεύουσα της «ΕΛΛΆΔΑΣ», ενώ σε έναν ταξιδιωτικό, ο όρος «ΑΘΗΝΑ» δηλώνει την περιοχή που βρίσκεται η «ΑΚΡΟΠΟΛΗ». Ένας ανθρώπινος χρήστης αντιλαμβάνεται ότι οι δύο αυτοί όροι, που προέρχονται από άλλες πηγές, είναι ταυτόσημοι. Ο υπολογιστής όμως, δεν μπορεί να καταλήξει, από μόνος του, στο ίδιο συμπέρασμα. Για να αντιμετωπιστεί αυτό το πρόβλημα, μέσω της RDF και της SKOS, μπορούμε να δηλώσουμε διάφορες σχέσεις σύνδεσης μεταξύ των δύο όρων.

Η διαδικασία αυτή ονομάζετε αντιστοίχιση όρων (Term alignment) και είναι το θεμέλιο για τη δημιουργία των Συνδεδεμένων Δεδομένων. Μέσω αυτής διαφορετικοί θησαυροί, μπορούν να δημιουργήσουν δεσμούς μεταξύ των όρων τους, οι οποίοι με την σειρά τους, δημιουργούν ένα μικρό η μεγάλο δίκτυο από Συνδεδεμένα Δεδομένα.

Αυτή τη στιγμή δεν υπάρχει κάποιο διαδεδομένο εργαλείο που να χρησιμοποιείτε για την δημιουργία συνδέσμων από διαφορετικά σύνολα. Όπως είδαμε και παραπάνω

κάθε θησαυρός μπορεί να έχει διαφορετική δομή και αυτό δυσκολεύει στην ανάπτυξη μιας ενιαίας αυτόματης μεθοδολογίας. Στο πλαίσιο της διπλωματική, αποφασίστηκε να χρησιμοποιήσουμε το εργαλείο αντιστοίχισης όρων Amalgame.

Το Amalgame μας επιτρέπει να εισάγουμε διάφορους θησαυρούς και στη συνέχεια, να βρούμε τους κοινούς όρους και να δημιουργήσουμε τις συνδέσεις, ανάλογα με την εννοιολογική σχέση που έχουν. Ένα χαρακτηριστικό του Amalgame είναι ότι δεν εισάγει νέα πληροφορία στα δύο υποψήφια λεξιλόγια αλλά κατασκευάζει ένα νέο αρχείο που περιέχει τους όρους που αντιστοιχήθηκαν και την σχέση που έχουν. Ακόμα δίνει την δυνατότητα στον χρήστη να διαλέξει διάφορες τεχνικές αντιστοίχισης όρων. Λεπτομέρειες για την διαδικασία αντιστοίχισης θα δούμε στις παρακάτω περιπτώσεις χρήσης του.

Στον πολιτιστικό τομέα, οι όροι ορίζονται από ετικέτες που εμπλουτίζονται με ένα περιεχόμενο. Αυτό το περιεχόμενο αποτελείται από:

- Από την ετικέτα του όρου και την γλώσσα που την περιγράφει.
- Το όνομα της ιδιότητας που χρησιμοποιεί ο όρος μας.
- Την θέση του πόρου μέσα στον θησαυρό .

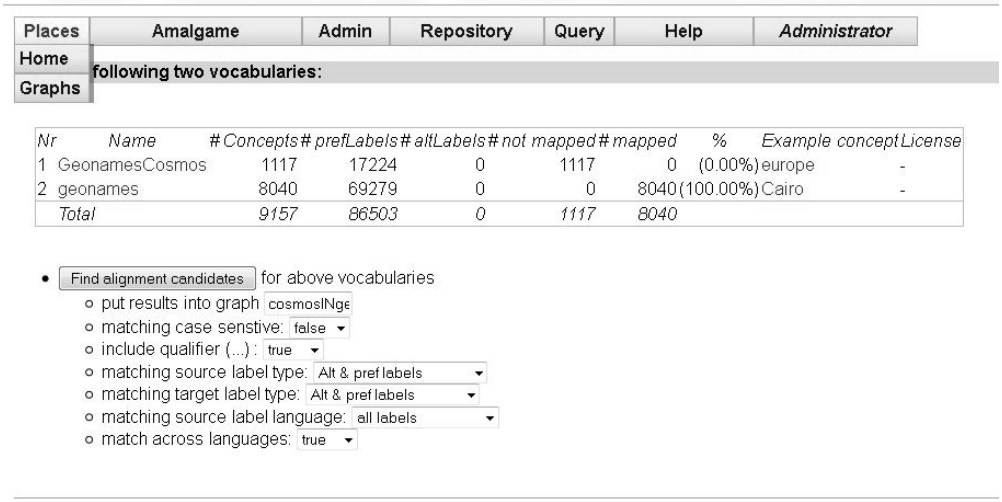
Επιπλέον για την αντιστοίχιση πρέπει να βρούμε το λεξιλόγιο-στόχος που θα εξερευνήσουμε για πιθανές συνδέσεις. Για παράδειγμα αν έχουμε ετικέτες που περιγράφουν μέρη είναι πιο πιθανόν να έχουμε μεγάλα ποσοστά αντιστοίχισης σε ένα γεωγραφικό λεξιλόγιο, όπως το TGN ή το Geonames. Η επιτυχία που θα έχει μια αντιστοίχιση εξαρτάται από το συγκεκριμένο λεξιλόγιο και την συλλογή όρων που θα χρησιμοποιηθεί.

Στην περίπτωση μας, εισάγαμε, σε ένα τοπικό εξυπηρετητή του Amalgame, τους θησαυρούς που μετατράπηκαν προηγουμένως σε SKOS αλλά και ένα σύνολο από άλλους θησαυρούς που θα ήταν οι στόχοι της αντιστοίχισης μας. Αρχικά, θα εφαρμόσουμε τη σύνδεση στους δύο γεωγραφικούς μας θησαυρούς, μετά στις δύο μορφές του «Θεματικού Συστήματος Αναφοράς» που αναπτύξαμε και τέλος τους θησαυρούς μας με εξωτερικά σύνολα όρων.

6.1 Γεωγραφικά

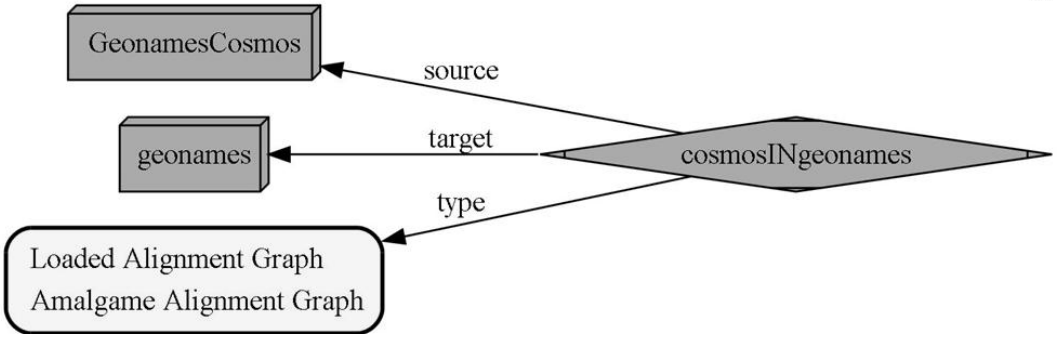
Στην 5.1 φτιάξαμε έναν γεωγραφικό θησαυρό (με όνομα Geonames) σε SKOS με βάση τα δεδομένα που εξορύξαμε από την βάση του Geonames ενώ στην 5.3 πήραμε έναν τοπικό γεωγραφικό θησαυρό και τον μετατρέψαμε σε SKOS((με όνομα GeonamesCosmos). Όπως είδαμε, ο θησαυρός του Geonames είναι συνδεδεμένος με

τα Συνδεδεμένα Ανοιχτά Δεδομένα και γι' αυτό θέλουμε να βρούμε τους αντίστοιχους όρους μεταξύ των δύο θησαυρών μας.



Εικόνα 5.3.2.1 Εικόνα από το Amalgame για επιλογή φίλτρων αντιστοίχησης

Στην παραπάνω εικόνα βλέπουμε το εργαλείο σύγκρισης του Amalgame. Το πρώτο λεξιλόγιο - πηγή έχει σύνολο 1117 όρους με 17224 γλωσσικές ετικέτες σύνολο, ενώ ο στόχος έχει 8040 όρους με 69279 γλωσσικές ετικέτες. Επίσης βλέπουμε ότι ο στόχος είναι ολόκληρος χαρτογραφημένος με εξωτερικές πηγές. Το αποτέλεσμα της αντιστοίχισης θα αποθηκευτεί σε ένα νέο αρχείο με το όνομα cosmosINGeonames. Στην συγκεκριμένη περίπτωση θα χρησιμοποιήσουμε την χαρτογράφηση με βάση όλες τις γλωσσικές ετικέτες της πηγής. Η δυνατότητα αυτή θα αυξήσει την εγκυρότητα και την αποτελεσματικότητα της αντιστοίχισης μας.



Εικόνα 5.3.2.2 Γράφος του cosmosINGeonames

Το τελικό αρχείο cosmosINGeonames έχει 529 <Concepts> της πηγής αντιστοιχισμένα στο Geonames, που είναι το 47,36% του θησαυρού GeonamesCosmos. Δημιουργήθηκαν 43249 τριάδες RDF που σχημάτισαν συνολικά 7208 συνδέσεις. Το αρχείο μας περιλαμβάνει όσες περιοχές βρέθηκαν και στους δύο

θησαυρούς και είχαν τουλάχιστον μια γλωσσική ετικέτα κοινή. Ακόμα, οι 7208 συνδέσεις που αναφέραμε οφείλονται στην αντιστοίχιση που έγινε μεταξύ κάθε κοινής γλωσσικής ετικέτας. Για να γίνει κατανοητό αυτό, αν επιλέξουμε να κάνουμε αντιστοίχιση μεταξύ μόνο των όρων της αγγλικής γλώσσας, χωρίς να βρει κοινή πληροφορία σε άλλες γλωσσικές ετικέτες, τότε το αποτέλεσμα θα ήταν 473 κοινά <Concepts> με 969 συνδέσεις. Το γεγονός αυτό, επιβεβαιώνει αυτό που αναφέραμε προηγουμένως ότι αν έχουμε πολύγλωσσους θησαυρούς αυξάνεται η αποτελεσματικότητα της αντιστοίχισης μας.

Το νέο αρχείο έχει στοιχεία της μορφής

```
<align:map>
   <align:Cell>
     <align:entity1 rdf:resource="http://www.image.ntua.gr/geography/10784"/>
     <align:entity2 rdf:resource="http://sws.Geonames.org/2800866/"/>
     <amalgame:evidence>
       <rdf:Description
          amalgame:method="exact_label"/>
     </amalgame:evidence>
   </align:Cell>
 </align:map>
<align:map>
   <align:Cell>
     <align:entity1 rdf:resource="http://www.image.ntua.gr/geography/11049"/>
     <align:entity2 rdf:resource="http://sws.Geonames.org/2973385/"/>
     <amalgame:evidence>
       <rdf:Description
          amalgame:method="exact_label"/>
     </amalgame:evidence>
   </align:Cell>
 </align:map>
```

Στην αρχή του εγγράφου ορίζεται το σχήμα του amalgame

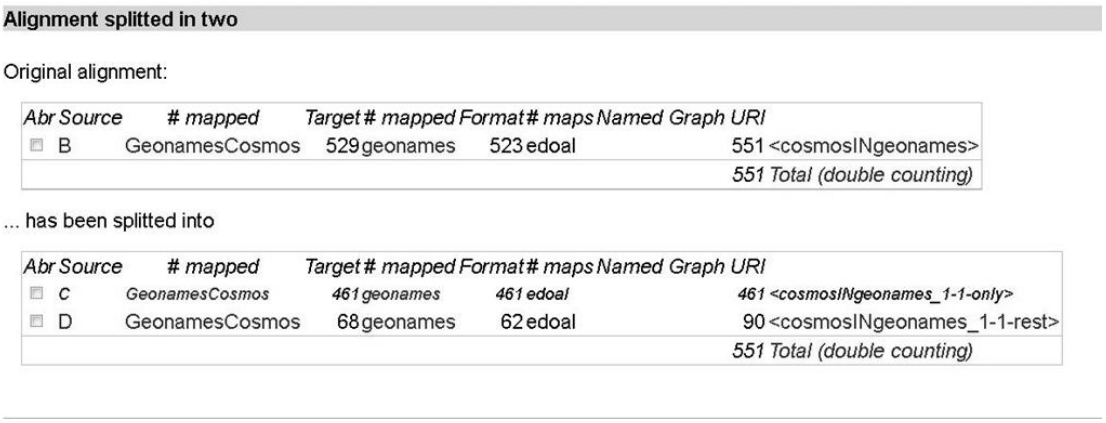
```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE rdf:RDF [
```

```
   <!ENTITY align
'http://knowledgeweb.semanticweb.org/heterogeneity/alignment#'>
   <!ENTITY amalgame 'http://purl.org/vocabularies/amalgame#'>
   <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>]>
<rdf:RDF
   xmlns:align="&align;"
   xmlns:amalgame="&amalgame;"
   xmlns:rdf="&rdf;">
```

το παραπάνω παράδειγμα, μας δείχνει ότι ο πόρος της πηγής με αναγνωριστικό http://www.image.ntua.gr/geography/10784 συνδέεται με τον πόρο στον στόχο http://sws.Geonames.org/2800866/ μέσω της ετικέτας exact_label, που μας ενημερώνει ότι πρόκειται για ταυτόσημους όρους.

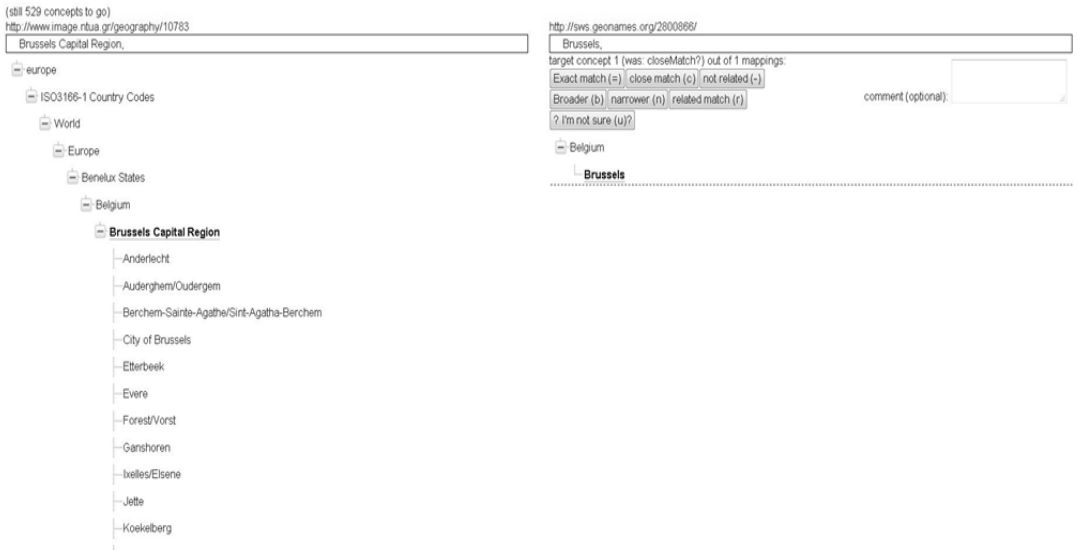
Επόμενο βήμα είναι η ανάλυση των αποτελεσμάτων μας. Θα αναλύσουμε τον θησαυρό – πηγή σε concepts που βρήκαν σύνδεση και σε αυτά που δεν είχαν. Το GeonamesCosmos (unmapped), μας βοηθά να κατανοήσουμε γιατί βρέθηκε σύνδεση μόνο στο 47,36%. Ο θησαυρός στόχος με αυτόν της πηγής, αν και γεωγραφικοί, έχουν διαφορετική φιλοσοφία για την περιεχόμενη πληροφορία. Ο GeonamesCosmos περιέχει πολλά ονόματα περιοχών, νησιών κτλπ, σε αντίθεση με τον στόχο που έχει τα ονόματα των σημαντικότερων πόλεων.

Μετά χωρίζουμε το cosmosINGeonames σε δύο σύνολα αντιστοίχισης. Στο πρώτο περιλαμβάνονται όσα <concept> της πηγής έχουν μόνο ένα στόχο και στο δεύτερο όσα έχουν πολλαπλούς. Στην εικόνα που ακολουθεί φαίνεται ότι 461 <concept> ανήκουν στην κατηγορία 1-1 και 68 στην κατηγορία 1-n.



Εικόνα 5.3.2.3 Χωρισμός του cosmosINGeonames στα δύο σύνολα.

Για να ελέγξουμε ποιο αναλυτικά την εγκυρότητα των αποτελεσμάτων, θα πάρουμε ένα τυχαίο δείγμα 25 συνδέσεων.



Εικόνα 5.3.2.4 Εργαλείο ελέγχου των αποτελεσμάτων

Στα αριστερά, έχουμε τον όρο της πηγής και στα δεξιά τους όρους του στόχου. Μέσω της εφαρμογής αυτής ελέγχουμε τη σχέση που έχουν αυτοί οι δύο όροι μεταξύ τους. Από τις 25 συνδέσεις που αξιολογήσαμε οι 23 ήταν ταυτόσημες (exact match) , 1 ήταν κοντινή (close match) και 1 ήταν μη σχετιζόμενη καθόλου. Το σφάλμα που βρήκαμε, οφειλόταν στο Amalgame, που χρησιμοποιεί απλό κώδικα για την συσχέτιση λέξεων με βάση την ρίζα της λέξης. Έτσι, την πόλη στην Γαλλία με όνομα Valence την ταύτισε με την πόλη της Ισπανίας Valencia. Σε προγράμματα αυτόματης αντιστοίχισης, που μας νοιάζει να έχουμε γρήγορη και απλή σύγκριση περιεχομένου, άρα και απλό αλγόριθμο, τέτοια σφάλματα είναι λογικό να υπάρχουν και ο μοναδικός τρόπος εξάλειψη τους είναι η ανθρώπινη παρέμβαση.

Τέλος, όσων αφόρα το αρχείο cosmosINGeonames, θα εξάγουμε από το Amalgame , ένα γράφο σε skos:exactMatch αφού σύμφωνα με το δείγμα μας είναι η αντιπροσωπευτική σχέση μεταξύ των συνδέσεων μας. έτσι στη νέα του μορφή θα έχει στοιχεία:

```
<rdf:Description rdf:about="http://www.image.ntua.gr/geography/12281">
  <skos:exactMatch rdf:resource="http://sws.Geonames.org/1512440/"/>
</rdf:Description>
<rdf:Description rdf:about="http://www.image.ntua.gr/geography/12278">
  <skos:exactMatch rdf:resource="http://sws.Geonames.org/1218197/"/>
</rdf:Description>
```

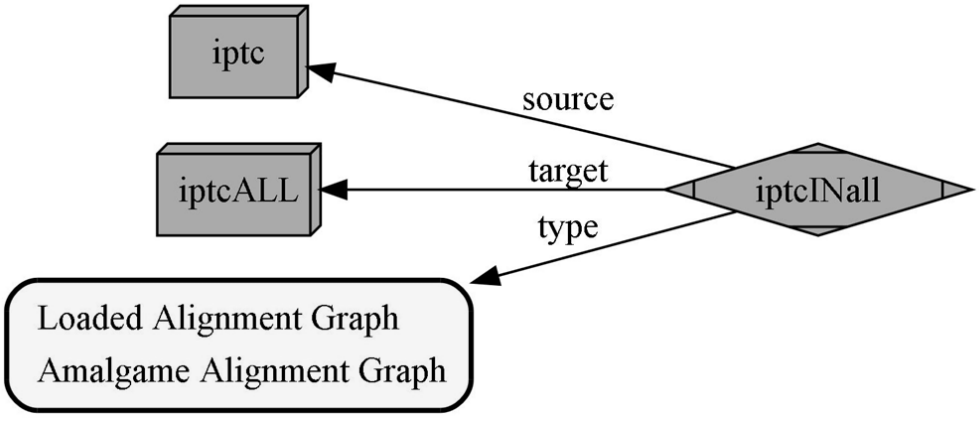
Στο συγκεκριμένο παράδειγμα χρήσης, οι όροι μας ήταν γεωγραφικά ονόματα, οπότε η σύνδεση των περισσότερων, από αυτούς που βρέθηκαν και στους δύο θησαυρούς, με το skos:exactMatch ήταν η αναμενόμενη.

6.2 Σύνδεση τηλεοπτικών θησαυρών

Στο 5^{O} κεφάλαιο σχηματίσαμε τον τηλεοπτικό θησαυρό του «Θεματικού Συστήματος Αναφοράς» . Στο πλαίσιο της μετατροπής του σε SKOS, αντιστοιχίσαμε τους όρους με τους υπάρχοντες στο IPTC, τους ορίσαμε με βάση το URI του IPTC κρατώντας όμως στοιχεία τοπικά, όπως τα χαρακτηριστικά της XML και τις ετικέτες γλώσσας. Εφαρμόσαμε, σε εκείνο το σημείο μια μορφή χειρωνακτικής αντιστοίχισης, αφού διαχωρίσαμε τα στοιχεία για τα οποία βρήκαμε ταυτόσημο όρο με αυτά που δεν είχαν αντιστοίχιση. Στη συνέχεια μετατρέψαμε ξανά ολόκληρο τον θησαυρό σε SKOS κάνοντας χρήση δικών μας Dereferenceable URI.

Όπως αντιλαμβανόμαστε, το πρώτο αρχείο είναι υποσύνολο του δεύτερου. Χρησιμοποιούν ίδιες γλωσσικές ετικέτες με την διαφορά ότι τα στοιχεία του πρώτου έχουν τα URI, μίας εξωτερικής βάσης δεδομένων, αυτής του IPTC. Έτσι, αν κάνουμε αντιστοίχιση των λεξιλογίων, τα δικά μας URI θα συνδεθούν με αυτά του IPTC και θα έχουμε δημιουργήσει ένα δικό μας σύνολο Συνδεδεμένων Δεδομένων. Αν στο μέλλον οι όροι του IPTC, συνδεθούν με άλλους θησαυρούς θα αποκτήσουν έμμεσα και τα δικά μας URI, σύνδεση.

Ως πηγή μας θα ορίσουμε το μικρότερο από τα δύο σύνολα το iptc (με βάση τα URI του IPTC) και ως στόχο το iptcALL. Το νέο αρχείο με τις συνδέσεις ονομάζεται iptcINall και περιέχει 527 όρους δηλαδή το σύνολο του iptc, όπως ήταν αναμενόμενο αφού πρόκειται για υποσύνολο. Από το iptcALL αντιστοιχήθηκε το 69,30% . τα υπόλοιπα είναι τα δεδομένα που είχαν μεταφερθεί στο IPTCdump.



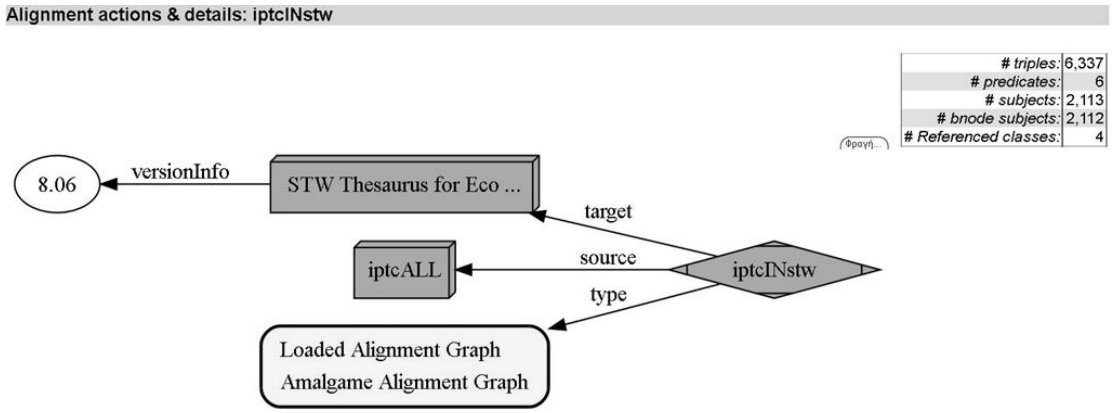
Εικόνα 5.3.2.1 Γράφος του iptcINall

Το συγκεκριμένο παράδειγμα χρήσης δεν θεωρούμε ότι χρειάζεται περεταίρω ανάλυση αφού και τα δύο λεξιλόγια είναι γνωστά μας. Μετατρέπουμε τις συνδέσεις σε skos:exactMatch και οι συνδέσεις έχουν πλέον την παρακάτω μορφή:

```
<rdf:Description rdf:about="http://cv.iptc.org/newscodes/subjectcode/07012000">
  <skos:exactMatch rdf:resource="http://www.image.ntua.gr/iptc/5095"/>
</rdf:Description>
<rdf:Description rdf:about="http://cv.iptc.org/newscodes/subjectcode/07008000">
  <skos:exactMatch rdf:resource="http://www.image.ntua.gr/iptc/5077"/>
</rdf:Description>
```

6.3 Σύνδεση με εξωτερικά σύνολα

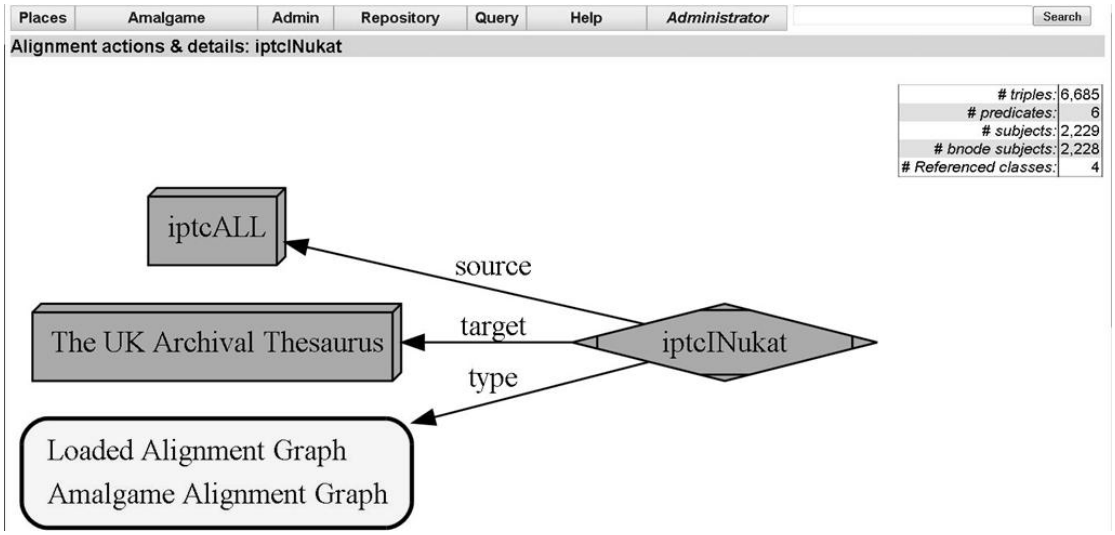
Ως τώρα εξετάσαμε συνδέσεις μεταξύ των γνωστών μας θησαυρών. Η διαδικασία γίνεται πιο σύνθετη όταν συνδέουμε μεγάλα σύνολα δεδομένων με εξωτερικούς θησαυρούς. Σε αυτήν την ενότητα, θα δοκιμάσουμε διάφορες αντιστοιχίσεις όρων, με πηγή τον δικό μας iptcALL. Να υπενθυμίσουμε εδώ ότι ο θησαυρός αυτός έχει 798 <concepts>. Θησαυροί στόχοι θα είναι μια σειρά από σύνολα δεδομένων, τα οποία υπάρχουν στον Ιστό και είναι ελεύθερη η διανομή τους με Creative Commons. Επιλέξαμε στόχους από διάφορες θεματολογίες ώστε να δούμε διάφορα θέματα που προκύπτουν σε κάθε περίπτωση. Επειδή δημιουργήθηκαν επτά αντιστοιχίσεις, θα τις περιγράψουμε σύντομα με εικόνες των γράφων τους και ένα πίνακα με τα στοιχεία που περιέχουν:



Εικόνα 5.3.2.1 Γράφος του iptcINstw

Στόχος είναι ο STW Thesaurus for Economics ένας θησαυρός με όρους οικονομικούς στα αγγλικά και γερμανικά. Είναι συνδεδεμένος με την DBpedia, οπότε έμμεσα οι όροι που θα συνδεθούν θα έχουν δεσμό με την βάση της.

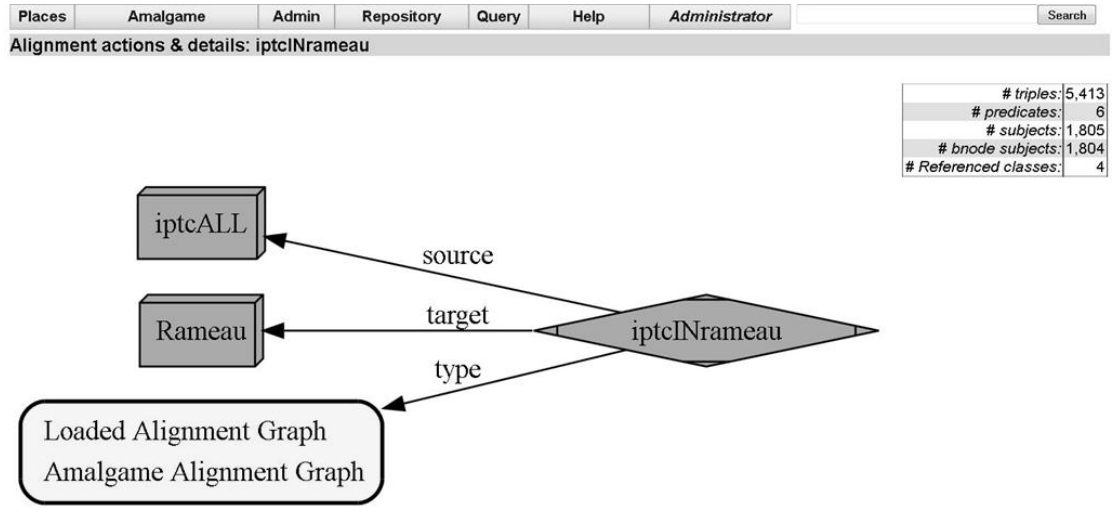
Αποτέλεσμα: 322 όροι του iptcALL συνδέθηκαν με το STW, δημιουργώντας 2112 συνδέσμους και 6337 τριάδες RDF.



Εικόνα 5.3.2.2 Γράφος του iptcINukat

Στόχος ο είναι ο UKAT UK Archival Thesaurus ένας θεματικός θησαυρός που δημιουργήθηκε για την υποστήριξη και την αρχειοθέτηση του τομέα αρχείων στην Αγγλία.

Αποτέλεσμα: 429 όροι του iptcALL συνδέθηκαν με το UKAT, δημιουργώντας 2228 συνδέσμους και 6685 τριάδες RDF.

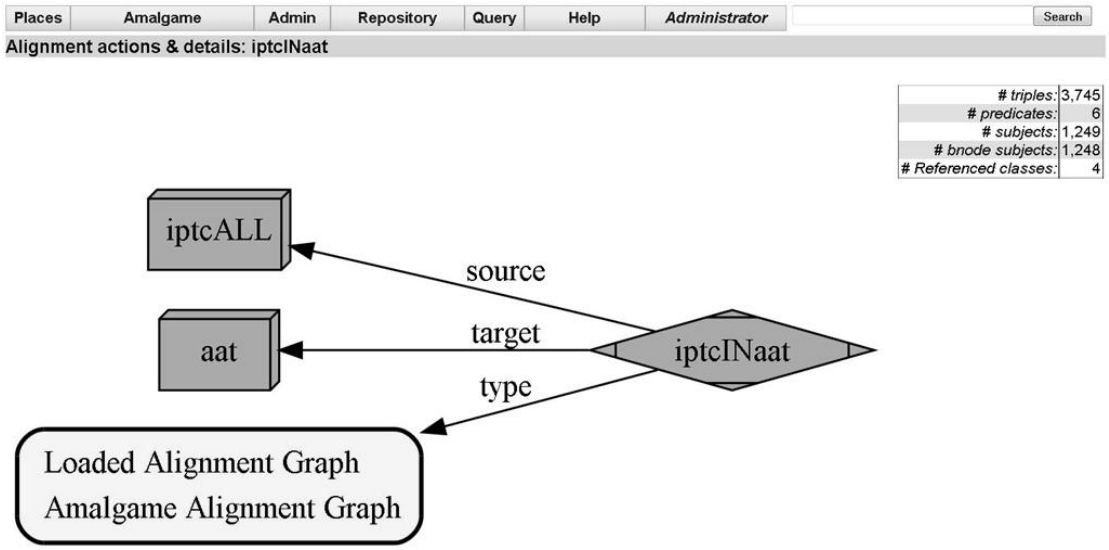


Εικόνα 5.3.2.3 Γράφος του iptcINrameau

Στόχος είναι ο RAMEAU French National Library's subject headings ένας θεματικός θησαυρός που δημιουργήθηκε για την υποστήριξη και την αρχειοθέτηση της εθνικής βιβλιοθήκης της Γαλλίας. Βρίσκεται στα Ανοιχτά Συνδεδεμένα Δεδομένα και

συνδέετε με τον LCSH και τον SWD, των αντίστοιχων θησαυρών για τις βιβλιοθήκες της Γερμανίας και των ΗΠΑ.

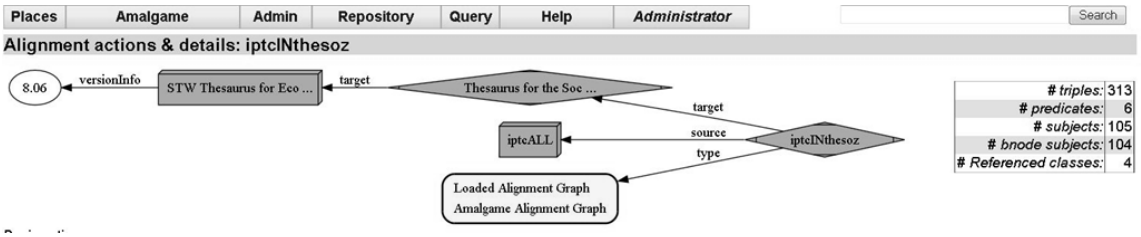
Αποτέλεσμα: 459 όροι του iptcALL συνδέθηκαν με το RAMEAU, δημιουργώντας 1804 συνδέσμους και 5413 τριάδες RDF.



Εικόνα 5.3.2.4 Γράφος του iptcINaat

Στόχος ο είναι ο AAT The Art & Architecture Thesaurus ένας θησαυρός με όρους σχετικούς με την τέχνη και την αρχιτεκτονική.

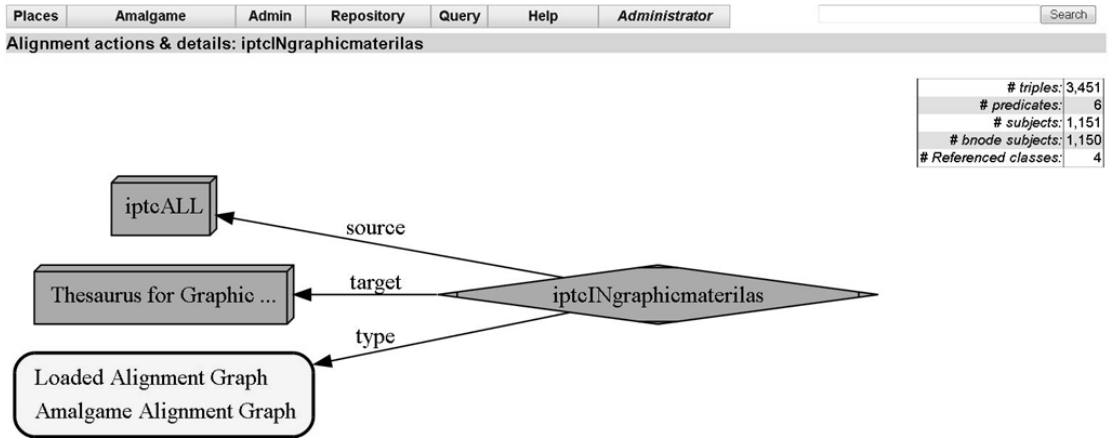
Αποτέλεσμα: 294 όροι του iptcALL συνδέθηκαν με το AAT, δημιουργώντας 1248 συνδέσμους και 3745 τριάδες RDF.



Εικόνα 5.3.2.5 Γράφος του iptcINthesoz

Στόχος ο είναι ο TheSoz Thesaurus for the Social Sciences ένας θησαυρός με περιεχόμενο από τις πολιτικές επιστήμες και οι όροι του είναι στα γερμανικά και στα αγγλικά.

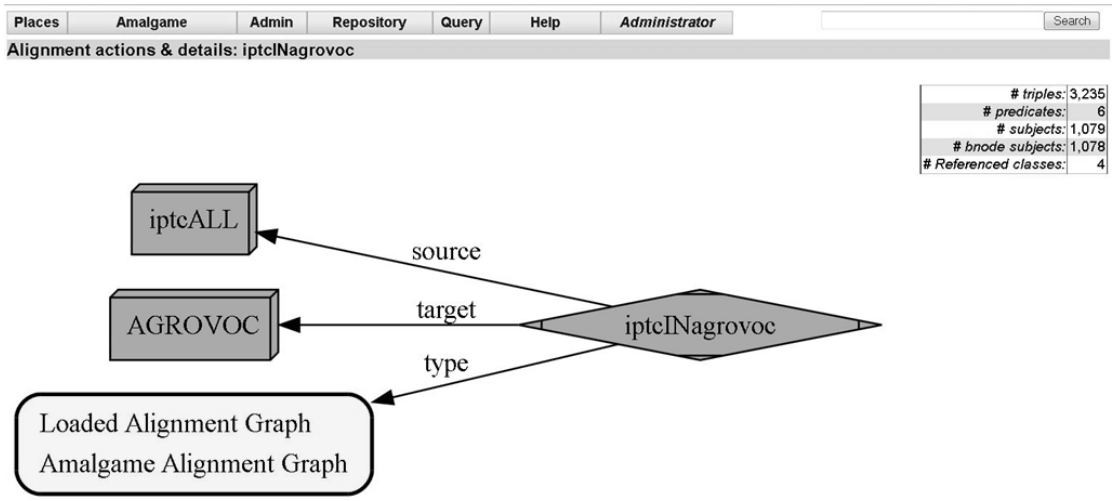
Αποτέλεσμα: 14 όροι του iptcALL συνδέθηκαν με το TheSoz, δημιουργώντας 104 συνδέσμους και 313 τριάδες RDF.



Εικόνα 5.3.2.6 Γράφος του iptcINgraphicmaterials

Στόχος ο είναι ο TGM Thesaurus for Graphic Materials της βιβλιοθήκης του κογκρέσου των ΗΠΑ. Περιέχει όρους για την αρχειοθέτηση φωτογραφιών, πινάκων και άλλων οπτικών υλικών. Είναι συνδεδεμένος με αυτόν του RAMEAU.

Αποτέλεσμα: 218 όροι του iptcALL συνδέθηκαν με το TGM, δημιουργώντας 1150 συνδέσμους και 3451 τριάδες RDF.



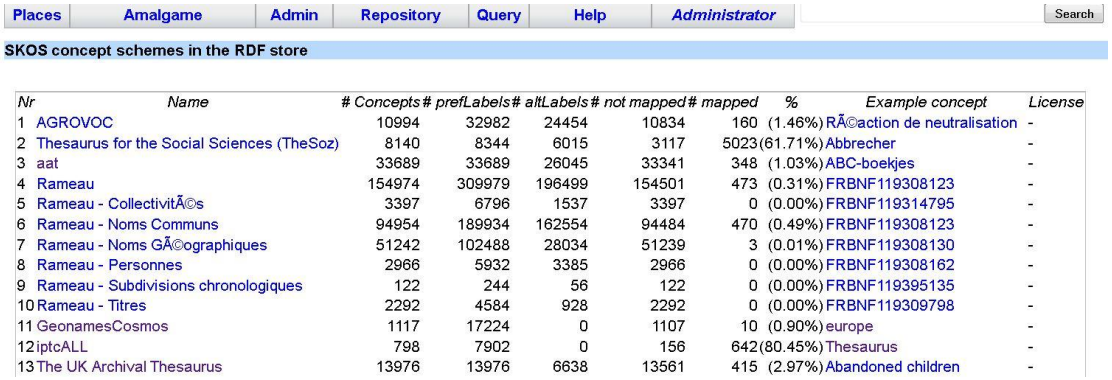
Εικόνα 5.3.2.7 Γράφος του iptcINagrovoc

Στόχος ο είναι ο AGROVOC Agricultural Thesaurus. Πρόκειται για τον μεγαλύτερο αγροτικό θησαυρό που περιέχει περίπου 20 γλώσσες. Είναι συνδεδεμένος με τους εξής θησαυρούς: CAT, NAL, SWD, GEMET.

Αποτέλεσμα: 158 όροι του iptcALL συνδέθηκαν με το AGROVOC, δημιουργώντας 1078 συνδέσμους και 3235 τριάδες RDF.

Τις παραπάνω αντιστοιχίσεις τις πραγματοποιήσαμε κάνοντας ταίριασμα μεταξύ των prefLabels της πηγής και των alt &prefLabels του στόχου. Όπου ήταν διαθέσιμες και άλλες γλώσσες η αντιστοίχιση δημιουργήθηκε μεταξύ όλων των γλωσσικών ετικετών,

αφού όπως είδαμε αυξάνεται η αντιστοίχιση αλλά και η εγκυρότητα των αποτελεσμάτων.



Nr	Name	# Concepts	# prefLabels	# altLabels	# not mapped	# mapped	%	Example concept	License
1	AGROVOC	10994	32982	24454	10834	160	(1.46%)	RÃ©action de neutralisation	-
2	Thesaurus for the Social Sciences (TheSoz)	8140	8344	6015	3117	5023	(61.71%)	Abbrecher	-
3	aat	33689	33689	26045	33341	348	(1.03%)	ABC-boekjes	-
4	Rameau	154974	309979	196499	154501	473	(0.31%)	FRBNF119308123	-
5	Rameau - CollectivitÃ©s	3397	6796	1537	3397	0	(0.00%)	FRBNF119314795	-
6	Rameau - Noms Communs	94954	189934	162554	94484	470	(0.49%)	FRBNF119308123	-
7	Rameau - Noms GÃ©ographiques	51242	102488	28034	51239	3	(0.01%)	FRBNF119308130	-
8	Rameau - Personnes	2966	5932	3385	2966	0	(0.00%)	FRBNF119308162	-
9	Rameau - Subdivisions chronologiques	122	244	56	122	0	(0.00%)	FRBNF119395135	-
10	Rameau - Titres	2292	4584	928	2292	0	(0.00%)	FRBNF119309798	-
11	GeonamesCosmos	1117	17224	0	1107	10	(0.90%)	europe	-
12	iptcALL	798	7902	0	156	642	(80.45%)	Thesaurus	-
13	The UK Archival Thesaurus	13976	13976	6638	13561	415	(2.97%)	Abandoned children	-

Εικόνα 5.3.2.8 Amalgame τελικά στατιστικά λεξιλογίων

Στην εικόνα 6.3.8 βλέπουμε ότι το 80,45%, δηλαδή 642 όροι του iptcALL, συνδέθηκαν με τους υπόλοιπους θησαυρούς. Δημιουργήσαμε ένα μικρό σύνολο από Συνδεδεμένα Δεδομένα. Οι συνδέσεις βρίσκονται στην μορφή παρουσίασης που χρησιμοποιεί το Amalgame:

```
<align:map>
  <align:Cell>
    <align:entity1 rdf:resource="http://www.image.ntua.gr/iptc/5311"/>
    <align:entity2 rdf:resource="http://zbw.eu/stw/descriptor/16298-4"/>
    <amalgame:evidence>
      <rdf:Description
        amalgame:method="exact_label"/>
    </amalgame:evidence>
  </align:Cell>
</align:map>
```

Για να εξεταστεί η εγκυρότητα των αποτελεσμάτων πήραμε ένα τυχαίο δείγμα, από κάθε αντιστοίχιση, με 10 συνδέσεις, τις αξιολογήσουμε μέσω του εργαλείου αξιολόγησης και τα αποτελέσματα τα παρουσιάζουμε στον παρακάτω πίνακα:

Πίνακας 6-1 Αξιολόγηση αντιστοιχίσεων μέσω δειγμάτων τους

Όνομα Γράφου	Exact match	Close match	Not related	I am not sure	broader	narrower	Related match
iptcINagrovoc	7	2				1	
iptcINgraphicmaterials	8				1		1
iptcINthesoz	9					1	
iptcINaat	10						
iptcINrameau	9					1	
iptcINukat	10						
iptcINstw	10						

Τα αποτελέσματα που είχαμε, ήταν παραπάνω από ικανοποιητικά και για να το επιβεβαιώσουμε επιλέξαμε ξανά ένα τυχαίο σύνολο του iptcINukat αυτή τη φορά με 30 διαφορετικές συνδέσεις:

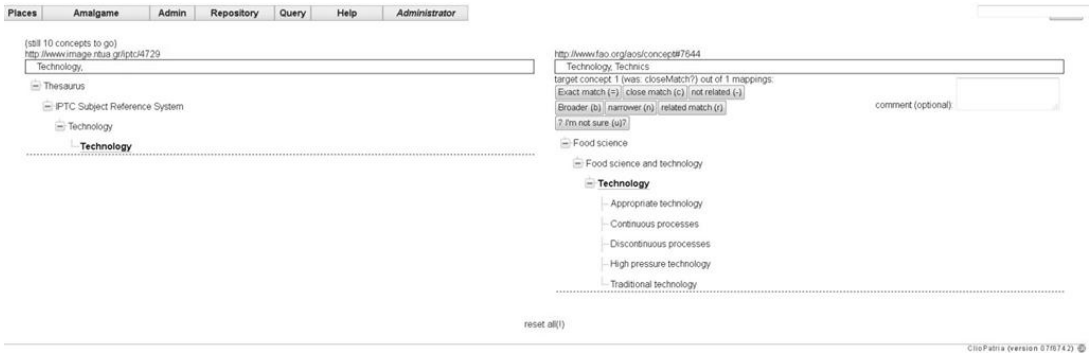
Πίνακας 6-2 Αξιολόγηση αντιστοίχισης του iptcINukat μέσω μεγαλύτερου δείγματος.

Όνομα Γράφου	Exact match	Close match	Not related	I am not sure	broader	narrower	Related match
iptcINukat 2	28	2					

Έτσι στα περισσότερα σύνολα που ευθυγραμμίσαμε θα μπορούμε να θεωρήσουμε την σύνδεση skos:exactMatch και να κάνουμε εξαγωγή των συνόλων αυτών στην μορφή:

```
…
<rdf:Description rdf:about="http://www.image.ntua.gr/iptc/4890">
  <skos:exactMatch rdf:resource="http://www.ukat.org.uk/thesaurus/concept/7714"/>
</rdf:Description>
<rdf:Description rdf:about="http://www.image.ntua.gr/iptc/5419">
  <skos:exactMatch rdf:resource="http://www.ukat.org.uk/thesaurus/concept/5479"/>
</rdf:Description>…
```

Κάνοντας την παραπάνω ανάλυση σε θησαυρούς όπως ο Rameau, που έχει τους όρους μόνο στην γαλλική γλώσσα, βλέπουμε το πλεονέκτημα της χρήσης πολύγλωσσων θησαυρών και την αντιστοίχιση να γίνεται μεταξύ όλων των γλωσσικών ετικετών. Όταν όμως η θεματολογία των θησαυρών είναι διαφορετική τότε ερχόμαστε αντιμέτωποι με το παρακάτω πρόβλημα:



Εικόνα 5.3.2.9 Amalgame εργαλείο ελέγχου βλέπουμε τον όρο «Technology»

Στο λεξιλόγιο του iptcALL ο όρος «Technology» περιγράφει γενικά τον χώρο της τεχνολογίας. Αντίθετα στον AGROVOC που είναι αγροτικός θησαυρός ο όρος με το ίδιο όνομα «Technology» αφορά μόνο την τεχνολογία που είναι σχετική με την επιστήμη της διατροφής. Ενώ η ονομασία τους ταυτίζεται, ο ένας όρος είναι υποσύνολο του άλλου και δεν θα συνδέονται μέσω του skos:exactMatch, όπως θεωρεί το πρόγραμμα, αφού αυτή η ετικέτα απαιτεί απόλυτη εννοιολογική ταύτιση μεταξύ των όρων που ενώνει. Αντίθετα, ένας ανθρώπινος χειριστής, βλέποντας την ιεραρχία που μας παρουσιάζει το εργαλείο αξιολόγησης του Amalgame, αντιλαμβάνεται την διαφορά και θα συνδέσει τους δύο όρους είτε με το skos:closeMatch είτε με το skos:narrower. Εδώ ακριβώς εντοπίζεται η αδυναμία όλων των προγραμμάτων αυτόματης δημιουργίας συνδέσεων, αφού για να αντιληφθεί ένα πρόγραμμα την εννοιολογική διάφορα που είδαμε, απαιτούνται πολύπλοκοι αλγόριθμοι.

6.4 Τελικό συμπέρασμα

Η δημιουργία ενός συνόλου Συνδεδεμένων Δεδομένων δεν είναι απλή διαδικασία. Ο χειριστής, πρέπει να γνωρίζει καλά το περιεχόμενο των θησαυρών που θα συνδέσει αλλά και το εργαλείο που θα χρησιμοποιήσει.

Αν οι θησαυροί έχουν κοινή θεματολογία είναι ποιο πιθανό να σχηματιστούν περισσότερες συνδέσεις και επίσης, αν ο στόχος μας έχει πολλούς όρους, τότε τα ποσοστά συνδέσεων στην πηγή θα είναι περισσότερα. Ένα εργαλείο όπως το

Amalgame, που παρέχει αρκετές επιλογές αλλά και μια διαφάνεια στην όλη διαδικασία, διευκολύνει αρκετά την σύνδεση μεταξύ των συνόλων. Παρατηρήθηκε ότι πολυγλωσσικοί θησαυροί δημιουργούν περισσότερες και ποιοτικότερες συνδέσεις, αφού η αντιστοίχιση γίνεται μεταξύ πολλαπλών πηγών. Τέλος, είδαμε ότι η επιτυχία των αντιστοιχίσεων εξαρτάται από το συγκεκριμένο λεξιλόγιο και τους όρους που χρησιμοποιεί, καθώς αυτοί ποικίλουν. Όμως, είναι αρκετά πιθανό να πετύχουμε αντιστοίχιση 50 έως 80% των όρων μας, με άλλα λεξιλόγια χωρίς την προηγμένη επεξεργασία της φυσικής γλώσσας, αλλά με τη χρήση του περιεχομένου και των ετικετών των όρων μας.

7

Πλατφόρμες και προγραμματιστικά εργαλεία

7.1 Ruby

Η γλώσσα που επιλέχθηκε, όπου χρειαζόταν δημιουργία κώδικα, ήταν η Ruby. Η Ruby είναι μια απλή και δυναμική γλώσσα. Ο προγραμματισμός που απαιτεί είναι πιο εύκολος, πιο γρήγορος και πιο απλός από άλλες γλώσσες και ο κώδικας της είναι μικρός σε έκταση. Έτσι δίνει ευελιξία και διευκολύνει σημαντικά τον προγραμματιστή. Θεωρείται η πιο δυνατή δυναμική metaprogramming γλώσσα. Με την ruby μπορείς να κάνεις ότι και με τις άλλες γλώσσες, με την διαφορά ότι με τη ruby δίνεται έμφαση στο πόσο εύκολο είναι να γράψεις ένα πρόγραμμα και πόσο ευανάγνωστο είναι. Η Ruby είναι αντικειμενοστραφής: κάθε τύπος δεδομένων είναι αντικείμενο, συμπεριλαμβανομένων και των κλάσεων και των τύπων που άλλες γλώσσες θεωρούν βασικούς. Κάθε συνάρτηση είναι μια μέθοδος. Οι τιμές με όνομα (μεταβλητές) πάντα είναι αναφορές σε αντικείμενα, και όχι τα ίδια τα αντικείμενα.

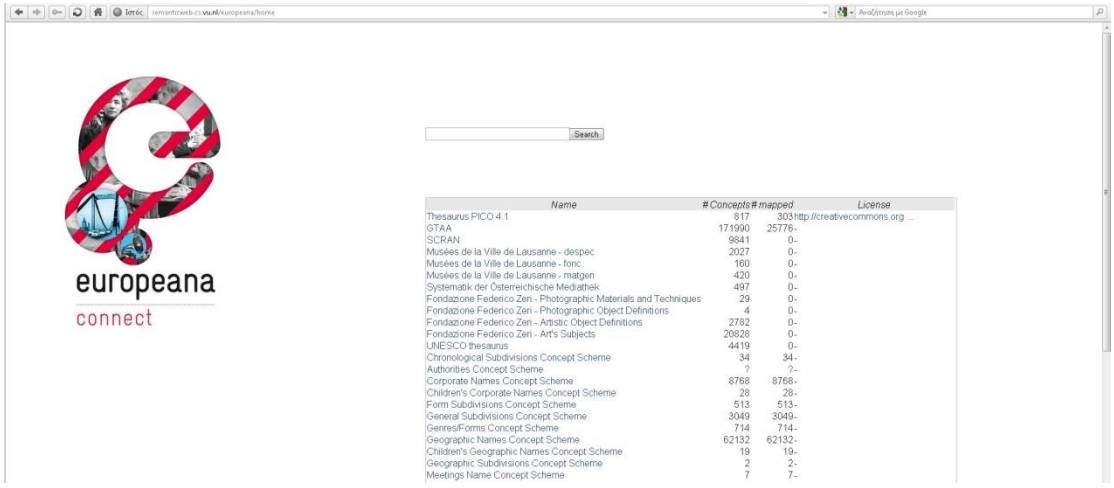
Σημαντικό στοιχείο της Ruby,είναι η πολύ εύκολη επεκτασιμότητα της με επιπλέον βιβλιοθήκες, μέσω του διαχειριστής πακέτων RubyGems. Η εγκατάσταση τους είναι απλή διαδικασία αφού δεν απαιτεί παρέμβαση του. Στην διπλωματική αυτή χρησιμοποιήσαμε τις βιβλιοθήκες REXML και open-uri. Η πρώτη είναι ένας ισχυρός XML επεξεργαστής που μας επιτρέπει να δημιουργούμε και εντολές με το Xpath ενώ η δεύτερη, δίνει την δυνατότητα να ανοίγουμε και να διαβάζουμε απευθείας τα http/https/ftp URL σαν απλό αρχείο. Λεπτομέρειες για τον τρόπο λειτουργίας κάθε κώδικα παρουσιάστηκαν στην αντίστοιχη περιοχή που τον χρησιμοποιήσαμε.

7.2 Amalgame

Το Amalgame (AMsterdam ALignment GenerAtion MEtatool) (49) είναι μια εφαρμογή που στηρίχθηκε στο περιβάλλον της ClioPatria (50), που χρησιμοποιεί SWI-Prolog και σχεδιάστηκε στο πλαίσιο των έργων του PrestoPrime και του EuropeanaConnect. Παρουσιάστηκε τον Φεβρουάριο τους 2011 από τους M. Hildebrand (VUA), J. van Ossenbruggen (VUA),V. de Boer (VUA). Πρόκειται για ένα εργαλείο με στόχο την εύρεση, την αξιολόγηση και τη διαχείριση συνδέσεων μεταξύ λεξιλογίων. Συνδυάζει υπάρχουσες τεχνικές και μεθόδους αντιστοίχισης όπως αυτές που αναπτύχθηκαν στο πλαίσιο της Ontology Alignment Evaluation Initiative (OAEI), στην οποία διαφορετικές μεθόδους μπορούν να συνδυαστούν με τη χρήση ροής εργασιών εγκατάστασης. Ο εξυπηρετητής του Amalgame διαθέτει:

- Μια λειτουργία ώστε να γίνεται σύνθεση κατά τη ροή εργασίας, όπου διάφορες γεννήτριες συσχετίσεων μπορούν να χρησιμοποιηθούν. Τα σύνολα χαρτογράφησης που προκύπτουν μπορούν να χρησιμοποιηθούν σαν είσοδο για μεθόδους φιλτραρίσματος, για άλλες γεννήτριες συσχετίσεων ή να συνδυαστούν σε επικαλυπτόμενα σύνολα.
- Μια λειτουργία στατιστικών, όπου θα εμφανίζονται οι στατιστικές για τα συσχετιζόμενα σύνολα .
- Ένα εργαλείο αξιολόγησης, όπου υποσύνολα των συσχετίσεων μπορούν να αξιολογηθούν με το χέρι.

Ένα από τα βασικά συστατικά του Amalgame είναι η διεπαφή αξιολόγησης. Ο χρήστης μπορεί να την χρησιμοποιήσει για να αναλύσει γρήγορα και αποτελεσματικά αλλά και να αξιολογήσει τις παραγόμενες συσχετίσεις.



Εικόνα 5.3.2.1 Κεντρική σελίδα του Amalgame

Στην διπλωματική αυτή, η τελική σύνδεση των θησαυρών μας, έγινε με την χρήση του εργαλείου αυτού. Το Amalgame είναι ένα πολύ πρόσφατο εργαλείο που ακόμα αναπτύσσεται. Η περιορισμένη δοκιμή του από πραγματικούς χρήστες, μας οδήγησε στο να παρουσιάσουμε αναλυτικά την χρήση που του κάναμε, στο κεφάλαιο 6, αλλά και μιας σύνοψης των δυνατοτήτων που μας δίνει αλλά και των προβλημάτων που αντιμετωπίσαμε.

7.2.1 Αδυναμίες Amalgame

Οι δημιουργοί του Amalgame στόχευαν αρχικά σε θησαυρούς πολιτιστικής κληρονομιάς, οι οποίοι συνήθως περιλαμβάνουν από 1000 έως και 100000 όρους. Όταν προσπαθήσαμε να εισάγουμε κάποιον θησαυρό μεγαλύτερου μεγέθους (π.χ. το descriptors των NYT ή το wordnet-wordsensesandwords), είτε ο χρόνος φόρτωσης τους ήταν μεγάλος είτε δεν φορτωνόταν καθόλου. Επίσης, σε μεγάλα μεγέθη πληροφορίας, παρατηρήσαμε αύξηση της απαιτούμενης μνήμης που χρησιμοποιεί το πρόγραμμα που το τρέχει. Αν μοναδική χρήση που θα γίνει στο εργαλείου είναι η σύνδεση πολιτιστικών θησαυρών δεν θα υπάρξει πρόβλημα, αν όμως θέλουμε να συνδέσουμε τα δεδομένα μας με τον Ιστό των Συνδεδεμένων Δεδομένων και με μεγάλα θεμελιακά σύνολα του, όπως το wordnet και την dbpedia, ο χειριστής πρέπει να καταφύγει σε άλλες λύσεις αφού τα σύνολο αυτά είναι αδύνατο να φορτωθούν στο Amalgame.

Επίσης το Amalgame είναι επιλεκτικό στο είδος των θησαυρών που μπορεί να αναγνωρίσει και να διαχειριστεί. Κοινό χαρακτηριστικό όσων θησαυρών φορτώσαμε είναι ότι βρίσκονταν σε συγκεκριμένη μορφή SKOS. Αν μάλιστα, σε ορισμένες δοκιμές, έλειπε η αρχική ετικέτα του <skos:ConceptScheme>, το εργαλείο δεν εμφάνιζε στην βιβλιοθήκη του τον θησαυρό μας, ώστε να μπορέσουμε να τον αντιστοιχίσουμε με άλλους. Αντίθετα, εμφανιζόταν στην ενότητα «Named graphs in the RDF store» σαν απλές τριάδες που δεν μπορούσαμε να χρησιμοποιήσουμε. Το ίδιο συνέβαινε και στις περιπτώσεις που δεν αναγνωριζόταν η δομή του αρχείου. Για παράδειγμα το iptcCodes, που είναι η επίσημο μεταφορά του ενός υποσύνολου του IPTC σε SKOS, δεν καταφέραμε να το διαβάσει. Εδώ δηλαδή ξανά βλέπουμε, μία αδυναμία που εμποδίζει το Amalgame, να γίνει ελκυστικό για ποιο ευρεία χρήση και να ξεφύγει από τα πλαίσια των πολιτιστικών θησαυρών.

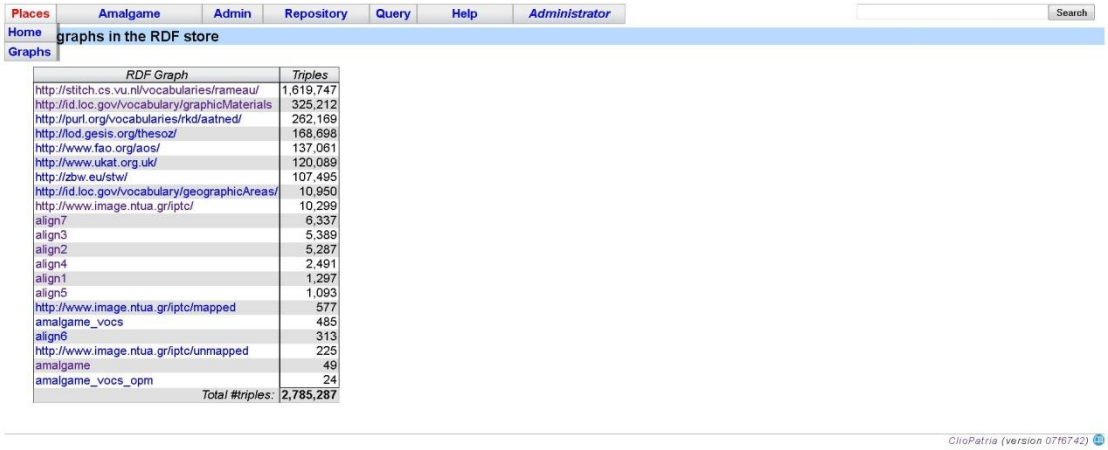
Μία τεχνική λεπτομέρεια που μας δημιούργησε πρόβλημα ήταν ότι, αν κάναμε επανεκκίνηση του εξυπηρετητή που φιλοξενούσε το Amalgame, η πληροφορία μας

δεν αποθηκευόταν, δεν αναγνώριζε το GIT, που απαιτείτε για τις συγκρίσεις και για αυτό χρειαζόταν εγκατάσταση του, από την αρχή.

Τέλος, η διαδικασία, της αντιστοίχισης, δεν είναι αυτοματοποιημένη. Η επιλογή, η σύνθεση και ο συνδυασμός των εργαλείων είναι υπό την ευθύνη του χειριστή. Ο χειριστής πρέπει, να έχει γνώση της διαδικασίας που θα ακολουθήσει και τι θέλει να πετύχει, ώστε να γίνουν τα απαραίτητα βήματα. Με λίγα λόγια, το Amalgame δεν είναι ένα εύκολο εργαλείο για κάποιον χωρίς γνώσεις, που θέλει αυτόματα να μετατρέψει τους θησαυρούς του.

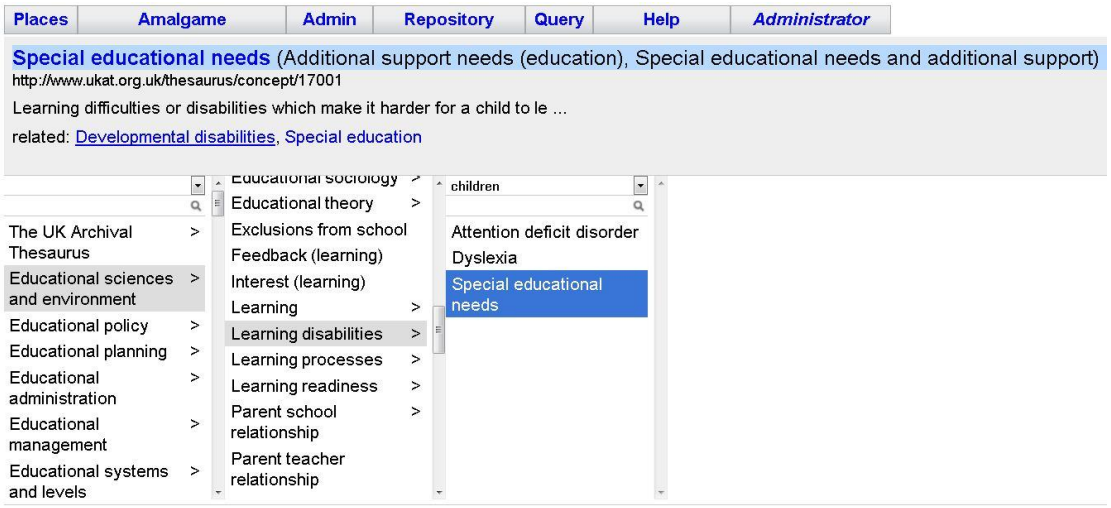
7.2.2 Δυνατότητες Amalgame

Ανεξάρτητα από τις αδυναμίες του, το Amalgame, είναι ένα αρκετά χρήσιμο εργαλείο. Αν τα σύνολα που θέλουμε να συνδέσουμε δεν ανήκουν στις παραπάνω κατηγορίες, που δημιουργούν πρόβλημα, τότε το αποτέλεσμα είναι ικανοποιητικό και γρήγορο. Η μη αυτοματοποιημένη διαδικασία αντιστοίχισης αναφέρθηκε ως αδυναμία, όμως γίνεται πλεονέκτημα αν ο χειρίστης γνωρίζει καλά τους προς σύγκριση θησαυρούς και εξοικειωθεί με τις διαδικασίες που του παρέχει. Όμως, η ροής εργασίας για μια αντιστοίχιση παρέχει διαφάνεια στην όλη διαδικασία. Το Amalgame σχεδιάστηκε με την προοπτική ο χρήστης του, να κατασκευάζει διαδραστικά μια ροή εργασίας. Μια ροή περιλαμβάνει selectors, για να ορίζουμε ποιοι από τους όρους θα χρησιμοποιηθούν από τα λεξιλόγια της πηγή και του στόχου. matchers, που ορίζουν ποιες μεταβλητές θα χρησιμοποιηθούν για την αντιστοίχιση των όρων. Partitioners για να χωρίζονται τα σύνολα από τις αντιστοιχίσεις (χωρισμός σε 1-1 ή 1-n είδαμε στο 6 κεφάλαιο) και mergers για να δημιουργούμε την ένωση υποσυνόλων. Το εργαλείο ανάλυσης για την εξερεύνηση των αντιστοιχίσεων Filters, για να διαχωρίσουμε τα αποτελέσματα μιας αντιστοίχισης. Όλα τα παραπάνω εργαλεία διευκολύνουν τον χειριστή, να έχει γνώση της ανάλυσης και των αποτελεσμάτων της, αυξάνοντας την ανάκληση και την ακρίβεια αυτών. Η διεπαφή χρήστη του Amalgame, είναι σε αρχικό στάδιο ανάπτυξης αλλά δεν δημιουργεί προβλήματα αφού είναι αρκετά απλή και εύκολη στην πλοήγηση. Την όποια παρουσιάζουμε στην συνέχεια:



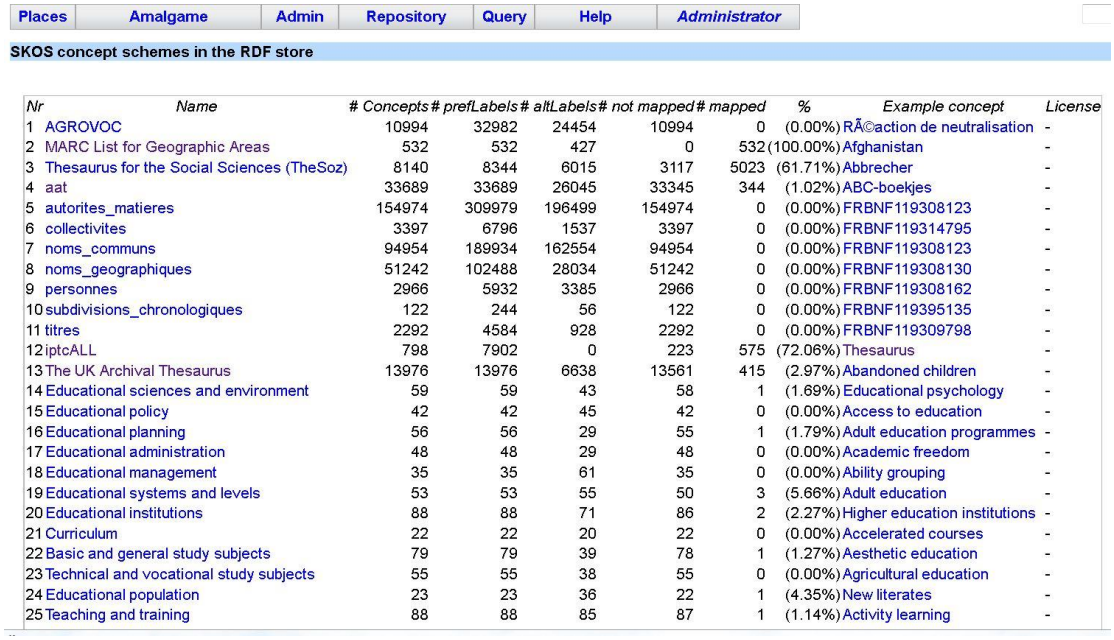
Εικόνα 7.2.2.1 Γράφοι που έχουμε φορτώσει

Εικόνα **7.2.2.1** Places →Graphs περιλαμβάνονται όλοι οι RDF γράφοι με το βασικό τους URI(αν τους εισάγαμε εμείς) ή το όνομα τους(αν είναι αποτέλεσμα μιας αντιστοίχισης) και τον αριθμό των RDF τριάδων που περιέχουν.



Εικόνα 7.2.2.2 Πλοήγηση στους θησαυρούς

Εικόνα **7.2.2.2** Amalgame →Voc browser περιλαμβάνονται όλοι θησαυροί, στους οποίους μπορούμε, μέσω της πλατφόρμα, να πλοηγηθούμε στα δεδομένα τους. Στο πάνω μέρος μας εμφανίζει τις λεπτομέρειες του όρου που επιλέξαμε.



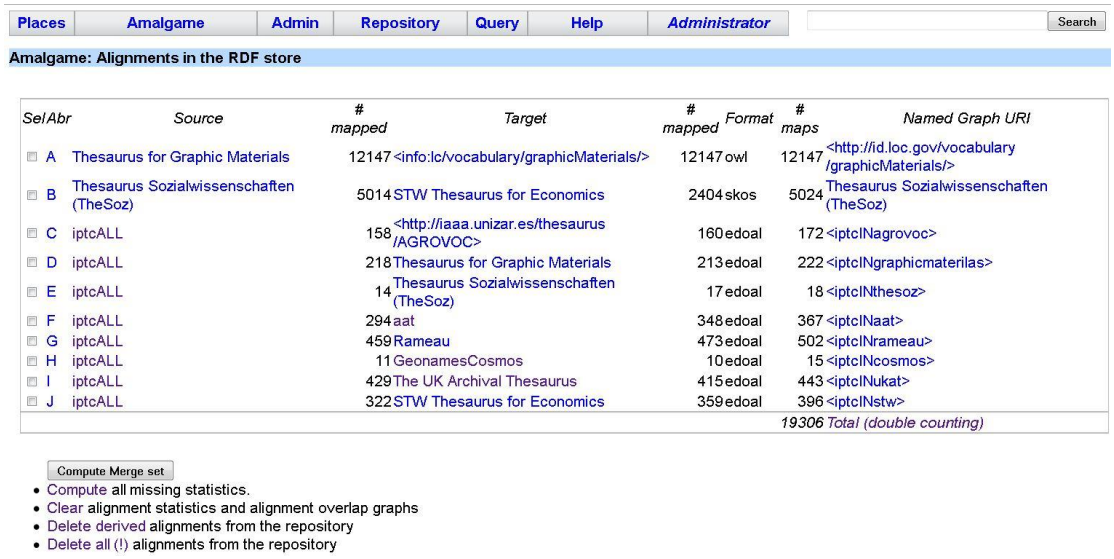
Places | Amalgame | Admin | Repository | Query | Help | Administrator

SKOS concept schemes in the RDF store

Nr	Name	# Concepts	# prefLabels	# altLabels	# not mapped	# mapped	%	Example concept	License
1	AGROVOC	10994	32982	24454	10994	0	(0.00%)	RÃ©action de neutralisation	-
2	MARC List for Geographic Areas	532	532	427	0	532	(100.00%)	Afghanistan	-
3	Thesaurus for the Social Sciences (TheSoz)	8140	8344	6015	3117	5023	(61.71%)	Abbrecher	-
4	aat	33689	33689	26045	33345	344	(1.02%)	ABC-boekjes	-
5	autorites_matieres	154974	309979	196499	154974	0	(0.00%)	FRBNF119308123	-
6	collectivites	3397	6796	1537	3397	0	(0.00%)	FRBNF119314795	-
7	noms_communs	94954	189934	162554	94954	0	(0.00%)	FRBNF119308123	-
8	noms_geographiques	51242	102488	28034	51242	0	(0.00%)	FRBNF119308130	-
9	personnes	2966	5932	3385	2966	0	(0.00%)	FRBNF119308162	-
10	subdivisions_chronologiques	122	244	56	122	0	(0.00%)	FRBNF119395135	-
11	titres	2292	4584	928	2292	0	(0.00%)	FRBNF119309798	-
12	iptcALL	798	7902	0	223	575	(72.06%)	Thesaurus	-
13	The UK Archival Thesaurus	13976	13976	6638	13561	415	(2.97%)	Abandoned children	-
14	Educational sciences and environment	59	59	43	58	1	(1.69%)	Educational psychology	-
15	Educational policy	42	42	45	42	0	(0.00%)	Access to education	-
16	Educational planning	56	56	29	55	1	(1.79%)	Adult education programmes	-
17	Educational administration	48	48	29	48	0	(0.00%)	Academic freedom	-
18	Educational management	35	35	61	35	0	(0.00%)	Ability grouping	-
19	Educational systems and levels	53	53	55	50	3	(5.66%)	Adult education	-
20	Educational institutions	88	88	71	86	2	(2.27%)	Higher education institutions	-
21	Curriculum	22	22	20	22	0	(0.00%)	Accelerated courses	-
22	Basic and general study subjects	79	79	39	78	1	(1.27%)	Aesthetic education	-
23	Technical and vocational study subjects	55	55	38	55	0	(0.00%)	Agricultural education	-
24	Educational population	23	23	36	22	1	(4.35%)	New literates	-
25	Teaching and training	88	88	85	87	1	(1.14%)	Activity learning	-

Εικόνα 7.2.2.3 Στατιστικά λεξιλογίων που είναι φορτωμένα στην βάση

Εικόνα **7.2.2.3** Amalgame →Voc stats περιλαμβάνονται όλοι θησαυροί, που έχουμε εισάγει. Εκτός από το όνομα του, έχουμε τον αριθμό των <concepts>, πόσες ετικέτες <prefLabels> και <altLabels> περιέχει και παρέχει έναν όρο δείγμα από το περιεχόμενο του. Επίσης, αφού κάνουμε αντιστοίχιση μεταξύ των όρων μας, εμφανίζοντα στοιχεία για το ποσοστό αυτών που αντιστοιχήθηκαν αλλά και τον ακριβή αριθμό τους.



Places | Amalgame | Admin | Repository | Query | Help | Administrator | Search

Amalgame: Alignments in the RDF store

Sel	Abr	Source	# mapped	Target	# mapped	Format	# maps	Named Graph URI
☐	A	Thesaurus for Graphic Materials	12147	<info:lc/vocabulary/graphicMaterials/>	12147	owl	12147	<http://id.loc.gov/vocabulary/graphicMaterials/>
☐	B	Thesaurus Sozialwissenschaften (TheSoz)	5014	STW Thesaurus for Economics	2404	skos	5024	Thesaurus Sozialwissenschaften (TheSoz)
☐	C	iptcALL	158	<http://iaaa.unizar.es/thesaurus/AGROVOC>	160	edoal	172	<iptclNagrovoc>
☐	D	iptcALL	218	Thesaurus for Graphic Materials	213	edoal	222	<iptclNgraphicmaterilas>
☐	E	iptcALL	14	Thesaurus Sozialwissenschaften (TheSoz)	17	edoal	18	<iptclNthesoz>
☐	F	iptcALL	294	aat	348	edoal	367	<iptclNaat>
☐	G	iptcALL	459	Rameau	473	edoal	502	<iptclNrameau>
☐	H	iptcALL	11	GeonamesCosmos	10	edoal	15	<iptclNcosmos>
☐	I	iptcALL	429	The UK Archival Thesaurus	415	edoal	443	<iptclNukat>
☐	J	iptcALL	322	STW Thesaurus for Economics	359	edoal	396	<iptclNstw>
							19306	Total (double counting)

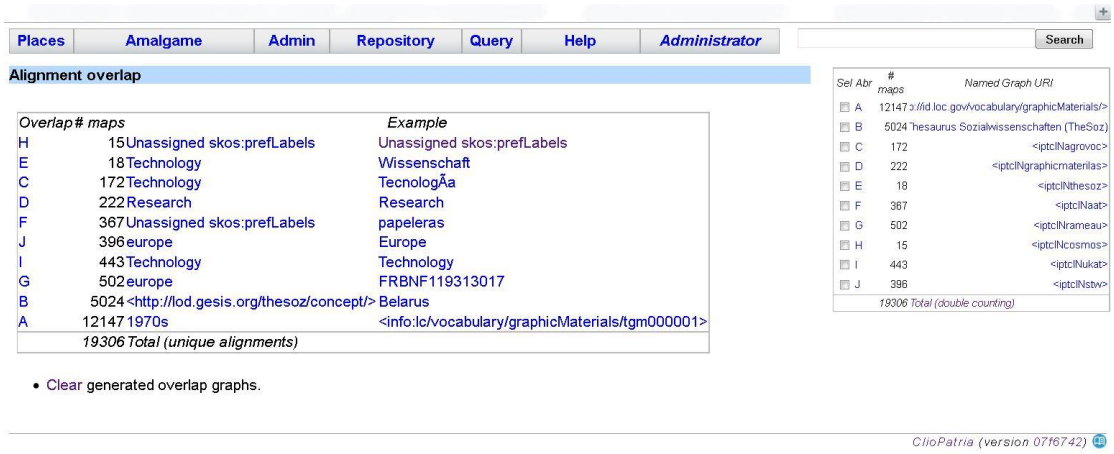
Compute Merge set

- Compute all missing statistics.
- Clear alignment statistics and alignment overlap graphs
- Delete derived alignments from the repository
- Delete all (!) alignments from the repository

Εικόνα 7.2.2.4 Στατιστικά ευθυγραμμίσεων που έχουν γίνει μεταξύ λεξιλογίων

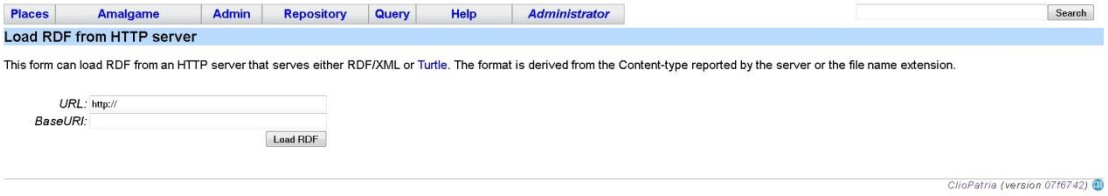
Εικόνα **7.2.2.4** Amalgame →Map stats παρουσιάζει όλες τις αντιστοιχίσεις που έγιναν με το όνομα τους, εμφανίζει την πηγή, τον στόχο, πόσα <concept>

αντιστοιχήθηκαν στο καθένα αλλά και τον συνολικό αριθμό των αντιστοιχίσεων που περιλαμβάνει ο γράφος.

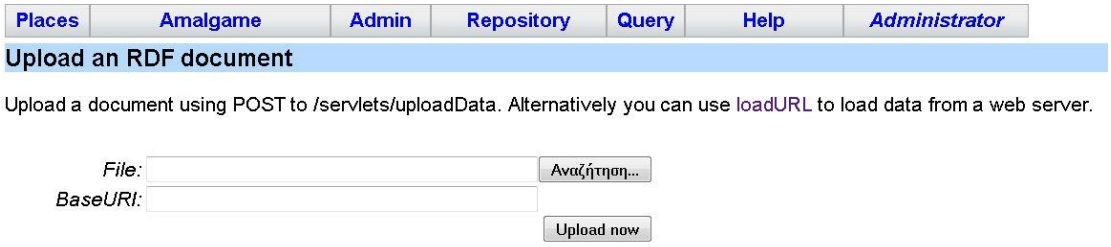


Εικόνα 7.2.2.5 Στατιστικά μοναδικών αντιστοιχίσεων

Εικόνα **7.2.2.5** Amalgame →Map overlap παρουσιάζει δεξιά όλες τις αντιστοιχίσεις που έγιναν με το όνομα τους, και υπολογίζει πόσες μοναδικές αντιστοιχίσεις έγιναν. Οι παρακάτω δείχνουν το περιβάλλον που μας επιτρέπει να ανεβάσουμε στην βιβλιοθήκη του ένα αρχείο που υπάρχει είτε τοπικά είτε στον Ιστό.



Εικόνα 7.2.2.6 Repository Load local file



Εικόνα 7.2.2.7 Repository Load from HTTP

Εικόνα 7.2.2.8 Χώρος ερωτήσεων SPARQL

Εικόνα **7.2.2.8** Query, μπορούμε να κάνουμε αναζήτηση στη βάση του Amalgame, μέσω του ερωτήσεων σε SPARQL (στο πλαίσιο της διπλωματικής δεν χρησιμοποιήσαμε αυτό το κομμάτι).

Τέλος, να αναφέρουμε ότι, οι δημιουργοί του αναπτύσσουν ένα σύνολο εξαρτημάτων, που θα ενσωματωθούν στην εφαρμογή. Γενικά θα λέγαμε ότι το Amalgame είναι ένα καλό εργαλείο, με καλή απόδοση και απλό στη χρήση, με συγκεκριμένες όμως δυνατότητες και με ειδικευμένο πεδίο εφαρμογής. Το πρόγραμμα, που χρειάζεται για την εγκατάσταση του, είναι διαθέσιμο με την μορφή ανοιχτού κώδικα.

8

Επίλογος

Στο Κεφάλαιο αυτό συνοψίζουμε την παρουσίαση της παρούσας διπλωματικής εργασίας και προτείνουμε μερικές ιδέες για μελλοντική έρευνα .

8.1 Σύνοψη και συμπεράσματα

Η διπλωματική αυτή εργασία είχε σκοπό να εξετάσει την διαδικασία μεταφοράς των πολιτιστικών θησαυρών στο χώρο του Σημασιολογικού Ιστού. Αφορμή υπήρξαν τα προβλήματα που αντιμετωπίζουν οι φορείς που τους διαχειρίζονται κατά την μετατροπή των θησαυρών σύμφωνα με τους κανόνες που απαιτεί ο Σημασιολογικός Ιστός, καθώς και κατά την σύνδεση τους με άλλα εξωτερικά σύνολα δεδομένων. Τα δύο αυτά ζητήματα εξετάστηκαν μέσα από μια σειρά παραδειγμάτων όπου εφαρμόστηκε η αντίστοιχη μεθοδολογία.

Πρώτο στάδιο ήταν η μετατροπή των θησαυρών από μορφή XML σε SKOS. Είδαμε ότι αν θέλουμε να εμπλουτίσουμε τη βάση μας με δεδομένα από μία άλλη, θα πρέπει αρχικά να βρούμε τρόπο ώστε να κάνουμε λήψη αυτών και στην συνέχεια να γίνει η μετατροπή τους. Αντιμετωπίσαμε επίσης την περίπτωση που θέλουμε να κρατήσουμε τα στοιχεία του τοπικού θησαυρού, χωρίς όμως να αναλάβουμε να τα ορίσουμε με κάποιο μοναδικό URI. Τέλος, είδαμε την ολοκληρωμένη διαδικασία μεταφοράς πλήρους θησαυρού σε SKOS. Το βασικό συμπέρασμα που έχουμε από όλη την διαδικασία είναι ότι ο διαχειριστής του θησαυρού, οφείλει να μελετήσει αναλυτικά την πληροφορία που χρειάζεται στην νέα σημασιολογική μορφή του, να εξετάσει τις τεχνικές δυνατότητες του και να καταλήξει στην μεθοδολογία που θα ακολουθήσει. Δυστυχώς είδαμε στην πράξη ότι η σημερινή, μη τυποποιημένη μορφή θησαυρών

καθιστά εξαιρετικά δύσκολη τη δημιουργία μιας εφαρμογής που θα μετατρέπει αυτόματα σε SKOS κάθε θησαυρό. Επομένως, το πρόγραμμα που αναπτύξαμε για την ολοκληρωμένη μεταφορά του θησαυρού σε SKOS, δεν μπορεί να εφαρμοστεί σε θησαυρούς που ακολουθούν διαφορετική δομή.

Στο δεύτερο στάδιο, εξετάσαμε την σύνδεση των θησαυρών μας είτε μεταξύ τους είτε με εξωτερικά σύνολα με χρήση του εργαλείου Amalgame. Παρατηρήσαμε ότι η διαδικασία δημιουργίας αντιστοιχίσεων μεταξύ των θησαυρών απαιτεί από το χειριστή να γνωρίζει πολύ καλά το περιεχόμενο του θησαυρού-πηγής, ώστε να επιλέξει κατάλληλα τον στόχο. Όσο πιο σχετικοί εννοιολογικά είναι οι όροι του στόχου με της πηγής, τόσο περισσότερες και ποιοτικότερες συνδέσεις θα έχουμε. Το ίδιο θα ισχύει, αν το εργαλείο που χρησιμοποιούμε εκμεταλλεύεται την πολυγλωσσικότητα του θησαυρού. Η διαδικασία αντιστοιχίσεως λεξιλογίων δεν μπορεί να γίνει πλήρως αυτοματοποιημένα αλλά χρειάζεται ανθρώπινη επίβλεψη των αποτελεσμάτων. Ο χώρος των Συνδεδεμένων Δεδομένων μας παρέχει μια σειρά από ετικέτες σημασιολογικής ομοιότητας, για περιγραφή της σχέσης που έχουν οι προς σύνδεση όροι. Ωστόσο, η χρήση αυτών των ετικετών απαιτεί να γίνεται με απόλυτη βεβαιότητα, γιατί διαφορετικά θα βρεθούμε με ένα σύνολο από λανθασμένες συνδέσεις. Ένας απλός αλγόριθμος αντιστοίχισης όπως του Amalgame, μας δίνει καλή ευστοχία στα αποτελέσματα αλλά η αξιολόγηση των αποτελεσμάτων από κάποιον χειριστή παραμένει αναγκαία.

8.2 Μελλοντικές επεκτάσεις

Παρόλο που έχουν αναπτυχθεί συγκεκριμένοι οδηγοί και κανόνες για την δημοσίευση δεδομένων στον Ιστό, η εφαρμογή τους συναντά αρκετά εμπόδια. Στην ενότητα αυτή, προτείνουμε μια σειρά από επεκτάσεις στην έρευνα, που θα διευκόλυναν την διαδικασία μεταφοράς των πολιτιστικών θησαυρών στο Σημασιολογικό Ιστό:

- η αναλυτική παρουσίαση της διαδικασίας μετατροπής σε SKOS, που ακολουθήθηκε σε υπάρχοντες θησαυρούς αλλά και των εργαλείων που χρησιμοποιήθηκαν για την επίτευξη της. Έτσι, οι διαχειριστές θησαυρών, που δεν έχουν την απαραίτητη πείρα στις τεχνολογίες του Σημασιολογικού Ιστού, θα μπορούν να βρουν κάποια περίπτωση που να αφορά θησαυρό, παρόμοιας δομής με τον δικό τους. Βασιζόμενοι σε αυτή, θα εξοικειωθούν γρηγορότερα με την απαιτούμενη διαδικασία.

- η ανάπτυξη και η διάδοση εργαλείων ώστε να δημοσιεύονται εύκολα στον Ιστό οι όροι, με dereferenceable URI, χωρίς να χρειάζεται ο χρήστης κάποιον επιπλέον τεχνολογικό εξοπλισμό. Αυτό θα δώσει μεγαλύτερη ανεξαρτησία στους κατόχους των θησαυρών, που δεν θα εξαρτώνται από εξωτερικά σύνολα όπως στην περίπτωση του IPTC.
- για τον Σημασιολογικό Ιστό η ύπαρξη μεγάλου όγκου δεδομένων σε μορφή SKOS δεν είναι αρκετή, αφού απαιτείται και η σημασιολογική σύνδεση τους. Στον τομέα αυτό, η έρευνα οφείλει να βοηθήσει και να εξελίξει τα εργαλεία αναζήτησης και αξιολόγησης των όμοιων πόρων. Οι αλγόριθμοι που θα χρησιμοποιούνται θα είναι πιο σύνθετοι. Η ιεραρχία, η περιγραφή και τα άλλα πεδία που περιγράφουν έναν όρο, πρέπει να ληφθούν υπόψη. Στόχος είναι η αυτοματοποίηση της διαδικασίας, με την ελαχιστοποίηση της ανθρώπινης παρέμβασης, δίνοντας όμως ποιοτικές και εννοιολογικά σωστές συνδέσεις.
- τρόπους ώστε οι συνδέσεις αυτές να είναι σταθερές στο χρόνο και να μην μεταβάλλονται τα URI των όρων τους. Όταν κάποιος διαχειριστής εκθέσει τον θησαυρό του στο σύννεφο των Συνδεδεμένων Δεδομένων και έπειτα είτε τον εξαφανιστεί από το σύννεφο είτε αλλάξει τα URI να δείχνουν άλλους πόρους, τότε θα σπάσει ένας κρίκος στην αλυσίδα και οι θησαυροί που έχουν αντιστοιχηθεί με αυτόν θα δείχνουν σε ένα κενό σύνολο ή σε μη σχετική πληροφορία. Αντιλαμβανόμαστε λοιπόν ότι είναι πολύ σημαντικό για τον Σημασιολογικό Ιστό, η σταθερότητα των URI.
- αύξηση της ζήτησης και της χρησιμοποίησης των Συνδεδεμένων Δεδομένων. Στον πολιτιστικό χώρο είδαμε ότι οργανωμένες κινήσεις, όπως της Europeana, οδήγησαν στην παραγωγή σημαντικού υλικού και εξέλιξαν την διαδικασία της μεταφοράς ευρωπαϊκών θησαυρών στον χώρο του Σημασιολογικού Ιστού. Με αυτό τον τρόπο, θεωρούμε ότι η ανάπτυξη νέων προγραμμάτων, η αυτοματοποίηση της διαδικασίας και η αύξηση των εφαρμογών που θα καταναλώνουν πληροφορία αυτής της μορφής, θα οδηγήσουν στην αύξηση του ενδιαφέροντος από μικρότερους χρήστες που έχουν στην διάθεση τους θησαυρούς.

9

Βιβλιογραφία

1. Tim Berners-Lee, James Hendler and Ora Lassila. The Semantic Web. *Scientific American Magazine.* 17 May 2001.

2. Consortium, World Wide Web. [Ηλεκτρονικό] 2007. http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#%2824%29.

3. Tim Bray, Jean Paoli, C. M. Sperberg-McQueen,Eve Maler, François Yergeau. Extensible Markup Language (XML) 1.0. [Ηλεκτρονικό] 26 November 2008. http://www.w3.org/TR/REC-xml/.

4. F. Manola, E. Miller. RDF Prime. [Ηλεκτρονικό] 10 February 2004. http://www.w3.org/TR/rdf-primer/.

5. Γ. Αντωνίου, F. van Harmelen. *Εισαγωγή στο σημασιολογικό ιστό (δεύτερη αμερικανική έκδοση).* s.l. : Κλειδάριθμος, 2009.

6. Beckett, Dave. RDF/XML Syntax Specification (Revised). [Ηλεκτρονικό] 10 February 2004. http://www.w3.org/TR/REC-rdf-syntax/.

7. Berners-Lee, T. Notation 3. [Ηλεκτρονικό] 3 September 2006. http://www.w3.org/DesignIssues/Notation3.html.

8. D. Beckett, T. Berners-Lee. Turtle - Terse RDF Triple Language. [Ηλεκτρονικό] 14 January 2008. http://www.w3.org/TeamSubmission/turtle/.

9. J. Grant, D. Beckett. RDF TestCases. [Ηλεκτρονικό] 10 February 2004. http://www.w3.org/TR/rdf-testcases.

10. Frank Manola, Eric Miller. Defining RDF Vocabularies: RDF Schema. [Ηλεκτρονικό] 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/#rdfschema.

11. Christian Bizer, Tom Heath, Tim Berners-Lee. Linked Data: Principles and State of the Art. [Ηλεκτρονικό] 23 April 2008. http://www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdf.

12. C. Bizer, T. Heath, T. Berners-Lee,. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS).* 2009.

13. Berners-Lee, T. Linked Data. [Ηλεκτρονικό] 27 July 2006. http://www.w3.org/DesignIssues/LinkedData.html.

14. Berners, Tim. [Ηλεκτρονικό] February 2009. http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html.

15. C. Bizer, R. Cyganiak, T. Heath. How to publish Linked Data on the Web. [Ηλεκτρονικό] 2007. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.

16. I. Jacobs, N. Walsh. Architecture of the World Wide Web. [Ηλεκτρονικό] 2004. http://www.w3.org/TR/webarch/.

17. Lewis, Rhys. Dereferencing HTTP URIs . [Ηλεκτρονικό] 31 May 2007. http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14.

18. L. Sauermann, R. Cyganiak. Cool URIs for the Semantic Web. [Ηλεκτρονικό] 3 December 2009. http://www.w3.org/TR/cooluris/.

19. LinkingOpenData. [Ηλεκτρονικό] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

20. Linked Open Data Around the Clock. [Ηλεκτρονικό] http://latc-project.eu/.

21. PlanetData project. [Ηλεκτρονικό] http://www.planet-data.eu/.

22. Linking Open Data 2 . [Ηλεκτρονικό] http://lod2.eu/Welcome.html.

23. Documentation -- Guidelines for the establishment and development of multilingual thesauri ISO 5964:1985. [Ηλεκτρονικό] http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?ics1=01&ics2=140&ics3=20&csnumber=12159.

24. Thesauri in IT. [Ηλεκτρονικό] http://en.wikipedia.org/wiki/Thesaurus.

25. Index term. [Ηλεκτρονικό] http://en.wikipedia.org/wiki/Subject_heading.

26. About GEMET - GEneral Multilingual Environmental Thesaurus. [Ηλεκτρονικό] http://www.eionet.europa.eu/gemet/about.

27. EuroVoc, πολύγλωσσος θησαυρός της Ευρωπαϊκής Ένωσης . [Ηλεκτρονικό] http://eurovoc.europa.eu/drupal/?q=el/abouteurovoc.

28. About the TGN. [Ηλεκτρονικό] http://www.getty.edu/research/tools/vocabularies/tgn/about.html.

29. AGROVOC. [Ηλεκτρονικό] http://aims.fao.org/website/AGROVOC-Thesaurus/sub.

30. Art & Architecture Thesaurus® Online . [Ηλεκτρονικό] http://www.getty.edu/research/tools/vocabularies/aat/index.html.

31. The Friend of a Friend (FOAF) project. [Ηλεκτρονικό] http://www.foaf-project.org/.

32. Dublin Core Metadata Element Set. [Ηλεκτρονικό] http://dublincore.org/documents/dces/.

33. What is WordNet. [Ηλεκτρονικό] http://wordnet.princeton.edu/wordnet/.

34. About GeoNames. [Ηλεκτρονικό] http://www.geonames.org/about.html.

35. GeoNames Ontology. [Ηλεκτρονικό] http://www.geonames.org/ontology/documentation.html.

36. DBpedia i. [Ηλεκτρονικό] http://wiki.dbpedia.org/About.

37. DBpedia 3.6. [Ηλεκτρονικό] http://blog.dbpedia.org/2011/01/17/dbpedia-36-released/.

38. The IPTC. [Ηλεκτρονικό] http://www.iptc.org/site/Home/About/.

39. The IPTC News Exchange Format Standards. [Ηλεκτρονικό] http://www.iptc.org/site/News_Exchange_Formats/.

40. NewsCodes. [Ηλεκτρονικό] http://www.iptc.org/cms/site/index.html?channel=CH0103.

41. Europeana: think culture. [Ηλεκτρονικό] http://europeana.eu/portal/aboutus.html.

42. Strategic Plan 2011-2015. [Ηλεκτρονικό] 10 March 2011. http://version1.europeana.eu/c/document_library/get_file?uuid=c4f19464-7504-44db-ac1e-3ddb78c922d7&groupId=10602.

43. About EUscreen. [Ηλεκτρονικό] http://www.euscreen.eu/beta/about.html.

44. Antoine Isaac, Ed Summers. SKOS Simple Knowledge Organization System Primer. [Ηλεκτρονικό] 18 August 2009. http://www.w3.org/TR/skos-primer/.

45. Alistair Miles, Sean Bechhofer. SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL). [Ηλεκτρονικό] 18 Augus 2009. http://www.w3.org/TR/skos-reference/skos-xl.html.

46. Miles, Alistair. Quick Guide to Publishing a Thesaurus on the Semantic Web. [Ηλεκτρονικό] 17 May 2005. http://www.w3.org/TR/2005/WD-swbp-thesaurus-pubguide-20050517/.

47. Mark van Assem, Veronique Malais, Alistair Miles, Guus Schreiber. A Method to Convert Thesauri to SKOS. [Ηλεκτρονικό] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.9668&rep=rep1&type=pdf .

48. A. Miles, N. Rogers, D. Beckett. Migrating Thesauri to the Semantic. [Ηλεκτρονικό] 2004. http://www.w3.org/2001/sw/Europe/reports/thes/8.8/.

49. M. Hildebrand, J. van Ossenbruggen, V. de Boer. V. de Boer (VUA). [Ηλεκτρονικό] 18 February 2011. http://semanticweb.cs.vu.nl/lod/prestoprimeD421/paper.pdf .

50. G. Schreiber, A. Amin, M. van Assem and all. MultimediaN E-Culture Demonstrator. In The Semnantic Web. [Ηλεκτρονικό] 2006. http://cliopatria.swi-prolog.org/home.

10

Παραρτήματα

Παράρτημα Α : Παράμετροι του GEONAMES SEARCH WEBSERVICE

ΠΑΡΑΜΕΤΡΟΣ	ΤΙΜΗ	ΠΕΡΙΓΡΑΦΗ
q	string (q,name or name_equals required)	αναζήτηση σε όλα τα χαρακτηριστικά ενός τόπου: τοπωνύμιο, το όνομα της χώρας, την ήπειρο, admin κώδικες
name	string (q,name or name_equals required)	τοπωνύμιο μόνο
name_equals	string (q,name or name_equals required)	ακριβές όνομα τόπου
name_startsWith	string (optional)	τοπωνύμιο ξεκινά με δεδομένους χαρακτήρες
maxRows	integer (optional)	το μέγιστο αριθμό των γραμμών στο έγγραφο που

		επιστρέφονται από την υπηρεσία. Η προεπιλογή είναι 100, η μέγιστη επιτρεπόμενη τιμή είναι 1000.
startRow	integer (optional)	Χρησιμοποιείται για τα αποτελέσματα σελιδοποίησης. Αν θέλουμε να πάρουμε τα αποτελέσματα 30 έως 40, startRow = 30 και maxRows = 10. Η προεπιλογή είναι 0.
country	string : country code, ISO-3166 (optional)	Η προεπιλογή είναι όλες οι χώρες. Η παράμετρος της χώρας μπορεί να χρησιμοποιηθεί πάνω από μία φορά, π.χ.: country = FR & country = GP
countryBias	string (option)	Οι εγγραφές από το countryBias εμφανίζονται πρώτες στον κατάλογο
continentCode	string : continent code : AF,AS,EU,NA,OC,SA,AN (optional)	περιορίζει την αναζήτηση του τοπωνύμιου σε συγκεκριμένη ηπείρο.
adminCode1, adminCode2, adminCode3	string : admin code (optional)	Κωδικός διοικητικής υποδιαίρεσης
featureClass	character A,H,L,P,R,S,T,U,V (optional)	featureclass(es) (προεπιλογή = all feature classes); μπορεί να χρησιμοποιηθεί πάνω από μία φορά.
featureCode	string (optional)	featurecode(s) (προεπιλογή = all feature codes); μπορεί να

		χρησιμοποιηθεί πάνω από μία φορά.
lang	string ISO-636 2-letter language code; en,de,fr,it,es,... (optional)	προεπιλογή είναι η αγγλική
type	string xml,json,rdf	τον τύπο μορφής του επιστρεφόμενου εγγράφου, προεπιλογή = xml
style	string SHORT,MEDIUM,LONG,FULL (optional)	Μέγεθος επιστρεφόμενης πληροφορίας. προεπιλογή = MEDIUM
isNameRequired	boolean (optional)	Τουλάχιστον ένας από τους όρους αναζήτησης πρέπει να είναι μέρος της ονομασία αυτής.
tag	string (optional)	αναζήτηση για τοπωνύμια με την καθορισμένη ετικέτα
operator	string (optional)	προεπιλογή είναι 'AND', ενώ με το 'OR' δεν απαιτούνται όλοι οι όροι να καλύπτονται από την απάντηση.
charset	string (optional)	προεπιλογή = 'UTF8'
fuzzy	float (optional)	προεπιλογή = '1'